"Le Sélectionneur Français" 1996 (47), **65-69**

LE POINT SUR LE SEQUENÇAGE DU GENOME NUCLEAIRE D'*ARABIDOPSIS THALIANA.*

Alain LECHARNY et Martin KREIS
CNRS-Université
Laboratoire de Biologie du Développement des Plantes
Institut de Biotechnologie des Plantes
Bât. 630
Université de Paris-Sud
91405 ORSAY Cedex

1 - INTRODUCTION

En septembre 1993 a débuté le séquençage systématique du génome nucléaire d'une plante supérieure, *Arabidopsis thaliana*. L'*Arabidopsis* (100 Mpb et 5 paires de chromosomes) prenait ainsi toute sa place parmi les espèces modèles pour l'étude des génomes eucaryotes, après la levure *Saccharomyces cerevisiae* (dont le génome, 14,4 Mpb et 16 chromosomes, est séquencé en entier) (WILLIAMS N. 1995) et à côté du nématode *Caenorhabditis elegans* (100 Mpb et 5 paires de chromosomes) (SULSTON J. *et al.* 1992) et du poisson *Fugu rubipes* (400 Mpb réparties en 22 à 24 paires de chromosomes) (BRENNER S. *et al.* 1993). Dans sa première phase, le projet de séquençage du génome d'*Arabidopsis* contient deux volets complémentaires, le séquençage d'un fragment du chromosome 4 et la production d'étiquettes de RNA messagers.

Pourquoi concentrer à l'heure actuelle une certaine quantité de nos moyens dans ce domaine sur *A. thaliana* comme modèle des plantes dicotylédones ?

Au-delà des considérations de coût et d'évolution des techniques qui seront volontairement ignorées, quatre raisons principales peuvent être invoquées. La première, c'est la connaissance d'un génome végétal dans ses différents niveaux de structure et dans leurs relations (DEAN, C. et SCHMIDT, R., 1995). La deuxième, c'est l'existence combinée au sein d'un groupe comme les dicotylédones de la similitude et de la synténie des séquences codant pour un même produit. Il est établi que, même chez des organismes phylogénétiquement éloignés, nombre de gènes présentent de fortes similitudes et qu'il est donc possible de trouver assez rapidement le gène homologue à un gène d'*A. thaliana* chez une autre dicotylédone. Par ailleurs, la synténie qui est la conservation des relations de proximité entre les gènes dans des blocs au sein des chromosomes facilitera la localisation chromosomique d'un gène chez d'autres espèces. Ceci est déjà vérifié pour *Brassica napus*. La troisième raison, c'est que la différence de taille entre les génomes ne tient pas dans le nombre de gènes mais dans le nombre de séquences répétées "dispersant" les gènes. Ces séquences répétées occupent 200

fois plus de place dans le génome de base des blés (taille du génome: 16 500 Mpb et 7 chromosomes) que dans celui d'*A. thaliana* et 70 fois pour le pois (7 200 Mpb). Très souvent, la même organisation des gènes se retrouve dans les différentes plantes : le nombre des exons et la position des introns sont respectés mais ces derniers sont beaucoup plus petits chez *A. thaliana* que chez le maïs ou le pois, par exemple. Enfin, les moyens informatiques dont on dispose pour décrypter les séquences anonymes et, en particulier, les programmes adaptés à *A. thaliana*, font des progrès rapides, augmentant notre capacité à prédire la fonction à partir de la séquence.

La connaissance complète des gènes d'une espèce modèle permettra de repérer les gènes qui pourraient être spécifiques d'une espèce et dont on ignore actuellement la quantité. Si connaître l'ensemble des gènes d'une espèce ne suffit évidemment pas à comprendre son génome, cela fournit cependant la base de départ nécessaire pour étudier leurs régulations et leurs interactions dans toute leur complexité. Dans ce contexte général, il est important de garder en tête le fait qu'il existe un bénéfice réciproque entre les différents programmes de séquençage systématique de génomes. De plus, il y a un aspect fédérateur des résultats produits par un programme de séquençage. Ainsi, une séquence produite dans un laboratoire utilisant *A. thaliana* verra sa signification accrue du fait des données obtenues par le programme de séquençage systématique.

2 - LE PROGRAMME EUROPEEN: EUROPEAN SCIENTISTS SEQUENCING ARABIDOPSIS (E.S.S.A.)

A l'instar du séquençage du génome de la levure (LEVY J. 1994), la Communauté Européenne a été la première instance à soutenir, à travers un consortium d'une douzaine de laboratoires, le séquençage d'un chromosome entier. A cette époque (fin 1992) seule la cartographie physique du chromosome 4, effectuée au John Innes Institut (JII en Grande Bretagne) était suffisamment avancée pour permettre un tel projet (HWANG *et al.* 1991). Les trois premières années de ce projet doivent permettre d'établir la faisabilité du programme et la capacité , en terme de quantité de séquence, des différents laboratoires. La taille du génome d'*A. thaliana* est estimée à 100 Mpb réparties sur 5 chromosomes. Le chromosome 4, environ 21,5 Mpb, est couvert par 4 contigs de YACs ("Yeast Artificial Chromosomes" ou chromosomes artificiels de levure) dont trois sont de grande taille (>4 Mpb). L'objectif est le séquençage, en trois ans, de 2 Mpb contiguës au milieu du bras inférieur du chromosome 4. Cette région contient un certain nombre de loci affectant la floraison (AG, FCA, ...) et un locus de sensibilité à l'arabinose (ARA1). Le projet s'articule autour d'un coordinateur (M. BEVAN, JII) et comprend un laboratoire responsable de la répartition et de la fourniture des cosmides à séquencer, d'un laboratoire informatique et des laboratoires chargés du séquençage. Ces derniers sont répartis en Grande-Bretagne, en Allemagne, en France, en Espagne et en Grèce. Le laboratoire informatique (Martinsried Institute for Protein Sequences, MIPS) est spécialisé dans l'assemblage des ressources sur le génome d'*A. thaliana*, l'analyse des étiquettes de RNA messagers et des séquences génomiques. Il développe pour l'ensemble de la communauté un outil de présentation graphique des résultats qui sera accessible au public après validation des séquences. La participation de la France, soutenue par le Groupement de Recherches et d'Etudes sur les Génomes (Dr. P. SLONIMSKI) s'est répartie entre 4 laboratoires. L'un de ces laboratoires (M. KREIS à Orsay) séquence 75 kpb du

chromosome 4 dans la région du locus FCA, les trois autres (M. DELSENY à Perpignan, B. LESCURE à Toulouse et R. MACHE à Grenoble) séquencent chacun 75 kpb autour de loci d'intérêt pour leurs laboratoires et pouvant se situer sur d'autres chromosomes. Par ailleurs, ces mêmes laboratoires, plus ceux de H. HOFTE (Versailles), de G. GIRAUDAT (Gif sur Yvette), de C. GIGOT et de J. FLECK (Strasbourg) ont entrepris un travail de caractérisation de séquences étiquettes (EST) d'un tissu donné, différent pour chacun. Ces ESTs correspondent à de courts fragments (300 à 600 nt) de séquences transcrites. Elles permettent de remonter aisément aux gènes correspondants, fournissent les cDNAs sans l'effort parfois important du clonage ciblé et sont utilisées dans les programmes de cartographie actuels.

A l'issu des deux premières années du programme E.S.S.A., 1 Mpb ont été séquencées et sont au stade de vérification et d'assemblage au MIPS.

3 - QU'AVONS NOUS APPRIS?

3.1 - nouveaux gènes

Plus de 130 gènes nouveaux dont la fonction est prédite par similitude avec des gènes connus chez d'autres organismes ont été séquencés. L'appariement se fait avec des gènes d'autres plantes mais aussi avec des gènes d'autres organismes comme les bactéries, la levure, les animaux et l'homme. Environ 40 autres gènes (20%) détectés par analyse informatique ne présentent aucune ressemblance avec des gènes connus. Les bases de données internationales contiennent pourtant plus de 450 000 séquences et 320 millions de nucléotides. L'importance de ce résultat, 170 gènes nouveaux, est claire si l'on se rappelle qu'avant le démarrage du programme de séquençage systématique du génome d'*A. thaliana* il n'y avait que 300 séquences nucléotidiques d'*A. thaliana* et 100 séquences protéiques dans les bases de données. Les gènes sans aucun homologue connu, dont le nombre devrait se réduire avec l'avancée des différents programmes de séquençage (88 000 ESTs distinctes pour l'homme; ADAMS M. D. *et al.*, 1995), constituent un objet d'étude particulièrement attractif pour l'avenir.

3.2 - structure des gènes

La structure des gènes varie entre des gènes ne présentant pas d'intron et ceux extrêmement morcelés avec parfois plus de 30 exons dont certains peuvent être très courts. La densité moyenne des gènes dans la région du locus FCA est de 1 gène tous les 5 kpb, avec une distance intergènique de 0,3 à 10 kpb. Compte tenu de la taille du génome, ce nombre semble indiquer qu'il doit y avoir de 20 à 25 000 gènes codant pour des protéines chez *A. thaliana*. A titre de comparaison, le nombre total de gènes par génome haploïde de l'homme est estimé à un peu moins de 80 000 (ANTEQUERA, F. and BIRD, A. 1993).

3.3 - distribution et expression

Des résultats préliminaires suggèrent une distribution non aléatoire des niveaux de transcription et la possibilité d'un regroupement des gènes présentant un niveau d'expression semblable. Il y a là un champ de recherches prometteur pour la compréhension du contrôle de l'expression des gènes.

3.4 - ESTs

Un quart des gènes prédits ont un appariement avec soit une EST d'*A. thaliana* soit une EST d'autres plantes. Plus de 6 000 ESTs non redondantes provenant de 10 banques de cDNAs ont été produites par le consortium français (HÖFTE *et al.*, 1993; COOKE *et al.*, 1996). A celles-ci s'ajoute un nombre à peu près équivalent de séquences américaines (NEWMAN *et al.*, 1994). Encore une fois, la comparaison de ces nombres avec ceux déjà cités des séquences connues avant le démarrage de ce programme est instructive. Le nombre élevé d'appariements avec des ESTs montre combien la complémentarité des approches génomiques et EST peut être fructueuse, d'autant plus que la conservation de petits blocs de séquences entre les gènes de céréales et ceux des dicotylédones est suffisante pour utiliser le nombre rapidement croissant de séquences obtenues par le programme riz (essentiellement japonais) pour la prédiction des gènes d'*A. thaliana* et réciproquement. L'analyse des ESTs d'*A. thaliana* et du riz (HAVUKKALA I. *et al.*, 1995) a montré qu'il existe des éléments génétiques anciens qui peuvent servir de marqueurs universels et communs à tous les génomes végétaux et être utilisés en cartographie comparée. L'utilité de cette approche est attestée par la demande en clones des laboratoires du monde entier : plus de 1 000 rien que pour le projet EST français.

Dans le cas de gènes ne présentant pas d'appariements avec des gènes connus, il est encore possible d'obtenir des informations quant à la localisation cellulaire (pariétale, membranaire ou compartimentale), grâce aux connaissances de plus en plus précises des séquences d'adressage des protéines.

3.5 - rétrotransposons

La région FCA du chromosome 4 a révélé des séquences du type rétrotransposon COPIA à intervalle d'environ 150 kpb. La comparaison des taux de recombinaison, dont l'estimation est en cours actuellement, pour l'ensemble de la région séquencée, avec la position des séquences, permettra d'évaluer leur rôle dans les remaniements chromosomiques. De même, l'étude de l'apport en éléments régulateurs pour l'activité des gènes par ces rétrotransposons est maintenant accessible.

3.6 - gènes et mutants

Associé à la cartographie génétique, le séquençage systématique a permis de désigner des gènes candidats responsables de phénotypes mutants connus.

4 - L'avenir

Les trois années du programme E.S.S.A. s'achèvent à la fin de l'année 1996. Les Européens séquenceront le bras inférieur du chromosome 4 (SCHMIDT R. *et al.*, 1995), soit 11,4 Mpb, pour l'an 2 000. Les Américains vont commencer le séquençage du bras supérieur du même chromosome, soit 3,5 Mpb, au cours de la même période. Le nombre de gènes sur le chromosome 4 est maintenant estimé à 3 200. Pour ces deux projets, le financement est assuré mais d'autres verront certainement le jour dans un avenir proche. Actuellement, la cartographie physique (contigs de YACs) est finie pour le chromosome 2 (laboratoire de H. GOODMAN, U.S.A.) et est très avancée pour le chromosome 1 (laboratoire de J. ECKER, U.S.A.) et le chromosome 5 (JII, G.B.). On estime que, pour l'an 2 000, quarante pour cent du génome d'*A. thaliana* seront séquencés grâce à un effort commun au sein du Projet

International sur le Génome d'*Arabidopsis* dont le but principal est la découverte de tous les gènes de cette espèce ainsi que de leur fonction.

Journée de l'A.S.F. du 1er février 1996

BIBLIOGRAPHIE

ADAMS, M. D. *et al.* -1995 - Initial Assessment of Human Gene Diversity and Expression Patterns Based upon 83 Million Nucleotides of cDNA Sequence. Nature 377 SUPP.:3-17

ANTEQUERA, F. et BIRD, A. - 1993 - Number of CpG Islands and Genes in Human and Mouse. Proc. Natl. Acad. Sci. USA 90: 11995-11999

BRENNER, S. *et al.* - 1993 - Characterization of The Pufferfish (*Fugu*) Genome as A Compact Model Vertebrate Genome. *Nature* 366: 265-268

COOKE, R. *et al.* - 1996 - Further Progress towards A Catalogue of all *Arabidopsis* Genes : Analysis of A Set of 5000 Non-Redundant ESTs. *Plant J.* 9: 101-124

DEAN, C. and SCHMIDT, R. - 1995 - Plant Genomes : A Current Molecular Description. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 46: 395-418

HWANG, I. *et al.* - 1991 - Identification and Map Position of YAC Clones Comprising One-Third of The *Arabidopsis* Genome. *Plant J.* 3: 367-374

HAVUKKALA, I., ICHIMURA, H., NAGAMURA, Y. and SASAKI, T. -1995 - Rice Genome Analysis by Integration of Sequencing and Mapping Data. *J. Biotechnol.* 41: 139-148

HÖFTE, H. *et al.* - 1993 - An Inventory of 1152 Expressed Sequences Tags Obtained by Partial Sequencing of cDNA from *Arabidopsis thaliana. Plant J.* 4: 1051-106

LEVY, J. - 1994 - Sequencing The Yeast Genome : An International Achievement. *Yeast* 10: 1689-1706

NEWMAN, T. *et al.* - 1994 - Genes Galore : A Summary of Methods for Accessing Results from Large-Scale Partial Sequencing of Anonymous *Arabidopsis* cDNA Clones. *Plant Physiol.* 106: 1241-1255

SCHMIDT, R. *et al.* -1995 - Physical Map and Organization of *Arabidopsis thaliana* Chromosome 4. *Science* 270: 480-483

SULSTON, J. *et al.* -1992 - The *C. elegans* Genome Sequencing Project : A Beginning. *Nature* 356: 37-41

WILLIAMS, N. -1995 - Genetics Research - Closing in on The Complete Yeast Genome Sequence. *Science* 268 : 1560-1561