

UNIVERSITE DE NEUCHÂTEL

FACULTE DES SCIENCES

LE SOJA (Glycine max) : ETUDE DE L'ACCUMULATION DES  
PROTEINES DE RESERVE DES GRAINES - ANALYSE DU GENOME  
CHLOROPLASTIQUE

Thèse présentée à la Faculté des Sciences par

Albert Spielmann

licencié es sciences, orientation biologie expérimentale de  
l'Université de Neuchâtel

A MES PARENTS

## REMERCIEMENTS

Ce travail de thèse a été effectué dans le laboratoire de l'Institut de Botanique de l'Université de Neuchâtel sous la direction du Professeur E. Stutz.

Ces recherches ont bénéficié de l'appui financier de la Société d'Assistance Technique pour produits Nestlé S.A.

Pour le soutien constant et les nombreux conseils prodigués tout au long de ce travail, je tiens à exprimer ma plus profonde reconnaissance à M. le Prof. E. Stutz.

Je tiens à remercier tout particulièrement M. le Dr P. Schürmann pour ses nombreux conseils judicieux, ainsi que pour avoir accepté de faire partie du jury de cette thèse.

Je remercie M. le Dr C. Schlienger et M. le Dr D. Pillay pour avoir accepté de juger ce travail.

J'aimerais finalement remercier chaleureusement tous les collaborateurs et amis du laboratoire de biochimie, et particulièrement Isabelle Howald pour son excellente assistance technique et Etienne Roux pour ses nombreuses discussions enrichissantes.

A cet instant subtil où l'homme se retourne sur sa vie, Sisyphe, revenant vers son rocher, contemple cette suite d'actions sans lien qui devient son destin, créé par lui, uni sous le regard de sa mémoire et bientôt scellé par sa mort. Ainsi, persuadé de l'origine tout humaine de tout ce qui est humain, aveugle qui désire voir et qui sait que la nuit n'a pas de fin, il est toujours en marche. Le rocher roule encore. Je laisse Sisyphe au bas de la montagne ! On retrouve toujours son fardeau. Mais Sisyphe enseigne la fidélité supérieure qui nie les dieux et soulève les rochers. Lui aussi juge que tout est bien. Cet univers désormais sans maître ne lui paraît ni stérile ni futile. Chacun des grains de cette pierre, chaque éclat minéral de cette montagne pleine de nuit, à lui seul forme un monde. La lutte elle-même vers les sommets suffit à remplir un coeur d'homme. Il faut imaginer Sisyphe heureux.

Albert Camus, Le mythe de Sisyphe

## TABLE DES MATIERES

|                                                                                                                                            |    |
|--------------------------------------------------------------------------------------------------------------------------------------------|----|
| <u>INTRODUCTION GENERALE</u>                                                                                                               | 1  |
| <u>PREMIERE PARTIE : ETUDE DE L'ACCUMULATION DES PROTEINES DE<br/>RESERVE DES GRAINES DE SOJA DURANT LE<br/>DEVELOPPEMENT DE LA GRAINE</u> | 3  |
| <u>A.INTRODUCTION</u>                                                                                                                      | 3  |
| 1.Caractéristiques générales des protéines de réserve des<br>graines de Légumineuses                                                       | 3  |
| 2.Cas du soja                                                                                                                              | 5  |
| <u>B.MATERIEL ET METHODES</u>                                                                                                              | 6  |
| 1.Matériel biologique                                                                                                                      | 7  |
| 2.Extraction des protéines                                                                                                                 | 7  |
| 3.Gradients de densité de saccharose                                                                                                       | 8  |
| 4.Electrophorèse sur gel de polyacrylamide                                                                                                 | 8  |
| <u>C.RESULTATS ET DISCUSSION</u>                                                                                                           | 9  |
| 1.Evolution du poids des gousses et des graines                                                                                            | 9  |
| 2.Extraction des protéines                                                                                                                 | 13 |
| 3.Identification des protéines de réserve                                                                                                  | 13 |
| 4.Evolution quantitative des protéines au cours<br>de la maturation                                                                        | 15 |
| 5.Evolution qualitative des protéines au cours<br>de la maturation                                                                         | 17 |
| 5.1 Fraction 2 S                                                                                                                           | 17 |
| 5.2 Fraction 6 S                                                                                                                           | 19 |
| 5.3 Fraction 7 S                                                                                                                           | 22 |
| 5.4 Fraction 11 S                                                                                                                          | 22 |
| 6.Discussion                                                                                                                               | 25 |
| <u>D.CONCLUSION</u>                                                                                                                        | 28 |

|                                                                                                               |    |
|---------------------------------------------------------------------------------------------------------------|----|
| <u>DEUXIEME PARTIE : ANALYSE DE L'ADN CHLOROPLASTIQUE DE SOJA</u>                                             | 29 |
| <u>I.INTRODUCTION</u>                                                                                         | 29 |
| <u>II.ISOLATION, PURIFICATION ET CARACTERISATION PHYSICO-CHIMIQUE DE L'ADN CHLOROPLASTIQUE DE SOJA</u>        | 33 |
| <u>A.INTRODUCTION</u>                                                                                         | 33 |
| <u>B.MATERIEL ET METHODES</u>                                                                                 | 34 |
| 1.Matériel biologique                                                                                         | 34 |
| 2.Isolation de l'ADN chloroplastique                                                                          | 34 |
| 3.Détermination de la densité de flottation de l'ADNct de soja                                                | 36 |
| <u>C.RESULTATS ET DISCUSSION</u>                                                                              | 36 |
| 1.Isolation et purification de l'ADNct de soja                                                                | 36 |
| 2.Densité de flottation de l'ADNct de soja                                                                    | 37 |
| <u>III.CONSTRUCTION D'UNE CARTE PHYSIQUE ET GENETIQUE DE L'ADNct DE GLYCINE MAX</u>                           | 39 |
| <u>A.INTRODUCTION</u>                                                                                         | 39 |
| <u>B.MATERIEL ET METHODES</u>                                                                                 | 40 |
| 1.Endonucléases de restriction                                                                                | 40 |
| 2.Electrophorèse sur gel d'agarose                                                                            | 41 |
| 3.Transferts d'ADN                                                                                            | 41 |
| 4.Marquage de l'ADN par "nick-translation"                                                                    | 42 |
| 5.Hybridation ADN-ADN                                                                                         | 42 |
| 6.Electroéluion                                                                                               | 43 |
| 7.Nomenclature                                                                                                | 43 |
| <u>C.RESULTATS ET DISCUSSION</u>                                                                              | 44 |
| 1.Nombres et tailles des fragments d'ADNct après digestion avec les endonucléases Sal I, Pst I et Sma I       | 44 |
| 2.Construction d'une carte de restriction Sal I/Pst I/Sma I                                                   | 46 |
| 2.1 Evidence et localisation de deux régions répétitives inversées portant les gènes pour les ARNr 23S et 16S | 48 |

|                                                                                                                                                                                                         |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.2 Localisation du gène codant pour une protéine<br>membranaire de Mr 32'000 (photogène)-Autres évidences<br>de l'existence de 2 régions répétitives inversées                                         | 52 |
| 2.3 Localisation du gène codant pour la grande sous-unité<br>de la ribulose 1,5-bisphosphate carboxylase ( <u>rbc L</u> )                                                                               | 56 |
| 2.4 Construction d'une carte complète Sal I/Pst I/Sma I                                                                                                                                                 | 58 |
| 2.5 Localisation du gène codant pour le facteur<br>d'élongation Tu ( <u>tuf A</u> )                                                                                                                     | 62 |
| 3.Conclusions                                                                                                                                                                                           | 68 |
| IV. TECHNIQUE DU CLONAGE - CREATION D'UNE BANQUE DE CLONES<br>DE L'ADN CHLOROPLASTIQUE DE SOJA                                                                                                          | 72 |
| <u>A. INTRODUCTION</u>                                                                                                                                                                                  | 72 |
| <u>B. MATERIEL ET METHODES</u>                                                                                                                                                                          | 74 |
| 1. Matériel biologique                                                                                                                                                                                  | 74 |
| 2. Clonage                                                                                                                                                                                              | 75 |
| 3. Analyse des colonies                                                                                                                                                                                 | 75 |
| 3.1 Méthode modifiée de Grunstein et Hogness                                                                                                                                                            | 75 |
| 4. Isolement de plasmides                                                                                                                                                                               | 77 |
| <u>C. RESULTATS ET DISCUSSION</u>                                                                                                                                                                       | 77 |
| 1. Choix des enzymes de restriction pour le clonage<br>de fragments d'ADNct de soja                                                                                                                     | 77 |
| 2. Clonage de fragments d'ADNct porteurs du <u>psb A</u>                                                                                                                                                | 78 |
| 3. Création d'une banque de clones de l'ADNct de soja                                                                                                                                                   | 84 |
| 3.1 Fragments Hind III                                                                                                                                                                                  | 85 |
| 3.2 Fragments BamHI                                                                                                                                                                                     | 85 |
| 3.3 Fragments Pst I                                                                                                                                                                                     | 85 |
| 4. Conclusions                                                                                                                                                                                          | 87 |
| V. SEQUENCAGE DE LA REGION DE L'ADNct DE SOJA PORTANT LE <u>psb A</u><br>ET LE <u>trn H</u> , AINSI QUE DES REGIONS CONTENANT LES 2 EXTREMITES<br>DES REGIONS REPETITIVES INVERSEES ADJACENTES A LA LSC | 89 |
| <u>A. INTRODUCTION</u>                                                                                                                                                                                  | 89 |
| <u>B. MATERIEL ET METHODES</u>                                                                                                                                                                          | 91 |
| 1. Préparation de l'ADN pour le séquençage                                                                                                                                                              | 91 |
| 2. Techniques de séquençage                                                                                                                                                                             | 91 |

|                                                                                                                            |      |
|----------------------------------------------------------------------------------------------------------------------------|------|
| <u>C.RESULTATS</u>                                                                                                         | 96   |
| 1.Choix des enzymes de restriction                                                                                         | 96   |
| 2.Analyse de la région séquencée                                                                                           | 98   |
| 3.Analyse de la séquence du <u>psb</u> A                                                                                   | 100  |
| 4.Analyse de la séquence du <u>trn</u> H et des séquences faisant la jonction entre la LSC et les IR                       | 103  |
| <u>D.DISCUSSION</u>                                                                                                        | 109  |
| 1.Analyse de la séquence du <u>psb</u> A                                                                                   | 109  |
| 1.1 Taille du produit primaire de traduction                                                                               | 109  |
| 1.2 Comparaison de la structure primaire du photogène de <u>G.max</u> avec celles de <u>S.oleracea</u> et <u>N.debneyi</u> | 111  |
| 1.3 Analyse des séquences des régions adjacentes au <u>psb</u> A                                                           | 113  |
| 1.3.1 Analyse de la région 5'                                                                                              | 113  |
| 1.3.2 Analyse de la région 3'                                                                                              | 114  |
| 2.Analyse de la séquence se trouvant en aval du <u>psb</u> A, contenant le <u>trn</u> H et l'extrémité de l'IR adjacente   | 116  |
| 2.1 Analyse de la séquence du <u>trn</u> H                                                                                 | 116  |
| 2.2 Analyse des séquences des 2 extrémités des IR adjacentes à la LSC                                                      | 119  |
| <u>E.CONCLUSION</u>                                                                                                        | 120  |
| <u>F.RESULTATS DE DERNIERE MINUTE</u>                                                                                      | 121a |
| 1.Analyse des séquences faisant la jonction entre les IR et la LSC                                                         | 121a |
| <u>VI.DISCUSSION FINALE</u>                                                                                                | 122  |
| <u>VII.BIBLIOGRAPHIE</u>                                                                                                   | 127  |

## ABREVIATIONS, UNITES ET NOMS TRIVIAUX

|                                 |                                                                                                      |
|---------------------------------|------------------------------------------------------------------------------------------------------|
| A                               | Absorbance                                                                                           |
| ADN(DNA)                        | acide désoxyribonucléique                                                                            |
| ARN(RNA)                        | acide ribonucléique                                                                                  |
| ARNm                            | ARN messenger                                                                                        |
| ARNr                            | ARN ribosomique                                                                                      |
| ARNt                            | ARN de transfert                                                                                     |
| ATP                             | (désoxy) adénosine-5'-triphosphate                                                                   |
| <u>atp</u> B                    | gène codant pour la sous-unité $\beta$ de l'ATP-synthétase                                           |
| Bleu de<br>bromophénol          | acide $\alpha, \alpha$ -bis(3,5-dibromo-4-hydroxyphényl)- $\alpha$ -<br>hydroxy-o-toluène-sulfonique |
| bp                              | paires de bases                                                                                      |
| ct                              | chloroplastique                                                                                      |
| DCMU                            | 3-(3,4-dichlorophényl)-1,1-diméthylurée                                                              |
| ddNTP                           | didéoxyribonucléotides (mélange)                                                                     |
| dNTP                            | désoxyribonucléotides (mélange)                                                                      |
| DNase                           | désoxyribonucléases                                                                                  |
| EDTA                            | sel disodique de l'acide éthylène diamine tétra-<br>acétique                                         |
| EtBr                            | bromure d'éthidium = bromure de 3,8-diamino-6-éthyl-<br>5-phénanthridium                             |
| IR                              | régions répétitives d'un génome chloroplastique                                                      |
| kbp                             | millier de paires de bases                                                                           |
| LSC                             | grande région unique d'un génome chloroplastique                                                     |
| LSU                             | grande sous-unité de la ribulose 1,5-bisphosphate<br>carboxylase                                     |
| Md                              | millions de daltons                                                                                  |
| Mr                              | masse moléculaire relative apparente                                                                 |
| ORF                             | open reading frame = séquence de nucléotides<br>ne possédant pas de codon stop                       |
| P                               | radical phosphate                                                                                    |
| photogène                       | protéine thylacoïdale de Mr 32'000                                                                   |
| <u>psb</u> A (P <sub>32</sub> ) | gène codant pour le photogène                                                                        |

|                     |                                                                                                                                                                |
|---------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <u>rbc</u> L        | gène codant pour la grande sou-unité de la ribulose 1,5-bisphosphate carboxylase                                                                               |
| rDNA                | gène codant pour les ARNr                                                                                                                                      |
| RNase               | ribonucléase                                                                                                                                                   |
| rpm                 | tours par minute                                                                                                                                               |
| SDS                 | dodécylsulfate de sodium (laurylsulfate)                                                                                                                       |
| SDS-PAGE            | gel de polyacrylamide contenant du SDS                                                                                                                         |
| SSC                 | solution standard de citrate ( $1 \times \text{SSC} = 0.15 \text{ M NaCl}, 0.015 \text{ M citrate trisodique}$ )                                               |
| TE                  | tampon Tris-HCl 10 mM, EDTA 1 mM, pH 7.9                                                                                                                       |
| TNE                 | tampon Tris-HCl 50 mM, NaCl 100 mM, EDTA 5 mM, pH 7.9                                                                                                          |
| Tris                | Tris (hydroxyméthyl) aminométhane                                                                                                                              |
| <u>trn</u> H        | gène codant pour l'ARNt <sup>His</sup>                                                                                                                         |
| <u>tuf</u> A        | gène codant pour le facteur d'élongation Tu                                                                                                                    |
| wobble (hypothesis) | hypothèse qui admet qu'un même ARNt peut reconnaître plusieurs codons, en raison d'une certaine souplesse dans l'appariement au niveau du troisième nucléotide |

L'homme sait enfin qu'il est seul dans l'immensité indifférente de l'Univers d'où il a émergé par hasard. Non plus que son destin, son devoir n'est écrit nulle part. A lui de choisir entre le Royaume et les ténèbres.

Jacques Monod, Le hasard et la nécessité

### INTRODUCTION GENERALE

Depuis la compréhension du langage de la vie par la découverte du code génétique et des mécanismes de régulation de la biosynthèse des protéines, les biologistes peuvent maintenant non seulement lire l'information génétique grâce à la mise au point de méthodes de séquençage de l'ADN, mais aussi l'écrire par le développement de machines à synthétiser de l'ADN.

Ils sont ainsi capables de reprogrammer les cellules vivantes pour en étudier le fonctionnement et mieux comprendre les secrets de la génétique.

Un des problèmes majeurs à résoudre est la régulation de l'expression des gènes. Vu la grande masse d'information génétique complexe présente chez les organismes vivants, même les plus simples, ce problème est loin d'être résolu.

Il y a de multiples façons de l'aborder, dont en voici quelques-unes :

- 1) étudier des systèmes génétiques simples, comme les phages, les virus ou les bactéries
- 2) étudier l'expression des gènes dans des tissus très spécialisés, qui ne synthétisent en grande quantité que 2 ou 3 produits différents (p.ex. protéines de réserve des graines chez les plantes)
- 3) analyser des systèmes génétiques relativement simples présents dans des organismes complexes, c'est-à-dire les génomes mitochondrial ou chloroplastique

Le sujet de cette thèse était au départ d'étudier l'expression des gènes codant pour les protéines de réserve de soja. Mais assez rapidement, nous nous sommes rendus compte qu'un tel problème était beaucoup trop complexe pour être abordé dans le cadre d'un travail de thèse, d'autant plus que plusieurs groupes de recherche étaient déjà très avancés dans ce domaine. Nous avons donc décidé de changer de sujet de recherche et d'analyser le génome chloroplastique de soja.

Pour ces raisons, ce travail de thèse est divisé en deux parties. La première partie est consacrée à l'étude de l'accumulation des protéines de réserve dans les graines de soja durant le développement de la graine, alors que la deuxième partie traite de l'analyse de l'organisation structurale et génétique du génome chloroplastique de soja.

## PREMIERE PARTIE

### ETUDE DE L'ACCUMULATION DES PROTEINES DE RESERVE DES GRAINES DE SOJA DURANT LE DEVELOPPEMENT DE LA GRAINE

#### A. INTRODUCTION

L'étude des protéines de réserve des graines est d'un intérêt double pour le biologiste. D'une part, ces protéines sont des composants très importants pour la nutrition humaine et animale, d'autre part leur localisation dans des tissus spécifiques, c'est-à-dire les parties des graines comme les cotylédons, l'épicotyle, l'hypocotyle et la radicule offre un système attrayant pour l'étude de l'expression différentielle des gènes.

Ce système est d'autant plus intéressant que la plus grande partie des protéines de réserve consiste en 2 ou 3 types majeurs de protéines, et que leur accumulation a lieu durant une période relativement courte pendant la maturation de la graine.

#### 1. Caractéristiques générales des protéines de réserve des graines de Légumineuses

- Elles font partie de la classe des globulines et sont presque toutes des glycoprotéines (Derbyshire et al, 1976; Millerd, 1975).
- Elles forment le 20 à 25 % de la matière sèche des graines (Derbyshire et al, 1976) [plus de 40% pour le soja (Bils et Howell, 1963)], alors qu'en comparaison, les feuilles n'ont que 3 à 5% de protéines par poids de matière sèche, et constituent le 40 à 80% des protéines cotylédonaire (Derbyshire et al, 1976; Millerd, 1975; Hill et Breidenbach, 1974a).

- De nombreuses études effectuées sur les protéines de réserve de différentes Légumineuses ont montré qu'elles étaient constituées de 2, voir 3 grands types majeurs (Derbyshire et al, 1976) :
  - a. Les protéines qui sédimentent sur gradients de densité de saccharose avec un coefficient de sédimentation de 11-12S, regroupées sous le terme général de légumine (glycinine pour le soja) et dont la masse moléculaire varie entre 300'000 et 400'000 daltons.
  - b. Les protéines qui ont un coefficient de 7-8S, regroupées sous le terme général de vicilline ( conglycinine pour le soja), dont la masse moléculaire est comprise entre 100'000 et 200'000 daltons.
  - c. Les protéines de coefficient de sédimentation 2-3 S et environ 15 S, présentes en plus faible quantité que les deux premiers types cités.

Bien qu'une grande quantité de recherches ait été effectuée pour mieux caractériser ces protéines, leurs propriétés physico-chimiques ne sont pas encore connues avec certitude. Nous pouvons cependant en définir quelques caractéristiques :

- les globulines 11 S sont formées d'un certain nombre de sous-unités acides et basiques, de masse moléculaire approximative de 40'000 et 20'000, respectivement.
- les globulines 7 S sont formées de plusieurs types majeurs de sous-unités, dont les masses moléculaires vont de 20'000 à 80'000.

## 2. Cas du soja

On y trouve 3 composants majeurs :

- a. Un composant 11 S, appelé glycinine, qui est la protéine de réserve la plus abondante du soja, et peut représenter plus de 15% du poids sec de la graine (Millerd, 1975; Derbyshire et al, 1976). Elle a une masse moléculaire d'environ 350'000 daltons et est formée de 12 sous-unités (Badley et al, 1975), subdivisées en 6 sous-unités acides ( $M_r \approx 33'000$  et  $38'000$ ) et 6 sous-unités basiques ( $M_r \approx 22'000$ ).

Cependant, le nombre exact des sous-unités acides et basiques reste controversé, et pourrait changer suivant les variétés de soja (Barton et al, 1982).

- b. Une fraction 7 S, composée de  $\beta$  et de  $\gamma$ -conglycinine (cette dernière ne formant que le 10% de la  $\beta$ -conglycinine). Représentant le 10% environ du poids sec de la graine, la  $\beta$ -conglycinine est formée par différentes combinaisons de 3 sous-unités principales,  $\alpha'$  ( $M_r \approx 80'000$ ),  $\alpha$  ( $M_r \approx 75'000$ ) et  $\beta$  ( $M_r \approx 52'000$ ) (Thanh et Shibasaki, 1976), dont la masse moléculaire totale varie de 180'000 à 210'000 (Derbyshire et al, 1976). Les sous-unités  $\alpha'$  et  $\alpha$  montrent un haut degré de similitude dans leur contenu en acides aminés (Thanh et Shibasaki, 1977) et dans leur schéma de protéolyse partielle (Beachy et al, 1980b).

- c. Une fraction 2 S, appelée  $\alpha$ -conglycinine, formée de 2 composés majeurs de  $M_r \sim 30'000$  et  $\sim 20'000$ . Cette fraction forme moins du 5% des protéines de réserve. Elle contient par contre un certain nombre d'autres protéines de type enzymatique, comme par exemple un cytochrome c et des inhibiteurs de la trypsine (Wolf, 1970; Meinke et al, 1981)

L'isolation, la purification et les propriétés physico-chimiques de ces protéines ont été étudiées par de nombreux groupes. Pour de plus amples informations, consultez les revues de Millerd, (1975) et de Derbyshire et al, (1976).

D'autre part, l'étude de la biosynthèse de ces protéines est maintenant très avancée. Citons à propos du soja, les travaux de Beachy et al, (1980a, 1980b), de Schuler et al, (1982a, 1982b) et de Goldberg (1980).

Dans cette première partie de thèse, nous décrivons l'accumulation des protéines solubles des graines de soja durant le développement de la graine.

Un tel type de recherche a déjà été effectué par Hill et Breidenbach, (1974b), alors qu'au moment où nous finissions cette étude, des observations similaires ont été faites par Gayler et Sykes (1981) et par Meinke et al, (1981).

## B.MATERIEL ET METHODES

De nombreuses méthodes d'isolement des protéines solubles des graines sont décrites dans la littérature (Wolf et Briggs, 1956; Millerd, 1975; Derbyshire et al, 1976; Hill et Breidenbach, 1974a). Nous avons choisi une méthode dérivée de celle de Wolf et Briggs (1956), avec comme seule étape de purification, un passage sur gradients de densité de saccharose.

Cette méthode, si elle est très simple et rapide, ne permet cependant que de séparer les différentes protéines par rapport à leur coefficient de sédimentation.

## 1. Matériel biologique

Les graines de soja (Glycine max, variété Maple Arrow) nous ont été aimablement fournies par la Société d'Assistance Technique pour Produits Nestlé S.A. Les graines sont d'abord imbibées dans de l'eau distillée à température ambiante durant 6 à 10 h à l'obscurité, puis sont ensuite semées dans des bacs (ou des pots) contenant de la vermiculite. Les plantes de soja sont cultivées en chambre de culture sous des conditions contrôlées, avec une photopériode de 15 h de lumière (intensité lumineuse  $2.5 \text{ mW/cm}^2 \cdot \text{sec}$ ,  $25^\circ\text{C}$ , 70% d'humidité relative) et 9 h d'obscurité ( $20^\circ\text{C}$ , 85% d'humidité relative). A partir de la deuxième semaine, les plantes reçoivent 3 fois par semaine une solution nutritive selon Siegenthaler et Depéry, (1976).

Après 4 à 5 semaines, les premières fleurs apparaissent. Durant toute la période de formation des fruits, les gousses de 2-4 mm sont marquées séparément. Cette longueur est atteinte 2 à 3 jours après la floraison.

Les graines ont été récoltées périodiquement. Le poids frais des gousses et des graines a été déterminé immédiatement après la récolte. Les graines ont ensuite été conservées à  $-70^\circ\text{C}$ .

Le poids sec des graines a été obtenu par séchage à  $60^\circ\text{C}$  durant 36 h.

## 2. Extraction des protéines

Les graines sont broyées dans un mortier en présence de poudre de quartz jusqu'à l'obtention d'un broyat homogène. Les protéines sont extraites comme décrit par Wolf et Briggs (1956) et par Hill et Breidenbach (1974a), dans du tampon phosphate 0.035 M ( $0.0325 \text{ M K}_2\text{HPO}_4$ ,  $0.0026 \text{ M KH}_2\text{PO}_4$ , pH 7.6,  $0.4 \text{ M NaCl}$ ,  $0.014 \text{ M } \beta$ -mercaptoéthanol,  $0.02\% \text{ NaN}_3$ ,  $\mu = 0.5$ ) durant 30 min à  $20^\circ\text{C}$  avec un rapport 1:15 entre le poids sec (g) des graines et le volume (ml) de tampon. L'extrait obtenu est filtré à travers 4 couches de gaze, puis centrifugé à  $31'000 \times g$  pendant 15 min. Le surnageant est récupéré et utilisé pour la suite des expériences.

Le contenu en protéines de l'extrait filtré, du surnageant et du culot provenant de la centrifugation à  $31'000 \times g$ , ainsi que des fractions formant les pics des gradients, a été déterminé par la méthode de Lowry et al., (1951).

### 3. Gradients de densité de saccharose

Le surnageant  $31'000 \times g$  a été placé directement sur des gradients exponentiels de densité de saccharose, 10-30% (p/v) en concentration, 38 ml en volume, préparé avec le tampon d'extraction.

1 ml de surnageant  $31'000 \times g$  a été déposé sur chaque gradient et ceux-ci ont été centrifugés 40 h à  $27'000 \text{ rpm}$  ( $100'000 \times g$ ) à  $20^\circ\text{C}$  dans une ultracentrifugeuse Beckman L5-65 (rotor SW-27).

Les gradients ont été divisés en fractions de 1 à 1,3 ml et la quantité de protéines déterminée par la mesure de  $1'A_{280}$  et  $1'A_{260}$  (Layne, 1957).

Les fractions correspondant aux pics de protéines ont été réunies, leur contenu en protéines déterminé, puis conservées à  $-20^\circ\text{C}$ .

### 4. Electrophorèse sur gel de polyacrylamide

Tous les échantillons devant être analysés par électrophorèse ont été préalablement dialysés une nuit contre un tampon phosphate 0.035 M ( $0.0325 \text{ M K}_2\text{HPO}_4$ ,  $0.0026 \text{ M KH}_2\text{PO}_4$ , pH 7.6,  $0.014 \text{ M } \beta$ -mercaptoéthanol,  $0.02\% \text{ NaN}_3$ , 5% saccharose).

Le saccharose a été ajouté pour maintenir les protéines en solution. Nous avons en effet remarqué que ces protéines, lorsqu'elles sont congelées et décongelées, ont tendance à précipiter, notamment les globulines 11 S (la précipitation à froid est d'ailleurs une méthode d'isolation de ces protéines).

Les protéines ont été séparées par électrophorèse sur gel de poly-acrylamide sous des conditions dénaturantes par la méthode de Laemmli et Favre (1973). Les gels de concentration, contenant 6% d'acrylamide, ont été faits dans du tampon 0.125 M Tris-HCl pH 6.8, 0.1% SDS, 7mM  $\beta$ -mercaptoéthanol. Les gels de séparation, contenant 12% d'acrylamide, ont été faits dans du tampon 0.75 M Tris-HCl pH 8.8, 0.1% SDS, 7 mM  $\beta$ -mercaptoéthanol. L'électrophorèse a eu lieu dans du tampon 0.02 M Tris-glycine pH 8.6, 0.1% SDS, 14 mM  $\beta$ -mercaptoéthanol. Les gels ont été colorés par du bleu de Coomassie selon Chrambach et al., (1967).

### C. RESULTATS ET DISCUSSION

Nous avons étudié l'accumulation des protéines à partir du 13e jour après la floraison (DAF); des études précédentes ont en effet montré qu'avant ce stade, très peu de protéines sont présentes (Bils et Howell, 1963; Hill et Breidenbach, 1974a).

Pour nous assurer que les conditions de culture que nous avons utilisées étaient proches des conditions naturelles, nous avons pris comme contrôle des graines mûres, cultivées en milieu naturel (valeurs dénommées TE ou MS sur les différentes figures).

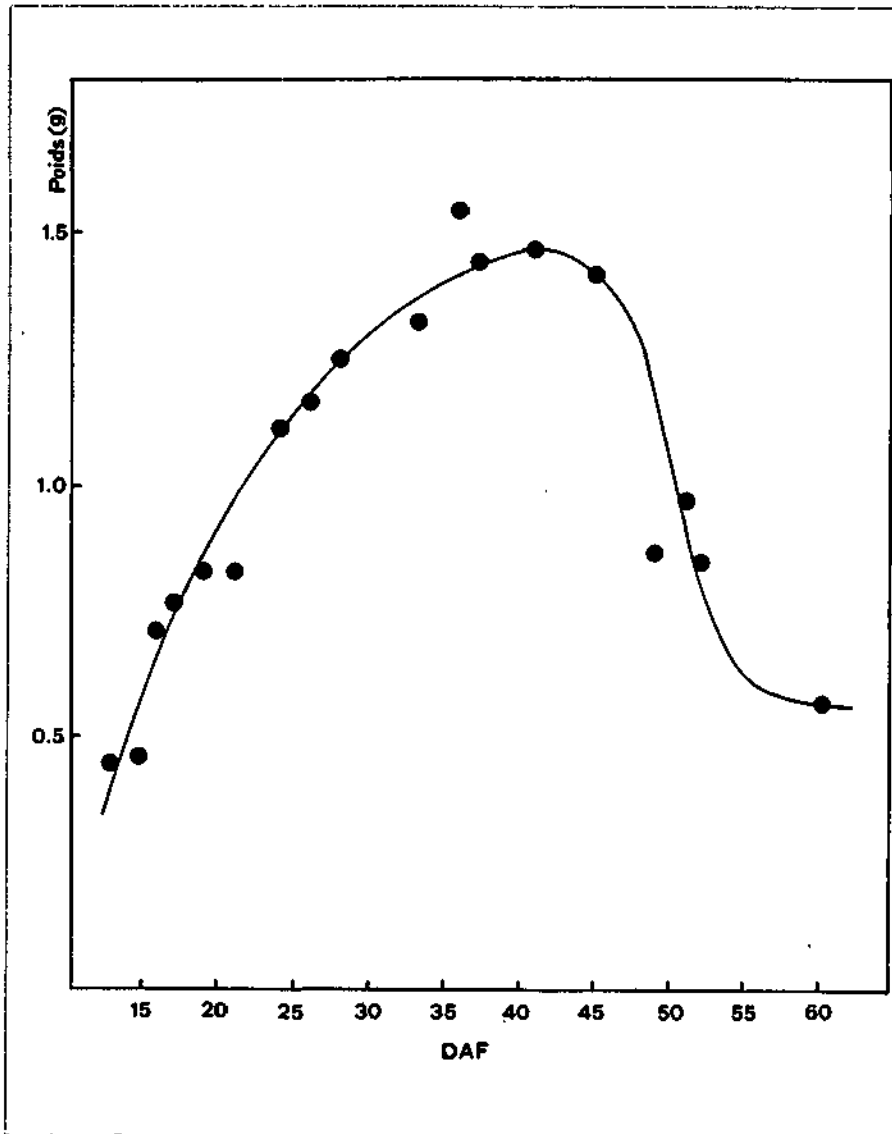
#### 1. Evolution du poids des gousses et des graines

La figure 1 montre l'évolution du poids frais/gousse. Le maximum ( 1.5 g) est atteint entre le 35e et le 45e jour. A partir du 50e jour environ, le poids va diminuer rapidement, pour se stabiliser vers le 55e jour ( 0.6 g).

Le poids frais/graine subit une évolution parallèle à celle du poids frais/gousse (figure 2). Au maximum de leur poids (35e -45e jour), les graines pèsent près de 400 mg; au moment de la maturation, leur poids va chuter de moitié.

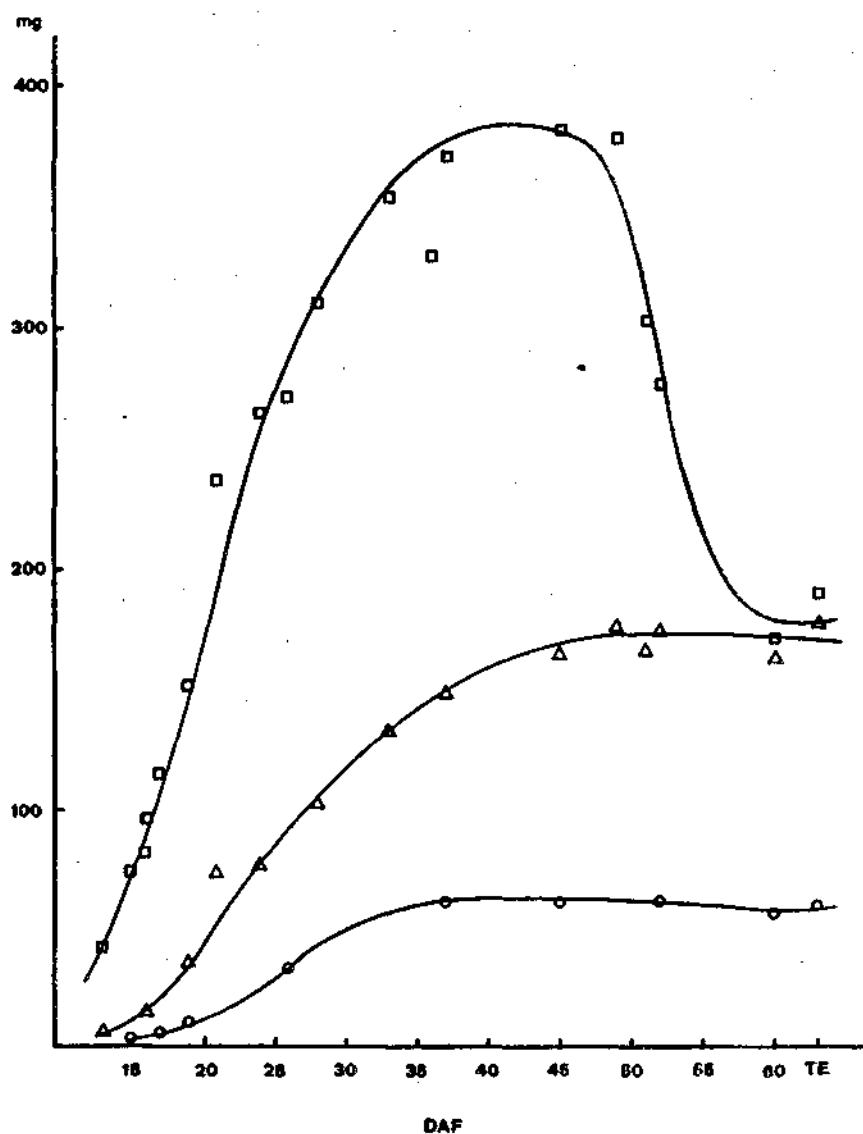
Figure 1 : Evolution du poids frais/gousse durant le développement des graines (moyenne d'au moins 10 gousses/DAF)

---



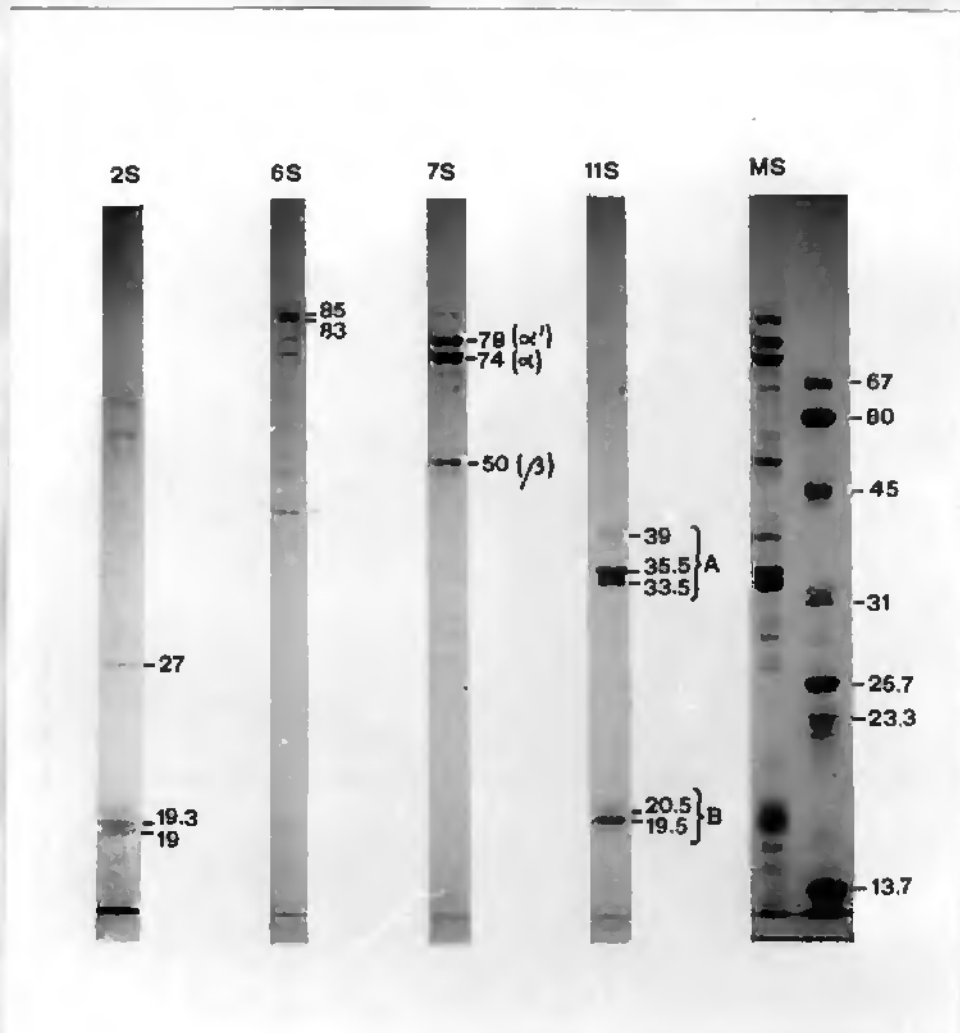
DAF : jours après la floraison

Figure 2 : Evolution du poids de la graine et de la quantité de protéines solubles/graine durant le développement (moyenne d'au moins 20 graines/DAF)



DAF : jours après la floraison  
□ : poids frais/graine  
△ : poids sec/graine  
○ : protéines solubles/graine  
TE : graines mûres

Figure 3 : Profils des protéines solubles des graines mûres  
obtenus sur SDS-PAGE 12%



2S, 6S, 7S et 11S : profils des polypeptides provenant des fractions  
pic des gradients

MS : profils des protéines totales solubles des graines  
mûres (surnageant 31'000  $\times$  g)

A : sous-unités acides de la glycinine

B : sous-unités basiques de la glycinine

Les masses moléculaires apparentes ont été estimées par rapport aux  
protéines standards suivantes : albumine sérique de boeuf, 67'000;  
catalase, 60'000; ovalbumine, 47'000; DNase I, 31'000; chymotrypsinogène  
25'700; trypsine, 23'300; RNase, 13'700. (en daltons)

Contrairement au poids frais/graine, qui augmente très rapidement dès les premiers stades du développement de la graine, le poids sec/graine augmente faiblement jusqu'au 20e jour environ, (figure 2), pour ensuite s'accroître fortement jusqu'au 45e jour, puis se stabiliser. Sous nos conditions de culture, la maturation de la graine est complète au 60e jour environ. Des résultats similaires ont été décrits par Hill et Breidenbach(1974b) et Gayler et Sykes (1981).

## 2. Extraction des protéines

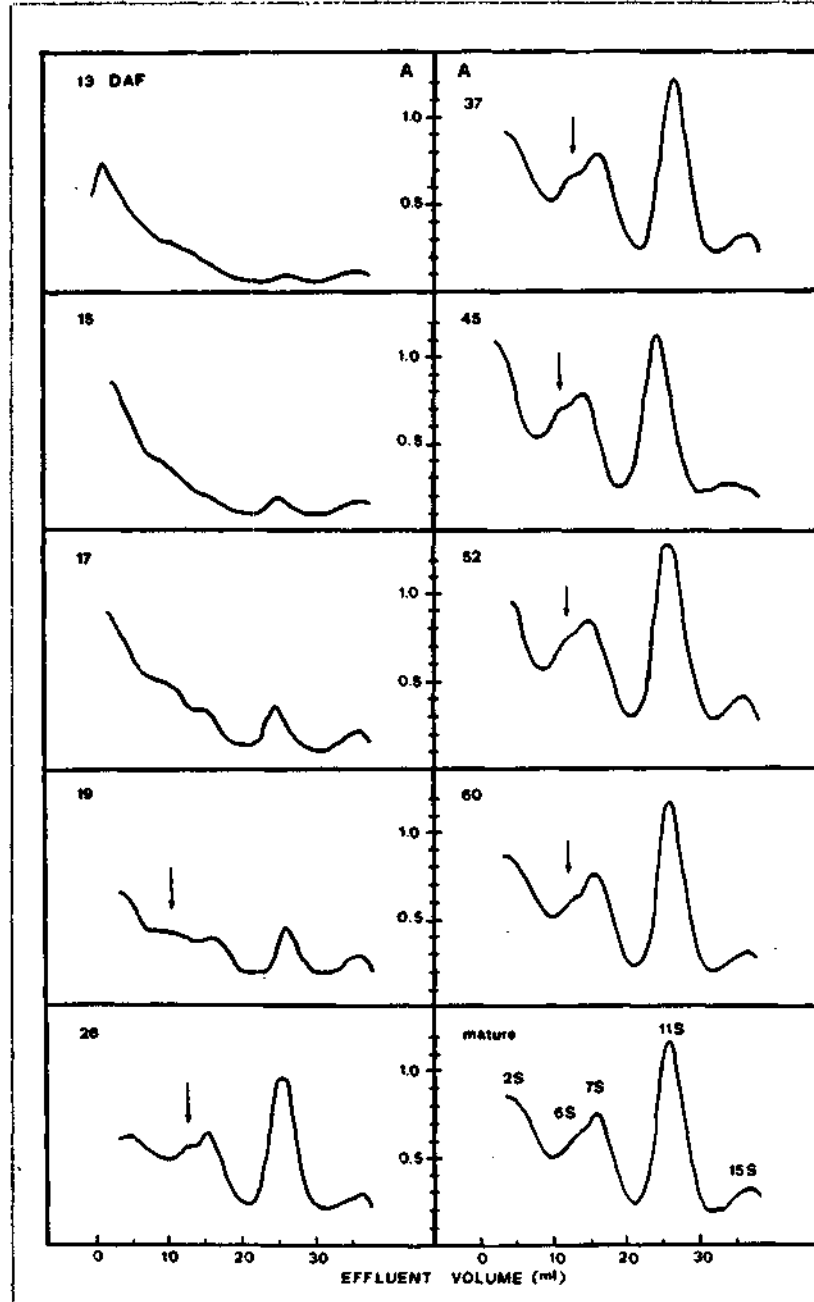
Le pourcentage d'extraction des protéines (rapport entre la quantité de protéines extraites et celle retrouvée dans le surnageant 31'000 × g) a varié entre 55 et 70% pour les stades jeunes (jusqu'au 26e jour) pour atteindre 80 à 90% pour les stades plus âgés. Un phénomène identique est décrit par Bils et Howell (1963), alors qu'au contraire, Hill et Breidenbach(1974a et 1974b) ont toujours obtenu, sauf pour le stade 12 jours, une valeur de 90%.

Cette augmentation du pourcentage d'extraction en fonction de la maturité de la graine pourrait être due à la présence, durant les stades juvéniles de la graine, de polypeptides en voie de formation qui seraient moins solubles sous les conditions expérimentales utilisées.

## 3. Identification des protéines de réserve

Nous avons étudié l'évolution des protéines formant les 3 pics principaux sur gradients de densité de saccharose, c'est-à-dire 2 S, 7 S et 11 S, ainsi que celle d'un pic dénommé ici 6 S (figure 4), qui est formé par plusieurs polypeptides qui ne sont pas caractéristiques du pic 7 S. La nature exacte de ces polypeptides n'est cependant pas déterminée, bien que certaines protéines différentes de la conglycinine soient présentes dans la fraction 7 S, notamment certains enzymes (Wolf, 1970).

Figure 4 : Profils des protéines solubles totales (surnageant 31'000 × g) obtenus sur gradients de densité exponentiels de saccharose



DAF : jours après la floraison

A : absorbance qui est exprimée par  $1.55 A_{280} - 0.76 A_{260}$  (Layne, 19

↓ : indique le pic 6S

La figure 3 montre les différents profils de polypeptides obtenus pour chaque fraction pic des gradients à partir de graines mûres. Les masses moléculaires apparentes estimées pour les sous-unités 2 S, 7 S et 11 S sont d'une manière générale en accord avec celles de la littérature (Hill et Breidenbach, 1974a; Thanh et Shibasaki, 1977; Draper et Catsimpoolas, 1977; Badley et al, 1975; Derbyshire et al, 1976).

#### 4. Evolution quantitative des protéines au cours de la maturation

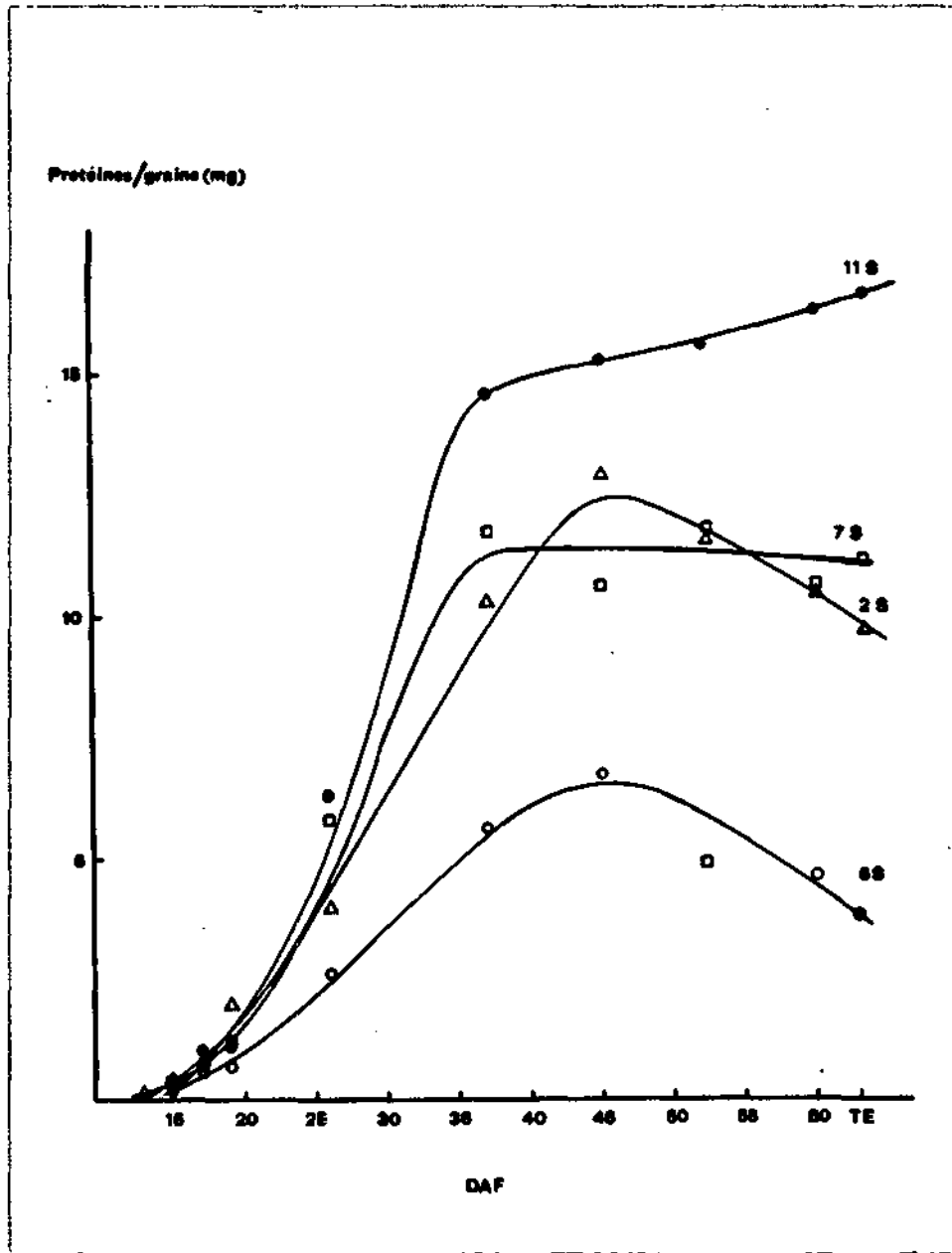
La figure 4 représente les profils de protéines obtenus sur gradient de densité de saccharose par la mesure de la différence d'absorbance à 280 nm et 260 nm (Layne, 1957). On remarque qu'au début du développement, les protéines sont surtout distribuées dans la partie supérieure du gradient et qu'au 37e jour, le profil obtenu est tout à fait comparable à celui des graines mûres, démontrant ainsi qu'à ce stade, la formation et l'accumulation des protéines de réserve sont achevées (voir aussi figure 2).

La figure 5 montre l'évolution quantitative des protéines de chaque fraction des gradients au cours de la maturation de la graine.

Nous voyons que les globulines 7 S et 11 S s'accumulent rapidement jusqu'au 35e jour environ, et qu'ensuite leur quantité reste à peu près stable jusqu'à la maturation.

Au contraire, la quantité de protéines des fractions 2 S et 6 S a tendance à diminuer lors de la maturation. Cette diminution pourrait s'expliquer par le fait que ces fractions contiennent certains enzymes, qui ne seraient plus nécessaires et donc absents dans la graine mûre.

Figure 5 : Evolution de la quantité de protéines de chaque fraction pic des gradients durant la formation de la graine



DAF : jours après la floraison

TE : graines mûres

## 5. Evolution qualitative des protéines au cours de la maturation

Si la méthode d'isolation utilisée, c'est-à-dire la séparation des protéines par rapport à leur coefficient de sédimentation, est simple et rapide, elle ne permet cependant pas une isolation complète des différentes protéines; il y a en effet toujours pour chaque fraction, une faible contamination, tout particulièrement pour les fractions 6 S et 7 S.

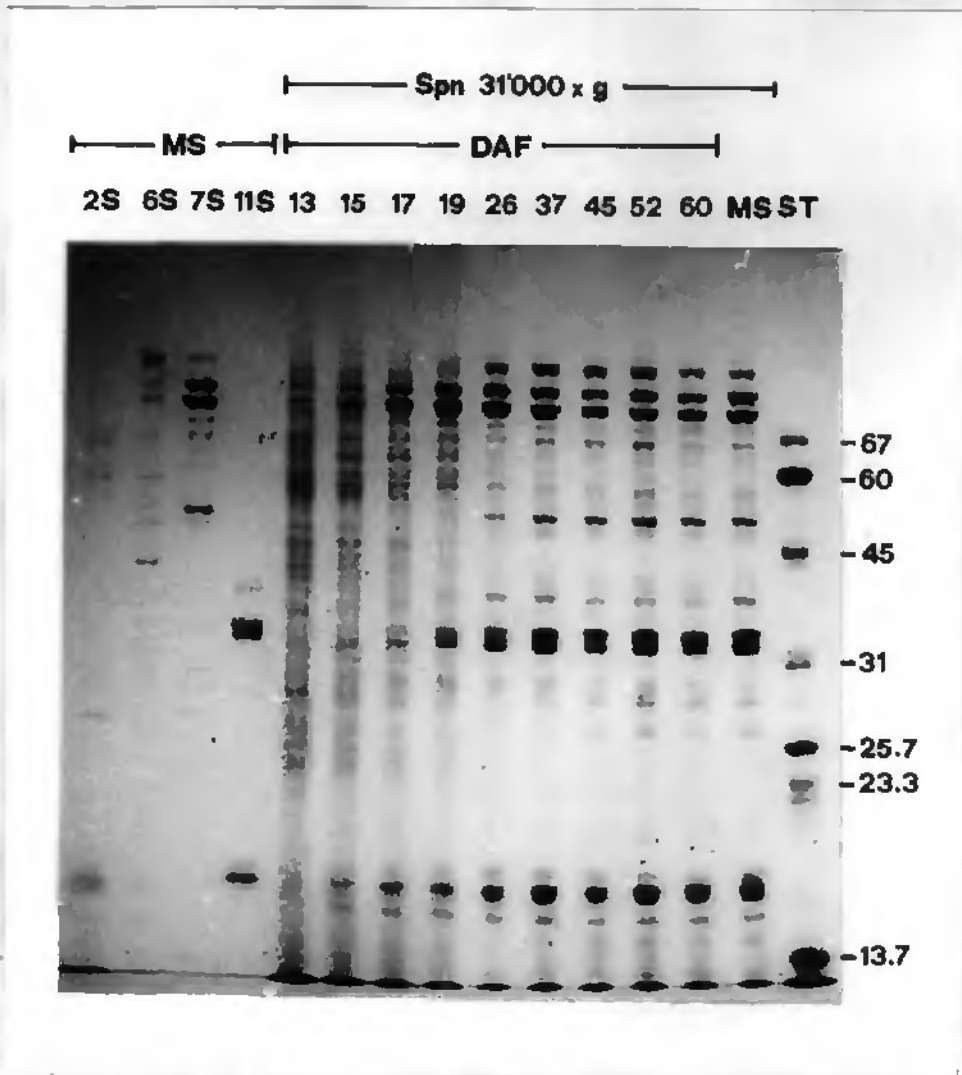
Les analyses par électrophorèse sur gel de polyacrylamide des polypeptides présents lors des différents stades de développement (figure 6) montrent que c'est durant une période d'environ 2 semaines, comprise entre le 13<sup>e</sup> et le 26<sup>e</sup> jour, qu'a lieu le plus grand remaniement au niveau de la synthèse et de l'accumulation des polypeptides qui donneront naissance aux protéines typiques des graines mûres. Nous voyons en effet sur la figure 6 que le profil général des protéines solubles des graines (surnageant 31'000 × g) ne varie pratiquement plus à partir du 26<sup>e</sup> jour. Ces observations sont comparables avec celles faites par Gayler et Sykes (1981).

Dans le but de déterminer si les polypeptides de chaque fraction subissaient une accumulation différentielle durant le développement de la graine, nous avons étudié l'évolution de chaque fraction isolée.

### 5.1 Fraction 2 S

L'évolution qui a lieu dans le profil des protéines de cette fraction semble paradoxale, puisque c'est lors des stades très avancés de la graine que des changements se produisent. En effet, nous voyons sur la figure 7 que jusqu'au 26<sup>e</sup> jour, 2 bandes de Mr 19'000 et 19'300 sont présentes, alors qu'entre le 26<sup>e</sup> et le 37<sup>e</sup> jour apparaissent une bande de Mr 27'000 et un ou plusieurs polypeptides migrant avec le front.

Figure 6 : Evolution du profil électrophorétique des protéines solubles totales (surnageant 31'000 × g) durant le développement de la graine (SDS-PAGE 12%)



Spn 31'000 × g : surnageant 31'000 × g

MS : graines mûres

ST : protéines standards (voir figure 3)

DAF : jours après la floraison

La disparition de la bande de Mr 19'300 dans les stades avancés de la graine est sans doute due à la quantité d'échantillon déposé sur le gel.

La nature exacte de ces différents polypeptides est difficile à préciser, mais nous pouvons supposer que la protéine de Mr 27'000 et une des bandes de Mr ~19'000 sont de l' $\alpha$ -conglycinine (Derbyshire et al, 1976), alors que les autres bandes dénotent sans doute la présence de certains enzymes, comme par exemple du cytochrome c (Wolf, 1970).

## 5.2 Fraction 6 S (figure 8)

Sur la figure 4, nous voyons que le profil des protéines du surnageant 31'000 x g, obtenu sur gradient de saccharose, montre à côté du pic 7 S un second pic chevauchant partiellement le pic 7 S, et qui reste présent durant tout le développement de la graine.

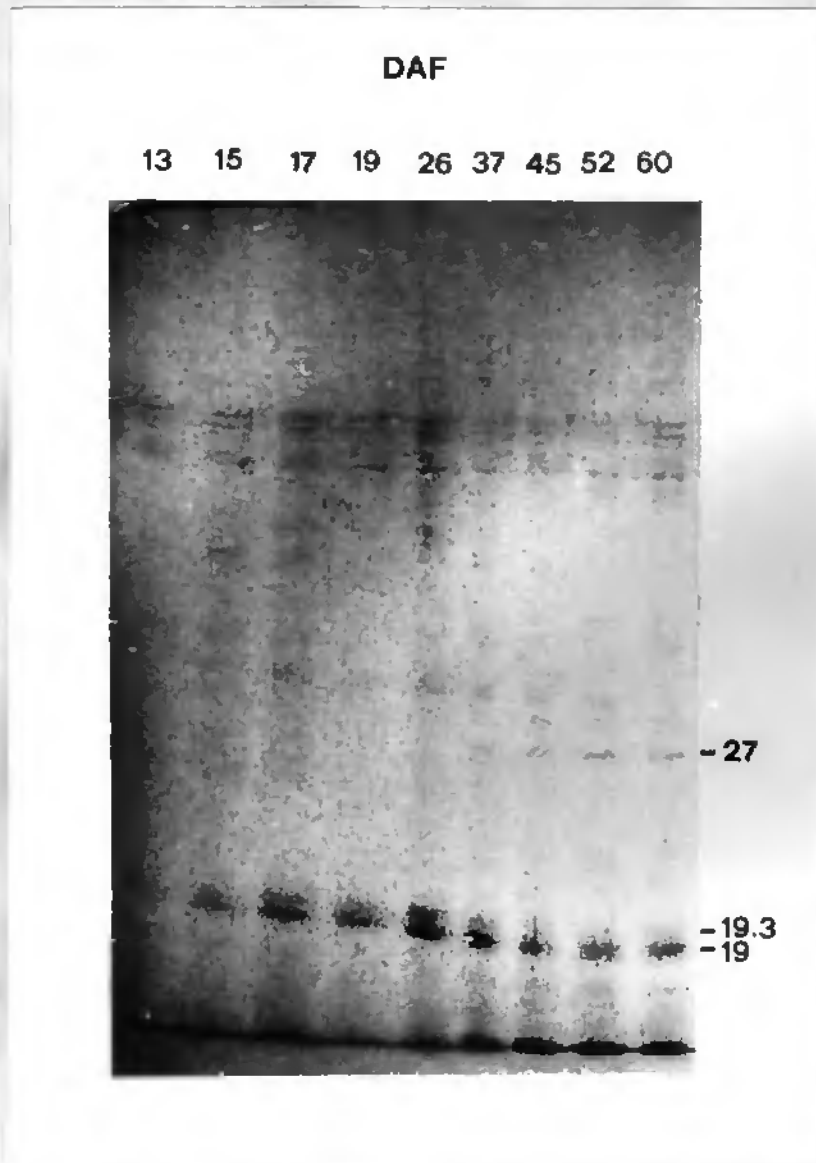
Cette fraction, que nous avons appelée 6 S, n'apparaît pas dans le travail de Hill et Breidenbach (1974a).

Elle est formée principalement de deux polypeptides de Mr 85'000 et 83'000, de fonctions inconnues. Cependant, des résultats récents (Meinke et al, 1981; Barton et al, 1982) démontrent qu'une de ces 2 bandes pourrait correspondre à une uréase.

Ces 2 polypeptides apparaissent ensemble entre le 19e et le 26e jour, et sont toujours présents dans les graines mûres. Ils suivent ainsi l'évolution des protéines des fractions 7 S et 11 S (voir ci-dessous).

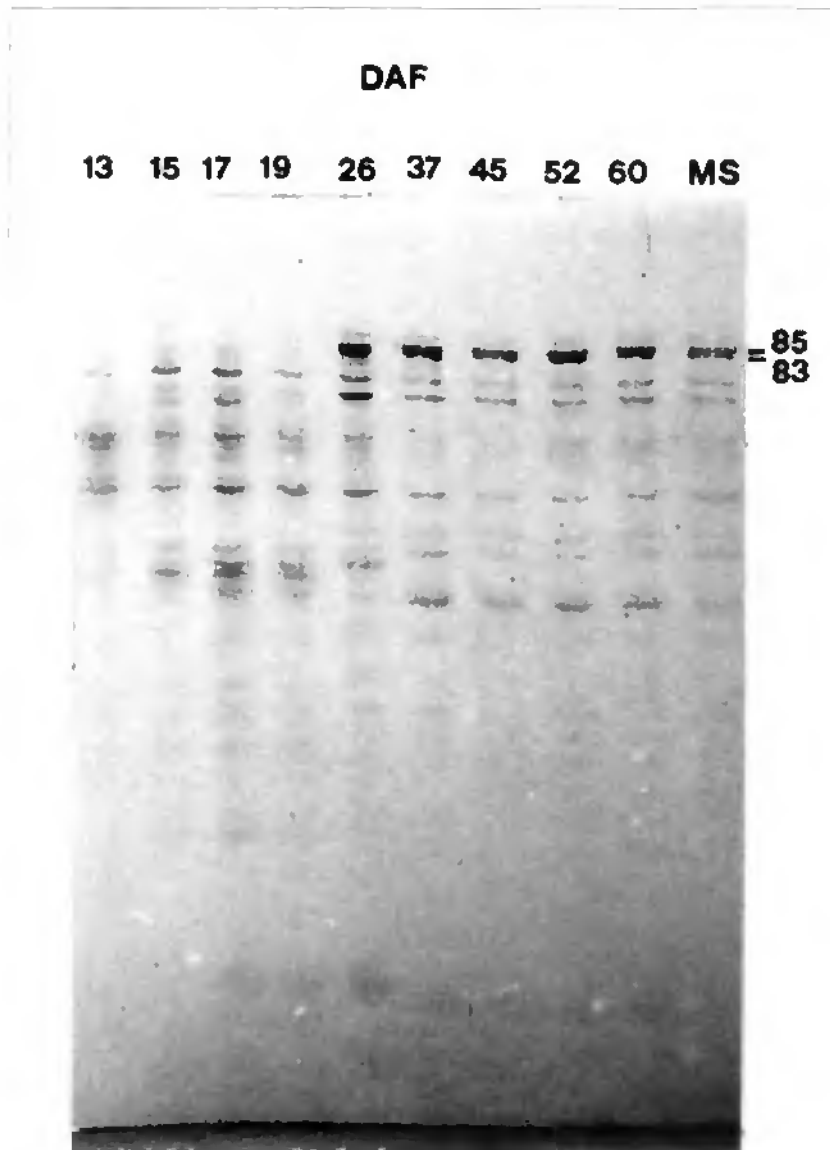
Figure 7.: Evolution du profil électrophorétique de la fraction 2S  
durant le développement de la graine (SDS-PAGE 12%)

---



DAF : jours après la floraison

Figure 8 : Evolution du profil électrophorétique de la fraction 6S  
durant le développement de la graine (SDS-PAGE 12%)



DAF : jours après la floraison

### 5.3 Fraction 7 S

La figure 9 montre que dans cette fraction, très peu de protéines sont présentes au 13<sup>e</sup> jour. A partir du 15<sup>e</sup> jour apparaît la sous-unité  $\alpha$ , alors que les sous-unités  $\alpha'$  et  $\beta$  commencent à s'accumuler seulement depuis le 26<sup>e</sup> jour. Cette apparition temporelle différentielle des 3 sous-unités de la  $\beta$ -conglycinine a également été observée par Hill et Breidenbach (1974b) et par Gayler et Sykes (1981).

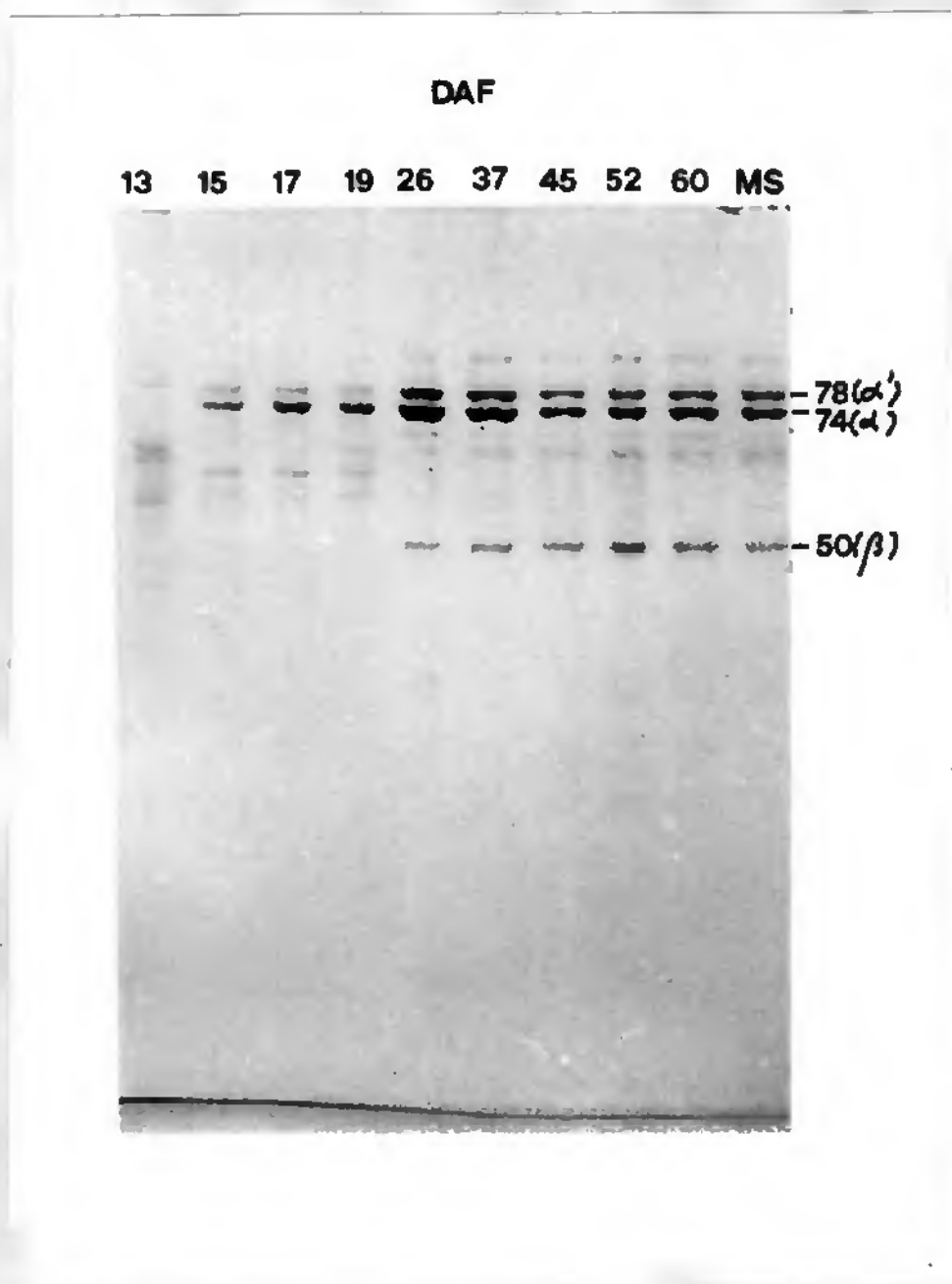
Ce résultat est intéressant, dans la mesure où Thanh et Shibasaki (1978) ont démontré l'existence de 6 isomères de la  $\beta$ -conglycinine; chaque isomère est un trimère formé de la combinaison unique des 3 sous-unités ( $\alpha\alpha\alpha$ ,  $\alpha\alpha\alpha'$ ,  $\alpha\alpha\beta$ ,  $\alpha\alpha'\beta$ ,  $\alpha\beta\beta$ ,  $\alpha'\beta\beta$ ). L'apparition précoce de la sous-unité  $\alpha$  pourrait s'expliquer parce que c'est elle qui est susceptible de former le plus de trimères et qu'elle est donc présente en plus grande quantité que les autres.

### 5.4 Fraction 11 S

Contrairement aux observations faites par Hill et Breidenbach (1974b) la fraction 11 S subit également des différences temporelles dans sa composition en polypeptides durant le développement de la graine (figure 10). En effet, jusqu'au 17<sup>e</sup> jour, un polypeptide de Mr 83'000 est nettement visible, et sa disparition à partir du 19<sup>e</sup> jour correspond à l'apparition des protéines majeures de la glycinine dans les graines mûres.

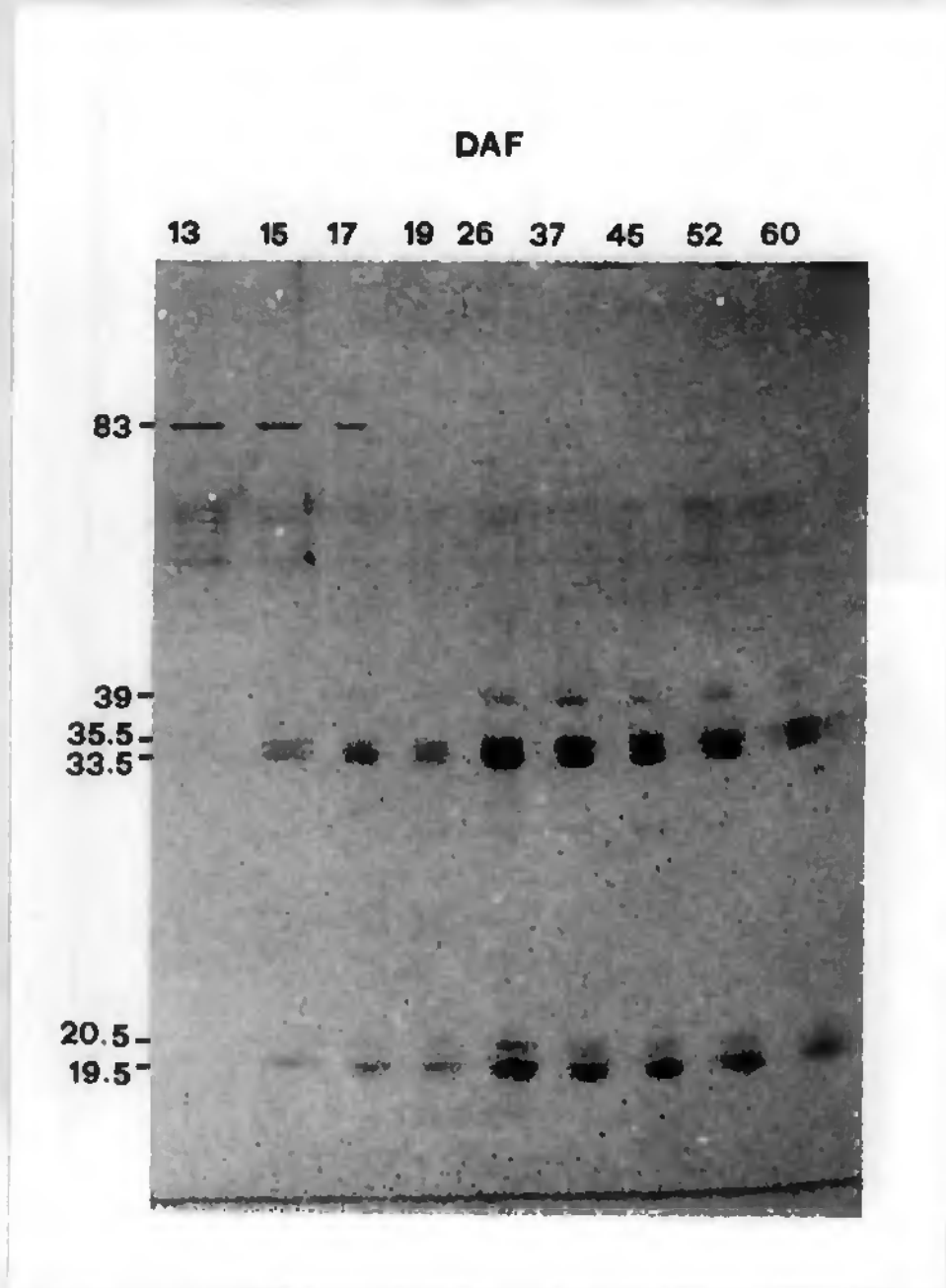
Nous pouvons de plus démontrer des différences temporelles dans l'accumulation des sous-unités acides et basiques. En effet, la bande de Mr 39'000 apparaît plus tard que les autres et il en est de même pour celle de Mr 20'500. Une telle observation a d'ailleurs été faite récemment par Meinke et al., (1981).

Figure 9 : Evolution du profil électrophorétique de la fraction 7S  
durant le développement de la graine (SDS-PAGE 12%)



DAF : jours après la floraison

Figure 10 : Evolution du profil électrophorétique de la fraction 11S durant le développement de la graine (SDS-PAGE, 12%)



DAF : jours après la floraison

Le fait que ces 2 bandes semblent disparaître lors de la maturation de la graine est sans doute un artefact dû à la quantité d'échantillon déposée sur le gel.

Il a été établi que les sous-unités de la glycinine étaient synthétisées sous la forme de précurseurs dont les tailles varient de 59'000 à 63'000 daltons (Barton et al, 1981). Ceci nous a conduit à supposer que le polypeptide de Mr 83'000 n'était pas un précurseur de la glycinine.

Cependant, Barton et al, (1982) ont identifié récemment l'ARNm qui code pour la glycinine dans les graines de soja. Or, cet ARNm a une taille d'environ 2100 nucléotides et serait donc capable de coder pour une protéine de Mr ~77'000, valeur qui est proche de 83'000. Il est par conséquent assez difficile de trancher actuellement entre ces deux hypothèses.

## 6. Discussion

Il est possible de diviser la formation et l'accumulation des protéines de réserve des graines de soja en 5 phases principales :

- I. 0-13 jours : Très peu de protéines synthétisées. Cette période est essentiellement caractérisée par le développement de la gousse, alors qu'au niveau de la graine, elle correspond à un stade de divisions cellulaires (Bils et Howell, 1963). En fait, au 13e jour, moins de 5% de la matière sèche est présente, et moins de 3% des protéines ont été formées.
- II. 13-26 jours : Formation des protéines de réserve. Peu de protéines se trouvent déjà sous leur structure définitive.

- III. 26-37 jours : Accumulation très rapide et importante des protéines.  
La fin de cette phase correspond à l'arrêt de la synthèse des protéines. L'étude qualitative et quantitative du contenu en protéines montre que celui-ci est semblable à celui des graines mûres.
- IV. 37-50 jours : La quantité de protéines reste constante. La graine se prépare à entrer en dessiccation.
- V. 50-60 jours : Phase de dessiccation, durant laquelle le poids frais/graine chute de moitié, alors que la quantité de protéines va diminuer légèrement, diminution qui est surtout marquée pour les fractions 2 S et 6 S.

Examinons plus en détail les phases II et III :

- II. C'est la période de formation des différents polypeptides qui donneront naissance aux protéines complètes des graines mûres. Contrairement aux résultats de Hill et Breidenbach (1974b), où seules les sous-unités des globulines 7 S subissent une accumulation différentielle au cours du temps, nous avons pu démontrer que ce phénomène est général, et que non seulement les globulines 7 S obéissent à ce phénomène, mais aussi les globulines 2 S, 6 S et 11 S.
- D'un point de vue génétique, il serait intéressant de se demander si la régulation différentielle des gènes codant pour ces protéines a lieu à un niveau transcriptionnel ou post-transcriptionnel. Meinke et al., (1981) ont démontré que la régulation de l'expression de ces gènes avait lieu essentiellement à un niveau transcriptionnel, bien que la présence d'ARNm codant pour les sous-unités  $\beta$  de la  $\beta$ -conglycinine durant les premiers stades de développement de la graine indique que des contrôles au niveau de la traduction pourraient restreindre la synthèse de ces sous-unités à des stades spécifiques du développement.

D'autre part, Schuler et al, (1982a et b) ont démontré que la  $\beta$ -conglycinine était codée par une famille de gènes étroitement apparentés, et que, tandis que les séquences des parties codantes pour les sous-unités  $\alpha'$  et  $\alpha$  étaient assez proches l'une de l'autre, les parties 3' non traduites de ces gènes différaient notablement.

L'accumulation différentielle de ces polypeptides pourrait ainsi être due, soit à l'efficacité de la traduction des ARNm, soit à la stabilité de ces ARNm.

La régulation de l'expression des gènes codant pour les protéines de réserve des graines de soja semble donc avoir lieu à la fois au niveau de la transcription et de la traduction (Meinke et al, 1981; Schuler et al, 1982a et b; Gayler et Sykes, 1981).

- III. L'accumulation des protéines devient très importante à partir du 20e jour environ, et durant la période d'un peu plus de 2 semaines comprise entre le 20e et le 37e jour, le poids sec/graine va quadrupler ( $\sim 40$  mg  $\rightarrow$   $\sim 160$  mg), alors que la quantité de protéine/graine va s'accroître énormément, pour passer de 10 à 60 mg, ce qui représente une moyenne de 3 mg de protéine formée par graine et par jour.

Finalement, à maturité, une graine de soja va contenir plus de 30% (par rapport à la matière sèche) de protéines solubles, dont le 80% au moins est constitué par des protéines de réserve.

#### D.CONCLUSION

Cette étude nous a permis de définir avec précision une période, comprise entre le 20e et le 40e jour après la floraison, durant laquelle la biosynthèse des protéines de réserve est très importante, et où l'on assiste à de grands remaniements dans la composition en protéines dans les graines de soja.

Ces observations nous auraient sans doute été très utiles lors d'une analyse de la population des ARNm présents durant le développement de la graine. Malheureusement, pour les raisons invoquées dans l'introduction générale, nous avons été contraint de changer de sujet de recherche.

L'analyse du génome chloroplastique de soja constitue donc la deuxième partie de ce travail.

## DEUXIEME PARTIE : ANALYSE DE L'ADN CHLOROPLASTIQUE DE SOJA

### I. INTRODUCTION

La vie sur terre dépend étroitement du processus photosynthétique, qui a lieu dans les chloroplastes des plantes. Ces organites cellulaires, comme les mitochondries, possèdent leur propre système génétique, c'est-à-dire les quatre composants essentiels nécessaires à l'expression de l'information génétique : ADN, ADN polymérase, ARN polymérase et une machinerie de synthèse de protéines (Ellis, 1981).

Cependant, des expériences génétiques (Kung, 1977) et des expériences de marquages à l'aide d'isotopes radioactifs en présence d'antibiotiques, qui inhibent spécifiquement soit la synthèse des protéines dans les organelles, soit dans le cytoplasme, ont montré que la biogenèse et la fonction du chloroplaste dans la cellule végétale dépendaient fortement de l'activité du génome nucléaire (Ellis, 1975).

Cette semi-autonomie peut être clairement illustrée par l'exemple de la ribulose 1,5-bisphosphate carboxylase (Highfield *et al*, 1978), qui est une enzyme chloroplastique essentielle pour la photosynthèse. Les petites sous-unités de cette enzyme sont codées par le génome nucléaire et synthétisées sous la forme de précurseurs au niveau des polyribosomes cytoplasmiques. Les précurseurs polypeptidiques sont ensuite transférés dans les chloroplastes, traités et assemblés avec les grandes sous-unités qui sont codées par le génome chloroplastique et synthétisées à l'intérieur du chloroplaste. La formation et la fonction du chloroplaste résultent donc de l'interaction complexe entre le système génétique nucléaire et celui du chloroplaste.

Le génome chloroplastique se trouve sous la forme d'une molécule circulaire d'ADN double brin, ayant une densité moyenne de  $1.697 \text{ g/cm}^3$  (Tewari, 1979), ce qui correspond à un contenu en G+C d'environ 37 %. Cette molécule est présente en 10 à 60 copies identiques/chloroplaste (Bedbrook et Kolodner, 1979). Suivant les espèces végétales, sa longueur varie de 40 à 45  $\mu\text{m}$ , ceci correspondant à une masse moléculaire de 70 à 100 Md (Bedbrook et al, 1979), ou à 120'000 à 170'000 bp.

Les analyses de ce génome par des endonucléases de restriction ont montré que la plupart des ADNct des plantes supérieures ont une organisation structurale semblable, c'est-à-dire deux régions répétitives inversées (d'environ 23'000 bp) séparées par une petite région unique et une grande région unique (Bedbrook et al, 1979).

Il y a également une similitude dans l'organisation génétique, puisque les gènes pour les ARN ribosomiques sont localisés dans les régions répétitives inversées, les gènes pour les ARN de transfert dans les régions répétitives et la grande région, alors que les gènes codant pour des protéines se trouvent disséminés partout sur l'ADN, mais avec des localisations identiques pour un même gène.

Il existe cependant des cas où l'organisation structurale et génétique est différente de celle décrite ci-dessus, et notamment :

- parmi les algues, chez l'euglène (Euglena gracilis), où il n'y a pas de régions répétitives inversées, et où les gènes pour les ARN ribosomiques sont arrangés en tandem (Jenni et Stutz, 1978).
- dans la famille des Légumineuses, chez le pois (Pisum sativum) et la fève (Vicia faba), où une des régions répétitives est manquante (sans doute perdue au cours de l'évolution) (Koller et Delius, 1980; Palmer et Thompson, 1981a).

Jusqu'à présent, les gènes localisés sur le génome chloroplastique couvrent moins du 20% de la capacité codante totale de cet ADN, d'où l'intérêt d'étudier l'organisation génétique de cette molécule pour tenter de préciser son rôle en tant que lieu d'information autonome, ainsi que les interactions entre les trois centres d'information de la cellule, à savoir le noyau, le chloroplaste et la mitochondrie. Une telle étude est d'autant plus intéressante qu'elle peut également nous amener à une meilleure compréhension de la régulation de l'expression des gènes.

Dans la deuxième partie de cette thèse, nous décrivons l'isolation, la purification et la caractérisation physique et génétique de l'ADNct de soja (variété Maple Arrow).

Le soja (Glycine max L., Merrill) fait partie de la famille des Légumineuses, tribu des Phaseolae, sous-tribu des Glycinae (Tauber, 1891). Du fait de son importance nutritionnelle, cette plante est cultivée depuis très longtemps par l'homme et il existe aujourd'hui une multitude de variétés, qui sont susceptibles de se développer sous différents climats.

Nous avons choisi d'étudier le génome chloroplastique de cette plante pour les raisons suivantes :

- le laboratoire de biochimie de l'Université de Neuchâtel travaille depuis de nombreuses années sur une espèce végétale inférieure, l'algue unicellulaire Euglena gracilis. Il nous paraissait important d'élargir ce champ de recherche en étudiant une plante supérieure.
- le soja fait partie des Légumineuses, et comme déjà mentionné plus haut, il a été démontré que les réarrangements de séquences dans le génome chloroplastique des espèces de cette famille sont beaucoup plus fréquents qu'ailleurs.

Dans le chapitre II, nous présentons l'isolation et la caractérisation physico-chimique de l'ADNct de soja. Le chapitre III traite de la construction d'une carte physique de cet ADNct par l'utilisation d'enzymes de restriction, ainsi que de la construction d'une carte génétique avec la localisation de plusieurs gènes.

Le chapitre IV est consacré aux techniques de clonage et à la création d'une banque de clones de l'ADNct de soja. Dans le chapitre V, nous étudions la structure fine de deux gènes (photogène et ARNt<sup>His</sup>) par l'analyse de leur séquence nucléotidique. Le bilan général de ce travail de thèse est présenté au chapitre VI.

## II. ISOLATION, PURIFICATION ET CARACTERISATION PHYSICO-CHIMIQUE DE L'ADN CHLOROPLASTIQUE DE SOJA

---

### A. INTRODUCTION

Depuis les premières isolations d'ADNct chez les euglènes (Edelman et al, 1964; Manning et al, 1971), de nombreuses publications sont apparues qui traitent de l'isolation d'ADNct à partir d'une grande variété de plantes (Tewari, 1979).

Les principales difficultés résident d'une part dans l'obtention de chloroplastes non contaminés par des mitochondries, des cellules ou des noyaux, d'autre part dans l'obtention de molécules d'ADNct circulaires intactes. D'une façon générale, deux grands types de méthodes sont décrites dans la littérature (Tewari, 1979) :

#### 1. Méthode comprenant un traitement avec une DNase et plusieurs extractions au phénol

---

- Isolation des chloroplastes par centrifugation différentielle et par gradients de densité
- Elimination de l'ADN nucléaire contaminant par traitement avec une DNase
- Dégradation des protéines par traitement avec une protéase
- Dénaturation et élimination des résidus polypeptidiques par plusieurs extractions au phénol
- Elimination de l'ARN par traitement avec une RNase, suivie à nouveau d'un traitement avec une protéase et d'extractions au phénol
- La solution d'ADN est finalement dialysée et l'ADN concentré par précipitation à l'alcool

## 2. Méthode sans traitement avec une DNase, et purification de l'ADN sur gradient de chlorure de césium (CsCl)

-----

- Isolation des chloroplastes par centrifugation différentielle et par gradients de densité
- Elimination de l'ARN par traitement avec une RNase
- Dégradation des protéines par traitement avec une protéase
- Purification de l'ADN sur gradients de CsCl
- La solution d'ADN est ensuite dialysée et l'ADN concentré par précipitation à l'alcool

## B. MATERIEL ET METHODES

### 1. Matériel biologique

Les conditions de culture du soja sont les mêmes que celles décrites dans la première partie de ce travail. Les feuilles âgées de 2 à 3 semaines ont été récoltées après 36 à 48 h d'obscurité, afin d'éliminer l'amidon.

### 2. Isolation de l'ADN chloroplastique

L'ADNct a été isolé suivant le principe de la deuxième méthode décrite ci-dessus.

Dans une isolation typique, 300 g de feuilles sont homogénéisées par plusieurs broyages brefs (3 à 5 sec) dans 1,2 l de tampon A à 4°C (0.2 M Tris-HCl, 0.2 M EDTA, 0.3 M D-mannitol, 7 mM  $\beta$ -mercapto-éthanol, pH 9.0). L'homogénat est filtré à travers 6 couches de gaze et centrifugé à 100 x g durant 5 min, pour éliminer les débris végétaux et une partie des cellules entières et des noyaux. Le surnageant est récupéré et centrifugé 15 min à 1500 x g. Le culot brut de chloroplastes est doucement resuspendu dans environ 10 ml de tampon B (0.05M Tris, 0.02 M EDTA, 0.3 M saccharose, pH 8.0) et déposé ensuite sur 2 couches de saccharose (volume : 35 ml; concentration : 30-50%; préparé dans du tampon C : 0.05 M Tris, 0.02 M EDTA, pH 8.0).

Les tubes sont centrifugés 30 min à  $1500 \times g$ . La fraction contenant les chloroplastes est récupérée, diluée avec environ 300 ml de tampon C et recentrifugée 15 min à  $1500 \times g$ . Le culot final de chloroplastes est resuspendu dans du tampon C jusqu'à un volume total de 11.2 ml. Afin de lyser complètement les chloroplastes pour qu'ils libèrent leur ADN, nous ajoutons 2.8 ml de tampon C contenant 10% de sarcosinate de sodium. Après 5 min d'incubation à  $0^{\circ}\text{C}$ , la lyse des chloroplastes est contrôlée au microscope et le mélange est centrifugé 5 min à  $1500 \times g$ . Le surnageant est traité avec de la RNase A ( $50 \mu\text{g/ml}$ ) pendant 1 h à température ambiante, puis avec de la protéinase K ( $200 \mu\text{g/ml}$ ) pendant 30 min à  $37^{\circ}\text{C}$ . Nous ajoutons alors 0.78 g/ml de CsCl et après 1 h 30 à  $0^{\circ}\text{C}$ , le mélange est centrifugé 30 min à  $16000 \times g$  (précipitation et élimination d'une partie des résidus polypeptidiques).

Le surnageant est déposé sur un coussin de CsCl (volume : 1.5 ml; 1.009 g/ml CsCl,  $n_D=1.4032$ ,  $200 \mu\text{g/ml}$  EtBr). Les tubes sont centrifugés dans un rotor Beckman Ti 65 à 39000 rpm pendant 16 h à  $20^{\circ}\text{C}$ . La bande d'ADN est récoltée sous lumière U.V. (366 nm) et l'EtBr est éliminé par 2 à 3 extractions avec un volume égal d'isopentyl-alcool (saturé avec 10 mM EDTA, pH 8.0).

L'ADN est ensuite dialysé contre  $3 \times 1 \text{ l}$  de  $0.1 \times \text{SSC}$  pendant 24 h à  $4^{\circ}\text{C}$ , et concentré par précipitation à l'alcool (30 min,  $-70^{\circ}\text{C}$ ). Le précipité d'ADN est récolté par centrifugation (20 min,  $16000 \times g$ ) et le culot est lavé 1  $\times$  avec de l'éthanol 80% froid.

Après séchage sous vide, le culot d'ADN est redissout dans  $50 \mu\text{l}$  de tampon TE. D'habitude, 25 à  $30 \mu\text{g}$  d'ADNct sont isolés à partir de 300 g de feuilles.

### 3. Détermination de la densité de flottation de l'ADNct de soja

La densité de flottation de l'ADNct a été déterminée par une centrifugation à l'équilibre dans une solution de CsCl d'indice de réfraction 1.6992 pendant 22 h (25°C) à 40000 rpm, avec comme référence de l'ADN de Micrococcus lysodeikticus (1.731 g/cm<sup>3</sup>).

## C. RESULTATS ET DISCUSSION

### 1. Isolation et purification de l'ADNct de soja

Au début de ce travail de thèse, nous avons essayé d'isoler l'ADNct de soja en testant plusieurs méthodes, comme celles décrites dans l'introduction. Des analyses par digestion de l'ADNct avec des enzymes de restriction (voir chap. III) ont montré que cet ADNct était toujours passablement contaminé par de l'ADN nucléaire et qu'il était souvent très dégradé, c'est-à-dire de faible poids moléculaire.

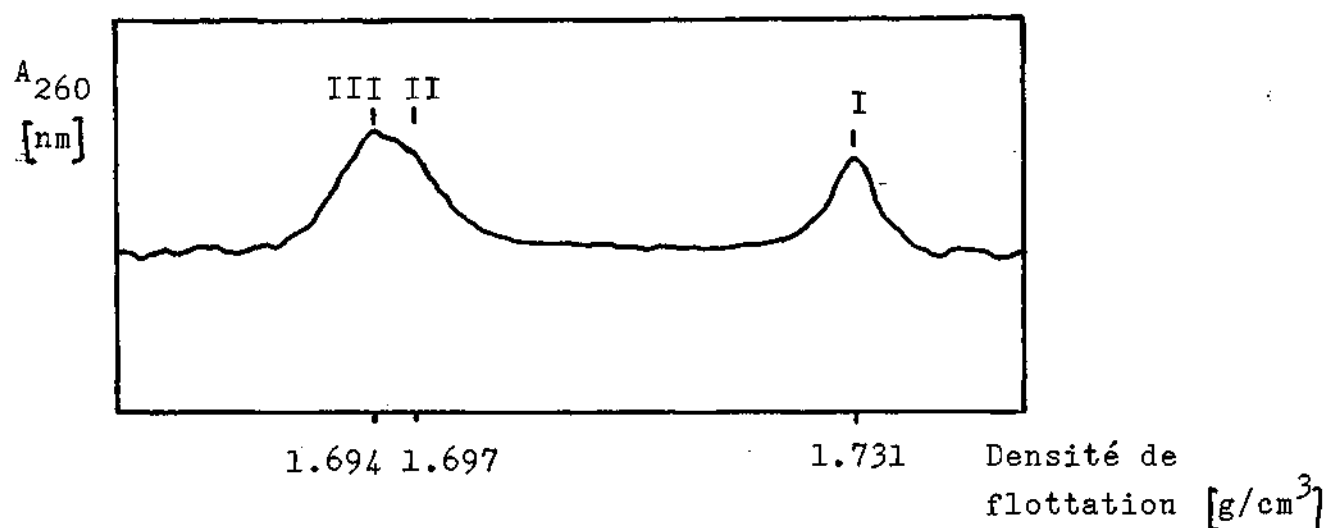
Pour remédier à ces problèmes, nous avons alors choisi une méthode d'isolation sans traitement avec une DNase et sans extractions au phénol, mais avec purification des chloroplastes sur coussins de saccharose et purification de l'ADN sur gradients de CsCl. De plus, pour éviter la dégradation de l'ADN par des nucléases endogènes, nous avons augmenté la concentration en Tris et en EDTA dans le tampon de broyage (tampon A).

Ces modifications nous ont permis d'obtenir de l'ADNct relativement peu contaminé par de l'ADN nucléaire et d'assez grande taille (40-60 kbp) pour qu'il soit possible de l'analyser par digestion avec des enzymes de restriction. Cependant, la qualité de l'ADNct isolé ainsi était toujours variable d'une préparation à l'autre.

## 2. Densité de flottation de l'ADNct de soja

L'ADNct de soja montre sur gradient de densité de CsCl une bande relativement large, avec deux pics principaux se chevauchant (figure 1). Ces faits suggèrent que l'ADNct de cette préparation est d'une part de taille hétérogène et d'autre part fortement contaminé par de l'ADN nucléaire.

Figure 1 : Détermination de la densité de flottation de l'ADNct de soja



I : *Micrococcus lysodeikticus*

II : ADNct de soja

III : ADN nucléaire de soja

Les valeurs moyennes obtenues pour ces deux pics sont de  $1.697 \pm 0.002$  g/cm<sup>3</sup> pour le pic II et  $1.694 \pm 0.002$  pour le pic III.

La valeur du pic II est en accord avec celles trouvées pour d'autres ADNct de plantes supérieures (Tewari, 1979). D'ailleurs, Synenki *et al.*, (1978), ont également trouvé cette valeur pour l'ADNct de soja. La valeur du pic III est comparable à celle déterminée pour l'ADN nucléaire d'autres plantes (Tewari, 1979).

Le pic III est plus important que le pic II, ce qui démontre que cette préparation est fortement contaminée par de l'ADN nucléaire. Cette expérience de détermination de densité n'a été faite qu'une seule fois, et ceci au début de ce travail de thèse, c'est-à-dire à l'époque où nous n'avions pas encore trouvé les conditions optimales de purification de l'ADNct.

Quoiqu'il en soit, les résultats obtenus ne diffèrent pas de ceux trouvés chez d'autres plantes de la famille des Légumineuses, comme Pisum sativum et Vigna radiata (Tewari, 1979). Cependant, seule l'analyse de l'ADNct de soja par digestion avec des enzymes de restriction nous permettra de savoir si celui-ci est semblable par sa taille à celui de l'haricot (150'000 bp) ou à celui du pois (120'000 bp).

### III.CONSTRUCTION D'UNE CARTE PHYSIQUE ET GENETIQUE DE L'ADNct DE GLYCINE MAX.

#### A.INTRODUCTION

Le génome chloroplastique se trouve sous la forme d'une molécule circulaire d'ADN, dont la taille varie entre 120'000 et 170'000 bp, suivant son origine. Pour déterminer son organisation structurale, il faut être capable de fragmenter cette molécule en petits morceaux et de séparer les fragments ainsi obtenus les uns des autres. Ceci est possible grâce à l'utilisation des enzymes de restriction.

Ces endonucléases de restriction sont des enzymes isolés à partir de bactéries. Ils sont capables de reconnaître une séquence spécifique d'ADN (séquence formée généralement de 4 à 8 nucléotides, toujours organisée en palindrome) et de couper la molécule d'ADN, le plus souvent à l'endroit même de cette séquence, parfois quelques dizaines de bases en amont ou en aval de la séquence de reconnaissance.

Les fragments d'ADN ainsi obtenus sont séparés par électrophorèse sur gel (agarose ou polyacrylamide), et leur taille est estimée par rapport à des standards.

Pour construire une carte de restriction d'un ADN quelconque, il faut d'abord tester un certain nombre d'enzymes de restriction disponibles sur le marché, pour ensuite sélectionner ceux qui coupent cet ADN le moins souvent possible. En combinant les résultats obtenus par des digestions simples et mixtes, il est possible d'obtenir une série d'informations nous permettant d'ordonner certains fragments les uns par rapport aux autres. Cependant, pour pouvoir construire une carte de restriction complète de l'ADNct de soja, il faut avoir recours à d'autres techniques.

Une approche rapide aboutissant à la localisation de sites de restriction consiste à séparer les plus grands fragments dans un gel d'agarose à bas point de fusion et ensuite de recouper chacun des fragments obtenus par un second enzyme (Herrmann et al, 1980). Cette approche n'a pas été efficace dans notre cas, d'une part parce qu'il était difficile de bien séparer certains fragments les uns des autres, d'autre part parce que ces mêmes fragments n'étaient pas toujours de qualité suffisante (dégradation de l'ADNct de certaines préparations).

Nous avons par conséquent construit une carte de restriction de l'ADNct de soja par hybridation de contact entre des sondes radioactives d'ADNct exogènes de fonction génétique connue et la copie sur filtre de différents profils de restriction de l'ADNct.

Une telle approche est possible grâce à la très grande homologie de séquences de certaines parties structurales des gènes chloroplastiques (Fluhr et Edelman, 1981 ). Cette méthode a en plus l'avantage de nous permettre de construire en même temps une carte génétique, par la localisation immédiate de ces gènes. D'autre part, nous avons cloné plusieurs fragments d'ADNct de soja (voir chapitre IV), qui ont non seulement été utilisés comme sonde, mais encore analysés dans des expériences de digestions avec un deuxième ou un troisième enzyme.

## B.MATERIEL ET METHODES

### 1.Endonucléases de restriction

Les enzymes de restriction utilisés dans le cadre de ce travail ont été obtenus soit de Boehringer, Mannheim, soit de New England Biolabs, Beverly.

Les digestions simples de l'ADNct (0.5 à 1.5  $\mu$ g) ont été faites dans des volumes de réaction de 20  $\mu$ l, sous les conditions recommandées par les fournisseurs.

Les digestions doubles ou triples ont été faites principalement de deux façons :

- l'ADN est incubé avec les deux (ou trois) enzymes différents en même temps
- l'ADN est d'abord digéré avec le premier enzyme, puis le deuxième est ajouté après ajustement des conditions de réaction

Les réactions sont stoppées en ajoutant 0.2 volume d'une solution contenant : 2 M saccharose, 5% Ficoll 400, 0.2 M EDTA, 0.05% de bleu de bromophénol.

## 2. Electrophorèse sur gel d'agarose

Dans ce procédé, les molécules d'ADN sont séparées dans un système tamponné (Tris-acétate 40mM, Na-acétate 20 mM, EDTA 1 mM, pH 7.8) sur un support inerte formé d'agarose polymérisé.

Les électrophorèses ont lieu soit sur des gels verticaux (20 × 15 × 0.2 ou 0.3 cm), soit sur des gels horizontaux (25 × 20 × 0.4 cm).

Suivant la taille des fragments à séparer, la concentration d'agarose a varié de 0.5% à 1.3%. Pour les gels verticaux, la migration se faisait pendant 12 à 15 h à 2-3 Volts/cm, et pendant 15 à 24 h à 1-2 Volts/cm pour les gels horizontaux.

A la fin de la migration, les gels sont colorés dans une solution de bromure d'éthidium (0.5 µg/ml) et photographiés sous lumière U.V. Comme références de poids moléculaire, nous avons utilisé l'ADN des phages Lambda et T<sub>7</sub>, et leurs fragments de restriction Hind III et Hpa I, respectivement. (Nosikov et Sain, 1977; McDonell et al, 1977).

## 3. Transferts d'ADN

Cette méthode (Southern, 1975) permet de transférer des fragments d'ADN d'un gel sur une feuille de papier nitrocellulose (Millipore, HAWP, 304F0, 0.45 µm) aboutissant à l'obtention d'une copie exacte du gel.

L'ADN est dénaturé par immersion du gel dans une solution de NaOH 0.5 M, NaCl 1.5 M, puis neutralisé dans une solution de Tris-HCl 0.5 M, NaCl 3 M, pH 7.0. Le transfert a lieu durant 12 à 15 h par passage d'une solution de  $20 \times$  SSC à travers le gel. La feuille de nitrocellulose (=filtre) est ensuite dessalée, séchée à l'air, puis cuite à  $80^{\circ}\text{C}$  sous vide pendant 2 h.

La réplique du gel obtenu sur le filtre peut alors être utilisée dans des expériences d'hybridation.

#### 4. Marquage de l'ADN par "nick-translation"

Cette méthode (Rigby et al, 1977) est basée sur l'utilisation de l'ADN polymérase I (enzyme de Kornberg) de Escherichia coli, qui possède une activité de polymérisation en  $5' \rightarrow 3'$ , alors qu'elle a également une activité de dégradation exonucléotidique en  $3' \rightarrow 5'$ .

Elle peut donc à partir d'un trou ("nick") dans une molécule d'ADN, synthétiser un nouveau brin complémentaire au brin intègre. Si le mélange réactionnel contient un désoxyribonucléotide radioactif, l'ADN obtenu est alors radioactif. Il est utilisé ensuite comme sonde dans des expériences d'hybridation.

#### 5. Hybridation ADN-ADN

Cette technique est notamment utilisée pour étudier le degré de complémentarité entre deux chaînes nucléotidiques ou pour localiser sur une molécule d'ADN quelconque une séquence spécifique, correspondant à celle de la sonde utilisée.

Le principe de cette méthode est basé sur le fait que, sous certaines conditions (température, concentration en formamide), deux molécules d'ADN simples brins peuvent s'apparier (s'hybrider) partiellement ou totalement, suivant leur degré d'homologie de séquence.

Pratiquement, nous procédons comme suit :

Un filtre sur lequel se trouve la réplique de profils de restriction est hybridé durant 12 à 15 h à 42°C ou 37°C dans une solution contenant : 5 × SSC, 50 ou 15% formamide, 0.02% polyvinylpyrrolidone, 0.02% Ficoll, 0.02% albumine sérique de boeuf (Denhardt, 1966) et 100 µg/ml d'ADN soniqué de thymus de veau ou de sperme de hareng. Nous ajoutons ensuite 0.3 à 1.0 µg d'ADN radioactif (sonde) et l'hybridation a lieu dans des sacs en plastique scellés pendant 24 à 48 h à 42 ou 37°C.

Après l'hybridation, le filtre est lavé plusieurs fois à température ambiante et à 65°C avec du 2 × SSC. Il est ensuite séché et recouvert d'un film X-ray (Kodak XAR-5). L'exposition se fait à température ambiante durant 1 à 30 jours.

## 6. Electroélution

C'est une des méthodes qui permet de ressortir d'un gel d'agarose ou de polyacrylamide un fragment d'ADN préalablement séparé d'autres fragments par électrophorèse.

En pratique, une quantité assez importante d'ADN, provenant généralement d'un plasmide, est digérée par une enzyme de restriction, puis les fragments obtenus sont séparés par électrophorèse. Le fragment intéressant est excisé du gel, la poche ainsi créée dans le gel étant agrandie et tapissée par une membrane de tube à dialyse; le fragment d'ADN est remis en place et recouvert de tampon. Par électrophorèse, l'ADN sort de l'agarose et il peut être récupéré dans le tampon dans lequel baigne le morceau de gel.

Cette solution d'ADN est ensuite extraite avec un mélange de phénol-chloroforme 1:1 (saturé avec du TNE 1×) et l'ADN est concentré par précipitation à l'éthanol.

## 7. Nomenclature

Dans l'ordre de leur taille décroissante, les fragments d'ADN obtenus par digestion simple avec Sal I, Pst I et Sma I sont appelés Sal-A, Sal-B, etc..., Pst-A, Pst-B, etc...

Les fragments obtenus après digestion mixte avec Sal I/Pst I sont appelés SP-A, SP-B, etc..., avec Sal I/Sma I, SSm-A, SSm-B et ainsi de suite pour les autres digestions mixtes.

Si, dans une digestion, une bande contient plusieurs fragments d'ADN de mobilité électrophorétique identique, ceux-ci sont appelés  $A_1$  et  $A_2$ , etc... s'ils font partie des régions répétitives inversées, ou A et A', etc... s'ils ne font pas partie des régions répétitives inversées.

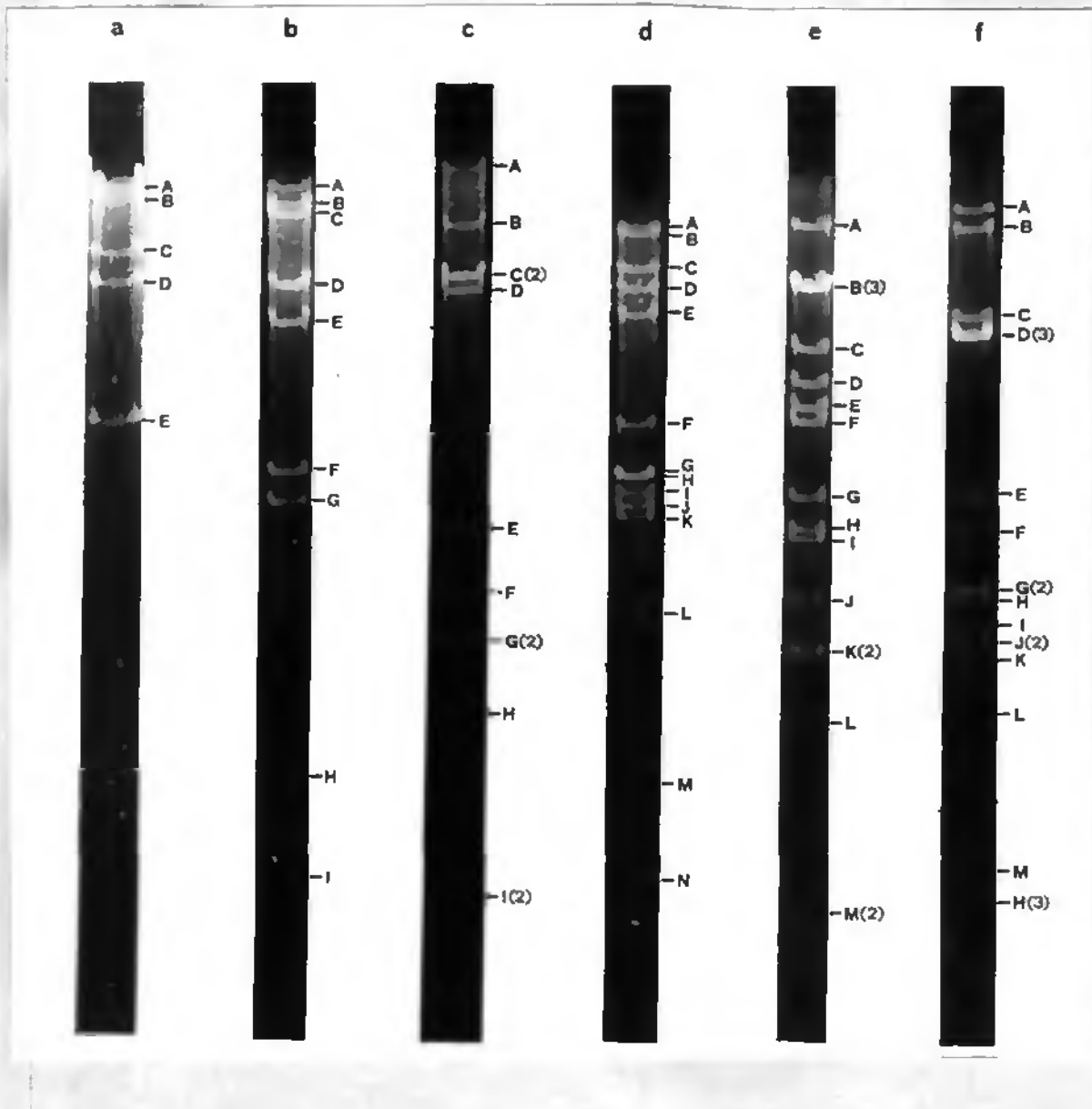
### C.RESULTATS ET DISCUSSION

#### 1.Nombre et taille des fragments d'ADNct après digestion avec les endonucléases de restriction Sal I, Pst I et Sma I

Pour construire une carte physique de l'ADNct de soja, nous avons choisi, après avoir testé une quinzaine d'enzymes de restriction différents, les 3 enzymes suivants :

| Nom<br>(Origine)                  | Séquence de reconnaissance<br>: : endroit de coupure | Coupures dans<br>l'ADNct de soja |
|-----------------------------------|------------------------------------------------------|----------------------------------|
| Sal I<br>(Streptomyces albus G)   | G <sup>+</sup> TCGAC<br>CAGCT <sup>+</sup> G         | 5                                |
| Pst I<br>(Providencia stuartii)   | CTGCA <sup>+</sup> G<br>G <sup>+</sup> ACCTC         | 9                                |
| Sma I<br>(Serratia marcescens Sb) | CCC <sup>+</sup> GGG<br>GGG <sup>+</sup> CCC         | 12                               |

Figure 1 : Profils électrophorétiques de l'ADNct de soja obtenus après digestions avec Sal I, Pst I et Sma I



Les nombres entre parenthèses indiquent les bandes contenant 2 ou 3 fragments de mobilité électrophorétique identique sous les conditions expérimentales utilisées. Sal I(a), Pst I(b), Sma I(c), Sal I/Pst I(d), Sal I/Sma I(e), Pst I/Sma I(f).

Nous avons choisi ces 3 enzymes essentiellement parce qu'ils ne coupaient pas souvent l'ADNct de soja et que les fragments ainsi obtenus étaient de tailles raisonnables pour pouvoir être séparés les uns des autres.

La figure 1 montre les 6 profils électrophorétiques obtenus après digestion de l'ADNct total de soja par Sal I, Pst I, Sma I, Sal I/Pst I, Sal I/Sma I, Pst I/Sma I. Dans toutes les digestions, les molécules d'ADN sont partiellement dégradées, créant un bruit de fond visible surtout dans la partie supérieure du gel. Cependant, il est possible de distinguer clairement toutes les bandes.

Certaines des bandes montrent une stoechiométrie proche de 2 (p. ex. Sma-C) ou de 3 (p. ex. SSm-B), estimée par rapport à leur intensité de coloration (figure 1). Ces bandes contiennent soit des fragments identiques provenant des régions répétitives inversées, soit des fragments de tailles identiques ou très semblables, et qui ne sont pas séparés sous ces conditions d'électrophorèse.

La liste des fragments obtenus après digestions simples et mixtes de l'ADNct de soja par Sal I, Pst I et Sma I est représentée sur la table 1. La taille du génome chloroplastique total, calculée à partir de la somme des fragments des diverses digestions est de l'ordre de 150-152 kbp; ceci correspond à la taille du génome chloroplastique d'espèces de Légumineuses qui possèdent deux régions répétitives inversées (Palmer et Thompson, 1981a)

## 2. Construction d'une carte physique de restriction Sal I/Pst I/Sma I

Nous avons utilisé 5 sondes d'ADN exogènes, provenant soit du génome chloroplastique d'Euglena gracilis, soit de celui de l'épinard (Spinacia oleracea) pour construire une carte physique de l'ADNct de soja et pour localiser les gènes portés par ces sondes.

Table 1 : Nomenclature et taille (en kbp) des fragments de restriction de l'ADNct de soja

| Sal I               | Pst I               | Sma I               | SP            | SSm                          | Psm                  |
|---------------------|---------------------|---------------------|---------------|------------------------------|----------------------|
| A 58.8 <sup>a</sup> | A 41.1 <sup>a</sup> | A 55.4 <sup>a</sup> | A 27.0        | A 27.0                       | A 30.0(Pst-C)        |
| 8 44.9 <sup>a</sup> | 8 32.8 <sup>a</sup> | B 25.1              | B 26.0        | B 16.8(3x)(Sma-C, 2x)(Sal-D) | B 25.1(Sma-8)        |
| C 21.3              | C 30.0 <sup>a</sup> | C 16.8(2x)          | C 18.1        | C 12.3                       | C 13.7(Pst-E)        |
| D 16.8              | D 16.8              | D 15.7              | D 16.8(Pst-D) | D 10.6                       | D 12.6(3x)           |
| E 8.9               | E 13.7              | E 5.3               | E 14.1        | E 9.7                        | E 6.2(Pst-G)         |
|                     | F 7.2               | F 4.1               | F 8.9(Sal-E)  | F 8.9(Sal-E)                 | F 5.3(Sma-E)         |
|                     | G 6.2               | G 3.3(2x)           | G 7.2(Pst-F)  | G 6.5                        | G 4.2(2x)            |
|                     | H 1.9               | H 2.4               | H 7.1         | H 5.6                        | H 4.1                |
|                     | I 1.0               | I 0.9(2x)           | I 6.6         | I 5.3(Sma-E)                 | I 3.6                |
|                     |                     |                     | J 6.2(Pst-G)  | J 4.1(Sma-F)                 | J 3.3(2x)(Sma-G)     |
|                     |                     |                     | K 6.0         | K 3.3(2x)(Sma-G)             | K 3.1                |
|                     |                     |                     | L 4.0         | L 2.4(Sma-H)                 | L 2.4(Sma-H)         |
|                     |                     |                     | M 1.9(Pst-H)  | M 0.9(2x)(Sma-I)             | M 1.0(Pst-I)         |
|                     |                     |                     | N 1.0(Pst-I)  |                              | N 0.9(3x)(Sma-I, 2x) |
|                     |                     |                     |               |                              | O 0.1 <sup>b</sup>   |

Les nombres entre parenthèses indiquent les bandes contenant 2 ou 3 fragments de mobilité électrophorétique identique sous les conditions expérimentales utilisées. Dans les doubles digestions, les fragments qui ne sont pas coupés par le second enzyme sont indiqués entre parenthèses.

a : La taille de ces grands fragments a été calculée à partir des tailles des sous-fragments correspondants, plutôt que par leur mobilité électrophorétique.

b : Ce fragment a été vu sur une digestion Sma I du fragment cloné Pst-H.

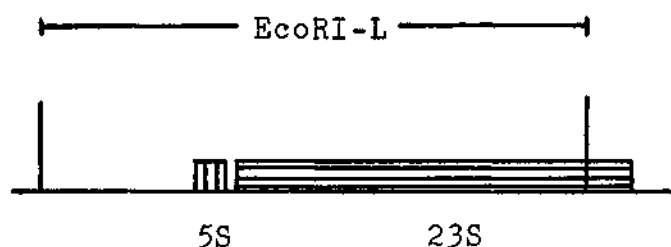
## 2.1. Evidence et localisation de deux régions répétitives inversées portant les gènes pour les ARN ribosomiques 23S et 16S

---

La sonde utilisée pour localiser les gènes des ARNr 23S est le fragment EcoRI-L (3.2 kbp) provenant de l'ADNct d'Euglena gracilis. Ce fragment contient environ les deux tiers du gène codant pour l'ARNr 23S (région 3') et la totalité du gène de l'ARNr 5S (Jenni et Stutz, 1978; Hallick et al, 1978) (figure 2)

Figure 2 : Représentation schématique du fragment EcoRI-L de l'ADNct d'E. gracilis

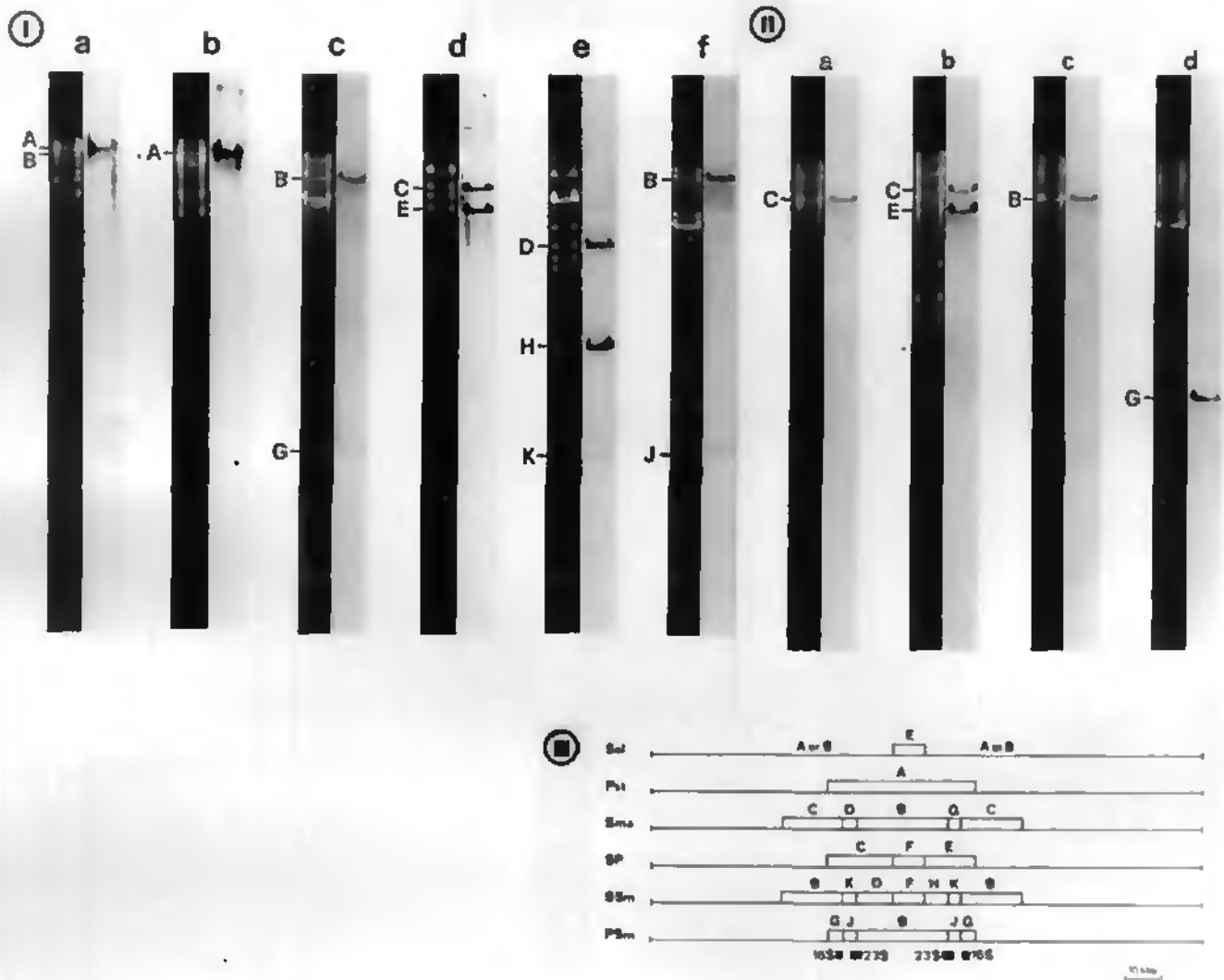
---



La figure 3.I montre les résultats des hybridations de cette sonde avec 6 profils de restriction de l'ADNct de soja. L'interprétation de ces données est représentée sur la figure 3.III pour chaque étape. Les résultats peuvent être résumés comme suit :

- 1) La sonde hybride avec Sal-A (58.8 kbp) et Sal-B (44.9 kbp) (figure 3.I,a) et avec un seul fragment Pst I, Pst-A (41.1 kbp) (profil b). Par conséquent, Pst-A doit être à cheval sur Sal-A et Sal-B.
- 2) La sonde hybride avec les fragments SP-C (18.1 kbp) et SP-E (14.1 kbp) (profil d). La longueur de ces fragments combinés est de 32.2 kbp, ce qui est environ 9 kbp plus court que la taille de Pst-A. Le seul fragment Sal-I pouvant compléter ce trou (et qui n'est pas coupé par Pst-I) est Sal-E (8.9 kbp). Ce fragment se trouve par conséquent entre Sal-A et Sal-B (fig.3.II).

Figure 3 : Localisation des ARNr 23S et 16S sur l'ADNct de soja



I : Hybridation d'une sonde portant les rDNA 23S et 5S avec des profils de restriction de l'ADNct de soja . Pour chaque digestion, les fluorographies (à gauche) et les autoradiographies (à droite) sont alignées. Sal I(a), Pst I(b), Sma I(c), Sal I/Pst I(d), Sal I/Sma I (e), Pst/Sma I(f). Les bandes qui s'hybrident sont marquées par les lettres correspondantes.

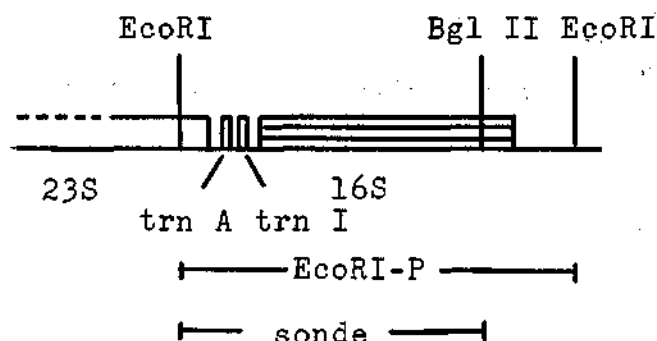
II : Hybridation d'une sonde portant les rDNA 16S avec des profils de restriction de l'ADNct de soja. Sma I(a), Sal I/Pst I(b), Sal I/Sma I (c), Pst I/Sma I(d).

III : Représentation schématique des résultats; Les fragments qui s'hybrident sont indiqués; les fragments de mobilité électrophorétique identique sont nommés de la même façon (ex. : Sma-G). Les rDNA 23S et 16S sont localisés et espacés approximativement, en assumant que la taille de ces gènes est de 2'800 et 1'500 bases, respectivement.

- 3) La sonde hybride avec le fragment Sma-B (25.1 kbp) (profil c). Ce fragment n'est pas coupé par Pst I, mais coupé par Sal I (table 1) et produit 2 fragments d'hybridation positive SSm-D (10.6 kbp) et SSm-H (5.6 kbp) (profil e). Le fragment Sma-B est donc composé en son centre de Sal-E, et dans ses extrémités de SSm-D et SSm-H (fig.3.III)
- 4) Sur le profil Sma I (c), une bande formée de 2 fragments (Sma-G<sub>1</sub> et G<sub>2</sub>) hybride faiblement. Nous pouvons placer ces deux fragments de chaque côté de Sma-B. Si cette position est correcte, des fragments de même taille devraient hybrider sur les profils SSm (e) et PSm (f), ce qui est le cas. Donc les fragments Sma-G<sub>1</sub> et G<sub>2</sub> sont équivalents aux fragments SSm-K<sub>1</sub> et K<sub>2</sub>, et aux fragments Psm-J<sub>1</sub> et J<sub>2</sub>.
- 5) La longueur de la partie de la sonde qui porte les gènes des ARNr est d'au moins 3 kbp. Selon les résultats ci-dessus, les gènes des ARNr 23S se trouvent à chaque extrémité de Sma-B, et à cheval sur les fragments Sma-G<sub>1</sub> et G<sub>2</sub> (fig 3.III.). Par conséquent il existe 2 régions portant l'information pour les ARNr 23S, une sur Sal-A et une sur Sal-B.
- 6) Les gènes pour les ARNr 5S se trouvent probablement à proximité de l'extrémité 3' des ARNr 23S, mais nous n'avons pas d'évidence pour cette localisation, puisque nous n'avons pas utilisé de sonde spécifique aux ARNr 5S.

La deuxième sonde utilisée est un fragment EcoRI-Bgl II (sous-fragment de EcoRI-P) du génome chloroplastique d'E.gracilis et portant environ 80% du gène de l'ARNr 16S (partie 3'), comme schématisé sur la figure 4 (Jenni et Stutz, 1978).

Figure 4 : Représentation schématique d'une partie de la région ribosomique chez E.gracilis



Cette sonde a été hybridée avec plusieurs profils de restriction de l'ADNct de soja (figure 3.II.). Nous pouvons résumer ces résultats ainsi :

- 1) Les fragments SP-C et SP-E hybrident avec l'ARNr 16S (profil b). Comme les ARNr 23S se trouvent également sur ces fragments, nous pouvons en déduire que ces gènes sont très proches l'un de l'autre, comme c'est le cas chez les autres plantes..
- 2) Sur le profil a, nous voyons que la double bande Sma-C hybride avec la sonde, alors que ni Sma-B, ni Sma-G<sub>1</sub> et G<sub>2</sub> ne s'hybrident avec l'ARNr 16S. Les fragments Sma-G<sub>1</sub> et C<sub>2</sub> ne peuvent donc être placés qu'à la suite des fragments Sma-G<sub>1</sub> et G<sub>2</sub>, respectivement. (figure 3.III)
- 3) Sur le profil SSm (c), une seule bande s'hybride avec cette sonde. Cette bande contient 3 fragments de mobilité électrophorétique identique (table 1). Deux de ces fragments, SSm-B<sub>1</sub> et B<sub>2</sub> doivent être équivalents à Sma-G<sub>1</sub> et C<sub>2</sub>.
- 4) Le profil PSm (figure 3.II,d) montre qu'une seule bande, formée de 2 fragments (PSm-G<sub>1</sub> et G<sub>2</sub>), s'hybride avec cette sonde. Ces deux fragments ne peuvent être ailleurs qu'à la suite de PSm-J<sub>1</sub> et J<sub>2</sub> (figure 3.III).

Ces résultats confirment l'existence de deux régions répétitives inversées, qui sont formées au minimum des fragments Sma-G<sub>1</sub> + C<sub>1</sub> et Sma-G<sub>2</sub> + C<sub>2</sub>.

- 5) La partie structurale du gène de l'ARNr 16S est supposée être d'environ 1'500 bp, comme chez les autres plantes (Schwarz et Kössel, 1980); ces gènes peuvent donc être placés comme représenté sur la figure 3.III.

## 2.2 Localisation du gène codant pour une protéine membranaire de Mr 32'000 (photogène) - Autres évidences de l'existence de deux régions répétitives inversées

---

Tous les génomes chloroplastiques étudiés jusqu'à maintenant portent à proximité d'une des régions inversées un gène codant pour une protéine thylacoïdale de Mr 32'000; ceci a notamment été démontré chez l'épinard (Driesel et al, 1980), le maïs (Link et al, 1978) et Spirodela oligorrhiza (Edelman et Reisfeld, 1980)

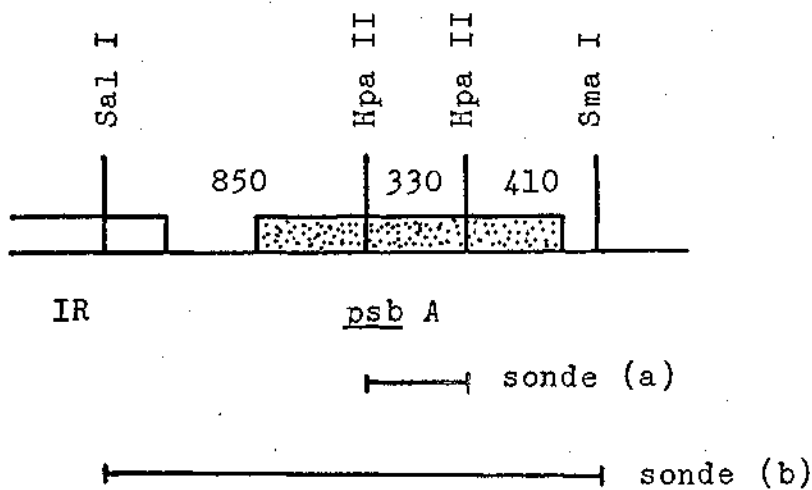
Cette protéine est caractérisée par sa synthèse très élevée dans les tissus verts de la plante et par son taux rapide de turnover (Edelman et Reisfeld, 1980). Elle joue un rôle dans le flux d'électrons provenant du photosystème II (Mattoo et al, 1981) et est responsable de la sensibilité de nombreuses plantes à certains herbicides (Pfister et al, 1981).

Pour la recherche de ce gène (appelé ici psb A) chez le soja, nous avons utilisé comme matériel de base, le plasmide pSocB511 (Driesel et al, 1980), qui est formé d'un fragment Bam HI de 5.4 kbp, inséré dans le pBR322, et qui porte le psb A de l'épinard. Ce plasmide nous a été aimablement fourni par le Dr. H. Bohnert.

A partir de ce plasmide, nous avons construit les sondes suivantes : (figure 5)

- 1) La sonde (a), spécifique du psb A, est un fragment Hpa II-Hpa II de 330 bp, portant la partie centrale du psb A
- 2) La sonde (b), qui est un fragment Sal I-Sma I de 1.6 kbp, contenant la totalité du psb A et environ 200 bp de l'extrémité d'une des régions répétitives inversées.

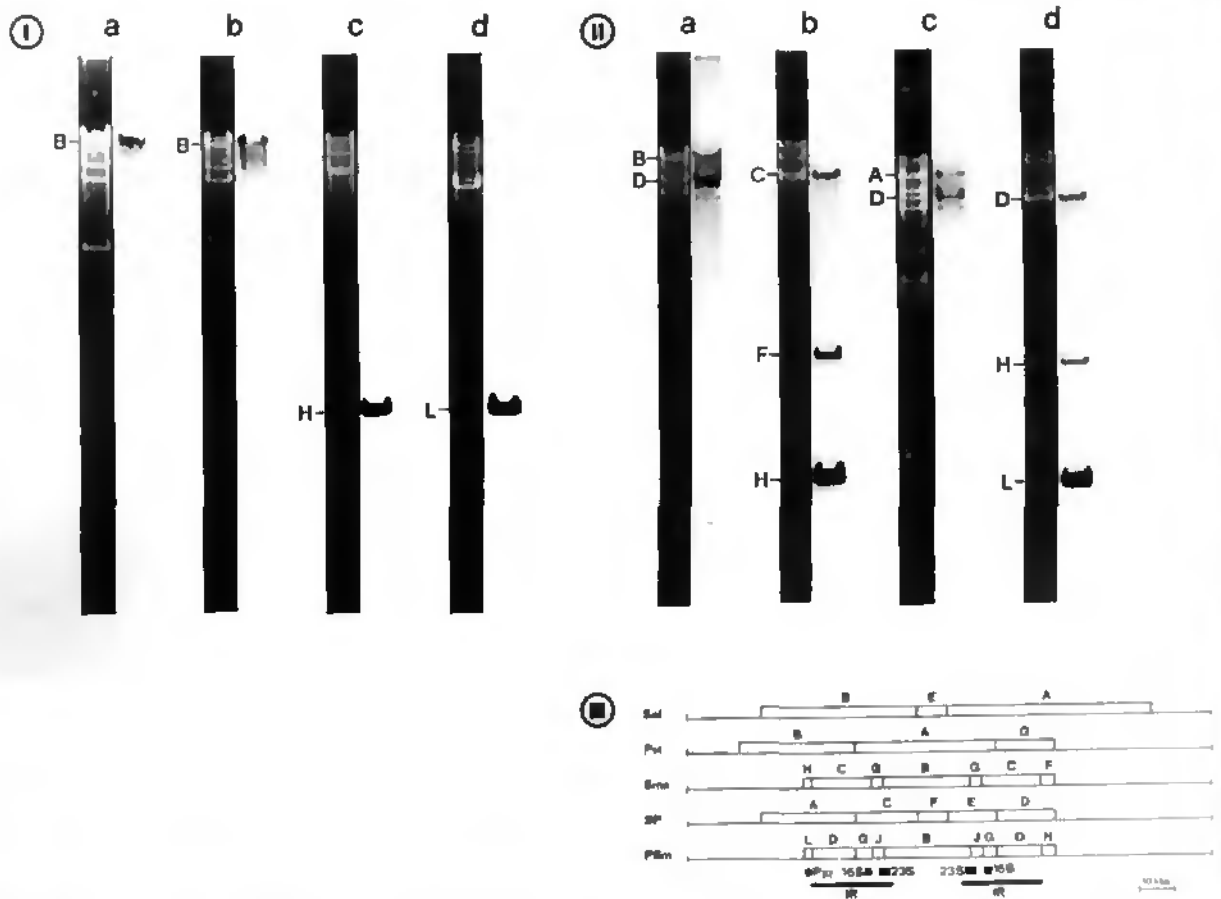
Figure 5 : Représentation schématique de la région portant le psb A chez l'épinard (Driesel et al, 1980)



Ces deux sondes devraient nous permettre d'une part de localiser le psb A, d'autre part de confirmer l'existence de 2 régions répétitives inversées, si l'on admet que ces séquences ( IR ) sont suffisamment conservées entre ces 2 espèces.

La figure 6 montre les résultats obtenus avec les sondes (a) et (b), alors que l'interprétation de ces données est représentée sur la figure 6.III.

Figure 6 : Localisation du psb A - Autres évidences de l'existence de 2 régions répétitives inversées



- I : Hybridation de la sonde(a) (voir texte) portant une partie du psb A d'épinard avec des profils de restriction de l'ADNct de soja. Pour le concept, voir figure 3. Sal I(a), Pst I(b), Sma I(c), Pst I/Sma I(d).
- II : Hybridation de la sonde(b) (voir texte) portant partiellement le psb A d'épinard et l'extrémité de l'IR adjacente avec des profils de restriction de l'ADNct de soja. Pst I(a), Sma I(b), Sal I/Pst I(c), Pst I/Sma I(d).
- III : Pour le concept, voir figure 3. Le psb A est ici dénommé P<sub>32</sub>; les régions répétitives inversées (IR) sont indiqués par des barres noires.

Nous pouvons résumer ces résultats comme suit :

- 1) La sonde (a) s'hybride avec Sal-B (figure 6.I,a), Pst-B (6.I,b) et Sma-H (6.I,c). Nous pouvons en déduire que ces 3 fragments ont des séquences en commun et peuvent être placés comme représenté sur la figure 6.III.
- 2) La sonde (b) s'hybride également avec Pst-B (6.II,a) et Sma-H (6.II,b), mais elle s'hybride en plus avec les fragments Pst-D (6.II,a), Sma-C<sub>1</sub>, C<sub>2</sub> et F (6.II.b). Nous pouvons conclure que ces hybridations supplémentaires sont dues à la présence des séquences provenant de la région répétitive inversée. Par conséquent, nous pouvons placer Pst-D à côté de Pst-A et Sma-F à côté de Sma-C<sub>2</sub>, comme représenté sur la figure 6.III.
- 3) Les profils PSm des figures 6.I et 6.II confirment ces résultats, c'est-à-dire que la sonde (a) s'hybride à un seul fragment (PSm-L), alors que la sonde (b) s'hybride avec 3 fragments supplémentaires, PSm-D<sub>1</sub>, D<sub>2</sub> et H, qui peuvent être placés comme représenté sur la figure 6.III.
- 4) Le profil c (SP) de la fig.6.II. montre que 2 bandes, SP-A et SP-D s'hybrident avec la sonde (b). SP-D correspond à Pst-D (table 1). SP-A porte donc le psb A et se trouve compris à l'intérieur de Pst-B (fig. 6.III.)
- 5) Nous pouvons donc confirmer l'existence de régions répétitives inversées et estimer leur longueur. Elles sont composées de Sma-G<sub>1</sub> + Sma-C<sub>1</sub> (Sma-G<sub>2</sub> + Sma-C<sub>2</sub>), plus une petite partie de Sma-H (Sma-F). La longueur d'une des régions inversées est donc d'environ 22 à 23 kbp.  
Le fragment Sma-H contient donc entièrement le psb A et l'extrémité d'une des régions inversées.

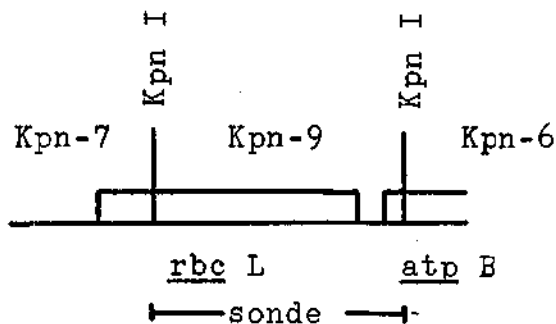
### 2.3 Localisation du gène codant pour la grande sous-unité de la ribulose 1,5-bisphosphate carboxylase (rbc L)

---

La sonde utilisée ici est le fragment Kpn-9 (1,6 kbp) provenant du génome chloroplastique d'épinard, et portant presque entièrement le rbc L, ainsi qu'une partie du gène de la sous-unité  $\beta$  de l'ATP-synthétase (atp B), (Westhoff et al., 1981), comme représenté sur la figure 7. Ce fragment nous a été aimablement fourni par le Prof. R.G. Herrmann.

Figure 7 : Représentation schématique de la région codant pour le rbc L et le atp B chez l'épinard.

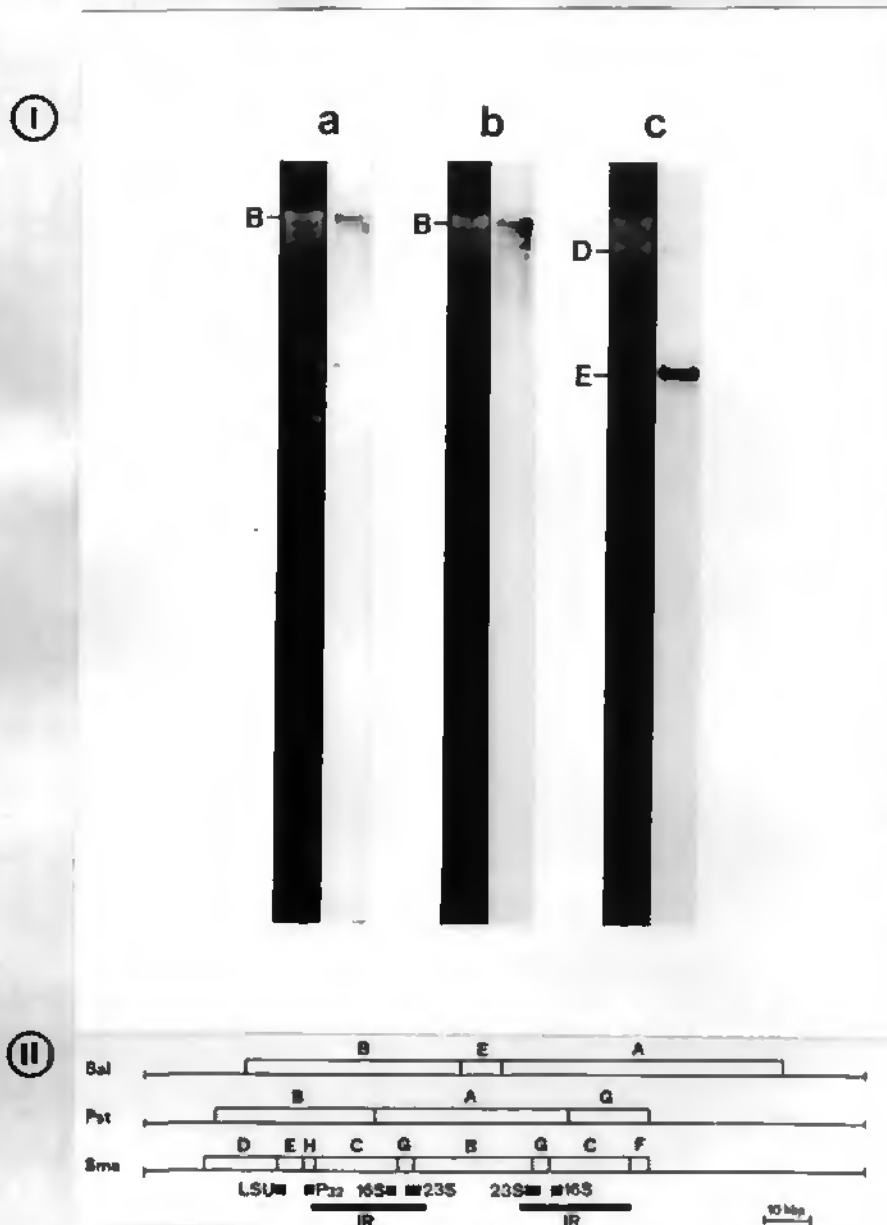
---



Les résultats de ces hybridations sont représentés sur la figure 8.1. La sonde hybride avec le fragment Sal-B (profil a), Pst-B (b) et les fragments Sma-D et Sma-E (c). Par conséquent, le rbc L est principalement localisé sur Sma-E (très fort signal), mais chevauche également Sma-D (faible signal). Nous ne pouvons cependant pas exclure que l'hybridation avec Sma-D soit due à l'atp B.

Nous pouvons ainsi placer les fragments liés Sma-E et Sma-D à côté de Sma-H (fig. 8.II). L'orientation correcte de ces fragments (Sma-D + Sma-E) est rendue possible parce que seul Sma-D est coupé par Sal I et Pst I, alors que Sma-E ne l'est pas (fig.1 et table 1).

Figure 8 : localisation du gène codant pour la LSU (rbc L)



I : Hybridation d'une sonde portant une partie du gène de la LSU chez l'épinard avec des profils de restriction de l'ADNct de soja. Sal I(a), Pst I(b), Sma I(c).

II : Pour le concept, voir figure 3. Le rbc L est appelé ici LSU.

Le rbc L est donc très proche (environ 5 kbp) du psb A, ce qui est assez exceptionnel, puisque dans la majorité des génomes chloroplastiques étudiés jusqu'à maintenant, ces deux gènes sont éloignés de plus de 50 kbp (voir discussion).

Cependant, Palmer et Thompson (1982) ont montré récemment que dans le cas du génome chloroplastique de deux Légumineuses, Vigna radiata et Pisum sativum, le rbc L est aussi très proche du psb A. Cet arrangement semble donc être spécifique des ADNct de Légumineuses.

#### 2.4 Construction d'une carte de restriction complète Sal-Pst-Sma

En consultant la table 1 et en comparant la liste des fragments avec ceux déjà localisés, nous voyons qu'il reste à placer :

- les fragments Sal-C et D
- " " Pst-C, E, F, G, H et I
- " " Sma-A, I et I'

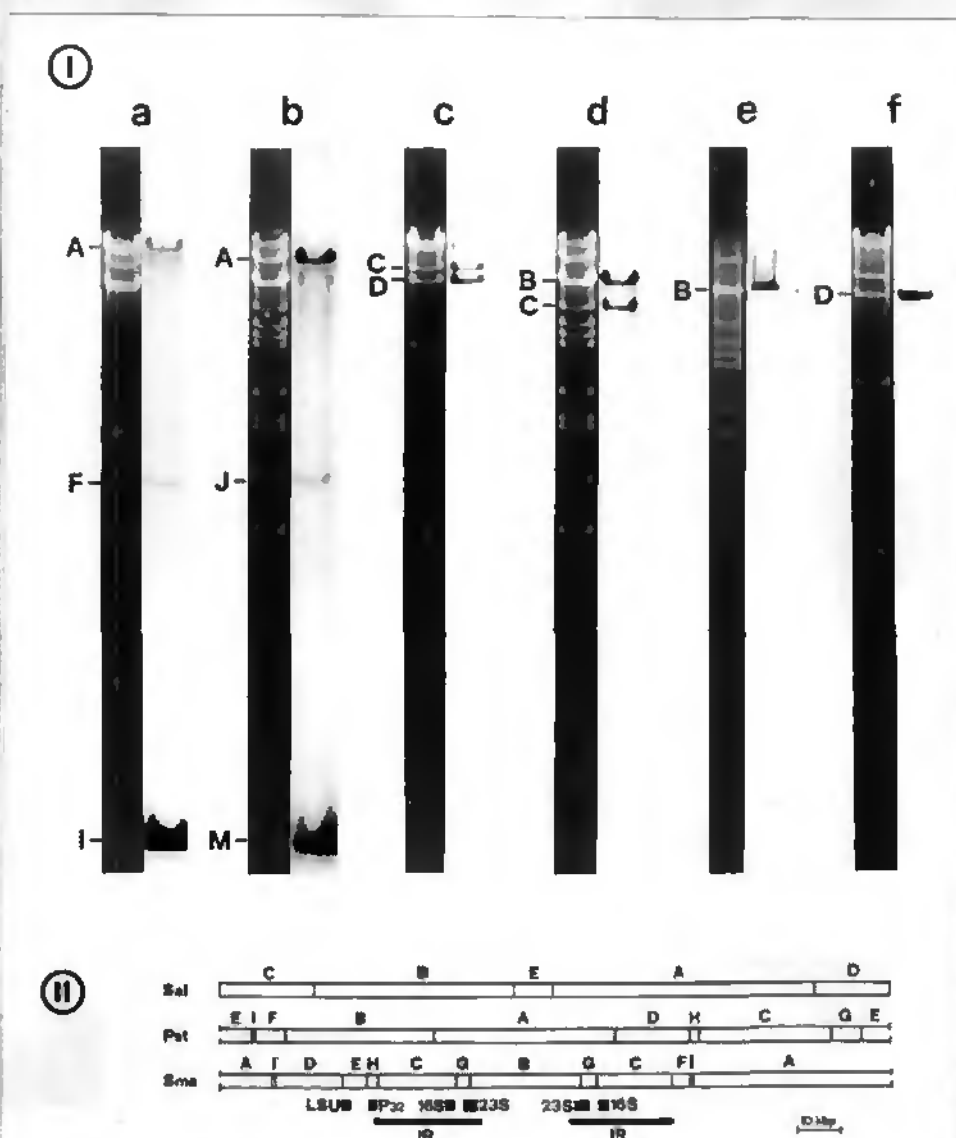
Afin de les localiser, certains de ces fragments (Pst-E, G, H et I) ont été clonés dans le pBR322 (chap. IV) et employés, soit directement dans des expériences de digestions mixtes, soit comme sonde radioactive dans des expériences d'hybridation.

Les expériences de digestions ont montré que Pst-G et I n'avaient pas de sites pour Sal I et Sma I, alors que Pst-E (13.7 kbp) est coupé par Sal I, donnant les sous-fragments SP-G<sub>1</sub> (7.1 kbp) et SP-H (6.6 kbp).

Pst-H est coupé deux fois par Sma I, produisant les sous-fragments Sma-I (0.9 kbp), PSm-N (0.9 kbp) et PSm-O (0.1 kbp). Ces informations sont nécessaires, mais elles ne nous permettent pas encore de placer ces fragments sur la carte physique.

La carte de restriction Sma I a pu être complétée en utilisant le fragment Pst-H comme sonde dans des expériences d'hybridation, comme représenté sur la figure 9.I.

Figure 9 : Construction d'une carte complète Sal I/Pst I/Sma I



I : Hybridation de certains fragments d'ADNct de soja clonés avec des profils de restriction de l'ADNct total de soja. Pst-H hybridé avec des digestions Sal I(a) et Sal I/Sma I(b); Pst-E hybridé avec des digestions Sal I(c) et Sal I/Sma I(d); Pst-I hybridé avec une digestion Sal I/Sma I(e); Pst-G hybridé avec une digestion Sal I(f).

II : Pour le concept, voir figure 3.

Pst-H s'hybride avec Sma-A, F et I (profil a), ainsi qu'avec SSm-A, J et M (b). Puisque Sma-F est déjà localisé, nous pouvons placer Pst-H à côté de Sma I, qui lui-même est suivi de Sma-A (fig.9.II.). Ces localisations sont en accord avec les observations faites ci-dessus, c'est-à-dire que Pst-H est coupé 2 fois par Sma I. La bande Sma-I contient 2 fragments de mobilité électrophorétique identique. Il reste par conséquent à localiser un dernier fragment Sma I, Sma-I', qui est placé par déduction entre Sma-A et Sma-D.

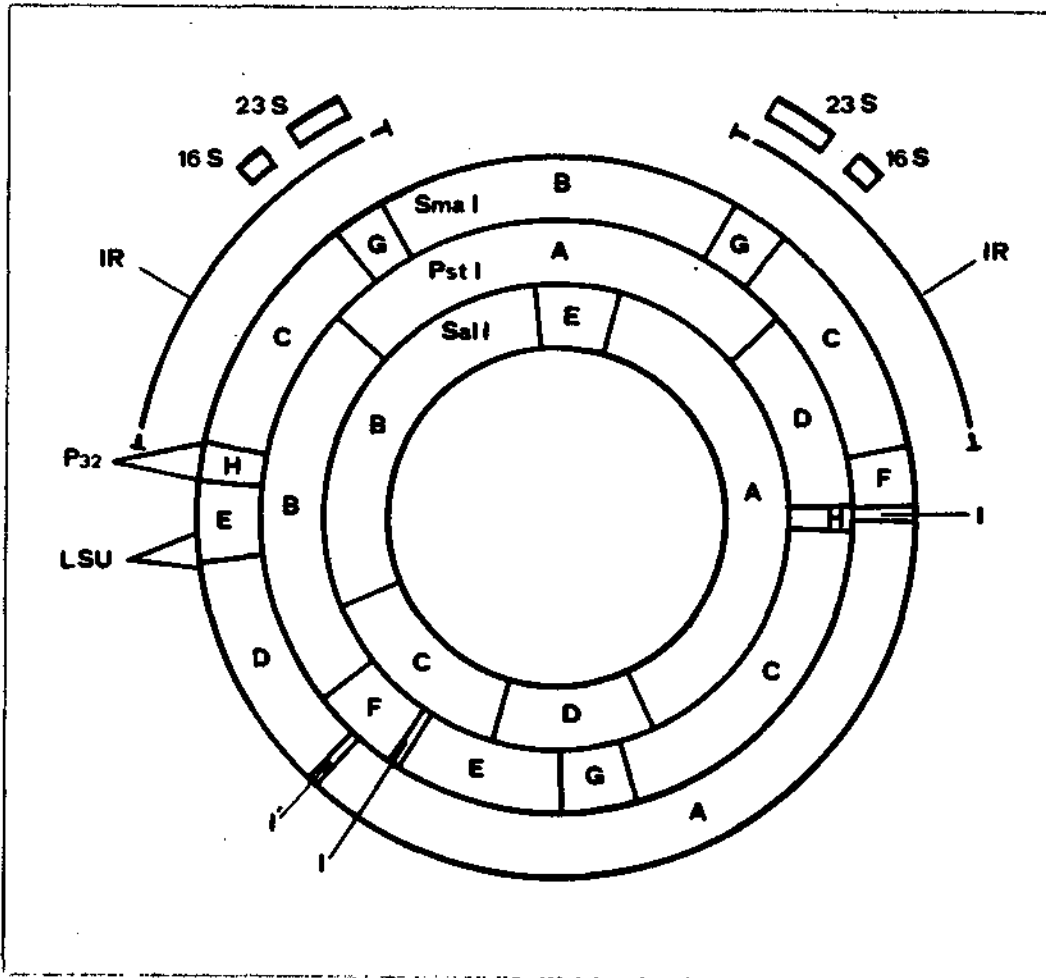
La carte Sal I peut maintenant être complétée. Sal-C est coupé par Sma I (table 1, fig.1), alors que Sal-D ne l'est pas. La seule façon de placer ces 2 fragments est celle représentée sur la fig.9.II c'est-à-dire Sal-D à côté de Sal-A et Sal-C à côté de Sal-B.

Il reste encore 5 fragments Pst I à localiser (Pst-C, E, F, G, I). Parmi ces fragments, seul Pst-F est coupé par Sma I (fig.1, table 1). Puisqu'il ne reste que 2 sites Sma I disponibles, éloignés l'un de l'autre de 0.9 kbp, Pst-F ne peut être placé qu'à côté de Pst-B (fig.9.II). Ceci confirme en même temps la position de Sma-I'.

Dans le but de placer Pst-E, qui est coupé par Sal I, nous l'avons hybridé avec un profil Sal I (fig.9.I,c). Pst-E s'hybride avec Sal-C et Sal-D, ce qui nous permet de le placer à cheval sur ces deux fragments (fig.9.II;). De plus, Pst-E s'hybride avec les fragments SSm-B et SSm-C (fig.9.I,d). Cette observation est nécessaire pour localiser le petit fragment Pst-I, qui a aussi été utilisé comme sonde : il s'hybride seulement avec SSm-C (fig.9.I,e). SSm-C contient des parties de Pst-F et Pst-E, et Pst-I doit par conséquent être placé entre Pst-F et Pst-E (fig.9.II.)

Nous avons finalement hybridé Pst-G contre un profil Sal I (fig.9.I,f). Pst-G réagit avec Sal-D et nous pouvons ainsi le placer à côté de Pst-E. Ceci laisse assez d'espace pour placer le grand fragment Pst-C entre Pst-H et Pst-G. Pst-C est coupé par Sal I, produisant les 2 sous-fragments SP-B et SP-L (table 1, fig.1).

Figure 10 : Carte de restriction circulaire Sal I/Pst I/Sma I de l'ADNct de soja



23 S : gènes codant pour les ARNr 23S  
 16 S : " " " " " 16S  
 P<sub>32</sub> : gène codant pour le photogène  
 LSU : gène codant pour la LSU  
 IR : régions répétitives inversées

Ainsi, tous les fragments Sal I, Pst I et Sma I sont placés sur le génome chloroplastique de soja, comme représenté sur la figure 10.

## 2.5 Localisation du gène codant pour le facteur d'élongation Tu (tuf A)

---

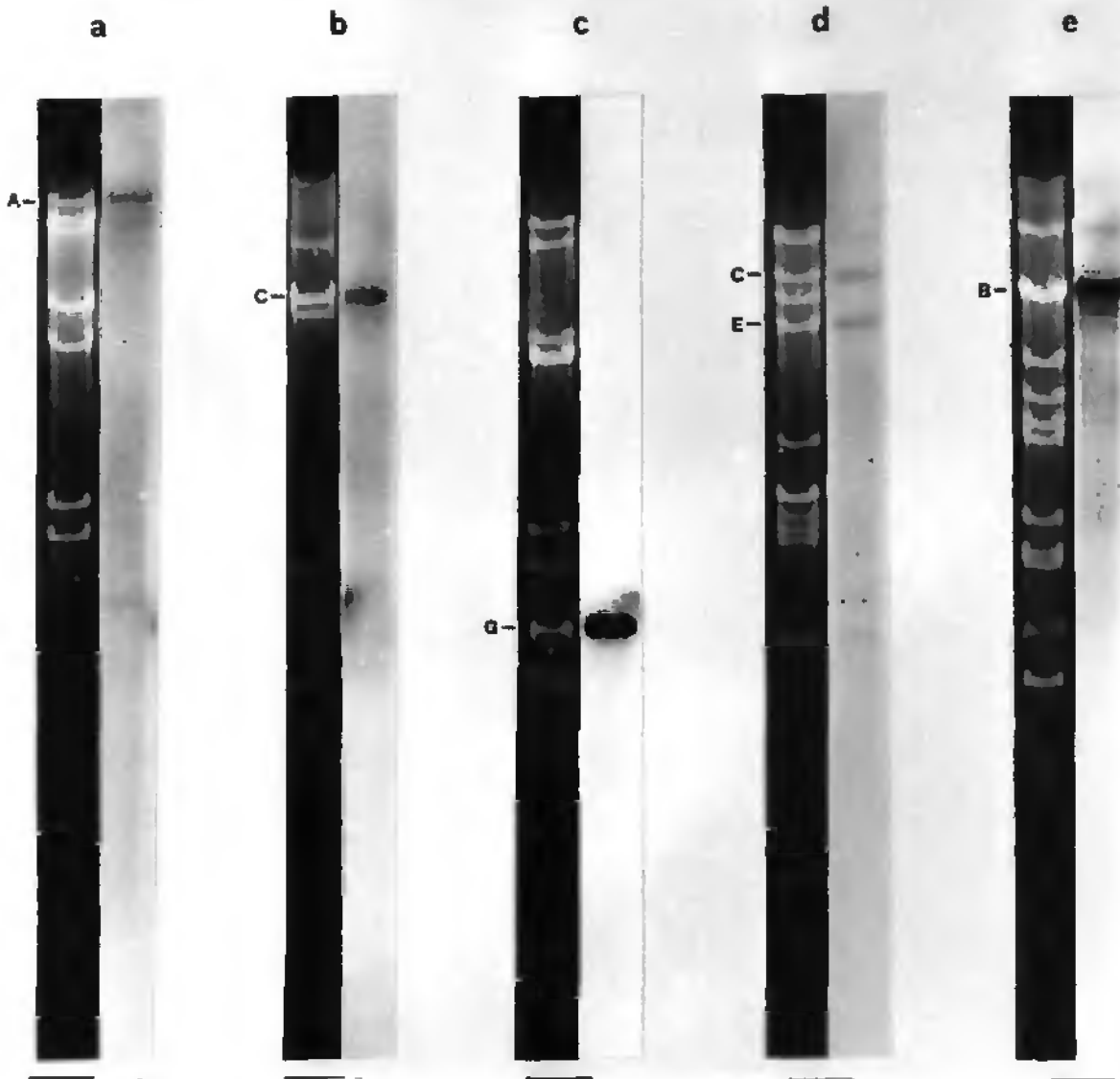
Le facteur d'élongation Tu est responsable de la fixation de l'aminoacyl-tRNA au ribosome durant la synthèse des protéines. Il a été démontré qu'au moins deux facteurs d'élongation chloroplastiques, dont le Tū, étaient codés au niveau du génome chloroplastique (Ciferri et al, 1979).

Plus récemment, le tuf A chloroplastique a été localisé chez Chlamydomonas reinhardtii (Watson et Surzycki, 1982) et chez Euglena gracilis (Passavant et al, 1983). Chez l'euglène, il se trouve entièrement sur le fragment EcoRI-N (3.0 kbp), et a la particularité d'être sous la forme d'un gène segmenté, c'est-à-dire divisé en parties codantes (exons) et non-codantes (introns) (Montandon et Stut 1983).

Nous avons utilisé ce fragment EcoRI comme sonde radioactive dans le but de localiser ce gène sur le génome chloroplastique de soja. Les résultats des hybridations de ce fragment avec plusieurs profils de restriction de l'ADNct de soja sont représentés sur la figure 11, et résumés sur la figure 12.

Le fragment EcoRI-N s'hybride avec le fragment Pst-A (fig.11.I,a), ainsi qu'avec les fragments Sma-C<sub>1</sub> et C<sub>2</sub> (11.I,b) et PSm-G<sub>1</sub> et G<sub>2</sub> (11.I,c). Ce gène est donc localisé à chaque extrémité du fragment Pst-A, c'est-à-dire à l'intérieur des fragments PSm-G<sub>1</sub> et PSm-G<sub>2</sub> (fig.12). Les résultats des hybridations obtenus avec les doubles digestions SP (11.I,d) et SSm (11.I,e) confirment cette localisation. Le tuf A est donc présent 2 fois sur l'ADNct de soja et se trouve à moins de 2 kbp du rDNA 16 S (fig.12)

Figure 11.I : Localisation du gène du facteur d'élongation Tu (tuf A)



Hybridation d'une sonde portant totalement le tuf A d'euglène avec des profils de restriction de l'ADNct de soja.  
Pst I(a), Sma I(b), Pst I/Sma I(c), Sal I/Pst I(d), Sal I/Sma I(a).

Figure 11.II : Localisation du gène du facteur d'élongation Tu (tuf A)

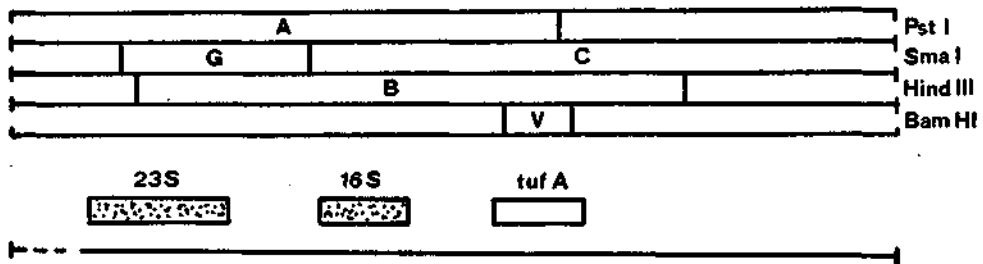


Hybridation d'une sonde portant totalement le tuf A d'euglène avec des profils de restriction de l'ADNct de soja. Dans les digestions simples, les fragments qui s'hybrident sont indiqués par les lettres correspondantes; dans les digestions mixtes, ils sont indiqués par une flèche. BamHI(a), BamHI/Pst I(b), Hind III(c) Hind/Sma I(d), Hind III/Pst I(e).

Grâce aux hybridations présentées sur la figure 11.II., nous pouvons localiser plus précisément ce gène, notamment parce qu'il est porté par les fragments Bam HI -V<sub>1</sub> et V<sub>2</sub> (1.2 kbp; 11.II,a), et que ceux-ci sont coupés par Pst I (11.II,b); le tuf A doit ainsi être à cheval sur les sites Pst I formant les extrémités de Pst-A (fig.12).

Les résultats du profil (c) nous montrent que les fragments Hind III-B<sub>1</sub> et B<sub>2</sub> (9.0 kbp) portent entièrement ces gènes. Les hybridations des profils (d) et (e) nous permettent de placer les fragments Hind-B<sub>1</sub> et B<sub>2</sub> par rapport aux fragments Sma-G<sub>1</sub>, G<sub>2</sub> et Pst-A, comme représenté sur la figure 12.

Figure 12 : Représentation schématique d'une des régions codant pour le tuf A chloroplastique de soja



Remarque : Seuls les sites Hind III et Bam HI connus sont notés sur ce schéma

Bien que les résultats ci-dessus démontrent clairement la localisation du tuf A chloroplastique chez le soja, il faut émettre les réserves suivantes :

Plusieurs autres fragments d'ADN ont été utilisés comme sonde pour localiser ce gène; ces fragments étaient :

- (a) Un fragment Sma I de 243 bp, codant pour la partie 5' et centrale du tuf A d'Escherichia coli; la sonde commence en position 388 et se termine en position 630 (Yokota et al, 1980).
- (b) Un fragment EcoRI d'environ 1.2 kbp, codant pour la partie 3' du tuf A de Chlamydomonas reinhardtii (Watson et Surzycki, 1982)
- (c) Un sous-fragment Hind III-EcoRI (1.4 kbp) provenant du fragment EcoRI-N d'E.gracilis, et portant la partie 3' de ce gène (Montandon et Stutz, 1983).

Aucun de ces fragments n'a donné de réponse positive lors d'hybridations avec différents profils de restriction de l'ADNct de soja. Cependant, des résultats récents (Montandon et Stutz, 1983) ont montré que la séquence nucléotidique de la région 3' de ce gène était assez peu conservée entre E.coli et E.gracilis, alors que la région 5' est plus constante.

Pour tenter d'expliquer nos résultats d'hybridation, nous pouvons émettre les hypothèses suivantes :

- 1) Ce gène est absent du génome chloroplastique du soja, et donc les hybridations positives du fragment EcoRI-N d'E.gracilis sont dues à une autre séquence portée par ce fragment.

2) Ce gène est présent sur l'ADNct de soja, et il se trouve dans les régions répétitives inversées, à côté des rDNA 16S.

Nous pouvons alors postuler que :

- les sondes (b) et (c) n'ont pas hybridé parce que la partie 3' du tuf A est très peu conservée entre d'une part G.max et C.rheinhardii, d'autre part G.max et E.gracilis
- il y a eu hybridation avec EcoRI-N total parce que la partie 5' de ce gène est mieux conservée entre G.max et E.gracilis
- la sonde (a) n'a pas hybridé parce que la partie 5' de ce gène diverge trop entre G.max et E.coli

Nous pensons que la troisième hypothèse est la meilleure, d'autant plus qu'il semble que chez l'épinard, le tuf A soit localisé dans les régions répétitives inversées (Ciferri, communication personnelle).

Des expériences sont actuellement en cours pour éclaircir ce problème. Quoiqu'il en soit, ces résultats confirment la carte de restriction déjà établie et nous permettent en plus de localiser deux fragments Hind III ( $B_1$  et  $B_2$ ) et deux fragments Bam HI ( $V_1$  et  $V_2$ ) (fig 12).

#### Correction de la carte de restriction établie

D'après des résultats récents de Pillay (communication personnelle), la bande Sma-I contient non pas 2, mais 3 fragments de mobilité électrophorétique identique; 2 de ces fragments se trouvent dans les IR, entre Sma-C et Sma-G, et le troisième fragment est localisé à l'intérieur de Pst-H. Sma-A est donc directement juxtaposé à Sma-D.

### 3. Conclusions

D'après notre propre expérience, il est difficile de purifier les chloroplastes de soja avec un bon rendement, ainsi que d'obtenir de l'ADNct intact (circulaire).

Les analyses effectuées avec des enzymes de restriction ont montré que l'ADNct isolé était régulièrement dégradé, ayant une taille moyenne de 50 à 60 kbp. Cependant ceci n'interfère pas avec l'analyse de ce génome, mais la taille réduite de l'ADNct (ainsi que sa contamination par de l'ADN nucléaire) contribuent à la formation d'un bruit de fond visible sur les gels.

Néanmoins, la construction d'une carte de restriction Sal I-Pst I-Sma I a pu être réalisée sans équivoque, et ceci grâce à l'utilisation de sondes d'ADN exogènes de fonction génétique définie, permettant une identification rapide des fragments d'ADN analogues sur le génome chloroplastique de soja. Cette approche s'est révélée très pratique et sera certainement plus fréquente à l'avenir, puisqu'il devient de plus en plus apparent que les parties structurales des gènes chloroplastiques possèdent un haut degré d'homologie de séquence (Fluhr et Edelman, 1981 ; Palmer et Thompson, 1982).

Dans la famille des Légumineuses, le soja, de par l'organisation générale de son génome chloroplastique, est très semblable à Vigna radiata (Palmer et Thompson, 1981a) et à Phaseolus vulgaris (Palmer et al., 1983). Par contre, il diffère passablement du génome chloroplastique de Vicia faba (Koller et Delius, 1980) et de Pisum sativum (Palmer et Thompson, 1981a).

Il est d'ailleurs intéressant de relier ces données biochimiques avec la taxonomie classique, puisque l'on remarque (voir table 2), que les espèces faisant partie de la tribu des Phaseolae ont toutes un ADNct de grande taille (150 kbp), possédant deux régions répétitives inversées, alors que les espèces de la tribu des Viciae ont un ADNct de taille réduite (120 kbp), sans région répétitive inversée.

Parmi les familles des Angiospermes, celle des Légumineuses ne semble pas être d'origine plus ancienne que les autres; nous pouvons donc admettre qu'il y a eu perte d'une des régions inversées dans la tribu des Viciées (Palmer et Thompson, 1982). Or il apparaît que les régions répétitives inversées stabilisent le génome chloroplastique, et que la perte d'une de ces régions accroît considérablement les réarrangements de séquences de ces génomes (Palmer et Thompson, 1982).

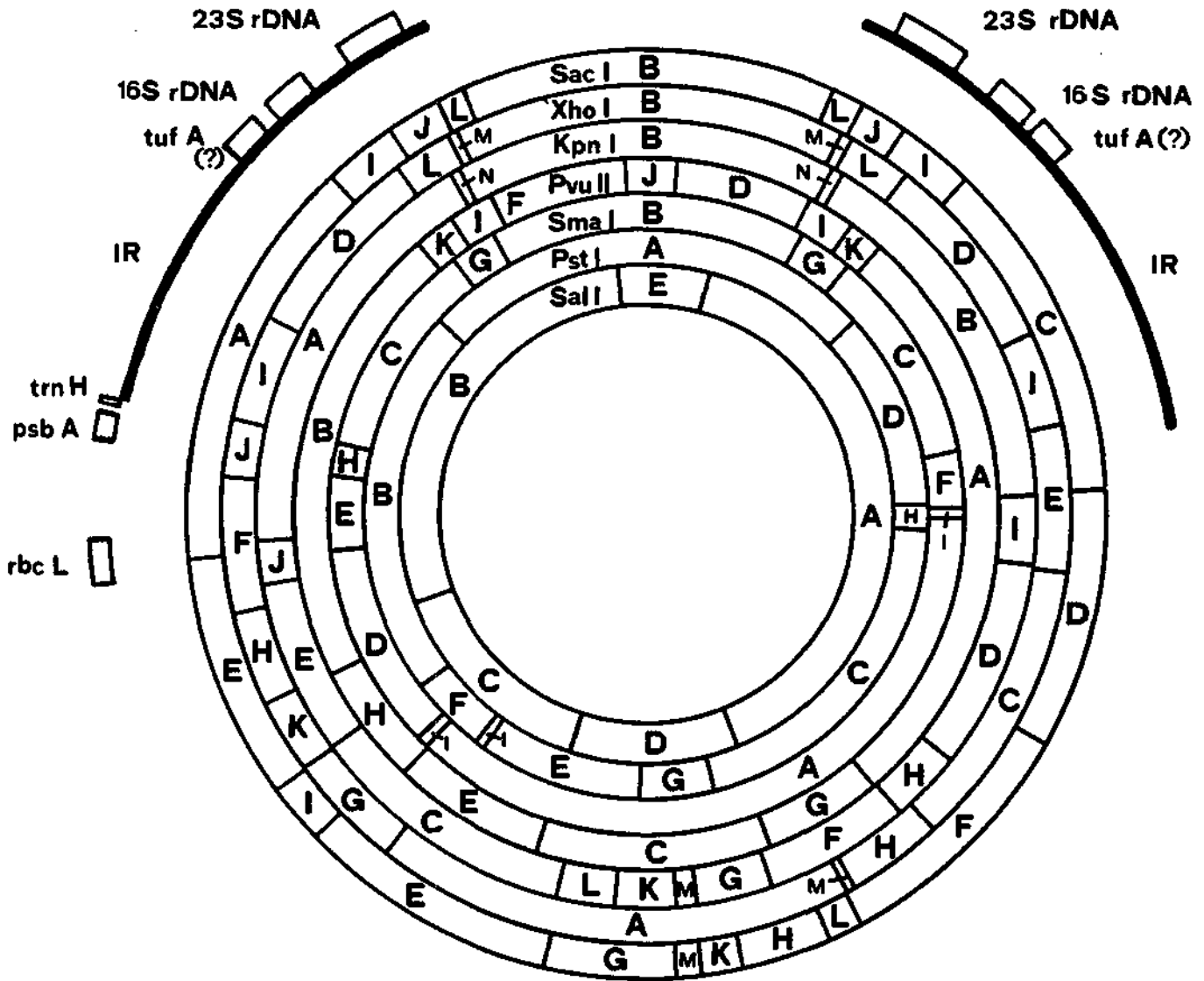
Table 2 : Relation entre la classification taxonomique de certaines espèces de Légumineuses et l'organisation de leur ADNct

| <u>Tribu</u> | <u>Sous-tribu</u> | <u>Espèce</u> | <u>Organisation de l'ADNct</u> |
|--------------|-------------------|---------------|--------------------------------|
| Phaseolae    | Phaseolinae       | P.vulgaris    | 150 kbp ; IR                   |
| "            | "                 | V.radiata     | 150 kbp ; IR                   |
| "            | Glycinae          | G.max         | 150 kbp ; IR                   |
| Viciae       | -                 | V.faba        | 120 kbp ; pas d'IR             |
| "            | -                 | P.sativum     | 120 kbp ; pas d'IR             |

Note : IR = région répétitive inversée

L'organisation structurale et génétique du génome chloroplastique des Légumineuses se distingue par le fait que le rbc L est très proche du psb A (moins de 5 kbp), alors que chez les autres espèces de plantes supérieures étudiées jusqu'à maintenant, comme le maïs (Bedbrook et al., 1979), l'épinard (Driesel et al., 1980), le tabac (Seyer et al., 1981), la moutarde (Link, 1981) et Spirodela oligorrhiza (Van Ee et al., 1982), le rbcL se trouve à plus de 50 kbp du psb A.

Figure 13 : Corrélation des cartes de restriction Sal I/Pst I/Sma I et Sac I/Xho I/Kpn I/Pvu II de l'ADNct de Glycine max.



trn H : ARNt<sup>His</sup> (sa localisation est expliquée au chap.V.)

Ce réarrangement de gènes provient de l'inversion d'un grand segment d'ADN (50 kbp) qui semble avoir eu lieu uniquement chez les Légumineuses, bien que de telles inversions, mais beaucoup plus petites, existent aussi chez le maïs (Palmer et Thompson, 1982).

Il serait intéressant d'étudier les conséquences du rapprochement du psb A et du rbc L chez les Légumineuses, et notamment au niveau de la transcription de ces deux gènes, surtout que ceux-ci sont très fortement transcrits (Mattoo et al, 1981).

Il faut encore mentionner qu'il a été démontré récemment (Palmer, 1983) que les séquences des régions uniques du génome chloroplastique de Phaseolus vulgaris pouvaient exister sous deux orientations différentes, par rapport aux régions répétitives. Cette bipolarité des séquences des régions uniques ne peut être démontrée que par l'emploi d'un enzyme de restriction qui ne coupe pas les régions répétitives, mais coupe asymétriquement les régions uniques. Chez Glycine max, l'enzyme Sal I a ces propriétés, mais comme il coupe la petite région unique 2 fois, une telle polarité est difficile à prouver, d'autant plus qu'il faudrait avoir à disposition de l'ADNct de très bonne qualité, pour être capable de séparer sur gel des fragments de très grande taille (50 à 60 kbp), et qui différencieraient les uns des autres de 3 à 4 kbp seulement. Nous ne pouvons cependant pas exclure l'existence d'un tel phénomène chez G.max.

Finalement, au moment où cette carte de restriction Sal I/Pst I/Sma I de l'ADNct de soja a été établie, une carte de restriction Sac I/Xho I/Kpn I/Pvu II de ce même génome a été publiée par Palmer et al (1983), où cependant seuls les gènes des ARN ribosomiques ont été localisés. Des analyses comparatives nous ont permis de corréler ces deux cartes, comme représenté sur la figure 13.

#### IV. TECHNIQUE DU CLONAGE - CREATION D'UNE BANQUE DE CLONES DE L'ADN CHLOROPLASTIQUE DE SOJA

---

##### A. INTRODUCTION

L'analyse détaillée d'un fragment d'ADN n'est possible que si l'on possède ce fragment à l'état pur (non contaminé par d'autres fragments) et en quantité suffisante; ceci est réalisable grâce à la technique du clonage.

Le principe général est le suivant : (figure 1)

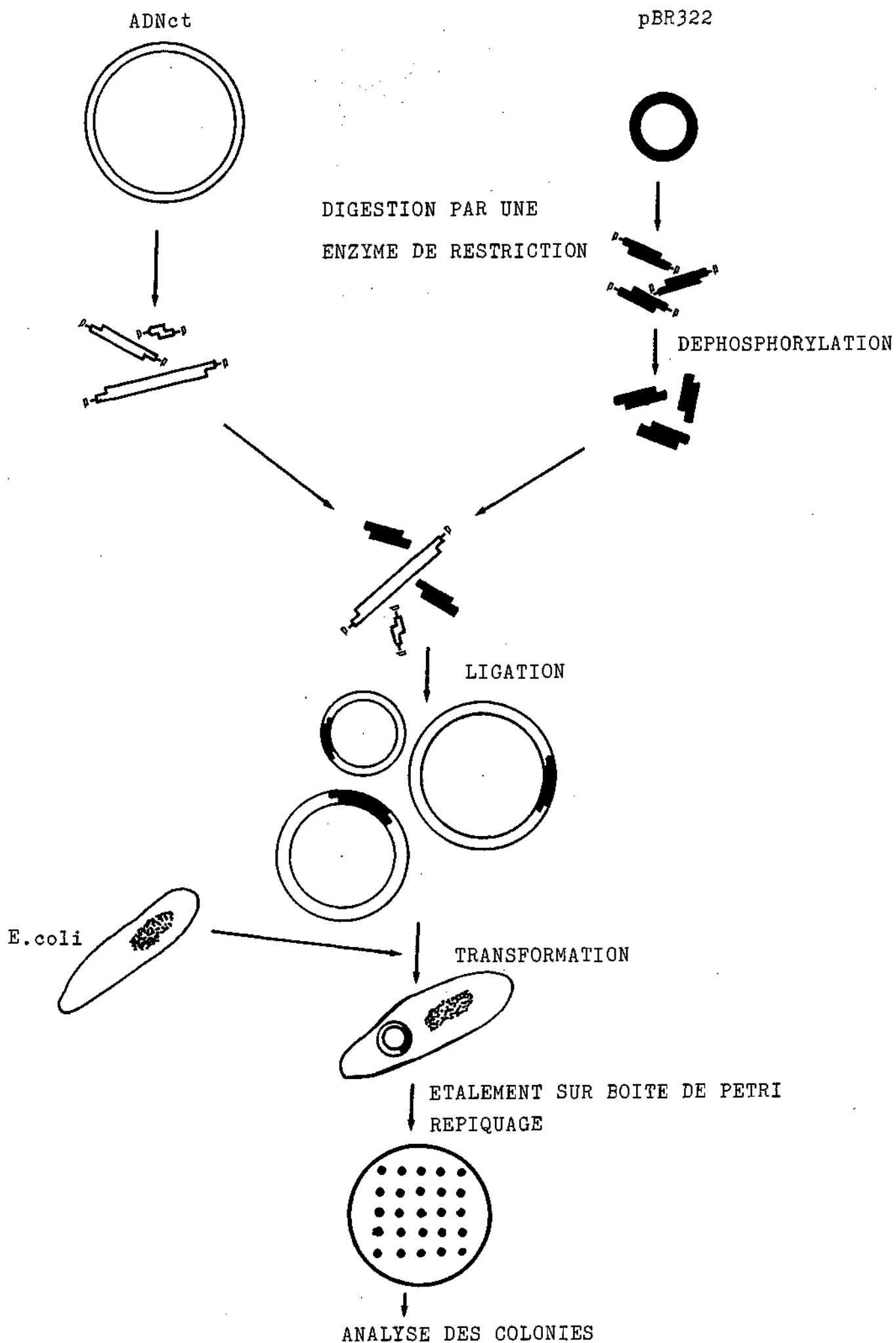
La plupart des bactéries possèdent, en plus de leur génome "chromosomal", une molécule d'ADN circulaire, appelée plasmide, se répliquant d'une manière autonome dans le cytoplasme de la bactérie.

La technique du clonage consiste à insérer un fragment de restriction d'ADN exogène dans un plasmide (qui sert de vecteur), puis de transformer la bactérie hôte avec cette molécule recombinée. Le clone ainsi formé peut être mis en culture, puis l'ADN plasmidial est isolé et purifié, et finalement le fragment étranger incorporé (= insertion) est séparé du vecteur.

N'importe quel plasmide ne peut cependant pas être utilisé comme vecteur. Il doit en effet avoir 3 caractéristiques essentielles :

- posséder au moins un site unique de restriction pour une enzyme
- pour être capable de sélectionner les bactéries transformées, le vecteur doit posséder au moins un marqueur, c'est-à-dire un gène codant pour un facteur quelconque qui n'est pas présent dans la bactérie sans plasmide. Ce facteur est souvent un gène qui confère à la bactérie la résistance à un antibiotique.
- être de taille raisonnable et être capable de se répliquer très rapidement.

Figure 1 : Représentation schématique des principales étapes du clonage dans le pBR322

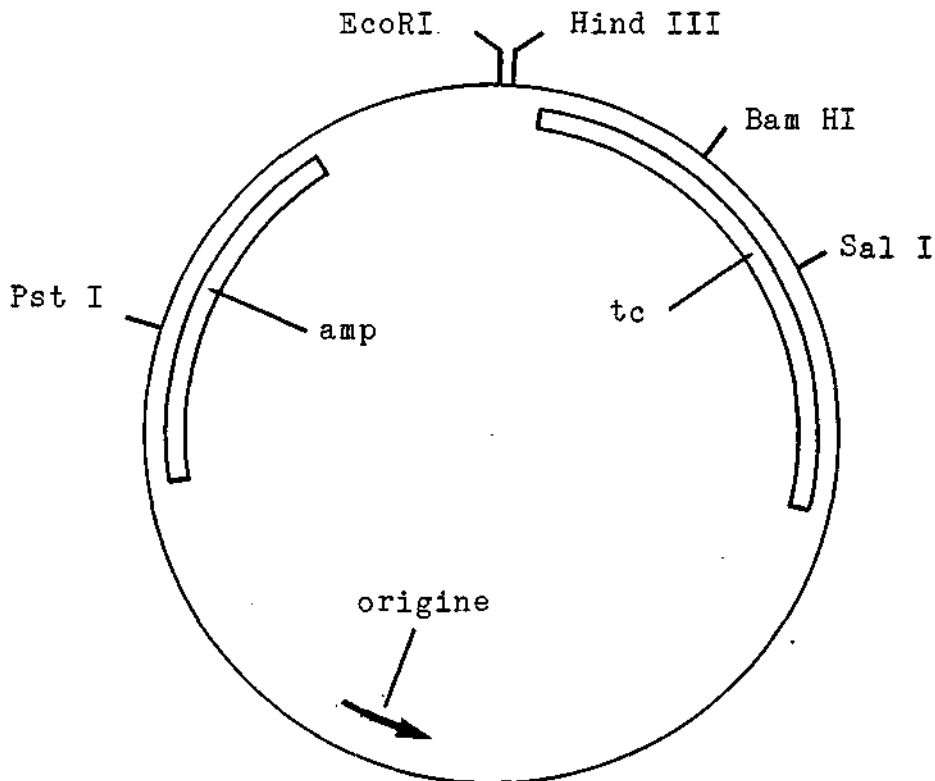


## B. MATERIEL ET METHODES

### 1. Matériel biologique

La bactérie hôte est une souche dérivée d'Escherichia coli K12, dénommée HB-101 (Boyer, 1969). Le plasmide utilisé comme vecteur est le pBR322, qui possède deux gènes de résistance à des antibiotiques (ampicilline = amp et tétracycline = tc), ainsi que plusieurs sites uniques de restriction, comme représenté sur la figure 2 (Sutcliffe, 1978).

Figure 2 : Représentation schématique de l'organisation génétique du pBR322



Taille : 4362 bp

## 2. Clonage

Le pBR322 a été préparé selon la méthode décrite par Ullrich et al, (1977). La ligation du vecteur avec des fragments d'ADNct a lieu pendant 12 à 14 h à 0°C sous les conditions recommandées par les fournisseurs de la T<sub>4</sub>-DNA ligase (Boehringer, Mannheim ou New England Biolabs, Beverly). La transformation de la bactérie hôte a lieu selon Mandel et Higa (1970).

## 3. Analyse des colonies

Après repiquage selon une grille d'identification, les colonies sont analysées selon deux méthodes :

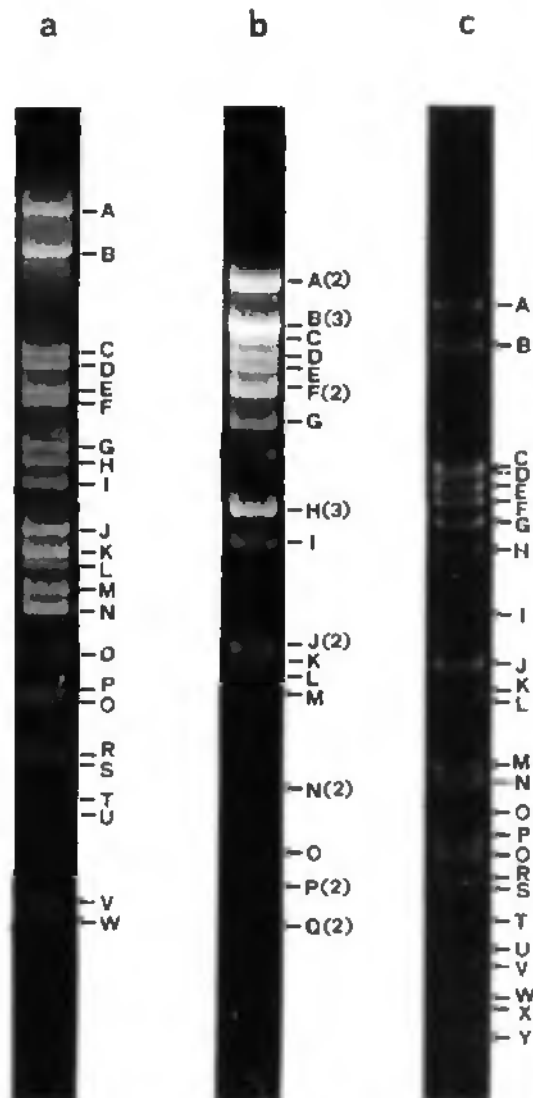
- celle dite de lyse alcaline (Maniatis et al, 1982), dérivée de celle de Birnboim et Doly(1979).
- une méthode modifiée de celle décrite par Grunstein et Hogness (1975)

### 3.1 Méthode modifiée de Grunstein et Hogness

Les bactéries sont repiquées sur des filtres Millipore (HCWG 047S1, 0.7  $\mu$ m) stériles placés dans des boîtes de Petri sur du milieu de culture contenant un antibiotique (20  $\mu$ g/ml). Les boîtes sont incubées 18 à 24 h à 37°C. Les bactéries sont lysées et l'ADN dénaturé en déposant les filtres sur du papier buvard imbibé de NaOH 0.5 N pendant 10 min. Les filtres sont ensuite transférés sur du papier buvard imbibé avec une solution de Tris-HCl 1 M, pH 7.5 (5 min), puis avec une solution de Tris-HCl 0.5 M, NaCl 0.5 M (10 min).

Finalement, les filtres sont déposés sur du papier buvard imprégné de 1 x SSC et recouverts d'une solution de protéinase K (2 mg/ml dans 1 x SSC). Après 15 min, les filtres sont immergés dans de l'éthanol (1 x), puis dans du chloroforme (3 x) et enfin dans une solution de NaCl 0.3 M. Les filtres sont séchés à l'air et cuits à 80°C sous vide pendant 2 h. L'analyse de l'ADN fixé sur le filtre a lieu selon la technique d'hybridation décrite au chapitre III.

Figure 3 : Profil électrophorétique de l'ADNct de soja obtenu après digestion avec BamHI, Hind III et EcoRI.



Les nombres entre parenthèses indiquent les bandes contenant 2 ou 3 fragments de mobilité électrophorétique identique sous les conditions utilisées. BamHI(a), Hind III(b), EcoRI(c).

#### 4. Isolement des plasmides

Les plasmides sont isolés selon une méthode tirée d'un cours SKMB, de Billeter et Weber (1978). Cette méthode est décrite en détail dans la thèse de Rutti (1982).

### C. RESULTATS ET DISCUSSION

#### 1. Choix des enzymes de restriction pour le clonage de fragments d'ADNct de soja

Nous avons utilisé la technique du clonage dans le but principal de cloner des fragments d'ADNct porteurs d'information génétique connue, comme par exemple le rbc L ou le psb A.

Cependant, au niveau du génome chloroplastique de soja, ces gènes se trouvent soit sur des fragments Sal I ou Pst I trop grands pour être clonés dans le pBR322, soit sur des fragments Sma I non clonables directement dans ce plasmide.

Si nous voulons utiliser le pBR322 comme vecteur, nous devons avoir à disposition des fragments de restriction Hind III, Bam HI ou Eco RI (un seul site de reconnaissance dans le pBR322 pour chacun de ces enzymes).

Nous avons digéré l'ADNct de soja par ces 3 enzymes, et les profils de restriction obtenus sont représentés sur la figure 3. Nous avons choisi de cloner les fragments Bam HI et Hind III pour les raisons suivantes :

- EcoRI coupe trop souvent l'ADNct de soja, produisant ainsi des fragments de trop petite taille ; de plus, le site EcoRI du pBR322 ne se trouve pas dans un gène de résistance à un antibiotique, d'où la difficulté de sélectionner les bactéries recombinantes.

- les enzymes BamHI et Hind III coupent respectivement l'ADNct de soja >35 fois et >25 fois, avec une taille maximum de fragment de 20.0 kbp pour BamHI et 11.0 kbp pour Hind III (figure 3). Tous les fragments BamHI et Hind III sont donc théoriquement clonables dans le pBR322.
- les sites BamHI et Hind III se trouvent dans le gène de résistance à la tétracycline (Hind III : dans la région de régulation de ce gène) (figure 2).

## 2. Clonage de fragments d'ADNct porteurs du psb A

Pour localiser le psb A sur les profils de restriction BamHI et Hind III, ainsi que détecter les colonies recombinantes positives, nous avons utilisé les techniques de marquage et d'hybridation décrites au chapitre III. Nous avons utilisé comme matériel de départ le plasmide pSocB511 qui contient une insertion BamHI de 5.4 kbp portant le psb A de l'épinard (Driesel *et al*, 1980). A partir de ce plasmide, nous avons construit les 2 sondes suivantes : (figure 4)

Figure 4 : Représentation schématique de la région codant pour le psb A chez l'épinard

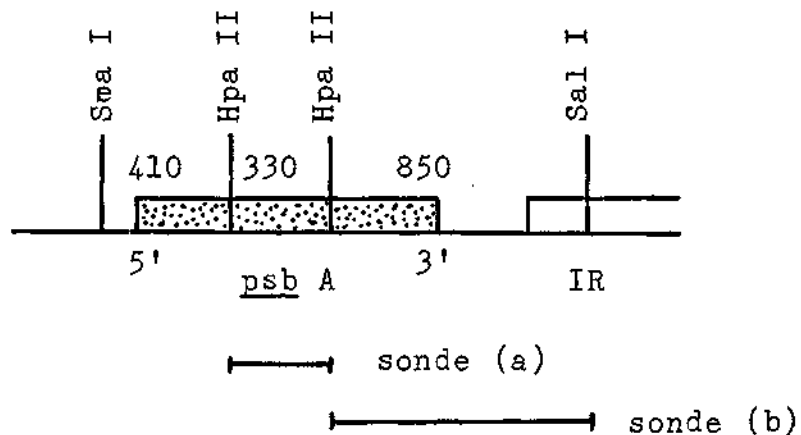
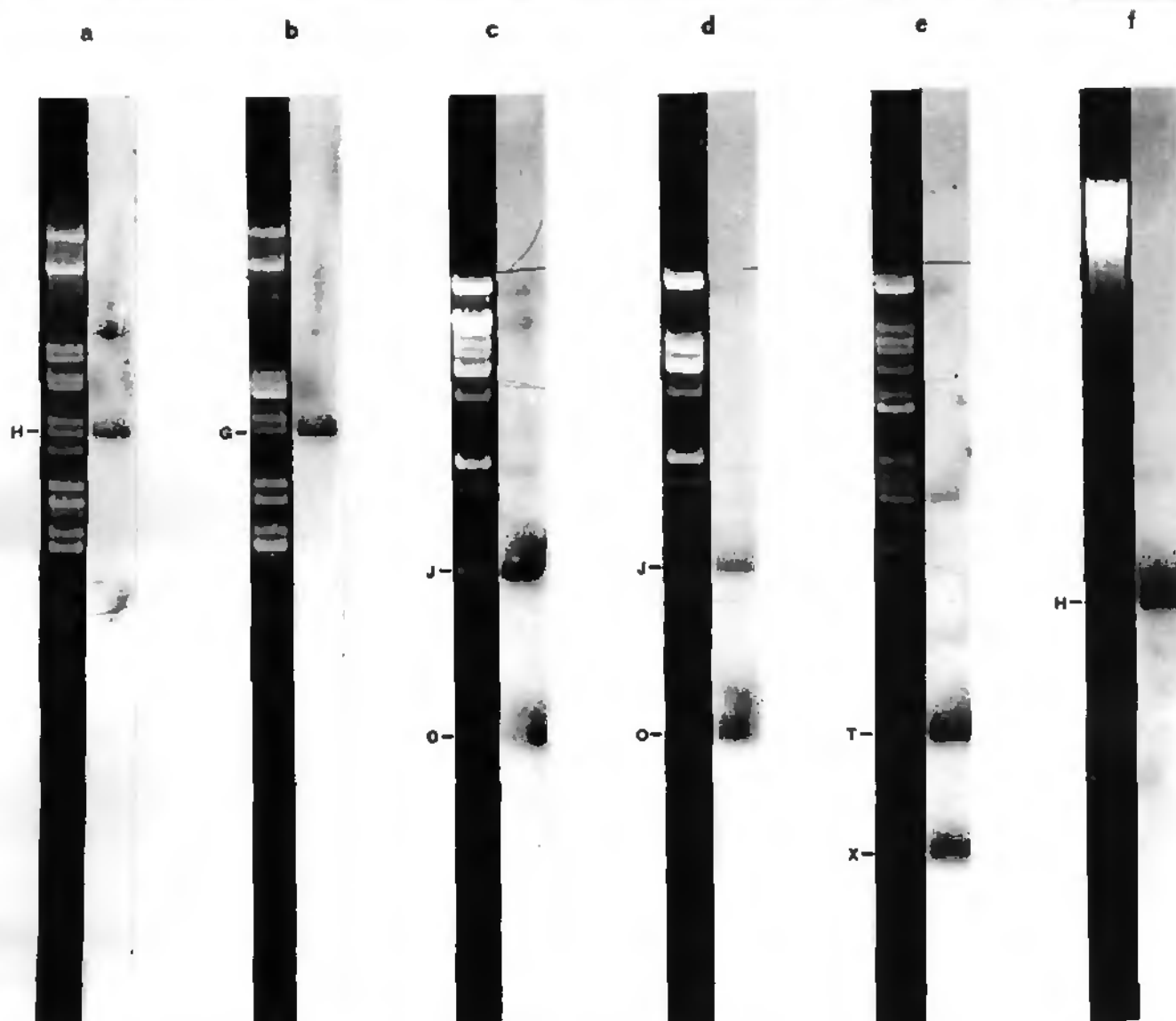


Figure 5 : Hybridation d'une sonde d'ADN portant une partie du psb A d'épinard avec des profils de restriction de l'ADNct de soja



La sonde utilisée est la sonde (a) (voir texte).

BamHI(a), BamHI/Pst I(b), Hind III(c), Hind III/Pst I(d), Hind III/Sma I (e), Sma I(f).

- la sonde (a) est un fragment Hpa II - Hpa II de 330 bp, se trouvant à l'intérieur du gène
- la sonde (b) est un fragment Sal I - Hpa II de 850 bp, qui porte la partie 3' du psb A, ainsi que 200 bp environ de l'extrémité d'une des régions inversées

Les résultats des hybridations de la sonde (a) sont représentés sur la figure 5 et schématisés sur la figure 6. Nous pouvons résumer ces résultats ainsi :

- le psb A se trouve d'une part entièrement compris sur les fragments Bam-H (5.2 kbp) (profil a) et Sma-H (2.4 kbp) (profil f), d'autre part à cheval sur les fragments Hind-J (2.8 kbp) et Hind-O (1.4 kbp) (profil c)
- les profils b et d montrent que Bam-H, Hind-J et Hind-O ne sont pas coupés par Pst I
- sur le profil e, nous voyons que le fragment Hind-O n'est pas coupé par Sma I, alors que Hind-J l'est et produit 2 sous-fragments Hind/Sma de 2.0 et 0.8 kbp; seul le sous-fragment de 0.8 kbp s'hybrise avec le psb A.

Figure 6 : Représentation schématique de la région de l'ADNct de soja portant le psb A

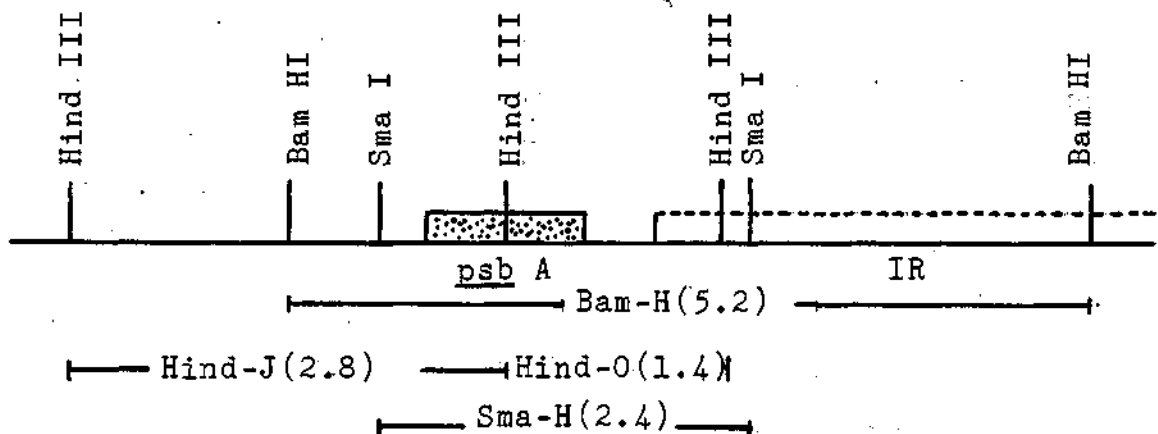
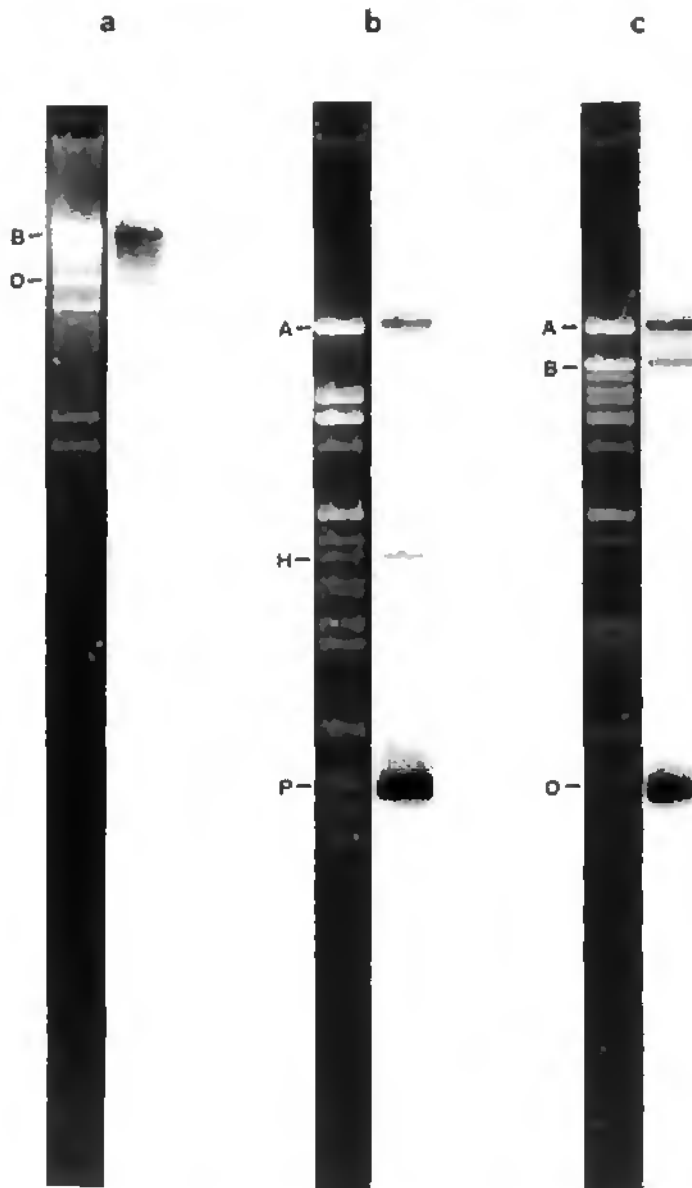


Figure 7 : Hybridation d'une sonde d'ADN portant partiellement le psb A d'épinard et l'extrémité de l'IR adjacente avec des profils de restriction de l'ADNct de soja



La sonde utilisée est la sonde (b) (voir texte).  
Pst I(a), Hind III/Pst I(b), Hind III(c).

Pour confirmer ces résultats, la sonde (b) a été hybridée avec les profils de restriction Pst I, Hind/Pst et Hind III. Les résultats de ces hybridations sont représentés sur la figure 7.

Nous pouvons en tirer les conclusions suivantes :

- le profil c de la figure 7 montre que :
  - 1) la sonde (b) s'hybride avec le fragment Hind-0, comme c'était le cas de la sonde (a), mais ne s'hybride pas avec Hind-J. Ceci confirme les résultats obtenus avec la sonde (a), et nous permet de dire que l'extrémité 3' du gène se trouve uniquement sur le fragment Hind-0. Sachant que la taille de ce gène est d'environ 1'200 bp (Palmer et al, 1982), nous pouvons placer le site Hind III pratiquement au milieu du gène (figure 6).
  - 2) deux bandes supplémentaires s'hybrident avec la sonde (b); ces bandes correspondent au double fragment Hind-A (11.0 kbp) et au triple fragment Hind-B (9.0 kbp). Comme la sonde (b) contient également environ 200 bp de l'extrémité d'une des régions répétitives inversées, nous pouvons conclure que les hybridations des fragments Hind-A et Hind-B sont dues à cette séquence.
- le profil a (figure 7) confirme ces résultats, à savoir que Pst-B et Pst-D s'hybrident avec la sonde (b); or seul Pst-B s'hybride avec le psb A (voir chap. III pour rappel)
- le profil b (figure 7) nous montre que la double bande Hind-A n'est pas coupée par Pst I, alors qu'au contraire le fragment Hind-B qui s'hybride l'est; seul un des sous-fragments Hind/Pst (Hind/Pst H, 4.0 kbp) s'hybride avec la sonde (b).

La seule façon d'expliquer ces résultats est de supposer qu'un des fragments Hind-A est juxtaposé au fragment Hind-0, alors que l'autre fragment Hind-A est juxtaposé à un des fragments Hind-B; comme tous ces fragments ont hybridé avec une des extrémités d'une des régions répétitives inversées, il est logique de placer les fragments liés Hind-A + Hind-B sur la région répétitive inversée opposée à celle qui est proche du psb A (figure 9).

Par commodité, nous appellerons dorénavant la région répétitive inversée se trouvant à côté du psb A, IR-1, alors que celle qui lui est opposée sera nommée IR-2. Les doubles fragments portant totalement des séquences des régions répétitives seront appelés  $A_1$  et  $A_2$ , etc...

Nous savons maintenant que le fragment Bam-H (5.2 kbp) porte entièrement l'information du psb A. Nous avons alors cloné la totalité des fragments de restriction BamHI de l'ADNct de soja. Plusieurs centaines de colonies positives ont été testées par la méthode de Grunstein et Hogness (1975); les sondes (a) et (b) ont été utilisées pour cette recherche.

Or les résultats ont toujours été négatifs avec la sonde (a), alors qu'avec la sonde (b), nous avons uniquement sélectionné le fragment Bam-C (8.0 kbp). Ce dernier ne peut que se trouver à cheval sur l'une des extrémités de l'IR-2.

Nous avons effectué plusieurs fois le clonage de la totalité des fragments BamHI de l'ADNct de soja. Nous avons également élué spécifiquement le fragment Bam-H à partir d'un gel d'agarose à bas point de fusion (Herrmann et al, 1980). Les résultats ont également toujours été négatifs.

Ceci n'est pas surprenant dans la mesure où de tels cas ont été décrits dans la littérature, et notamment chez le pois; chez cette espèce, Palmer et Thompson (1981b) ont été incapables de cloner un fragment Pst I de 12.2 kbp. Ces auteurs ont suggéré que certaines séquences d'ADN devaient posséder des propriétés qui les rendent réfractaires au processus du clonage.

Nous avons donc été contraints de cloner le psb A par l'intermédiaire des fragments de restriction Hind III.

Comme pour les fragments BamHI, tous les fragments Hind III de l'ADNct de soja ont été clonés dans le pBR322, et nous avons testé plusieurs centaines de colonies par la méthode de Grunstein et Hogness (1975).

Les sondes (a) et (b) ont été utilisées pour cette recherche et nous avons pu sélectionner un certain nombre de colonies positives.

L'analyse de ces colonies par la méthode modifiée de Birnboïm et Doly (1979) nous a permis de sélectionner 4 plasmides recombinants qui possédaient un ou plusieurs des fragments d'ADN recherchés.

Ces plasmides sont les suivants :

pGmcH 16 : Hind-A (11.0), Hind-J (2.8)  
pGmcH 179 : Hind-J (2.8)  
pGmcH 209 : Hind-B' (9.0), Hind-O (1.4), Hind-P (1.3)  
pGmcH 69 : Hind-A (11.0), Hind-? (~0.3)

Dans le but de confirmer que les fragments Hind-J des plasmides pGmcH 16 et pGmcH 179, ainsi que le fragment Hind-O du plasmide pGmcH 209 sont bien ceux recherchés, des profils de restriction de ces plasmides ont été hybridés contre la sonde (a). Les résultats ont démontré que seul le fragment Hind-J du pGmcH 179 et le fragment Hind-O du pGmcH 209 portent le psb A.

Le fragment Hind-O du pGmcH 209 a été isolé et sous-cloné dans le pBR322. Le nouveau plasmide recombinant ainsi créé est le pGmcH 133.

La région de l'ADNct de soja portant l'information génétique pour le photogène a ainsi été clonée sous la forme de 2 fragments Hind III, Hind-J (pGmcH 179) et Hind-O (pGmcH 133).

### 3. Création d'une banque de clones de l'ADNct de soja

Tout en recherchant les fragments Hind III et BamHI porteurs du psb A, nous avons également isolé et localisé d'autres fragments de restriction, et nous avons pu ainsi commencer la formation d'une banque de clones de l'ADNct de soja.

### 3.1 Fragments Hind III

Seuls 4 fragments Hind III clonés ont été localisés sur la carte de restriction de l'ADNct de soja; il s'agit des fragments Hind-A<sub>1</sub> et A<sub>2</sub> (11.0 kbp), Hind-B<sub>1</sub> (9.0 kbp), Hind-J (2.8 kbp) et Hind-O (1.4 kbp). Leur localisation est représentée schématiquement sur la figure 9. Les deux fragments Hind-B<sub>1</sub> et B<sub>2</sub>, porteurs de l'information génétique de l'EF-Tu (voir chap. III), n'ont pas pu être clonés, mais sont cependant localisés (figure 9).

### 3.2 Fragments BamHI

Plusieurs fragments BamHI ont été clonés dans le pBR322; il s'agit de :

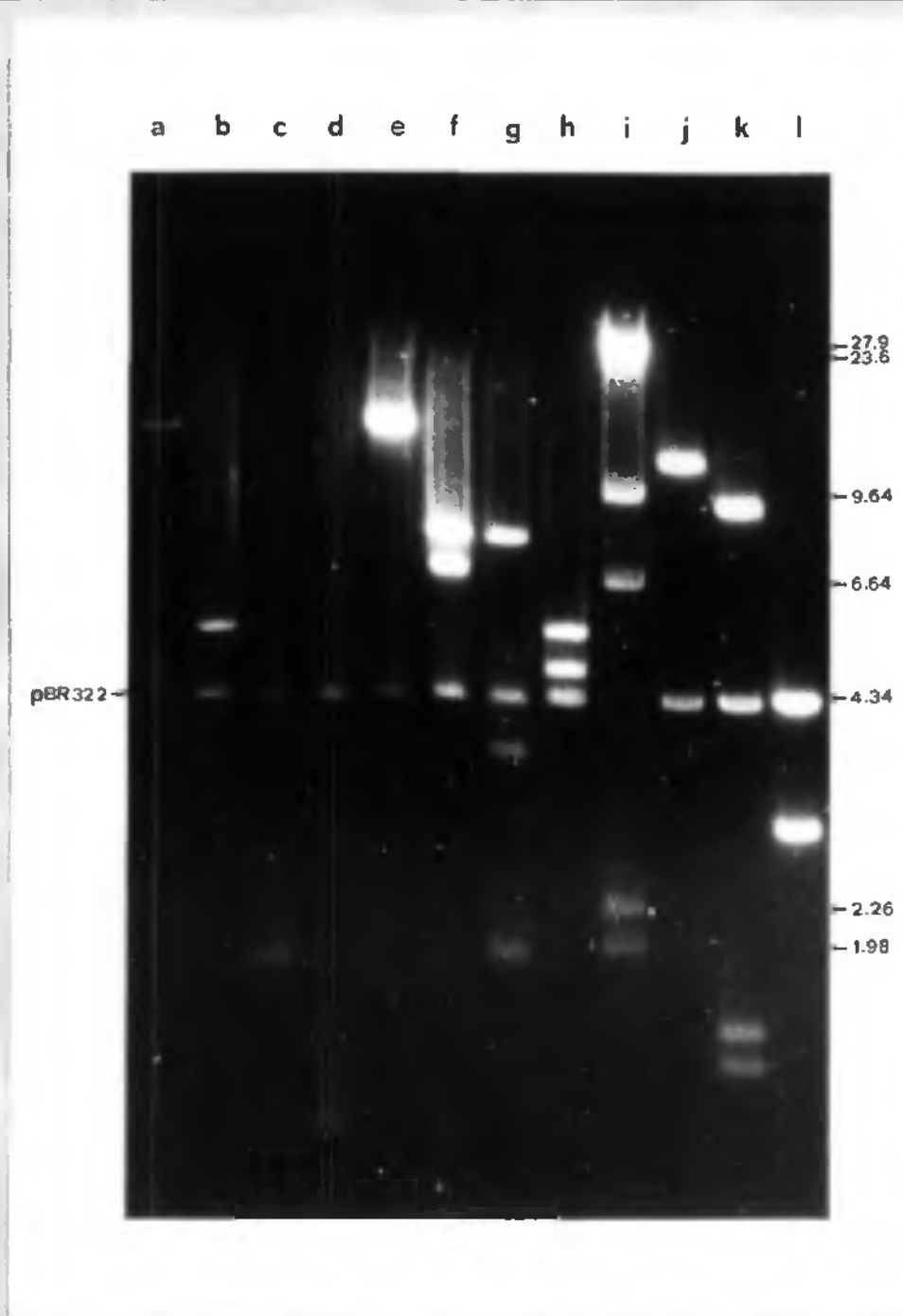
|          |   |                                              |
|----------|---|----------------------------------------------|
| pGmcB 15 | : | <u>Bam-B</u> (15.0)                          |
| pGmcB 14 | : | <u>Bam-C</u> (8.0); <u>Bam-E</u> (6.8)       |
| pGmcB 81 | : | <u>Bam-D</u> (7.5); Bam-L (3.7); Bam-S (1.9) |
| pGmcB 31 | : | <u>Bam-G</u> (5.4); <u>Bam-I</u> (4.7)       |

Seuls les 6 fragments soulignés ont été localisés. Leur position sur l'ADNct de soja a été déterminée soit par digestion avec différents enzymes de restriction, soit par hybridation. Le fragment Bam-A n'a pas été cloné, mais a pu être localisé par déduction des résultats précédents. La localisation de ces fragments est représentée sur la figure 9.

### 3.3 Fragments Pst I

Nous avons également essayé de cloner les fragments de restriction Pst I, bien que seuls les 6 plus petits fragments (Pst-D à Pst-I) soient clonables dans le pBR322. L'analyse des colonies positives a révélé que seuls 4 de ces fragments étaient clonés, à savoir :

Figure 8 : Electrophorèse sur gel de recombinants Pst I, BamHI et Hind III de l'ADNct de soja



Les recombinants ont été digérés par les enzymes de restriction correspondants. La bande correspondant au vecteur (pBR322) est indiquée. Le profil (i) est une digestion du phage Lambda par Hind III; les tailles des différents fragments de restriction obtenus sont indiquées (en kbp).

pGmcP E(a), pGmcP G(b), pGmcP H(c), pGmcP I(d), pGmcB 15(e), pGmcB 14(f), pGmcB 81(g), pGmcB 31(h), Lambda digéré par Hind III(i), pGmcH 69(j), pGmcH 209(k), pGmcH 179(l).

|         |   |              |
|---------|---|--------------|
| pGmcP E | : | Pst-E (13.7) |
| pGmcP G | : | Pst-G (6.2)  |
| pGmcP H | : | Pst-H (1.9)  |
| pGmcP I | : | Pst-I (1.0)  |

Ces fragments ont déjà été localisés sur le génome chloroplastique de soja (chapitre III). Nous n'avons pas réussi à retrouver les fragments Pst-D (peut-être trop grand pour être clonable) et Pst-F (peut-être pour les mêmes raisons que le fragment Bam-H)

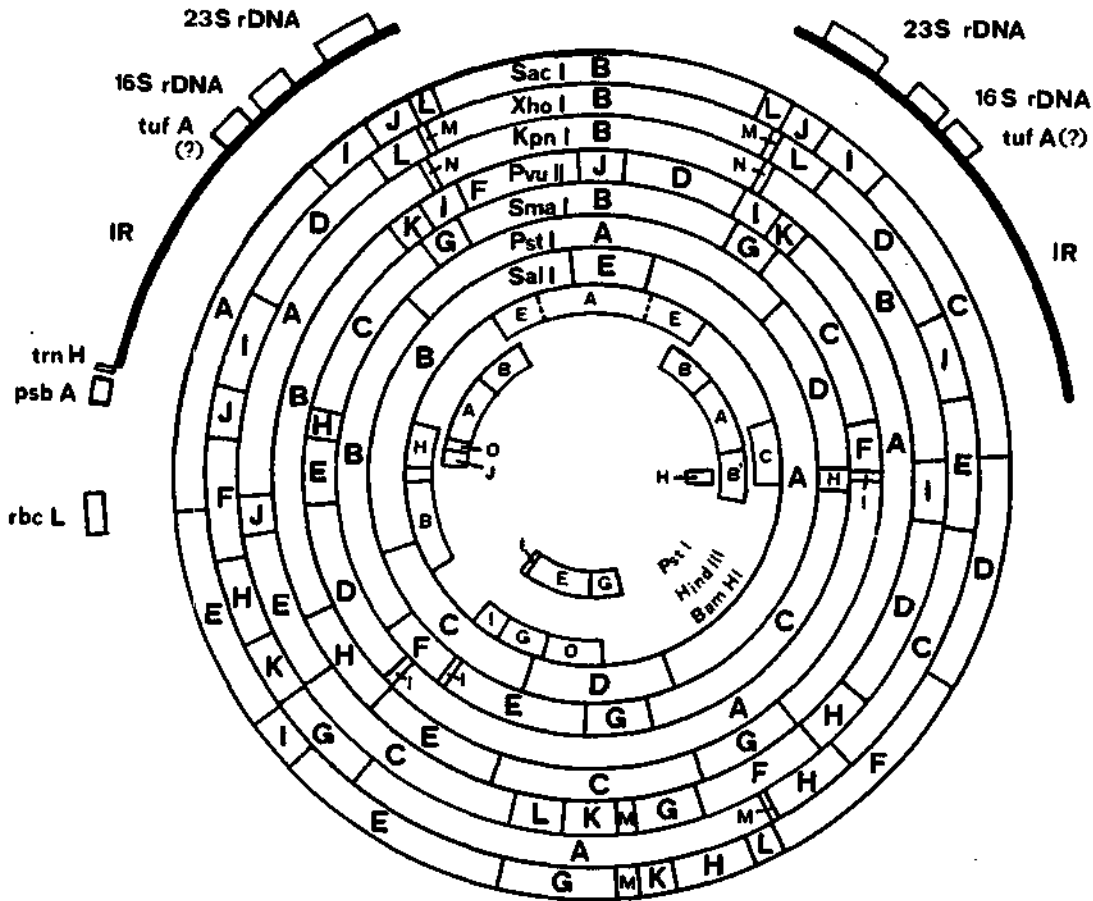
#### 4. Conclusions

Plus du 53% (~80 kbp) du génome chloroplastique du soja a été cloné dans le pBR322 sous la forme de fragments de restriction BamHI, Hind III et Pst I (figures 8 et 9). Ainsi, les régions portant les gènes pour le photogène et la grande sous-unité de la ribulose 1,5-bisphosphate carboxylase sont disponibles sous forme de clones.

Nous avons malheureusement été incapables de cloner plusieurs fragments de taille raisonnable, et notamment les fragments Hind-B<sub>1</sub> et B<sub>2</sub> (porteurs de l'information génétique de l'EF-Tu, de l'ARNr 16S et d'une partie de l'ARNr 23S), et Bam-H (photogène).

Il est possible que certaines séquences d'ADN présentes sur ces fragments inhibent une ou plusieurs fonctions du plasmide, comme par exemple sa faculté de répllication (Palmer et Thompson, 1981b).

Figure 9 : Localisation de certains fragments BamHI, Hind III et Pst I de l'ADNct de soja clonés dans le pBR322



Fragments BamHI :

- les fragments Bam-B, C, D, I et G sont clonés et localisés.
- le fragment Bam-E est cloné, mais ne peut être localisé avec certitude; il est à cheval sur l'extrémité d'une des IR, soit sur Sal-A, soit sur Sal-B.
- le fragment Bam-A n'a pas été cloné, mais peut être placé à cheval sur Sal-A, B et E; sa localisation précise dépend de celle de Bam-E.

Fragments Hind III:

- les fragments Hind-A<sub>1</sub>, A<sub>2</sub>, B', O et J ont été clonés et localisés.
- les fragments B<sub>1</sub> et B<sub>2</sub> ne sont pas clonés, mais seulement localisés.

Fragments Pst I :

- les fragments Pst-E, G, H et I ont été clonés et localisés.

V. SEQUENCAGE DE LA REGION DE L'ADN CHLOROPLASTIQUE DE SOJA  
PORTANT LES GENES DU p<sub>sb</sub> A ET DU trn H, AINSI QUE DES  
REGIONS CONTENANT LES 2 EXTREMITES DES REGIONS REPETITIVES  
INVERSEES-ADJACENTES A LA GRANDE REGION UNIQUE

---

A. INTRODUCTION

Une des caractéristiques des chloroplastes des plantes supérieures est de synthétiser à la lumière une protéine de Mr 32'000, en quantité presque aussi abondante que la LSU. Cependant cette protéine de Mr 32'000 ne s'accumule pas, ayant un turnover 50 à 60 fois plus rapide que celui de la LSU (Edelman et Reisfeld, 1980; Mattoo et al, 1981). En outre, cette protéine est caractérisée par l'absence de résidu lysine dans sa composition en acides aminés (Edelman et Reisfeld, 1980). Chez le maïs (Grebanier et al, 1978) et Spirodela oligorrhiza (Edelman et Reisfeld, 1980), il a été mis en évidence l'existence de précurseurs de Mr 34'500 et 33'500, respectivement.

Plus récemment (Mattoo et al, 1981), il a été démontré que cette protéine de Mr 32'000 pouvait être identifiée d'une part comme le "bouclier protéique" décrit par Renger (1976) engagé dans la régulation du flux d'électrons par l'intermédiaire du complexe B plastoquinone-protéine, d'autre part comme responsable de la sensibilité des chloroplastes à des herbicides comme le DCMU et les triazines (Pfister et al, 1981).

Comme cette protéine est synthétisée en abondance dans les chloroplastes de toutes les plantes étudiées jusqu'à maintenant, la sensibilité ou la résistance aux herbicides n'est pas due à la présence ou à l'absence de cette protéine, mais plutôt à sa structure primaire, qui ferait que les herbicides sont susceptibles ou non de se fixer sur cette protéine (Steinback et al, 1981).

A cause du rôle important de cette protéine membranaire, il nous a paru intéressant de connaître sa séquence de nucléotides (pour ensuite en déduire celle d'acides aminés), et c'est pour cette raison que nous avons décidé de séquencer le gène de cette protéine chez le soja.

Le psb A a été localisé sur le génome chloroplastique de plusieurs plantes, notamment chez le maïs (Bedbrook et al, 1978), l'épinard (Driesel et al, 1980), Spirodela (Van Ee et al, 1982), la moutarde (Link, 1981), le pois et Vigna radiata (Palmer et al, 1982), l'euglène (Keller et al, 1982) et Chlamydomonas reinhardtii (Rochaix, 1981); il se trouve dans la grande région unique, très proche de l'extrémité d'une des régions inversées (exceptions faites de l'euglène et de C. reinhardtii); la localisation de ce gène chez le soja a été expliquée au chapitre III.

En outre, ce gène a déjà été séquencé chez deux espèces, c'est-à-dire Spinacia oleracea et Nicotiana debneyi (Zurawski et al, 1982). Ils ont démontré que, dans les 2 cas, ce gène donnait naissance à une protéine formée de 353 acides aminés, correspondant à une taille de 38'950 daltons. La taille de ce produit primaire de traduction dépasse celle estimée sur la base de la mobilité électrophorétique du photogène (Mr 32'000 à 36'000). Ces auteurs ont estimé par conséquent que la traduction pourrait commencer depuis un second codon d'initiation ATG, éloigné de 108 nucléotides du premier ATG, ce qui réduirait la taille du produit primaire de traduction d'environ 4'000 daltons.

D'un autre côté, la très grande homologie de séquence existant entre ces 2 gènes provenant de plantes différentes suggère que la traduction commence depuis le premier ATG.

D'autre part, Swamy et Pillay (1982) ont récemment identifié un ARNt<sup>His</sup> chloroplastique chez le soja, sans pour autant en localiser le gène. Dans le cas de l'épinard, le trn H est localisé entre le psb A et l'extrémité de la région répétitive adjacente (Bohnert et al, 1982).

Nous avons donc également séquencé cette région et avons de plus déterminé le début des régions répétitives en séquençant la région d'ADN correspondant à l'autre côté de la région unique. Ceci nous a permis de localiser précisément le psb A et le trn H par rapport à la région répétitive inversée adjacente.

## B. MATERIEL ET METHODES

### 1. Préparation de l'ADN pour le séquençage

Les conditions de digestion avec des enzymes de restriction et la méthode du clonage sont décrites aux chapitre III et IV, respectivement. Les sous-fragments d'ADN obtenus par digestion des fragments Hind-J, Hind-O et Hind-B' ont été séparé par électrophorèse sur gel de polyacrylamide (20 × 20 × 0.05 cm; 5 à 7.5% en acrylamide). L'électrophorèse a lieu dans du tampon TBE ( Tris 50 mM, acide borique 50 mM, EDTA 2mM, pH 8.3) durant 3-4 h à 250-300 Volts.

### 2. Techniques de séquençage

Deux méthodes de séquençage ont été utilisées :

#### 1) La méthode de Maxam et Gilbert (1980)

Principe : (figure 1)

Après déphosphorylation, les fragments d'ADN à séquencer sont marqués radioactivement en extrémité 5' ( $\gamma$ -ATP<sup>\*</sup>; polynucléotide-kinase de T<sub>4</sub>). Les fragments sont ensuite séparés les uns des autres par électrophorèse et les bandes d'ADN sont repérées par autoradiographie. L'ADN est récupéré (élution, précipitation), puis dénaturé et les deux simples brins complémentaires sont séparés par électrophorèse. L'ADN est visualisé et récupéré comme précédemment.

Chaque échantillon est alors divisé en quatre parts, et chacune de ces parties subit un traitement chimique particulier, qui va modifier spécifiquement chaque base. Un traitement à la pipéridine permet de cliver l'ADN simple brin à l'endroit de chaque modification. La multitude de fragments obtenus est alors séparés par électrophorèse sur gel de polyacrylamide dénaturant (urée). Il est alors possible de visualiser les bandes d'ADN radioactives par autoradiographie et de lire ainsi la séquence nucléotidique dans le sens 5'→3' (puisque seuls les fragments d'ADN possédant encore le radical phosphate radioactif en 5' sont visibles).

## 2) Méthode de Sanger et Coulson (1980)

Principe : (figure 2)

Cette technique est basée sur la propriété de polymérisation en 5'→3' du fragment "Klenow" de l'ADN-polymérase I de E.coli. En effet, en présence d'une amorce (primer) et d'un ADN simple brin (template), cet enzyme est capable de synthétiser le brin complémentaire au template.

Si nous ajoutons dans le mélange de réaction, en plus des désoxyribonucléotides (dont un est radioactif), des didéoxyribonucléotides (qui bloquent l'élongation de la chaîne d'ADN), nous sommes à même de faire de multiples synthèses partielles d'un brin d'ADN complémentaire au template.

Après dénaturation, ces molécules d'ADN monocaténaïres radioactives sont alors séparées par électrophorèse (comme 1).

En outre, cette méthode est rendue très simple par l'utilisation des propriétés du phage M13, dont la forme infectieuse est monocaténaire. Comme ce phage peut être utilisé comme vecteur pour du clonage, il suffit de cloner dans ce phage les fragments d'ADN que l'on veut séquencer.

Ensuite, la forme monocaténaire est directement utilisée comme template par le fragment "Klenow" de l'ADN-polymérase I de E.coli.

FIGURE 1 : REPRESENTATION SCHEMATIQUE DU SEQUENCAGE DE FRAGMENTS  
D'ADN PAR LA METHODE DE MAXAM ET GILBERT

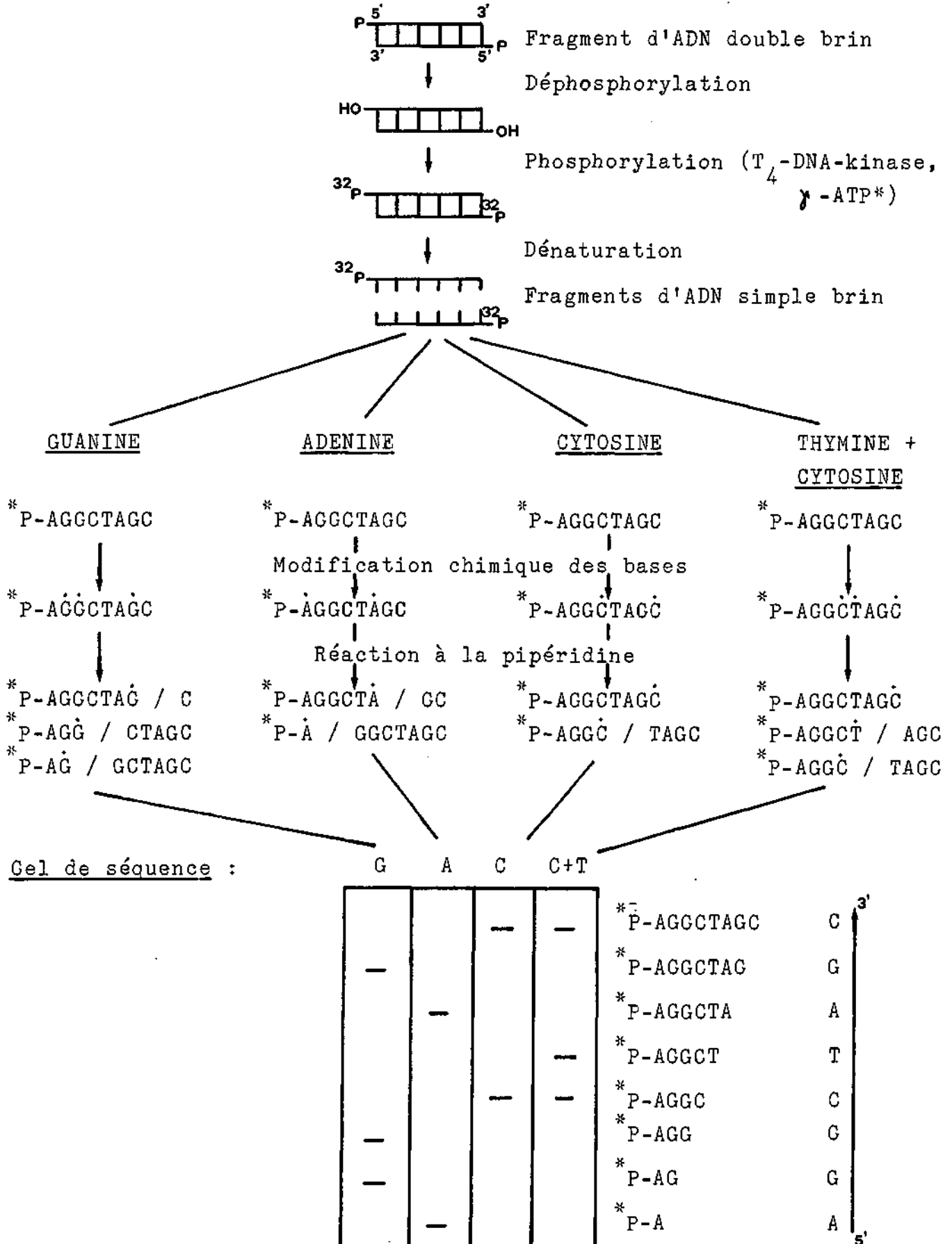
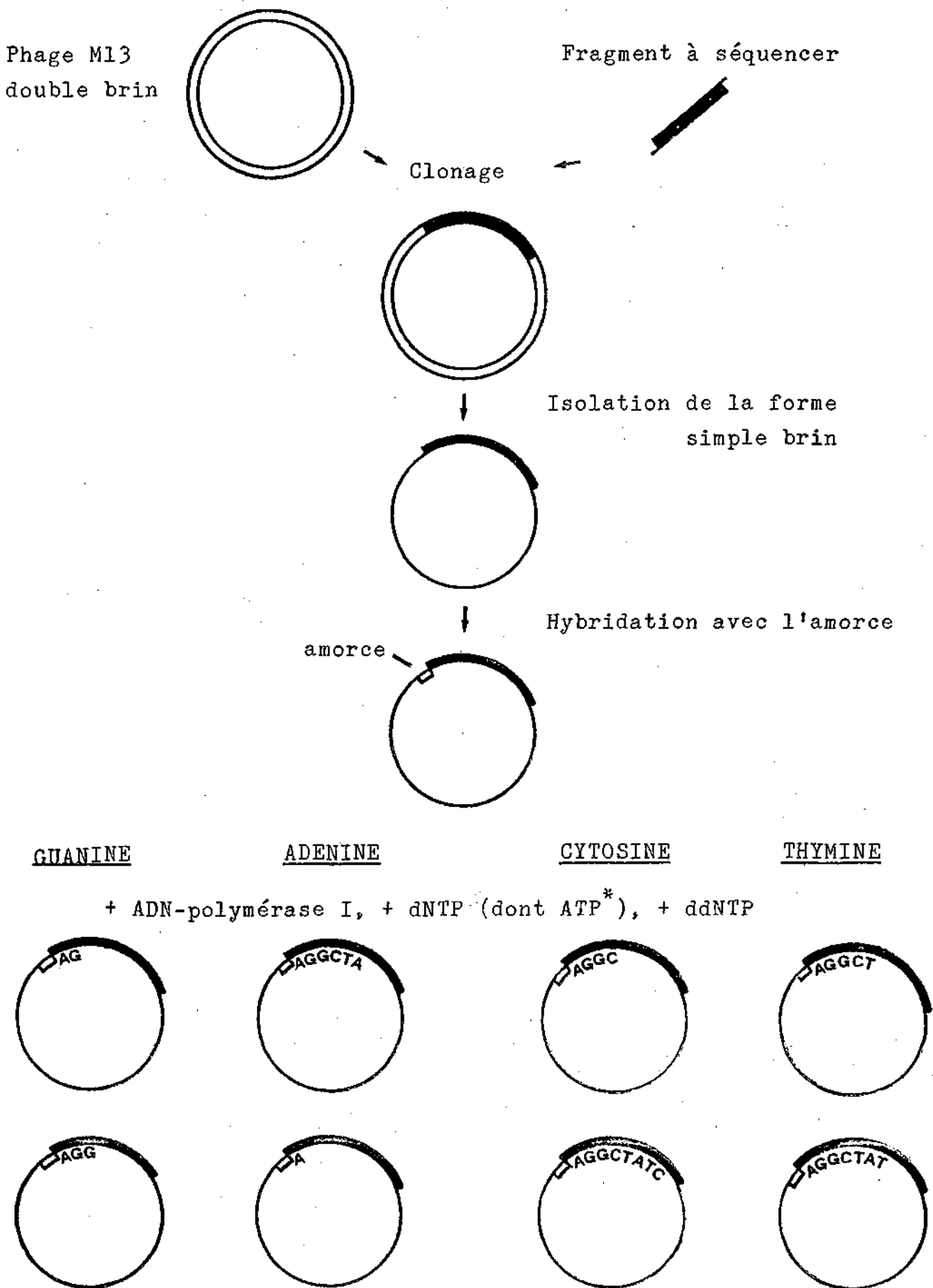
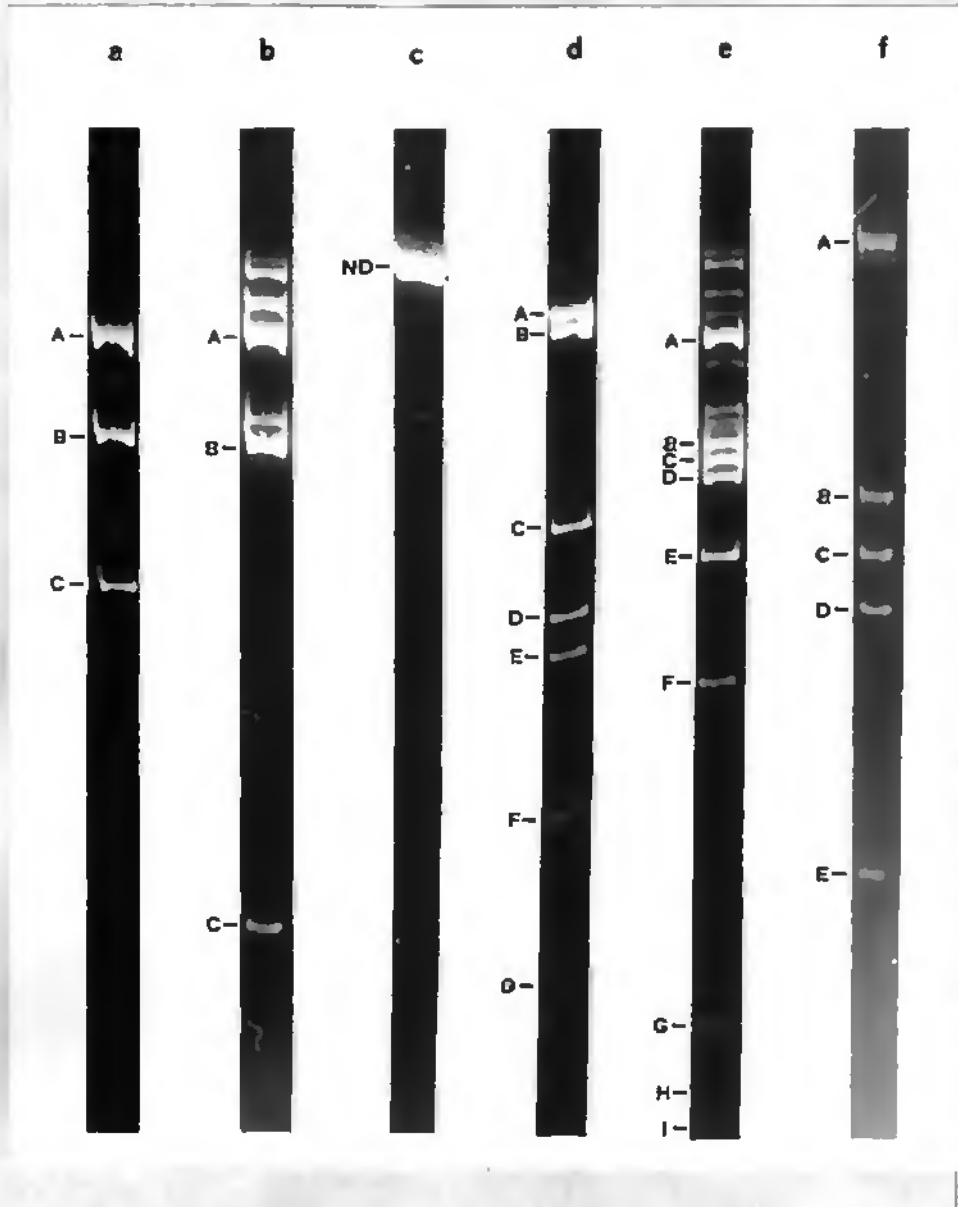


FIGURE 2 : REPRESENTATION SCHEMATIQUE DU SEQUENCAGE DE FRAGMENTS D'ADN PAR LA METHODE DE SANGER ET COULSON



Dénaturation et séparation sur gel de séquence

Figure 3 : Profil électrophorétique des fragments Hind/Sma 0.8 et Hind-0 (portant le psb A) après digestions avec Hinf I, Sau3A et Hae III.



Electrophorèse sur gel de polyacrylamide, 6% en acrylamide, 20 × 20 × 0.1 cm.

Hind/Sma 0.8 : Hinf I (a), Sau3A (b), Hae III (c).

Hind-0 : Hinf I (d), Sau3A (e), Hae III (f).

Les bandes qui ne sont pas indiquées par une lettre proviennent de digestions partielles.

## C. RESULTATS

### 1. Choix des enzymes de restriction

Nous savons que le psb A se trouve à cheval sur les fragments Hind-J (2.8 kbp) et Hind-O (1.4 kbp); cependant, seul le sous-fragment Hind/Sma de 0.8 kbp provenant de Hind-J porte ce gène. Nous avons donc la situation représentée sur la figure 6 du chapitre IV, page 80.

Comme nous ne savons pas exactement où se trouvent le psb A et le trn H sur ces deux fragments, nous avons séquencé la totalité du fragment Hind-O et du sous-fragment Hind/Sma de 0.8 kbp.

Ces fragments étant trop grands pour être séquencés intégralement, il faut les recouper en plus petits morceaux par des enzymes de restriction.

Les fragments Hind-O et Hind/Sma 0.8 kbp ont été isolés du reste du plasmide par digestion, séparation sur gel et élution (voir chap. IV).

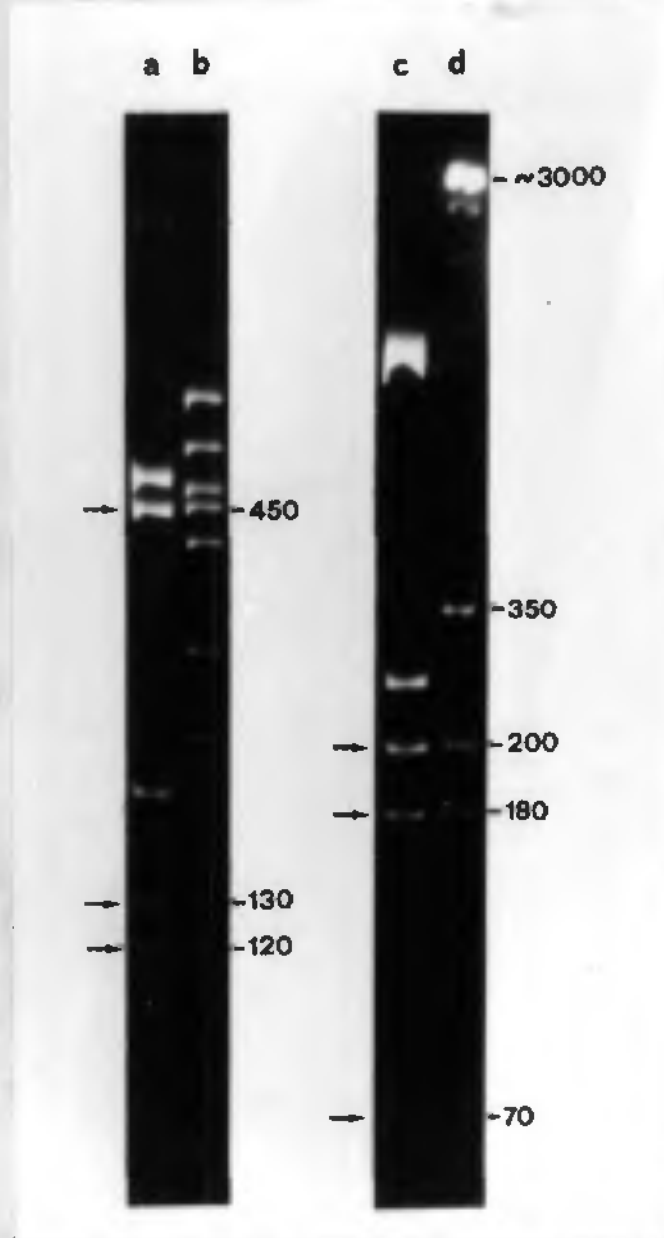
Nous avons ensuite testé plusieurs enzymes de restriction, susceptibles de couper assez fréquemment ces 2 fragments. Notre choix s'est finalement porté sur les 3 enzymes suivants : Hinf I, Sau3A et Hae III.

Les résultats de ces digestions sont représentés sur la figure 3 et résumés sur la table 1.

Table 1 : Taille (en bp) des sous-fragments de restriction Hinf I, Sau3A et Hae III obtenus à partir des fragments Hind/Sma 0.8 et Hind-O.

| <u>Hind/Sma 0.8</u> |              |                | <u>Hind-0</u> |              |                |
|---------------------|--------------|----------------|---------------|--------------|----------------|
| <u>Hinf I</u>       | <u>Sau3A</u> | <u>Hae III</u> | <u>Hinf I</u> | <u>Sau3A</u> | <u>Hae III</u> |
| A 420               | A 460        | ND 800         | A 500         | A 460        | A 800          |
| B 280               | B 290        |                | B 450         | B 280        | B 280          |
| C 150               | C 60         |                | C 190         | C 265        | C 225          |
|                     |              |                | D 130         | D 255        | D 180          |
|                     |              |                | E 120         | E 200        | E 65           |
|                     |              |                | F 90          | F 135        |                |
|                     |              |                | G 60          | G 40(2x)     |                |
|                     |              |                |               | H 20         |                |
|                     |              |                |               | I 10         |                |

Figure 4 : Profil électrophorétique des fragments Hind-0 (jonction IR-1 - LSC) et Hind/Pet 3.9 (jonction IR-2 - LSC) obtenu après digestions avec Hinf I et Hae III.



Electrophorèse sur gel de polyacrylamide, 5% en acrylamide, 20 × 20 × 0.1 cm.

Les flèches indiquent les fragments communs aux deux profils de restriction.

Hind-0 digéré par Hinf I(a); Hind/Pet 3.9 digéré par Hinf I(b);  
Hind-0 digéré par Hae III(c); Hind/Pst 3.9 digéré par Hae III(d).

Pour savoir d'une part où se situent exactement les régions répétitives inversées par rapport au psb A, d'autre part si ces régions sont homologues à 100%, nous avons également séquencé l'extrémité de la région inversée opposée (IR-2).

La localisation et le clonage des fragments de restriction Hind III et BamHI portant l'extrémité de l'IR-2 sont décrits au chapitre IV. Le sous-fragment Hind/Pst de 3.9 kbp, résultant de la digestion par Pst I du fragment Hind-B', a été utilisé comme matériel de départ (figure 5,D). Afin d'établir une carte de restriction de ce sous-fragment, ainsi que de localiser avec précision l'extrémité de l'IR-2, nous avons digéré en parallèle les fragments Hind-0 (IR-1) et Hind/Pst 3.9 (IR-2) par Hinf I et Hae III (figure 4). Les résultats obtenus montrent que :

- dans les 2 digestions, il y a au plus 3 bandes communes aux 2 profils de restriction, c'est-à-dire les bandes de 450, 130 et 120 bp pour Hinf I et les bandes de 200, 180 et 70 bp pour Hae III.

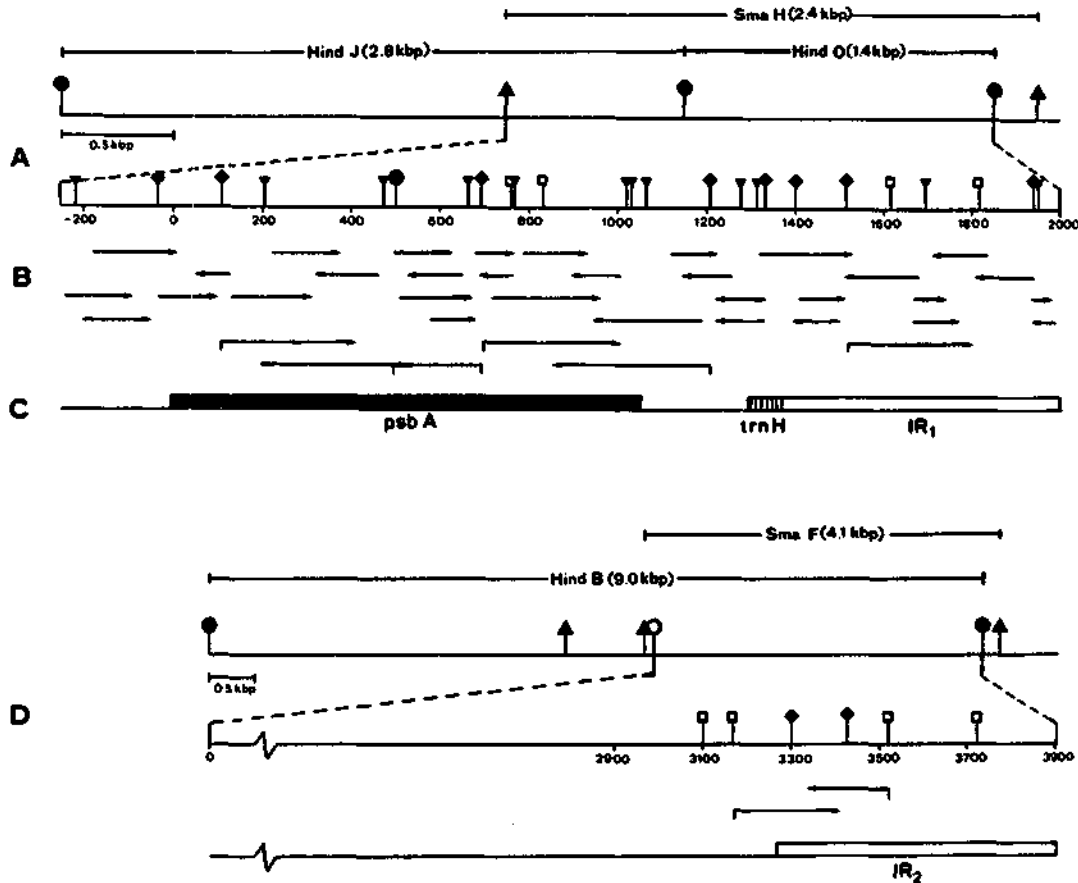
Nous pouvons donc en déduire que la fin des IR est entièrement contenue dans les fragments Hinf I 450 et 120, et les fragments Hae III 200 et 180 (figure 5,D). En analysant le profil Hae III du sous-fragment Hind/Pst 3.9 (figure 4,d), nous voyons que seuls 2 des 3 fragments non localisés (fragments de 3.0 kbp et 350 bp) peuvent être juxtaposés au fragment Hae III 200. A cause des digestions partielles présentes sur le profil Hae III (figure 4,d), nous en avons déduit que le fragment Hae III 350 était juxtaposé au fragment Hae III 200, comme représenté sur la figure 5,D.

## 2. Analyse de la région séquencée

La séquence de nucléotides des fragments Hind-0 et Hind/Sma 0.8 a été déterminée comme représenté sur la figure 5,A-B. L'analyse de cette séquence nous a montré que :

- 1) si nous essayons de traduire chaque brin de cette séquence, nous trouvons une seule ORF (open reading frame = séquence ne possédant pas de codons stop) assez longue pour coder une protéine de Mr 32'000.
- 2) à moins de 300 bp de la fin du psb A se trouve une région codant pour le trn H.

Figure 5 : Carte de restriction et stratégie de séquençage de la région de l'ADNct de soja portant le psb A, le trn H et les extrémités des IR



A : Carte de restriction du fragment Sma-H.

B : Stratégie de séquençage. Les flèches représentent les portions des fragments de restriction qui ont été séquencées sans ambiguïté.

→ : séquençage selon Maxam et Gilbert (1980)

→ : séquençage selon Sanger et Coulson (1980)

→ : brin non-codant (identique à l'ARN); ← : brin codant

C : Localisation des parties structurales du psb A et du trn H, ainsi que d'un segment de l'IR-1.

D : Carte de restriction partielle du fragment Hind/Pst 3.9, avec la stratégie de séquençage et la localisation de l'IR-2.

Les chiffres sur les échelles se réfèrent à ceux donnés sur les figures 6, 7 et 9.

● : Hind III, ↑ : Sma I, ∟ : Sau3A, ◆ : Hinf I, ▢ : Hae III, ○ : Pst I.

- 3) le fragment Hae III de 350 bp, devant contenir l'extrémité de l'IR-2, a également été séquencé. Ces données nous ont permis de localiser avec précision l'extrémité des IR, comme représenté sur la figure 5,C-D.
- 4) l'analyse de la séquence de l'IR-1 en aval du trn H a montré qu'elle pourrait contenir la fin d'un gène de protéine.

### 3. Analyse de la séquence du psb A

La figure 6 représente la séquence de nucléotides du psb A et celles des régions adjacentes à ce gène. La séquence d'acides aminés, obtenue par déduction de la séquence de nucléotides de la partie structurale du psb A, est représentée au-dessous de la séquence de nucléotides. Par comparaison, les séquences de nucléotides des régions correspondantes de l'ADNct de Spinacia oleracea et Nicotiana debneyi sont également indiquées.

Les résultats de cette étude comparative peuvent être résumés ainsi :

- 1) La partie codante du psb A du soja est identique en taille à celle du psb A de S.oleracea et N.debneyi, c'est-à-dire qu'elle donne naissance à une protéine formée de 353 acides aminés.
- 2) Il y a 73 (78) nucléotides différents entre la partie codante du psb A du soja et celle de S. oleracea (N.debneyi).
- 3) Mis à part 3 exceptions, les différences de nucléotides apparaissent en position "wobble", et seules 2 mutations produisent des changements dans la composition en acides aminés : l'acide aspartique remplace l'acide glutamique (position 347) et l'isoleucine remplace la thréonine (position 351).
- 4) La lysine est également absente
- 5) Environ 130 nucléotides de la région 5' adjacente sont fortement conservés (80 à 85% d'homologie). Cette région possède des sites potentiels de régulation (promoteurs).

FIGURE 6

-----

S.o. CCAATTTT TAATAC TGAAT CAT ATTCTTT AG TTTCTTTTC  
G.m. CCACTAGAAATTAATATATGAAATTCATATATAGAATTAT AGCG  
N.d. CCACTAGC ATAT CCAATCTA ATTCTCTGTAG C AAGTC

-220 AATTCACTTCCAT -200 TCCAC -180 AAAAAATT -160 CTGA TA GA T CTC -140 ACTAGATA TGGTTGACAGGG  
ATCCATAGAAAATAAACCCAT TTTTAATGAATGAATGAGCTA AAAAACTAAATTCGGATAAATCTAAATAAGAGCATACTTACTATGGATATGCTAT TGGTTGACAGCTGG  
CGTATTTTCCGAC AACTCCAT TAAAAATT TGAATA GA T CTA CATACACCTTGGTTGACAGAG  
TCTTGACA

-100 -80 -60 -40 -20 -10  
CATATAAGCATGTTATACTGTTGAATAACANT CTTTCAA TTCTATTCTAGCGAATTTCGTGCTTGGGAGTCCCTGATGATTAATTAATTAACCAAGATTTTACG  
TATATAAGCATGTTATACTGTTGAATAACAG TCTCAAAATTTCTAATTCATAGATAAATTTGCTGCTTGGGAGTCCCTGATGATTAATTAACCAAGATTTTACG  
TATATAAGCATGTTATACTGTTGAATAAAGCCTTCTTTTCTA TTTTATTGTAGAAAACACTAGTCTCTTGGGAGTCCCTGATGATTAATTAACCAAGATTTTACG  
tg TATAAT

90  
ATG ACT GCA ATT TTA GAG AGA CGC GAG AGC GAA AGC CTA TGG GGT CGC TTC TGT AAC TGG ATA ACC AGC ACC GAA AAT CGT CTT TAC ATT  
Met Thr Ala Ile Leu Glu Arg Arg Glu Ser Glu Ser Leu Trp Gly Arg Phe Cys Asn Trp Ile Thr Ser Thr Glu Asn Arg Leu Tyr Ile  
10 20 30

180  
CGA TGG TTT GGT GTT TTC ATG ATT CGT ACT TTA TTG ACC GCC ACT TCT GTA TTT ATT ATC GCT TTT ATT GCT GCC CCT CCA GTA CAT ATT  
Gly Trp Phe Gly Val Leu Met Ile Pro Thr Leu Leu Thr Ala Thr Ser Val Phe Ile Ile Ala Phe Ile Ala Ala Pro Pro Val Asp Ile  
40 50 60

270  
CAT GGT ATT CGT GAG CCT GTT TCT GGA TCT CTA CTT TAT GGA AAC AAT ATC ATT TCT GGT GCC ATT ATT CCT ACT TCT GCC GCT ATA GGT  
Asp Gly Ile Arg Glu Pro Val Ser Gly Ser Leu Leu Tyr Gly Asn Asn Ile Ile Ser Gly Ala Ile Ile Pro Thr Ser Ala Ala Ile Gly  
70 80 90

360  
TTG CAC TTT TAT CCT ATT TGG GAA GCG GCA TCT GTT GAT GAA TGG TTA TAC AAC GCC GGT CCT TAT GAA CTA ATT GTT CTA CAC TTC TTA  
Leu His Phe Tyr Pro Ile Trp Glu Ala Ala Ser Val Asp Glu Trp Leu Tyr Asn Gly Gly Pro Tyr Glu Leu Ile Val Leu His Phe Leu  
100 110 120

450  
CTT GGT GTA GGT TGC TAC ATG GCG CGT GAG TGG GAA CTT AGT TTT CGT TTG GGT ATG GGT CCT TGG ATT GCT GTT GCA TAT TCA GCT CCT  
Leu Gly Val Ala Cys Tyr Met Gly Arg Glu Trp Glu Leu Ser Phe Arg Leu Gly Met Arg Pro Trp Ile Ala Val Ala Tyr Ser Ala Pro  
130 140 150

540  
GTT GCA GCC GCT ACT GCT GAT TTC TTG ATG TAT CCT ATT GGA CAG GGA AGC TTT TCA GAT GGT ATG CCT CTA GGA ATT TCA GGT ACT TTC  
Val Ala Ala Ala Thr Ala Val Phe Leu Ile Tyr Pro Ile Gly Glu Gly Ser Phe Ser Asp Gly Met Pro Leu Gly Ile Ser Gly Thr Phe  
160 170 180

630  
AAT TTT ATG ATT GTA TTT CAG GCT GAG CAT AAT ATT CTT ATG CAT CCA TTT CAC ATG TTA GGT GTA GGT GGT GTA TTC GCG GCG TCG CTA  
Asn Phe Met Ile Val Phe Glu Ala Glu His Asn Ile Leu Met His Pro Phe His Met Leu Gly Val Ala Gly Val Phe Gly Gly Ser Leu  
190 200 210

720  
TTC AGT GCT ATG CAT GGT TCG TTG GTA ACT TCT AGT TTG ATC AGG GAA ACC ACA GAA AAT GAA TCT GCT AAT GAA GGT TAC ACA TTT GGT  
Phe Ser Ala Met His Gly Ser Leu Val Thr Ser Ser Leu Ile Arg Glu Thr Thr Glu Asn Glu Ser Ala Asn Glu Gly Tyr Arg Phe Gly  
220

|                                                                                                                         |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
|-------------------------------------------------------------------------------------------------------------------------|--|--|--|--|--|--|--|--|--|--|--|--|--|-----|--|--|--|--|--|--|--|--|--|-----|--|--|--|------|------|--|--|--|------|--|--|--|------|--|--|--|------|--|--|--|
|                                                                                                                         |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  | 810  |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
|                                                                                                                         |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| CAA GAG GAA GAA ACC TAT AAT ATT GTA GCT GCT CAT GGT TAT TTT GGC CGA TTC ATC TTC CAA TAT CCA AGT TTC AAC AAT TCT CGT TCT |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| C C C T C G                                                                                                             |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| Gln Gln Glu Glu Thr Tyr Asn Ile Val Ala Ala His Gly Tyr Phe Gly Arg Leu Ile Phe Gln Tyr Ala Ser Phe Asn Asn Ser Arg Ser |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| 250                                                                                                                     |  |  |  |  |  |  |  |  |  |  |  |  |  | 260 |  |  |  |  |  |  |  |  |  | 270 |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
|                                                                                                                         |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  | 900  |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| TTA CAT TTC TTC TTA GCT GCT TGG GCT GTA GTA GGT ATT TCG TTT ACC GCT TTA GGT ATC ACC ACT ATC GCT TTC AAC TTA AAT GGT TTC |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| C C C C C                                                                                                               |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| Leu His Phe Phe Leu Ala Ala Trp Pro Val Val Gly Ile Trp Phe Thr Ala Leu Gly Ile Ser Thr Met Ala Phe Asn Leu Asn Gly Phe |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| 280                                                                                                                     |  |  |  |  |  |  |  |  |  |  |  |  |  | 290 |  |  |  |  |  |  |  |  |  | 300 |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
|                                                                                                                         |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  | 990  |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| AAT TTC AAC CAA TCC GTA GTT CAT AGT CAA GGT GGT GTA ATT AAT ACC TGG GCT GAT ATT ATT AAC CGA GCT AAC CTT GGT ATG GAA GTA |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| T C C C T C T                                                                                                           |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| Asn Phe Asn Gln Ser Val Val Asp Ser Gln Gly Arg Val Ile Asn Thr Trp Ala Asp Ile Ile Asn Arg Ala Asn Leu Gly Met Glu Val |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| 310                                                                                                                     |  |  |  |  |  |  |  |  |  |  |  |  |  | 320 |  |  |  |  |  |  |  |  |  | 330 |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
|                                                                                                                         |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  | 1050 | 1070 |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| ATG CAT GAA CGT AAT GCT CAT AAT TTC GCT CTA GAT CTA GCT GCG ATC GAC GCT CCA TCT ATT AAT CGA TAA AATTTGGATCTTAAGC TAGATT |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| C C C T A CA                                                                                                            |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| Met His Glu Arg Asn Ala His Asn Phe Pro Leu Asp Leu Ala Ala Ile Asp Ala Pro Ser Ile Asn Gly TER                         |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| 340                                                                                                                     |  |  |  |  |  |  |  |  |  |  |  |  |  | 350 |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
|                                                                                                                         |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  | 1090 | 1110 |  |  |  | 1130 |  |  |  | 1150 |  |  |  | 1170 |  |  |  |
| AGATGACTTAGT CAAAGTAAAGGGGGCAGTC CCGA TTCTTGACAAAACAACAANTCGGTTATTGCTCCTTTACT ACTAC TAGTTATAAATA                        |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| AGATC TTTT CAAAGTAAAGG CAATA TCAACTTTTTTGA TATTCGCCCTTTACTTTTTATTATTGATTAGTAATCTTTTTATTATAAATA                          |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |
| ATAGGAGGTTTTGAAAAGCAAG GAG CAATANTCAT TTCTTGTCTATCAAGACGG TCCTATTG TCCTTT CTITT TTT CTTTTTATT                           |  |  |  |  |  |  |  |  |  |  |  |  |  |     |  |  |  |  |  |  |  |  |  |     |  |  |  |      |      |  |  |  |      |  |  |  |      |  |  |  |      |  |  |  |

Figure 6 : Séquence de nucléotides du psb A et de ses régions adjacentes

Seul le brin non-codant (identique à l'ARN) est indiqué.

Les séquences correspondant à celles du psb A de S.oleracea (S.o.) et N.debneyi (N.d.) sont alignées avec celle de G.max (G.m.); dans la partie structurale, seules les déviations de la séquence de soja sont données; dans les régions adjacentes, les séquences complètes sont marquées et alignées pour avoir un maximum d'homologie. Les régions homologues à 100% sont encadrées.

La séquence d'acides aminés déduite de celle de nucléotides est indiquée; la première méthionine est assimilée à la position 1.

Les sites potentiels de régulation dans la région 5', ainsi que les régions susceptibles de former une structure secondaire en épingle à cheveux dans la région 3', sont soulignés.

Les séquences consensus des promoteurs de type procaryotique (régions -10 et -35; Sprinzl et Gauss, 1982) sont données.

- 6) L'homologie de séquence devient très faible à partir de la position -130 environ, ceci étant surtout dû à de multiples insertions et délétions.
- 7) La région 3' adjacente montre une homologie de 63% (57%) avec celle de S.oleracea (N.debneyi). Elle peut donner naissance à un ARNm susceptible de former une structure secondaire en "épingle à cheveu" (positions 1094-1115 et 1121-1143) similaire à celles proposées pour les régions correspondantes chez S.oleracea et N.debneyi. Ces structures pourraient servir de signal de terminaison de la transcription.

La table 2 montre les fréquences d'utilisation des différents codons du psb A de G.max et S.oleracea. Il est intéressant de noter que l'usage des codons est assez semblable entre ces 2 espèces, avec la présence d'un même codon stop (TAA) et l'absence des 7 mêmes codons (en ne tenant pas compte des codons pour la lysine).

#### 4. Analyse de la séquence du trn H et des séquences faisant la jonction entre la grande région unique et les régions répétitives inversées

Le génome chloroplastique de soja contient deux régions répétitives inversées (IR), séparées par une grande région unique (~84 kbp) et une petite région unique (~22 kbp) (chapitre III). Nous savons d'autre part que chez l'épinard, le trn H est localisé entre le psb A et l'extrémité de l'IR adjacente (Bohnert et al, 1982). Dans le but de situer exactement ces 2 gènes par rapport au début de l'IR-1, nous avons également séquencé le début de l'IR-2, comme représenté sur la figure 5.

Les résultats de ce séquençage sont représentés sur la figure 7. Il y a 319 nucléotides entre le codon stop du psb A et le début de l'IR-1. A l'intérieur de cette séquence, nous avons trouvé le gène codant pour l'ARNt<sup>His</sup><sub>GUG</sub> (trn H); il débute en position 1301, c'est-à-dire à 238 nucléotides du codon stop du psb A et a une taille de 76 nucléotides.

Table 2 : Comparaison de l'usage des codons dans la séquence  
de l'ARNm codant pour le psb A de G.max et S.oleracea  
(valeurs entre parenthèses; Zurawski et al,1982)

|   | U   |         | C   |         | A   |         | G   |           |
|---|-----|---------|-----|---------|-----|---------|-----|-----------|
| U | Phe | 12 (9)  | Ser | 11 (10) | Tyr | 8 (7)   | Cys | 1 (2) U   |
|   | Phe | 14 (17) | Ser | 3 (5)   | Tyr | 4 (5)   | Cys | 1 (0) C   |
|   | Leu | 9 (9)   | Ser | 3 (0)   | Ter | 1 (1)   | Ter | 0 (0) A   |
|   | Leu | 8 (7)   | Ser | 0 (2)   | Ter | 0 (0)   | Trp | 10 (10) G |
| C | Leu | 6 (6)   | Pro | 12 (9)  | His | 7 (4)   | Arg | 8 (8) U   |
|   | Leu | 0 (0)   | Pro | 0 (0)   | His | 3 (6)   | Arg | 2 (3) C   |
|   | Leu | 8 (9)   | Pro | 3 (6)   | Gln | 4 (5)   | Arg | 2 (1) A   |
|   | Leu | 0 (0)   | Pro | 0 (0)   | Gln | 2 (1)   | Arg | 0 (0) G   |
| A | Ile | 23 (18) | Thr | 8 (12)  | Asn | 14 (12) | Ser | 5 (6) U   |
|   | Ile | 7 (10)  | Thr | 7 (3)   | Asn | 8 (10)  | Ser | 5 (4) C   |
|   | Ile | 2 (3)   | Thr | 1 (2)   | Lys | 0 (0)   | Arg | 2 (2) A   |
|   | Met | 12 (12) | Thr | 0 (0)   | Lys | 0 (0)   | Arg | 1 (1) G   |
|   | Val | 8 (9)   | Ala | 23 (25) | Asp | 7 (6)   | Gly | 21 (24) U |
|   | Val | 0 (0)   | Ala | 4 (2)   | Asp | 1 (1)   | Gly | 4 (2) C   |
|   | Val | 13 (12) | Ala | 5 (6)   | Glu | 14 (16) | Gly | 7 (6) A   |
|   | Val | 0 (0)   | Ala | 3 (2)   | Glu | 6 (5)   | Gly | 1 (1) G   |

Figure 7 : Séquence de nucléotides de la région du trn H, de ses régions adjacentes et de l'extrémité des IR-1 et IR-2.

1200  
TCTTTTATTTATAAATATTTATACATAAGTTTTTGATTCTTTTCGCGATTCTTTTAGCATTTTTCTATCTT L5C-1  
tcTTGACa tg TAaAT

1250 1300  
AAAAGGAAAAAGAATGATAACGAACGAAAGGATAGAATTTTATATATAGATCATTTTTACATAGTATAAGGGC L5C-1

1350  
GGATCTAGCCAAGTGGATCAAGGCAGTGGATTGTGAATCCACCATCGCGGGGTTCAATTCCCGTCGTTGCGCCCA L5C-1  
GGCCAATCATTGTAGGTATAATGCTAGATGCTCTGGACCAAGTTATTATTATATCTTTTTCCGCTTTTGTG L5C-2

3250

1400 1450  
TTAAGTTTATTTATTTTCTTAATAAATGATTCCGTACAAAAGGATTTTTTTTTTAGTGAACGTGTCACAGTTAA IR-1  
TTAAGTTTATTTATTTTCTTAATAAATGATTCCGTACAAAAGGATTTTTTTTTTAGTGAACGTGTCACAGTTAA IR-2

3300 3350

1500  
TTACTCCTTTTTCTTGTAAGACGAAGAAACAATTTCTATTTTCTCTACTATTTAGTACGACGACGAAGAATCA IR-1  
TTACTCCTTTTTCTTGTAAGACGAAGAAACAATTTCTATTTTCTCTACTATTTAGTACGACGACGAAGAATCA IR-2

3400

1550 1600  
AATTATCACTATATTTCTTCCTTTTCTACTTCTTCTTCCAAGTGCAGGAAAACCCCAAGGAGTTGCGGGTTTT IR-1  
AATTATCACTATATTTCTTCCTTTTCTACTTCTTCTTCCAAGTGCAGGAAAACCCCAAGGAGTTGCGGGTTTT IR-2

3450 3500

TTTCTACCAATTGGGGCC  
TTTCTACCAATTGGGGCC IR-1  
IR-2

Seul le brin non-codant (identique à l'ARN) est indiqué.

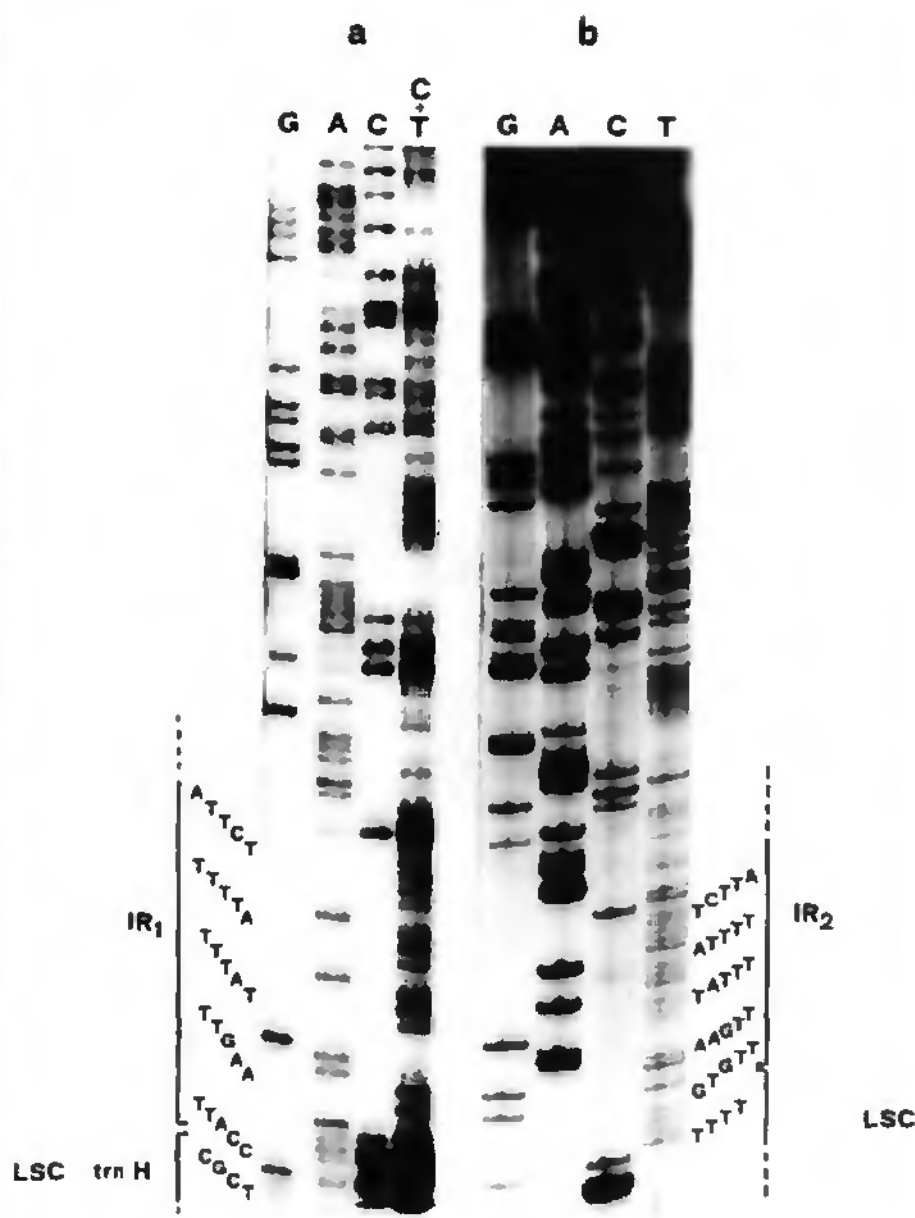
La partie structurale du trn H, ainsi que les séquences formant les IR sont encadrées. La numérotation des nucléotides est en continuité avec celle de la figure 6. Remarquer que le psb A et le trn H ont la même polarité.

Les séquences consensus des promoteurs de type procaryotique sont donnés. Pour l'IR-1, le brin ayant la même polarité que le trn H est indiqué, alors que par définition, c'est le brin opposé qui est indiqué pour l'IR-2; les 71 premiers nucléotides de la LSC-2 sont également donnés.

LSC-1 : extrémité de la grande région unique adjacente à l'IR-1

LSC-2 : " " " " " " " " à l'IR-2

Figure 8 : Autoradiographie de gels comportant les séquences de l'extrémité des IR-1 et IR-2



- a : Séquence partielle (selon Maxam et Gilbert, 1980) du fragment Sau3A-A (provenant du fragment Hind-O, voir figure 5), avec la jonction entre l'IR-1 et la grande région unique (LSC).
- b : Séquence partielle (selon Sanger et Coulson, 1980) du fragment Hae III 350 (provenant du fragment Hind/Pst 3.9, voir figure 5)
- G : guanine; A : adénine; C : cytosine; T : thymine

Figure 9 : Séquence de nucléotides de l'extrémité de l'IR-1

```

                                1650
GGCCCTCCCTTCACCACCCCATGCCGATGGTCTACGGGCTTCATAACTACTCCTCTTACTACAGGACGCTTAC

                                1700
CTA GCC AAC ATT TGA TCC GGC TCT ACC AAA TTT TTC TCG TTT ACT CCA ACA TTT CCC
GAT CGG TTG TAA ACT AGC CCG AGA TCG TTT AAA AAG ACC AAA TGA GGT TGT AAA GGG
TER Gly Val Asn Ser Gly Ala Arg Gly Phe Lys Glu Pro Lys Ser Trp Cys Lys Gly

                                1800
ACT TGT CCG ACT GTT GCC GAG CAG TTT TTG GAT ATC AAA CGG ACC TCC CCA GAA GGT
TGA ACA GGC TGA CAA CGG CTC CTC AAA AAC CTA TAG TTT GCC TGG AGG GGT CTT CCA
Ser Thr Arg Ser Asn Gly Leu Leu Lys Gln Ile Asp Phe Pro Gly Gly Trp Phe Thr

                                1850
AAT TTT AAA TGT GGC CGA TTT CCC CTC TTT TGC AAT TAG TTT CGC TAC AGC ACC TGC
TTA AAA TTT ACA CCG CCT AAA GGC GAG AAA ACG TTA ATC AAA CGC ATG TCG TCG ACG
Ile Lys Phe Thr Ala Ser Lys Gly Glu Lys Ala Ile Leu Lys Ala Val Ala Gly Ala

                                1900
TGC TCT AGC TAA TTG TCC ACC CTT TCC GAG TCT CAT TTC TAT GTT ATG TAT TGC CGT
ACG AGA TCG ATT AAC AGC TCG GAA AGC CTC ACA CTA AAG ATA CAA TAC ATA ACG GCA
Ala Arg Ala Leu Gln Gly Gly Lys Gly Leu Thr Ile Glu Ile Asn His Ile Ala Thr

                                1950
GCC TAA GGC CAT ATC GGT TGA AGT AGA TTC TTC TTT TTG ATC AAT CAA AAC CTC TTC
CGG ATT CCC GTA TAG CCA ACT TCA TCT AAG AAG AAA AAC TAG TTA GTT TTG GAG AAG
Gly Leu Pro Met Asp Thr Ser Thr Ser Glu Glu Lys Gln Asp Ile Leu Val Glu Glu

CCA AAC TGT AC
GGT TTG ACA TG
Trp Val Thr

```

La numérotation des nucléotides est en continuité avec celle de la figure 7.

La séquence de nucléotides complémentaires au brin codant, ainsi que celle d'acides aminés déduites, sont indiquées.

Remarquer que la polarité de cette ORF est opposée à celle du psb A et du trn H.

La physionomie générale de ce gène est semblable à la majorité des autres gènes d'ARNt chloroplastiques déjà séquencés chez d'autres organismes. Ce gène ne possède pas d'introns et ne code pas pour le triplet CCA présent à l'extrémité de l'ARNt fonctionnel. Comme pour les ARNt cytoplasmiques, ce triplet est donc ajouté lors d'un mécanisme post-transcriptionnel.

D'autre part, 67 nucléotides des 2 extrémités de la grande région unique, ainsi que 250 nucléotides de chaque IR ont été séquencés; ces séquences sont alignées et comparées sur la figure 7, alors que la figure 8 montre les photographies des gels à partir desquels ces séquences ont été lues. Nous voyons (figure 7) que les 2 IR ont une homologie parfaite sur tout le segment analysé (250 bp). Les 2 extrémités de la grande région unique n'ont rien de commun; les séquences présentées sur la figure 7 n'ont aucune ORF intéressante.

Finalement, en plus des 250 premiers nucléotides séquencés pour chaque extrémité des IR, nous avons également séquencé 365 nucléotides supplémentaires de l'IR-1 (figure 9).

L'analyse de cette séquence a montré qu'elle pourrait contenir la fin d'un gène codant pour une protéine. Ce gène aurait une polarité opposée à celle du psb A et du trn H, et se terminerait par le codon stop TAG en position 1692. Sa taille serait d'au moins 288 nucléotides (96 acides aminés), et il pourrait ainsi coder pour une protéine de  $M_r > 10'000$  (figure 9).

D'autres évidences expérimentales seraient bien entendu nécessaires pour confirmer l'existence de ce gène.

## D. DISCUSSION

### 1. Analyse de la séquence du psb A

La physionomie générale du psb A du soja est très semblable à celle du psb A de S.oleracea et N.debneyi (Zurawski et al, 1982). La longueur de la région transcrite (distance entre les sites les plus probables d'initiation et de terminaison de la transcription) est d'environ 1.2 kbp, ce qui est en accord avec la taille de l'ARN majeur qui interagit avec le psb A (Palmer et al, 1982). Ces auteurs comparent les transcripts du psb A de diverses espèces d'angiospermes, dont le soja et l'épinard.

Ils ont observé que pour le soja, deux produits de transcription mineurs, de 1.0 et 0.2 kb , interagissent aussi avec le psb A. Ils pensent cependant que ces ARN mineurs proviennent d'une dégradation de l'ARN majeur de 1.2 kb ,, le site spécifique de coupure se trouvant dans la partie 3' de la région codante.

Si nous comparons les séquences du psb A du soja et de l'épinard autour de la position 900 (figure 6), où une coupure préférentielle pourrait avoir lieu, nous voyons qu'il n'y a pas de différences particulières; le chloroplaste de soja doit donc contenir une ou des ribonucléase(s) qui diffèrent dans leur spécificité de celle des chloroplastes d'épinard.

### 1.1 Taille du produit primaire de traduction

La taille du produit de traduction de ce gène est nettement supérieure à celle estimée sur la base de la mobilité électrophorétique du photogène chez d'autres plantes (32'000-35'000 daltons). Le même cas se présente chez S.oleracea et N.debneyi (Zurawski et al, 1982)

Comme il existe un second codon d'initiation de traduction 108 nucléotides après le premier codon, ces auteurs ont postulé que la traduction pourrait commencer au niveau du second ATG, ce qui donnerait une protéine de Mr 34'600. Cette valeur serait beaucoup plus en accord avec celle estimée sur la base de la mobilité électrophorétique de cette protéine. Ils ont cependant conclu que la traduction commence au niveau du premier codon ATG, vu la très grande homologie des séquences dans la région codante de ce gène chez S.oleracea et N.debneyi (>95%). Ceci est renforcé par le fait que les séquences d'acides aminés de cette protéine chez ces deux espèces sont identiques.

Comme la partie N-terminale du photogène du soja est homologue à 100% avec celles de S.oleracea et N.debneyi, nous pouvons admettre que les 36 premiers codons sont traduits et donc nécessaires au premier produit de traduction.

Cependant, McIntosh (1983) a très récemment démontré que le psb A de Zea mays et Amaranthus hybridus code pour une protéine de 317 acides aminés seulement. Selon ces résultats préliminaires, il semble possible que les 36 premiers acides aminés ne sont pas indispensables à la protéine fonctionnelle de Mr 32'000.

D'autres évidences expérimentales sont cependant nécessaires pour résoudre ce problème. Zurawski et al (1982) discutent d'ailleurs de plusieurs possibilités pour expliquer cette contradiction.

## 1.2 Comparaison de la structure primaire du photogène de Glycine max avec celles de Spinacia oleracea et Nicotiana debneyi

---

Il y a une très grande homologie de séquence nucléotidique entre ces 3 espèces (figure 6; table 3). Il y a 73 nucléotides différents entre G.max et S.oleracea, et 78 entre G.max et N.debneyi, correspondant à une homologie d'environ 93%, alors qu'il n'y a que 47 nucléotides différents entre S.oleracea et N.debneyi (homologie >95%).

Cependant, contrairement à la situation chez S.oleracea et N.debneyi, où les 47 modifications sont silencieuses, chez G.max, 3 de ces modifications provoquent le changement de 2 acides aminés. Ces 3 modifications se trouvent dans la région COOH-terminale de la protéine, la première en position 1040 (acide aminé 347; C→A⇒Asp→Glu), la deuxième et la troisième en position 1051 et 1052 (acide aminé 351; T→C et T→A⇒Ile→Thr) (figure 6).

Jusqu'à présent, la séquence du photogène est connue chez 3 espèces. La caractéristique commune de ces séquences est l'absence de résidu lysine; d'autre part, la structure primaire de cette protéine est fortement conservée, puisqu'elle est identique chez S.oleracea et N.debneyi, et diffère seulement de 2 acides aminés chez G.max.

Il est donc intéressant de se demander si ces 2 changements chez le soja seraient suffisant pour influencer le rôle physiologique de cette protéine, et notamment son comportement face à des substances utilisées comme herbicides.

A ce jour, aucune variété de soja n'est connue pour être résistante à de tels produits. Nous pouvons donc admettre que ces 2 changements ne modifient pas radicalement le rôle physiologique du photogène, surtout que dans le premier cas, il y a simplement remplacement d'un acide aminé acide (Glu) par un autre acide aminé acide (Asp), et que

Table 3 : Comparaison des séquences du psb A chez G.max (G.m), S.oleracea (S.o) et N.debneyi (N.d)

|                        | Partie 5'     |               | Partie structurale        |                             |  |               | Partie 3' |
|------------------------|---------------|---------------|---------------------------|-----------------------------|--|---------------|-----------|
|                        | Homologie (%) | Homologie (%) | Nucléotides<br>différents | Acides aminés<br>différents |  | Homologie (%) |           |
| G.m - S.o              | 52.0          | 93.2          | 73                        | 2                           |  | 63.0          |           |
| G.m - N.d              | 54.0          | 92.6          | 78                        | 2                           |  | 57.0          |           |
| S.o - N.d <sup>a</sup> | 75.0          | 95.0          | 47                        | 0                           |  | 38.0          |           |

a : d'après Zurawski et al, 1982

dans le deuxième cas (Ile au lieu de Thr), la seule différence importante entre ces 2 acides aminés est la présence d'un radical hydroxyle sur la thréonine.

Cependant, le fait que ces 2 modifications soient très proches l'une de l'autre (4 acides aminés), à l'extrémité COOH-terminale de la protéine, pourrait au contraire nous laisser supposer qu'elles ont une influence sur certaines propriétés du photogène.

Vu le peu d'information concernant le rôle physiologique du photogène, il est difficile de trancher entre ces 2 hypothèses. Il serait par contre très intéressant de comparer la structure primaire du photogène entre des espèces végétales résistantes aux herbicides et des espèces sensibles à de tels produits.

### 1.3 Analyse des séquences des régions adjacentes au psb A

#### 1.3.1 Analyse de la région 5'

Il a été démontré que les promoteurs reconnus par l'ARN-polymérase de Escherichia coli étaient constitués de régions dont les séquences étaient fortement conservées; elles sont localisées environ 10 nucléotides ("région -10" = "Pribnow box") et 35 nucléotides ("région -35" = site de reconnaissance) en amont du site de départ de la transcription (Hawley et McClure, 1983). Ces séquences consensus sont les suivantes :

"Région -10"    tg TAtAaT

"Région -35"    ttcTTGACa

(les lettres en majuscules indiquent une fréquence d'apparition d'au moins 54% et celles en minuscules d'au moins 39%)

Ces séquences ont également été trouvées en amont de plusieurs gènes chloroplastiques ou d'autres origines (Hawley et McClure, 1983; Steinmetz et al, 1983). Comme nous pouvons le voir sur la figure 6, ces séquences sont aussi présentes dans le cas du psb A. Pour le soja, elles débutent en position -94 (TATACT) et -118 (TTGACA). Chez S. oleraceus, elles se trouvent en position -96 et -120, et chez N. debneyi, en position -98 et -122.

Il est frappant de constater que :

- 1) les séquences des régions -10 et -35 sont absolument identiques pour le psb A de ces 3 espèces
- 2) elles se retrouvent chez les 3 espèces à la même distance du site d'initiation de la traduction ( $-96 \pm 2$  et  $-120 \pm 2$ )
- 3) l'intervalle qui sépare ces 2 séquences dans les 3 cas est non seulement parfaitement identique en taille (24 nucléotides), mais encore homologue à plus de 90%
- 4) Entre ces 3 espèces, l'homologie des séquences allant depuis le codon ATG jusqu'aux signaux de reconnaissance est de plus de 85%, alors que cette valeur tombe à 35% en amont de la "région -35".

(voir conclusion)

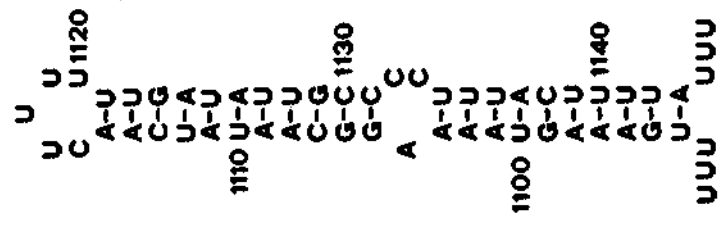
Il faut finalement mentionner qu'en position -37, le tétranucléotide -GGAG- est complémentaire à la séquence caCCTCc présente à l'extrémité 3' des ARNr 16S de type procaryotique (Schwarz et Kössel, 1979) et pourrait fonctionner comme site de fixation de l'ARNm sur le ribosome (Shine et Dalgarno, 1974; Sherer et al, 1980).

### 1.3.2 Analyse de la région 3'

L'analyse de la région 3' du psb A du soja (figure 6) montre qu'elle peut coder pour un ARNm susceptible de former une structure secondaire en "épingle à cheveux", comme représentée sur la figure 10. Cette structure commence en position 1094 et se termine en position 1143 (plus précisément : 1094 - 1115  $\wedge$  1121 - 1143).

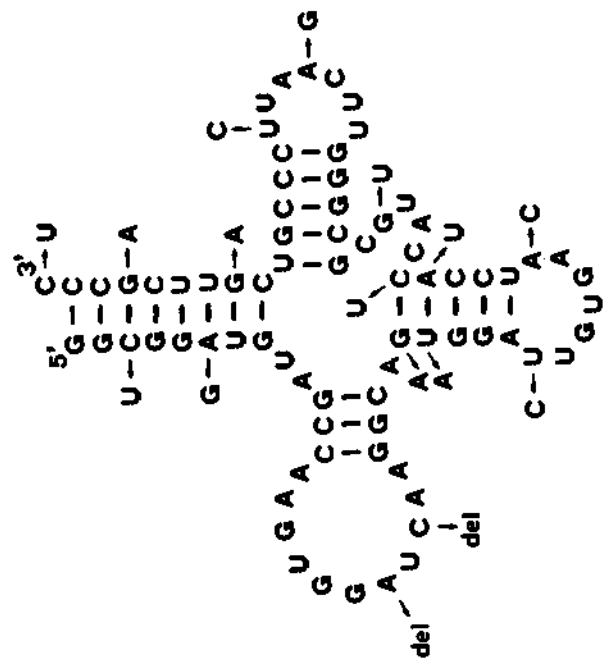
De telles structures apparaissent au niveau des sites procaryotiques de terminaison de la transcription (Rosenberg et Court, 1979) et il a été démontré qu'elles étaient nécessaires pour cette terminaison (Zurawski et al, 1981). Ces auteurs ont notamment montré que dans le cas du rbc L chez l'épinard, la transcription s'arrête quelques bases après cette structure en "épingle à cheveux".

Figure 10 : Structure secondaire en épingle à cheveux formée par l'extrémité 3' de l'ARNm codant pour le photogène de soja



La numérotation des nucléotides est la même que celle de la figure 6.

Figure 11 : Structure en feuille de trèfle (non modifiée) de l'ARNt<sup>His</sup><sub>GUG</sub> du chloroplaste de soja



Cette structure est déduite de la séquence de nucléotides du trn H.  
La guanine à l'extrémité 5' est considérée comme la position 1.

Les flèches indiquent les nucléotides présents à cette position dans l'ARNt<sup>His</sup><sub>GUG</sub> du chloroplaste d'euglène (Hollingsworth et Hallick, 1982).

Dans le cas du psb A du soja, nous savons que l'ARNm transcrit de ce gène à une taille d'environ 1'200 nucléotides (Palmer et al, 1982) et que la transcription doit commencer quelques bases après la "Pribnow box" (position -90 environ). Nous pouvons donc en déduire que la transcription doit s'arrêter environ 100 nucléotides après le codon d'arrêt de la traduction (TAA), c'est-à-dire en position 1150 environ, ce qui correspond bien avec la fin de la structure en "épingles à cheveux" (figure 6).

Si nous comparons la région 3' du psb A chez G.max, S.oleracea et N.debneyi en alignant les séquences pour avoir une homologie maximale (figure 6), nous obtenons 63% d'homologie entre G.max et S.oleracea, et 57% entre G.max et N.debneyi (table 3). Malgré cette homologie relativement faible, les séquences des régions 3' du psb A de S.oleracea et N.debneyi peuvent aussi donner naissance à des ARNm susceptibles de former des structures secondaires en "épingles à cheveux" (figure 6).

## 2. Analyse de la séquence se trouvant en aval du psb A, contenant le trn H et l'extrémité de la région inversée adjacente

### 2.1 Analyse de la séquence du trn H

La présence d'un trn H à proximité de la fin d'une des régions répétitives inversées a été mis en évidence sur les génomes chloroplastiques de S.oleracea (Bohnert et al. 1982) et Phaseolus vulgaris (Mubumbila et al, 1983). Chez Zea mays, le trn H a été localisé à l'intérieur des régions répétitives inversées ; il se trouve donc en 2 exemplaires sur le génome chloroplastique de cette plante et est situé de plus à cheval sur un gène de polarité inverse qui est transcrit en un ARN de 1.6 kbp (Schwarz et al, 1981).

La partie structurale du trn H du soja est entièrement comprise dans la grande région unique, alors que sa partie 3' (contenant un site potentiel de terminaison de la transcription) se trouve à l'intérieur de la région inversée adjacente (figure 7).

La séquence de ribonucléotides déduite de celle du trn H est identique à 100% avec celle de l'ARNt<sup>His</sup> du chloroplaste de maïs. Toutes les positions hautement conservées (Sprinzl et Gauss, 1982) sont maintenues dans la séquence de ce trn H. Sur la figure 11, nous comparons les ARNt<sup>His</sup><sub>GUG</sub> chloroplastiques de soja et d'Euglena gracilis (Hollingsworth et Hallick, 1982). Il y a environ 80% d'homologie; dans les 2 cas, la région bicaténaire D est raccourcie et la boucle variable est de même taille. Le triplet terminal en 3' CGA n'est pas codé sur les gènes, une propriété que les ARNt chloroplastiques partagent avec les gènes eucaryotiques (Sprinzl et Gauss, 1982). Les 3 trn H chloroplastiques séquencés jusqu'à maintenant codent tous pour l'ARNt<sup>His</sup><sub>GUG</sub>. Selon de multiples expériences d'hybridation en utilisant plusieurs génomes chloroplastiques différents (pour les références, voir Bohner et al, 1982), aucun autre trn H n'a été identifié, c'est-à-dire qu'un gène possible pour un ARNt<sup>His</sup><sub>AUG</sub> n'a pas encore été détecté. Néanmoins nous pouvons voir sur la table 2 que le codon CAT (CAU) est fréquemment utilisé pour la traduction du photogène du soja et de l'épinard. Ceci est en accord avec la possibilité d'un flottement dans l'appariement au niveau du troisième nucléotide d'un codon (wobble hypothesis).

Les 3 trn H chloroplastiques séquencés jusqu'à présent ont chacun un environnement différent. Le trn H du soja se trouve directement placé à la suite d'une des régions répétitives inversées; il est éloigné de 241 nucléotides du codon stop du psb A et est transcrit dans le même sens que le photogène. En amont de la partie structurale du trn H, en positions 1193 et 1198, nous trouvons des séquences que l'on peut qualifier de séquences promoteurs. En aval, à l'intérieur de l'IR, il y a une séquence susceptible de former une structure secondaire en épingle à cheveux (figure 7).

Le trn H de maïs est situé à l'intérieur des IR et se trouve à cheval sur un gène de protéine de polarité inverse (Schwarz et al, 1981). Le trn H d'Euglena gracilis est le second gène d'un ensemble (cluster de gènes d'ARNt arrangés en tandem (Hollingsworth et Hallick, 1982), qui sont probablement co-transcrits et sous le contrôle d'un seul opérateur. La régulation de la transcription de ces trn H doit être différente dans les 3 cas, ce qui est très intéressant.

Finalement, si nous comparons les séquences de plusieurs isoaccepteurs correspondants (ARNt<sup>His</sup><sub>GUG</sub>) provenant d'autres organismes (table 4), nous voyons que l'homologie est totale avec l'isoaccepteur du chloroplaste de maïs, élevée avec celui du chloroplaste d'euglène (79%), mais devient relativement faible pour les 3 autres isoaccepteurs.

Nous pouvons résumer ces résultats ainsi :

- 1) L'homologie de séquences entre les ARNt<sup>His</sup> chloroplastiques est élevée; ceci est aussi vrai pour d'autres ARNt chloroplastiques, notamment l'ARNt<sup>Phe</sup>, l'ARNt<sup>Thr</sup> et l'ARNt<sup>Met</sup> (Steinmetz et al, 1983)
- 2) Ce gène possède à la fois des caractéristiques procaryotiques (65% d'homologie avec E.coli) et eucaryotiques (le triplet CCA en 3' est ajouté après la transcription).

Table 4 : Homologies de séquences (en %) entre les ARNt<sup>His</sup><sub>GUG</sub> de différentes origines et celui du chloroplaste du soja

|                 | <u>ct maïs</u> <sup>1</sup> | <u>ct euglène</u> <sup>2</sup> | <u>E.coli</u> <sup>3</sup> | <u>mt levure</u> <sup>3</sup> | <u>drosophile</u> <sup>3</sup> |
|-----------------|-----------------------------|--------------------------------|----------------------------|-------------------------------|--------------------------------|
| ct <u>C.max</u> | 100                         | 79                             | 65                         | 49                            | 45                             |

ct : chloroplastique

mt : mitochondrial

1. Schwarz et al, 1981

2. Hollingsworth et Hallick, 1982

3. Sprinzl et Gauss, 1982

## 2.2 Analyse des séquences des 2 extrémités des IR adjacentes à la grande région unique

---

Des expériences d'hybridations croisées ont montré que les régions répétitives inversées des génomes chloroplastiques étaient apparentées, révélant de considérables homologies de séquences (Palmer et Thompson, 1982). Nos résultats de séquençage définissent pour la première fois le début exact de l'IR-1 et de l'IR-2 sur le génome chloroplastique d'une angiosperme. Une comparaison avec l'IR du chloroplaste de Zea mays révèle qu'un petit segment de 32 nucléotides est homologue à 91% avec une partie de l'ADNct de soja (position 1510-1542). L'homologie de séquence continue probablement au-delà de ces positions (Schwarz et al, 1981). Il est intéressant de constater que cette homologie se situe à des endroits différents à l'intérieur de deux IR d'origine différente. Cette observation (ainsi que la localisation différente du trn H) montre que durant l'évolution, non seulement les régions uniques ont subi des réarrangements de séquences, mais aussi les régions relativement bien conservées formant les IR.

En outre, jusqu'à présent, aucun gène fonctionnel codant pour une protéine n'a réellement été mis en évidence à l'intérieur des IR des différents ADNct étudiés. Chez le soja cependant, les résultats des expériences décrites au chapitre III suggèrent fortement que le tuf A est localisé à l'intérieur des IR.

D'autre part, l'ORF mise en évidence à l'intérieur de l'IR-1 doit probablement correspondre à la fin d'un gène. En effet, elle est trop longue (près de 300 nucléotides) pour que son existence soit simplement due à un arrangement hasardeux de nucléotides. (voir discussion finale

## E. CONCLUSION

Il est intéressant de constater qu'il existe une très grande homologie dans l'organisation générale d'un gène codant pour une protéine (photogène) chez 3 espèces phylogénétiquement assez éloignées (Leguminosae, Solanaceae et Chenopodiaceae); cette homologie se retrouve aussi bien dans la partie structurale du gène, que dans les régions de régulation, et notamment au niveau de la "leader sequence" (séquence comprise entre les promoteurs et le codon d'initiation), ainsi que dans les séquences portant les signaux de reconnaissance et d'attachement des ARN polymérases.

Ces faits suggèrent que la structure primaire de cette protéine est assez rigide et doit être importante pour son incorporation et son orientation dans la membrane thylacoïdale. Hoffman-Falk et al., (1982) ont étudié les profils de protéolyse partielle de cette protéine in situ chez plusieurs espèces d'angiospermes et Chlamydomonas reinhardtii. Ils ont toujours obtenu des profils identiques et ont suggéré que les interactions photogène-membrane thylacoïdale (qui dépendent de la structure tertiaire de la protéine) sont semblables chez toutes ces espèces.

Par comparaison, la structure primaire d'une autre protéine codée par le génome chloroplastique, la LSU, est plus souple (Zurawski et al., 1981). En effet, la divergence entre les séquences de nucléotides de ce gène chez l'épinard et le maïs est de 16%, provoquant une différence de 10% dans la composition en acides aminés.

Une autre enzyme chloroplastique soluble, la glycéraldéhyde-3-phosphate déshydrogénase, a également une structure très variable suivant son origine (Hoffman-Falk et al., 1982).

Les contraintes dans la structure primaire des protéines membranaires semblent donc beaucoup plus fortes que pour des protéines stromatiques.

D'un autre côté, nous avons vu que l'organisation structurale de 2 gènes chloroplastiques différents (psb A et trn H) montre une certaine similitude. Dans les 2 cas, il a été possible de mettre en évidence la présence de régions de régulation de type procaryotique, aussi bien au niveau des sites d'initiation de la transcription, qu'au niveau des sites de terminaison de la transcription.

Il est particulièrement intéressant de noter que le trn H possède à la fois des caractéristiques procaryotiques et eucaryotiques; ceci suggère que le chloroplaste est une entité génétique définie, différente à la fois des systèmes procaryotiques et eucaryotiques, mais partageant avec ceux-ci certaines propriétés (Steinmetz et al, 1983). Cette hypothèse est renforcée par la découverte récente de l'existence de gènes chloroplastiques fragmentés (Stiegler et al, 1982; Montandon et Stutz, 1983); cette organisation structurale des gènes est caractéristique des systèmes eucaryotiques.

## F. RESULTATS DE DERNIERE MINUTE

### Analyse des séquences faisant la jonction entre les IR et la LSC

Au moment où nous avons terminé la rédaction de cette thèse, nous nous sommes aperçus que chez le tabac, Sugita et Sugiura (1983) avaient très récemment démontré l'existence d'un gène codant pour une protéine ribosomique appelée CS19 ( $M_r = 10'000$ ; 92 acides aminés); or ce gène est localisé sur l'ADNct de tabac à l'extrémité de l'IR-2.

Nous avons alors comparé les séquences de ces régions chez le soja et le tabac et avons trouvé qu'elles étaient très semblables. Ce gène existe donc aussi chez le soja, et il se trouve à cheval sur l'IR-2 et la LSC. La séquence des 140 premiers nucléotides de ce gène (ainsi que celle d'acides aminés déduite) est présentée sur la figure 12.

D'autre part, comme mentionné au chapitre V, le trn H de Zea mays est à cheval sur un gène de polarité inverse; Le début de la séquence d'acides aminés de ce gène est connue (Schwarz et al, 1981). Une comparaison avec la séquence d'acides aminés du CS19 de tabac montre que, chez le maïs également, cette protéine est codée à l'extrémité des IR.

Cependant, comme c'est le cas pour le trn H, le CS19 se trouve à l'intérieur des IR et est donc présent à double exemplaires chez le maïs. La figure 14 montre le début de la séquence du CS19 chez E.coli, le tabac, le soja et le maïs.

D'un autre côté, Sugita et Sugiura (1983) ont également démontré que ce gène (CS19) se trouvait entre deux autres ORF, de même polarité que le CS19. L'ORF située en amont du CS19 se termine à 60 nucléotides du début de ce gène. Chez le soja, cette ORF existe aussi. Nous n'avons séquencé que les 240 premiers nucléotides de l'IR-2, alors que 536 nucléotides ont été séquencés dans l'IR-1.

Comme l'homologie de séquence est parfaite entre les 240 premiers nucléotides de chaque IR, il est logique de penser que ces séquences sont identiques au-delà de ces positions. En admettant cela, nous avons donc la séquence totale du gène se trouvant en amont du CS19. Il code pour une protéine de 52 acides aminés, correspondant à une Mr d'environ 5700 (figure 12). Il n'est cependant pas possible de prouver que ce gène est fonctionnel (par commodité, nous l'appellerons ORF-A).

En outre, 26 nucléotides en amont de l'ORF-A, se trouve la fin de l'ORF décrite au chapitre V (nous appellerons cette ORF, ORF-B).

Finalement, Sugita et Sugiura (1983) ont également démontré l'existence d'une troisième ORF (que nous appellerons ORF-C) à 40 nucléotides en aval du CS19; nous pouvons raisonnablement supposer qu'elle existe aussi chez le soja.

En résumé, nous avons donc la situation suivante : (figure 13)

Figure 13 : Représentation schématique de la région de l'ADNct de soja faisant la jonction entre la LSC et l'IR-2

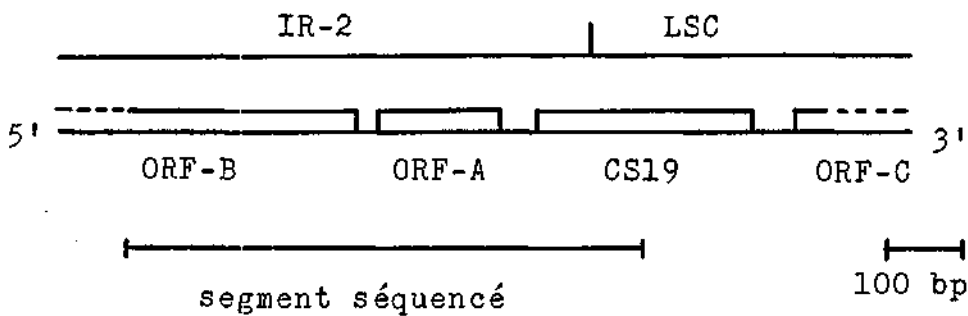


Figure 12 : Séquence de nucléotides des régions faisant la jonction  
entre la LSC et les IR-1 et IR-2

---

La numérotation des nucléotides correspond à celle donnée sur les figures 5, 6 et 7.

Pour les séquences faisant partie de la LSC-1 et de l'IR-1, seul un des brins est indiqué (celui qui est identique au trn H).

Pour les séquences faisant partie de la LSC-2 et de l'IR-2, les deux brins sont indiqués.

Les régions comportant des ORF sont données; les séquences d'acides aminés déduites de celles de nucléotides sont indiquées sous les séquences de nucléotides (par le code à une lettre).

La partie structurale du trn H est encadrée.

Le début des IR est marqué par un crochet (position 1380 : IR-1 ; position 3280 : IR-2).

LSC-1 et LSC-2 : comme figure 7

CS19, ORF-A et ORF-B : voir texte

Figure 12 : Séquence de nucléotides des régions faisant la jonction entre la LSC et les IR-1 et IR-2

|                                                                             |      |       |
|-----------------------------------------------------------------------------|------|-------|
| 1250                                                                        | 1300 |       |
| AAAAGCAAAAAACAATGATAACGAACGAAAGGATAGAATTTTATATATAGATCATTTTTACATAGTATAAGGGC  |      | LSC-1 |
| 1350                                                                        |      |       |
| GGATGTAGCCAAGTGGATCAAGGCAGTGGATTCTGAATCCACCATTGGCGCGCTTCAATTECCGTCGTTCCGCCA |      | LSC-1 |
| 3250                                                                        |      |       |
| GGCCAATCATTGTAGCTATAATCGTAGATCGTCTGGACCAACTTATTATTATATCTTTTTCCGCTTTTCTC     |      | LSC-2 |
| CCGCTTAGTAACATCCATATTACCATCTACGAGACCTGCTTCAATAATAATATAGAAAAAGCCGAAAACAC     |      |       |
| G I M T P I I T S T R S W T I I I D K E A K T                               |      |       |
| 1400                                                                        | 1450 |       |
| TTAAGTTTATTTATTTTCTTAATAAATGATTCCGTACAAAAGGATTTTTTTTTTACTGAACGTCTCACAGTTAA  |      | IR-1  |
| 3300                                                                        | 3350 |       |
| TTAAGTTTATTTATTTTCTTAATAAATGATTCCGTACAAAAGGATTTTTTTTTTACTGAACGTCTCACAGTTAA  |      | IR-2  |
| AATTCAAATAAATAAAAAAGAATTATTTACTAAGCGATGTTTTCTAAAAAAAATCACTTCCACAGTCTCAATT   |      |       |
| N L K N I K R L L H N A V F P N K R L S R T M                               |      |       |
| 1500                                                                        |      |       |
| TTACTCCTTTTTCTTGTAAGACCAAGAAACAATTTCTATTTTCTCTACTATTTAGTACGACGACCAAGAATCA   |      | IR-1  |
| 3400                                                                        |      |       |
| TTACTCCTTTTTCTTGTAAGACCAAGAAACAATTTCTATTTTCTCTACTATTTAGTACGACGACCAAGAATCA   |      | IR-2  |
| AATCAGCAAAAAGAACATTTCTGCTTCTTGTAAGATAAAGAGATGATAAATCATGCTGCTGCTTCTTACT      |      |       |
| TER K T R R R L I L                                                         |      |       |
| 1550                                                                        | 1600 |       |
| AATTATCACTATATTTCTTCTTTTCTACTTCTTCTTCCAAGTGCAGGAAAACCCCAAGCAGTTCCCGCTTTT    |      | IR-1  |
| 3450                                                                        | 3500 |       |
| AATTATCACTATATTTCTTCTTTTCTACTTCTTCTTCCAAGTGCAGGAAAACCCCAAGCAGTTCCCGCTTTT    |      | IR-2  |
| TTAATAGTCAATAAAGAAGCAAAAAGATGAAGAAGAAGGTTCCAGCTCTTTTCCGCTTCTCAACCCCAAAA     |      |       |
| K D S Y K K R R R S R R G L A P F G W P T A P K                             |      |       |
| 1650                                                                        |      |       |
| TTTCTACCAATTGGGGCC                                                          |      | IR-2  |
| 3550                                                                        |      |       |
| TTTCTACCAATTGGGGCCCTCCCTTCACCACCCCATGCGGATGGCTACGGCGCTTCATAACTACTCCTCTTAC   |      | IR-1  |
| AAAGATGGTTAACCCCGGGAGGGAAGTGGTGGGGGTACGGCTACCAGATGCCCCAAGTATTGATGAGGAGAATC  |      |       |
| K R C I P A R C E C C C H P H D V P N M                                     |      |       |
| 3600                                                                        |      |       |
| TACAGGACGCTTACCTAGCCAACATTTAGATCCGGCTCTACCTAAATTTTTCTCGTTTACTCCAACATTTCCCA  |      | IR-1  |
| ATGCTCTGCGAATGATCGGTTGTAATCTAGCCCCGAGATGGATTTAAAAAGACCAATGAGGTTGTAAGGGCT    |      |       |
| TER C F K E P R S W C K C                                                   |      |       |
| 3700                                                                        |      |       |
| CTTGTCGGACTGTTGCGGACGAGTTTTTGCATATCAAACGGACCTCCCCAGAACGTAATTTTAAATCTGCCCCA  |      | IR-1  |
| GAACAGGCTGACAACGGCTCGTCAAAAACCTATAGTTTGGCTGGAGGGCTCTTCCATTAAAAATTTACACCGGCT |      |       |
| S T R S N G L L K D I D F P C G W F T I R F T A S                           |      |       |
| 3750                                                                        |      |       |
| TTTCCCTCTTTTGCATTAGTTTCCGTACAGCACCTGCTGCTCTAGCTAATTGTCACCCCTTTCCGAGTCTGA    |      | IR-1  |
| AAAGGGGAGAAAACGTTAATCAAACGATGCTGTCAGCAGCAGATCGATTAAACAGTCCGAAAGGCTCACACT    |      |       |
| R C E K A I L K A V A C A A R A L Q C G K G L T I                           |      |       |
| 3850                                                                        |      |       |
| TTTCTATGTTATGTATTGCGCTGCCCTAAGGCCATATCGGTTGAAGTAGATTCTTCTTTTTCATCAATCAAACC  |      | IR-1  |
| AAAGATACAATACATAACGGAACGGATTCCCTATAGCCAACTTCATCTAAGAAGAAAACTAGTTAGTTTTCG    |      |       |
| E I N H I A T C L P M D T S T S E E K Q D I L V                             |      |       |
| 3900                                                                        |      |       |
| TCTTCCCAAAGTGTAC                                                            |      | IR-1  |
| AGAAGGCTTTGACATG                                                            |      |       |
| E E W V T                                                                   |      |       |

trn H

CS19

ORF-A

ORF-B

Il est frappant de constater que toutes ces ORF sont de polarité identique, suggérant que si elles correspondent réellement à des gènes fonctionnels, ceux-ci pourraient être transcrits en même temps sous la forme d'un même ARNm polycistronique. De telles situations sont fréquentes chez les gènes procaryotiques, et notamment dans le cas de l'opéron epc d'E.coli, lequel contient les gènes codant pour dix protéines ribosomiques différentes (Ceretti et al, 1983).

Il serait évidemment intéressant d'étudier ce problème chez le soja, d'autant plus que chez le maïs, l'ARNm codant pour le CS19 a une taille d'environ 1.6 kbp, alors que la taille de la partie structurale du CS19 est de 276 nucléotides (Sugita et Sugiura, 1983).

Cette observation renforce encore l'hypothèse de l'existence d'un ARNm polycistronique.

Toutes ces observations nous font penser que , même si les séquences des IR-1 et IR-2 sont identiques, à l'endroit de jonction des IR et de la LSC (et sans doute aussi de la SSC), ces séquences peuvent coder pour des gènes différents de polarité inverse et partiellement à cheval les uns sur les autres.

Figure 14 : Comparaison du début de la séquence d'acides aminés du CS19 de E.coli, de tabac, de soja et de maïs

---

|                     |                                                  |
|---------------------|--------------------------------------------------|
| E.coli <sup>a</sup> | PRSLKKGPFIDLHLLKKVEKAVESGDKKPLRTWSR-STIFPDRMIG   |
| Tabac <sup>a</sup>  | MTPRSLKKNPFVANHLLKKIDKLNTKAEKEIIVTWSRASTIIP-RMIG |
| Soja                | MT-RSLKKNPFVANHLLRKINKLNTKAEKDIIITWSRTSTIIP-TMIG |
| Maïs <sup>b</sup>   | MT-R-KKTNPFVARHLA-KIEKVNMGKEKEIIVTWS             |

a : Sugita et Sugiura (1983)

b : Schwarz et al (1981)

## VI. DISCUSSION FINALE

Pour revenir brièvement sur le problème de l'expression des gènes des protéines de réserve des graines, il faut encore mentionner que cette étude est très avancée chez 3 espèces de Légumineuses, c'est-à-dire le soja (Schuler et al, 1983a et b), le pois (Lycett et al, 1983) et le haricot (Sun et al, 1981; Slightom et al, 1983).

Chez ces 3 espèces, le composé majeur 7S ( $\beta$ -conglycinine chez le soja; vicilline chez le pois; phaseoline chez le haricot) est formé de plusieurs sous-unités différentes. Or il a été démontré que les sous-unités du composé 7S d'une de ces espèces (p.ex. les sous-unités  $\alpha'$ ,  $\alpha$  et  $\beta$  de la  $\beta$ -conglycinine du soja) montrent une similitude dans leur profil de protéolyse partielle et leur contenu en acides aminés.

De plus, la plupart des gènes codant pour ces protéines ont été partiellement ou totalement séquencés (voir références ci-dessus). Or une analyse comparative de ces séquences a montré qu'il existe non seulement une très grande homologie de séquence entre 2 sous-unités formant le composé 7S d'une même espèce, mais encore que ces séquences sont hautement conservées dans les sous-unités du composé 7S d'origine différente, c'est-à-dire entre la  $\beta$ -conglycinine, la vicilline et la phaseoline (Lycett et al, 1983).

Ces observations suggèrent qu'à l'intérieur d'un groupe phylogénétique donné (Légumineuses), la structure primaire de ces protéines est assez rigide et par conséquent nécessaire à leur fonction.

Si l'on admet que ces protéines sont dégradées spécifiquement par des enzymes protéolytiques lors de la germination de la graine, nous pourrions supposer que les régions où les séquences d'acides aminés sont bien conservées sont des endroits privilégiés, reconnus par ces enzymes protéolytiques.

Le grand problème à résoudre reste cependant la régulation de l'expression de ces gènes. En effet, s'il était possible d'influencer directement l'expression de ces gènes, il suffirait alors d'allonger de quelques jours seulement la période de biosynthèse des protéines de réserve, pour obtenir des graines beaucoup plus riches en protéines.

L'analyse de l'organisation structurale et génétique du génome chloroplastique de soja nous a permis de répondre à un certain nombre de questions. Nous avons notamment pu confirmer que le génome chloroplastique des Légumineuses était très semblable à celui d'autres angiospermes mais qu'il en différait cependant par l'inversion d'un grand segment d'ADN dans la grande région unique.

Malheureusement, plusieurs problèmes restent encore en suspens. Un de ces problèmes est le rôle exact que jouent les régions répétitives inversées dans les génomes des plantes supérieures.

D'après Palmer et Thompson (1982), ces régions semblent considérablement stabiliser le génome chloroplastique en empêchant les réarrangements de séquences qui ont lieu chez les espèces qui en sont dépourvues, comme par exemple le pois.

Cette hypothèse est confirmée par l'analyse d'un autre génome chloroplastique, aussi dépourvu d'IR, celui d'Euglena gracilis. En effet, suivant les souches d'euglènes analysées, l'ADNct de cet organisme semble avoir subi passablement de réarrangements de séquences, comprenant des duplications et des délétions de certains fragments d'ADN (Koller et Delius, 1982a).

D'autre part, parallèlement à cette caractéristique (absence d'IR), les génomes chloroplastiques de Pisum sativum (Palmer et al, 1982b) et d'Euglena gracilis (Jenni et al, 1981; Schlunegger et al, 1983) ont un autre point commun, c'est-à-dire la présence d'une petite région (2-3 kbp) de taille variable, qui pourrait être un lieu privilégié de petites délétions et insertions.

Or chez Euglena gracilis, cette région est très proche de l'origine de répllication de l'ADNct (Koller et Delius, 1982b).

L'existence de telles régions dans les génomes chloroplastiques possédant des IR n'a jusqu'à présent pas été démontrée. Nous avons effectué des expériences d'hybridations croisées entre la zone variable d'euglène et l'ADNct total de soja et de pois; les résultats de ces expériences préliminaires ont toujours été négatifs.

Une autre observation intéressante concernant les IR est celle faite par Palmer (1983). Cet auteur a en effet démontré que les régions uniques de l'ADNct de Phaseolus vulgaris pouvaient exister sous deux orientations différentes par rapport aux IR. L'existence d'un tel phénomène implique la nécessité d'une coupure spécifique de la molécule d'ADNct à l'intérieur des IR.

Comme expliqué au chapitre III, il n'est pas possible de démontrer l'existence d'un tel phénomène chez le soja, mais il doit également avoir lieu chez cette espèce, et est sans doute caractéristique des génomes chloroplastique possédant des IR. Actuellement, les conséquences de cette recombinaison intramoléculaire ne sont pas connues, bien qu'une situation semblable existe dans un plasmide de levure (Broach, 1982).

La question qui se pose après toutes ces observations est de savoir jusqu'à quel point tous ces phénomènes ne seraient pas plus ou moins liés, et ceci notamment par le fait que l'existence de région(s) de taille variable semble aller de pair avec l'absence d'IR (et inversement). Comme chez l'euglène, l'origine de répllication est très proche de la zone variable, il serait intéressant de localiser l'origine de répllication de l'ADNct de soja, pour pouvoir comparer ces 2 régions.

Un autre problème qui reste à élucider est la localisation du tuf A sur l'ADNct de soja. Des expériences sont actuellement en cours dans notre laboratoire pour confirmer les résultats décrits au chapitre III

La connaissance de la séquence de ce gène serait particulièrement intéressante, dans la mesure où actuellement, le seul tuf A chloroplastique séquencé est celui d'E.gracilis (Montandon et Stutz, 1983), organisme chez lequel ce gène est formé d'introns et d'exons.

Le problème suivant est celui du début de la traduction du psb A. Comme ceci a déjà été discuté au Chapitre V, nous n'y reviendrons que brièvement.

Dans l'éventualité d'un début de traduction à partir du deuxième codon ATG, il serait intéressant de se demander pour quelle(s) raison(s) la séquence se trouvant entre les 2 codons ATG est si fortement conservée chez 3 espèces assez éloignées phylogénétiquement.

Nous pourrions par exemple imaginer que cette séquence est une relique d'un gène plus ancien. Une telle hypothèse pourrait être vérifiée par l'analyse de la séquence de ce gène chez des organismes photosynthétiques primitifs, comme les Cyanobactéries. Or très récemment, Mulligan et al (1983) ont démontré que chez la Cyanobactérie Fremelia diplosiphon, ce gène était présent en double exemplaire et que la séquence de sa partie structurale était très semblable à celle du psb A de l'épinard (>90%). Cette observation suggère que ce gène n'a subi que très peu de modifications durant l'évolution. A l'inverse, des résultats préliminaires ont montré que le rbc L de différentes Cyanobactéries paraissait beaucoup moins homologue au rbc L des plantes supérieures (Mulligan et al, 1983).

Ces faits confirment que la structure primaire d'une protéine membranaire est beaucoup plus rigide que celle d'une protéine stromatique.

Un autre sujet de réflexion est le rôle exact du photogène dans la résistance de certaines plantes à des herbicides comme le DCMU ou les triazines. Dans cette optique, il serait particulièrement intéressant de comparer la structure primaire du photogène provenant de plantes dont le comportement face aux herbicides est différent.

Comme le psb A de Zea mays (plante qui est résistante à certains de ces herbicides) a été séquencé récemment (McIntosh, 1983), et que de plus un gène conférant la résistance au DCMU a été cloné dans un plasmide de la Cyanobactérie Anacystis nidulans (Golden et al, 1983), ce problème va sans doute être rapidement éclairci.

L'analyse de l'organisation structurale et génétique du génome chloroplastique de soja nous a amené à mieux connaître le système génétique de cet organe cellulaire.

Il est admis actuellement que les chloroplastes et les mitochondries des organismes vivants supérieurs sont les descendants éloignés de bactéries symbiotiques. Cette hypothèse est confirmée par les résultats obtenus récemment grâce aux techniques de biologie moléculaire.

Il n'en reste pas moins que durant l'évolution, le chloroplaste s'est individualisé et est devenu une entité génétique unique, ayant à la fois des caractéristiques procaryotiques et eucaryotiques.

## VII. BIBLIOGRAPHIE

- Badley RA, Atkinson D, Hauser H, Oldani D, Green JP et Stubbs JM, (1975), *Biochim.Biophys.Acta*, 412:214-228.
- Barton KA, Thompson JF, Madison JT, Rosenthal R, Jarvis NP et Beachy R, (1982), *J.Biol.Chem.* 257:6089-6095.
- Beachy R, Barton KA, Madison JT, Thompson JF et Jarvis N (1980a) In: *Genome organization and Expression in Plants*, (C.Leaver,ed.), Plenum Press, New-York, pp. 273-281
- Beachy R, Barton KA, Thompson JF et Madison JT (1980b) *Plant Physiol.* 65:990-994
- Bedbrook JR, Link G, Coen D, Bogorad L et Rich A (1978) *Proc.Natl.Acad.Sci.USA* 75:3060-3064
- Bedbrook JR, Kolodner R (1979) *Ann.Rev.Plant.Physiol* 30:593-560
- Bedbrook JR, Coen DM, Beaton AR, Bogorad L et Rich A (1979) *J.Biol.Chem.* 254:905-910
- Billeter M et Weber H (1978) *Cours SKMB, Institut für Molekularbiologie I, Universität Zürich*
- Bils RF et Howell RW (1963) *Crop Sci.* 3:304-308
- Birboim HC et Doly J (1979) *Nucl.Acids Res.* 6:1513-1523
- Bohnert HJ, Crouse EJ et Schmitt JM (1982) *Encyclopedia of Plant physiology, New Series, vol 14B, Parthier et Boulter D, eds, Springer Verlag, Berlin-Heidelberg*
- Boyer (1969) *J.Mol.Biol.* 41:459-472
- Broach JR (1982) *Cell* 28:203-204
- Chrambach A, Reisfeld RA, Wyckoff M et Zaccari J (1967) *Anal.Biochem.* 20:150
- Cerretti DP, Dean D, Davis GR, Bedwell DM et Nomura M (1983) *Nucl.Acids Res.* 11:2599-2615

- Ciferri O, Di Pasquale G et Tiboni O (1979)  
Eur.J.Biochem. 102:331-335
- Denhardt DT (1966)  
Biochem.Biophys.Res.Comm. 23:641-646
- Derbyshire E, Wright DJ et Boulter D (1976)  
Phytochemistry 15:3-24
- Draper M et Catsimpoolas N (1977)  
Phytochemistry 16:25-27
- Driesel AJ, Speirs J et Bohnert HJ (1980)  
Biochim.Biophys.Acta 610:297-310
- Edelman M, Cowan CA, Epstein MT, Schiff JA (1964)  
Proc.Natl.Acad.Sci. USA 52:1214-1219
- Edelman M et Reisfeld A (1980)  
In : Genome Organization and Expression in Plants, C.Leaver, ed,  
Plenum Press, New-York, pp 353-362
- Ellis RJ (1975)  
Phytochemistry 14:89-93
- Ellis RJ  
Ann.Rev.Plant Physiol. 32:111-137
- Fluhr R et Edelman M (1981)  
Nucl.Acids Res 9:6841-6853
- Gayler KR et Sykes GE (1981)  
Plant Physiol. 67:958-961
- Goldberg RB (1980)  
In : Sixth EMBO Annual Symposium, "Molecular biologists look at  
green plants", EMBL, Heidelberg
- Golden SS, Vann C et Sherman LA (1983)  
In : Sixth International Congress of Photosynthesis, 1:103
- Grebanier AE, Coen DM, Rich A et Bogorad L (1978)  
J.Cell Biol. 78:734-746
- Grunstein M et Hogness DS (1975)  
Proc.Natl.Acad.Sci. USA 72:3961-3965
- Hallick RB, Gray PW, Chelm BK, Rushlow KE et Orozco EM (1978)  
In : Chloroplast Development, Akoyunoglou G et al, eds, Elsevier,  
Amsterdam, pp 619-622
- Hawley DK et McClure WR (1983)  
Nucl.Acids Res. 11:2237-2255

- Herrmann RG, Whitfield PR et Bottomley W (1980)  
Gene B:179-191
- Highfield PE et Ellis RJ (1978)  
Nature 271:240-424
- Hill JE et Breidenbach RW (1974a)  
Plant Physiol. 53:742-746
- Hill JE et Breidenbach RW (1974b)  
Plant Physiol. 53:747-751
- Hoffman-Falk H, Mattoo AK, Marder JB, Edelman M et Ellis RJ (1982)  
J.Biol.Chem. 257:4583-4587
- Hollingsworth MJ et Hallick RB (1982)  
J.Biol.Chem. 257:12795-12799
- Jenni B et Stutz E (1978)  
Eur.J.Biochem. 88:127-134
- Jenni B, Fasnacht M et Stutz E (1981)  
FEBS Lett. 125:175-179
- Keller M, Rutti B et Stutz E (1982)  
FEBS Lett 149:133-137
- Koller B et Delius H (1980)  
Mol.Gen.Genet 178:261-269
- Koller B et Delius H (1982a)  
Mol.Gen.Genet 188:305-308
- Koller B et Delius H (1982b)  
Eur.Mol.Biol.Org.J. 1:995-998
- Kung SD (1977)  
Ann.Rev.Plant Physiol. 28:401-437
- Laemmli UK et Favre M (1973)  
J.Mol.Biol. 80:575-599
- Layne E (1957)  
In : Methods in Enzymology 3:447-454
- Link G (1981)  
Nucl.Acids Res 9:3681-3694

- Link G, Coen DM et Bogorad L (1978)  
Cell 15:725-731
- Lowry OH, Rosebrough, Farr AL et Randall RJ (1951)  
J.Biol.Chem. 193:265-275
- Lycett GW, Delauney AJ, Gatehouse JA, Gilroy J, Groy RRD et Boulter (1983)  
Nucl.Acids Res. 11:2367-2380
- Mandel M et Higa A (1970)  
J.Mol.Biol. 53:159-162
- Maniatis T, Fritsch EF et Sambrook J (1982)  
In : Molecular cloning : a laboratory manual. Cold Spring Harbor Laboratory, pp 366-367
- Manning JE, Wolstenholme DR, Ryan RS, Hunter JA et Richards OC (1971)  
Proc.Natl.Aca.Sci.USA 68:1169-1173
- Mattoo AK, Pick U, Hoffman-Falk M et Edelman M (1981)  
Proc.Natl.Aca.Sci.USA 78:1572-1576
- Maxam AM et Gilbert W (1980)  
In : Methods in Enzymology 65:499-560
- McDonell MW, Simon MN et Studier FW (1977)  
J.Mol.Biol. 110:119-146
- McIntosh L (1983)  
J.Cell.Biochem. s 7B,295
- Meinke DW, Chen J et Beachy RN (1981)  
Planta 153:130-139
- Millerd A (1975)  
Ann.Rev.Plant Physiol. 26:53-72
- Montandon PE et Stutz E (1983)  
Nucl.Acids Res (in press)
- Mubumbila M, Gordon KHJ, Crouse EJ, Burkard G et Weil JH (1983)  
Gene 21:257-266

- Mulligan B, Lin C, Schultes N et Bogorad L (1983)  
In : Sixth international Congress on Photosynthesis 1:101
- Nosikov VV et Sain B (1977)  
Nucl.Acids Res. 4:2263-2274
- Palmer JD (1983)  
Nature 301:92-93
- Palmer JD, Singh GP et Pillay DTN (1983)  
Mol.Gen.Genet. 190:13-19
- Palmer JD et Thompson WF (1981a)  
Proc.Natl.Aca.Sci.USA 78:5533-5537
- Palmer JD et Thompson WF (1981b)  
Gene 15:21-26
- Palmer JD et Thompson WF (1982)  
Cell 29:537-550
- Palmer JD, Edwards H, Jorgensen RA et Thompson WF (1982)  
Nucl.Acids Res. 10:6819-6832
- Passavant CW, Stiegler GL et Hallick RB (1983)  
J.Biol.Chem. 258:693-695
- Pfister K, Steinback KE, Gardner G et Arntzen CJ (1981)  
Proc.Natl.Aca.Sci.USA 78:981-985
- Renger G (1976)  
Biochim.Biophys.Acta 440:287-300
- Rigby PWJ, Dieckmann M, Rhodes C et Berg P (1977)  
J.Mol.Biol. 113:237-251
- Rochaix JD (1981)  
Experientia 37:323-332
- Rosenberg M et Court M (1979)  
Annu.Rev.Genet. 13:319-353
- Rutti B (1982)  
Thèse de doctorat, Université de Neuchâtel, pp. 26-28
- Sanger F, Coulson AR, Barrel BG, Smith AJH et Roe BA (1980)  
J.Mol.Biol. 143:161-178

- Schlunegger B, Fasnacht M, Stutz E, Koller B et Delius H (1983)  
Biochim.Biophys.Acta 739:114-121
- Schuler MA, Schmitt ES et Beachy R (1982a)  
Nucl.Acids Res 10:8225-8244
- Schuler MA, Ladin BF, Pollaco JC, Freyer G et Beachy R (1982b)  
Nucl.Acids Res 10:8245-8261
- Schwarz Z et Kössel H (1979)  
Nature 279:520-522
- Schwarz Z et Kössel H (1980)  
Nature 283:739-742
- Schwarz Z, Setsuko OJ, Steinmetz AA et Bogorad L (1981)  
Proc.Natl.Aca.Sci.USA 78:3423-3427
- Seyer P, Kowallik KV et Herrmann RG (1981)  
Current Genet. 3:189-204
- Sherer GFE, Walkinshaw MD, Arnott S et Morré DJ (1980)  
Nucl.Acids Res. 8:3895-3907
- Shine J et Dalgarno L (1974)  
Proc.Natl.Aca.Sci.USA 71:1342-1346
- Siegenthaler PA et Depéry F (1976)  
Eur.J.Biochem. 61:573-580
- Slightom SM, Sun SM et Hall TC (1983)  
Proc.Natl.Aca.Sci.USA 80:1897-1901
- Southern EM (1975)  
J.Mol.Biol. 98:503-517
- Sprinzl M et Gause DH (1982)  
Nucl.Acids Res. 10:r1-r55
- Steinback KE, McIntosh L, Bogorad L et Arntzen GJ (1981)  
Proc.Natl.Aca.Sci.USA 78:7463-7467
- Steinmetz AA, Krebbers ET, Schwarz Z, Gubbins EJ et Bogorad L (1983)  
J.Biol.Chem. 258:5503-5511
- Sugita M et Sugiura M (1983)  
Nucl.Acids Res. 11:1913-1918

- Stiegler GL, Matthews HH, Bingham SE et Hallick RB (1982)  
Nucl.Acids Res. 10:3427-3444
- Sun SM, Slightom JL et Hall TC (1981)  
Nature 289:37-41
- Sutcliffe JG (1978)  
Proc.Natl.Aca.Sci.USA 75:3737-3741
- Synenki RM, Levings CS et Shah DM (1978)  
Plant Physiol. 61:460-464
- Swamy GS et Pillay DTN (1982)  
Plant Sci.Lett. 25:73-84
- Tauber B (1891)  
Leguminosae : In : Engler A, Prantle B (eds). Die natürlichen  
Pflanzenfamilien, vol III. Engelmann, Leipzig, pp 70-385
- Tewari KK (1979)  
In : Nucleic Acids In Plants, vol I, Hall TC and Davies JW, eds,  
CRC Press, pp 41-108
- Thanh VH et Shibasaki K (1976)  
Biochim.Biophys.Acta 439:326-338
- Thanh VH et Shibasaki K (1977)  
Biochim.Biophys.Acta 490:370-384
- Thanh VH et Shibasaki K (1978)  
J.Agric.Food Chem. 26:692-695
- Ullrich A, Shine J, Chirgwin J, Pictet R, Tischer E, Rutter WJ et  
Goodman HH (1977)  
Science 196:1313-1319
- Van Ee JH, Vos YJ, Bohnert HJ et Planta RJ (1982)  
Plant Mol.Biol. 1:117-131
- Watson JC et Surzycki SJ (1982)  
Proc.Natl.Aca.Sci.USA 79:2264-2267
- Westhoff P, Nelson N, Bünemann H et Herrmann RG (1981)  
Current Genet. 4:109-120

- Wolf WJ (1970)  
J.Agr.Food Chem. 18:969-976
- Wolf WJ et Briggs (1956)  
Arch.Biochem.Biophys. 63:40-49
- Yokota T, Sugisaki H, Takanami M et Kasiro Y (1980)  
Gene 12:25-31
- Zurawski G, Perrot B, Bottomley W et Whitfeld PR (1981)  
Nucl.Acids Res. 9:3251-3270
- Zurawski G, Bohnert JH, Whitfeld PR et Bottomley (1982)  
Proc.Natl.Aca.Sci.USA 79:7699-7703

UNIVERSITE DE NEUCHÂTEL  
LABORATOIRE DE BIOCHIMIE

FACULTE DES SCIENCES  
PROFESSEUR E. STUTZ

LE SOJA (Glycine max) : ETUDE DE L'ACCUMULATION  
DES PROTEINES DE RESERVE DES GRAINES - ANALYSE  
DU GENOME CHLOROPLASTIQUE

---

Forme réduite

de la

THESE

présentée à la Faculté des Sciences  
de l'Université de Neuchâtel pour  
obtenir le grade de docteur ès sciences  
option biochimie et biologie moléculaire

par

ALBERT SPIELMANN

# IMPRIMATUR POUR LA THÈSE

*Le soja (Glycine max.): Etude de l'accumulation  
des protéines de réserve des graines - Analyse  
du génome chloroplastique*

*de Monsieur Albert Spielmann*

---

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel,  
sur le rapport des membres du jury,

*Messieurs E. Stutz, P. Schürmann, C. Schlienger  
(La Tour-de-Peilz) et D.T.N Pillay (Windsor,  
Ontario)*

autorise l'impression de la présente thèse.

Neuchâtel, le *13* janvier 1984

Le doyen:

*H. Beck*  
H. Beck

## PUBLICATIONS

- A.Spielmann, P.Schürmann and E.Stutz (1982)  
Gel electrophoretic Characterisation of protein fractions  
from soybean during seed development.  
Plant Science Letters 24 : 137-145
- A.Spielmann, W.Ortiz and E.Stutz (1983)  
The soybean chloroplast genome : Construction of a circular  
site map and location of DNA regions encoding the genes for  
rRNAs, the large subunit of ribulose-1,5-bisphosphate  
carboxylase and the 32 KD protein of the photosystem II  
reaction center.  
Mol Gen Genet 190 : 5-12
- A.Spielmann and E.Stutz (1983)  
Nucleotide sequence of soybean chloroplast DNA regions which  
contain the psb A and trn H genes and cover the ends of the  
large single copy region and one end of the inverted repeats.  
Nucl Acids Res 20 : 7157-7167

## GEL ELECTROPHORETIC CHARACTERISATION OF PROTEIN FRACTIONS FROM SOYBEAN DURING SEED DEVELOPMENT

A. SPIELMANN, P. SCHÜRMANN and E. STUTZ\*

*Laboratoire de Biochimie, Université de Neuchâtel, Chantemerle 20, 2000 Neuchâtel (Switzerland)*

(Received July 21st, 1981)

(Revision received August 14th, 1981)

(Accepted September 7th, 1981)

---

### SUMMARY

While monitoring the accumulation of soybean proteins during seed development we observed a major protein fraction of unknown function sedimenting approx. at 6S and somewhat slower than the well defined 7S fraction. We show that the 6S fraction is composed of the two major polypeptides of apparent  $M_r$  83 500 and 85 000. This protein fraction appears later in seed development than do the 7S (conglycinin) and 11S (glycinin) proteins. Furthermore we observe at early stages of seed development in the 11S fraction a  $M_r$  83 000 polypeptide which disappears at about the same time when the two major classes of subunits of glycinin appear. Precursors of glycinin subunits of  $M_r$  59 000–63 000 were reported to exist. This most likely excludes the  $M_r$  83 000 polypeptide to be a precursor of glycinin polypeptides.

---

### INTRODUCTION

Mature soybeans contain a complex mixture of storage proteins which are usually subdivided according to their sedimentation behavior into an 11S fraction (glycinin) a 7S fraction ( $\beta$ - and  $\gamma$ -conglycinin) and a 2S fraction ( $\alpha$ -conglycinin) (for reviews, see Refs 1 and 2). Each of these fractions yields a heterogeneous polypeptide pattern upon electrophoresis under denaturing conditions indicating that during seed maturation many polypeptides must be synthesized and assembled into complex proteins.

Hill and Breidenbach [3,4] have shown that the 2S fraction (2.2S in their nomenclature) appears early in seed formation while the 7S (7.5S) and the

---

\*To whom correspondence should be sent.

Abbreviations: SDS, sodium dodecyl sulphate; MS, mature seeds.

11S (11.8S) fractions accumulate later in seed maturity. From their studies it became clear that the 7S fraction contained three protein components which appeared at different times but the 11S and 2S proteins seemed to develop in a rather uniform manner, i.e., without temporal differences in the accumulation of its components.

Using a similar approach we have analysed the polypeptide composition of the three major protein fractions during seed development but using sodium dodecyl sulphate (SDS) polyacrylamide gel electrophoresis. In extension of published work [3,4] we show that in addition to the 7S also the 11S fraction undergoes temporal differences in polypeptide composition and we observe a protein fraction appearing late in seed development which sediments close to the 7S fraction; however, its polypeptide composition is definitely different from that of the 7S fraction.

## MATERIALS AND METHODS

### *Plant material*

Sterilized soybean seeds (*glycine max*, var. Maple Arrow) were planted after 6 h of soaking in the dark, into pots with vermiculite and grown under controlled conditions with a diurnal rhythm of 15 h light (25°C, 70% relative humidity) and 9 h dark (20°C, 85% relative humidity). From the second week on the plants received 3 times a week a nutrient solution according to Siegenthaler and Depéry [5]. Seeds were harvested at given times, weighed and stored at -80°C.

### *Preparation of the 31 000 × g supernatant*

The extraction procedure was essentially as published [3,6] except that the ground seeds (harvested at different days after flowering) were extracted for 30 min, 20°C in 15 vol. (w/v) of buffer A; 0.035 M potassium phosphate (pH 7.6), 0.4 M NaCl, 0.014 M β-mercaptoethanol, 0.02% NaN<sub>3</sub>, μm = 0.5). The extract was filtered through 4 layers of cheesecloth, centrifuged for 15 min at 31 000 × g (Sorvall RC-5B, SS-34 rotor). Proteins in the pellet, the 31 000 × g supernatant and in the filtered extract were determined according to Lowry et al. [7].

### *Sucrose gradients*

The supernatant proteins were analysed in exponential sucrose density gradients, 10–30% (w/v), 38 ml in buffer A. Centrifugation was for 40 h, 20°C, Beckman L5-65, SW-rotor. Fractions of 1–1.3 ml were taken and proteins photometrically monitored at 280 and 260 nm [8]. Proteins in the peak fractions were determined according to Lowry et al. [7].

### *Gel electrophoresis [9]*

All samples were dialyzed overnight at 4°C against buffer B (0.035 M potassium phosphate (pH 7.6), 0.014 M β-mercaptoethanol, 0.02% NaN<sub>3</sub>,

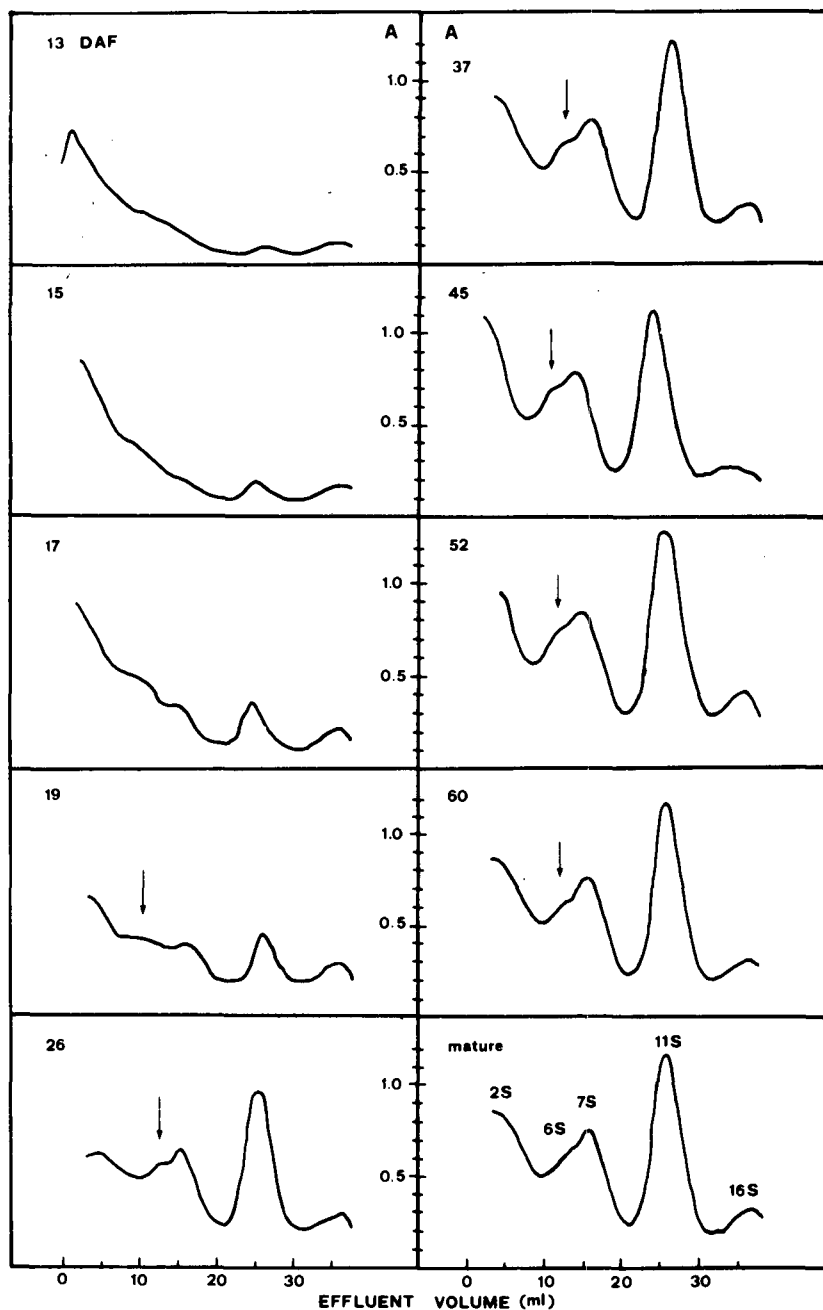


Fig. 1. Sucrose gradient profiles of the 31 000 supernatants. Harvest days after flowering (DAF) are given for each sample starting with the 13th day (top left) to mature seeds (bottom right). Absorbance (A) is expressed as  $1.55 A_{280} - 0.76 A_{260}$  according to Layne [8]. The arrow points towards the 6S peak. This is clearly seen as shoulder of the 7S peak in the 26th day profile and thereafter. The nomenclature of the peaks 2S, 7S and 11S is as published [1].

5% sucrose). Gel concentrations: stacking gel, 6% acrylamide, in 0.125 M Tris-HCl (pH 6.8), 0.1% SDS, 0.007  $\beta$ -mercaptoethanol. Separating gel, 12% acrylamide in 0.75 M Tris-HCl (pH 8.8), 0.1% SDS, 0.007 M  $\beta$ -mercaptoethanol. Electrophoresis buffer, 0.02 M Tris-glycine (pH 8.6), 0.1% SDS, 0.014 M  $\beta$ -mercaptoethanol. Coomassie blue staining was according to Charmbach et al. [10].

## RESULTS AND DISCUSSION

We have analyzed, in analogy to the work of Hill and Breidenbach [3], the 31 000  $\times g$  supernatant from developing and mature seeds in sucrose density gradients (Fig. 1), using however exponential instead of linear sucrose

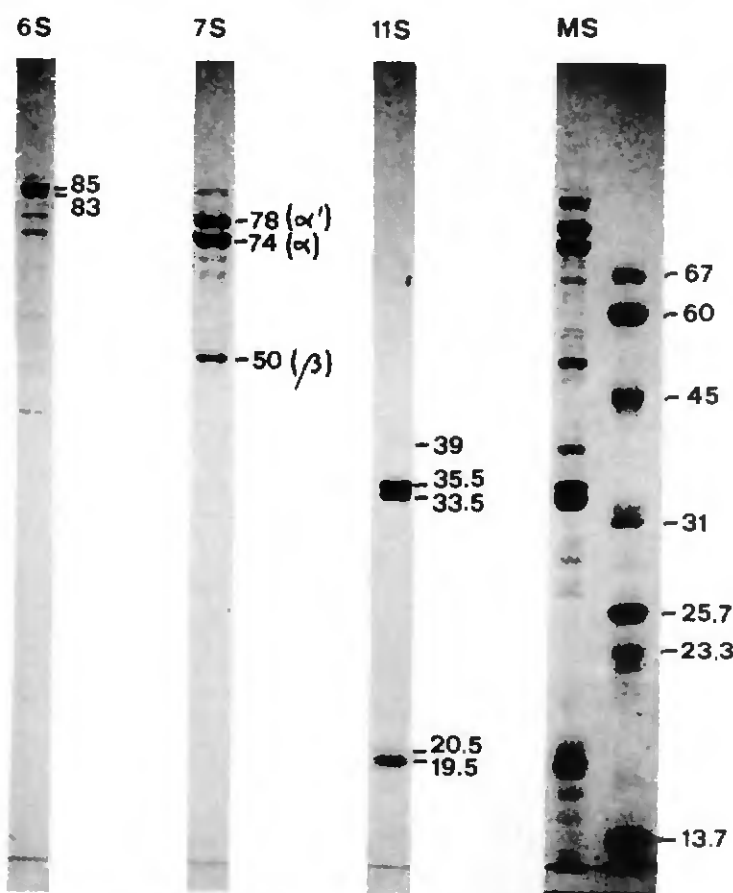


Fig. 2. Photograph of stained gels. About 3–5  $\mu$ g of proteins of respectively, the 6S, 7S and 11S peak fraction were layered on gels, electrophoresed and stained as specified in Materials and Methods. MS is the profile of the supernatant from mature seeds. For calibration we used bovine serum albumin (67 000), catalase (60 000), ovalbumin (47 000), DNase I (31 000), chymotrypsinogen (25 700), trypsin (23 000) and RNase (13 700). Apparent  $M_r$  are given in kilodaltons.

gradients (10–30%, w/v). The sedimentation profile from the mature seed supernatant reveals the presence of the three major protein fractions, i.e., 2S, 7S and 11S, as expected and reported [3]. As already known the 2S protein appears early in development while the 7S and 11S proteins accumulate later in seed development (compare the profiles of the 19th with the 26th day). We noticed, however, that the 7S peak was skewed towards the slower sedimenting side (arrow) and the shoulder (6S) persisted throughout seed development.

### DAF

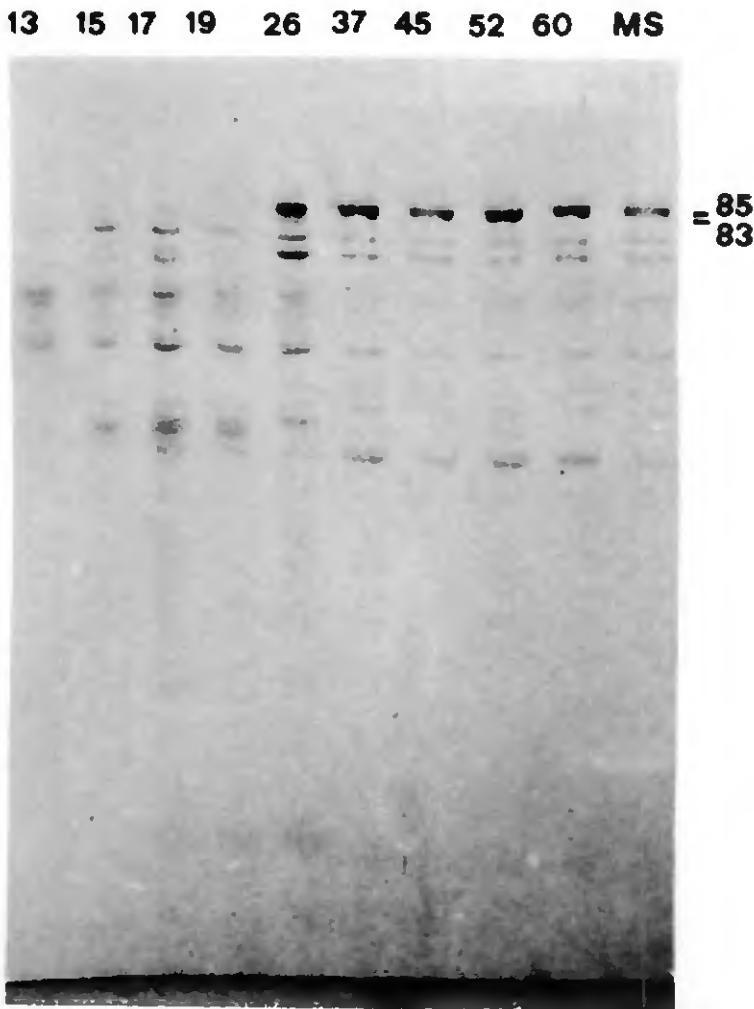


Fig. 3. Photographs of gels from the 6S peak fractions. Supernatants from seeds harvested 13–60 days after flowering (DAF) were electrophoresed as specified. Otherwise as given in legend to Fig. 2.

We analyzed the supernatants from mature and developing seeds under dissociating conditions by gel electrophoresis [9]. In Fig. 2 we show the stained protein pattern from mature seeds (MS) and from each of the major fractions, however, we separately analyzed the 6S and 7S fractions (the 2S fraction is not included in further experiments). The 7S and 11S protein patterns are in essence as reported [3,11–13]. We see, e.g., in the 7S fraction the three distinct polypeptides called  $\alpha'$ ,  $\alpha$  and  $\beta$  [12] and in the 11S fraction we also see several polypeptides (glycinin) in the range  $M_r$  33 000–35 000 and  $M_r$  19 000. However, the 6S fraction which is composed of two major polypeptides of  $M_r$  83 000 and 85 000 has not been described in soybean to

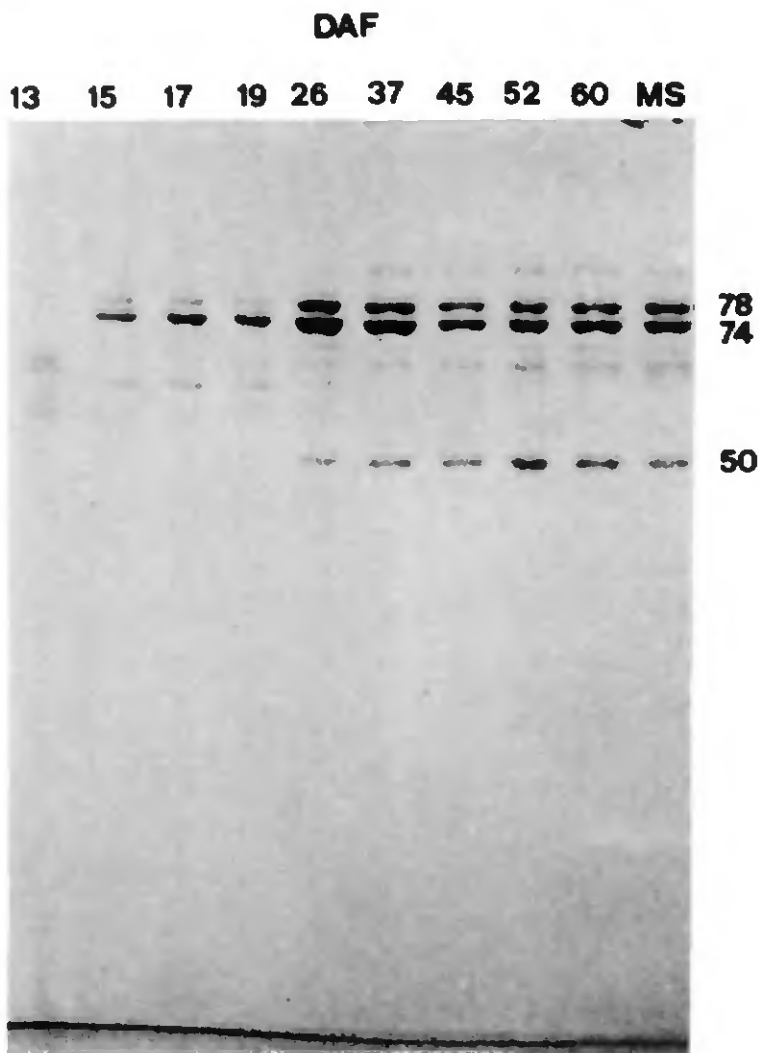


Fig. 4. Photographs of gels from the 7S peak fractions (see legend to Fig. 3).

the best of our knowledge, although Thanh and Shibasaki [12] mention that the 7S may be composed of at least five electrophoretic components.

In order to see the temporal appearance of the  $M_r$  83 000 and 85 000 polypeptides during seed development we analyzed the polypeptide profiles of the 6S fraction at various stages of seed development (Fig. 3). For comparative reasons we also show the profile from the 7S fraction (Fig. 4) and 11S fraction (Fig. 5). Under our experimental conditions the  $M_r$  83 000 and 85 000 polypeptides accumulate relatively late being first seen 26 days after flowering, while e.g., the 7S  $\alpha$ -polypeptide is already present in the 15th day

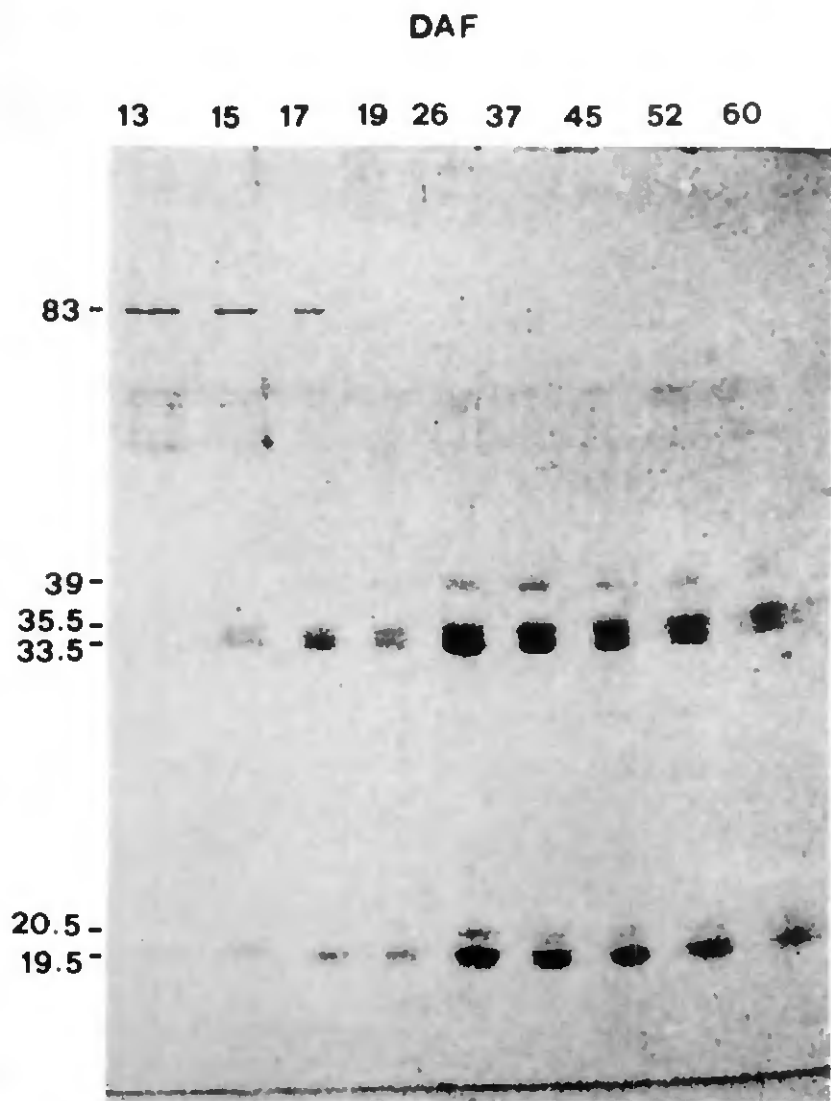


Fig. 5. Photographs of gels from the 11S peak fractions (see legend to Fig. 3).

supernatant similar to the proteins of the 11S fraction. The 6S proteins are present at all later stages including the dry seed stage. We do not know the function of these two proteins, but their quantitative contribution as possible seed (storage) protein cannot be neglected.

Our data are in line with those reported by Hill and Breidenbach [4] with respect to the 7S fraction, i.e., the 7S fraction undergoes temporal differences in protein composition the  $\alpha$ -subunit being first to appear followed by the  $\alpha'$ - and  $\beta$ -subunits (Fig. 4). The same was observed and reported most recently by Gayler and Sykes [15]. However, the 11S fraction, contrary to Hill and Breidenbach's observation also undergoes temporal differences in protein composition as shown in Fig. 5. At early stages there is a strong polypeptide band of  $M_r$  83 000 which disappears when the two major protein components of glycinin appear which migrate in the range  $M_r$  19 500–20 500 (basic polypeptides) and  $M_r$  33 500–39 000 (acidic polypeptides), respectively. We may even argue that in each of these two subfractions there are temporal differences in polypeptide composition, e.g., the  $M_r$  39 000 polypeptide appears later than the others do and the same holds for the  $M_r$  20 500 polypeptide.

It seems to be established that glycinin subunits are synthesized as long precursors [16], the precursor sizes being in the range  $M_r$  59 000–63 000. This most likely excludes the 83 000 polypeptide to be a precursor of glycinin. Nevertheless, its existence in the 11S fraction has to be taken into account in studies on the control of transcription and translation during seed development [17,18].

#### ACKNOWLEDGEMENTS

This research was supported in part by Société d'Assistance Technique pour Produits Nestlé S.A. and Fonds National Suisse de la Recherche Scientifique (3.452.79). This work is part of a doctoral program carried out by A.S.

#### REFERENCES

- 1 A. Millerd, *Annu. Rev. Plant Physiol.*, 26 (1975) 53.
- 2 E. Derbyshire, D.J. Wright and D. Boulter, *Phytochemistry*, 15 (1976) 3.
- 3 J.E. Hill and R.W. Breidenbach, *Plant Physiol.*, 53 (1974) 742.
- 4 J.E. Hill and R.W. Breidenbach, *Plant Physiol.*, 53 (1974) 747.
- 5 P.A. Siegenthaler and F. Depéry, *Eur. J. Biochem.*, 61 (1976) 573.
- 6 W.J. Wolf and D.R. Briggs, *Arch. Biochem. Biophys.*, 63 (1956) 40.
- 7 O.M. Lowry, N.J. Rosebrough, A.L. Farr and R.J. Randall, *J. Biol. Chem.*, 193 (1951) 265.
- 8 E. Layne, in: S.R. Colowick and N.O. Kaplan (Eds.), *Methods in Enzymology*, Vol. 3, Academic Press, New York, 1957, p. 447.
- 9 U.K. Laemmli and N. Favre, *J. Mol. Biol.*, 80 (1973) 575.
- 10 A. Chrambach, R.A. Reisfeld, M. Wyckoff and J. Zaccari, *Anal. Biochem.*, 20 (1967) 150.
- 11 V.M. Thanh and K. Shibasaki, *Biochim. Biophys. Acta*, 439 (1976) 326.

- 12 V.M. Thanh and K. Shibasaki, *Biochim. Biophys. Acta*, 490 (1977) 370.
- 13 M. Draper and N. Catsimpoalas, *Phytochemistry*, 16 (1977) 25.
- 14 R.A. Badley, D. Atkinson, M. Hauser, D. Oldani, J.P. Green and J.M. Stubbs, *Biochim. Biophys. Acta*, 412 (1975) 214.
- 15 K.R. Gayler and G.F. Sykes, *Plant Physiol.*, 67 (1981) 958.
- 16 K.A. Barton, J.F. Thompson, J.T. Madison, N.P. Jarvis and R.N. Beachy, *Plant Physiol.*, S67 (1981) 46.
- 17 R.N. Beachy, K.A. Barton, J.T. Madison, J.F. Thompson and N. Jarvis, in: C. Leaver (Ed.), *Genome Organization and Expression in Plants*, Plenum Press, New York, 1980, p. 273.
- 18 R. Beachy, K.A. Barton, J.F. Thompson and J.T. Madison, *Plant Physiol.*, 65 (1980) 990.

# The Soybean Chloroplast Genome: Construction of a Circular Restriction Site 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

Albert Spielmann, William Ortiz\*, and Erhard Stutz

Laboratoire de Biochimie, Université de Neuchâtel, Ch. de Chantemerle 18, CH-2000 Neuchâtel, Switzerland

**Summary.** Chloroplast DNA was isolated in preparative amounts from purified chloroplasts from soybean (*Glycine max.*). This DNA upon digestion with the restriction enzymes *Sall*, *Pst*I, and *Sma*I yields 5, 9, and 12 fragments, respectively. The total genome size is between 150–152 kilobase pairs. A complete circular *Sall*-*Pst*I-*Sma*I restriction site map has been constructed. The circular genome contains two inverted repeats of about 21–23 kbp separated by a single copy region of about 24 kbp. The rDNA region is part of the inverted repeat. The genes for the large subunit of the ribulose 1,5-bisphosphate carboxylase is localized about five kbp away from the gene coding for the 32 KD protein of the photosystem II reaction center. The close proximity of these two genes seems to be characteristic for chloroplast genomes of the legume family.

## Introduction

The chloroplast genomes of several members of the legume family have been physically analysed on a molecular level, namely broad bean (*Vicia faba*) (Koller and Delius 1980), pea (*Pisum sativum*) (Kolodner and Tewari 1979; Chu et al. 1981) and mung bean (*Vigna radiata*) (Palmer and Thompson 1981). The chloroplast genome of soybean (*Glycine max.*) is not yet on that list although this plant is subject to extensive genetic research due to its economic importance as a major crop plant. This situation may be due partly to difficulties in obtaining a good yield of chloroplasts suitable for DNA extraction. Synenko et al. (1978) mention in a paper on soybean mitochondrial DNA that they also had isolated soybean chloroplast DNA. However, the yield seemed to be very low and the only physical parameter reported is the buoyant density of 1.697 g/cm<sup>3</sup> (CsCl), which corresponds to about 35% G + C.

The circular chloroplast genomes of broad bean and pea are composed of about 120 kbp while the size of the mung bean chloroplast genome is about 150 kbp. This large

size difference is mainly due to loss of a repeated segment only seen so far in chloroplast genomes of legume plants of the tribe *vicieae* (Palmer and Thompson 1982). In all other cases of higher plants studied including corn (Bedbrook and Bogorad 1976), spinach (Kolodner and Tewari 1979), tobacco (Jurgenson and Bourque 1980; Fluhr and Edelman 1981a), *Spirodela* (Van Ee et al. 1980), petunia (Bovenberg et al. 1981), mustard (Link et al. 1981), *Oenothera* (Gordon et al. 1981), wheat (Bowman et al. 1981), daffodil and nasturtium (Thompson et al. 1981), mung bean (Palmer and Thompson 1981), cucumber (Palmer 1982) the general anatomy is rather similar and two inverted repeats in the range of 20–25 kbp separate a small and a large single copy region.

Soybean is related to mung bean and belongs to the sub-tribe *phaseoleae-glycineae* (Taubert 1891). Therefore, we expect that the soybean chloroplast genome in its general anatomy resembles the chloroplast genome of mung bean rather than that of pea or broad bean. We have constructed the complete *Sall*-*Pst*I-*Sma*I restriction site map and on the circular genome of about 150 kbp located two inverted repeats containing an rDNA region. We also placed the genes for the large subunit protein of the ribulose 1,5-bisphosphate carboxylase and for a 32 KD thylakoid protein of the photosystem II reaction center on the circular map. With this map available it should become possible to include the chloroplast genome in general and applied genetic studies of the soybean plant and extend comparative studies on base sequence divergence within the legume family as already successfully done by Palmer and Thompson (1982).

## Materials and Methods

Soybean plants (*Glycine max.*, var. Maple Arrow) were grown in vermiculite under controlled conditions with a diurnal rhythm of 15 h light (25° C, 70% relative humidity) and 9 h dark (20° C, 85% relative humidity). From the second week on three times a week the plants received a nutrient solution according to Siegenthaler and Depéry (1976). Leaves 2–3 weeks old were collected after 36 h to 48 h of darkness to eliminate starch.

**Preparation and Analysis of Chloroplast DNA.** CpDNA was isolated from purified chloroplasts essentially following the procedure of Kolodner and Tewari (1975).

The cpDNA was digested with the restriction endonu-

\* Present Address: University of California, Department of Plant and Soil Biology, 108, Hilgard Hall, Berkeley, CA 94720, USA

Offprint requests to: A. Spielmann

**Abbreviations:** Kbp: kilobase pairs; cpDNA: chloroplast DNA; LSU: large subunit; P32: 32 kilodalton thylakoid protein of the reaction center of photosystem II; EtBr: ethidium bromide; SSC: standard saline buffer (0.15 M NaCl, 0.015 M Na-citrate)

**Table 1.** Nomenclature and size in kbp of restriction fragments from soybean chloroplast DNA

| Sal                 | Pst                 | Sma                 | SP                      | SSm                                                     | PSm                                |
|---------------------|---------------------|---------------------|-------------------------|---------------------------------------------------------|------------------------------------|
| A 58.8 <sup>a</sup> | A 41.1 <sup>a</sup> | A 55.4 <sup>a</sup> | A 27.0                  | A 27.0                                                  | A 30.0 ( <i>Pst</i> -C)            |
| B 44.9 <sup>a</sup> | B 32.8 <sup>a</sup> | B 25.1              | B 26.0                  | B 16.8 (3 ×) ( <i>Sma</i> -C)<br>(2 ×) ( <i>Sal</i> -D) | B 25.1 ( <i>Sma</i> -B)            |
| C 21.3              | C 30.0 <sup>a</sup> | C 16.8 (2 ×)        | C 18.1                  | C 12.3                                                  | C 13.7 ( <i>Pst</i> -E)            |
| D 16.8              | D 16.8              | D 15.7              | D 16.8 ( <i>Pst</i> -D) | D 10.6                                                  | D 12.6 (3 ×)                       |
| E 8.9               | E 13.7              | E 5.3               | E 14.1                  | E 9.7                                                   | E 6.2 ( <i>Pst</i> -G)             |
|                     | F 7.2               | F 4.1               | F 8.9 ( <i>Sal</i> -E)  | F 8.9 ( <i>Sal</i> -E)                                  | F 5.3 ( <i>Sma</i> -E)             |
|                     | G 6.2               | G 3.3 (2 ×)         | G 7.2 ( <i>Pst</i> -F)  | G 6.5                                                   | G 4.2 (2 ×)                        |
|                     | H 1.9               | H 2.4               | H 7.1                   | H 5.6                                                   | H 4.1                              |
|                     | I 1.0               | I 0.9 (2 ×)         | I 6.6                   | I 5.3 ( <i>Sma</i> -E)                                  | I 3.6                              |
|                     |                     |                     | J 6.2 ( <i>Pst</i> -G)  | J 4.1 ( <i>Sma</i> -F)                                  | J 3.3 (2 ×) ( <i>Sma</i> -G)       |
|                     |                     |                     | K 6.0                   | K 3.3 (2 ×) ( <i>Sma</i> -G)                            | K 3.1                              |
|                     |                     |                     | L 4.0                   | L 2.4 ( <i>Sma</i> -H)                                  | L 2.4 ( <i>Sma</i> -H)             |
|                     |                     |                     | M 1.9 ( <i>Pst</i> -H)  | M 0.9 (2 ×) ( <i>Sma</i> -I)                            | M 1.0 ( <i>Pst</i> -I)             |
|                     |                     |                     | N 1.0 ( <i>Pst</i> -I)  |                                                         | N 0.9 (3 ×) ( <i>Sma</i> -I) (2 ×) |
|                     |                     |                     |                         |                                                         | O 0.1 <sup>b</sup>                 |

*Sal*, *Pst* and *Sma* indicate DNA fragments obtained by digestion with the restriction endonucleases *Sal*I, *Pst*I, *Sma*I, respectively; SP, SSm, and PSm indicate DNA fragments obtained by digestion with *Sal*I + *Pst*I, *Sal*I + *Sma*I and *Pst*I + *Sma*I, respectively. Numbers in parenthesis indicate bands containing two or three fragments of equal gel electrophoretic mobility under the experimental conditions used. In the three columns with fragments from double digestions, fragments not cut by the second enzyme are given in parenthesis.

<sup>a</sup> The size of these large fragments was calculated from the sizes of the corresponding subfragments, rather than from the gel electrophoretic mobilities

<sup>b</sup> This fragment was seen in a *Sma* digest of the cloned fragment *Pst*-H (not shown)

cleases *Sal*I, *Pst*I and *Sma*I as directed by the suppliers (Boehringer, Mannheim, FRG). Vertical agarose slab gels (20 × 15 × 0.3; 0.5–1.3% agarose depending on fragment size) were prepared in 40 mM Tris-acetate, 20 mM Na-acetate, 1 mM EDTA, pH 7.8. Gels were run for 14–18 h at 0.6 to 1 mA/cm, colored with EtBr and photographed under UV-light.

The DNA from the phages T7 and lambda, and their *Hpa*II and *Hind*III fragments, respectively, were used as molecular weight standards.

cpDNA was transferred from agarose gels onto nitrocellulose filters (Millipore, HAWP 304FO, 0.45 µm) according to Southern (1975). DNA fragments were labelled by nick-translation according to Rigby et al. (1977). Hybridization was according to Denhardt (1966) and X-ray films were exposed for 2–30 days at room temperature.

**DNA Probes Carrying Defined Genes.** The 23S rDNA was the *Euglena gracilis* chloroplast *Eco*RI-L fragment which was cloned into pBR322 in our laboratory (B. Schlunegger). The *Eco*RI-L fragment has been mapped and characterized as published (Jenni and Stutz 1978; Orozco et al. 1980).

The 16S rDNA was the *Euglena gracilis* chloroplast *Bgl*II-*Eco*RI fragment EG 19 (Jenni and Stutz 1978), obtained from *Eco*RI-P which was cloned into pBR322 in our laboratory (B. Schlunegger).

The P32 gene probe was a kind gift of Dr. H. Bohnert (EMBL, Heidelberg) who provided us with the pSoc B511 which carries an insert (5.4 kbp) encoding the P32 gene in the *Bam*HI site of pBR322 (for mapping of this gene on spinach chloroplast DNA see Driesel et al. 1980).

The LSU gene probe was a kind gift of Prof. R.G. Herrmann, University of Düsseldorf. We received the fragment Kpn-9 (1.6 kbp) which essentially carries the LSU gene [for mapping of this gene on spinach chloroplast DNA see Whitfeld and Bottomley (1980)].

## Results

### 1. Number and Size of DNA Fragments after Digestion with the Endonucleases *Sal*I, *Pst*I, and *Sma*I

Total soybean cpDNA was digested with the restriction endonucleases *Sal*I, *Pst*I, and *Sma*I and combinations of the three enzymes. The gel electrophoretic analysis yielded the results given in Table 1. The total genome size is in the range of 150–152 kbp, which is within the size range of those legume chloroplast genomes carrying two inverted repeats (Palmer and Thompson 1981). In the three columns with the fragments from double digestions we indicate equivalences with fragments from the first digestion.

### 2. Evidence for and Physical Mapping of Two Inverted Repeats Carrying rDNA Sequences

We constructed the restriction site map essentially by contact hybridization using filter blots of different fragment patterns and radioactively-labeled exogenous DNA probes of known genetic functions.

As a first probe we used the cloned *Eco*RI-L fragment from the *Euglena gracilis* chloroplast genome which carries about two thirds of the 23S rRNA gene (3' part) and the adjacent total 5S rRNA gene [Orozco et al. 1980]. The results of the hybridization of the probe with filter blots of six fragment patterns are shown in Fig. 1, I (a–f). The interpretation of the data is graphically presented in Fig. 1, III for each step. The results can be summarized as follows (numbers in parenthesis following a fragment name are kbp):

1) The 23S rRNA probe hybridizes with *Sal*-A (58.8) and *Sal*-B (44.9) as shown in Fig. 1, I (panel a) and with a single *Pst* fragment, *Pst*-A (41.1), (panel b). Therefore *Pst*-A must share parts with *Sal*-A and *Sal*-B.

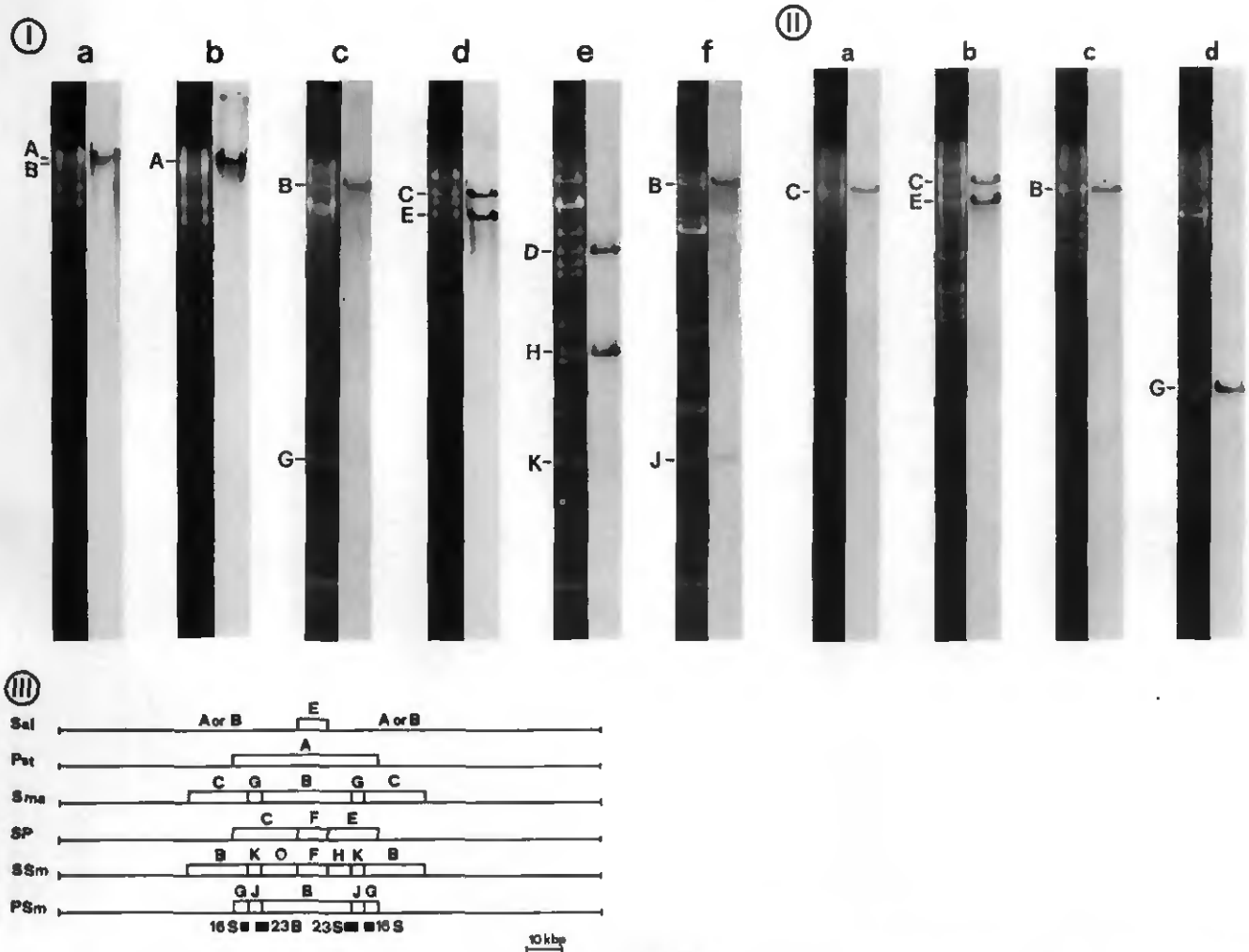


Fig. 1. I. Hybridization of a 23S plus 5S rDNA probe to restriction fragments from total chloroplast DNA. For each cpDNA digest the fluorograph (left side) and the autoradiograph (right side) are aligned. *SalI* (a), *PstI* (b), *SmaI* (c), *SalI* plus *PstI* (d), *SalI* plus *SmaI* (e), *PstI* plus *SmaI* (f). The hybridizing bands are marked by the corresponding letter. II. Hybridization of a 16S rDNA probe to restriction fragments from total chloroplast DNA. *SmaI* (a), *SalI* plus *PstI* (b), *SalI* plus *SmaI* (c), *PstI* plus *SmaI* (d). III. Schematic presentation of the results on a linearized map; hybridizing fragments are shown as boxes. The relative position of a fragment can be tentative (e.g. top line) or definite as discussed in the text. Hybridizing fragments of identical mobility are identically labeled, e.g. *SmaI* G. In the text we used the labels  $G_1$  and  $G_2$  indicating that the fragments belong to repeated regions arbitrarily called 1 and 2. The 23S and 16S rRNA (dark boxes) are approximately positioned and spaced assuming the size of the genes to be about 2800 and 1500 bases, respectively

2) The probe hybridizes with the SP fragments C (18.1) and E (14.1) (panel d). The combined length of these two fragments is 32.2 kbp, which is about 9 kbp shorter than the length of *PstI*-A. There is only one *SalI* fragment that can fill the gap having the matching size and is not cut by *PstI*, namely, *SalI*-E (8.9) (see Table 1). This fragment is therefore placed between *SalI*-A and *SalI*-B, as shown in Fig. 1, III.

3) The probe hybridizes with the fragment *SmaI*-B (25.1) (panel c). This fragment is not cut by *PstI*, but is cut by *SalI* (see Table 1) and yields two hybridizing fragments SSm-D (10.6) and SSm-H (5.6) as seen in panel E. The *SmaI*-B fragment is therefore composed of the center piece *SalI*-E, and the two flanking fragments SSm-D and H, as shown in Fig. 1, III.

4) In panel c a band estimated to be composed of two fragments ( $G_1$  and  $G_2$ ) hybridizes with the probe. We have placed them tentatively contiguous to *SmaI*-B. If this position is correct, fragments of equal size and stoichiometry

have to appear and hybridize in the SSm pattern (panel e) and in the PSm pattern (panel f). Indeed the sizes of the smallest hybridizing band in the *SmaI* pattern (panel c) and in the SSm pattern (panel e) and PSm pattern (f) are identical, i.e. PSm-B and the double fragments *SmaI*- $G_1$  and  $G_2$  are equivalent to SSm- $K_1$  and  $K_2$  and to PSm- $J_1$  and  $J_2$  (3.3).

5) The length of the hybridizing probe, which contains about two thirds of the 23S rRNA gene (3' part) and the entire 5S gene, is at most 3 kbp. According to the results shown we have to place it in tandem with the two fragments *SmaI*-B and G, as shown in Fig. 1, III. It means that two rDNA regions exist, one on *SalI*-A and the other on *SalI*-B. The non hybridizing *SalI*-E fragment lies in between.

6) The 5S gene is probably in the proximity of the 3' end of the 23S or 4.5S rRNA gene but we have no direct evidence since we did not use a separate 4.5S nor 5S rRNA probe.

As a second probe we used an *EcoRI*-*BglII* fragment

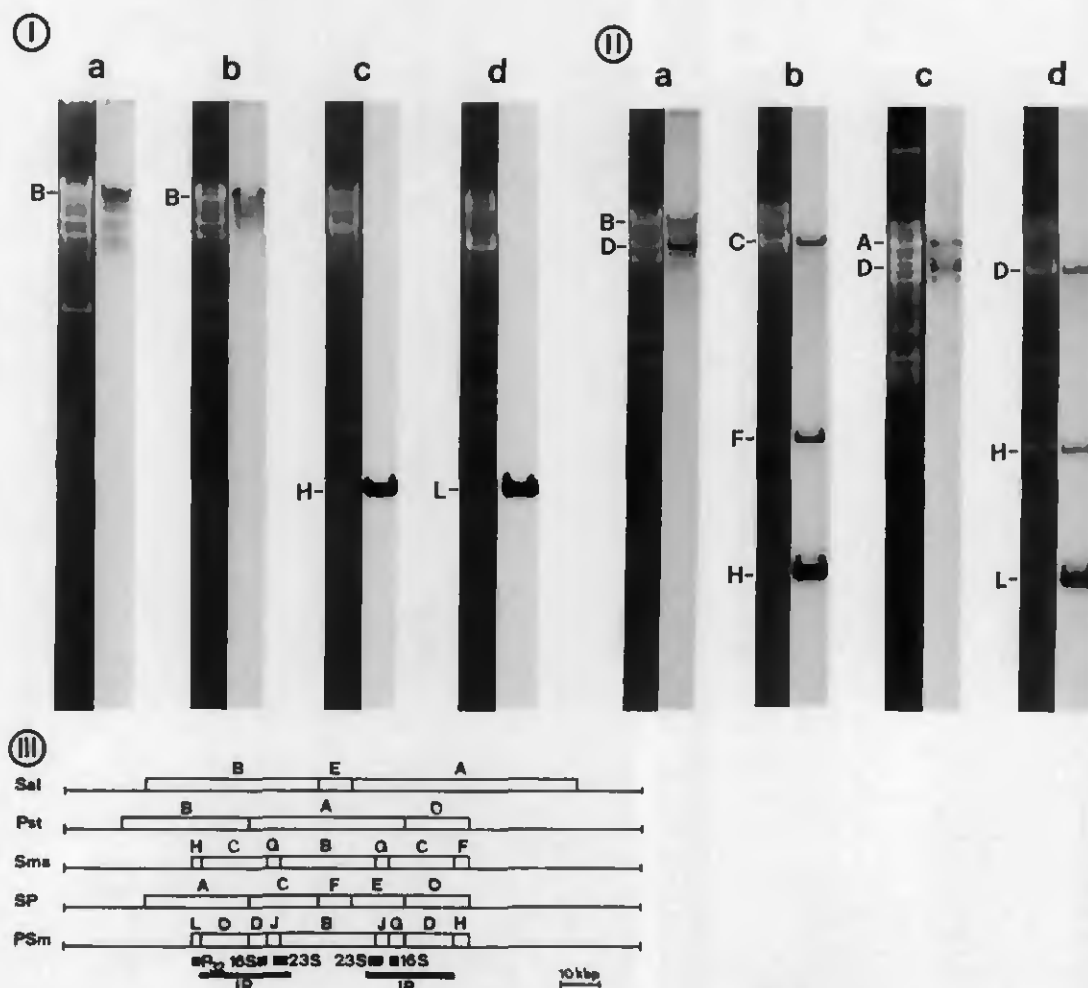


Fig. 2. Hybridization of DNA probe (a) carrying part of the P32 gene and DNA probe (b) carrying the P32 gene and a terminal part of the repeated segment to restriction fragments from total chloroplast DNA. I: For the concept see Fig. 1, I, *Sal*I (a), *Pst*I (b), *Sma*I (c), and *Pst*I plus *Sma*I (d); II: *Pst*I (a), *Sma*I (b), *Sal*I plus *Pst*I (c) and *Pst*I plus *Sma*I (d). The correlation of an autoradiographic signal with a particular fragment is difficult to achieve, e.g., in panel Ia and IIa. It is only possible when consulting the original larger staining patterns and autoradiographs; III: For the concept see Fig. 1, III. The size of the inverted repeats is indicated by a heavy line (IR)

from *Euglena gracilis* chloroplast DNA which carries about 80% of the 16S rRNA gene (3' part) [Jenni and Stutz 1978]. This probe was hybridized with several different filter blots as depicted in Fig. 1, II.

1) In Fig. 1, II (panel a) we see that a double band (*Sma*-C<sub>1</sub> and C<sub>2</sub>) interacts with the probe but neither *Sma*-B nor *Sma*-G<sub>1</sub> and G<sub>2</sub> hybridize with 16S rRNA. We tentatively place the hybridizing *Sma*-C<sub>1</sub> and C<sub>2</sub> fragments next to the G<sub>1</sub> and G<sub>2</sub> fragments, as shown in Fig. 1, III.

2) The SP fragments C and E hybridize with the 16S rRNA probe (panel b) as they did with the 23S rRNA probe, meaning that the two rRNA genes are relatively close together.

3) Only one band of the SSm pattern hybridizes with this probe (panel c). This band contains three fragments of equal gel electrophoretic mobility (see Table 1), two of which (SSm-B<sub>1</sub> and B<sub>2</sub>) must be equivalent to the fragments *Sma*-C<sub>1</sub> and C<sub>2</sub>. They are integral parts of the fragments *Sal*-A and *Sal*-B.

4) The SSm pattern (profile d) shows a single band interacting with the probe (PSm-G<sub>1</sub> and G<sub>2</sub>). This result is in line with the restriction site map established so far and

can only be obtained if two inverted repeats exist consisting at least of the fragments G<sub>1</sub> plus C<sub>1</sub> and G<sub>2</sub> plus C<sub>2</sub>.

5) The structural part of the 16S gene is expected to be about 1500 bp as in other cases [Schwarz and Kössel 1980] and can be placed as shown in Fig. 1, III.

### 3. Location of the P32-Gene and Further Evidence for Two Inverted Repeats

All higher plant chloroplast genomes studied so far, in the proximity of one of the inverted repeats contain a gene coding for a 32 KD thylakoid protein of the photosystem II reaction center (P32-gene). As a working hypothesis we may assume that this is also the case on the soybean chloroplast genome. We used as starting material for the DNA probes the clone pSoc B511, which has a 5.4 kbp insert carrying the P32 gene from spinach (see Materials and Methods). From this we constructed the two following hybridization probes: probe (a) is a 330 bp DNA segment cut by *Hpa*II from within the P32 gene and probe (b) is a 1.6 kbp SSm fragment containing the entire P32 gene and about 200 bp of the adjacent end of the large inverted

repeat of the spinach chloroplast genome. These two probes should allow location of the P32 gene and at the same time give further evidence for the existence of the inverted repeats assuming that the corresponding sequences are sufficiently preserved. In Figs. 2, I and 2, II the hybridization data with probe (a), and probe (b), respectively, are displayed and an interpretation of both series of experiments is combined in Fig. 2, III. The results can be summarized as follows:

1) Probe (a) hybridizes with *Sal*-B [Fig. 2, I (panel a)], *Pst*-B (panel b) and *Sma*-H (panel c).

2) Probe (b) hybridizes to *Pst*-B and *Pst*-D (Fig. 2, II) (panel a) and to *Sma*-H, *Sma*-F and *Sma*-C<sub>1</sub> and C<sub>2</sub> (panel b). A first conclusion is that the fragments *Sal*-B, *Pst*-B and *Sma*-H have sequences in common and can therefore be placed as shown in Fig. 2, III. Furthermore, we see that probe (b) hybridizes to additional *Pst* and *Sma* fragments, due to the additional sequences from the inverted repeat. As a consequence we can place *Pst*-D next to *Pst*-A and *Sma*-F next to *Sma*-C<sub>2</sub>. *Sma*-H, which carries the P32 gene, must be next to *Sma*-C<sub>1</sub> as shown in Fig. 2, III.

3) The PSm patterns in Fig. 2, I and Fig. 2, II fully support this mapping result, i.e., probe (a) hybridizes to a single fragment (PSm-L) while the probe (b) hybridizes in addition to PSm-H and a strong band containing the fragments PSm-D<sub>1</sub> and D<sub>2</sub> which can be placed as shown in Fig. 2, III.

4) If the restriction map is correct then the probe (b) should hybridize to two bands in the SP pattern, namely SP-A which carries the P32 gene and parts of the inverted repeat and SP-D which is equivalent in length to *Pst*-D and carries part of the inverted repeat. This is seen in Fig. 2, II (panel c) and is graphically presented in Fig. 2, III.

5) We can now estimate the length of the inverted repeat region which is composed of *Sma*-G<sub>1</sub> and C<sub>1</sub> (*Sma*-G<sub>2</sub> and C<sub>2</sub>) plus a small part of *Sma*-F and consequently a small part of *Sma*-H. This means that *Sma*-H contains the P32 gene and the beginning of the inverted repeat. The total length of the inverted repeat is therefore about 21–23 kbp and the single copy region in between is somewhat smaller than *Sma*-B<sub>1</sub> and therefore certainly not larger than 25 kbp.

#### 4. Location of the LSU-Gene Close to the P32-Gene

As a fourth probe we used a cloned fragment (Kpn-9, Westhoff et al. 1981) from the spinach chloroplast genome which carries the gene for the large subunit of the 1,5 bisphosphate carboxylase-oxygenase. We hybridized it with filter blots of the restriction patterns shown in Fig. 3, I. The probe hybridized with *Sal*-B (profile a), *Pst*-B (profile b) and the *Sma* fragments D and E, as shown in pattern c. Accordingly, the LSU gene is mainly located on *Sma*-E (strong radioactive signal) and probably overlaps the adjacent *Sma*-D fragment (weak signal). We cannot exclude, however, that the beta-subunit of the coupling factor CF<sub>1</sub> which is next to the LSU gene on Kpn-9 [Westhoff et al. 1981] overlaps *Sma*-D. This uncertainty does not affect the restriction site mapping. Given these hybridization results and the restriction site map so far established, we can place the linked fragment *Sma* (D+E) contiguous to *Sma*-H. Proper orientation of the *Sma* (D+E) fragment becomes possible as displayed in Fig. 3, II when considering that *Sma*-D but not *Sma*-E is cut by *Sal*I as well as by *Pst*I (see Table 1). This places the LSU gene in close vicinity to the P32 gene,

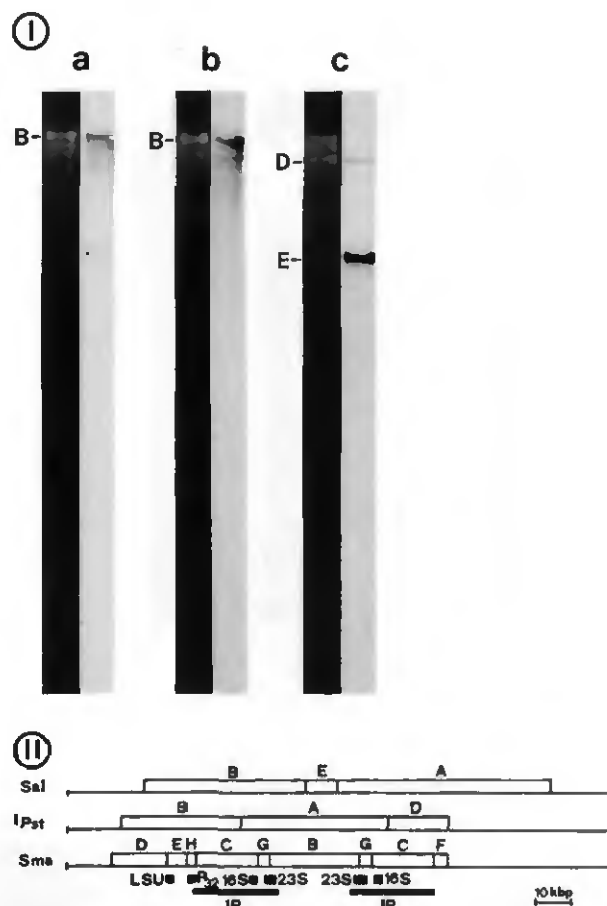


Fig. 3. Hybridization of the LSU-gene to restriction fragments from total chloroplast DNA. I: *Sal*I (a), *Pst*I (b), *Sma*I (c). The *Sma*-D band yields a weak radioactive signal, which is clearly visible on the original film: II: For the concept see Fig. 1, III

which is quite an exceptional gene arrangement, since in most chloroplast genomes studied so far the two genes are more than 50 kbp apart (see Discussion). However, quite recently it was shown by Palmer and Thompson (1982) that in the mung bean chloroplast genome the LSU gene is also very close to the P32 gene. Such a gene arrangement seems to be specific for legume chloroplast genomes.

#### 5. Construction of a Complete Circular *Sal*I-*Pst*I-*Sma*I Map

Consulting Table 1 and comparing the listed fragments with those already mapped (Fig. 3, II) reveals that the fragments *Sal*-C and D, *Pst*-C, E, F, G, H and I and the fragments *Sma*-A and I (double fragment) remain to be mapped. To place these fragments on the genome map, we cloned some of them, i.e., *Pst*-E, G, H, and I into pBR322 and used these fragments either directly in digestion experiments or as radioactive probes in hybridization experiments. Direct digestion experiments revealed (not shown) that *Pst*-G and I have not sites for *Sal*I nor for *Sma*I, whereas *Pst*-E is cut with *Sal*I yielding two fragments SP-G<sub>1</sub> and SP-H. *Pst*-H is cut twice by *Sma*I yielding fragments of 0.9 kbp (*Sma*-I), 0.9 kbp (PSm-N) and 0.1 kbp (PSm-O). This information is not sufficient to place any of the remaining fragments on the map.

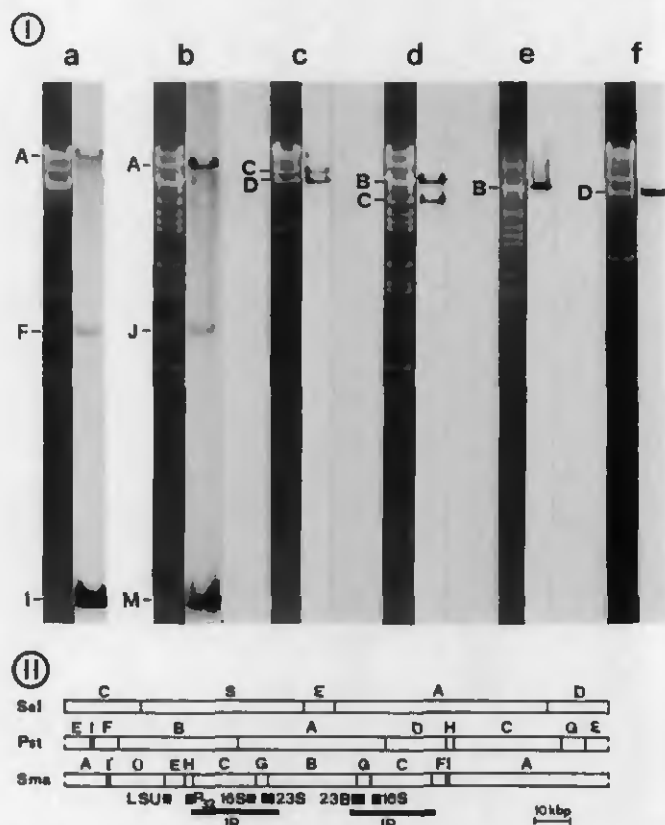


Fig. 4. Hybridization of selected cloned soybean chloroplast fragments with restriction fragments from total chloroplast DNA. I: *Pst*-H hybridized with *Sal*I digests (a), *Sal*I plus *Sma*I digests (b); *Pst*-E hybridized with *Sal*I digest (c), *Sal*I plus *Sma*I digest (d); *Pst*-I hybridized with *Sal*I plus *Sma*I digest (e); *Pst*-G hybridized with *Sal*I digest (f)

The *Sma*I restriction site map could be completed using the labeled *Pst*-H as hybridization probe, as shown in Fig. 4, I. This probe hybridizes with *Sma*-A, F and I (profile a) and with SSm-A, J and M (profile b). Since the *Sma*-F is already positioned on the map we can place the strongly hybridizing *Sma*-I next to *Sma*-F and the weakly hybridizing *Sma*-A next to *Sma*-I, as shown in Fig. 4, II. This necessitates placing *Pst*-H next to *Pst*-D, which correlates with the observation mentioned above, that *Pst*-H is twice cut by *Sma*I. The *Sma*-I band contains two fragments of equal mobility. Therefore, there remains a last, small *Sma* fragment to be mapped. This fragment (*Sma*-I') is tentatively placed between *Sma*-A and D.

The *Sal*I map can now easily be completed. *Sal*-C is cut by *Sma*I (see Table 1) whereas *Sal*-D is not cut. Therefore, the only way to place the two *Sal* fragments is as presented in Fig. 4, II, i.e., *Sal*-D is next to *Sal*-A and *Sal*-C is next to *Sal*-B.

There remain five *Pst* fragments to be mapped (*Pst*-C, E, F, G, I). Among these five fragments *Pst*-F is the only one cut by *Sma*I (see Table 1). Since only two *Sma*I sites remain available that are only about 900 bp apart (*Sma*I'), we must place *Pst*-F next to *Pst*-B (Fig. 4, II). This result supports at the same time the correct placing of *Sma*-I' (see above). To place *Pst*-E, which is cut with *Sal*-I, we hybridized it with the *Sal* fragment pattern (Fig. 4, I, panel

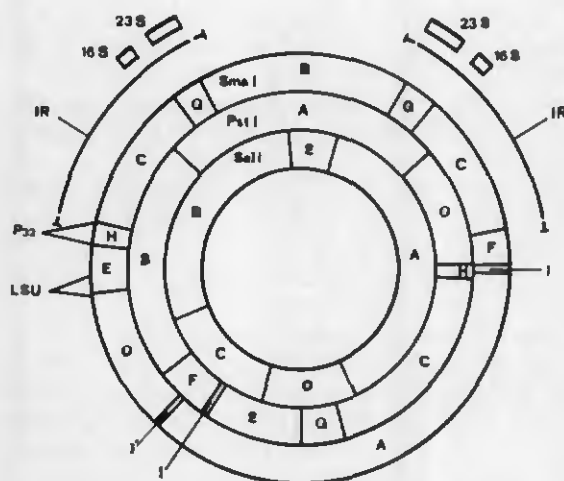


Fig. 5. Circular *Sal*I-*Pst*I-*Sma*I map of soybean chloroplast DNA, showing the two inverted repeats with the approximate size and position of the regions encoding 16S and 23S rRNA, the LSU gene and the P32 gene

c). *Pst*-E hybridized with *Sal*-C and D, which allows it to be placed as shown in Fig. 4, II. *Pst*-E also strongly hybridizes with SSm-B and C (panel d). This observation is crucial for mapping the small *Pst*-I fragment, which was also used as a radioactive probe: it hybridized only with SSm-C (panel e). SSm-C contains parts of *Pst*-F and E. Therefore *Pst*-I must be placed between *Pst*-E and F, as shown in Fig. 4, II.

Finally we hybridized *Pst*-G to the *Sal* fragment pattern (panel f). *Pst*-G interacts with *Sal*-D and therefore we can place it next to *Pst*-E. This leaves enough space to place the large fragment *Pst*-C between *Pst*-H and E. *Pst*-C is cut by *Sal*I, yielding two SP fragments, SP-B and L of 26 and 4 kbp, respectively (see Table 1). Thereby all *Sal*, *Pst*, and *Sma* fragments are placed on the circular chloroplast genome of soybean as depicted in Fig. 5.

## Discussion

According to our own experience, soybean plants are not a very suitable source from which to obtain a good yield of pure chloroplasts and it is also difficult to obtain covalently closed, circular DNA in preparative amounts. The DNA was always degraded, having an average size of about 60–70 kbp. The buoyant density profiles in CsCl had a peak at 1.696 g/cm<sup>3</sup> and usually a shoulder at 1.694 g/cm<sup>3</sup> indicating slight contamination by nuclear DNA. This did not interfere too strongly with the genome analysis, but both the reduced DNA size and the slight contamination contributed to the considerable background seen in the gels, especially in the upper part of the fragment patterns. The construction of an unequivocal *Sal*I-*Pst*I-*Sma*I restriction site map was possible nevertheless due to the accumulated knowledge on chloroplast genomes from other plants. Particularly helpful was the use of exogenous DNA probes of defined genetic function that allowed rapid identification of analogous DNA fragments on the soybean chloroplast genome. This approach will certainly be used more frequently in the future, especially since it becomes increasingly apparent (Fluhr and Edelman 1981b; Palmer and

Thompson 1982) that the structural parts of chloroplast genes show a high degree of sequence homology.

The circular chloroplast genome of soybean corresponds in overall size and general anatomy to that of mung beans. Such a result is in line with its taxonomic position within the legume family. As a member of the sub-tribe *phaseoleae-glycimae* it is closer to the mung bean (tribe *phaseoleae*) than to, e.g., the broad bean (tribe *viciae*) which contains a chloroplast genome of reduced size having lost a repeated segment during evolution [Palmer and Thompson 1982].

Particularly interesting is the finding that the LSU gene is only a few kbp from the P32 gene, as found in mung bean and pea (Palmer and Thompson 1982) but contrary to all other cases studied like corn (Bedbrook et al. 1979), mustard (Link 1981), tobacco (Seyer et al. 1981), spinach (Driesel et al. 1980; Westhoff et al. 1981) and duckweed (Van Ee et al. 1982) where the two genes are at least 50 kbp apart. The P32 gene [Mattoo et al. 1981] is heavily transcribed as is the LSU gene and it will be of great interest to study on a transcriptional level the consequences resulting from such a large spatial gene rearrangement. It seems that the relative proximity of the two genes is a characteristic of legume chloroplast genomes.

A comparison of the *SalI*-*PstI*-*SmaI* restriction site map of soybean with that of mung bean does not reveal any similarities, except that in both cases the 23S rRNA gene contains an *SmaI* site and the rDNA leader part a *PstI* site at seemingly identical positions. To measure the degree of sequence homology/divergence between the two chloroplast genomes, cross-hybridization experiments will be necessary as already successfully done with other members of the legume family (Palmer and Thompson 1982).

**Acknowledgments.** We are grateful to Xin Minzhong for technical and C. Bachmann for clerical help. We received financial support from the Swiss National Research Foundation (3,452.79 to E.S.) and from Nestlé Products Technical Assistance Co., Ltd.

## References

- Bedbrook JR, Bogorad L (1976) Endonuclease recognition sites mapped on *Zea mays* chloroplast DNA. *Proc Natl Acad Sci USA* 73:4309-4313
- Bedbrook JR, Coen DM, Beaton AR, Bogorad L, Rich A (1979) Location of the single gene for the subunit of ribulose biphosphate carboxylase on the maize chloroplast chromosome. *J Biol Chem* 254:905-910
- Bovenberg WA, Kool AJ, Nijkamp HJJ (1981) Isolation, characterization and restriction endonuclease mapping of *Petunia hybrida* chloroplast DNA. *Nucleic Acids Res* 9:503-517
- Bowman CM, Koller B, Delius H, Dyer TA (1981) A physical map of wheat chloroplast DNA showing the location of the structural genes for the rRNAs and the large subunit of ribulose 1,5-bisphosphate carboxylase. *Mol Gen Genet* 183:93-101
- Chu NM, Oishi KK, Tewari KK (1981) Physical mapping of the pea chloroplast DNA and localisation of the rRNA genes. *Plasmid* 6:279-292
- Denhardt DT (1966) A membrane-filter technique for the detection of complementary DNA. *Biochem Biophys Res Commun* 23:641-646
- Driesel AJ, Speirs J, Bohnert HJ (1980) Spinach chloroplast mRNA for a 32,000 dalton polypeptide. Size and location on the physical map of the chloroplast DNA. *Biochim Biophys Acta* 610:297-310
- Fluhr R, Edelman M (1981a) Physical mapping of *Nicotiana tabacum* chloroplast DNA. *Mol Gen Genet* 181:484-490
- Fluhr R, Edelman M (1981b) Conservation of sequence arrangement among higher plant chloroplast DNAs: molecular cross hybridization among *solanaceae* and between *nicotiana* and *spinacia*. *Nucleic Acids Res* 9:6841-6853
- Gordon KHJ, Crouse EJ, Bohnert HJ, Herrmann RG (1981) Restriction endonuclease cleavage site map of chloroplast DNA from *Oenothera parviflora*. *Theor Appl Genet* 59:281-296
- Jenni B, Stütz E (1978) Physical mapping of the rDNA region of *Euglena gracilis* chloroplast DNA. *Eur J Biochem* 88:127-134
- Jurgenson JE, Bourque DP (1980) Mapping of rRNA genes in an inverted repeat in *Nicotiana tabacum* chloroplast DNA. *Nucleic Acids Res* 8:3505-3516
- Koller B, Delius H (1980) *Vicia faba* chloroplast DNA has only one set of rRNA genes as shown by partial denaturation mapping and R-loop analysis. *Mol Gen Genet* 178:261-269
- Kolodner R, Tewari KK (1975) The molecular size and conformation of the chloroplast DNA from higher plants. *Biochim Biophys Acta* 402:372-390
- Kolodner R, Tewari KK (1979) Inverted repeats in chloroplast DNA from higher plants. *Proc Natl Acad Sci USA* 76:41-45
- Link G (1981) Cloning and mapping of the chloroplast DNA sequences for two mRNAs from mustard (*Sinapis alba* L.). *Nucleic Acids Res* 9:3681-3694
- Link G, Chambers SE, Thompson JA, Falk H (1981) Size and physical organization of chloroplast DNA from mustard (*Sinapis alba* L.). *Mol Gen Genet* 181:454-457
- Mattoo AK, Pick U, Hoffman-Falk H, Edelman M (1981) The rapidly metabolized 32,000 - dalton polypeptide of the chloroplast is the "proteinaceous shield" regulating photosystem II electron transport and mediating diuron herbicide sensitivity. *Proc Natl Acad Sci USA* 78:1572-1576
- Orozco EM, Gray PW, Hallick RB (1980) *Euglena gracilis* chloroplast rRNA transcription units. The location of tRNA, 5S, 16S and 23S rRNA genes. *J Biol Chem* 255:10991-10996
- Palmer JD (1982) Physical and gene mapping of chloroplast DNA from *atriplex triangularis* and *cucumis sativa*. *Nucleic Acids Res* 10:1593-1605
- Palmer JD, Thompson KF (1981) Rearrangements in the chloroplast genomes of mung bean and pea. *Proc Natl Acad Sci USA* 78:5533-5537
- Palmer JD, Thompson WF (1982) Chloroplast DNA rearrangements are more frequent when a large inverted repeat is lost. *Cell* 29:537-550
- Rigby PWJ, Dieckmann M, Rhodes C, Berg P (1977) Labelling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. *J Mol Biol* 113:237-251
- Schwarz Zs, Kössel H (1980) The primary structure of 16S rDNA from *Zea mays* chloroplast is homologous to *E. coli* 16S rRNA. *Nature* 283:739-742
- Seyer P, Kowallik KV, Herrmann RG (1981) A physical map of *nicotiana tabacum* plastid DNA including the location of structural genes for rRNAs and the large subunit of ribulose biphosphate carboxylase/oxygenase. *Current Genet* 3:189-204
- Siegenthaler PA, Depéry F (1976) Influence of unsaturated fatty acids in chloroplasts. *Eur J Biochem* 61:573-580
- Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503-517
- Sytenko R, Levings CS, Shah DM (1978) Physico-chemical characterization of mitochondrial DNA from soybean. *Plant Physiol* 61:460-464
- Taubert B (1981) *Leguminosae*. In: Engler A, Prantle B (eds) *Die natürlichen Pflanzenfamilien*, vol III. Engelmann, Leipzig, pp. 70-385
- Thompson JA, Hausmann P, Knoth R, Link G, Falk H (1981) Electron microscopical localization of the 23S and 16S rRNA genes within an inverted repeat of two chloroplast DNAs. *Current Genet* 4:25-28

- Van Ec, JH, Vos YJ, Planta RJ (1980) Physical map of chloroplast DNA of *Spirodela oligorrhiza*; analysis by the restriction endonucleases Pst I, Xho I and Sac I. *Gene* 12:191-200
- Van Ec, JH, Vos YJ, Bohnert HJ, Planta RJ (1982) Mapping of genes on the chloroplast DNA of *Spirodela oligorrhiza*. *Plant Mol Biol* 1:117-131
- Westhoff P, Nelson N, Bünemann H, Herrmann RG (1981) Localization of genes for coupling factor subunits on the spinach plastid chromosome. *Current Genet.* 4:109-120
- Whitfeld PR, Bottomley W (1980) Mapping of the gene for the large subunit of ribulose biphosphate carboxylase on spinach chloroplast DNA. *Biochem Intern* 1:172-178

Communicated by J. Schell

Received August 23 / November 4, 1982

#### Note Added in Proof

The restriction site map given here can be correlated with that given in the adjacent paper of Palmer et al. (1983) considering the relative position of the two inverted repeats and the P32 gene which maps within the 34 kbp PvuII fragment.

**Nucleotide sequence of soybean chloroplast DNA regions which contain the *psb A* and *trn H* genes and cover the ends of the large single copy region and one end of the inverted repeats**

Albert Spielmann and Erhard Stutz\*

Laboratoire de Biochimie, Université de Neuchâtel, Chantemerle 18, CH-2000 Neuchâtel, Switzerland

Received 11 August 1983; Revised and Accepted 26 September 1983

**ABSTRACT**

The soybean chloroplast *psb A* gene (photosystem II thylakoid membrane protein of Mr 32 000, lysine-free) and the *trn H* gene (tRNA<sup>His</sup><sub>GUG</sub>), which both map in the large single copy region adjacent to one of the inverted repeat structures (IR1), have been sequenced including flanking regions. The *psb A* gene shows in its structural part 92% sequence homology with the corresponding genes of spinach and *N. debneyi* and contains also an open reading frame for 353 aminoacids. The aminoacid sequence of a potential primary translation product (calculated Mr, 38 904, no lysine) diverges from that of spinach and *N. debneyi* in only two positions in the C-terminal part. The *trn H* gene has the same polarity as the *psb A* gene and the coding region is located at the very end of the large single copy region. The deduced sequence of the soybean chloroplast tRNA<sup>His</sup><sub>GUG</sub> is identical with that of *Zea mays* chloroplasts. Both ends of the large single copy region were sequenced including a small segment of the adjacent IR1 and IR2.

**INTRODUCTION**

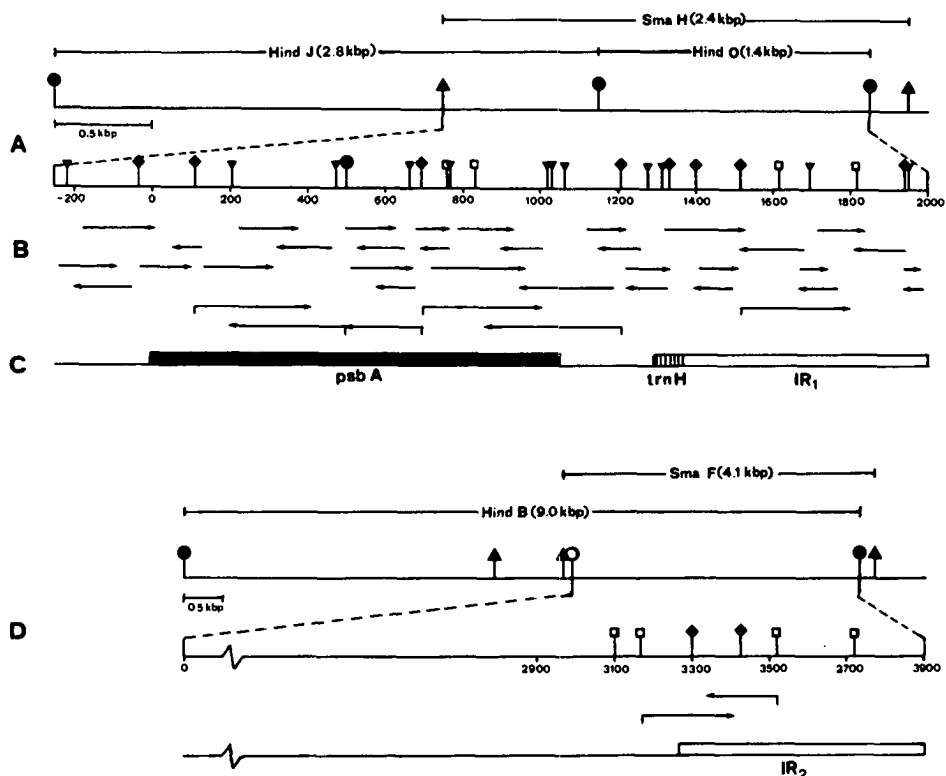
We have recently mapped the *psb A* gene on the soybean (*Glycine max.*) chloroplast genome in close vicinity of one of the inverted repeat regions (1). This gene codes for the so-called Mr 32 000 rapidly labeled photosystem II thylakoid membrane protein (2,3) which may be involved in the binding of urea and triazine herbicides (4,5). Zurawski et al. (6) sequenced the *psb A* gene region from *Spinacia oleracea* and *Nicotiana debneyi* chloroplast DNA and they observed in both cases identical open reading frames for 353 aminoacids equivalent to a protein of Mr 38 950. The size of this primary translation product surpasses the size of the rapidly labeled thylakoid membrane protein (Mr 32 000 to 36 000); the authors therefore argued that translation might start at a second ATG in the same reading frame, reducing thereby the size of the translation product by about 4 000. On the other hand, the total sequence identity of the reading frame in both types of chloroplast DNA strongly suggested that the entire reading frame must be functional.

In view of these open questions and considering the functional importance of this gene product it seemed warranted to sequence the psb A region of the soybean chloroplast genome. Representatives of the legume family are known to contain chloroplast genomes which have undergone relative to other angiosperm chloroplast genomes considerable DNA rearrangements (7,8). Therefore, size and fine anatomy of the soybean chloroplast psb A gene region might be different from those reported and answer some of the open questions.

Swamy and Pillay (9) recently identified soybean chloroplast tRNA<sup>His</sup> without, however, mapping the corresponding gene. In case of spinach the trn H gene maps between the psb A gene and the inverted repeat (10). We included in our sequence studies the corresponding DNA segment and furthermore determined the beginning of the inverted repeats by sequencing the corresponding DNA region on the other side of the large single copy region. This allowed to exactly position both the psb A and trn H gene relative to one of the inverted repeats, which are structural hallmarks of most higher plant chloroplast genomes.

### MATERIALS AND METHODS

Isolation of soybean chloroplast DNA, restriction sites analysis and mapping of the psb A gene on the circular chloroplast genome have been described (1). For a more detailed restriction site mapping and sequencing of the relevant region, we cloned HindIII fragments of total chloroplast DNA into pBR322 and tested the clones with both, a nick-translated (11) HpaII fragment obtained from the clone pSoc B511 (1,3) which carries 330 nucleotides of the spinach chloroplast psb A gene (probe a) and a HpaII-SalI fragment (850 bp) which carries the 3' end of the psb A gene including about 200 bases of the adjacent inverted repeat (probe b). A HindIII fragments of 2.8 kb (HindIII-J) interacted only with probe a, a HindIII fragment of 1.4 kb (HindIII-0) interacted with probe a and b and a HindIII fragment of 9.0 kb (HindIII-B) interacted only with probe b. Further mapping experiments allowed to place the three HindIII fragments as shown in Fig. 1A,D. For the sequencing experiments we used the entire fragments HindIII-0, the HindIII-SmaI subfragment of HindIII-J (0.8 kb) and a HindIII-PstI subfragment (3.9 kb) of HindIII-B (Fig. 1,A,D). DNA fragments were isolated and purified as described (12). For DNA sequencing we used both current methods (13,14) and as specified (12). Enzymes were purchased from Boehringer-Mannheim and New



**Fig. 1.** Restriction sites map and strategy used to sequence the *psb A* and *trn H* gene region of the soybean chloroplast DNA. A. Restriction sites of SmaI-H fragment (1). B. Arrows represent portions of the 5' end labeled fragments from which unambiguous sequences could be established (13);  $\rightarrow$  RNA-like strand,  $\leftarrow$  coding strand; arrows with a short vertical line indicate regions sequenced according to (14). C. Location of structural parts of the *psb A* gene, *trn H* gene and parts of the inverted repeat [IR1]. D. Restriction sites map of SmaI-F fragment and position of parts of the inverted repeat [IR2].  $\nabla$  HindIII,  $\uparrow$  SmaI,  $\nabla$  Sau3A,  $\nabla$  HinfI,  $\square$  HaeIII,  $\circ$  PstI. Numbers on scales in A and D refer to nucleotide sequence position given in Figs 2 and 3.

England Biolabs and used following the instructions of the supplier.  $^{32}$ P-ATP was from Amersham.

## RESULTS

### 1. Sequence analysis of the *psb A* gene

In Fig. 1 we show a stretch of the soybean circular chloroplast DNA which carries the *psb A* gene. In Fig. 2 we give the nucleotide sequence of the

[illegible]

[illegible]

**Fig. 2.** Nucleotide sequence of the psb A gene and flanking regions. Only the RNA like strand is given starting with the 5' position. Aligned with the soybean chloroplast psb A gene region (G.m.) are the corresponding sequences of Spinacia oleracea (S.o.) and Nicotiana debneyi (N.d.); within the structural part only deviations from the soybean sequence are given; within the flanking parts the entire sequences are shown and aligned such as to maximize sequence homology. Homologous regions are boxed. The deduced aminoacid sequence is given; the first methionine of the open reading frame is taken as position 1. Potential regulatory sites in the 5' flanking and a potential stem-loop structure in the 3' flanking part are underlined. Prokaryotic promoter consensus sequences for the -35 and -10 region (15) are given.

psb A gene and its flanking regions along with the deduced amino acid sequence for the structural part of the gene. For comparative reasons we add the nucleotide sequences of the corresponding parts of the S. oleracea and N. debneyi chloroplast DNA. The results of this comparative study can be summarized as follows : 1) The coding part of the soybean psb A gene is

identical in length to that of S. oleracea and N. debneyi, i.e., the reading frame is also open for a maximum of 353 aminoacids. 2) The soybean psb A gene diverges in its coding part at 73 (78) nucleotide positions from the S. oleracea (N. debneyi) gene. 3) With only three exceptions, the nucleotide differences occur in the wobble position, and only two mutations lead to a change in the aminoacid composition : aspartic acid replaces glutamic acid (position 347) and isoleucine replaces threonine (position 351). 4) Codons for lysine are absent in both cases. 5) About 130 nucleotide positions of the 5' flanking part are highly conserved, the sequence homology being in the range of 80 to 85%. This region contains potential promotor sites. Within this region the soybean sequences shares once a pentanucleotide gap with the N. debneyi sequence and once a heptanucleotide gap with the S. oleracea sequence. 6) Within the positions -130 to -271 sequence homology is very low especially due to multiple short insertion/deletions. 7) The 115 positions of the 3' flanking part are homologous, respectively, to 63% and 57% with the S. oleracea and N. debneyi counterparts. This includes gaps required for maximal sequence alignment. 8) A stem and loop structure with 21 basepairs can be formed [positions 1094 - 1115  $\curvearrowright$  1121 - 1143] similar to but not identical with the stem and loop structures proposed for the corresponding sequences of S. oleracea and N. debneyi (6). They may serve as transcription termination signals.

## 2. Sequence analysis of the trn H gene and of a segment overlapping the large single copy region and the inverted repeats

The circular soybean chloroplast genome contains two inverted repeats which separate a large (84 kb) from a small (24 kb) single copy region (1,8). The psb A gene maps close to one of the inverted repeats (1) the identified trn H gene (9) was not located yet. We anticipated, however, that the trn H gene would map between the psb A gene and the IR1, as seen in other angiosperm chloroplast genomes (10). In order to exactly locate the relative map positions of these genes and study the fine anatomy we sequenced the gap between the psb A gene and the beginning of the IR1, including a small part of the IR2. To identify the beginning of the IR1 it was necessary to sequence the corresponding DNA segment on the other side of the large single copy region which was mapped as shown in Fig. 1D. The sequence results are given in Fig. 3. The gap between the terminator codon of the psb A coding region and the first nucleotide of the inverted repeat is 319 positions (nucleotide positions

**Fig. 3.** Nucleotide sequence of the trn H gene region, its flanking parts and parts of the inverted repeats IR1 and IR2. Only the RNA like strand is shown. The structural part of the trn H gene and IR1 and IR2 are framed. Note that the psb A and trn H gene have the same polarity. Counting of nucleotide positions is continuous from Fig. 2, note the overlap. For IR1 the strand with the polarity of the trn H gene is given. For IR2 the sequence of the opposite strand (by definition) is given including 70 positions of the large single copy region adjacent to IR2.

Seventy positions of both ends of the large single copy region and 250

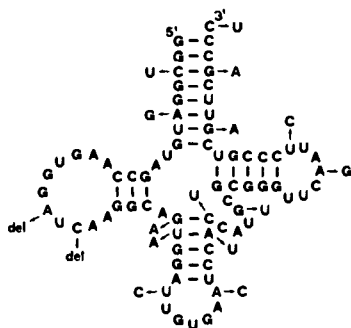


Fig. 4. Cloverleaf structure (unmodified) of soybean chloroplast tRNA<sup>His</sup><sub>GUG</sub> as deduced from the trn H sequence. The 5' terminal G is taken as position 1. Arrows point towards nucleotides present at that position in Euglena gracilis tRNA<sup>His</sup><sub>GUG</sub> (22); del : deletion.

positions of the IR1 and IR2 are aligned and compared in Fig. 3. IR1 and IR2 show perfect sequence homology for the entire analysed segment (250 positions). The two ends of the single copy region have nothing in common within the analysed stretch (70 positions), i.e., the structural part of the trn H gene is certainly not duplicated in this genome contrary to the situation in Zea mays (16). The size and function of an open reading frame which starts at position 3348 within IR2 (Fig. 3) and continues in the large single copy region (coding strand) is presently under investigation.

## DISCUSSION

### The psb A gene

The soybean chloroplast psb A gene is in its structural properties essentially identical to that of S. oleracea and N. debneyi (6). The length of the transcribed region (distance between the most likely transcription initiation and termination) is well within 1.2 kb, what agrees with the size of the major RNA which interacted with a psb A gene probe (17). These authors compared the psb A gene transcripts of several angiosperm chloroplasts including soybean and spinach. They observed for soybean, but not for spinach, that two minor transcription products of about 1.0 and 0.2 kb also interacted with the psb A DNA probe. They thought, however, that these minor RNAs were specific degradation products of the 1.2 kb RNA, the specific cleavage site being on the 3' terminal part of the coding region. A sequence comparison between the soybean and spinach psb A gene around nucleotide position 900

(Fig. 2), where preferential cleavage would occur in case of soybean, reveals no particular differences, i.e., the soybean chloroplast must contain a ribonuclease(s) which differ in specificity from that of spinach chloroplasts.

Hoffman-Falk et al. (18) studied the Mr 32 000 thylakoid membrane protein from several angiosperms and the alga Chlamydomonas reinhardtii. They found extensive similarities at levels of precursor maturation, membrane orientation and primary structure. This is in accordance with the psb A sequence data published sofar. There is no doubt that constraints on the primary structure of the entire translated region are very rigid, permitting only point mutations (with few exceptions) in the wobble position as shown now in chloroplast genomes of representatives from three distant plant families (Chenopodiaceae, Solanaceae, Leguminosae). Of particular importance is the observation that also in case of soybean the psb A gene codes for 353 aminoacids, the N-terminal part being to 100% homologous with that of S. oleracea and N. debneyi. This strongly suggests that the first 36 codons of the open reading frame are translated and essential in the primary translation product. However, McIntosh (19) reported very recently that the psb A genes from Zea mays and Amaranthus hybridus encode a protein of 317 aminoacids only. According to this preliminary report, it seems possible that the first 36 aminoacids are not required for a functional 32 kd thylakoid membrane protein. More analytical data, however, are necessary to obtain a clear picture concerning the size difference between the coding region and the gene product. Zurawsky et al. (6) discussed several possibilities to explain this discrepancy.

#### The trn H gene and the terminal part of the inverted repeats

A trn H gene was mapped adjacent to one of the inverted repeats on the chloroplast genomes of Spinacia oleracea (10) and Phaseolus vulgaris (20). For Zea mays the trn H gene was mapped within the inverted repeats and it was shown that the trn H gene slightly overlaps with a gene of opposite polarity which is transcribed into a RNA of 1.6 kb (16). The soybean trn H gene sits with its structural part right at one end of the large single copy region and a potential transcription termination site is located on the inverted repeat. The ribonucleotide sequence deduced from the trn H gene is to 100% identical with that of Zea mays chloroplasts. All highly conserved positions (21) are maintained in the trn H sequence. In Fig. 4 we compare the soybean with the Euglena gracilis chloroplast tRNA<sup>His</sup><sub>GUG</sub> (22) as deduced from the corresponding

structural genes. There is about 80% sequence homology. In both cases the D-stem is shortened and the extra loops are of identical length. The 3' terminal CCA are not encoded in the genes, a property chloroplast tRNA genes seem to share with eukaryotic tRNA genes (21). All three chloroplasts trn H genes sequenced so far code for tRNA<sup>His</sup><sub>GUG</sub>. According to extensive hybridization experiments using a variety of chloroplast genomes (for references, see 10) no second trn H gene was identified, i.e., a possible gene for tRNA<sup>His</sup><sub>AUG</sub> is still undetected. Nevertheless, we can see (Fig. 2) that the codon CAT (CAU) is frequently used in translating both the spinach and soybean Mr 32 000 thylakoid membrane protein, this in agreement with the wobble hypothesis.

The three chloroplasts trn H genes sequenced so far have a different genetic environment. The soybean trn H gene is at one end of the large single copy region proximate to and with the same polarity as the psb A gene. Its transcription termination site is most likely part of the inverted repeat. The maize trn H gene is integral part of the inverted repeat, occurs therefore twice per circular genome and it overlaps slightly with a protein coding gene of opposite polarity (16). Finally the Euglena gracilis trn H gene is the second gene in a cluster of five tandemly arranged tRNA genes (22), which most likely are co-transcribed and under the control of one promoter. trn H gene transcription regulation must be different in the three types of chloroplasts, a point of considerable interest.

It is known from cross hybridization experiments that the inverted repeats of chloroplast genomes are structurally related, displaying considerable sequence homology (7). Our sequence results define for the first time the exact beginning of IR1 and IR2 on a angiosperm chloroplast genome. A comparison with the Zea mays chloroplast IR (16) reveals that a short stretch of 32 positions is to 91% homologous with a soybean chloroplast DNA segment (position 1510-1542). Most likely, sequence homology between the two IRs continues beyond the sequenced positions (16). It's noteworthy that the short homologous parts map at different places within the two kinds of inverted repeats. In case of soybean this short segment starts 138 positions inside of the beginning of the IR, for maize it is most likely about 0.5 kb away from the IR start. This and the different location of the trn H gene shows that during evolution not only the single copy region underwent DNA rearrangements but also the rather conservative inverted repeats region.

ACKNOWLEDGEMENTS

We received financial support from the Swiss National Science Foundation (3.183.82 to E.S.) and from Nestlé Products Technical Assistance Co., Ltd. We are grateful to Ch. Bachmann for secretarial and B. Schlunegger and I. Howald for technical assistance. E. Roux gave us helpful advices for DNA sequencing and computer analysis. This report represents part of a Ph.D. thesis (A.S.) submitted to the Science Department of the University of Neuchâtel.

\*To whom reprint requests should be addressed.

REFERENCES

1. Spielmann, A., Ortiz, W. and Stutz, E. (1983) *Mol. Gen. Genet.* 190, 5-12.
2. Bedbrook, J.R., Link, G., Coen, D.M., Bogorad, L. and Rich, A. (1978) *Proc. Natl. Acad. Sci. USA* 75, 3060-3064.
3. Driesel, A.J., Speirs, J. and Bohnert, H.J. (1980) *Biochim. Biophys. Acta* 610, 297-310.
4. Mattoo, A.K., Pick, U., Hoffman-Falk, H. and Edelman, M. (1981) *Proc. Natl. Acad. Sci. USA* 78, 1572-1576.
5. Steinback, K.E., McIntosh, L., Bogorad, L. and Arntzen, C.J. (1981) *Proc. Natl. Acad. Sci. USA* 78, 7463-7467.
6. Zurawski, G., Bohnert, H.J., Whitfield, P.R. and Bottomley, W. (1982) *Proc. Natl. Acad. Sci. USA* 79, 7699-7703.
7. Palmer, J.D. and Thompson, W.F. (1982) *Cell* 29, 537-550.
8. Palmer, J.D., Singh, G.P. and Pillay, D.T.N. (1983) *Mol. Gen. Genet.* 190, 13-19.
9. Swamy, G.S. and Pillay, D.T.N. (1982) *Plant Sci. Lett.* 25, 73-84.
10. Bohnert, H.J., Crouse, E.J. and Schmitt, J.M. (1982) *Encyclopedia of Plant Physiology, New Series*, vol. 14B, Parthier, B. and Boulter, D., eds, Springer-Verlag Berlin-Heidelberg, pp. 475-530.
11. Rigby, P.W.J., Dieckman, H., Rhodes, C. and Berg, P. (1977) *J. Mol. Biol.* 113, 237-251.
12. Graf, L., Roux, E., Stutz, E. and Kössel, H. (1982) *Nucleic Acids Res.* 10, 6369-6381.
13. Maxam, A.M. and Gilbert, W. (1980) *Methods in Enzymology*, vol. 65, pp. 499-560, Grossman, L. and Moldave, K., eds, Academic Press, New York.
14. Sanger, F., Coulson, A.R., Barrel, B.G., Smith, A.J.H. and Roe, B.A. (1980) *J. Mol. Biol.* 143, 161-178.
15. Hawley, D.K. and McClure, W.R. (1983) *Nucleic Acids Res.* 11, 2237-2255.
16. Schwarz, Z., Jolly, S.O., Steinmetz, A.A. and Bogorad, L. (1981) *Proc. Natl. Acad. Sci. USA* 78, 3423-3427.
17. Palmer, J.D., Edwards, H., Jorgensen, R.A. and Thompson, W.F. (1982) *Nucleic Acids Res.* 10, 6819-6832.
18. Hoffman-Falk, H., Mattoo, A.K., Marder, J.B. and Edelman, M. (1982) *J. Biol. Chem.* 257, 4583-4587.
19. McIntosh, L. (1983) *J. Cell. Biochem.* 27B, 295.
20. Mubumbila, M., Gordon, K.H.J., Crouse, E.J., Burkard, G. and Weil, J.H. (1983) *Gene* 21, 257-266.
21. Sprinzl, M. and Gauss, D.H. (1982) *Nucleic Acids Res.* 10, r1-r56.
22. Hollingsworth, M.J. and Hallick, R.B. (1982) *J. Biol. Chem.* 257, 12795-12799.

