MEDECINE/SCIENCES 2003; 19: ••••-•••

Science, médiatisation et politique : le séquençage du génome du riz

Michel Delseny

> Près de quatre ans après le lancement d'un consortium public international, l'*International Rice Genome Sequencing Project* (IRGSP), le génome du riz est pratiquement complètement séquencé. Il s'agit du deuxième génome végétal, après celui d'*Arabidopsis thaliana*, et l'on peut espérer qu'il est représentatif des génomes de céréales. Cependant, pas moins de quatre séquences, plus ou moins élaborées, ont été réalisées indépendamment, résultant d'une intense compétition économico-politico-médiatique. On peut s'interroger sur l'efficacité et l'impact de cette façon de conduire des projets de grande envergure. Cet article retrace le déroulement du séquençage du génome du riz et analyse ce que nous commençons à en apprendre. <

UMR 5096 Cnrs-IRD-UP, Génome et Développement des Plantes, Université de Perpignan, 66860 Perpignan Cedex, France.
delseny@univ-perp.fr

Le séquençage des génomes de plante a autant d'intérêt, voire plus, que celui des génomes microbiens ou animaux. En effet, les plantes sont responsables du développement de toute vie sur Terre. Elles assurent la production de l'oxygène indispensable à la respiration et contribuent à la fixation du gaz carbonique de l'air sous forme de matière organique, grâce à la photosynthèse. À côté de ces rôles fondamentaux, elles sont un élément essentiel de la nutrition animale et humaine. L'analyse du génome des plantes est un objectif gigantesque compte tenu de leur diversité et de la taille de leur génome qui dépasse souvent largement celle du génome humain. Après l'achèvement du génome de l'espèce modèle *Arabidopsis thaliana* en 2000, l'année 2002 est l'année du génome du riz.

(➜) m/s 2002, n°5, p. 625

Pourquoi séquencer le génome du riz ?

Si le séquençage des 130 Mpb du génome d'*Arabidopsis thaliana* [1] a servi à démontrer la faisabilité du séquençage d'un génome de plante, il est clair que tous les gènes de plante n'y sont pas représentés et qu'il est nécessaire d'acquérir d'autres informations, en particulier sur les génomes de céréales dont trois espèces, le riz, le blé et le maïs, représentent la fraction majeure de la ration alimentaire humaine et animale. Le riz a le plus petit génome avec « seulement » 430 Mpb, à comparer avec les 2800 du maïs et les 16000 du blé. L'intérêt de disposer d'une séquence complète et précise est évident : c'est le seul moyen de disposer d'un catalogue complet des gènes et de pouvoir les identifier à l'aide de programmes prédictifs. L'ordre précis des gènes est une information essentielle pour les sélectionneurs et, de plus, il est apparemment bien conservé entre différentes céréales. Enfin, le clonage de gènes d'intérêt agronomique repose bien souvent sur des stratégies de clonage positionnel qui requièrent une carte génétique et physique précise.

Dès 1993, un projet de séquençage du génome complet du riz a été lancé au Japon. La première phase visait à mettre en place les outils nécessaires (cartes génétiques et physiques, banques de clones ADNc, YAC, BAC, séquençage d'étiquettes ou EST [*expressed sequence tag*]) et à démarrer le séquençage du génome. Fin 1997, grâce aux travaux réalisés principalement au Japon et aux États-Unis, à l'initiative des Fondations Rockefeller (➜) et Novartis, les principaux outils étaient disponibles. Un consortium public international fut alors constitué pour séquencer complètement le génome du riz, animé par le responsable du programme japonais, Takuji Sasaki, et par Ben Burr, l'un des généticiens moléculaires du maïs les plus respectés aux États-Unis.

Un consortium public international, l'*International Rice Genome Sequencing Programme* (IRGSP)

L'objectif du consortium est de séquencer complètement le génome avec un degré de finition et d'assemblage du même niveau que celui du séquençage d'*Arabidopsis* [2]. Les premières activités de l'IRGSP ont surtout consisté à faire du *lobbying* pour disposer des financements nécessaires et mobiliser des équipes pour réaliser ce projet. Dès le début, la communauté végétaliste française a revendiqué l'étude du chromosome 12, qui comprenait des régions intéressantes pour plusieurs équipes, et a obtenu le soutien de principe du Génoscope pour participer à ce projet.

La stratégie a été de séquencer clone par clone après avoir établi des cartes physiques et un recouvrement minimal des clones BAC à séquencer, permettant ainsi une répartition du travail par chromosome. Le séquençage proprement dit n'a véritablement commencé que fin 1999, après un gros effort de cartographie physique [3, 4] et l'obtention des financements nécessaires. Six et trois des douze chromosomes ont été attribués respectivement au Japon et aux États-Unis. La France, la Chine et Taiwan ont chacun la responsabilité d'un chromosome, alors que la Thaïlande, la Corée et l'Inde ont quelques zones réservées. La séquence établie par l'IRGSP est libre d'accès et mise à jour régulièrement en fonction de l'avancement du projet (http://dna.affrc.go.jp).

Un enjeu économique : l'intérêt des firmes privées

Un premier coup de théâtre est survenu avec l'annonce, en avril 2000, que la société Monsanto avait séquencé la quasi-totalité du génome du riz selon une stratégie proche de celle du consortium et sur la même variété, Nipponbare ! En raison de sa ressemblance avec le génome des autres céréales et de l'importance de sa culture dans le monde (150 millions d'hectares, soit 3 fois la surface de la France), le riz constitue un enjeu économique fantastique : le premier industriel qui identifie et brevette une série de gènes d'intérêt agronomique majeur dans une céréale a de grandes chances d'augmenter ses parts de marché. L'objectif des firmes est alors de repérer au plus vite les gènes intéressants, sans trop se soucier de la qualité ou de la finition du reste des séquences. De fait, malgré l'annonce très médiatisée que 95 % du génome était couvert, la séquence produite par Monsanto s'assemble en un peu plus de 52000 *contigs* (*contiguous clones* : fragments clonés d'ADN se chevauchant) couvrant effectivement 259 Mpb. Fallait-il alors poursuivre sans se soucier de cette annonce ou bien valait-il mieux négocier et tenter d'intégrer les résultats de Monsanto pour accélérer le projet public ?

L'annonce faisait état de la mise en accès des données aux équipes académiques, sous réserve de signature d'engagements très contraignants en termes de propriété industrielle, et de la mise à disposition de l'IRGSP sans cette contrainte, mais avec des modalités pratiques difficilement compatibles avec les objectifs et obligations du consortium [5]. Cette annonce a failli faire éclater l'IRGSP, l'accord avec Monsanto ayant été négocié directement avec les responsables de l'IRGSP sans que les autres membres n'aient été informés ! Un certain nombre de bailleurs de fonds, dont le gouvernement français, ont eu la tentation de renoncer au projet. Fort heureusement, les financements japonais et américains avaient été débloqués et le projet public a démarré. L'accord avec Monsanto a alors été renégocié de telle sorte que les données brutes et les clones soient réellement accessibles.

C'est alors qu'une deuxième compagnie, Syngenta, annonça en février 2001 qu'elle avait fait mieux et obtenu 99 % du génome de la même variété. Cette annonce heureusement n'a pas eu d'effets aussi néfastes que celle de Monsanto. Là encore, on peut mesurer l'écart entre l'annonce médiatique et la réalité puisque la séquence publiée un an plus tard [6] n'est encore organisée qu'en 42000 *contigs* couvrant 389 Mpb du génome. La stratégie utilisée est celle prônée par Craig Venter pour le génome humain : un *shotgun* complet (le génome est découpé en fragments qui sont insérés de façon aléatoire dans un vecteur et séquencés ; les séquences contiguës sont reconstruites par recouvrement). L'équipe de Syngenta a cependant réalisé un ancrage partiel de sa séquence sur la carte génétique. L'accès aux données est également soumis à des accords de confidentialité.

Un enjeu politique

Bien qu'un laboratoire chinois participe à l'IRGSP en séquençant le chromosome 4 de la variété Nipponbare et après avoir entrepris celui du même chromosome d'une variété Guangluai 4, le gouvernement chinois a décidé de financer sa propre entreprise de séquençage complet du génome d'une autre variété chinoise, 93-11, l'un des parents du riz super-hybride. Il faut rappeler ici que l'on distingue deux grands types de riz : les types *indica*, représentés majoritairement en Chine et dans les zones tropicales, et les types *japonica*, présents au Japon et dans les zones plus tempérées (*voir Encadré*). La stratégie utilisée est également un *shotgun* complet et la séquence est publiée simultanément à celle obtenue par Syngenta. Elle est organisée en 127000 *contigs* [7] qui ne sont pas ancrés génétiquement. On ne peut cependant qu'être frappé par la rapidité avec laquelle le projet a été mené : 74 jours pour réaliser la totalité des 4,62 millions de lectures. Cela est

révélateur de l'enjeu politique: faire comprendre aux Japonais et au reste du monde que, sur une espèce aussi stratégique pour le pays que le riz, il fallait désormais compter avec la Chine. Cet aspect politique de s'afficher comme leader de la génomique du riz est confirmé, si besoin était, par la participation active de Jiang Zemin, le Président chinois, à la cérémonie d'ouverture, en présence de nombreux ambassadeurs, de l'*International Rice Genetics Congress* le 16 septembre 2002 à Pékin. D'emblée, la séquence chinoise a été libre d'accès (http://btn.genomics.org.cn/rice).

Une accélération du programme public

La conséquence de ces péripéties a été l'accélération du programme de l'IRGSP qui, fin 2001, et après accord avec Monsanto, s'est fixé comme objectif de produire un brouillon beaucoup plus élaboré dès la fin 2002. Plus récemment, Syngenta a aussi conclu un accord de transfert de ses données, permettant à l'IRGSP de réaliser des économies et de gagner un temps précieux pour la mise à disposition publique d'une séquence de bien meilleure qualité. L'objectif est en passe d'être atteint puisque l'IRGSP vient de publier les séquences quasi complètes des chromosomes 1 et 4 [8, 9] réparties respectivement en 9 et 8 *contigs* et qu'une conférence de presse annonçant le succès du projet a eu lieu le 18 décembre 2002.

Qu'avons-nous appris?

Malgré le caractère encore incomplet des données publiées, plusieurs informations intéressantes sont disponibles et permettent d'avoir une vue générale du génome du riz. Nous avons aussi un aperçu de la tâche qui nous attend pour l'étape de finition de la séquence, et pour l'étape suivante: son analyse fonctionnelle.

Une première observation porte sur le nombre de gènes prédits: l'estimation oscille entre 32000 et 55600 gènes pour les premiers brouillons et entre 57000 et 62500 gènes pour les extrapolations à partir des chromosomes 1 et 4. Les différences entre les fourchettes en disent long tant sur l'intérêt de disposer de séquences précises et correctement assemblées que sur l'état d'annotation de la séquence et l'efficacité toute relative des programmes de prédiction. Cet exercice est encore compliqué par l'observation inattendue d'un gradient dans la composition des exons des gènes de riz, les exons côté 5' étant plus riches en GC que ceux du côté 3'. Il est d'ores et déjà clair qu'un programme exhaustif de séquençage d'ADNc complet sera indispensable pour annoter correctement ce génome.

La comparaison avec les gènes prédits d'*Arabidopsis* montre que 80 à 85 % d'entre eux ont un homologue dans le riz. En revanche, près de la moitié de ceux prédits dans le riz n'ont apparemment pas d'homologue dans *Arabidopsis*, ni avec quelque espèce que ce soit dans les bases de données. Ce résultat, un peu surprenant, s'explique sans doute en partie par la qualité médiocre des brouillons et des annotations, mais indique peut-être aussi l'existence de gènes spécifiques des céréales. Des séquences homologues à 98 % des gènes connus dans le blé, l'orge et le maïs sont retrouvées dans le riz, confirmant l'intérêt du riz pour identifier les gènes des autres céréales. Un peu plus de 8000 gènes communs à *Arabidopsis* et au riz sont clairement absents des génomes de la drosophile, de la levure, du nématode ou des bactéries: ils seraient donc caractéristiques des végétaux.

Du point de vue de l'organisation, il semble bien que le génome du riz soit, comme celui d'*Arabidopsis*, assez largement dupliqué. Cependant, les résultats ne peuvent être analysés de façon exhaustive du fait que la séquence est inachevée et incomplètement assemblée. Contrairement à ce que suggéraient des études antérieures, il apparaît que des gènes de riz sont conservés sur des blocs de quelques centaines de kpb, dans le même ordre que les gènes d'*Arabidopsis* [6, 10]. Dans chacun de ces blocs, de nombreux gènes présents chez *Arabidopsis* semblent avoir été remplacés par d'autres gènes dans le riz, ou inversement. La synténie (conservation, entre deux espèces, de l'organisation des gènes) est bien meilleure lorsque l'on compare le riz avec le maïs, mais n'est sans doute pas aussi parfaite que le laissaient croire les premières publications, et devra être validée chaque fois que l'on voudra l'utiliser. Là encore, une séquence précise est indispensable pour exploiter la synténie.

L'une des caractéristiques des génomes de céréales est leur richesse en séquences répétées. Ces éléments constituent un handicap considérable pour l'assemblage final des séquences déterminées en *shotgun* et elles ont en fait été écartées dans les assemblages très partiels produits par cette méthode. Le groupe de Syngenta indique avoir identifié 38 Mpb de longues séquences répétées et environ 150 Mpb de séquences plus courtes, alors que 104 Mpb de séquences répétées déterminées par les Chinois ont dû être masquées lors de l'assemblage. Au total, 213 familles d'éléments répétés, majoritairement des rétrotransposons, ont été identifiées, mais ce chiffre est sans doute sous-estimé. Le séquençage des chromosomes 1 et 4 révèle aussi l'existence de nombreux gènes présents en copies organisées en tandem, difficiles à assembler correctement après un *shotgun*.

L'existence d'une séquence de référence, celle de Nipponbare, permet d'aligner précisément les séquences des autres variétés et de déceler les polymorphismes. Environ 16 % des deux séquences Nipponbare et 93-11 ne

peuvent être alignés [7], sans doute du fait de l'insertion d'éléments transposables à des sites différents dans chacune des variétés, ce qui est confirmé par l'analyse du chromosome 4 [9]. Sur ce chromosome, la fréquence des *single nucleotide polymorphism* (SNP) entre *japonica* et *indica* est de 1/268 paires de bases. Le caractère incomplet des séquences *shotgun* est souligné par le fait que 22 % des séquences du chromosome 1 de Nipponbare ne sont pas retrouvées dans les assemblages de la variété 93-11.

En conclusion, tout comme l'histoire du séquençage du génome humain, celle du génome du riz illustre à la fois les dangers et les risques que font courir, à la poursuite des programmes publics et à la qualité des résultats, les effets d'annonces d'un certain nombre d'acteurs de la génomique. Malgré ou grâce à ces péripéties, la communauté scientifique disposera d'une séquence de référence de bonne qualité plus tôt que prévu. Elle dispose en outre d'éléments de comparaison entre différentes variétés qui vont grandement faciliter le clonage de gènes d'intérêt. Au total, on doit retenir que la médiatisation et la compétition économique et politique ont accéléré le séquençage du génome du riz. La supériorité du séquençage clone par clone demeure écrasante pour l'interprétation des séquences. Elle ne doit cependant pas cacher l'ampleur de l'effort d'annotation et d'analyse fonctionnelle qui reste à faire. Nous disposons déjà de nombreuses informations et des outils précieux pour analyser d'autres génomes de céréales plus complexes. La comparaison encore limitée avec d'autres génomes, dont celui d'*Arabidopsis*, suscite de nombreuses questions largement insoupçonnées il y a seulement deux ans : quelle est l'étendue des duplications génomiques ? Quand ont-elles eu lieu ? Comment se sont différenciés les gènes apparemment spécifiques des végétaux ? Pourquoi certains gènes présents dans *Arabidopsis* ne le sont pas dans le riz, et inversement ? Quel est l'impact adaptatif de ces quelques gènes ? Des questions dont les réponses devraient apporter des informations tout à fait importantes sur l'évolution des génomes des plantes. ◊

RÉFÉRENCES

1. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000 ; 408 : 796-815.
2. Sasaki T, Burr B. International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr Opin Plant Biol* 2000 ; 3 : 138-41
3. Chen MS, Presting G, Barbazuk WB, *et al*. An integrated physical and genetic map of the rice genome. *Plant Cell* 2002 ; 14 : 537-45.
4. Wu J, Maehara T, Shimokawa T, *et al*. A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* 2002 ; 14 : 525-35.
5. Bary G. The use of the Monsanto draft rice genome sequence in research. *Plant Physiol* 2001 ; 125 : 1164-5.
6. Goff SA, Ricke D, Lan TH, *et al*. A draft sequence of the rice genome (*Oryza sativa* L SSP *japonica*). *Science* 2002 ; 296 : 92-100.
7. Yu J, Hu S, Wang J, *et al*. A draft sequence of the rice genome (*Oryza sativa* L SSP *indica*). *Science* 2002 ; 296 : 79-92.
8. Sasaki T, Matsumoto T, Yashimoto K, *et al*. The genome sequence and structure of rice chromosome 1. *Nature* 2002 ; 420 : 312-6.
9. Feng Q, Zhang Y, Hao P, *et al*. Sequence and analysis of rice chromosome 4. *Nature* 2002 ; 420 : 316-20
10. Salse J, Piegu B, Cooke R, Delseny M. Synteny between *Arabidopsis* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res* 2002 ; 30 : 2316-28.

SUMMARY

Science, communication and policy: sequencing the rice genome

Nearly 4 years after launching the International Rice Genome Sequencing Project (IRGSP), the rice genome sequence is almost completed. This is the second plant genome after *Arabidopsis thaliana* and one expect that it is more representative of other cereal genomes. Indeed, no more than 4 sequences have been independently reported as a result of a tough competition between economy, politics and media. The efficiency and impact of this way of managing a large scale project is questionable. This paper reports the various phases in sequencing rice genome as well as what we start to learn. ◊

TIRÉS À PART

M. Delseny

DE LA DIVERSITÉ À REVENDRE

Il existe deux espèces de riz cultivé, à côté d'une vingtaine d'espèces sauvages apparentées. La culture la plus importante est celle d'*Oryza sativa*, avec 150 millions d'hectares, *Oryza glaberrima* étant limité à l'Afrique de l'ouest. Les riz ont été domestiqués il y a environ 9 000 ans de part et d'autre de l'Himalaya et ont donné plusieurs dizaines de milliers de variétés traditionnelles relativement bien adaptées à des conditions écologiques très différentes. En effet, les zones de culture vont des régions équatoriales aux régions tempérées, s'étageant du niveau de la mer jusqu'à 2 800 mètres d'altitude, poussant dans des milieux arides, en milieu irrigué, ou encore dans des zones inondées. Cette capacité d'adaptation reflète une variabilité génétique importante. Ce grand nombre de variétés est également associé à des qualités culinaires très distinctes, sélectionnées par les populations.

Les variétés de riz se répartissent très schématiquement en deux grandes catégories pauvrement interfertiles : les types *indica*, qui se sont différenciés au sud de l'Himalaya, et les types *japonica*, qui se sont différenciés au nord. Classiquement, les *indica* sont plutôt des riz à grains longs, alors que les *japonica* sont plutôt à grains ronds. Les riz aromatiques (Basmati, Thaï) sont des *indica*. Les *japonica* sont principalement cultivés au Japon. Le choix du consortium IRGSP s'est porté sur la variété Nipponbare, une variété *japonica* élite au Japon dans les années 1980. En revanche, l'immense majorité des Chinois consomme des variétés *indica*, et c'est donc sur une variété de ce type que s'est porté le choix de génotypage dans ce pays.