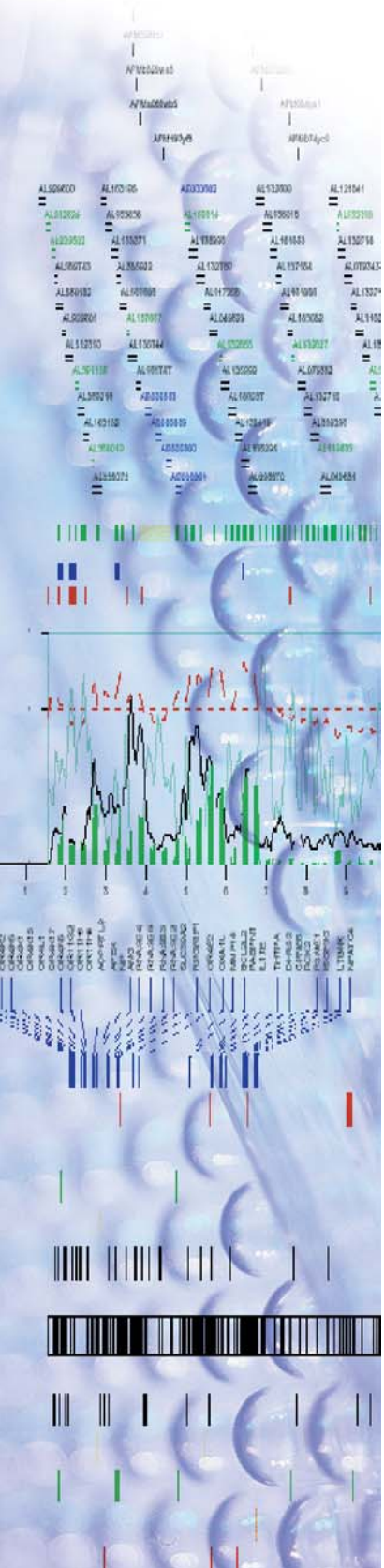
L'annotation du génome humain

8

l'exemple du chromosome 14

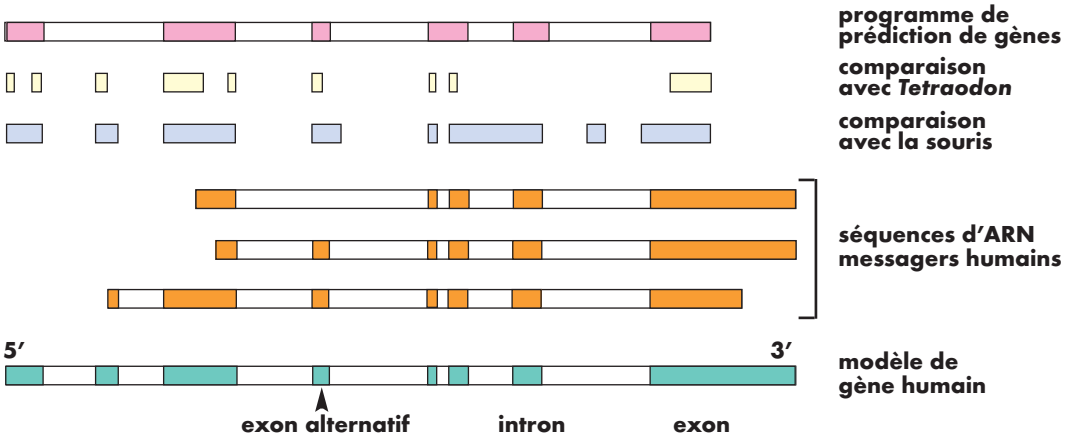


Le séquençage du génome humain a été entrepris en premier lieu pour constituer un inventaire des gènes humains. C'est cet objectif qui justifie les efforts consacrés à l'obtention d'une séquence de qualité, seule à même de révéler la structure des gènes sans ambiguïté. Car la séquence n'est que le point de départ, il faut encore la faire parler, "l'annoter", pour y localiser les gènes. Le Genoscope a effectué ce travail sur la séquence du chromosome 14 humain, le chromosome dont il avait la charge au sein du "Projet Génome Humain".

L'annotation n'a rien de simple, car chez l'homme comme chez les autres mammifères, les gènes occupent moins de 10% de l'ADN. Et encore faut-il compter avec leur morcellement en introns et exons : ces derniers, qui correspondent à la partie biologiquement significative des gènes, représentent moins de 3% du génome et ne sont pas faciles à délimiter *(voir la fiche* Interpréter les séquences*)*. Par exemple, les 24 exons du gène codant la neurexine 3, séparés par de très grands introns, sont dispersés sur près de 1,5 million de nucléotides le long de la séquence du chromosome 14 !

Nous utilisons au Genoscope une procédure en deux temps pour rechercher les gènes. Tout d'abord, nous comparons la séquence du chromosome 14 aux séquences entières ou partielles d'ARN messagers et de protéines appartenant à l'homme ou à d'autres vertébrés, et figurant dans diverses bases de données. Par des procédés informatiques d'alignement des séquences homologues, des limites entre introns et exons sont proposées et des "modèles de gènes" sont définis le long du chromosome.

D'autres informations sont intégrées à ce stade, mais ne servent pas à la définition des modèles : en particulier, l'alignement des séquences génomiques partielles de la souris et du poisson *Tetraodon nigroviridis avec la séquence du chromosome 14 (voir la fiche* Comparer les génomes*)* révèle des régions conservées au cours de l'évolution entre l'homme et ces animaux, indices de régions codantes. On procède aussi à la prédiction des gènes à partir des seules caractéristiques de la séquence, par des moyens informatiques. *(suite au dos)*

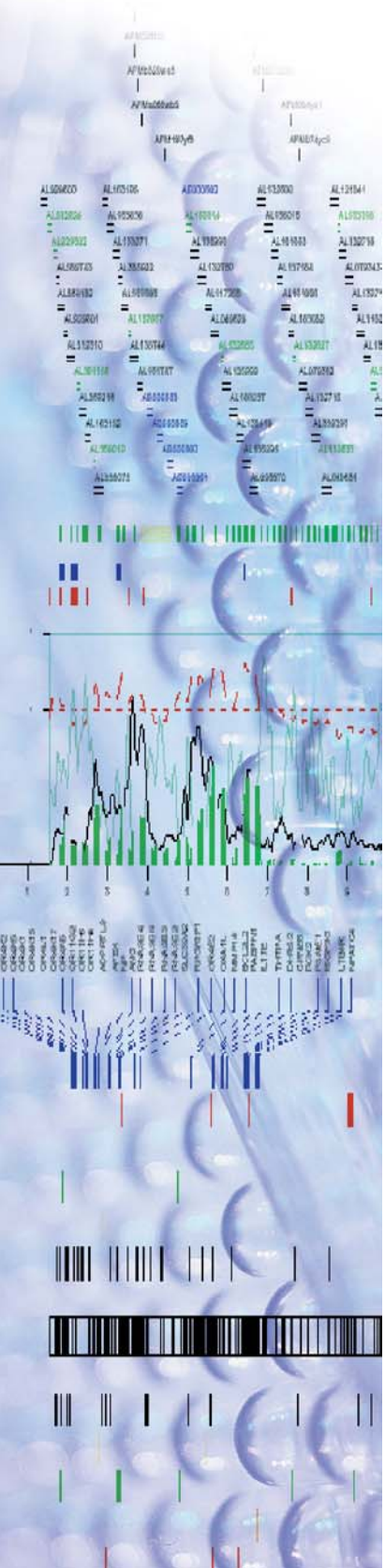


Un premier modèle de gène humain sur la séquence du chromosome 14 est défini à partir de séquences partielles d'ARN messagers, les molécules issues de la transcription du gène. Les comparaisons avec d'autres génomes ainsi que le programme de prédiction de gènes permettent ici d'étendre le modèle de gène vers l'amont, et confortent l'existence d'un exon "alternatif", absent de certains ARN messagers. En revanche, ils ne repèrent pas toujours correctement les frontières intron-exon, peuvent "rater" des exons ou, au contraire, en prédire faussement.

L'annotation du génome humain

8

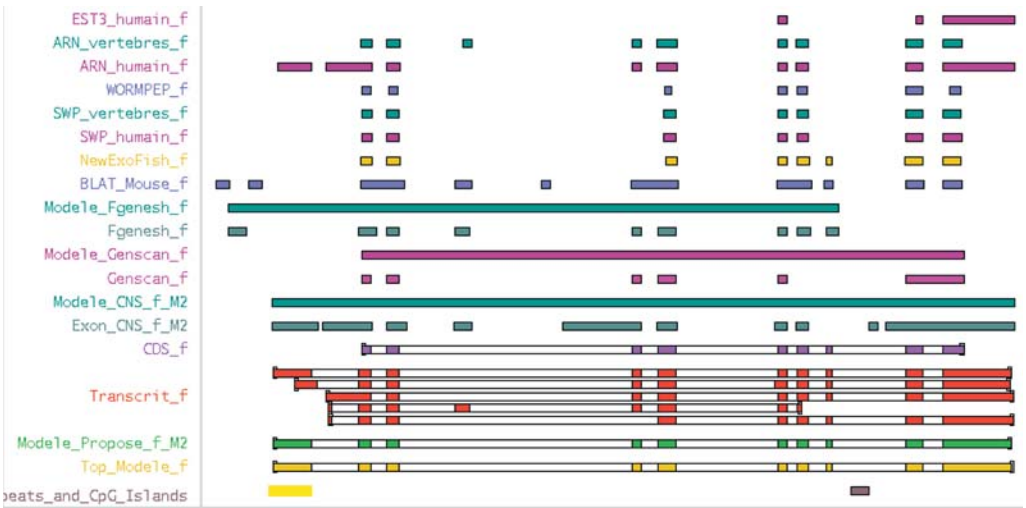
l'exemple du chromosome 14 (suite)



Dans un second temps, les modèles de gènes sont examinés par un annotateur humain qui utilise toutes les données annexes pour les corriger et les valider. L'annotateur peut par exemple décider d'étendre ou de rajouter un exon, de fusionner deux modèles de gènes en un gène unique ou au contraire de scinder un modèle en deux. Selon le nombre d'indices qui viennent l'étayer, et notamment l'identification d'une séquence codante, le modèle est alors classé dans diverses catégories. Ainsi, sur les 850 gènes annotés, 506 étaient déjà connus. Parmi les autres, nouvellement répertoriés sur le chromosome 14, près de la moitié nécessiteront un travail expérimental pour être validés. Enfin, plusieurs centaines de pseudogènes, vestiges non fonctionnels de gènes, ont été identifiés.

L'annotation de la séquence du génome humain aura de nombreuses retombées dans les décennies à venir. Les plus importantes seront d'ordre scientifique et médical, et elles-mêmes donneront lieu à toute une série de nouvelles applications. Mais de nombreuses années de recherche seront nécessaires avant que l'on engrange ces fruits de la séquence. Il est toutefois un domaine où le bénéfice est immédiat, celui de la recherche des gènes impliqués dans les maladies dites "héréditaires" ou "génétiques". Très souvent, des études génétiques ont permis de localiser le gène responsable dans un intervalle sur un chromosome. Grâce à la séquence annotée, on peut aujourd'hui se reporter directement à l'inventaire des gènes compris dans l'intervalle en question, et retenir ceux qui ont le plus de chances d'être impliqués dans la pathologie. Il ne reste alors plus qu'à rechercher des mutations dans ces gènes "candidats" chez les sujets malades *(voir la fiche* Exploiter la séquence du génome I*)*.

Ainsi, la séquence annotée du chromosome 14 a d'ores et déjà permis l'association de six gènes à des maladies génétiques, dont une forme de paraplégie spastique. Des dizaines d'autres devraient suivre sur le chromosome 14, et plusieurs milliers sur le génome entier. La connaissance de ces gènes permettra de mettre au point des tests diagnostic à partir de l'ADN. Pour les maladies les plus graves, le diagnostic génétique peut être pratiqué avant la naissance dans les familles à risque. L'identification du gène responsable d'une maladie peut aussi permettre de comprendre le mécanisme physiologique de son apparition et donc, dans certains cas, d'explorer de nouvelles possibilités thérapeutiques.



Un gène en cours d'annotation par le Genoscope sur le chromosome 14.