m/s SYNTHÈSE

médecine/sciences 2001 ; 17 : 829-35

Séquençage des génomes eucaryotes : Arabidopsis, le quatrième élément

Alain Bernot
Nathalie Choisne
Marcel Salanoubat

Avec le séquençage des trois premiers génomes eucaryotes (levure, nématode et drosophile), nous avons eu accès au catalogue des gènes présents dans trois grands groupes du vivant. Les gènes de deux autres grands groupes étaient jusqu'à présent absents de ce catalogue : les végétaux et les deutérostomiens. Le séquençage du génome d'*Arabidopsis* et la première version de la séquence du génome humain ont comblé cette lacune. Certains membres du groupe des protistes, actuel parent pauvre, sont en cours de séquençage et la détermination de la séquence d'un génome complet de l'un d'entre eux verra le jour très bientôt. Nous nous sommes attachés, dans cet article, à donner un aperçu d'un certain nombre de données issues du séquençage du premier génome de plante en mettant son analyse dans le contexte des autres génomes déjà séquencés.

Arabidopsis thaliana est un modèle important pour les plantes à fleurs. L'analyse de la séquence de ce génome nous donne accès à une quantité considérable d'informations, qui permet d'établir les bases du fonctionnement des plantes, mais aussi d'enrichir la compréhension des processus conservés chez l'ensemble des eucaryotes. En particulier, un catalogue de tous les gènes qui assurent le cycle de vie d'une plante a été proposé. Il est probable que ce catalogue est imparfait, car les gènes qui le composent sont essentiellement fondés sur des prédictions et, dans la grande majorité des cas, leur fonction demeure hypothétique. Ce catalogue permet néanmoins d'évaluer, pour la première fois, l'ampleur de la tâche qui reste à accomplir pour comprendre toute la complexité des processus qui régissent la vie d'un angiosperme.

De l'ombre à la lumière

La génétique formelle a commencé en 1866 par l'établissement des trois premières lois de la génétique par Mendel, qui réalisa l'essentiel de ses travaux sur le pois *(Pisum sativum)*. Par la suite, McClintock a mis en évi-

ADRESSES

A. Bernot, M. Salanoubat : Genoscope et Cnrs UMR-8030, 2, rue Gaston-Crémieux, 91057 Évry Cedex, France. N. Choisne : Genoscope et Inra-URGV, 2, rue Gaston-Crémieux, 91057 Évry Cedex, France.

dence, en 1950, un ensemble de phénomènes conduisant à l'instabilité de certains locus chez le maïs *(Zea mays)*, éléments qui ont été caractérisés 30 ans plus tard au niveau moléculaire [1]. En 1983, la première expérience de transformation de plante ayant un phénotype nouveau, apporté par l'ADN-T (ADN de transfert) d'une agrobactérie, a été réalisée chez le tabac *(Nicotiana tabacum)* [2]. Aujourd'hui, les trois espèces cultivées dont la production mondiale dépasse les 500 millions de tonnes par an sont le blé, le maïs et le riz. Dans le domaine horticole, le plus gros chiffre d'affaire de la profession est réalisé par les roses**.** Alors comment se fait-il que la première plante dont le génome ait été entièrement séquencé soit une obscure mauvaise herbe, *Arabidopsis thaliana* (arabette des dames), qui, il y a quelques dizaines d'années, n'était connue encore que par un petit nombre de botanistes et de généticiens ? La réponse est contenue dans une citation de Stuyvesant qui disait en substance que « le bon choix d'un organisme dépend de la nature du problème posé, du moment où le problème apparaît et des compétences de l'expérimentateur ». Quand, il y a quelques années, un certain nombre de chercheurs en génétique végétale ont entrevu le fantastique intérêt d'avoir la séquence complète d'un génome de plante, les capacités de séquençage étaient limitées. S'il avait déjà été démontré que le séquençage d'un génome d'environ 100 Mb était une tâche faisable (le séquençage du nématode *Caenorhabditis elegans* était en cours), personne ne pouvait imaginer, au sein de la communauté des végétalistes, le séquençage d'un génome ayant une taille de plusieurs milliers de mégabases comme celui du blé ou du maïs *(Tableau I)*. Il était donc nécessaire de choisir une plante dont le génome soit de petite taille. Toutefois, si cette condition était nécessaire, elle était loin d'être suffisante : le séquençage d'un génome n'est pas une fin en soi, mais un début qui doit être complété par un grand nombre d'analyses, pour lesquelles les facilités de « manipulation » de la plante choisie sont d'une importance considérable. De ce point de vue, *Arabidopsis thaliana* présente différents avantages tels qu'une taille réduite, un temps de génération très court, un nombre de graines très important, une facilité de transformation, et bien sûr la taille de son génome, la plus faible qui soit actuellement connue chez les plantes à fleurs [3]. Ce sont ces considérations qui, ajoutées à l'existence d'une génétique solide, ont conduit la communauté scientifique végétale à choisir *Arabidopsis* comme plante modèle, et, à ce titre, être le premier génome de plante entièrement séquencé.

La première initiative de séquençage fut européenne. Elle a débuté en 1994 sous l'impulsion d'un chercheur britannique, Mike Bevan, et elle fut suivie en 1996 par la création de l'*Arabidopsis Genome Initiative* (AGI) [4]. L'AGI, comprenant des représentants des six instituts ou consortiums internationaux impliqués dans le séquençage, a pris en charge l'organisation et l'intégralité du séquençage de cette plante. Ce projet s'est achevé en décembre 2000 [5-8].

Une stratégie de séquençage déjà démodée ?

Le génome d'*Arabidopsis* a été séquencé selon une stratégie « BAC à BAC » *(m/s 2000, n° 1, p. 10)* qui repose sur l'existence de cartes génétiques, physiques, et de données de cartographies supplémentaires (séquences d'extrémités de BAC – *bacterial artificial chromosome* –, construction de nouvelles cartes physiques fondée sur des données de profils d'enzyme de restriction de ces mêmes clones de BAC). Depuis 1996, les stratégies de séquençage des génomes de grande taille ont largement évolué, et la tendance actuelle semble aller vers une stratégie de séquençage aléatoire global du génome. Cette stratégie a déjà été utilisée par certaines entreprises privées américaines : *Celera* pour le séquençage de la drosophile [9] et de l'homme [10], *Syngenta* pour celui du riz. D'autres projets utilisant cette stratégie sont en cours, comme par exemple le séquençage du génome de la souris par *Celera*. Par ailleurs, le projet public de séquençage du génome de la souris par le Consortium international est lui réalisé grâce à une stratégie modifiée, combinant à la fois l'approche de séquençage aléatoire global du génome avec une approche de séquençage « BAC à BAC ». Si l'efficacité de la stratégie de séquençage aléatoire global du génome est loin d'être prouvée dans le séquençage du génome humain *(m/s 2001, n° 3, p. 287-9)*, elle semble l'être pour le séquençage de génomes à faible complexité comme la drosophile. Ainsi, la question de son utilisation a été posée dans le cas du séquençage d'*Arabidopsis* : au cours de la première réunion de l'AGI, un groupe américain avait proposé de séquencer environ un équivalent génomique de façon totalement aléatoire, en complément de la stratégie de séquençage « BAC à BAC ». Cette approche aurait eu l'avantage d'obtenir rapidement un inventaire partiel des gènes d'*Arabidopsis*. Cette proposition a été refusée principalement en raison de la difficulté liée à l'intégration des données engendrées par les deux approches. Les avantages d'une stratégie de séquençage aléatoire global peuvent se résumer en deux mots : coût et rapidité. Néanmoins, elle comporte des inconvénients qui sont essentiellement liés à la qualité de la séquence obtenue. Au moment de sa publication, la séquence du génome de la drosophile comportait 1 434 lacunes de séquence, à comparer aux 148 lacunes du génome de *C. elegans* [11] et aux quelques lacunes

Tableau I. Caractéristiques des génomes d'*Arabidopsis* et de végétaux économiquement importants.

Espèce	Taille du génome (Mb)	Nombre de chromosomes
Arabidopsis thaliana (arabette)	125	5 (2n = 2x = 10)
Lycopersicon esculentum (tomate)	1 000	12 (2n = 2x = 24)
Zea mays (maïs)	2 500	10 (2n = 2x = 20)
Oryza sativa (riz)	430	12 (2n = 2x = 24)
Triticum aestivum (blé)	16 000	7 (2n = 6x = 42)
Sorghum bicolor (sorgho)	800	10 (2n = 2x = 20)
Hordeum vulgare (orge)	5 200	7 (2n = 2x = 14)

(moins de 20) du génome d'*Arabidopsis*. Il est donc évident que ce qui est appelé la séquence d'un génome recouvre des réalités très différentes, intimement liées à la stratégie utilisée, et peut-être serait-il souhaitable pour plus de clarté de définir des « catégories » indiquant la qualité de finition des différents génomes. Par exemple, les 170 000 lacunes de séquence existant dans l'assemblage de la séquence du génome humain, version *Celera*, correspondent à environ 6 800 lacunes de séquence sur un génome de la taille d'*Arabidopsis*. Pourtant, dans les deux cas, le titre de la publication était en substance : *« The sequence of… »*.

Organisation du génome : une réputation de simplicité mise à mal

L'analyse des 115 409 949 nucléotides séquencés du génome d'*Arabidopsis* a permis de confirmer un certain nombre de résultats déjà connus : le faible nombre de séquences répétées, la localisation des organisateurs nucléolaires à proximité des centromères sur les chromosomes 2 et 4, celle des gènes codant pour les ARNr 5S au niveau des centromères des chromosomes 3, 4 et 5. Elle a surtout permis d'accéder à une image globale de la structure de chaque chromosome. Celle-ci est remarquablement conservée, avec de grandes régions euchromatiques, riches en gènes (1 gène tous les 4,5 kb) s'étendant des répétitions télomériques (de séquence 5'-CCCTAAA-3') jusqu'aux régions péri-centromériques/centromériques, pauvres en gènes. Par ailleurs, *Arabidopsis* est le premier organisme dont le séquençage des centromères a été réalisé et l'analyse de ces régions a mis en évidence la présence d'environ 200 gènes. Bien que beaucoup d'entre eux ne soient pas fonctionnels, une cinquantaine de ces gènes est exprimée, et 40 correspondent à des gènes uniques dans le génome. Curieusement, des séquences répétées télomériques se retrouvent à proximité des centromères : elles pourraient être dues à des réarrangements, tels que des inversions des bras chromosomiques. De plus, une insertion récente couvrant 620 kb d'ADN mitochondrial a été trouvée dans le centromère du chromosome 2, montrant un transfert récent d'ADN de cet organite vers le génome nucléaire [12].

La plus grande surprise de l'analyse du génome d'*Arabidopsis* vient de la grande quantité de duplications qui a été mise en évidence. Ainsi, 24 régions d'une taille supérieure à 100 kb se retrouvent dupliquées : elles recouvrent 65,6 Mb, soit 58 % du génome. L'origine de ces duplications à grande échelle est encore matière à débat et deux hypothèses évolutives ont été émises : l'une fait intervenir une forme tétraploïde ancestrale, l'autre propose plusieurs événements de duplication intervenus successivement [8-13]. La redondance du génome se manifeste aussi par la présence de familles de gènes qui est évidemment due, d'une part, aux duplications du génome, mais aussi, d'autre part, à la présence de gènes d'une même famille répétés en tandem (de 2 à 23 membres). Ainsi, 1 528 familles de gènes répétés en tandem ont été répertoriées. De ce point de vue, la simplicité supposée du génome d'*Arabidopsis*, qui avait été l'un des arguments de choix pour son inscription au club fermé des organismes modèles, était donc une illusion, et la caractérisation de ses gènes devra être menée en gardant à l'esprit la possibilité de redondance fonctionnelle.

Un grand nombre de gènes, mais une complexité comparable à d'autres eucaryotes

A la fin de l'année 1991, moins de 200 gènes de plantes avaient été identifiés. Aujourd'hui, avec le séquençage du génome d'*Arabidopsis*, le nombre total de gènes identifiés chez cette espèce est de 25 498, et leurs caractéristiques sont résumées dans le *Tableau II*. Le nombre de gènes prédits chez *Arabidopsis* est supérieur à ceux obtenus chez le nématode *C. elegans* (19 100) ou la drosophile (13 600), et légèrement inférieur à l'estimation faite pour l'homme (de l'ordre de 30 000 à 35 000) [14-16] (certaines caractéristiques de ces génomes sont présentées dans le *Tableau III* et la *figure 1*). Une telle comparaison doit néanmoins être prudente, en premier lieu en raison de la redondance du génome d'*Arabidopsis* : ainsi, si au lieu de comparer le nombre de gènes entre les différentes espèces, on compare le nombre de familles de gènes (singletons et familles composées de plusieurs gènes), on obtient 11 801 familles distinctes, ce qui est du même ordre de grandeur que pour la drosophile (10 736) ou *C. elegans* (14 177). Par ailleurs, les épissages alternatifs semblent peu fréquents chez *Arabidopsis* (moins de 5 % des gènes, contre 20 % à 55 % chez les métazoaires) [16, 17] : la diversité protéomique obtenue par ce biais est donc vraisemblablement faible. Une autre façon d'aborder la complexité du protéome est de s'intéresser aux domaines protéiques (domaines Inter-Pro) [18], et en particulier à la combinatoire des ces domaines présents dans les protéines des différents organismes. Aucune analyse exhaustive n'a été présentée jusqu'à ce jour, mais des résultats préliminaires obtenus à partir des 10 domaines les plus représentés pour chacun des cinq génomes d'eucaryotes déjà sequencés semblent montrer que la complexité des protéines d'*Arabidopsis* serait inférieure à celle des génomes animaux.

Le protéome : un eucaryote parmi les eucaryotes...

La comparaison des gènes présents chez *Arabidopsis* et chez les autres génomes déjà séquencés permet de mettre en évidence les processus biologiques conservés au cours de l'évo-

Tableau II. Caractéristiques des gènes d'*Arabidopsis*.

Nombre de paires de bases (pb) séquencées	115 409 949
Nombre de gènes	25 498
Nombre d'exons	132 982
Nombre moyen d'exons par gène	5,2
Taille moyenne d'un exon (pb)	251
Nombre d'introns	107 484
Taille moyenne d'un intron (pb)	170

Tableau III. Caractéristiques comparées des génomes de la levure, du nématode, de la drosophile, de l'arabette et de l'homme.

	Levure	Nématode	Drosophile	Arabette	Homme
Nombre de gènes	6 200	19 100	13 600	25 500	30 à 35 000
Nombre moyen d'exons par gène	1,06	5,5	4,2	5,2	8,8
Taille moyenne des gènes	1,4 kb	2,7 kb	3 kb	2,1 kb	27 kb
Fréquence des gènes	2 kb	5 kb	9 kb	4,5 kb	~ 125 kb

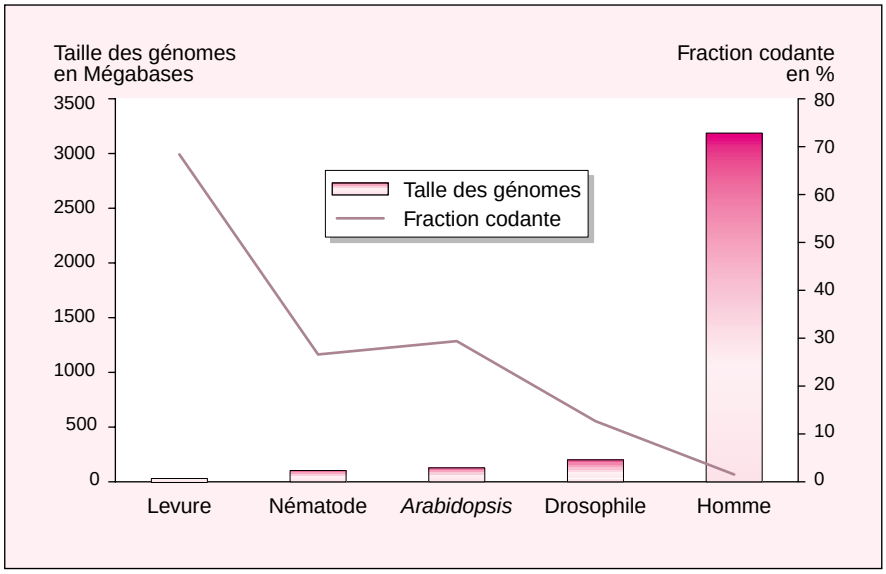

Figure 1. **Taille et fraction codante des différents génomes eucaryotes séquencés.**

lution *(figure 2)*, ainsi que les divergences spécifiques au règne végétal. L'expression des gènes chez *Arabidopsis* implique plus de 3 000 protéines, indiquant une complexité comparable à celle des autres génomes eucaryotes complètement séquencés. *Arabidopsis* possède cependant comme l'homme (et contrairement à la drosophile ou à *C. elegans*) un génome méthylé, ce qui ajoute un niveau supplémentaire au contrôle des gènes, la méthylation intervenant potentiellement par exemple dans des processus comme le *silencing* [19]. Les enzymes impliqués dans le processus de méthylation (ADN méthyltransférases) sont soit orthologues aux mammifères, soit spécifiques des plantes (comme par exemple la chromométhyltransférase [20]).

Le système de transcription d'*Arabidopsis* est typique de celui des eucaryotes. Ainsi, le génome code pour trois systèmes de transcription, comprenant les polymérases d'ARN de type I, II et III. Les protéines associées sont similaires à celles connues chez les autres eucaryotes dans le cas des polymérases d'ARN de type II et III. En revanche, les facteurs de transcription associés à la polymérase d'ARN de type I ne sont pas identifiés, hormis deux facteurs de régulation homologues de RRN3 (levure) et TTF-1 (souris). De façon plus surprenante, contrairement à tous les eucaryotes analysés à ce jour, *Arabidopsis* possède deux gènes codant pour les deux plus grandes sous-unités d'une « quatrième classe » de polymérase d'ARN, dont il reste à déterminer le rôle.

Le nombre de gènes impliqués dans le contrôle de la transcription est de l'ordre de 1 700, ce qui est deux fois et demi plus important que chez le nématode ou la drosophile, mais demeure proportionnel au nombre de gènes présents. Cette classe de gènes est la moins conservée : seulement 23 % des protéines impliquées présentent une homologie avec les autres eucaryotes (16 familles de gènes de ce type sont spécifiques des végétaux et certaines familles de facteurs de transcription connues chez les autres eucaryotes sont absentes chez *Arabidopsis*). Globalement, les gènes de réparation de l'ADN et de recombinaison *(RAR)* sont similaires à ceux identifiés chez d'autres espèces, bien que plusieurs gènes *RAR* soient uniquement présents chez *Arabidopsis*, tandis que d'autres identifiés chez les métazoaires sont absents.

L'acquisition, la distribution et la compartimentation de substrats ou d'énergie, vitales pour l'organisme, sont assurées par 600 systèmes de transport, chiffre comparable à celui du nématode (de l'ordre de 700). En ce qui concerne l'organisation cellulaire, *Arabidopsis* partage avec les autres eucaryotes les gènes codant pour les principaux composants du cytosquelette (à l'exception des gènes codant pour les filaments intermédiaires, qui n'existent pas chez les végétaux), ainsi que la plupart des gènes impliqués dans l'activité intracellulaire (trafic de vésicules, cycle cellulaire).

... mais avec des différences

Plusieurs domaines biologiques présentent évidemment des adaptations complètement distinctes de celles connues chez les métazoaires. La cytokinèse représente un exemple typique : alors qu'elle s'effectue chez les levures et les métazoaires par étranglement progressif depuis la surface vers le centre de la cellule, cette séparation est réalisée chez les végétaux par cloisonnement transversal à partir de vésicules issues de l'appareil de Golgi. De plus, ces deux types de division sont réglés par des protéines distinctes. Par ailleurs, pendant plus d'un milliard d'années d'évolutions séparées, métaphytes et métazoaires ont élaboré des voies de développement spécifiques. Par

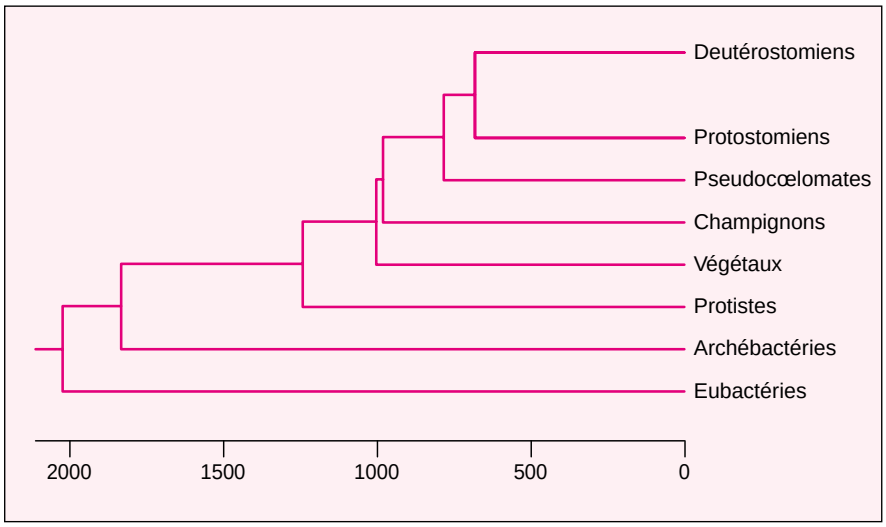

Figure 2. **Phylogénie des principaux groupes du vivant.** Trente-six génomes d'eubactéries ont été entièrement séquencés, 8 génomes d'archébactéries, un génome de végétal (Arabidopsis thaliana, Oryza sativa est en cours), plusieurs génomes de protistes sont en cours (Encephalitozoon cuniculi, Leishmania major...), un génome de champignon a été séquencé (S. cerevisiae, d'autres sont en cours), un génome de pseudocoelomate (C. elegans, C. briggsae en cours), un génome de protostomien (D. melanogaster), et le séquençage de l'homme est en cours. L'abscisse représente, en millions d'années, le temps qui nous sépare du présent (d'après [32]).

exemple, la différenciation antéro-postérieure est réalisée chez les métazoaires grâce aux facteurs de transcription à homéoboîte (région de 60 acides aminés qui interagit avec l'ADN), alors que la différenciation des sépales, pétales, étamines et carpelles est réalisée chez les métaphytes grâce aux gènes à boîte MADS (53 acides aminés, dont 9 conservés). Les gènes à boîte MADS et à homéoboîte existent aussi bien chez les métaphytes que chez les métazoaires, mais un certain nombre d'éléments suggèrent que les mécanismes qui conduisent à l'organisation spatiale sont apparus indépendamment chez ces deux groupes. Si l'identité des organes floraux repose essentiellement sur l'utilisation des gènes à boîte MADS, il faut noter que chez les plantes, les gènes à homéoboîte jouent également un rôle important. Par exemple les gènes à homéoboîte *KNOX* sont impliqués dans le maintien de la division et la prévention de la différenciation des cellules du méristème végétatif [21]. Concernant le développement et la communication intercellulaire, d'autres exemples montrent des solutions différentes adoptées par les plantes et les animaux au cours de l'évolution pour réaliser des fonctions apparentées [22].

Enfin, de nombreuses caractéristiques sont évidemment propres aux plantes, qui ne comportent qu'une quarantaine de tissus distincts, et dont l'organisation cellulaire est nettement différente de celle des métazoaires: présence de chloroplastes (essentiels pour la photosynthèse), organisation particulière des appareils de Golgi, paroi extracellulaire, importance des vacuoles, communication intercellulaire directe par l'intermédiaire de plasmodesmes. Avec le séquençage du génome nucléaire et des deux génomes cytoplasmiques [23, 24], *Arabidopsis* est devenu un organisme qui permet d'avoir une idée du nombre de gènes nécessaires pour le fonctionnement des organites. Environ 14 % et 11 % des gènes nucléaires codent pour des protéines ayant un peptide de ciblage indiquant respectivement une possible localisation chloroplastique et mitochondriale. Même si les caractéristiques essentielles du cytosquelette sont conservées avec les autres eucaryotes, un certain nombre de gènes codant pour des protéines du cytosquelette spécifiques des végétaux a été identifié. De plus, aucun gène n'est décelable qui coderait pour des protéines impliquées chez les animaux dans la liaison du cytosquelette à la matrice extracellulaire à travers la membrane. Le nombre de protéines impliquées dans le transport de l'eau est particulièrement élevé, ce qui rappelle l'importance de cette fonction dans cet embranchement. Dans le domaine des systèmes de signalisation, les plantes semblent avoir inventé des voies de transduction différentes des métazoaires. En particulier, certains systèmes de transduction utilisent des voies originales, évoquant à la fois les systèmes bactériens et animaux [25]. La majorité des plantes sont sessiles et les modifications de l'environnement se répercutent par des modifications de la physiologie, de la croissance (photomorphogenèse), qui peuvent être locales ou transmises grâce à des hormones: auxine, éthylène, acide abscissique, cytokinines dont il n'existe aucun apparenté chez les métazoaires. Finalement, les cellules végétales sont autotrophes (par opposition aux cellules animales hétérotrophes), et ne nécessitent pour vivre que des minéraux, de l'air, de la lumière et de l'eau. Ainsi, une large fraction du génome d'*Arabidopsis* code pour des enzymes impliquées dans des processus métaboliques. De ce point de vue, *Arabidopsis* contient un ensemble de gènes homologues à celui d'une cyanobactérie, *Synechocystis*. Néanmoins, le métabolisme d'*Arabidopsis* se différencie de celui de *Synechocystis* par la présence d'un grand nombre de gènes intervenant dans des processus n'appartenant qu'aux végétaux supérieurs. En particulier, environ 420 gènes sont impliqués dans la synthèse ou dans les modifications des parois, propres aux végétaux supérieurs. La redondance entre ces enzymes est importante, vraisemblablement liée à la diversité des substrats utilisés.

... et l'homme

Sur 289 gènes impliqués dans des maladies génétiques chez l'homme, 139 sont homologues d'un gène d'*Arabidopsis*, et 17 d'entre eux présentent une homologie plus élevée avec *Arabidopsis* qu'avec la drosophile ou le nématode. C'est par exemple le cas du gène *ATM (m/s 2000, n° 3,*

p. 414), impliqué dans l'ataxie-télangiectasie, et des gènes *BRCA1* et *BRCA2 (m/s 1997, n° 6-7, p .874)* qui interviennent tous les trois dans la réparation de l'ADN chez l'homme. D'un point de vue évolutif, l'homme est très éloigné d'*Arabidopsis*, tout comme il l'est de la levure. Cependant, dans certains cas, l'analyse d'homologues de gènes humains chez la levure a apporté des informations importantes pour la détermination de la fonction de ces gènes chez l'homme. Citons en particulier le cas du gène responsable de la mucoviscidose, pour lequel il a été montré chez la levure que la mutation la plus commune (ΔF508) *(m/s 1998, n° 12, p. 1413)*, qui conduit à une protéine de structure anormale, peut être corrigée par d'autres mutations de la protéine [26] ; ou bien le cas du gène responsable de l'ataxie de Friedreich *(m/s 1999, n° 11, p. 1314)*, pour lequel la localisation de l'homologue de la frataxine dans la mitochondrie a été déterminée chez la levure [27]. Dans le cas de maladies multifactorielles, des homologies entre l'homme et la plante ont été observées : un des gènes impliqués dans la maladie de Crohn a par exemple été récemment identifié [28]. Il s'agit de *NOD2* qui fait partie du système immunitaire inné, et il est étonnant de remarquer qu'un gène homologue, impliqué dans la résistance aux pathogènes bactériens, existe chez les plantes, d'autant plus étonnant qu'à ce jour moins de 10 gènes impliqués dans des maladies multifactorielles chez l'homme ont été identifiés. Il est donc probable que d'autres exemples seront identifiés dans le cas de gènes impliqués dans des processus communs à tous les eucaryotes, et pour lesquels l'identification de la fonction sera plus aisée chez les plantes que chez les animaux.

L'homme dépend des plantes (directement ou indirectement) pour sa respiration, son énergie, sa nourriture, ses acides aminés essentiels, ses vitamines, ainsi que pour un grand nombre de composés utilisés en pharmacologie (de l'aspirine au taxol). L'analyse des génomes d'autres végétaux à partir des données de séquençage d'*Arabidopsis* sera pour l'homme d'un intérêt considérable, en raison de l'importante consommation dont certains sont l'objet. La taille du génome de blé (16 000 Mb) est élevée, en raison de sa structure hexaploïde ; celle du génome du maïs est elle aussi importante (2 500 Mb), probablement en raison d'une duplication ancienne qui aurait été suivie d'un retour à l'état diploïde. De plus, la complexité de ces génomes est considérable, particulièrement en raison d'une présence importante de transposons. Pour ces deux espèces, le séquençage complet de leur génome ne semble pas être d'actualité. En revanche, le riz possède le plus petit génome aujourd'hui connu chez les monocotylédones et un projet public de séquençage du génome complet est en cours, les données de séquençage des projets privés *(Syngenta* et *Monsanto)* n'étant pas à disposition de la communauté scientifique (bien que les données de *Monsanto* soient accessibles aux groupes de séquençage participant au projet public).

L'identification de la fonction des gènes chez *Arabidopsis* a déjà eu des retombées chez les espèces cultivées, non seulement dans la même famille qu'*Arabidopsis*, à savoir les Brassicacées, mais aussi dans d'autres familles de dicotylédones et même chez des espèces monocotylédones. Par exemple, la forme mutée du gène *etr1*, qui confère chez *Arabidopsis* une insensibilité à l'éthylène, modifie chez la tomate le temps nécessaire à la maturation des fruits et chez le pétunia la sénescence florale [29]. De plus, les gènes de deux enzymes impliquées dans la production de lipides poly-insaturés ont été identifiés chez *Arabidopsis* et leurs homologues ont ainsi pu être connus chez le soja. L'obtention de variétés transgéniques de soja, pour lesquelles l'expression de l'un de ces gènes a été supprimée, a montré que la teneur en lipides poly-insaturés est passée de 60 % à 2 % au profit de celle des mono-insaturés qui a augmenté de 25 % à 85 % [30]. Ces exemples d'application des recherches menées initialement chez *Arabidopsis* reflètent les retombées attendues sur des plantes d'intérêt agronomique. L'importance de telles retombées dépend de l'effort consenti pour déterminer la fonction des protéines d'*Arabidopsis*. La NSF *(National Science Foundation)* a annoncé un budget de 25 millions de dollars par an, dans le but de réaliser ce travail d'ici 2010 [31]. Les financements n'étant pas extensibles à l'infini, il faudra équilibrer les moyens entre séquencer d'autres génomes de plantes et continuer à étudier la fonction des gènes de la plante modèle *Arabidopsis*. La détermination de la fonction de l'ensemble des gènes chez *Arabidopsis* peut avoir non seulement un impact en ce qui concerne l'agronomie et l'environnement, mais aussi sur la connaissance des autres génomes eucaryotes ■

RÉFÉRENCES

1. Doring HP, Starlinger P. Barbara McClintock's controlling elements : now at the DNA level. *Cell* 1984 ; 39 : 253-9.

2. Fraley RT, Rogers SG, Horsch RB, *et al.* Expression of bacterial genes in plant cells. *Proc Natl Acad Sci USA* 1983 ; 80 : 4803-7.

3. Meinke DW, Cherry JM, Dean C, *et al.* *Arabidopsis thaliana* : a model plant for genome analysis. *Science* 1998 ; 282 : 79-82662.

4. Bevan M, Toto P, Murphy G, *et al.* Objective : the complete sequence of a plant genome. *Plant Cell* 1997 ; 9 : 476-8.

5. Theologis A, Ecker JR, Curtis JP, *et al.* Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana. Nature* 2000 ; 408 : 816-20.

6. Tabata S, Kaneko T, Nakamura Y, *et al.* Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana. Nature* 2000 ; 408 : 823-6.

7. Salanoubat M, Lemcke K, Rieger M, *et al.* Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana. Nature* 2000 ; 408 : 820-2.

8. *The Arabidopsis Genone Initiative.* Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* 2000 ; 408 : 796-815.

9. Myers EW, Sutton GG, Delcher AL, *et al.* A whole-genome assembly of *Drosophila. Science* 2000 ; 287 : 2196-204.

10. Venter JC, Adams MD, Myers EW, *et al.* The sequence of the human genome. *Science* 2001 ; 291 : 1304-51.

11.The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans* : a platform for investigating biology. *Science* 1998 ; 282 : 2012-8.

12. Lin X, Kaul S, Rounsley S, *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana. Nature* 1999 ; 402 : 761-8.

RÉFÉRENCES

13. Vision TJ, Brown DG, Tanksley SD. The origins of genomic duplications in *Arabidopsis*. *Science* 2000 ; 290 : 2114-7.

14. Ewing B, Green P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* 2000 ; 25 : 232-4.

15. Crollius H, Jaillon O, Bernot A, *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat Genet* 2000 ; 25 : 235-8.

16. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001 ; 409 : 860-921.

17. Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. *Genome Res* 1999 ; 9 : 1288-93.

18. Apweiler M, Attwood TK, Bairoch A, *et al.* Interpro. *CCP11 Newsletter* 2000 ; 10 (http://www.HGMP.mrc.ac.uk/CCP11/mewsletter/vol3-4).

19. Morel J, Mourrain P, Beclin C, Vaucheret H. DNA methylation and chromatin structure affect transcriptional and post-transcriptional transgene silencing in *Arabidopsis*. *Curr Biol* 2000 ; 10 : 1591-4.

20. Finnegan EJ, Kovac KA. Plant DNA methyltransferases. *Plant Mol Biol* 2000 ; 43 : 189-201.

21. Byrne ME, Barley R, Curtis M, *et al.* Asymmetric leaves1 mediates leaf patterning and stem cell function in *Arabidopsis*. *Nature* 2000 ; 408 : 967-71.

22. Meyerowitz EM. Plants, animals and the logic of development. *Trends Cell Biol* 1999 ; 9 : M65-8.

23. Unseld M, Marienfeld JR, Brandt P, Brennicke A. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nat Genet* 1997 ; 15 : 57-61.

24. Sato S, Nakamura Y, Kaneko T, *et al.* Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res* 1999 ; 6 : 283-90.

25. Wurgler-Murphy S, Saito H. Two-component signal transducers and MAPK cascades. *Trends Biol Sci* 1997 ; 22 : 172-6.

26. Teem JL, Berger HA, Ostedgaard LS, *et al.* Identification of revertants for the cystic fibrosis delta F508 mutation using STE6-CFTR chimeras in yeast. *Cell* 1993 ; 73 : 335-46.

27. Babcock M, de Silva D, Oaks R, *et al.* Regulation of mitochondrial iron accumulation by Yfh1p, a putative homolog of frataxin. *Science* 1997 ; 276 : 1709-12.

28. Ogura Y, Bonen DK, Inohara N, *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001 ; 411 : 603-6.

29. Wilkinson JQ, Lanahan MB, Clark DG, *et al.* A dominant mutant receptor from *Arabidopsis* confers ethylene insensitivity in heterologous plants. *Nat Biotechnol* 1997 ; 15 : 444-7.

30. Glaser V. EC and US agencies fund large-scale *Arabidopsis* sequencing. *Nat Biotechnol* 1996 ; 14 : 696-7.

31. Somerville C, Dangl J. Plant biology in 2010. *Science* 2000 ; 290 : 2077-8.

32. Doolittle RF, Feng DF, Tsang S, *et al.* Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 1996 ; 271 : 470-7.

Summary

Sequencing of eukaryotic genomes : *Arabidopsis*, the fourth element

Arabidopsis thaliana is an important model for flowering plants. Analysis of the sequence of this genome is giving us access to a considerable quantity of data which will lead to an understanding of plant function, as well as important information about conserved processes in all eucaryotes. In particular, a catalogue of the genes which participate in the life cycle of a plant has been developed. This catalog is probably imperfect, because the genes that it contains are essentially based on predictions, and in the majority of cases, their function is hypothetic at best. Nevertheless, this catalog makes it possible, for the first time, to evaluate the scope of the work that remains to be done in order to comprehend the biological processes in the life of an angiosperm.

TIRÉS À PART

M. Salanoubat.