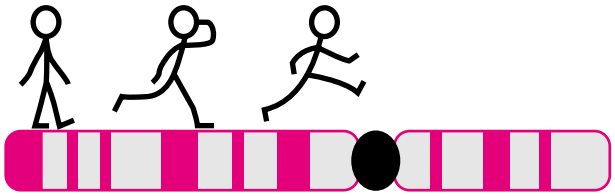
m/s

Chroniques génomiques



par Bertrand JORDAN

médecine/sciences 2001 ; 17 : 81-4

ADNc : les incontournables

Au tout début des années 1990, l'on peinait à séquencer une ou deux centaines de kilobases d'ADN humain, et les résultats scientifiques de tels travaux apparaissaient assez minces par rapport aux efforts et aux fonds investis [1].

Les débuts des EST

L'option des ADNc, leur déchiffrage partiel mais massif apparurent vite comme une alternative réaliste au séquençage intégral. Lancée en franc-tireur par Craig Venter [2], très largement médiatisée par le scandale que soulevèrent les tentatives de brevets sur ces séquences partielles, l'approche des EST *(expressed sequence tags)* allait jouer un rôle central dans l'exploration de notre génome tout comme dans la mise au point de produits nouveaux. L'accumulation de ces étiquettes fut entreprise dans le secteur public, les données étant déposées au fur et à mesure dans la base *ad hoc* dbEST (http://www.ncbi.nlm.nih.gov/dbET/). Elle s'effectua aussi au sein de groupes privés comme *Human Genome Sciences* ou *Incyte*, qui gardaient ces résultats pour eux ou en vendaient l'accès à des industriels de la pharmacie. D'autres entreprises, notamment Merck, choisirent de collaborer avec les laboratoires académiques et de financer l'obtention de données publiques. Le nombre d'EST répertoriés augmentait rapidement, atteignant (pour dbEST) 325 000 en janvier 1996 et 550 000 un an plus tard.

Redondant par principe (puisque l'on détermine les séquences partielles de clones pris au hasard dans des banques d'ADNc), cet ensemble était analysé par des systèmes dénommés *gene index** comparant toutes les séquences afin de les regrouper en *clusters* censés représenter chacun un transcrit : c'est à partir de ces données qu'avait été faite l'estimation d'environ 100 000 gènes humains aujourd'hui très discutée [3]. L'utilité des EST fut encore renforcée par la localisation d'un grand nombre d'entre eux sur notre génome : effectuée massivement grâce à l'emploi des hybrides d'irradiation, celle-ci devait aboutir en 1998 au positionnement de plus de 30 000 étiquettes [4]. Ces EST « pré-positionnés » devenaient ainsi une source gratuite de gènes-candidats pour les projets de génétique humaine, une fois passée l'indispensable étape de la localisation.

Le grand retour du séquençage

Pendant ce temps les techniques de séquençage s'amélioraient progressivement. Des réactifs plus stables, des « séquenceurs » plus performants, une exploitation informatique plus sophistiquée, et surtout une prise de conscience de la nécessité de planifier une opération de « Très Grand Séquençage » comme une entreprise industrielle, aboutissaient, malgré l'absence d'une révolution technique majeure, à rendre la lecture de mégabases d'ADN possible et presque abordable. Ces progrès devinrent évidents pour tous avec l'obtention en 1996 de la séquence complète de la levure. Les 13 mégabases déchiffrées constituaient de loin le plus important ensemble jamais obtenu, et montraient que de tels projets étaient devenus viables ; et l'utilité de ces données était attestée par la découverte de près de 3 000 gènes « nouveaux » (en plus des 3 000 déjà répertoriés) chez cet organisme que l'on croyait pourtant très bien connaître après des lustres d'études biochimiques et génétiques. La démonstration était faite que seul le séquençage intégral permettait d'accéder à l'ensemble des données géniques. Parallèlement, l'avancée rapide du séquençage du nématode (100 mégabases) donnait confiance dans notre capacité à passer une étape de plus et à aborder les 3 000 mégabases de l'ADN humain.

C'est ainsi que fut lancé à partir de 1997 le déchiffrage intégral de notre ADN. Réalisé principalement en Grande-Bretagne et aux États-Unis (avec une participation minoritaire de la France, du Japon et de l'Allemagne), aiguillonné par la concurrence avec l'initiative privée de Craig Venter et de l'entreprise *Celera*, il

** Il existe aujourd'hui plusieurs de ces systèmes, en général accessibles sur le réseau. En voici une liste :*
UNIGENE : http://www.ncbi.nlm.nih.gov/UniGene/index.html
TIGR : http://www.tigr.org/tdb/tgi.shtml
IMAGENE : http://image.llnl.gov/image/imagne/current/bin/search
STACK : http://www.sanbi.ac.za/Dbases.html
GENEXPRESS : http://idefix.upr420.vjf.cnrs.fr/IMAGE/Genexpress.html

donnait lieu en juin 2000 à l'annonce conjointe de la première séquence « brouillon » de notre génome. Annonce plus politique que scientifique, et séquence dont les critères de qualité sont assez mal définis (à l'exception des chromosomes 21 et 22 qui, eux, sont de qualité « finie ») ; il n'en reste pas moins qu'environ 90 % de notre génome est aujourd'hui déchiffré avec un taux d'erreur d'une fraction de pour cent, et que toutes ces données sont accessibles à quiconque dispose d'un accès Internet. Avancée considérable, dont on pourrait attendre une explosion de nouveaux résultats et la résolution de maintes controverses. Le moins que l'on puisse dire est que ce n'est pas toujours le cas, comme le montre l'incertitude actuelle sur le nombre de gènes humains dont l'estimation varie de moins de 30 000 à plus de 120 000 [3]...

Les EST continuent

C'est que l'interprétation de la séquence, et notamment la détection des gènes, posent des problèmes redoutables dans le cas de l'homme (et des mammifères en général). La complexité des structures géniques, la multiplicité d'exons souvent très petits, l'existence de nombreux pseudogènes, le caractère relativement flou des signaux d'épissage et encore plus des promoteurs... tout cela rend actuellement impossible une interprétation *a priori* fiable des données humaines, même lorsqu'il s'agit de séquence « finie » : les difficultés d'annotation des chromosomes 21 et 22 le montrent très clairement [5, 6]. Du coup, loin d'être devenue caduque, l'analyse des EST et plus généralement l'étude des ADNc continue de plus belle*. En témoigne notamment la progression du nombre de séquences nouvelles dans dbEST : 3 millions entre 1999 et 2000 (dont 1 million de séquences humaines), bien plus qu'au cours des cinq premières années de cette base de données. En fait, les informations obtenues grâce aux ADNc jouent actuellement un rôle indispensable dans l'annotation de la séquence humaine : c'est bien souvent en comparant les EST et la séquence génomique que l'on repère des exons, d'autant plus que de telles comparaisons s'accommodent assez bien de séquences de qualité « brouillon ». De plus, l'analyse des ADNc met aujourd'hui en évidence des phénomènes susceptibles de modifier profondément la compréhension de notre matériel génétique et même celle de l'évolution.

De très nombreuses séquences partielles d'ADNc continuent à être obtenues dans le cadre de projets EST, en ciblant le plus possible des tissus spécialisés dont les gènes spécifiquement exprimés n'ont pas encore été échantillonnés. De plus, un effort général est effectué pour obtenir et déchiffrer des ADNc complets *(full length)* représentant l'intégralité du transcrit. Un point général de ces travaux a été récemment effectué lors du colloque « Transcriptome 2000 » tenu à l'Institut Pasteur début novembre 2000**. Il a fourni la plupart des informations résumées dans cette chronique – qui n'est pas pour autant un compte rendu de cette réunion, au cours de laquelle bien d'autres thèmes ont été abordés.

La nouvelle vague des projets EST humains a débuté dès 1997, avec le programme CGAP du *National Cancer Institute* dont l'objectif était d'explorer les tissus tumoraux en obtenant des banques d'ADNc à partir de tumeurs ou de fragments microdisséqués. La fraction de séquences « nouvelles » (non homologues à des séquences déjà contenues dans dbEST) s'est avérée importante dans ces données, et a permis d'accroître très nettement la représentativité de dbEST. Ce type d'effort est poursuivi, tant dans le cadre de CGAP que pour différents tissus spécialisés. L'équipe de Bento Soares (Université d'Iowa, États-Unis), par exemple, effectue des soustractions successives de banques d'ADNc afin d'augmenter la proportion de séquences nouvelles. Et, bien que cet article soit centré sur les travaux portant sur l'homme, n'oublions pas les nombreux projets menés sur d'autres organismes pour lesquels les EST sont souvent la principale information génomique actuellement disponible...

La quête de l'ADNc *full length*

Depuis quelque temps, l'obtention de jeux importants d'ADNc complets, et leur séquençage, sont devenus le but de nombreuses équipes. La plupart des clones d'ADNc à partir desquels des EST ont été déterminés sont en effet courts et peu représentatifs de l'ARN messager dont ils dérivent. La majeure partie des banques séquencées ont été construites par l'équipe de Bento Soares, en utilisant un amorçage sur la queue polyA de l'ARN messager, suivi de deux traitements d'égalisation pour augmenter la proportion de séquences peu exprimées (parmi lesquelles se trouve la plus forte proportion de séquences « nouvelles »). Dans ces conditions la taille moyenne des *inserts*, qui se situent tous à l'extrémité 3' du transcrit, est de l'ordre de la kilobase. Or beaucoup d'ARN messagers ont une longueur de plusieurs milliers de bases, et la région 3' non codante mesure souvent plus d'une kilobase : on pourra s'en persuader en examinant un jeu pris au hasard de grands ADNc humains sur la base de données du *Kazusa Sequencing Institute* (Japon) (http://zearth.kazusa.or.jp/huge/). Les EST 3' et 5' obtenus à partir d'un tel clone peuvent donc ne contenir aucune séquence codante. Ils ne révèlent alors rien sur la nature de la protéine codée par le gène correspondant, et ne permettent aucune prédiction fonctionnelle. Pendant assez longtemps, la démarche généralement suivie a consisté à obtenir, « à l'unité », le clone d'ADNc complet à partir d'un EST jugé particulièrement intéressant en raison de

** En tout état de cause, les EST sont aujourd'hui le réactif presque obligatoire pour la construction de réseaux d'ADN sous forme de* macroarrays *ou de* microarrays. *Je ne discute pas ici cet aspect qui est très important, même si l'on peut penser que les réseaux seront à l'avenir fondés de plus en plus sur des oligonucléotides de synthèse [7].*

*** Transcriptome 2000, organisé par Charles Auffray, Bento Soares et Sumio Sugano à Institut Pasteur du 6 au 9 novembre 2000. Voir le site correspondant http://www.vjf.cnrs.fr/transcriptome/*

son profil d'expression et/ou de sa localisation chromosomique. Cela était réalisé en criblant des banques d'ADNc spécialisées, et en pratiquant différentes manœuvres d'extension (5' RACE, par exemple) à partir du clone existant et de l'ARNm d'un tissu judicieusement choisi.

La nouvelle tendance, en cours depuis quelque temps déjà, consiste à s'efforcer d'obtenir des banques contenant une forte proportion d'ADNc complets, puis à identifier les clones correspondant à ce critère grâce à un séquençage partiel, et enfin à déchiffrer intégralement celles des séquences qui sont à la fois *full length* et nouvelles. Les méthodes employées sont variées. Certaines équipes effectuent une sélection sur la taille de l'ARNm et/ou sur celle de l'ADNc après rétrotranscription (Bento Soares, Université d'Iowa, États-Unis; Omahu Ohara, *Kazusa DNA Research Institute*, Japon; Stefan Wiemann, *Deutsche Krebs Forschungs Zenter*, Heidelberg, Allemagne; Robert Strausberg, *National Cancer Institute*, Bethesda, États-Unis). D'autres utilisent la «coiffe» présente à l'extrémité 5' de l'ARN messager pour isoler les ARNm complets par un système de capture *(cap trapping)* biotine/avidine (Judy Margolin, Baylor, États-Unis; June Kawai, *Riken*, Tsukuba, Japon) ou pour procéder à la ligation en 5' d'un oligonucléotide qui se retrouve ensuite dans l'ADNc (Sumo Sugano, Université de Tokyo, *Nedo cDNA project* [MITI] Japon). Notre Génoscope mène également un travail de ce type en utilisant des banques d'ADNc produites par l'entreprise *Lifetech*.

Les équipes effectuent ensuite une séquence en 5' et en 3' afin de déterminer si l'ADNc est complet ou presque (au minimum, présence d'un ATG dans un contexte de séquence approprié), et s'il est nouveau (en tant qu'ADNc complet), avant de procéder à sa séquence complète. Le rendement dépend beaucoup des projets et des techniques: de quelques pour cent à la moitié de clones obtenus s'avèrent être complets. Il n'est d'ailleurs pas toujours évident de le déterminer, le critère de l'ATG et de son environnement n'étant pas très restrictif; par ailleurs, la lecture révèle souvent des amorçages internes qui, lors du clonage, ont eu lieu sur des régions internes riches en A et non sur la région polyA 3' terminale. Enfin le débrouillage des différents clones d'ADNc correspondant aux épissages alternatifs *(voir plus loin)* demande beaucoup de temps. Les différents projets annoncent avoir obtenu de quelques centaines à près de 20 000 séquences de clones *full length*. Il existe certainement de nombreux doublons dans ces travaux, d'autant plus que les séquences sont transmises aux bases de données avec un certain retard, et en tous cas après la fin du séquençage complet d'une série de clones. Il n'en reste pas moins que l'ensemble de ces travaux devrait fournir rapidement la séquence complète des transcrits pour les 20 000 à 25 000 gènes humains les plus approchables (parce que exprimés dans des tissus accessibles).

Généralité de l'épissage alternatif

La fréquence et la complexité des phénomènes d'épissage alternatif ressortent de manière évidente à l'examen de ces résultats. Ils se présentent sous de nombreuses formes: variation de la limite des exons (extension ou raccourcissement), déplacement du site d'initiation de la transcription ou du site de polyadénylation, exons ou introns cryptiques, saut d'exons, répétition d'exons... L'obtention de séquences d'ADNc *full length* révèle ces phénomènes, dans la mesure où des clones complets s'avèrent différents après séquençage tout en partageant d'importantes zones homologues à 100 %. La comparaison avec la séquence «brouillon» du génome joue un rôle important en permettant de montrer que ces formes correspondent bien au même gène. Réciproquement, bien sûr, cette comparaison définit avec précision l'ensemble des exons. On parle maintenant de 50 % de gènes présentant un épissage alternatif, et le comité de nomenclature de Hugo *(Human Genome Organisation)* commence à attribuer des noms spécifiques à ces différentes formes.... Dans de nombreux cas, l'épissage alternatif fausse les résultats des *gene indexes*. Pour plus du tiers des ADNc étudiés dans le projet du DKFZ (Heidelberg, Allemagne), les séquences complètes montrent que deux, trois ou même quatre *clusters* Unigene correspondent en fait au même gène avec des exons alternatifs. Ce sont là autant de gènes «en trop» dans les dénombrements fondés sur ce type d'analyse... Plusieurs exemples montrant cinq ou six épissages alternatifs du même gène ont été présentés, sans que l'on sache si toutes les formes sont biologiquement significatives. Cette question de la pertinence biologique est évidemment capitale, et les informations à ce sujet restent fragmentaires*. J'aurais personnellement tendance à penser que la majorité de ces épissages ont un rôle fonctionnel, et que ces phénomènes permettent de produire 2 ou 3xN protéines à partir de N gènes (chacun donnera à N la valeur qui lui convient**). Ces possibilités montrent l'intérêt que peut présenter pour l'organisme une structure morcelée des gènes. Elles en constituent peut-être une justification, qui s'ajoute à l'habituel argument évolutif: la construction facilitée de protéines multifonctionnelles par assemblage de domaines protéiques ayant évolué indépendamment. N'oublions pas les modifications post-traductionnelles qui, elles, peuvent donner naissance à plusieurs entités à partir de chaque séquence d'acides aminés...

D'autres surprises ?

Un dernier élément nouveau, encore imprécis et qualitatif, ressort de ces travaux: il semble que les cas de gènes présents à plusieurs exemplaires, sur le même chromosome ou sur des chromosomes différents, soient en train de se multiplier. Ces indications résultent de la comparaison entre les ADNc complets et la sé-

** Il semble – mais les indications sont encore fragmentaires – que les schémas d'épissage alternatifs diffèrent souvent entre l'homme et la souris. On peut en déduire que cela indique leur caractère artefactuel... ou au contraire que cela démontre leur rôle possible dans la différenciation des espèces !*

*** L'estimation faite par l'équipe du Génoscope à partir de la séquence «brouillon» couvrant près de 90 % de notre génome confirme celle publiée en juin 2000 dans* Nature Genetics *[8] et prenant en compte 42 % de cet ensemble: moins de 30 000 gènes.*

quence brouillon du génome humain ; elles sont encore fragmentaires, et peuvent dans certains cas être liées à des erreurs dans ce brouillon (séquences attribuées au mauvais chromosome ou à la mauvaise région) ; mais il se pourrait que le phénomène soit assez général et ajoute à la complexité de notre génome – tout comme à la difficulté de définir le résultat du *sweepstake* de *Cold Spring Harbor* ! (*Voir* http://www.ensembl.org/Genesweep/).

On le voit, la saga des EST n'est pas terminée. Loin de correspondre à une étape désormais caduque dans l'analyse du génome humain, l'étude des ADNc s'avère aujourd'hui indispensable à la compréhension des nouvelles données sur notre génome ; combinée avec la séquence brouillon, elle est en train de changer l'image que nous nous faisons de notre ADN et, sans nul doute, de nous aider à comprendre comment 30 000 gènes «seulement» peuvent rendre compte de la complexité de notre organisme... ■

RÉFÉRENCES

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

2. Adams MD, Kerlavage AR, Fleischmann RD, *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 1995 ; 377 ; 3-174.

3. Roest Crollius H, Jaillon O. Le nombre de genes dans le genome humain : les paris sont ouverts. *Med Sci* 2000 ; 16 : 988-90.

4. Deloukas P, Schuler GD, Gyapay G, *et al.* A physical map of 30 000 human genes. *Science* 1998 ; 282 : 744-6.

5. Dunham I, Shimizu N, Roe BA, *et al.* The DNA sequence of human chromosome 22. *Nature* 1999 ; 402 : 489-95.

6. Hattori M, Fujiyama A, Taylor TD, *et al.* The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* 2000 ; 405 : 311-9.

7. Jordan B. Jusqu'où iront les puces ? *Med Sci* 2000 ; 16 : 950-3.

8. Roest Crollius H, Jaillon O, Bernot A, *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat Genet* 2000 ; 25 : 235-8.

Bertrand Jordan

Marseille-Génopole, Parc scientifique de Luminy, case 901, 13288 Marseille Cedex 9, France.

TIRÉS À PART

B. Jordan.