

Stratégie de codage dans le système végétal

M. BOUDRAA

Institut d'Evolution moléculaire, Université Lyon I, F 69622 Villeurbanne cedex

Résumé

L'analyse statistique des séquences nucléotidiques végétales nous a permis de montrer que dans les cellules végétales coexistent deux stratégies de codage différentes : l'une nucléaire, l'autre chloroplastique. Nous ignorons s'il existe ou non une troisième stratégie chez les mitochondries, faute de données.

L'utilisation des codons est relativement homogène à l'intérieur de chaque type de génome (nucléaire et chloroplastique) : ceci est connu sous le nom « hypothèse du génome ».

Des ressemblances entre eubactéries et chloroplastes ont été évoquées et plusieurs données de la biologie moléculaire sont concordantes. Cependant, le présent travail montre que les gènes chloroplastiques ont un usage du code différent de celui d'*Escherichia coli*. Néanmoins, l'étude des changements évolutifs d'un gène donné chez différentes espèces montre que les changements observés ne sont pas en désaccord avec la théorie endosymbiotique de l'évolution.

Mots clés : Gènes de plantes et chloroplastes, hypothèse du génome, stratégie de codage, bases dégénérées.

Summary

Coding strategy variation in the plant system

Statistical analysis of nucleotide sequences of plants has allowed us to show that in the same cell two distinct coding strategies, nuclear and chloroplastic, coexist. It is unknown whether a third strategy exists for mitochondria, because but few genes have been sequenced. Each type of genome (nuclear and chloroplast) has a characteristic and relatively homogeneous codon usage throughout its genes. This is known as the « genome hypothesis ».

Eubacteria and chloroplasts have certain similar characteristics, on which some molecular biology data are in agreement. The present work shows that chloroplasts do not follow the bacterial scheme of codon usage. Nevertheless, a study of evolutionary changes in a particular gene sequenced in several species reveals that observed changes are not in disagreement with the symbiotic theory of evolution.

Key words : Plant and chloroplastic genes, genome hypothesis, coding strategy, degenerate bases.

I. Introduction

L'évolution moléculaire étudie les espèces par le biais de leurs macromolécules informatives (acides nucléiques et protéines), ce qui permet de nouvelles approches du processus évolutif. Une même protéine peut être théoriquement codée par un très grand nombre de séquences nucléotidiques. Le choix d'une séquence plutôt qu'une autre dépend de diverses contraintes : en premier lieu, l'organisation en triplets pour coder une protéine donnée. Le triplet utilisé pour coder un acide aminé donné dépend de l'organisme auquel appartient cette séquence.

La comparaison des séquences deux à deux (gène codant pour une même protéine chez différentes espèces) permet de retracer l'histoire évolutive du gène et, par conséquent, des organismes qui l'ont hébergé. Les gènes phylogénétiquement proches présentent une tendance similaire dans l'utilisation du code génétique (GRANTHAM & GAUTIER, 1980 ; GRANTHAM *et al.*, 1980 et 1981). Le choix entre codons synonymes (codant le même acide aminé) n'est pas arbitraire mais suit des règles qui semblent précises. Ainsi, le règne animal préfère les bases C et G en 3^e position des codons (GRANTHAM, 1980 ; GRANTHAM & GAUTIER, 1980 ; GRANTHAM *et al.*, 1980). Chez *E. coli* et la levure, on a montré que le choix entre codons synonymes favorise les codons appelant des ARNt fréquents dans la cellule et que l'importance de ce biais était d'autant plus grande que le gène était hautement exprimé (GRANTHAM *et al.*, 1981 ; IKEMURA, 1981 ; GOUY & GAUTIER, 1982). L'emploi du code génétique chez le bactériophage T7 paraît également être influencé par l'abondance des ARNt de l'hôte, particulièrement pour les gènes hautement exprimés (SHARP *et al.*, 1985). Toutefois cette tendance est moins forte chez les bactériophages que chez l'hôte (GRANTHAM *et al.*, 1985).

L'intérêt pour les biomacromolécules du monde végétal se développe, et certains résultats fondamentaux sont déjà acquis. Par exemple, on a montré l'existence d'introns dans les parties nucléotidiques codant des protéines végétales. Ces régions introniques obéissent au niveau de leurs jonctions à la règle de Chambon applicable à tous les gènes protéiques nucléaires des animaux : elles commencent toutes par le dinucléotide GT et se terminent par AG (SLIGHTOM, 1983). On a aussi déterminé chez les plantes, dans les parties 5' non traduites, les séquences régulatrices connues chez d'autres eucaryotes (SLIGHTOM, 1983 ; LYCETT *et al.*, 1983). De même, des séquences variantes du prototype AAUAAA (proposé comme signal de polyadénylation), localisées dans la partie 3' non traduite du gène, ont été trouvées chez les plantes (LYCETT *et al.*, 1983).

Quant aux gènes chloroplastiques, ils ont une nature procaryotique (SUBRAMAIAN *et al.*, 1983) dans leurs régions flanquantes 5' et 3'. L'arrangement de leurs gènes d'ARNr rappelle celui de *E. coli* chez lequel cependant, l'ARNr 4.5s est absent (TAKAIWA & SUGIURA, 1982). La région espaceur 16s-23s contient, comme un opéron d'*E. coli*, deux gènes codant pour deux ARNt (Ile et Ala).

Les gènes nucléaires et chloroplastiques des végétaux utilisent le code génétique universel : il n'y a aucune déviation connue chez les chloroplastes contrairement à ce qu'on observe chez les mitochondries.

LYCETT *et al.* (1983), en travaillant sur quelques séquences végétales, ont montré que, pour 5 acides aminés (Leu, Val, Ala, Gly et Thr), le codon préféré est toujours différent entre animaux et végétaux.

L'objet de ce travail est de décrire et de caractériser l'usage du code chez les végétaux supérieurs par des méthodes statistiques (analyse factorielle des correspondances, calcul de la fréquences des bases, test Chi-2 et comparaison de certains indices). Dans un premier temps, nous allons comparer les gènes végétaux entre eux. La deuxième partie sera consacrée aux comparaisons de chaque type de génome avec les données bibliographiques.

Nous étudions le comportement des séquences nucléaires, chloroplastiques et mitochondriales des végétaux vis-à-vis des 61 codons. Nous caractérisons le choix des codons pour chaque type de génome, nucléaire et organellaire. Plusieurs influences jouent sur le choix entre codons synonymes et c'est l'interaction de toutes ces influences qui est déterminante (GRANTHAM *et al.*, 1985 ; IKEMURA, 1985). Nous allons étudier certaines de ces influences en les comparant avec les données bibliographiques. En nous référant à la théorie endosymbiotique de l'évolution (MARGULIS, 1975) : les bactéries serviront de modèle de référence pour les organelles. Les gènes nucléaires de plantes seront discutés à la lumière de leurs homologues animaux.

II. Matériel et méthodes

A. Matériel

Le matériel d'étude est un ensemble de séquences nucléotidiques de plantes extrait de la banque GenBank Version 42 [système d'interrogation ACNUC (GOUY *et al.*, 1985)]. Il contient 20 séquences nucléaires de plantes et 17 séquences chloroplastiques de gènes protéiques (tabl. 1). Nous n'avons retenu qu'une séquence de chaque famille de gènes d'un même génome (certains gènes existent en plusieurs exemplaires chez un même individu, par exemple : la zéine). Lorsqu'un gène existe chez plusieurs espèces, seuls les séquences de quelques espèces sont représentées. Deux séquences mitochondriales seront traitées à titre indicatif.

B. Méthode (*)

1. L'analyse factorielle des correspondances (AFC)

L'AFC (BENZECRI, 1973) est une méthode multivariée qui vise à fournir une représentation graphique plus accessible du contenu d'un tableau de données. Cette méthode va nous permettre de traiter notre tableau de contingence croisant 39 séquences (les lignes) avec les 61 codons (les colonnes). Le résultat de l'analyse est une représentation de chaque séquence comme un point dans un espace multidimensionnel. La position de chaque point est fonction de la fréquence relative de chacun des 61 codons dans la séquence correspondante (GRANTHAM *et al.*, 1980 et 1981). La projection sur un plan de ce nuage de points multidimensionnel permet une visualisation simple des distances entre séquences. La méthode fournit le plan (dit plan factoriel) pour lequel la distorsion impliquée par cette projection est la plus faible. D'une façon analogue, les distances entre codons sont construites à partir de la variation de leurs fréquences dans les séquences. Les 2 plans factoriels (ARNm et codons) sont superposables.

(*) Tous les calculs ont été effectués par le système ANALSEQ de la banque ACNUC.

TABLEAU 1

Echantillon d'étude comprenant 20 séquences nucléaires, 17 chloroplastiques et 2 mitochondriales de plantes supérieures.

The sample studied, comprising 20 nuclear, 17 chloroplastic and 2 mitochondrial sequences of higher plants.

SEQ.	LONG.	Espèce	Protéine
Gènes nucléaires			
OAMA	1 317	<i>Hordeum vulgare</i>	Alpha-amylase, type A.
OAMB	1 284	<i>Hordeum vulgare</i>	Alpha-amylase, type B.
BNAP	537	<i>Brassica napus</i>	Napine, Protéine de réserve.
MACT	1 128	<i>Zea mays</i>	Actine.
MZEI	798	<i>Zea mays</i>	Zéine.
PP15	810	<i>Pisum sativum</i>	Protéine de fixation de la chlorophylle.
PRUB	543	<i>Pisum sativum</i>	rbcS.
PLEG	996	<i>Pisum sativum</i>	Légumine.
SGLY	2 455	<i>Glycin max</i>	Glycine.
PCHA	1 197	<i>Petroselinum hortense</i>	Chalacone synthétase.
PPOT	1 161	<i>Solanum tuberosum</i>	Potatine.
PLEC	741	<i>Phaseolus vulgaris</i>	Lectine.
PPHA	1 260	<i>Phaseolus vulgaris</i>	Phaséoline.
SPRO	264	<i>Glycin max</i>	Protéine inhibitrice de la protéase.
SLB2	438	<i>Glycin max</i>	Léghémoglobine.
SLB1	438	<i>Glycin max</i>	Léghémoglobine, pseudogène.
SLEC	858	<i>Glycin max</i>	Lectine.
TTHA	708	<i>Thaumatococcus daniellii</i>	Préthaumatine.
BGLI	957	<i>Triticum aestivum</i>	Gliadine, type alpha.
BHIS	411	<i>Triticum aestivum</i>	Histone H3.
Gènes chloroplastiques			
APSB	954	<i>Amaranthus hybridus</i>	Protéine de fixation de l'herbicide.
OATB	414	<i>Hordeum vulgare</i>	ATPase, sous-unité bêta.
ORUB	1 279	<i>Hordeum vulgare</i>	rbcL.
MATE	1 497	<i>Zea mays</i>	ATPase, sous-unité epsilon.
MRUB	1 428	<i>Zea mays</i>	rbcL.
SAII	1 062	<i>Sinapsis alba</i>	Protéine membranaire du photosystème II.
EASE	1 527	<i>Spinacia oleracea</i>	Chlorophylle P-680, apoprotéine alpha.
EPSB	1 062	<i>Spinacia oleracea</i>	Protéine membranaire de la thylakoïde.
ER19	345	<i>Spinacia oleracea</i>	Protéine ribosomique rps19'.
EPL2	861	<i>Spinacia oleracea</i>	Protéine ribosomique L2.
ES19	279	<i>Spinacia oleracea</i>	Protéine ribosomique S19.
TP32	1 062	<i>Nicotiana tabacum</i>	Protéine membranaire de la thylakoïde.
TS19	279	<i>Nicotiana tabacum</i>	Protéine membranaire cs19.
TRUB	1 434	<i>Nicotiana tabacum</i>	rbcL.
TPL2	801	<i>Nicotiana debenev</i>	Protéine membranaire L2.
MPS4	603	<i>Zea mays</i>	Protéine ribosomique CS4.
BLYF	963	<i>Triticum aestivum</i>	Cytochrome F.
Gènes mitochondriaux			
RCY2	783	<i>Oryza sativa</i>	Cytochrome oxidase, sous-unité II.
MCOB	1 167	<i>Zea mays</i>	Apocytochrome B.

SEQ. = séquence ; LONG : longueur (en nombre de bases).

rbc = ribulose biphosphate carboxylase ; L : grosse sous-unité ; S = petite sous-unité.

SEQ. = sequence ; LONG : length (in base number).

rbc = ribulose biphosphate carboxylase ; L : large subunit ; S = small subunit.

2. Calcul des indices

Pour l'étude d'une éventuelle relation entre la 3^e base du codon et les 2 précédentes nous utilisons 2 types d'indices :

- Le BC, dit indice du bon choix (GOUY & GAUTIER, 1982) :

$$BC = \frac{WWC + SSU}{WWY + SSY}$$

- où W (pour Weak) représente la base A ou U,
S (pour Strong) représente la base C ou G,
Y = les bases pyrimidiques, C ou U.

Les 2 types de bases W et S, correspondent respectivement aux énergies d'appariement les plus faibles et les plus fortes entre codon-anticodon. Des valeurs de l'ordre de 50 p. 100 de cet indice indiquent l'absence de biais dans l'utilisation du code génétique. Des valeurs très élevées montrent au contraire une forte tendance à l'établissement d'une énergie moyenne d'interaction codon-anticodon.

- Le triple indice WWC/WWY, SSC/SSY et MMC/MMY (GRANTHAM *et al.*, 1986).

MM = représente les dinucléotides mixtes, quand la 1^{re} base est de type W, la 2^e est S, et inversement. Dans l'hypothèse de l'optimisation de l'énergie d'interaction codon-anticodon, le rapport MMC/MMY servira de témoin, sa valeur sera intermédiaire entre la valeur élevée de WWC/WWY et la faible valeur de SSC/SSY. Ainsi ce triple indice tient compte des variations du (G + C) parmi différents gènes (voir ci-dessous).

III. Résultats

A. Etude des séquences végétales et comparaison des gènes nucléaires avec leurs homologues animaux

La répartition des 4 bases nucléotidiques n'est pas la même dans les différents génomes présents dans une cellule végétale. Dans les gènes nucléaires de l'échantillon, le contenu en G + C global est de 52 p. 100, celui en position III est de 58 p. 100. Par contre, les gènes chloroplastiques présentent des pourcentages du G + C de 41 p. 100 en toutes positions et de 30 p. 100 en 3^e position. Les 2 séquences mitochondriales dont on dispose ont des fréquences de G + C voisines de celles des gènes chloroplastiques. Les gènes nucléaires sont donc plus riches en G + C que ceux des organelles, en particulier en 3^e position. Comparés aux gènes de mammifères, les gènes nucléaires de plantes présentent une différence moins grande entre la composition totale et la composition en position III. Dans les 2 cas cependant, la position III est la plus riche en G + C. Il faut toutefois signaler que le nombre de séquences de plantes est faible. Cette tendance à préférer les bases dégénérées C et G a été observé dans un vaste échantillon de séquences de vertébrés (GRANTHAM, 1980 ; GRANTHAM *et al.*, 1985 et 1986) malgré une assez forte variabilité (IKEMURA, 1985).

L'utilisation de l'AFC a permis de distinguer les 2 grands groupes de codons utilisés par chacun des 2 types de génomes, nucléaire et organellaire. La figure 1 représente le plan factoriel des séquences. On observe selon l'axe horizontal une séparation entre les gènes nucléaires et organellaires. Cet axe, qui constitue le premier

facteur, exprime la source de variabilité la plus importante entre les différents ARNm de notre échantillon. Chacun des 2 groupes de séquences se caractérise par son affinité pour certains codons particuliers.

Sur le plan factoriel des codons (fig. 2), presque tous les codons du type NNS se rangent du côté des gènes nucléaires tandis que ceux du type NNW, se trouvent du côté chloroplastique et mitochondrial (N représente les 4 bases A, C, G ou U ; W, la base A ou U ; S, la base C ou G). Chaque type de gènes est donc caractérisé par le choix d'un type de codons.

L'étude du choix entre les 2 pyrimidines est un bon moyen pour se rendre compte de la sélection entre bases dégénérées, ce choix n'impliquant aucun changement d'acides aminés. Le rapport NNC/NNY (Y représentant la base C ou U) est de 60 p. 100 pour les gènes nucléaires, et de 30 p. 100 pour les gènes chloroplastiques ; les

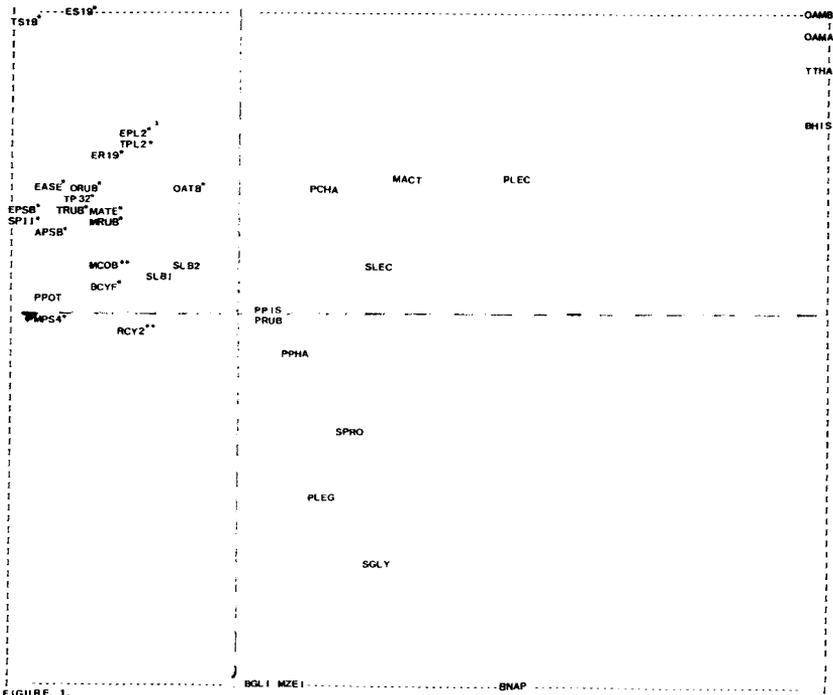


FIG. 1

Position des 39 séquences nucléaires, chloroplastiques et mitochondriales par analyse factorielle des correspondances, selon leur composition en codons.

Location of 39 sequences (nuclear, chloroplastic and mitochondrial) by factor analysis of correspondences of sequences according to their codon composition.

Les séquences chloroplastiques sont repérées par une étoile et les séquences mitochondriales par 2, les autres sont des gènes nucléaires. La signification des abréviations est donnée dans le tableau 1. *The chloroplast sequences are marked by an asterisk and mitochondrial by a double asterisk. The remaining entries are nuclear sequences. Signification of abbreviations is shown in table 1.*

valeurs pour CCN/YYN sont respectivement de 51 et 45 p. 100. Ce dernier rapport est en relation avec la distribution des bases pyrimidiques dans les différentes positions du codon, mais il dépend aussi de la composition des protéines en acide aminé proline codé par les codons CCN.

Le choix entre pyrimidines en position III se fait le plus souvent en faveur de la base C pour les gènes nucléaires de mammifères (GRANTHAM *et al.*, 1986). Les gènes nucléaires de plantes partagent cette caractéristique quand les 2 premières bases sont du type WW : le rapport WWC/WWY est de 67 p. 100, alors que celui de SSC/SSY est de l'ordre de 50 p. 100.

B. Comparaison des gènes organellaires et bactériens

Chez *E. coli*, il a été montré (GRANTHAM *et al.*, 1981 ; GOUY & GAUTIER, 1982) que pour les gènes hautement exprimés, le choix de la position III est conditionné par

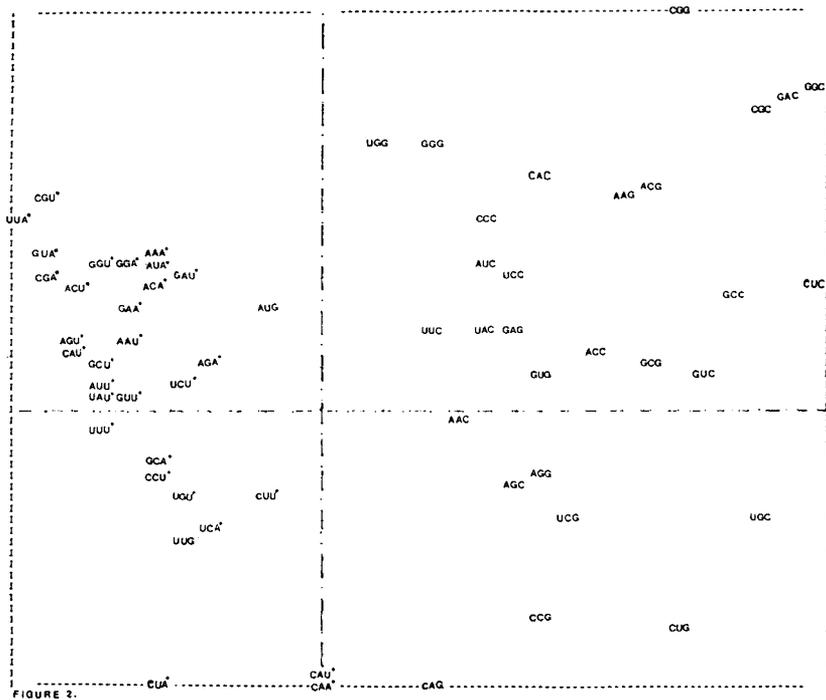


FIG. 2

Position des 61 codons par l'analyse factorielle des correspondances sur leurs fréquences dans les 39 séquences nucléaires, chloroplastiques et mitochondriales.

Location of the 61 codons by factor analysis of correspondences of their frequencies in the 39 nuclear, chloroplastic and mitochondrial sequences.

Cette figure se superpose avec la figure 1. Les codons se terminant par la base A ou U sont repérés par 1 étoile.

This figure can be surimposed on figure 1. Codons ending in A or U are marked by an asterisk.

la nature des 2 bases précédentes. La 3^e base préférée est C lorsque les 2 premières sont du type WW. Cette 3^e base est U si les 2 précédentes sont du type SS. Ceci a été interprété comme une tendance à l'établissement d'une énergie d'interaction moyenne entre codons et anticodons, afin d'assurer l'optimisation de la vitesse et de la fidélité de traduction de ces gènes (GROSJEAN *et al.*, 1978).

En nous référant à la théorie de l'endosymbiose, nous nous sommes intéressés aux gènes de protéines chloroplastiques abondantes en comparaison avec ceux d'*E. coli*. L'utilisation des codons à énergie d'appariement extrême (WWU et SSC) a été comparée, dans les séquences chloroplastiques, à celle des codons à énergie moyenne (WWC et SSU) qui leur sont synonymes. Pour ce faire, nous avons utilisé un test Chi-2 qui éprouve l'hypothèse d'indépendance entre les 2 premières bases (selon qu'elles sont WW ou SS) et la 3^e (selon qu'elle est C ou U). Il faut bien remarquer que le résultat du test ne dépend pas des fréquences en position III de C et U.

L'examen des séquences chloroplastiques des gènes hautement exprimés montre que c'est uniquement dans le cas de la *rbcL* (ribulose biphosphate carboxylase, grosse sous-unité), que les Chi-2 sont significatifs. Ils ne le sont pas pour d'autres protéines également abondantes dans la cellule (protéines ribosomiques et facteurs d'élongation) contrairement à ce qui a été observé chez *E. coli* (GOUY & GAUTIER, 1982). Nous nous sommes intéressés de très près à ce problème d'optimisation d'énergie d'appariement codon-anticodon pour notre échantillon, et plus particulièrement aux gènes chloroplastiques hautement exprimés du fait de leur forte homologie avec *E. coli* selon plusieurs critères. Nous avons étudié l'indice BC (GOUY & GAUTIER, 1982). Les valeurs de BC pour les gènes protéiques ribosomiques et les facteurs d'élongation des plantes supérieures et d'*Euglena gracilis* sont de l'ordre de 50 p. 100, ce qui indique qu'il n'y a pas de biais dans l'utilisation du code génétique. Les gènes de ces mêmes protéines chez *E. coli* et *Saccharomyces cerevisiae* présentent des valeurs nettement supérieures, se situant entre 75 et 81 p. 100. Des valeurs élevées de cet indice sont observées pour la *rbcL* des chloroplastes des plantes supérieures et d'*E. gracilis*, mais ces valeurs restent très faibles en comparaison avec la *rbcL* des cyanobactéries et de *Chlamydomonas reinhardtii* (tabl. 2).

Toutefois le BC peut être influencé par la composition en bases. Pour tourner cette difficulté, nous avons fait appel au triple indice WWC/WWY, SSC/SSY et MMC/MMY (GRANTHAM *et al.*, 1986). C'est encore uniquement dans le cas du gène protéique *rbcL* qu'on observe une forte optimisation de l'énergie d'appariement dans pratiquement toutes les espèces. On remarque que *C. reinhardtii* est un organisme unicellulaire qui possède un seul chloroplaste, tandis que les 5 autres espèces en ont plusieurs. La tendance à l'optimisation est très forte chez *C. reinhardtii* par rapport aux autres espèces (tabl. 2). *E. gracilis* est aussi un organisme unicellulaire, mais qui se comporte de la même façon que les plantes supérieures. IKEMURA (1985), a proposé la distinction entre organismes unicellulaires et organismes pluricellulaires, au lieu d'opposer les procaryotes aux eucaryotes par leurs usage du code. Dans notre cas, nous observons une différence entre *C. reinhardtii* et les cellules plurichloroplastiques. Nous disposons uniquement de la *rbcL*, gène partagé par plusieurs espèces dont une seule est monochloroplastique. L'étude d'autres gènes nous permettra peut être de généraliser cette distinction entre organismes monochloroplastiques et organismes plurichloroplastiques.

TABLEAU 2

*Etude de l'optimisation de l'énergie d'appariement codon-anticodon.
Study of optimisation of the energy of codon-anticodon binding.*

Protéine	Espèce	BC (%)	WWC/WWY (%)	SSC/SSY (%)	MMC/MMY (%)
Ribosomique	<i>Escherichia coli</i> (20 SEQ.)	75	77	31	45 *
	<i>Saccharomyces cerevisiae</i> (8 SEQ.)	81	75	14	65 **
	<i>Euglena gracilis</i> S12	47	44	50	30
	S7	46	27	09	25
	<i>Spinacia oleracea</i> L2	50	37	27	42
	S19	43	29	36	20
	<i>Nicotiana tabacum</i> Cs19	37	19	36	23
	L12	53	39	33	42
	<i>Zea mays</i> S4	49	18	19	27
	rbcL	<i>Anacystis nidulans</i>	70	88	42
<i>Anabaena 7120</i>		78	78	22	57 *
<i>Chlamydomonas reinhardtii</i>		89	74	02	33
<i>Euglena gracilis</i>		64	21	08	07
<i>Hordeum vulgare</i>		58	36	20	27
<i>Zea mays</i>		61	46	25	25
<i>Spinacia oleracea</i>		60	39	20	28
<i>Nicotiana tabacum</i>		64	45	18	23
Facteurs d'élongation	<i>Escherichia coli</i> (4 SEQ.)	79	87	23	49 *
	<i>Saccharomyces cerevisiae</i> (3 SEQ.)	77	75	19	45 **
	<i>Euglena gracilis</i>	46	20	20	16

Les valeurs des différents indices indiquent le degré d'optimisation de l'énergie d'appariement codon-anticodon pour les 3 types de protéines : protéines ribosomiques, grosses sous-unité de la ribulose bisphosphate carboxylase (rbcL) et facteurs d'élongation.

W = base A ou U

S = base C ou G

Y = base U ou C.

MM = doublets mixtes, quand l'une des 2 bases est de type W, l'autre est de type S.

$$BC = \frac{WWC + SSU}{WWY + SSY}; \text{ (GOUY \& GAUTIER, 1982).}$$

SEQ. = séquences ; * séquences procaryotiques ; ** gènes nucléaires de la levure ; les autres entrées représentent les gènes chloroplastiques.

The values of the different indices show the degree of optimisation of the energy of the codon-anticodon bond for the 3 types of protein.

IV. Discussion et conclusion

A l'intérieur de la cellule végétale coexistent donc 2 stratégies de codage différentes, l'une nucléaire et l'autre chloroplastique. Cela semble souvent le cas chez les organelles et les virus, comme la mitochondrie humaine (GRANTHAM *et al.*, 1983) et les virus humains (GRANTHAM *et al.*, 1985) qui n'imitent pas leurs hôtes, mais ont leur propre stratégie de codage.

Parmi les pyrimidines, les gènes d'animaux préfèrent la base C en position III. Chez les plantes, cette propriété est partagée par les gènes nucléaires, mais d'une façon beaucoup moins nette quand les deux premières bases sont du type SS. Les gènes chloroplastiques ont une nette préférence pour la base U sur C à cause du pourcentage de G + C plus faible de leurs génomes. Cependant, comme nous l'avons démontré, le gène de la *rbcl* (gène le plus exprimé), montre une tendance à l'optimisation de l'énergie d'appariement par le choix C/U en position III.

Plusieurs données de la biologie moléculaire sont en faveur de l'hypothèse endosymbiotique de l'évolution. Notre étude montre que, pour les mêmes gènes, l'usage du code est différent pour les chloroplastes et leurs ancêtres hypothétiques. Parmi les protéines hautement exprimées, le gène de la *rbcl* est le seul qui suit le schéma bactérien présenté par GOUY & GAUTIER (1982).

L'alignement des séquences deux à deux a permis de tracer la phylogénie de la séquence du gène *rbcl* (SHINOZAKI *et al.*, 1983 ; BOUDRAA, en préparation). La protéine est fortement conservée, mais de nombreux changements silencieux dans les gènes ont abouti à un changement radical de l'usage du code. Nous manquons actuellement de données concernant le pool d'ARNt chez les plantes mais, par analogie avec *E. coli* et la levure, nous pensons que les changements de bases dégénérées chez des espèces très éloignées (par exemple cyanobactéries et chloroplastes des plantes supérieures) seraient en relation avec des changements dans la composition en ARNt.

Reçu le 16 décembre 1985.

Accepté le 16 octobre 1986.

Remerciements

Je remercie C. GAUTIER, M. GOUY, R. GRANTHAM, P. PERRIN et J.L. PRATO pour leurs aides et critiques.

Références bibliographiques

- BENZECRI J.P., 1973. L'analyse des correspondances. In : *L'analyse des données*, vol. 2, 619 pp., Dunod, Paris.
- GOUY M., GAUTIER C., 1982. Codon usage in bacteria : correlation with gene expressivity. *Nucl. Acids Res.*, **10**, 7055-7074.
- GOUY M., GAUTIER C., MILLERET F., 1985. System analysis and nucleic acid sequence banks. *Biochimie*, **67**, 433-436.

- GRANTHAM R., 1980. Workings of the genetic code. *Trends Biochem. Sci.*, **5**, 327-331.
- GRANTHAM R., GAUTIER C., 1980. Genetic distances from mRNA sequences. *Naturwiss.*, **67**, 93-94.
- GRANTHAM R., GAUTIER C., GOUY M., MERCIER R., PAVE A., 1980. Codon catalog usage and the genome hypothesis. *Nucl. Acids Res.*, **8**, r49-r62.
- GRANTHAM R., GAUTIER C., GOUY M., JACOBZONE M., MERCIER R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucl. Acids Res.*, **9**, r43-r74.
- GRANTHAM R., GOUY M., GAUTIER C., 1983. The genome as unit of selection : evidence from molecular biology. In : GEISSELER E., SCHELER W. (ed.), *Darwin today*, 95-100. Academic-Varley, Berlin.
- GRANTHAM R., GREENLAND T., LOUAIL S., MOUCHIROUD M., PRATO J.L., GOUY M., GAUTIER C., 1985. Molecular evolution of viruses as seen by nucleic acid sequence study. *Bull. Inst. Past.*, **83**, 95-148.
- GRANTHAM R., PERRIN P., MOUCHIROUD D., 1986. Patterns in codon usage of different kinds of species. *Oxford Surveys in Evolution. Biol.*, **3**, 48-81.
- GROSJEAN H., SANKOFF D., MIN JOU W., FIEERS W., CEDERGREN R.J., 1978. Bacteriophage MS2 RNA : a correlation between the stability of the codon-anticodon interaction and the choice of code words. *J. Mol. Evol.*, **12**, 113-119.
- IKEMURA T., 1981. Correlation between the abundance of *E. coli* transfer RNAs and the occurrence of the respective codon in its protein genes : a proposal for asynonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, **151**, 389-409.
- IKEMURA T., 1985. Codon usage and tRNA content in unicellular and multicellular organism., *Mol. Biol. Evol.*, **2** (1), 13-34.
- LYCETT G.W., DELAUNEY A.J., CROY R.D., 1983. Are plant genes different ? *Febs Lett.*, **153**, 43-46.
- MARGULIS L., 1975. Symbiotic theory of the origin of eukaryotic organelles : criteria for proof. *Symp. Soc. Exp. Biol.*, **29**, 277-289.
- SHARP P.M., ROYERS M.S., MCCONNELL D.J., 1985. Selection pressures on codon usage in complete genome of bacteriophage T7. *J. Mol. Evol.*, **21**, 150-160.
- SHINOZAKI K., DENO H., KATO A., SUGIURA M., 1983. Overlap and cotranscription of genes for the beta and epsilon subunits of tobacco chloroplast ATPase. *Gene*, **24**, 147-155.
- SLIGHTOM J., 1983. Complete nucleotide sequence of French bean storage protein gene. *Proc. Natl. Acad. Sci., USA*, **80**, 1897-1901.
- SUBRAMAIAAN A.R., STEINMETZ A., BOGORAD L., 1983. Maize chloroplast DNA encodes a protein sequence homologous to the bacterial ribosome assembly protein S4. *Nucl. Acids Res.*, **11**, 5277-5286.
- TAKAIWA F., SUGIURA M., 1982. Nucleotide sequence of the 16S-23S spacer region in a rRNA gene cluster from tobacco chloroplast DNA. *Nucl. Acids Res.*, **10**, 2665-2676.