Enjeux du développement des recherches fondamentales

Qu'avons-nous appris en analysant le génome d'*Arabidopsis thaliana* ?

Michel Delseny

Le choix d'*Arabidopsis thaliana*, comme espèce végétale modèle pour analyser le génome des plantes, est justifié par la petite taille de son génome, 120 Mpb environ répartis en cinq paires de chromosomes, l'un des plus petits décrits chez les plantes. Cette plante *(photo 1)* appartient à la famille des crucifères, qui comprend plusieurs espèces cultivées économiquement importantes comme le colza, les choux ou les moutardes. Elle est essentiellement autogame et se prête bien à l'analyse génétique classique ou inverse. Elle se développe rapidement sans exiger beaucoup de place ni de soin, ce qui en fait un objet d'étude particulièrement intéressant dans un contexte académique [1]. Depuis 1989, la recherche sur *Arabidopsis thaliana* est coordonnée au niveau mondial et, depuis 1992, plusieurs projets, principalement européens, puis américains et japonais ont permis d'envisager le séquençage complet du génome de cette plante. Cet article est limité à ce que nous avons appris du séquençage systématique.

En fait, ces projets de séquençage comprennent deux grands volets : le séquençage rapide et partiel des gènes exprimés, sous la forme d'étiquettes, et le séquençage génomique complet. Fin 1991, moins de 200 séquences ADN complémentaires (copies d'ARN messagers ou ADNc, gènes et protéines) étaient connues chez *Arabidopsis*. La première tâche entreprise fut de séquencer systématiquement des ADNc de façon à dresser un catalogue des gènes exprimés. Dans un premier temps, il n'est pas nécessaire de séquencer la totalité de l'ADNc mais simplement quelques centaines de nucléotides à l'une des extrémités de façon à identifier l'ADNc par un morceau de séquence encore appelé étiquette. Le terme anglais EST *(Expressed sequence tag)* signifie étiquetage d'une séquence transcrite. Cette tâche a été partagée entre un consortium français et un consortium américain. La seconde tâche entreprise fin 1993 par le programme européen ESSA (European scientists sequencing *Arabidopsis*) a consisté à analyser la séquence de grands morceaux d'ADN chromosomique.

Photo. Plante adulte d'*Arabidopsis thaliana*, sept semaines après semis (photo : M.-F. Chanoué).

Séquençage d'EST *(Expressed sequence tags)*

Une EST est une séquence de 300-400 pb, déterminée une seule fois, portant sur l'extrémité 5' ou 3' d'un clone ADNc. En séquençant en aveugle un grand nombre d'EST, on obtient une représentation de la population d'ARN messager (ARNm) dans un tissu ou à un stade de développement donné.

Les stratégies américaine et française ont été légèrement différentes. Les Américains ont réalisé une banque d'ADNc unique à partir de mélanges d'ARNm préparés à partir des différents tissus de la plante, alors que les Français ont réalisé des banques spécialisées correspondant à leur centre d'intérêt

M. Delseny : Laboratoire de physiologie et biologie moléculaire des plantes, UMR 5545, CNRS, Université de Perpignan, 66860 Perpignan cedex, France.

Tirés à part : M. Delseny

biologique. Les Américains n'ont séquencé que l'extrémité 5' des clones alors que de très nombreux ADNc ont été séquencés aux deux extrémités par les Français. Le programme européen n'a financé que les séquences nouvelles, tandis qu'aucune contrainte de redondance n'a été imposée aux collègues américains.

L'effort conjoint des deux consortiums a ainsi permis d'obtenir près de 36 000 étiquettes en 5 ans. Du fait de la redondance on peut estimer qu'au moins la moitié des quelque 20 000 gènes du génome d'*Arabidopsis* ont ainsi été étiquetés [2].

Le séquençage en lui-même constitue une opération facile. La difficulté est de déchiffrer le message obtenu. La première analyse réalisée consiste à traduire la séquence de l'ADNc dans chacune des six phases de lecture possibles et à comparer la traduction aux séquences protéiques engrangées dans les bases de données. Les logiciels FASTA, BLAST et quelques autres permettent ce travail. Ainsi, si l'on ne considère que les séquences non redondantes, on peut observer que près de 40 % d'entre elles présentent une homologie significative avec une protéine déjà connue chez un autre organisme. Cela signifie que près de 60 % des clones séquencés et non redondants sont sans fonction connue. Ainsi, par simple séquençage partiel, on a pu obtenir une information de séquence et proposer une fonction potentielle pour près du quart des gènes d'*Arabidopsis*. Il s'agit de résultats moyens sur l'ensemble du programme et qui peuvent varier d'un tissu à l'autre.

Que peut-on faire avec les EST ?

Le premier renseignement tiré des EST est bien sûr la connaissance et l'identification potentielle d'un gène. Les premières applications mises en place par nos laboratoires ont surtout concerné la biologie. Chaque EST peut constituer une sonde pour analyser l'expression d'un gène. Lorsque les EST révèlent l'existence de familles multigéniques, il devient possible d'analyser leur expression différentielle et d'en rechercher les mécanismes de contrôle. Le nombre de gènes appartenant à des petites familles multigéniques a constitué l'une des surprises majeures du programme. Ainsi, on s'aperçoit que certaines protéines sont codées par plusieurs gènes de séquence très proche, mais dont l'expression est souvent spécialisée ou prédominante dans un type cellulaire, un tissu ou un organe donné. Cette observation suggère que les plantes se sont développées en dupliquant un nombre important de leurs gènes et en spécialisant leur séquence de régulation pour définir des expressions limitées dans le temps et l'espace.

Une deuxième application des EST a été de les utiliser pour faire la cartographie du génome d'*Arabidopsis*. La première stratégie utilisée a été la technique classique du RFLP *(Restriction fragment length polymorphism)* ou polymorphisme de longueur des fragments de restriction mais on peut aussi utiliser les informations de séquences pour définir des marqueurs PCR. La PCR *(réaction de polymérisation en chaîne)* est une méthode qui permet d'amplifier spécifiquement un fragment d'ADN délimité par deux séquences amorces. Cette stratégie a été particulièrement utile pour typer des clones YAC *(Yeast artificial chromosome)*, définir quelles EST étaient communes à plusieurs YAC, définir des contigs* et réaliser la carte physique des chromosomes. Celle-ci est maintenant quasiment achevée pour quatre des cinq chromosomes, et en voie de l'être pour le dernier [3, 4].

On doit se poser la question de savoir combien de gènes nous avons réellement identifiés. Cette question est abordée en regroupant toutes les EST qui présentent des régions chevauchantes identiques ou très proches de façon automatique. Ainsi les 30 000 EST peuvent être regroupées en 13 000 contigs [5], ce qui est sans doute une surestimation du nombre réel de gènes, puisque deux contigs peuvent correspondre à un même gène s'il n'y a pas, par exemple, d'EST dans la région centrale du gène chevauchant avec celles de la partie 5' ou de la partie 3'.

Nous avons essayé de préciser la situation dans le cas des EST correspondant à des protéines ribosomiques cytoplasmiques : nous avons ainsi reconstitué 106 ADNc distincts correspondant à 50 types de protéines différentes [6].

Une dernière application des EST est le repérage des gènes et en particulier des bordures entre exons et introns.

Séquençage génomique

Le séquençage génomique dans le cadre du programme ESSA 1 a porté sur environ 2,5 Mpb (Mega paires de bases, ou millions de paires de bases), dont 1,9 Mpb d'un seul tenant autour du locus *FCA*, un gène contrôlant la précocité de la floraison situé sur le chromosome 4.

Au total près de 450 gènes potentiels ont été entièrement séquencés [7]. Il reste cependant à confirmer que chacun de ces gènes est bien réel et actif.

Le programme a rencontré plusieurs types de difficultés. La première a été l'obtention de clones prêts à séquencer. Il a fallu d'abord organiser les contigs de phages ou de cosmides (dont le sous-clonage orienté est laborieux) avec un chevauchement minimal. Dans la phase actuelle du projet, ce sont des fragments d'ADN clonés dans des vecteurs BAC *(Bacterial artificial chromosome)*, ou PAC *(P1 Phage artificial chromosome)*, qui sont séquencés. Ces clones ont l'avantage de pouvoir insérer des fragments de 100 kpb, ce qui accélère considérablement le processus. Il a fallu apprendre à travailler d'une autre façon, quasi industrielle, et être capable de repérer les gènes et leurs limites dans les séquences obtenues. De très nombreux gènes sont en effet beaucoup plus morcelés que l'on ne s'y attendait. L'obtention préalable de 36 000 EST d'*Arabidopsis* et l'utilisation d'EST humaines ou du nématode *Caenorhabditis elegans* ont grandement facilité la tâche, et l'isolement d'ADNc complets correspondant aux gènes séquencés a permis d'améliorer les programmes de prédiction de gène. Selon les régions séquencées, de 30 à 60 % des gènes prédits présentent une identité de séquence avec une EST, ce qui confirme bien que le programme EST a probablement repéré environ la moitié des gènes actifs.

Les principaux enseignements sont que la densité de gènes est très élevée (1 gène tous les 4 à 5 kpb, ce qui ne laisse pas la place pour plus de 20 000 gènes.) Ces gènes ne sont pas groupés de façon fonctionnelle évidente, même si l'on observe des groupes de fonctions analogues. Les gènes sont morcelés avec souvent des exons de très petite taille. Le plus petit exon connu ne comprend que 9 pb et est présent dans un gène d'invertase. L'analyse détaillée de quelques gènes révèle l'existence de mécanismes d'épissage alternatif et de bordures d'introns atypiques [8], ce qui laisse entrevoir de nouvelles possibilités de régulation. Enfin, le séquençage génomique a révélé de nouveaux éléments transposables ainsi que des courtes duplications qui devraient être riches d'enseignements sur le mode d'évolution du génome. Depuis 1997 le séquençage génomique s'accélère. L'effort européen a stimulé les initiatives dans les autres pays et un consortium international AGI (*Arabidopsis* genome initiative) s'est constitué [9]. Il

* Contigs : ensemble de clones chevauchants définissant un fragment continu d'ADN.

comprend trois consortiums américains, le consortium européen, un consortium japonais et le Genoscope français. Fin 1997 environ 3,5 Mpb étaient disponibles dans les bases de données, il y en a en juillet 1998 près de 30. Les bases de données sont mises à jour chaque semaine et c'est un déluge d'informations qui assaille maintenant le chercheur.

Conclusion

Les cinq années écoulées ont vu se mettre en place une formidable machine technologique qui a fait faire un bond fulgurant à notre connaissance du génome d'*Arabidopsis*. Trois grandes orientations se dégagent.

• Achever le plus vite possible le séquençage complet du génome. Ce travail devait être terminé en 2003 ; en fait, on pense maintenant qu'il le sera à la fin de l'an 2000.

• Développer les méthodologies d'identification de fonction des gènes d'*Arabidopsis*. Il n'y a malheureusement pas de méthode miracle. Les stratégies vont de la comparaison régulière des séquences encore inconnues avec les bases de données en croissance exponentielle, à la création de collections de mutants d'insertion, en passant par un retour à la biochimie classique. Cette tâche nécessitera le travail de plusieurs générations de biologistes.

• Transférer au plus vite les acquis obtenus sur un génome modèle aux plantes d'intérêt économique. Cela va de l'utilisation directe de sondes d'*Arabidopsis* lorsqu'elles sont suffisamment conservées, à la mise en œuvre de stratégie PCR ou de cartographie comparée. Du point de vue des biotechnologies, les principaux apports du programme de séquençage d'*Arabidopsis* ont été de permettre un accroissement considérable de la connaissance des gènes de la plante et la mise à disposition, dans le domaine public, d'informations permettant de repérer près de la moitié d'entre eux. Les séquences publiées sont ainsi maintenant utilisables pour aller rechercher les gènes homologues chez les espèces cultivées et s'en servir pour faire du génie génétique ou pour évaluer, au travers de l'étude de la variabilité génétique, les relations séquence-fonction pour certains gènes importants.

De façon générale, ces programmes d'analyse du génome d'une espèce modèle ont permis de faire faire un bond en avant considérable dans l'analyse de grandes fonctions telles que la floraison, les régulations hormonales ou la perception et la transduction des signaux de l'environnement. Ces connaissances nouvelles constituent le tremplin sur lequel les biotechnologies végétales devraient prendre un nouvel essor.

Enfin, la conséquence de ces programmes est l'émergence d'une nouvelle façon de travailler et de faire de la biologie en utilisant les informations contenues dans des bases de données largement publiques. On peut prévoir que l'existence de toutes ces séquences va révolutionner nombre de domaines de la biologie en changeant les stratégies expérimentales et en mettant sur le marché de nouveaux outils de diagnostic et de sélection. Le programme international sur le génome d'*Arabidopsis* a montré qu'il était désormais possible d'acquérir des informations essentielles sur un génome de plante dans un laps de temps très limité. De nombreuses compagnies privées et organismes publics de recherche développent maintenant des programmes d'EST sur les principales espèces cultivées. Il y en a ainsi 26 000 sur le riz, mises dans le domaine public par les Japonais. Il y en a à peu près 450 000 chez le maïs, malheureusement inaccessibles car du domaine privé, et plusieurs milliers chez le soja, le cotonnier ou la vigne.

Dans le même temps, la communauté scientifique internationale envisage de séquencer des génomes plus importants. Les opérations sont déjà lancées sur le riz dont le génome, gros comme 4 fois celui d'*Arabidopsis*, pourrait être achevé en 2005-2006.

Ces données nouvelles favorisent l'émergence d'une nouvelle science, la génomique, qui regroupe les activités de cartographie, séquençage, bio-informatique mais qui surtout incite à appréhender les problèmes de façon globale et non plus gène par gène.

Ainsi la connaissance complète d'un génome va permettre d'évaluer globalement l'expression de l'ensemble des génomes d'un organisme en constituant des puces à ADN ou *DNA chips*. De telles puces sont déjà commercialisées pour la levure et *Escherichia coli*, dont les gènes sont entièrement séquencés, et sont en développement à partir des EST d'*Arabidopsis*.

Pour en savoir davantage sur le génome d'*Arabidopsis* on se reportera à quelques revues récentes [10, 11] ▩

Références

1. Somerville CR, Meyerowitz EM. In : Somerville CR, Meyerowitz EM, eds. *Arabidopsis*. New York : Cold Spring Harbor Laboratory Press, 1994 : 1 300 pages.

2. Delseny M, Cooke R, Raynal M, Grellet F. The *Arabidopsis thaliana* cDNA sequencing projects. *FEBS Letters* 1997 ; 403 : 221-4.

3. Schmidt R, West J, Love K, *et al.* Physical map and organisation of *Arabidopis thaliana* chromosome 4. *Science* 1995 ; 270 : 480-3.

4. Camilleri C, Lafleuriel J, Macadre F, *et al.* A YAC contig map of *Arabidopsis thaliana* chromosome 3. *Plant J 1998* ; 14 : 633-42.

5. Rounsley S, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JC, Kerlavage AR. The construction of *Arabidopsis* expressed sequence tags assemblies. *Plant Physiol* 1996 ; 112 : 1177-83.

6. Cooke R, Raynal M, Laudié M, Delseny M. Identification of members of gene families in *Arabidopsis thaliana* by contig construction from partial cDNA sequences : 106 genes encoding 50 cytoplasmic ribosomal proteins. *Plant J* 1997 ; 11 : 1127-40.

7. Bevan M, *et al.* Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. *Nature* 1998 ; 391 : 485-8.

8. Wu HJ, Gaubier Comella P, Delseny M, Grellet F, van Montagu M, Rouzé P. Non-cannonical introns are at least 10^9 years old. *Nature Genetics* 1996 ; 14 : 383-4.

9. Bevan M, Ecker J, Theologis S, *et al.* Objectif : the complete sequence of a plant genome. *Plant Cell* 1997 ; 9 : 476-8.

10. Delseny M, Cooke R, Comella P, Wu HJ, Raynal M, Grellet F. The *Arabidopsis thaliana* genome project. *C R Acad Sci Paris* 1997 ; 320 : 589-99.

11. Delseny M, Cooke R. The *Arabidopsis* nuclear genome. In : Spurr NK, Young BD, Bryant SP, eds. *ICRF handbook of genome analysis*. Blackwell Science Ltd., 1998 ; 2 : 761-87.

Résumé

Qu'avons-nous appris en analysant le génome d'*Arabidopsis thaliana* ?

M. DELSENY

Arabidopsis thaliana est une crucifère sauvage, dont le génome (l'un des plus petits connus chez les plantes) sert de modèle d'étude. Dans un premier temps, les copies ADN des ARN messagers ont été séquencées de façon partielle et sans sélection, de manière à produire une collection d'étiquettes représentatives de la fraction la plus abondamment transcrite du génome. Dans un deuxième temps, le séquençage du génome lui-même a été entrepris. Actuellement le quart de ce génome est connu et sa séquence devrait être achevée d'ici l'an 2000.

L'utilisation de ces séquences pour étudier la biologie d'*Arabidopsis* et pour réaliser le transfert de connaissance et de technologie en direction des espèces cultivées est discutée.

Summary

Analysis of the Arabidopsis thaliana *genome – what has been learned?*

M. DELSENY

Arabidopsis thaliana, *a wild crucifer with one of the smallest known plant genomes, is an interesting research model. cDNA copies of mRNA were first partially and randomly sequenced, with the aim of obtaining a collection of tags representative of the most abundant transcribed fraction of the genome. Genome sequencing was then undertaken. A quarter of this genome has now been sequenced, and the full sequencing should be completed by the year 2000.*

A discussion is presented on the use of these sequences to investigate the biology of Arabidopsis *and on the potential transfer of knowledge and technology to cultivated species.*

Cahiers Agricultures 1998 ; 7 : 459-62.