Organisation et évolution du génome des Angiospermes

by

Nicolas Carels

ISBN: 1-58112-112-1

DISSERTATION.COM

USA • 2000

Organisation et évolution du génome des Angiospermes

Dissertation.com
USA • 2000

ISBN: 1-58112-112-1

www.dissertation.com/library/1121121a.htm

Laboratoire de
Génétique Moléculaire
Institut Jacques Monod
Université de Paris VII – CNRS
75005 Paris

Laboratorio di
Evoluzione Molecolare
Stazione Zoologica Anton Dohrn
80121 Napoli
Italie

Thèse de Doctorat de l'Université Paris VI - Pierre et Marie Curie

UFR : Sciences de la Vie

Spécialité : Biologie, Diversité et Adaptation des Plantes Cultivées

Présentée par M. Nicolas CARELS

Pour obtenir les titres de Docteur de l'Université Paris VI et de Docteur en Biotechnologie de la Communauté Européenne (http://www.eurodoctor.it/)

Sujet de la thèse :

Organisation et évolution du génome des Angiospermes

Soutenance le 14 décembre 1999

Devant le jury composé de :

M. Jean-Luc	ROSSIGNOL	Président
M. Giorgio	BERNARDI	Directeur de Thèse
M. Marc	VAN MONTAGU	Rapporteur EDBT
M. William	MARTIN	Rapporteur EDBT
M. Francis	QUETIER	Rapporteur
M. Michel	DELSENY	Rapporteur
M. Gilbert	BOMPEIX	Examinateur

Résumé : Les relations génomiques dégagées dans ce travail font apparaître que l'ancêtre commun des graminées a subi une transition de la composition en bases qui se traduit par l'augmentation en GC dans tous les consitutants géniques et génomiques. Par opposition, cette transition n'a pas eu lieu dans la plupart des autres Monocots et Dicots. La transition *compositionnelle* renforce l'évidence de l'existence de deux classes de gènes chez les Angiospermes : les gènes pauvres et riches en GC. Les premiers sont caractérisés par la présence d'introns en plus grand nombre et de plus grande longueur que les derniers. Par ailleurs, les deux classes de gènes chez les graminées ont, sur base de l'étude des substitutions synonymes, des taux d'évolution différentes et sont associées à des plans de fonction différents.
La composition en bases est un facteur déterminant de la localisation des gènes dans le génome. Chez les graminées, la plupart des gènes sont concentrés dans un intervalle de composition en bases inférieur à 2%. Chez *Arabidopsis*, la composition en GC est plus élevée aux parties distales des chromosomes que dans leurs parties centrales. La composition en GC des gènes et des séquences codantes suit la même tendance. Le génome d'*Arabidopsis* peut être décomposé en 2 composantes principales, l'une pauvre en GC correspondant à la partie centrale des chromosomes, l'autre riche en GC, et quantitativement plus importante, correspondant aux parties distales des chromosomes.

Summary: The work presented here shows, by a detailed analysis of the genomic features of *Gramineae*, that their common ancestor did the experience of a compositional transition leading to the increase in GC level of all parts of the genome and of the genes. By contrast, this transition did not occur in a number of Dicots and Monocots. With regard to GC level of the genes, the compositional transition stresses the existence of two classes of genes in Angiosperms. The GC-poor genes are characterised by introns higher in average number and size compared to GC-rich genes. By analysing synonymous and non-synonymous substitutions, we showed that the evolution rates in the two classes of genes are different and correspond to different functional patterns.
The base composition is a determinant of gene location within the Angiosperm genomes. In *Gramineae*, most genes occur in genomic regions that have a narrow range of GC levels (≤ 2%). In *Arabidopsis*, the GC level of contigs is higher, on average, in the distal regions of chromosomes compared to their central regions. The GC levels of genes and coding sequences follow the same trend. The *Arabidopsis* genome is made up of two major components. One is GC-poor and corresponds to the central part of the chromosomes, the other, quantitatively higher, is in comparison GC-rich and corresponds to the distal regions of chromosomes.

Riassunto: Le caratteristiche del genoma di cui abbiamo discusso in questo lavoro mostrano che l'antenato comune delle *Gramineae* deve aver subito una transizione nella composizione in basi che ha fatto aumentare il livello di GC in tutti i costituenti genici e genomici. Al contrario, questa transizione non si riscontrava nella maggior parte delle Dicot. e Monocot. La transizione *composizionale* rinforza l'evidenza dell' esistenza di due classi di geni nelle Angiosperme. In media, i geni poveri in GC hanno introni più lunghi e in numero più grande comparato ai geni ricchi in GC. Peraltro, le due classi di geni nelle *Gramineae*, hanno, sulla base degli studi sulle sostituzioni sinonime, tassi di evoluzione diversi che corrispondono a gruppi funzionali diversi.
La composizione in basi è un fattore determinativo della localizzazione dei geni nel genoma. Nelle *Gramineae*, la maggior parte dei geni sono concentrati in un intervallo di composizione in basi inferiore a 2%. In *Arabidopsis*, il livello di GC è più alto nelle parte distale dei cromosomi che nella loro parte centrale. La composizione di GC dei geni e delle sequenze codificanti segue la stessa tendenza. Il genoma d'*Arabidopsis* può essere scomposto in due componenti principali: la componente povera in GC corrisponde alla parte centrale dei cromosomi; l'altra ricca in GC, e quantitativamente più importante, corrisponde alle parte distale dei cromosomi.

REMERCIEMENTS

Je tiens à remercier tout particulièrement le Dr. G. Bernardi pour l'accueil qu'il m'a réservé dans son laboratoire.

Tous mes remerciements vont aussi à l'AUPELF-UREF qui en m'octroyant une bourse d'1 an a orienté mon destin vers l'IJM à un moment où tout était possible.

Que le Dr. E.-M. Geigl soit remercié très chaleureusement pour les conseils techniques, toujours avisés, qu'elle a pu me donner dans certaines parties de ce travail.

Que M. O. Clay trouve ici l'expression de ma haute considération pour les conseils toujours désintéressés, mais adaptés et pour la générosité dont il a toujours fait preuve.

Que le Dr. K. Jabbari soit également chaudement remercié pour ses conseils et pour les discussions les plus diverses que nous avons pu avoir ensemble.

Que Mes. C. Plonquet et M. Brient soient toutes deux remerciées pour leur constant dévouement et leur bonne humeur.

Que le personnel du service technique de l'Institut et en particulier M. G. Lefèvre soient remerciés pour l'assistance qu'ils m'ont fournie dans la résolution des problèmes techniques liées à ce travail.

Que Me. V. René soit chaleureusement remerciée pour sa capacité à corriger efficacement ce manuscrit.

Je tiens, dans ces lignes, à remercier très particulièrement le Prof. D. de Vienne pour l'appui qu'il m'a prêté et surtout pour la gentillesse avec laquelle il m'a fourni des sondes adaptées à la localisation de gènes de maïs à un moment où vraiment il était capital pour moi d'en obtenir.

Je tiens à remercier tout particulièrement le personnel de la Stazione Zoologica Anton Dohrn dans son ensemble pour l'acceuil vraiment chaleureux qu'il ma reservé.

La FRASEMA via la personne de M. Ph. Carré fut notre pourvoyeur de graines de maïs. Sans difficulté, nous eûmes plusieurs kg de graines. Que M. Ph. Carré soit remercié très chaleureusement pour cette action gratuite et peut-être inoportunee dans l'emploi du temps d'un industriel.

Enfin, je souhaite remercier, dans ces lignes, tous ceux que je n'aurais pas nommé et qui m'ont aidé directement ou indirectement à réaliser ce travail.

Je parle? mais de quoi?
quel silence coud sur moi son suaire?
quel chemin où marcher?
je te le demande ô mouette
je te le demande ô mouette dérivant
dans le bleu de la mer...
qui prétend que je te questionnais?
qui a dit que je rêvais les vagues
et parlais à une mouette?
je n'y suis pour rien
je n'ai pas bougé
je n'ai soufflé mot...

Adonis, Mémoire du vent
Poèmes 1957-1990

Et pourtant, cher lecteur, viens voir ce que je vois!

TABLE DES MATIERES

II. RESULTATS et DISCUSSION

Préambule

Notre travail expérimental ayant les génomes des Angiospermes pour objet, nous les avons comparés, en tant qu'eukaryotes, à ceux des vertébrés pour lesquels un aperçu général est maintenant disponible.

Comme toute matière biologique et particulièrement génétique de cette portée, elle baigne implicitement dans le creuset de l'évolution. Dès lors, nous nous sommes attachés à nous servir de ses concepts pour dégager notre propos. Nous avons procédé de cette manière car, depuis Darwin, la recherche de mécanismes pouvant expliquer l'enchaînement des espèces au travers de leurs transformations morphologiques est toujours d'actualité. Il était, dès lors, impératif de présenter le théâtre où se joue notre pièce, (i.e., l'implication de l'organisation compositionnelle des eukaryotes dans les mécanismes de l'évolution). Aussi, avons-nous enchaîné Résultats à Introduction, directement, reportant Matériel et Méthodes dans Annexes.

Dans le développement de l'introduction, nous présentons les différentes entités qui participent à l'architecture des plans compositionnels. Nous montrerons également comment leur agencement peut participer à la notion de phénotype génomique et comment cette notion peut, à des niveaux taxinomiques supérieurs à l'espèce, permettre de classer les eukaryotes en groupes clairement identifiables dans leurs stratégies fonctionnelles.

But de la thèse

Nous avons cherché à montrer, dans cette étude, comment au cours de l'évolution, les génomes adoptent, en relation avec les contraintes auxquelles ils sont soumis, des stratégies dont les effets sont perceptibles dans leur organisation.

Nous nous sommes intéressés à la situation des Angiospermes et plus particulièrement des graminées en relation avec l'état des connaissances acquises chez les vertébrés.

Dans sa matière expérimentale, notre travail s'est employé à contribuer au dégagement et à la comparaison des similitudes, quant à l'organisation des plans de composition en bases, entre les eukaryotes végétaux de types Angiospermes et animaux de type vertébrés. Le but de cette démarche étant de présenter nos observations, relatives aux Angiospermes, dans le cadre du système de référence que constitue désormais les vertébrés. En effet, nous avons maintenant un ensemble important de connaissances quant à l'organisation du génome de ces derniers par rapport au critère de composition en bases de l'ADN.

I. INTRODUCTION

I.1 LES EUKARYOTES ET LEUR MILIEU

La conquête du milieu terrestre par les eukaryotes débute au Silurien (ère Primaire), il y a environ 415 millions d'années (MA) avec les premières plantes (*Lycopsida*) dont les représentants actuels sont les lycopodes et les sélaginelles. Ce n'est qu'à partir de -385 MA (Dévonien) que de nouveaux groupes végétaux se différencient. Il s'agit des fougères et des prêles. C'est à cette époque que l'on trouve les premières traces d'arthropodes terrestres (scorpions). Les vertébrés conquièrent l'élément terrestre aux environs de -360 MA avec les ancêtres des dipneustes, suivis par les amphibiens et les reptiles vers -340 MA (Carbonifère). Les premières Gymnospermes apparaissent vers -300 MA tandis que les premières Angiospermes vers -110 MA, c'est-à-dire au Crétacé inférieur (Fig. 1).

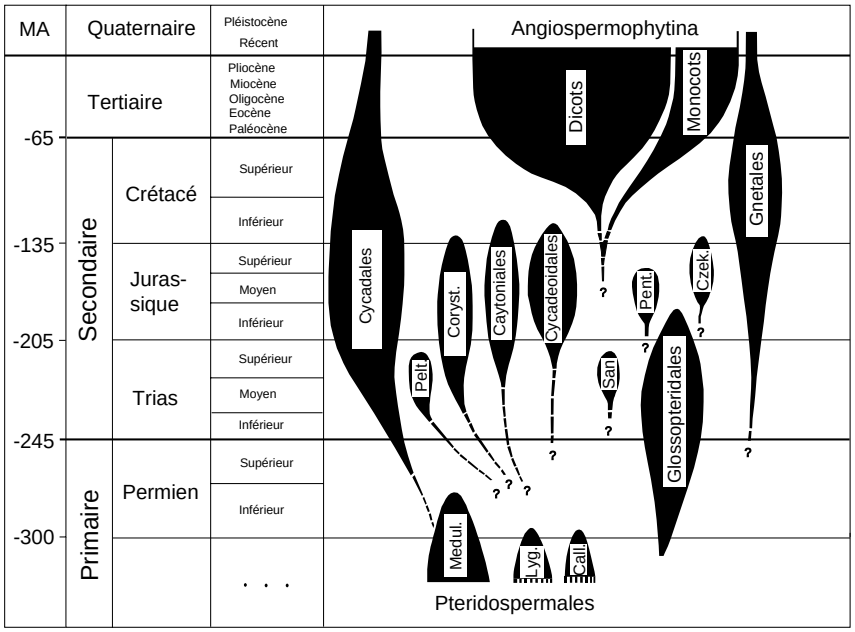

Fig. 1 : Résumé des relations évolutives des Angiospermes, Cycadophytes et Glossopterides et de leurs origines dans l'échelle des temps géologiques (modifié de Stewart et Rothwell, 1993). Medul. = *Medullosaceae* ; Lyg. = *Lyginopteridaceae* ; Call. = *Callistophytaceae* ; Pen. = Pentoxylales ; Czek. = Czekanowskiales ; San. = *Sanmiguelia* ; Coryst. = *Corystospermaceae* ; Pelt. = *Peltaspermaceae*.

Entre -295 et -200 MA se sont succédés d'innombrables groupes de reptiles mammaliens, de plus en plus semblables aux mammifères. Vers -200 MA plusieurs groupes de reptiles mammaliens, herbivores et carnivores, avaient acquis toutes les caractéristiques des mammifères. On situe, donc, l'apparition des premiers mammifères en coïncidence avec le début du

Jurassique, peu après (50 MA) la première crise d'extinction qui marque la fin de l'ère Primaire. Ce n'est que 135 MA plus tard que les mammifères pourront connaître leur véritable essor à l'occasion de la seconde crise planétaire qui marque la fin de l'ère Secondaire avec l'extinction des dinosaures.

Les mammifères placentaires, depuis l'aube de leur radiation, présentent un profil similaire. Des fragments épars de mammifères primitifs sont connus dès le début du Crétacé inférieur, mais leurs fossiles ne deviennent abondants qu'à partir du Crétacé supérieur et du début du Cénozoïque (Tertiaire). A partir de cette époque, les mammifères sont restés à peu près constants durant les 57 MA qui ont suivi (cf. Carroll, 1997). Il en va de même des oiseaux qui sont apparus, certes, un peu plus tard que les mammifères et indépendamment à partir d'autres lignées de reptiles (-150 MA), c'est-à-dire vers la fin du Jurassique.

Le Crétacé apparaît comme une véritable période charnière puisque c'est à cette époque que les oiseaux, les mammifères et les Angiospermes connaissent, tous, une importante vague de radiation.

Le profil d'évolution des plantes vasculaires est similaire à celui mentionné pour les oiseaux et les mammifères. En effet, elles sont représentées par un petit nombre de groupes majeurs qui apparaissent soudainement dans les archives fossiles et persistent sans changements fondamentaux durant les centaines de MA qui suivent.

Comme nous le verrons dans la suite de cet exposé, ces alternances dans l'évolution où se succèdent des périodes explosives d'inventions courtes et des périodes de stagnations longues (Gould et Eldredge, 1993) sont perceptibles au niveau de l'ADN (Bernardi, 1993b).

I.2 L'ORGANISATION DU GENOME DES EUKARYOTES

I.2.1 Le génome et l'ADN répété

Les séquences d'ADN répété qui n'ont pas ou peu de fonctions connues forment un large groupe. On y rencontre des segments courts d'ADN répété non-codant tels que ceux trouvés à l'intérieur des introns, aux alentours des gènes ou comme entretoises de ces derniers. On y trouve aussi des séquences en tandem hautement répétées et finalement des séquences moyennement répétées groupées ou dispersées et qui constituent jusqu'à 30 % de l'ADN génomique total (Dootlittle et Sapienza, 1980). L'ADN répété est généralement non transcrit (à l'exception des rétrotransposons) et localisé dans l'hétérochromatine. Toutes ces séquences répétées sont regroupées par

Orgel et Crick (1980) sous la définition d'ADN égoïste (*selfish DNA*). C'est-à-dire de l'ADN qui n'apporte pas de contribution spécifique au phénotype et qui se répand dans le génome par formation de copies additionnelles.

Dootlittle et Sapienza, (1980) ont suggéré que l'ADN moyennement répété forme une fraction trop large de la plupart des génomes eukaryotes pour pouvoir être conservé par sélection darwinienne, et que, par conséquent, il est vraisemblable que son intégrité a été conservée par transposition. En conséquence, l'ADN moyennement répété doit être considéré comme constitué d'éléments transposables ou de reliques plus ou moins dégénérées de ceux-ci ayant perdu leur activité transposable. Dootlittle et Sapienza, (1980) admettent que, s'il en est ainsi, les changements dans l'abondance et la gamme de divergence des familles de séquences d'ADN moyennement répétées ont été le résultat d'une sélection qui a agi au niveau du génotype plutôt qu'au niveau du phénotype. Une fois établies à l'intérieur du génome, les séquences d'ADN inutiles, qui auraient été perpétuées par activité transposable ou par tout autre mécanisme de *turnover* génomique, seraient difficiles à éliminer et, donc, vouées à une durée de vie prolongée.

La mobilisation des éléments transposables au travers de l'acte d'hybridation dans *D. melanogaster* peut entraîner une activité mutagène. De même un stress génomique, tels que ceux qui sont instaurés par les cycles de bris-fusion-pontage des chromosomes dans le maïs, peuvent aussi provoquer un regain d'activité d'éléments transposables jusque là en léthargie, et provoquer un pic d'activité mutagène (McClintock, 1984).

Un autre mécanisme de genèse de variabilité à l'intérieur des membres d'une famille de séquences est également concevable. Certains membres peuvent subir une contrainte sélective du fait de leur composition en base alors que d'autres sont libres d'évoluer. Seule une petite fraction des séquences répétées des grands génomes eukaryotes s'est vue attribuer une fonction biologique. Par ailleurs, une quantité de données comparatives de plus en plus importante implique que la majeure partie de l'ADN des eukaryotes, incluant la majorité des séquences répétées, doit être considérée comme ADN dérivé (*secondary DNA*). La réitération d'un ou de quelques membres d'une famille de séquences répétées divergentes par amplification dérivée (*secondary amplification*) est une caractéristique des grands génomes végétaux. Ce mécanisme conduit à une vitesse d'amplification élevée, à une composition hétérogène des familles de séquences répétées, et à un *turnover* rapide de celles-ci (Preisler et Thompson, 1981). Bien que l'élimination de rétroélément du type non-LTR n'ait pas été décrite, les séquences répétées pourraient aussi, dans certaines circonstances, être impliquées dans des

phénomènes de contraction génomique, par délétion entre séquences proches du point de vue homologie, entraînant la disparition des fragments d'ADN insérés entre elles (Smyth, 1991). Ainsi, Voytas *et al*. (1990) ont montré que, chez *Arabidopsis,* des crossing over sont susceptibles de se produire dans les familles de rétroéléments à LTR, entraînant leur réduction jusqu'à leur transformation en *solo*. Chez d'autres espèces, les mécanismes de recombinaison sont probablement à la base d'insertions de séquences (*Cin*1 chez le maïs - Shepherd *et al*., 1984; *del*1 chez *Lilium henryi* - Sentry et Smyth, 1985, 1989).

Les séquences en copies uniques pourraient être des reliques de très vieilles familles de séquences répétées qui auraient divergé à un point tel qu'elles n'ont plus d'homologie significative. Les événements d'amplification recyclent certaines de ces séquences aussi bien que certains membres de familles de séquences encore reconnaissables comme étant répétées. Dans le cadre de ces considérations, un génome peut être considéré grand lorsqu'il dépasse 10^9 pb (cf. Preisler et Thompson, 1981). Certaines céréales et le pois sont des plantes à grand génome. Ces deux types de plantes ont des séquences répétées à évolution rapide, mais diffèrent considérablement par la composition moyenne en bases de leurs génomes.

I.2.2 L'ADN égoïste

Dans une perspective sélectionniste, on s'attend à la disparition des séquences qui ne codent pas pour des fonctions cellulaires. Si des stratégies d'évitement des délétions spécifiques qui visent à faire disparaître ces séquences existent, elles pourraient subsister. La sélection opérant à l'intérieur du génome, indépendamment du phénotype de l'individu ou de l'adaptation de la population, va favoriser l'installation et le maintien de séquences d'ADN qui adoptent ces stratégies (Doolittle, 1980 ; Orgel *et al*., 1980; Sapienza et Doolittle, 1981 ; Doolittle, 1981). Cet argument et son corollaire selon lequel il n'y a pas à chercher d'autres explications pour justifier l'existence de l'ADN génomique dont ce type de comportement assure son maintien, peuvent paraître inattaquables. Cependant, ils sont stériles à partir du moment où ces stratégies ne sont pas disponibles ou si aucune séquence d'ADN génomique connue ne s'avère l'adopter (Doolittle, 1980).

La notion d'ADN égoïste (*selfish*) ne doit pas être assimilée à celle d'ADN *junk* qui est beaucoup plus réductrice (Orgel *et al*., 1980) et n'apporte finalement aucune proposition positive sur les implications de l'ADN ainsi désigné si ce n'est de révéler notre ignorance en regard de son éventuelle

fonction, qu'elle soit structurale, évolutive, ou autre. A ce titre, il est clair qu'un contenu élevé en séquences moyennement et hautement répétées doit avoir un effet régional sur la composition chromosomique. Le type même de séquences impliquées dans l'ADN répété, en conditionnant le mode de son évolution (i.e., duplication, crossing-over inégaux, transposition, évolution concertée) doit moduler cet effet de manière différente. Les exigences fonctionnelles en relation avec l'arrangement spatial des chromosomes dans le noyau (Thuriaux, 1977), la promotion de la recombinaison en relation avec la variabilité phénotypique, l'organisation centromérique, le temps de réplication, la régulation génique, constituent des éléments en équilibre avec les contraintes sélectives. La dissémination des familles de transposons et les mutations qui s'en suivent (Bohr et Wassermann, 1988), ainsi que celle des rétrotransposons et autres séquences répétées qui conduisent aux variations de taille du génome, sont régies par cet équilibre (Biradar *et al*., 1994). Comme on le sait, la taille du génome n'est pas corrélée à sa complexité. Cette observation est patente chez *Arabidopsis thaliana* (10^8 pb) (Pruitt, et Meyerowitz, 1986) et *Lilium henryi* (10^{11} pb) (Bennett et Smith, 1976), deux plantes à fleurs qui doivent contenir approximativement le même nombre de gènes, mais qui diffèrent par la quantité d'ADN de leur génome nucléaire d'un facteur 1.000. Cette variation de la taille du génome d'une espèce à l'autre est principalement due à l'expansion des parties non-codantes. Chez l'homme, approximativement 1% du génome est alloué à l'encodage des protéines. La situation est approximativement la même chez le maïs (Hake et Walbot, 1980) pour une taille génomique correspondant à celle de l'homme (2,5-3 10^9 bp). Au sein de la famille des *Poaceae*, la taille du génome des céréales illustre cette variabilité : *Hordeum vulgare* = 5 10^9 pb, *Triticum aestivum* = 1.7 10^{10} pb, *Oryza sativa* = 4 10^8 pb et *Sorghum bicolor* = 8 10^8 pb.

Les caractéristiques qui affectent la taille du génome peuvent aussi affecter la compétitivité interespèces dans leur milieu. Des variations importantes du contenu en ADN ont été observées chez le maïs (Biradar *et al*., 1994). Le contenu en ADN a souvent été corrélé avec des paramètres cellulaires tels que le volume nucléaire, le volume cellulaire, le cycle mitotique, la durée de la méiose et le nombre de chloroplastes par cellules de garde (cf. Biradar *et al*., 1994). Ces effets (*nucleotype effects*) ne sont pas limités au niveau cellulaire (Bennett, 1972). Ils déterminent aussi la vitesse de développement, la croissance de la plante entière, le poids des graines, etc. Chez les plantes supérieures, ces effets sont additifs à chaque cycle cellulaire, de telle sorte qu'ils se répercutent à tous les niveaux de la plante (Bennett, 1987, 1996). La taille du génome a été corrélée négativement avec la latitude

(Laurie et Bennett, 1985 ; Rayburn *et al.*, 1985). La quantité d'ADN est corrélée avec des variations d'altitude de 1.500 m dans les populations de maïs observées par Rayburn (1990) et Rayburn et Auger (1990). Bullock et Rayburn (1991) ont observé que la longueur de la saison de croissance est corrélée avec la taille du génome. Tous les paramètres de croissance et de production sont corrélés négativement avec la quantité d'ADN nucléaire. Ces corrélations démontrent à quel point les variations de taille du génome peuvent influencer les paramètres agronomiques et la conduite de la sélection. La corrélation négative entre la quantité d'ADN et la croissance de la plante, observée chez le maïs, n'est pas une règle générale. Chez *Allium*, *Vicia*, *Crepis* et certains légumes et même graminées, une relation positive à été observée entre ces deux facteurs (Bennett, 1972 ; Jones et Brown, 1976 ; Mowforth, 1985). La corrélation entre taille du génome et maturité a aussi été observée chez le soja (Graham *et al.*, 1994). Les corrélations ne sont pas, *a priori*, identiques ou même semblables dans toutes les espèces (Cavalini et Natali, 1991) en raison des effets différents que la taille du génome peut avoir sur différents caractères et dans différentes espèces (Bennett, 1987), sans compter les sensibilités particulières aux conditions de l'environnement.

Dans les hybrides dérivés d'un croisement entre *Nicotiana tabacum* et *Nicotiana otophora*, d'énormes chromosomes (méga-chromosomes) sont générés par amplification de segments intra-chromosomiques dans les noyaux somatiques. L'amplification intéresse principalement l'hétérochromatine. Par contre, les méga-chromosomes générés dans les hybrides de *N. tabacum* et de *Nicotiana plumbaginifolia* sont dus à l'amplification de segments d'euchromatine. Ces amplifications génome-dépendantes résultant probablement d'une nécessité de stabilisation de parité entre chromosomes homéologues, doivent avoir des répercussions sur l'équilibre physiologique de la plante (Hutchinson *et al.*, 1980).

En résumé, si l'accumulation de l'hétérochromatine n'est peut être pas un paramètre indispensable à l'épanouissement de l'individu, il ne peut cependant pas être considéré dans un sens péjoratif (*junk*), parce que cela revient à "une vision inadéquate de la relation entre fonction et nécessité" (Zuckerkandl et Hennig, 1995). La quantité minimum d'hétérochromatine est la quantité requise pour assurer la survie d'une population sous certaines conditions de l'environnement. La quantité optimale donnera à la population ses meilleures chances de survie en présence de nouveaux défis de l'environnement tels que la mise en compétition avec d'autres populations (Zuckerkandl et Hennig, 1995). Si l'information codante apparaît relativement constante dans les génomes eukaryotes, même si elle est

soumise à des changements qualitatifs, l'ADN non-codant doit jouer un rôle essentiel dans l'évolution du génome. Un "contexte génomique" favorable peut être l'instrument du succès dans la compétition entre 2 populations, et donc être sélectionné. En l'absence de compétition, la population dépourvue de ce contexte génomique peut néanmoins être "reproductivement suffisante".

L'ADN répété

Les éléments transposables de la levure et de la drosophile constituent la plus grande partie de l'ADN moyennement répété de ces organismes, à l'exclusion des familles multigéniques et de l'ADNr. Ils sont généralement de plusieurs milliers de paires de bases. Chez d'autres eukaryotes, on rencontre aussi, en parallèle, des séquences plus courtes avec une périodicité d'insertion (*interspersion*) plus élevée. L'exemple le plus connu est celui des *Alu*, longs de 300 pb et présents jusqu'à 300.000 exemplaires, chez l'homme. Les *Alu*I, membres de cette famille, représentent plus de la moitié de l'ADN humain moyennement répété (Rubin *et al.*, 1980 ; Jelinek *et al.*, 1980).

La plupart des études sur l'évolution de l'ADN répété en tandem (*highly repetitive* ou satellite) font appel au modèle de crossing-over inégaux pour expliquer leur amplification (Smith, 1976, 1978 ; John et Miklos, 1977 ; Walker, 1978 ; Bostock, 1980 ; Brutlag, 1980). Les répétitions en tandem, une fois générées, sont capables de se perpétuer bien que leur longueur puisse varier par le même mécanisme que celui qui les crée. De par les mécanismes qui leur sont inhérents, les crossing-over inégaux assurent le maintien de l'homogénéité des séquences. Il n'y a donc aucun besoin de faire appel à la notion de sélection pour expliquer l'homogénéité de séquence de leur produit d'amplification. La notion de sélection n'est applicable, dans ce contexte, que pour expliquer la représentation relative de certaines familles à partir du moment où leur amplification est favorisée par les mécanismes de crossing-over. Dans ce cas, on peut considérer ces séquences comme de l'ADN égoïste. Cependant, des évidences relatives à ces comportements sont difficiles à réunir (Dover, 1980 ; Christie et Skinner, 1980). Par ailleurs, on a pu montrer, dans certains cas, que les séquences répétées sont reconnues par des protéines (*binding protéines*) susceptibles de leur conférer une fonction, et donc de donner prise sur ces séquences à la sélection (Brutlag, 1980). Dans ces conditions, on peut difficilement leur attribuer la notion d'ADN égoïste.

Les éléments transposables

La transposition qui produit des copies fidèles et elles-mêmes transposables partout dans le génome, tout en préservant l'intégrité de la séquence originale, est un cas de stratégie spécifique adoptée par une séquence pour assurer sa pérennité dans le génome. Les éléments qui l'adoptent ne peuvent être éliminés du génome que par un événement hautement improbable de délétion simultanée intéressant toutes leurs copies, l'original y compris.

Tous les transposons bactériens ont adopté cette stratégie (Shapiro, 1979, ; Arthur et Sheratt, 1979 ; Kopecko, 1980 ; Calos et Miller, 1980 ; Harshey et Bukhari, 1981).

Les mécanismes d'excision et la réinsertion sans étape de duplication ne peuvent pas être considérés comme des stratégies spécifiques pour le maintien de la séquence, bien qu'ils correspondent à la définition de la transposition. Dans ce cas, en effet, l'amplification de la séquence ne peut survenir que par des phénomènes de recombinaison au hasard. Si l'amplification est dépendante de la réplication, ses produits doivent avoir le même âge quel que soit leur nombre de transposition. Ils auront donc le même niveau d'homologie et ne peuvent pas être divisibles en sous-familles à l'intérieur d'une superfamille (Doolittle et Sapienza, 1980).

Les *controlling elements* chez le maïs, par exemple, ne peuvent être considérés comme égoïstes (Peterson, 1981 ; Burr et Burr, 1981). Ce sont des éléments qui provoquent une variabilité génétique spécifique, par insertion intragénique induite par leur mouvement au sein d'un organisme isolé durant son cycle de vie , ou dans une population, en quelques générations. Par le fait de leur insertion intragénique, ces éléments peuvent très bien être maintenus par sélection (Doolittle et Sapienza, 1980).

Les rétrotransposons du type Ty-1-*like elements* correspondent à la notion d'ADN égoïste du fait de leur division en sous-familles à l'intérieur d'une superfamille (Williamson *et al.*, 1981 ; Kingsman *et al.*, 1981). Des éléments de ce type sont connus chez tous les eukaryotes et chez les graminées en particulier. Le traitement mathématique de ce type d'éléments a été étudié par Ohta et Kimura (1981). C'est aux éléments transposables que la notion d'ADN égoïste se prête le mieux. Néanmoins, ce genre de développement peut aussi être appliqué à d'autres catégories de séquences des génomes eukaryotes.

Les séquences uniques non-codantes

La fonction de beaucoup de séquences uniques non-codantes d'eukaryotes, espaceurs inclus (Federoff, 1979), reste à déterminer. On ne peut pas

considérer ces séquences comme égoïstes puisque leurs séquences ne sont généralement pas bien conservées et qu'il n'y a pas de protection évidente contre leur élimination. Certaines d'entre elles proviennent de duplications de gènes ou d'événements de transposition et dérive subséquente (Proudfoot et Maniatis, 1980).

I.2.3 Les isochores et le génome des vertébrés

Organisation compositionnelle du génome humain

Le terme "génome" a été créé, il y a trois quarts de siècles, par Hans Winkler (botaniste), pour désigner le jeu de chromosomes haploïdes des eukaryotes. Généralement, on se limite actuellement à la définition purement opérationnelle du génome, c'est-à-dire la somme des gènes et des séquences intergéniques. On peut toutefois penser que le génome est plus que la somme de ses parties. Ceci implique, comme nous l'avons vu plus haut, l'existence d'interactions structurelles et fonctionnelles entre la minorité des séquences codantes et la majorité des séquences non-codantes. Les propriétés compositionnelles du génome des vertébrés ont permis d'établir de telles relations. Au nombre de ces propriétés, on distingue : l'organisation du génome en mosaïque d'isochores, la corrélation compositionnelle entre fragments (ou molécules) d'ADN et séquences codantes, les corrélations compositionnelles entre séquences codantes et non-codantes, et, surtout, la distribution des gènes et les propriétés fonctionnelles qui lui sont associées (Bernardi, 1995).

Les isochores.

Notre laboratoire a montré que l'ADN est, chez les vertébrés, organisé en domaines (0,2-1,3 Mb, voir plus chez l'homme - Bernardi, 1989 ; Gardiner, 1990 ; De Sario, 1996), de composition en bases homogène, alternés au sein du génome (Macaya *et al.*, 1976; Bettecken *et al.*, 1992). L'alternance (Fig. 2) de ces domaines, donne lieu à une mosaïque compositionnelle (cf. Bernardi *et al.*, 1985; Bernardi, 1989, 1993a,b, 1995 pour une revue).

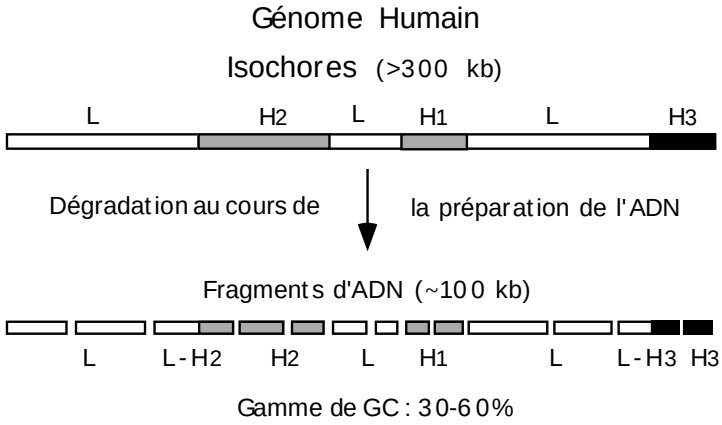

Fig. 2 : Schéma de l'organisation en isochores du génome humain. Le génome humain, représentatif du génome des mammifères, est une mosaïque de segments d'ADN (>300 kb), homogènes en composition et pouvant être divisés en familles pauvres en GC (L), riches en GC (H1 et H2), et très riches en GC (H3). Les isochores sont dégradées au cours de la préparation de l'ADN en fragments de l'ordre de 100kb. La gamme de GC couverte par les isochores du génome humain va de 30 à 60 % (cf. Bernardi, 1995).

Les caractéristiques de composition homogène de ces domaines leurs ont valu le nom d'*isochores* (Cuny *et al*., 1981). Elles furent découvertes, il y a 25 ans, chez les bovins (Filipski *et al*., 1973). Chez l'homme, ces isochores peuvent être regroupées en un petit nombre de familles compositionnelles couvrant une gamme de GC de 30 à 60% (Bernardi, 1985 ; Bernardi, 1995), chaque famille s'étalant sur environ 2% de GC.

Pendant la procédure d'extraction, l'ADN se brise en molécules de l'ordre de 50 à 200 kb sous l'effet des actions mécaniques et enzymatiques. Les fragments ainsi générés ont une taille inférieure, de plusieurs ordres de grandeur, aux isochores, de telle sorte que la plupart des fragments peuvent être attribués, sans ambiguïté, à telle ou telle famille d'isochores. Seuls les fragments (minoritaires) qui se trouvent à cheval sur une jonction entre deux isochores posent un problème d'attribution.

En pratique, pour réaliser le tri compositionnel des fragments, on procède, par ultracentrifugation, à l'équilibre (isopycnique) de l'ADN en Cs_2SO_4 en présence d'un ligand. Ce ligand, un acétate de mercure dont le nom abrégé est le BAMD (Zipper *et al.*, 1982), se fixe sur des séquences riches en AT, si bien que plus un fragment sera riche en AT, plus le complexe qu'il forme avec le BAMD aura une

densité de flottaison élevée, et plus il sera situé vers le fond du tube après centrifugation à l'équilibre. Le fractionnement compositionnel est réalisé par récolte de fractions de volume constant, au moyen d'une aiguille qui pénètre dans le gradient. Ce fractionnement correspond à un tri des fragments d'ADN sur base de leur composition en bases.

Les isochores du génome humain peuvent être divisées en deux groupes : le premier, représentant les deux tiers du génome, est composé par la famille d'isochores L, pauvres en GC (elle-même subdivisée en deux familles L1 et L2) ; le second, couvrant le tiers restant, comprend trois familles riches en GC (H1, H2 et H3) (Fig. 3).

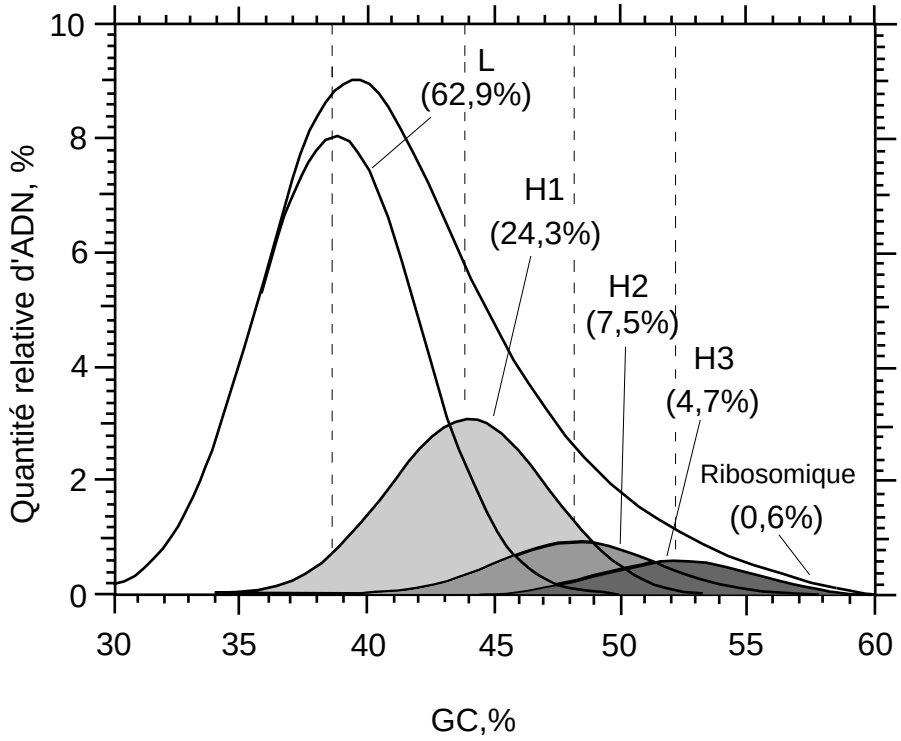

Fig. 3 : Décomposition du profil de l'ADN humain en CsCl, la contribution relative de chaque composante est indiquée par les nombres entre parenthèses, l'axe des ordonnées correspond à la quantité relative d'ADN (tiré de Zoubak *et al.*, 1996).

Plans compositionnels

Les distributions compositionnelles des longs fragments d'ADN (~100 Kb), des exons, des bases en troisième position des codons, et des introns, constituent des plans compositionnels caractéristiques (Bernardi, 1985 ; Bernardi, 1989 ; Bernardi, 1993), dont l'ensemble constitue le *phénotype génomique* (Bernardi, 1987), car ils diffèrent

non seulement entre les vertébrés à sang froid et à sang chaud, mais également entre mammifères et oiseaux, et entre plantes.

Corrélations compositionnelles

La composition en bases des exons et des isochores n'est pas indépendante (Bernardi, 1985 ; Bernardi, 1986). Il existe, en effet, des corrélations entre ces différents éléments (Fig. 4) bien que les séquences codantes ne représentent que 1 à 3% du génome.

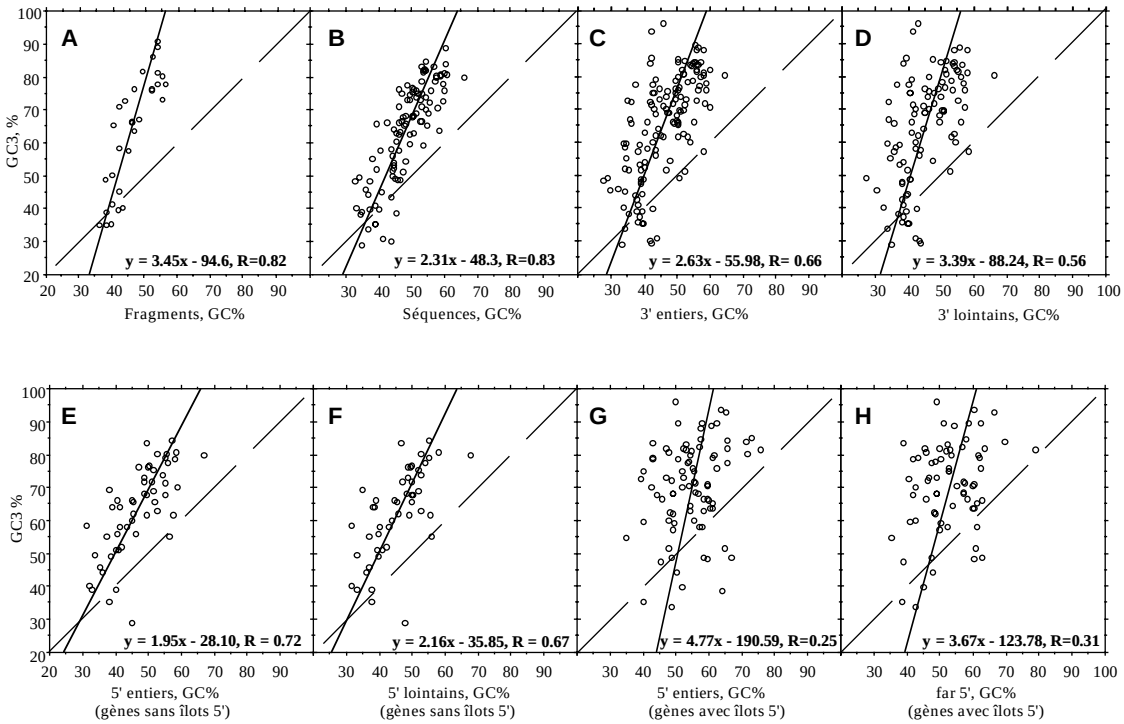

Fig. 4 : Corrélation entre le niveau de GC_3 des gènes humains et de la composition (A) des molécules d'ADN (fragments) des fractions compositionnelles dans lesquelles les gènes sont localisés ; (B) des longues séquences génomiques (≥ 10kb) contenant ces gènes ; (C) des séquences flanquantes en 3' ; (D) des séquences flanquantes en 3' à plus de 500 pb du codon stop ; (E) des séquences flanquantes en 5' des gènes non associés aux îlots CpG ; (F) des séquences flanquantes en 5' à plus de 500 pb du codon initiateur des gènes non associés aux îlots CpG ; (G) des séquences flanquantes en 5' des gènes associés aux îlots CpG ; et (H) des séquences flanquantes en 5' à plus de 500 pb du codon initiateur des gènes associés aux îlots CpG. Les équations de la régression orthogonale sont données pour chaque diagramme (cf. Clay *et al.,* 1996).

De plus, il y a une corrélation universelle (D'Onofrio, 1992) entre les niveaux de GC des différentes positions des codons ; par exemple, entre le niveau de GC en troisième position (GC_3) et celui des première et deuxième positions. L'ensemble de ces corrélations représente un *code génomique* (Bernardi, 1990 ; Bernardi, 1993)

dont l'expression dans le phénotype génomique est en relation avec les contraintes qui opèrent sur le génome.

Isochores et bandes chromosomiques

Les chromosomes métaphasiques humains présentent divers types de bandes selon le traitement auquel ils sont soumis (colorants, température, enzymes protéolytiques ou DNAses). Les bandes G(iemsa) positives représentent environ 50% des bandes chromosomiques, l'autre moitié étant constituée de bandes G(iemsa) négatives. Ces dernières coïncident avec les bandes R(everse), obtenues par coloration après dénaturation thermique. Quant aux bandes T(élomériques), identifiées par Dutrillaux, en 1973, elles forment le sous-ensemble des bandes R, à savoir les bandes plus résistantes à la dénaturation, et sont localisées, en grande majorité, au niveau des télomères.

Il est apparu que dans le génome humain, les isochores des familles H2 et H3 sont localisées dans les bandes T (Saccone *et al*., 1992, 1993), alors que les bandes R' (c'est-à-dire les bandes R à l'exclusion des bandes T) comprennent aussi bien des isochores riches en GC (famille H1) que pauvres en GC. Enfin, les bandes chromosomiques G sont formées, presque exclusivement, d'isochores pauvres en GC. Le fait que la majorité des gènes soit localisée au niveau des bandes T présente vraisemblablement un avantage fonctionnel (les télomères étant associés à la matrice et à l'enveloppe du noyau) quant à l'exportation des produits de transcription dans le cytoplasme où ils sont traduits par les ribosomes.

La concentration des gènes dans les isochores les plus riches en GC, qui représentent environ 5% du génome, est au moins 17 fois plus élevée que dans les isochores pauvres en GC qui représentent plus de 60% du génome. Ces isochores à très haute concentration en gènes présentent un intérêt tout particulier pour les raisons suivantes : (i) elles comprennent la très grande majorité des îlots CpG, qui sont des séquences contenant des signaux de régulation, et qui correspondent à une structure particulière de la chromatine (absence de l'histone H1, acétylation des histones H3 et H4, rareté des nucléosomes - Tazi et Bird, 1990) ; (ii) elles correspondent aux bandes T, qui sont des bandes chromosomiques résistantes à la dénaturation thermique (Dutrillaux, 1973), très riches en GC (Ambros, 1987). Ces bandes sont localisées dans une vingtaine de télomères et dans quelques régions internes des chromosomes métaphasiques (Saccone *et al*., 1992).