ADN complémentaires

9

une ressource pour l'annotation



Pour repérer les gènes le long de la séquence du génome, les programmes de prédiction automatique ne suffisent pas. On doit s'assurer que ces prédictions sont complètes et correspondent à de véritables gènes, des "instructions" pour le fonctionnement cellulaire. Ceci passe par l'isolement des produits de l'expression des gènes.

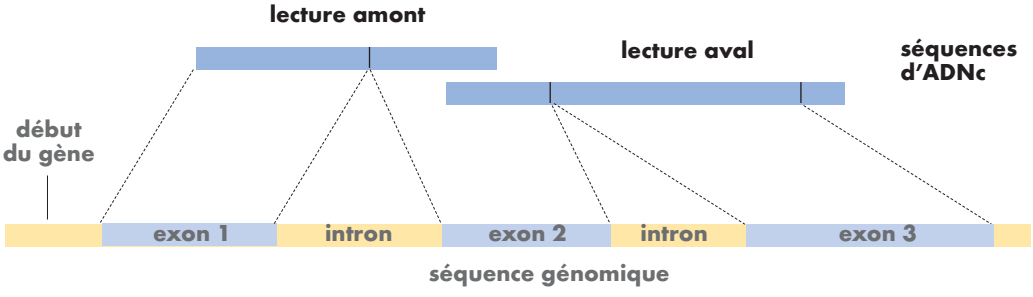
Lorsqu'un gène est actif, sa séquence d'ADN est copiée ("transcrite") en un autre type de molécule, nommé ARN. Les ARN transcrits à partir de la plupart des gènes sont qualifiés de messagers, car ils transportent l'information génétique jusqu'à des structures cellulaires qui la "traduisent" sous la forme de protéines, les véritables agents des fonctions cellulaires.

Chez les animaux et les plantes, les ARN qui viennent d'être transcrits subissent en général une maturation nommée épisssage : les parties biologiquement significatives des gènes, les exons, sont jointes par l'élimination des séquences intercalaires, les introns *(voir la fiche* Interpréter les séquences*)*. Les séquences des ARN messagers constituent donc une ressource idéale pour établir ou valider l'existence d'un gène actif, mais aussi pour préciser sa structure interne, c'est-à-dire les frontières entre introns et exons *(voir la fiche* Annotation du génome humain*)*.

Toutefois, l'ARN est une molécule fragile, que l'on ne sait pas manipuler comme l'ADN. Pour accéder aux séquences des ARN messagers, on procède donc à rebours du processus biologique : grâce à une enzyme trouvée notamment chez les virus, on copie les molécules simple brin d'ARN en molécules double brin d'ADN. Comme la séquence du premier brin d'ADN synthétisé est complémentaire du brin d'ARN messager, ces ADN "rétrotranscrits" sont dits complémentaires.

Depuis plus d'une décennie, de très nombreux ADN complémentaires (ADNc) ont été clonés à partir de tissus variés chez l'homme et les autres organismes faisant l'objet de programmes génomiques. On se contente souvent d'un séquençage partiel de ces clones, aux deux extrémités, pour obtenir des séquences nommées EST ("*expressed sequence tags*"). Plus d'une dizaine de millions sont aujourd'hui disponibles. La comparaison des EST avec la séquence génomique s'est avérée très utile pour découvrir ou confirmer la présence de gènes ou d'exons. Toutefois, les clones d'ADNc séquencés sont souvent incomplets vers l'amont du gène. Du coup, il est fréquent que les modèles de gènes définis sur le génome soient prolongés à la faveur d'une nouvelle annotation, lorsqu'on dispose de la séquence d'un ADNc véritablement complet.

(suite au dos)



Clone d'ADNc "pleine longueur" assemblé apres alignement sur la séquence génomique. La structure du gène correspondant est alors entièrement révélée.

ADN complémentaires (suite)

9



Afin d'améliorer l'annotation des génomes, le Genoscope s'est engagé dans plusieurs programmes de séquençage d'ADNc "pleine longueur", à l'extrémité amont complète. Des collections d'ADNc enrichies en clones pleine longueur ont été construites par la société Life Technologies à partir de différents tissus de l'être humain et de la plante *Arabidopsis thaliana* (collaboration avec l'Unité de Recherche en Génomique Végétale, INRA). D'autres collections d'ADNc pleine longueur ont été construites au Genoscope pour le poisson *Tetraodon nigroviridis* et pour le moustique anophèle, en collaboration avec l'institut Pasteur *(voir les fiches correspondantes)*.

L'intérêt de ces ADNc pleine longueur est qu'ils révèlent la structure complète des gènes, et notamment les régions amont transcrites mais non traduites. Ils révèlent également l'existence d'exons "alternatifs", tantôt présents, tantôt absents dans les ARNm. Le processus par lequel ces exons sont éliminés dans certains transcrits, nommé épissage différentiel, permet d'engendrer différents ARN messagers, et donc différentes protéines, à partir d'un même gène.

Nous avons mis au point au Genoscope une procédure générale d'exploitation des séquences d'ADNc. Concrètement, les séquences lues aux deux extrémités des clones d'ADNc sont comparées à la séquence génomique pour une première localisation. Une prédiction fine des frontières entre introns et exons est ensuite opérée, puis les "modèles de transcrits" couvrant les mêmes régions sont regroupés. Dans le meilleur des cas, les séquences lues en amont et en aval d'un même ADNc sont chevauchantes *(voir la figure au recto)*, et leur assemblage révèle donc la totalité du gène.

Ces ressources d'ADNc ont déjà livré des résultats substantiels quant à l'annotation des génomes de l'homme et de l'arabette. Le séquençage de 90 000 clones d'ADNc humains, issus de 9 tissus, nous a permis de confirmer expérimentalement de nombreux gènes prédits, d'étendre vers l'amont plus de 400 modèles de gènes et de définir plus de 500 nouveaux modèles de gènes. Bon nombre des nouveaux exons sont d'ailleurs confortés par l'existence de régions conservées entre l'homme et *Tetraodon (voir la fiche* Comparer les génomes*)*.

De même, le séquençage de plus de 30 000 clones d'ADNc d'*Arabidopsis* a permis de couvrir par un ADNc pleine longueur près de 2000 gènes prédits dans le génome de cette plante. Ces gènes ont pu ainsi être confirmés ou corrigés, avec l'aide supplémentaire apportée par les comparaisons entre les génomes d'*Arabidopsis* et du riz. Ces différentes ressources ont également révélé 165 nouveaux gènes dans des régions dénuées d'annotations. L'utilisation conjointe des ADNc et des comparaisons entre génomes, pour laquelle le Genoscope a développé une véritable expertise bioinformatique, a prouvé sa validité et devrait être étendue à l'annotation d'autres génomes.