Comparer les génomes pour identifier les gènes

10



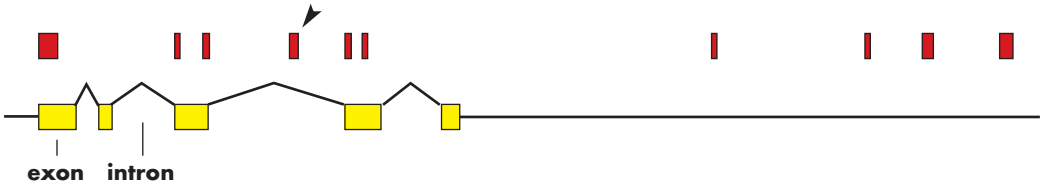




Le séquençage du génome humain, achevé en 2003, s'accompagne d'un travail d'identification des gènes humains qui devra quant à lui se poursuivre durant plusieurs années. On ne dispose pas en effet de programme qui prédise à coup sûr les frontières des gènes et leur structure en introns et exons chez les plantes et les animaux *(voir la fiche* Annotation*)*. On doit du coup faire appel à diverses ressources, telles que les séquences entières ou partielles des produits d'expression des gènes, pour "annoter" le génome *(voir la fiche* ADN complémentaires*)*.

Le repérage des gènes peut aussi grandement bénéficier des comparaisons de séquences d'espèces relativement éloignées. En effet, les séquences d'ADN sans rôle dans le codage des protéines – les introns et les séquences "intergéniques", situées entre les gènes – divergent plus vite au cours de l'évolution que les exons, "contraints" par leur signification fonctionnelle. L'intérêt de la comparaison dépend du degré de parenté des deux espèces comparées. Les mammifères et les poissons, par exemple, ont divergé depuis suffisamment longtemps pour que leurs introns et leurs séquences intergéniques diffèrent presque complètement. Toutefois, ces deux groupes de vertébrés restent suffisamment apparentés pour que leur collection de gènes soit à peu près la même : les séquences de nombreux exons, du fait de leur conservation, peuvent donc être repérées par la comparaison d'un génome de mammifère et d'un génome de poisson.

Si l'on compare à présent deux génomes d'espèces plus proches, par exemple ceux de l'homme et de la souris, on détectera plus d'exons, mais aussi davantage de régions non codantes conservées chez ces deux mammifères *(voir la fiche* Exploiter la séquence III*)*.



La procédure Exofish, utilisée sur la séquence annotée d'un chromosome humain avec les séquences de *Tetraodon*, ne détecte que certains des exons annotés du gène à gauche (les blocs rouges sont les régions conservées entre les deux génomes). En revanche, elle détecte au milieu du gène une séquence supplémentaire, qui est peut être un nouvel exon utilisé de façon alternative *(flèche)*. A droite, Exofish prédit un gène nouveau (ou une extension du gène de gauche) dans une région laissée vierge par l'annotation.

Nous avons mis au point une procédure informatique, nommée Exofish *(Exon FInding by Sequence Homology)*, pour ce type de comparaison. Exofish a d'abord été calibré pour la détection d'exons humains à partir de séquences génomiques du poisson *Tetraodon nigroviridis*. L'intérêt de ce génome compact, huit fois plus petit que le génome humain, est qu'une séquence de longueur donnée contient en moyenne huit fois plus d'exons *(voir la fiche* Tetraodon*)*. Au printemps 2000, nous avions séquencé le tiers du génome de ce poisson, ce qui nous a suffi pour repérer, dans l'ébauche du génome humain, une grande partie des gènes humains par au moins un exon. Nous avons pu de la sorte estimer le nombre de gènes humains à environ 30 000, une valeur bien inférieure à toutes les estimations antérieures. Cette estimation a été confortée depuis lors par les progrès de l'annotation. L'homme ne posséderait donc que le double environ des gènes de la mouche. *(suite au dos)*

Comparer les génomes pour identifier les gènes (suite)

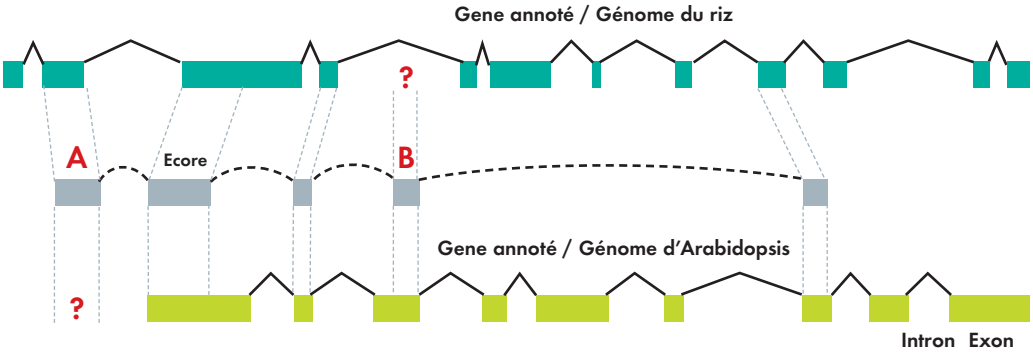
10







Nous avons ensuite appliqué Exofish dans une comparaison entre les séquences de *Tetraodon* et la séquence complète du chromosome 14 humain, achevée au Genoscope à l'automne 2002 ; nous avons ainsi disposé d'une ressource supplémentaire pour l'annotation de ce chromosome *(voir la fiche* Annotation du génome humain*)*. Nous avons aussi utilisé cet outil pour des comparaisons entre d'autres génomes : les séquences génomiques de la mouche drosophile ont par exemple servi à corriger et compléter l'annotation du génome d'un autre insecte diptère, le moustique anophèle *(voir la fiche* Anophèle*)*, et vice versa. De même, des comparaisons ont été effectuées au moyen d'Exofish entre les séquences génomiques du riz et de la plante modèle *Arabidopsis thaliana*, première plante dont le génome a été séquencé. Ces comparaisons servent à corriger l'annotation existante d'*Arabidopsis*, mais aussi à affiner la prédiction des gènes du riz dans l'effort d'annotation en cours sur la séquence génomique de cette céréale *(voir ci-dessous)*. Avec la disponibilité croissante de séquences de grands génomes, cette méthode de génomique comparative est appelée à être utilisée plus largement.



La procédure Exofish est actuellement utilisée au Genoscope pour compléter et corriger l'annotation de la séquence du génome d'*Arabidopsis thaliana* (vert clair) à l'aide des séquences de l'ébauche génomique du riz (vert foncé). Dans cet exemple, on constate que des séquences conservées entre les deux génomes (ecores, en gris) correspondent à certains exons d'un gène annoté chez *Arabidopsis*. La correspondance est d'autant plus significative que les ecores sont colinéaires aux exons. Toutefois, l'une des ecores (A) tombe en dehors du gène d'*Arabidopsis*. La possibilité qu'il s'agisse d'un exon non annoté d'*Arabidopsis* doit être explorée, par exemple à l'aide d'une collection d'ADN complémentaires "pleine longueur" séquencés au Genoscope.

On peut également s'intéresser, dans l'autre sens, à la localisation des ecores dans le génome du riz, actuellement en cours d'annotation. On constate que l'ecore A se trouve dans un exon prédit d'un gène de riz, ce qui rend plus probable qu'elle corresponde à un exon non annoté d'*Arabidopsis*. De façon symétrique, l'ecore B correspond à un exon d'*Arabidopsis*, mais tombe dans la séquence d'un intron chez le riz. Là encore, la possibilité qu'un exon du gène du riz n'ait pas été détecté est hautement vraisemblable.