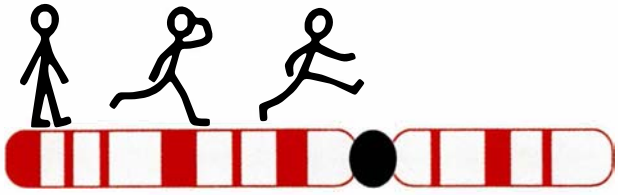
m/s

Chroniques génomiques



par Bertrand JORDAN

Le festival des ADNc

Un vieux débat

Depuis qu'il est question de séquencer l'ADN humain, la discussion fait rage : faut-il déchiffrer l'ADN génomique « tout venant » (qui d'après les estimations généralement admises, contient seulement quelques pour cent de séquences codantes) ou s'attaquer en priorité aux seuls gènes, représentés par les clones d'ADN complémentaires préparés à partir de l'ARN messager extrait de cellules ou de tissus ? Nous avons plusieurs fois évoqué ce déjà vieux débat, et présenté dans ces colonnes, il y a quelques mois, le « deuxième souffle » du séquençage génomique [1]. L'approche par l'ADNc a, elle aussi, beaucoup de partisans et fait l'objet de travaux menés avec des moyens notables. Plusieurs publications récentes dans *Nature Genetics* — qui s'impose actuellement comme la meilleure revue semi-spécialisée du domaine — nous donnent l'occasion de faire le point. Notre tour d'horizon englobera également les premiers articles sur ce sujet parus depuis l'été 1991, ainsi que les données plus récentes glanées à l'occasion du congrès *Human Genome*, à Nice, en octobre dernier, ou lors d'échanges avec les différents protagonistes de ce nouveau champ de recherches.

La construction même des banques d'ADNc — si elle est bien faite — assure que l'on va déchiffrer des séquences codantes et non des introns ou des séquences intergéniques. Aussi intéressants que soient ces derniers pour les partisans des isochores, voire les passionnés des fractales et de la « génétique numérique », leur signification biologique est obscure. Avantage, donc, aux stragégies fondées sur le séquençage d'ADNc. Mais le caractère à la fois partiel et redondant de la représentation des gènes dans toute banque d'ADNc est immédiatement venu tempérer l'enthousiasme des chercheurs. On sait en effet que chaque cellule n'exprime qu'une partie des cinquante ou cent mille gènes que renferme notre patrimoine génétique : la banque d'ADNc correspondante est donc forcément incomplète. De plus, les travaux de cinétique de réassociation de l'équipe de John Bishop, encore cités aujourd'hui bien que remontant à 1974 [2], ont clairement mis en évidence trois classes d'abondance parmi les ARN messagers d'une cellule. Les messagers très abondants comprennent un petit nombre d'espèces moléculaires dont chacune existe à des milliers ou des dizaines de milliers d'exemplaires par cellule ; ils constituent 10 à 20 % de la masse totale de l'ARN messager. Viennent ensuite les ARNs moyennement abondants, quelques centaines d'espèces différentes présentes, chacune, à plusieurs centaines d'exemplaires par cellule et représentant 30 à 40 % du total. Le reste correspond aux messagers rares, qui regroupent une ou même plusieurs dizaines de milliers d'espèces ; on en trouve une ou deux molécules par cellule, parfois même moins. Ces chiffres, obtenus par des méthodes qui paraissent aujourd'hui bien grossières, ont été pour l'essentiel confirmés par les études plus fines menées depuis le début de l'ère du clonage et restent la base de tout raisonnement.

Une approche systématique par les ADNc, visant à dresser par ce biais un inventaire des gènes humains, suppose l'analyse de très nombreux clones — pris au hasard dans une banque construite avec soin. L'optique est différente de celle du scientifique intéressé par une protéine particulière et qui en cherche le gène : il s'agit ici, au contraire, d'en examiner le plus possible. L'écueil qui menace, dès lors, le séquenceur d'ADNc est celui de retomber fréquemment sur les mêmes clones, ceux qui proviennent des messagers abondants et donc très représentés dans la banque. Son inventaire va, du coup, se limiter aux gènes les plus exprimés dans le tissu duquel il est parti — gènes qui sont souvent déjà connus : du fait même de leur forte expression, ils sont plus facilement repérables et clonables que ceux dont le messager est rare. C'est ainsi que le groupe de Peter Schimmel au MIT, qui s'était lancé dans cette aventure au début des années 1980, affichait un bilan assez décevant [3]. Ces chercheurs avaient établi une banque de muscle squelettique de lapin et séquencé plus de 150 clones tirés de cette collection : ils y trouvèrent, certes, les clones correspondant à 13 des 19 protéines musculaires connues, mais ne découvrirent aucun gène jusque-là inconnu... Cela a sans doute dissuadé beaucoup d'équipes de s'engager dans cette voie — avant que les premiers résultats du groupe de Venter ne remettent à la mode ce genre d'approche.

La fin des années 1980 : programmes nationaux, progrès de la « normalisation »

Le Programme Génome aux États-Unis, officiellement inauguré fin 1990, avait, en réalité, déjà débuté sous les auspices du DOE, du NIH et du HHMI, dès 1987-1988. Il ne comportait pas, avant 1991, de volet « ADNc », centré qu'il était sur les

cartes génétiques et physiques, sur l'amélioration des technologies et sur le lancement de quelques projets de séquençage génomique : à l'époque, Craig Venter proposait le séquençage de la bande q28 du chromosome X. Les programmes nationaux annoncés ou mis en route par plusieurs « petits » pays — Grande-Bretagne, Japon ou France — faisaient, eux, une part importante à ce type de recherche. Cela traduisait sans doute le désir d'occuper un « créneau » peu investi aux États-Unis, et l'impression que ce travail pouvait être effectué sans mettre en œuvre de très gros moyens. Des méthodes nouvellement élaborées semblaient aussi rendre ce travail plus fructueux. Nous faisons ici allusion à l'égalisation ou « normalisation » des banques d'ADNc — procédé dont l'objectif est d'arriver à ce que les différentes espèces moléculaires soient présentes à des fréquences comparables dans une banque donnée.

Le principe de ces méthodes est assez simple. On part d'une banque d'ADNc non normalisée, dont on prépare en masse les segments insérés. Ce mélange est dénaturé, puis placé dans des conditions où la réassociation des brins complémentaires peut avoir lieu. Bien entendu, les espèces moléculaires abondantes, représentées à de très nombreux exemplaires dans le mélange de départ, se réassocient rapidement puisque la probabilité de rencontre des deux brins est élevée en raison de leur forte concentration. Les espèces rares, elles, se réassocient beaucoup plus lentement. De sorte que, après un certain temps de réassociation, les espèces abondantes sont majoritairement converties en ADN bicaténaire, au contraire des espèces rares. Un passage sur colonne d'hydroxy-apatite (qui sépare ADN simple brin et double brin) donne alors un mélange de fragments d'ADN monocaténaires enrichi en séquences rares — ou, plus précisément, dans lequel l'abondance des séquences fréquentes a été diminuée.

La mise en œuvre de ce procédé est laborieuse. Les cycles de réassociations éliminent — c'est leur but — la majeure partie (en masse) de l'ADNc, amenant le chercheur à travailler par la suite avec des quantités d'ADN infinitésimales dont le maniement est délicat. C'est, en fait, l'inclusion d'une ou plusieurs étapes d'amplification par PCR dans les protocoles qui a permis d'aboutir à la construction de banques normalisées représentatives — publiées à peu près simultanément par une équipe japonaise [4] et par le laboratoire de Sherman Weissman à Yale (CO, USA) [5]. L'existence de ces banques, et la perspective de pouvoir appliquer les protocoles de « normalisation » ainsi définis à toute banque particulière, ont certainement rendu l'option ADNc plus attractive pour les décideurs scientifiques. Mais, curieusement, comme nous allons le voir, les travaux réalisés jusqu'ici n'ont guère fait appel à la normalisation. L'emploi de banques classiques, dans lesquelles l'abondance des clones reflète, plus ou moins, celle des molécules de messager dans la cellule de départ, s'avère très possible ; il est même préférable dans certains cas...

Les EST (ou « signatures », ou encore, « étiquettes ») entrent en scène

C'est des États-Unis qu'allait venir le premier résultat marquant. Ce fut une surprise : les travaux du groupe de Craig Venter, bien que conduits dans le cadre du NIH à Bethesda (MD, USA) n'étaient pas financés par le programme génome ; C. Venter lui-même n'était guère en odeur de sainteté auprès de Jim Watson, son directeur prestigieux — mais un peu caractériel... Et la controverse déclenchée par la demande de brevets aussitôt déposée par Venter et Reid Adler devait attirer l'attention sur ces études, bien au-delà des cercles scientifiques directement concernés. Nous n'évoquerons pas aujourd'hui cette controverse, déjà abondamment commentée.

Le premier article du groupe de Venter, paru en juin 1991, montrait l'efficacité de la méthode choisie [6]. Il s'agissait tout simplement de prendre, au hasard, des clones dans une banque d'ADNc (provenant du cerveau humain en l'occurrence), et de faire, sur chacun de ces clones, « un coup » de séquence : une seule lecture sur un séquenceur de type *Applied biosystems*, donnant deux cents à trois cents nucléotides de séquences fiables à 97 ou 99 %. Cette information apparemment fragmentaire se révèle extrêmement utile. Comparée aux séquences enregistrées dans les bases de données, elle permet d'établir si l'on est en présence d'un gène déjà connu, ou d'une nouvelle entité. Elle autorise aussi la définition d'amorces PCR pour l'isolement du gène correspondant sous forme de cosmide ou de YAC, ainsi que la détermination du chromosome d'origine grâce à un « panel » d'hybrides monochromosomiques. L'on a ainsi obtenu un STS *(sequence tagged site)* repérant une séquence exprimée : c'est ainsi que Venter a baptisé ces entités (EST, *expressed sequence tags*). Enfin — du moins dans certains cas — la traduction en acides aminés du court segment d'ADN lu peut donner quelques idées sur la protéine. Ces données sont, de plus, obtenues à un coût très raisonnable : les divers laboratoires pratiquant en grand cette méthode avancent un prix de revient de l'ordre de dix dollars par étiquette. Rappelons que le séquençage complet d'un ADNc de deux mille bases, au tarif optimiste d'un dollar la base, coûte deux cent fois plus... C'est ainsi qu'en quelques mois de travail, sans faire appel à des moyens techniques lourds, l'équipe avait séquencé plus de six cents clones pris au hasard dans une banque d'ADNc de cerveau. Trois cent trente sept de ces derniers étaient des entités jusque-là inconnues. Ce n'était pourtant que le galop d'essai puisque, un peu plus de six mois plus tard, la même équipe présentait deux mille trois cent soixante-quinze nouvelles séquences obtenues selon le même schéma opératoire [7]. On pouvait dès lors prévoir que cette connaissance — certes schématique et partielle — allait, à bref délai, pouvoir s'étendre à une fraction significative de nos cinquante ou cent mille gènes. D'autant que Craig Venter n'était pas seul à suivre cette voie...

Les « étiqueteurs » américains...

L'acteur principal, au Nouveau Monde, est naturellement Craig Venter. Comme il eut l'occasion de

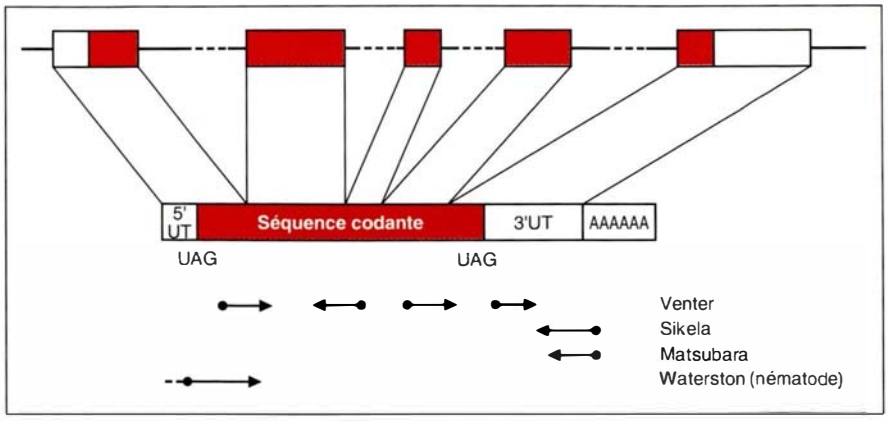


Figure 1. ***Les options de séquençage d'ADNc.*** *On a figuré un gène (en haut), l'ARN messager qui en dérive et la position des séquences effectuées par différentes équipes, celles de Craig Venter [6, 7], de James Sikela [10], Kenichi Matsubara [11] et Robert Waterson [14].*

l'expliquer lors du *Human Genome Meeting* de Nice en octobre 1992, il est venu à l'ADNc par déception devant le peu de rendement (pour l'ADN humain) du séquençage génomique : ses articles récemment parus sur le sujet [8] soulignent en effet les difficultés de l'identification des exons, et ne montrent, après 100 kilobases de séquence, que cinq gènes dont deux déjà connus... Au contraire, un travail équivalent fournit des centaines, sinon des milliers, d'*expressed sequence tags* (EST). Sur le plan technique, il a choisi, dans un premier temps, l'emploi de banques d'ADNc commerciales : elles se sont révélées de qualité médiocre, ce qui a entraîné un changement de tactique. Qu'elles soient commerciales ou « maison », les banques proviennent d'un amorçage au hasard sur l'ARN messager, c'est-à-dire que l'endroit séquencé est placé de façon aléatoire dans la séquence — le plus souvent codante en raison de la prépondérance de cette dernière dans l'ARN messager *(m/s n° 9, vol. 8, p. 966) (figure 1)*. Ce pari augmente la probabilité d'obtenir une information sur la structure de la protéine correspondante ; en revanche, elle peut aboutir à séquencer plusieurs morceaux du même ADNc, clonés dans différents plasmides — sans que l'on ne s'en aperçoive tant que ces séquences partielles ne se recouvrent pas. Dans une étude effectuée avec le groupe de Michael Polymeropoulos [9], une petite cinquantaine de ces EST ont été localisés sur le génome humain. Localisés est d'ailleurs un terme un peu fort, puisque, en fait, il ne s'agit que d'une « assignation » chromosomique, réalisée par réaction PCR (amorces définies à partir des séquences déjà déterminées) sur un jeu de cellules hybrides homme-souris, ou homme-hamster, contenant chacune un seul chromosome humain. L'information ainsi obtenue est par trop imprécise pour être utile en elle-même, mais elle permet d'entamer ensuite la localisation fine de chaque clone sur un *panel* spécifique du chromosome auquel il appartient.

Une autre équipe d'outre-Atlantique, celle de James Sikela [10], a publié des travaux similaires menés d'une façon un peu différente : ces auteurs ont choisi d'obtenir la séquence de l'extrémité 3' de l'ARN messager, en amorçant la synthèse de l'ADNc par de l'oligo dT et en le clonant dans un vecteur directionnel *(figure 1)*. Dans ce cas, les séquences portent presque toujours sur la région 3' non codante de l'ARN messager, et sont directement comparables, tant à l'intérieur du laboratoire qu'avec les autres groupes ayant pris la même option. De plus, la grande divergence de ces séquences non codantes entre l'homme et la souris, et leur fréquent polymorphisme, facilitent leur localisation ultérieure précise sur le génome. L'équipe a ainsi séquencé un peu plus de mille clones, dont plus de neuf cents semblent correspondre à des gènes jusque-là inconnus : il faut dire qu'il avait été procédé à une étape de « précriblage » pour éliminer les clones les plus abondants, parmi lesquels se retrouve la plus forte proportion de gènes déjà connus. Comme pour l'équipe de Venter, la localisation chromosomique est une étape limitante : une vingtaine seulement des clones séquencés a été localisée...

... l'approche japonaise...

Le programme Génome du ministère de l'Éducation Japonais (le « Monbusho ») incluait dès 1990 une composante ADNc notable. Je la vis en œuvre dans le laboratoire de Kenichi Matsubara au printemps 1991. C'était un petit projet, mené à l'époque par deux ou trois jeunes chercheurs et dont la conception m'avait fort intéressé ; les premiers résultats ont paru récemment dans *Nature Genetics* [11].

L'originalité de cette équipe est d'employer le séquençage, non seulement pour découvrir de nouveaux gènes, mais aussi pour évaluer le spectre d'expression d'un tissu donné. A cet effet, la banque est établie de façon à refléter le plus fidèlement possible la composition de l'ARN messager du tissu — en clonant volontairement des segments assez courts (deux à trois cents nucléotides) à partir de l'extrémité poly A du messager. Les clones sont alors pris au hasard et séquencés selon les modalités maintenant classiques du *cycle sequencing* suivi d'un passage sur séquenceur *Applied*. L'équipe répertorie alors les séquences « redondantes » (trouvées plus d'une fois) et celles qui sont « solitaires », les compare aux banques de données et effectue quelques contrôles pour vérifier que les valeurs de fréquence trouvées lors de l'analyse des clones reflètent bien la réalité au niveau de l'ARN messager...

Sur 982 clones ainsi séquencés, 468 appartiennent à cette catégorie des « solitaires » ; c'est là que l'on trouve les quatre cinquièmes des séquences nouvelles. Les clones redondants, eux, constituent la moitié de l'échantillon mais ne renferment que 173 espèces distinctes, dont trois très

fréquentes. La majorité de ces espèces est, elle aussi, nouvelle (135 sur 173). On voit que si l'on cherche des gènes encore inconnus, il vaut mieux sélectionner des séquences peu exprimées, peu représentées dans les banques : c'était d'ailleurs assez évident *a priori*, mais les chiffres précisent ce point. Finalement, l'on retrouve les trois classes de fréquences décrites jadis par l'équipe de Bishop : trois espèces constituent à elles seules 10 % du total des clones (donc, toutes choses égales par ailleurs, 10 % de la masse de l'ARN messager cellulaire) ; elles représentent, dans cette cellule, la classe des messagers « très abondants ». Les 170 autres séquences redondantes forment la classe « moyennement abondante » : 40 % du total ; les solitaires correspondent, quant à eux, aux messagers rares. Bien entendu, la frontière entre « moyennement abondant » et « rare » est floue et arbitraire : si l'équipe avait séquencé 2 000 ou 5 000 clones au lieu de 1 000, certains des solitaires seraient devenus redondants ! Les auteurs explorent d'ailleurs cette question en prenant quelques clones solitaires et en les utilisant comme sonde sur 8 800 ou 26 400 clones de la même banque. Dans cette expérience, certains des clones testés ne sont pas retrouvés : ils sont réellement rares. D'autres sont rencontrés de une à cinq fois, ce qui permet d'estimer leur abondance.

L'examen des clones reliés à des gènes déjà connus, une petite centaine d'espèces représentée par 214 clones, permet d'aller plus loin : dans la mesure où la fonction est, dans ce cas, connue ou soupçonnée, ces données définissent une sorte de carte transcriptionnelle *(figure 2)* de la cellule étudiée, qui est une lignée hépatocytaire. On voit ainsi apparaître les gènes impliqués dans la synthèse des protéines, une douzaine d'espèces constituant un quart des clones, les gènes liés à diverses autres fonctions dans le cytoplasme et les organites (une quarantaine d'espèces, un quart des clones également), et les protéines impliquées dans la sécrétion : un troisième quart des clones, contenant une quinzaine d'espèces distinctes. Le quatrième quart, quant à lui, se répartit entre protéines nucléaires, de membrane, du cytosquelette... Ce spectre, bien que partiel puisque la plupart des clones ne sont pas identifiés — donc pas rattachés à une fonction — donne néanmoins une bonne idée de ce qui se passe dans la cellule étudiée. Il n'est pas très surprenant qu'une cellule d'origine hépatique synthétise beaucoup de protéines sécrétées, mais il est intéressant de constater que près de 30 % des messagers moyennement abondants sont des transcrits spécifiques du foie. La spécificité tissulaire est donc nette : les partisans du criblage différentiel ou des banques soustraites, souvent attaqués sous prétexte que tous les gènes seraient plus ou moins transcrits dans tous les tissus, en tireront quelque réconfort. Quant à la forte présence des facteurs de synthèse protéique (entre autres, le facteur d'élongation 1 α, qui, avec 22 clones, remporte la palme de la séquence la plus représentée), elle est liée au fait que la cellule de départ est une lignée établie, et n'a pas été retrouvée dans une banque construite directement à partir de foie lors de travaux plus récents de la même équipe.

L'idée directrice de cet exercice, dont les premiers résultats sont prometteurs, est de répéter la détermination pour chacun des 200 types cellulaires qui suffiraient à constituer notre corps selon l'ouvrage classique de Bruce Alberts : « *Molecular Biology of the Cell.* » Cela devrait fournir un ensemble de spectres de transcription, dont l'interprétation deviendra de plus en plus riche au fur et à mesure que plus de séquences seront connues et que les gènes correspondants auront été étudiés. Les 200 000 séquences ainsi déterminées donneront à la fois une moisson de gènes et des informations biologiques précises. Notons qu'à 10 dollars le « coup de séquence », le montant total du projet s'établit à quelques millions de dollars, un chiffre raisonnable comparé au séquençage d'un seul chromosome humain qui, en l'état actuel des techniques, reviendrait à une ou plusieurs centaines de millions. Et un tel exercice sur les ADNc se prête aisément à un fonctionnement relativement décentralisé et collaboratif impliquant de nombreux laboratoires ; il peut ainsi jouer un rôle d'animation utile dans le monde japonais de la recherche.

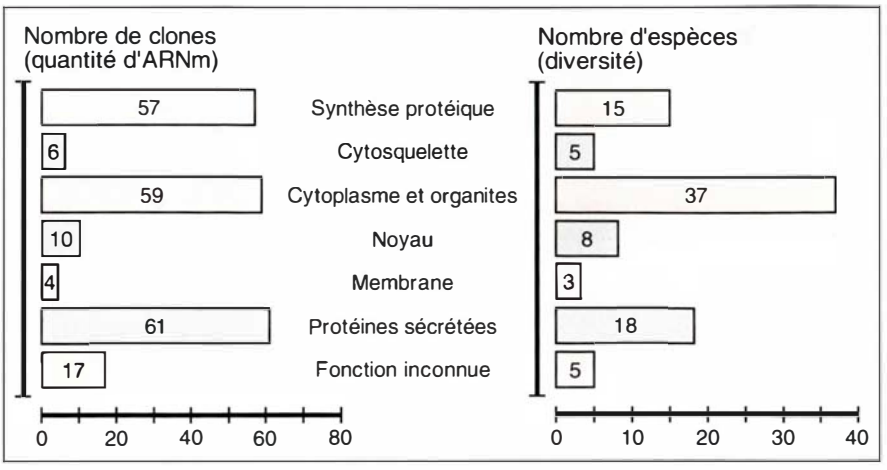


Figure 2. ***Spectre d'expression d'une lignée hépatique.*** *Certains des résultats publiés par le groupe de Kenichi Matsubara [11] sont ici représentés sous forme graphique : à gauche, le nombre de clones (sur les 214 identifiés) appartenant aux diverses catégories, et à droite, le nombre d'espèces différentes pour chacune d'elles. Soulignons que, par la force des choses, cette statistique ne porte que sur les séquences correspondant à des gènes déjà identifiés, et que ce spectre est donc susceptible de se modifier fortement dans le futur.*

... et les autres

Les travaux rapportés ci-dessus ne rendent pas compte de l'ensemble des entreprises de séquençage systémati-

que d'ADNc en cours de par le monde. Mentionnons le petit groupe des ADNc qui fonctionne dans le cadre du « Centre de ressources » du programme Génome britannique. Il a à son actif plus d'un millier de séquences nouvelles, bloquées pendant quelque temps par le dépôt de demandes de brevet effectuées à titre conservatoire par le *Medical Research Council* — à mon avis bien mal inspiré... Plus près de nous, le projet « Genexpress » de Charles Auffray (CNRS, Villejuif et Généthon, Évry, France), a produit huit mille séquences dont plus de trois mille sont uniques et nouvelles, et ont été enregistrées dans les banques de données EMBL et *Genbank*. Ce dépôt a été entouré d'une certaine solennité, et a donné lieu à une cérémonie, à l'UNESCO, destinée à frapper les esprits et à marquer l'opposition de ses protagonistes aux procédures de brevetage. D'autres programmes en cours, en Europe et ailleurs, permettent d'estimer que l'on dispose déjà de plus de vingt mille EST humains — nombre respectable, si on le rapproche du total estimé de nos gènes, cinquante à cent mille, et du nombre de gènes humains clonés à la fin des années 1980, moins de deux mille. Encore que ces divers chiffres ne soient pas tout à fait comparables...

Combien d'EST, combien de gènes ?

Il est instructif d'examiner le résultat des « croisements » auxquels se sont livrés certaines de ces équipes. Le groupe de Matsubara, par exemple, a comparé ses cinq cents « nouvelles » séquences aux deux mille six cents de Venter : vingt-trois seulement sont communes. Faut-il en déduire que les deux sous-ensembles (les banques d'ADNc employées) se recouvrent très peu, ou qu'elles correspondent à l'échantillonnage d'un capital de plusieurs centaines de milliers de gènes différents ? Non, car comme nous l'avons noté, les séquences de Matsubara proviennent en principe de l'extrémité 3' de l'ARN messager, alors que celles de Venter sont prises au hasard le long de la molécule *(figure 1)*. Le recouvrement réel pourrait aller jusqu'à vingt ou trente pour cent ; la comparaison des séquences de Sikela (elles aussi en 3') avec celles du groupe japonais sera de ce point de vue révélatrice. Bref, les données n'imposent pas pour le moment une remise en cause du chiffre généralement admis de cinquante à cent mille gènes chez l'homme ; et leur accumulation indique bien qu'une fraction substantielle de cet ensemble sera prochainement « étiquetée ».

Il serait naturellement souhaitable que, pour aider aux comparaisons, les laboratoires s'accordent sur la région à séquencer. Ce n'est pas le cas actuellement ; la zone qui présenterait le maximum d'avantages est la région 5' du messager puisque, compte tenu de la faible longueur, quand elles existent, des séquences 5' non codantes, elle donnerait des informations sur la structure de la protéine *(figure 1)*. Mais cela suppose des banques d'ADNc *full length*, dans laquelle chaque molécule de messager serait représentée par un transcrit inverse complet — ce qui est encore difficilement faisable aujourd'hui. On peut penser que les travaux d'étiquetage stimuleront le séquençage complet de clones d'ADNc, ce qui facilitera les recoupements nécessaires. En tout état de cause, l'archivage de ces séquences partielles doit être rapide afin de permettre les comparaisons entre laboratoires. Cet archivage est bel et bien réalisé pour les séquences de Venter : le droit des brevets américains — au contraire du droit français — autorise en effet la publication des résultats avant obtention des brevets. Mais le délai entre l'obtention d'une « signature » et son archivage est encore trop long, certaines informations recueillies par les auteurs (comme les homologies avec d'autres gènes qu'ils ont décelées en interrogeant les banques de données) sont perdues dans des bases comme EMBL, et tout le système doit être amélioré pour faciliter son accès.

La plus-value de la localisation chromosomique et de la génétique

Comme nous l'avons signalé, la localisation chromosomique des « étiquettes » ne suit pas, et de loin, leur obtention. Il est en effet relativement aisé de charger les produits de réaction de trente clones sur un appareil *Applied Biosystems* et d'en tirer trente séquences ; en revanche, la réalisation d'un nombre équivalent de *Southern blots* ou, pire, d'hybridations *in situ* représente un travail beaucoup plus lourd. Au surplus, la sensibilité de la méthode FISH *(fluorescent in situ hybridization)* est encore très « limite » pour des petites sondes. Pourtant, l'information de localisation chromosomique est capitale, et à plus d'un titre. Tout d'abord, c'est elle qui transforme l'EST en un repère qui peut servir à l'intégration des cartes physiques, génétiques et transcriptionnelles. Mais, de plus, la connaissance de la zone chromosomique d'où provient un EST, jointe à un minimum d'informations sur le tissu dans lequel il est exprimé, permet d'intéressants croisements avec les données de la génétique clinique. N'oublions pas que l'homme est l'organisme dont la pathologie est la mieux étudiée, que trois ou quatre mille maladies génétiques sont connues, que plusieurs centaines d'entre elles sont localisées sans que pour autant le gène en cause ait été isolé. Parmi les EST nouvellement déterminés, certains vont fatalement se situer dans le segment chromosomique où « doit » — d'après l'étude génétique — se trouver le gène responsable de telle ou telle maladie. Cela ne prouve pas que l'EST corresponde à ce gène ; mais il devient à tout le moins un candidat sérieux.

A l'heure actuelle, les méthodes qui permettent de placer des ADNc sur le génome restent lourdes et lentes. L'emploi d'hybrides somatiques ne définit que le chromosome d'origine, et coûte dix fois plus cher — d'après les estimations de Craig Venter — que l'obtention de la séquence ; l'hybridation *in situ* de ces petites sondes réclame une adaptation de la technique pour presque chaque clone, contrairement à la localisation de cosmides ou de YAC qui fonctionne en routine, dans des conditions définies une fois pour toutes. L'avenir, en fait, est aux « filtres polytènes »*, qui exploitent avec élégance les travaux de cartographie physique dans lesquels s'est récemment illustrée l'équipe de Daniel Cohen [12, 13] *(m/s n° 8, vol. 8, p. 881)*. Le concept

vient du nématode, dont la carte physique est pour l'essentiel terminée. Le groupe de John Sulston a pu ainsi réaliser des filtres sur lesquels sont disposés un peu plus de 900 YAC représentant l'ensemble des cent mégabases de ce génome. Puisque la carte physique, et donc la position de chacun des YAC sont connues, ces derniers ont pu être déposés sur le filtre dans l'ordre dans lequel ils se trouvent sur la carte des six chromosomes de *C. elegans*. Ainsi la simple hybridation d'une sonde — un ADNc par exemple — sur ce filtre va normalement révéler deux ou trois points adjacents : ce seront les deux ou trois YAC (recouvrants) qui contiennent la séquence correspondante. La position de ces points détermine le chromosome et la position de la séquence sur ce dernier, à une ou deux centaines de kilobases près. Elle indique aussi quels clones le chercheur doit demander à Sulston s'il souhaite examiner l'environnement de son ADNc ou rechercher d'autres gènes à proximité... Les projets de type EST sur le nématode font naturellement bon usage de cette possibilié. C'est ainsi que le groupe de Waterston a récemment positionné 670 ADNc : 606 n'ont posé aucun problème, les autres présentant des séquences répétées qui compliquent quelque peu l'opération [14].

Le chromosome 21 humain a été complétement cartographié, l'ensemble de notre génome est en passe de l'être ; rien ne s'oppose en principe à ce que des « filtres polytènes » humains soient préparés. Le projet « Genexpress » devrait logiquement en bénéficier, et ainsi intégrer ses ADNc dans la carte physique. Cela ne sera pas immédiat, car il reste à définir la position exacte des *contigs* obtenus par l'équipe de D. Cohen, à établir un « jeu minimum » de YAC (inutile d'en mettre trente mille sur les filtres, cinq ou six mille devraient suffire)... et à installer la logistique de fabrication et de distribution de ces filtres qui seront à coup sûr très demandés ! Mais existe-t-il une meilleure manière de montrer l'utilité des approches génomiques lourdes, et de répondre aux critiques qui prétendent que ces travaux ne présentent pas d'intérêt biologique ? ■

* *Le nom de « filtre polytène » fait référence à la drosophile. Dans cet organisme, les chromosomes polytènes des glandes salivaires ont depuis très longtemps facilité la localisation de gènes par hybridation* in situ : *ils sont en effet constitués d'environ 4 000 copies d'ADN répliqué sur place, les signaux obtenus sont donc très nets et la localisation aisée — comme avec les filtres auxquels, par analogie, John Sulston a donné leur nom.*

Bertrand R. Jordan
Directeur de recherche au Cnrs, responsable du groupe de génétique moléculaire humaine. CIML, Inserm/Cnrs, case 906, 13288 Marseille Cedex 9, France.

SOCIÉTÉ DE BIOLOGIE

Séance du 17 mars 1993

Morphogenèse et gènes régulateurs du développement

Françoise Dieterlen *(Institut d'Embryologie cellulaire et moléculaire, Nogent-sur-Marne)*
Introduction

Muriel Umbhauer *(Université P. et M. Curie, Paris)*
Régionalisation de l'expression de la ténascine en réponse aux inducteurs du mésoderme

Hubert Condamine *(Institut Pasteur, Paris)*
Homéogènes et développement du poisson zèbre

Patrick Charnay *(École Normale Supérieure, Paris)*
Analyse moléculaire de la segmentation du cerveau postérieur

Anne-Hélène Monsoro *(Institut d'Embryologie cellulaire et moléculaire, Nogent-sur-Marne)*
Quox 8, un homéogène impliqué dans l'établissement de l'organisation dorso-ventrale des vertébrés

Thierry Jaffredo *(Institut d'Embryologie cellulaire et moléculaire, Nogent-sur-Marne)*
Intégration et expression de vecteurs rétroviraux de type ALV chez l'embryon d'oiseau : implications pour le lignage cellulaire et le transfert de gènes *in vivo*

Marie-Aimée Teillet *(Institut d'Embryologie cellulaire et moléculaire, Nogent-sur-Marne)*
Transfert d'une épilepsie d'origine génétique par greffe embryonnaire de certaines vésicules cérébrales chez le poulet

La séance aura lieu à 16 h 30, au Collège de France, 11, place M. Berthelot, 75005 Paris (Salle 8)

RÉFÉRENCES

1. Jordan BR. Séquençage génomique : le deuxième souffle. *médecine/sciences* 1992 ; 8 : 854-7.

2. Bishop JO, Morton JG, Rosbach M, *et al.* Three abundance classes in HeLa cell messenger RNA. *Nature* 1974 ; 250 : 199-204.

3. Putney SD, Herlihy WC, Schimmel P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* 1983 ; 302 : 718-21.

4. Ko MSH. An « equalized cDNA library » by the reassociation of short double-stranded cDNAs. *Nucleic Acids Res* 1990 ; 18 : 5705-11.

5. Patanjali SR, Parimoo S, Weissman SM. Construction of a uniform-abundance (normalized) cDNA library. *Proc Natl Acad Sci USA* 1991 ; 88 : 1943-7.

6. Adams MD, Kelley JM, Gocayne JD, *et al.* Complementary DNA sequencing : expressed sequence tags and Human Genome project. *Science* 1991 ; 252 : 1651-6.

7. Adams MD, Dubnick M, Kerlavage A, *et al.* Sequence identification of 2,375 human brain genes. *Nature* 1992 ; 355 : 632-4.

8. Martin Gallardo A, McCombie WR, Gocayne JD, *et al.* Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nature Genet* 1992 ; 1 : 34-9.

9. Polymeropoulos MH, Xiao H, Glodek A, *et l.* Chromosomal assignement of 46 brain cDNAs. *Genomics* 1992 ; 12 : 492-6.

10. Khan AS, Wilcox AS, Polymeropoulos MH, *et al.* Single pass sequencing and physical and genetic mapping of human brain cDNAS. *Nature Genet* 1992 ; 2: 180-5.

11. Okubo K, Hori N, Matoba R, *et al.* Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet* 1992 ; 2 : 173-9.

12. Chumakov I, Rigault P, Guillou S, *et al.* A continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* 1992 ; 359 : 380-7.

13. Bellanne-Chantelot C, Lacroix B, Ougen P, *et al.* Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992 ; 70 . 1059-68.

14. Waterston R, Martin C, Craxton M, *et al.* A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genet* 1992 ; 1 : 114-23.

TIRÉS A PART

B.R. Jordan