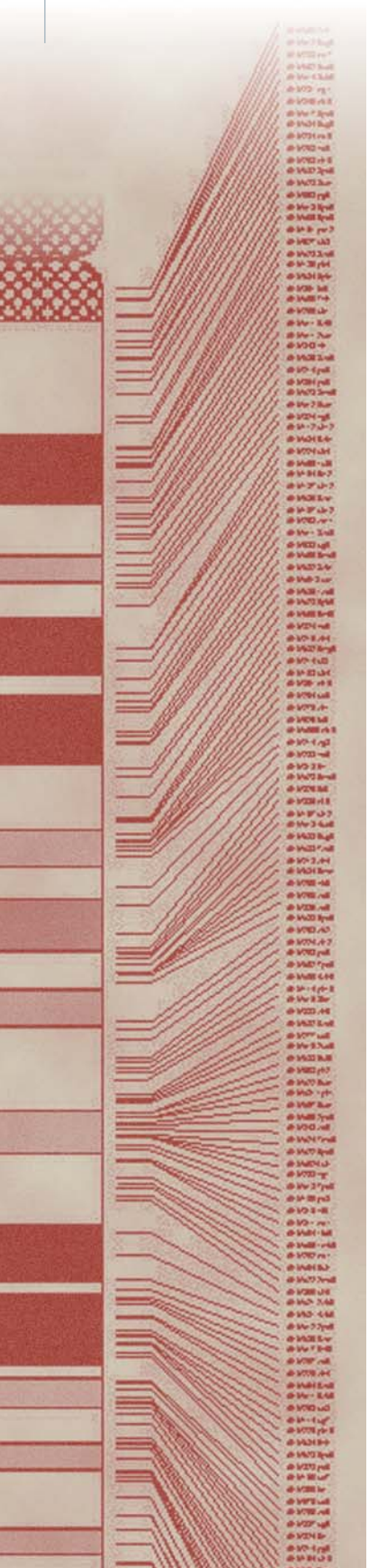
Exploiter la séquence du génome

III. Découvrir les séquences régulatrices



Depuis la fin du projet génome humain, les chercheurs travaillent à parfaire l'annotation de la séquence du génome humain *(voir la fiche correspondante)* : les gènes humains – tout au moins ceux qui codent des protéines – sont presque tous repérés et de mieux en mieux délimités, grâce aux comparaisons de séquences et aux ressources d'ADN complémentaires, mais aussi aux caractéristiques de la séquence génomique elle-même. Toutefois, l'inventaire des gènes humains, aussi complet soit-il, ne peut suffire à expliquer le fonctionnement des cellules et de l'organisme. Il est également très important d'élucider la façon concertée dont les gènes s'expriment, selon les circonstances, le stade de développement et le tissu considéré.

La transcription d'un gène en ARN messager - son "expression" - est sous le contrôle de séquences régulatrices variées, qui sont beaucoup plus difficiles à identifier que les gènes eux-mêmes. Nous ne connaissons en effet à ce jour aucune caractéristique qui permette de les identifier à partir de la seule séquence génomique, sans données expérimentales. Qui plus est, nous ne savons même pas où les chercher. Si les promoteurs, courtes séquences qui initient la transcription, se trouvent tout près de la partie amont des gènes, d'autres séquences régulatrices, comme les activateurs et les répresseurs, peuvent se trouver à une distance variable en amont ou en aval du gène ; certaines peuvent même être distantes de plus d'une centaine de milliers de paires de bases du gène régulé, et se nicher dans un intron d'un autre gène !

Caractériser les séquences régulatrices est donc une tâche difficile, mais indispensable pour comprendre le développement et le fonctionnement des organismes : ces séquences déterminent en effet l'intensité, le moment et le lieu de l'expression du gène, et la précision d'un tel contrôle spatial et temporel est cruciale au cours du développement ; de même, on a mis en évidence, pour nombre de fonctions cellulaires, l'orchestration précise de l'expression de nombreux gènes. Ceux-ci interagissent au sein de réseaux dont on commence seulement à percevoir la complexité. L'enjeu est ici de comprendre comment l'organisme humain se bâtit et fonctionne avec "seulement" 25 000 gènes.

C'est d'ailleurs au niveau de la régulation de l'expression des gènes qu'il faut rechercher la source de nos différences avec nos plus proches parents dans le monde animal. Nous partageons près de 99% de la séquence de notre génome avec le chimpanzé, et la majorité des protéines codées par nos gènes sont pratiquement identiques à celles de ce grand singe. Des différences aussi importantes que le développement du cortex cérébral chez l'être humain pourraient résulter de changements mineurs dans la séquence codante ou dans l'expression de quelques gènes : la modification d'un unique facteur de transcription – une protéine capable de se fixer à une séquence régulatrice – peut en effet modifier le profil d'expression de centaines de gènes "en aval".

Enfin, l'identification des séquences régulatrices revêt une grande importance en génétique médicale. On pense en effet qu'une grande partie des allèles qui confèrent une susceptibilité accrue aux maladies communes *(voir la fiche* Exploiter le génome II*)* ne correspondent pas à des changements de séquence d'une protéine : il s'agirait plutôt de polymorphismes situés en dehors de la séquence codante des gènes. Une telle variation, survenant dans une région régulatrice d'un gène, modifierait par exemple le dosage de la protéine codée par ce gène, avec pour conséquence de perturber la fonction accomplie par la protéine.

(suite au dos)

Exploiter la séquence du génome

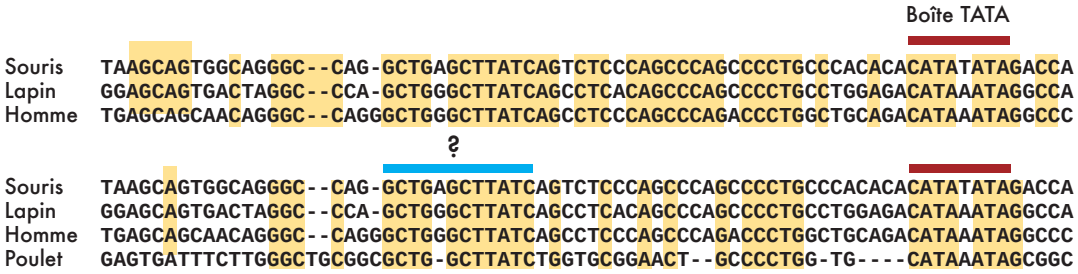
III. Découvrir les séquences régulatrices (suite)

Malheureusement, ces variations fonctionnelles ne sont pas faciles à distinguer des nombreux autres polymorphismes présents dans les régions non codantes : la mesure d'une association avec la maladie peut simplement signifier que le polymorphisme étudié est en déséquilibre de liaison avec le véritable facteur de risque *(voir de nouveau la fiche II)*. Là encore, la caractérisation des régions régulatrices faciliterait l'identification des variants fonctionnels.

Plusieurs méthodes expérimentales sont employées pour rechercher ces régions régulatrices. Aujourd'hui que l'on dispose de séquences complètes pour les génomes de l'homme et de plusieurs autres organismes, on peut également recourir à des stratégies bioinformatiques de plus en plus efficaces. La principale consiste à comparer la séquence génomique de l'homme à celle(s) d'une ou de plusieurs autres espèces, afin d'identifier les régions conservées en dehors des gènes. Au cours de l'évolution, ces régions auraient moins divergé que les séquences alentour, du fait de leur importance fonctionnelle dans la régulation des gènes ou bien dans l'organisation de l'ADN au sein du noyau de la cellule.

Les premières comparaisons effectuées entre l'homme et la souris ont été très encourageantes : elles ont permis de retrouver des séquences régulatrices déjà connues, et d'en découvrir de nouvelles. La difficulté est qu'une fraction importante des séquences non codantes conservées entre deux espèces aussi proches que l'homme et la souris n'a pas de signification fonctionnelle. Une solution pour identifier les séquences régulatrices sans ambiguïté est d'ajouter d'autres espèces dans la comparaison. Il peut s'agir d'autres mammifères, car une grande part des éléments de régulation est conservée chez l'ensemble des mammifères. Les chercheurs souhaitent toutefois disposer de la séquence génomique d'espèces apparentées à l'homme à des degrés plus divers *(voir la figure ci-dessous)*.

Un travail expérimental peut alors être entrepris pour confirmer que les séquences ainsi identifiées sont bien des séquences régulatrices. Une autre source d'information est l'étude de l'expression d'un grand nombre de gènes d'un même génome. De telles études sont aujourd'hui réalisables grâce aux "puces à ADN" : il s'agit de minuscules surfaces où sont échantillonnées les séquences des gènes identifiés lors de programmes génome. Ces puces servent à mesurer simultanément l'abondance relative de milliers d'ARN messagers dans différents échantillons. Certains gènes sont alors regroupés sur la base de profils d'expression similaires, et l'on peut rechercher, dans les séquences en amont de ces gènes co-régulés, la présence d'un ou de plusieurs motifs communs, cibles possibles d'un même facteur de transcription. On constitue ainsi une collection de motifs, utile pour identifier d'autres séquences régulatrices.



La comparaison des séquences d'une région en amont du gène ApoAI chez l'homme et deux autres mammifères ne livre pas beaucoup d'informations, du fait d'une trop grande conservation. En ajoutant le poulet, on rend les zones conservées moins nombreuses mais plus significatives, et l'on fait apparaître, en plus de la boite TATA (un élément du promoteur), un site possible de régulation. (d'après *Nat. Rev. Genet.* 2, 100)