m/s

Chroniques génomiques

par Bertrand JORDAN

médecine/sciences 1995 ; 11 : 273-6

La valse des étiquettes

Sydney Brenner fut le premier, il y a près de dix ans, à proposer d'aborder l'analyse du génome en privilégiant l'étude des ADNc. Dans un court article publié à l'été 1994, il commentait l'imbroglio déclenché par les tentatives de prise de brevets sur ces séquences, et concluait sur cette phrase caustique (traduction libre) : « Voici une exemple classique de la facilité avec laquelle ceux qui administrent la recherche arrivent à saboter la science et transforment ce qui aurait pu être une belle réalisation en un gâchis banal » [1].

Ces paroles acerbes, écrites au début de l'année 1994, sont plus que jamais d'actualité. L'idée de base des projets ADNc, entamés à partir de 1991 aux États-Unis, au Japon et en France, était que l'étiquetage rapide des séquences transcrites allait permettre de constituer à brève échéance un catalogue quasiment complet des gènes humains. Compte tenu des progrès de la cartographie génétique et physique, on pouvait espérer positionner sans trop de délai ces repères sur les cartes correspondantes. Dans notre pays, Généthon était particulièrement bien placé pour réussir cette intégration puisque s'y menaient simultanément carte génétique, carte physique et séquençage d'étiquettes. Ainsi, bien avant l'an 2000, la grande majorité des gènes humains seraient identifiés par une ou plusieurs séquences partielles et une position sur la carte dans un intervalle inférieur à une mégabase. De ces connaissances – naturellement accessibles à tous les chercheurs grâce à leur archivage dans les bases de données internationales – découleraient de rapides progrès en génétique médicale, et des conséquences importantes pour la biologie en général.

Aujourd'hui, TIGR *(The Institute for Genome Research)*, l'institut de recherche dirigé par Craig Venter et financé par la compagnie Human Genome Science (HGS), elle-même soutenue par la puissante firme Smith Kline Beecham, aurait effectué plus de 150 000 séquences partielles* de clones pris au hasard dans des banques d'ADNc construites à partir de divers tissus humains. Malgré les inévitables redondances (clones présents à plusieurs exemplaires, séquences correspondant à différentes régions du même ARN messager...), il semble vraisemblable que ces EST *(expressed sequence tags)* représentent la majorité des gènes humains, dont le nombre est maintenant évalué à 60 000 ou 70 000 [2]. Le problème est que ces informations, tout comme celles obtenues par Incyte, entreprise alliée au groupe Pfizer, ne sont pas accessibles. Elles n'ont pas été déposées dans les banques de données publiques, qui ne contiennent à l'heure actuelle qu'environ 35 000 séquences partielles d'ADNc humain. Notons ici qu'une des raisons pour lesquelles ces EST restent secrets est le fait qu'ils ne soient pas brevetés. En droit américain, la publication est tout à fait compatible avec les brevets ; s'ils avaient été accordés, TIGR pourrait mettre les séquences dans le domaine public tout en préservant ses droits. Au contraire, la publication d'un EST sans brevet est susceptible d'empêcher la protection ultérieure d'un produit ou procédé utilisant le gène correspondant. L'affaire des brevets n'est d'ailleurs nullement terminée puisque *HGS*, *Incyte* et sans doute d'autres tentent toujours d'en obtenir, avec des chances de succès qui ne sont pas négligeables. Selon certaines estimations, le débat juridique pourrait encore durer six à sept ans...

La controverse qui fait maintenant rage aux États-Unis – et dont les péripéties occupent chaque semaine une ou deux pages dans *Nature* et *Science* – tourne autour des conditions sous lesquelles TIGR se propose d'accorder aux chercheurs « académiques » l'accès aux séquences d'EST. Il n'est en effet plus question de les déposer purement et simplement auprès d'organismes publics comme Genbank ou EMBL, et TIGR a diffusé en octobre 1994 un texte de *Database Agreement*, que devrait signer tout chercheur souhaitant interroger la base des EST. Remarquons d'abord que le

* *Le chiffre surprendra peut-être. La plus grande fantaisie règne à ce sujet, on parle tantôt de 50 000, tantôt de 300 000 EST. En fait, il faut distinguer entre les séquences effectuées, celles qui ont déjà été analysées, les séquences uniques (puisque certains clones sont retrouvés de multiples fois). Il faut aussi faire la part d'une certaine « intox »... Il semble bien pourtant que TIGR et HGS aient effectivement réalisé au moins 150 000 séquences – ce qui ne signifie pas, naturellement, qu'ils aient répertorié 150 000 gènes différents !*

contrat – vingt pages de jargon juridico-administratif d'une lecture ardue – désigne comme interlocuteur HGS, compagnie commerciale, et non TIGR, institut privé mais en principe sans but lucratif. Il stipule que le contractant soumettra à *HGS*, au moins un mois à l'avance, toute « publication » écrite ou orale envisagée, et détaille dans quels cas cette dernière peut inclure des séquences tirées de la base de données de TIGR. Il définit surtout la marche à suivre pour tout résultat éventuellement brevetable, le délai pouvant alors être porté à 60 jours, et précise que l'institution du chercheur doit s'engager à donner à HGS une option exclusive (un droit de premier refus) sur tout brevet découlant de toute recherche ayant utilisé ces séquences. Le contrat se termine par une page de « réserves » déchargeant TIGR, HGS et Smith Kline Beecham de toute responsabilité quant à l'exactitude, la qualité, l'utilité et même la disponibilité au sens juridique des informations éventuellement fournies. L'objet de ce long texte peut en somme être résumé de façon très simple : « Human Genome Science et Smith Kline Beecham autorisent des chercheurs à accéder aux séquences de leur projet ADNc. Ces derniers pourront ainsi établir la fonction biologique ainsi que l'intérêt thérapeutique (et donc commercial) de certaines de ces séquences ; Smith Kline sera le destinataire privilégié de ces informations et leur exploitant exclusif. » Le luxueux dossier récemment diffusé à de nombreux chercheurs français par la direction médicale de Smith Kline Beecham ne contredit pas vraiment cette interprétation.

Est-ce, en fait, si scandaleux ? Smith Kline annonce un budget de cent millions de dollars pour le projet, montant qui représente plus de la moitié du programme Génome américain, et près de dix fois la dotation du GREG en 1994 (Groupement d'Études et de Recherches sur les Génomes) en France. Il faut certes le comparer au coût de développement d'un nouveau médicament, estimé récemment à plus de deux cents millions de dollars – pour des produits dont la plupart ne seront jamais mis sur le marché [3], mais la somme n'est pas dérisoire. Il est après tout normal que l'industriel attende un retour sur son investissement... Le plus grand scandale, c'est peut-être que le monde académique et les autorités qui dirigent les programmes génome aient mis aussi longtemps à découvrir les charmes de cette approche et se soient laissé prendre de vitesse par le privé !

Les prétentions de Smith Kline Beecham apparaissent pourtant excessives. Les séquences en elles-mêmes sont quasiment muettes: c'est la confrontation avec d'autres informations, dont l'obtention est généralement bien plus longue et plus coûteuse, qui leur donne la parole. La première étape, c'est naturellement la comparaison avec le contenu des grandes bases de données, EMBL ou GenBank. Les EST en ressortent classés comme « déjà connus », « apparentés » ou « nouveaux ». L'effectif de ces différentes catégories indique quel chemin reste à faire pour cataloguer ainsi l'ensemble des gènes exprimés : lorsque 99% des clones pris au hasard dans n'importe quelle banque ADNc révéleront une séquence déjà connue, nous saurons que l'inventaire est presque terminé... Viennent ensuite des travaux bien plus complexes. Comme dit Sydney Brenner dans l'article déjà cité, « chaque séquence inconnue est un projet de recherches »...

Parfois une hypothèse astucieuse braque les projecteurs sur un EST jusque-là anonyme. Le déjà classique travail de Bert Vogelstein en est un bel exemple (*m/s n° 11, vol. 10, p. 1178* [4]). Rappelons-en l'essentiel. Diverses indications biologiques avaient amené les auteurs à supposer qu'un défaut dans la réparation des mésappariements de l'ADN pouvait jouer un rôle dans certains cancers. On connaissait, chez la levure ainsi que chez *Escherichia coli*, des gènes appelés *mutL* et nécessaires à cette fonction, mais, dans l'espèce humaine, leur existence n'était qu'une supposition. La démarche conventionnelle aurait alors consisté à mettre en route des expériences visant à cloner le gène *mutL* humain. On pouvait employer la séquence connue dans les micro-organismes pour tenter un tel isolement par hybridation croisée, ou par PCR grâce à des amorces fortement dégénérées – mais l'entreprise restait acrobatique en raison de la faible conservation de séquence attendue entre les gènes d'organismes aussi éloignés dans l'évolution. Au lieu de cela, les auteurs ont procédé à une recherche d'homologie, sorte d'hybridation informatique dans laquelle la « sonde » était la séquence *mutL* bactérienne ou de levure, la banque criblée étant constituée par le jeu des EST accumulés à TIGR par Craig Venter. L'énorme avantage de cette méthode *in silicio* est sa capacité à révéler des similitudes subtiles, invisibles par hybridation ou par PCR. De plus, elle est rapide, fiable, peut être indéfiniment répétée en modifiant la rigueur des critères (comme l'on ajuste la « stringence » du lavage d'un *Southern blot*) et échappe à nombre d'aléas expérimentaux tout comme à l'emploi d'isotopes radioactifs... De fait, la comparaison devait identifier dans la banque de Venter trois EST codant (potentiellement) pour des protéines qui présentent une homologie faible mais reconnaissable avec la protéine MutL de levure ou de *E. coli*. Ces EST fournissaient des séquences d'ADN humain (et des clones ADNc) permettant de mettre en jeu tout l'éventail des méthodes usuelles. Il devenait possible de déterminer (par PCR sur un jeu d'hybrides somatiques) le chromosome sur lequel sont situés ces gènes humains, de cribler (classiquement cette fois) une banque de phages pour obtenir un clone plus grand autorisant la localisation précise par hybridation *in situ*, et de montrer ainsi qu'un de ces gènes, *hMLH1*, se trouvait précisément à un *locus* impliqué, d'après l'analyse génétique, dans le cancer héréditaire du côlon. Une analyse de mutations dans les familles en cause devait indiquer que cette séquence y est effectivement altérée. Son inactivation, interférant avec les processus de réparation de l'ADN, est donc presque certainement la responsable de l'affection...

Ainsi, le recoupement avec des renseignements obtenus dans un autre contexte peut subitement décupler l'intérêt d'une étiquette. Nos connaissances sur le métabolisme des

procaryotes et de quelques eucaryotes inférieurs sont précises et étendues ; à brève échéance, nous disposerons de la séquence complète du génome de la levure et donc de celle des 7 000 gènes que met en œuvre cet organisme. On peut donc imaginer de nombreuses recherches sur la base de ce modèle. On voit aussi quelle portée peut avoir cette démarche pour des firmes pharmaceutiques toujours à la recherche d'enzymes, de peptides ou d'hormones susceptibles de corriger un défaut du métabolisme – et d'ouvrir des marchés rémunérateurs ! Le groupe de Vogelstein avait négocié – à des conditions qui n'ont pas été rendues publiques – l'accès aux EST de TIGR, mais il est clair que ces connexions ne se réaliseront pleinement que si l'information est largement accessible afin que tout chercheur puisse sans hésitation tester son hypothèse, aussi farfelue soit-elle. De là l'importance cruciale de la libre disponibilité des séquences dans les bases de données publiques. Notons aussi – pour reprendre l'exemple du gène *hMLH1* – que la banque des EST de TIGR a sans aucun doute fourni une information cruciale et un outil difficilement remplaçable, mais que le travail accumulé en amont et en aval par les six autres laboratoires impliqués est considérable. L'idée de départ sur laquelle fut fondée la recherche d'homologie supposait une analyse pointue des résultats de nombreuses équipes ; la validation de l'un des EST trouvés a demandé son assignation puis sa localisation chromosomiques, l'obtention de clones lambda, YAC et P1, une étude de l'expression du gène dans différents tissus, et la séquence complète des régions codantes pour de nombreux individus appartenant à dix familles précédemment étudiées. Il ne semble pas raisonnable que le « fournisseur » de la séquence de l'EST se réserve le bénéfice exclusif d'un (très improbable) médicament anticancéreux qui découlerait de ce travail imposant réalisé pour l'essentiel dans le monde « académique » et à l'aide de fonds publics... Un prochain colloque organisé par l'Académie des Sciences (La propriété industrielle dans le domaine du Vivant, 26 et 27 janvier) doit contribuer à faire le point sur cette question, d'autant qu'y participeront des personnalités comme Reid Adler, qui avait été à l'origine de la première demande de brevets, ou Bill Haseltine, président de *HGS*.

L'histoire de *mutL* montre tout ce qu'apporte l'information de position. C'est en effet la connaissance (au moins approximative) de la localisation de l'EST qui l'élève au statut de gène candidat : elle permet de corréler la séquence (qui indique, directement ou par homologie, de quel type de protéine il peut s'agir) avec la carte génétique et cytogénétique humaine et son très riche inventaire de phénotypes et de maladies. N'oublions pas non plus la carte génétique détaillée de la souris et son vaste catalogue de mutants. Dans l'exemple cité ci-dessus, c'est bien le placement de l'EST qui a fait l'objet des premières investigations, et c'est sa position dans l'intervalle précédemment défini par l'analyse génétique des malades qui en a fait le gène à étudier en priorité. Conformément à la logique des programmes Génome – faire les choses en grand, de manière systématique et organisée, plutôt qu'au coup par coup dans le cadre de projets très ciblés – la localisation en masse des EST est d'actualité. Elle fait même l'objet de grandes manœuvres politico-financières [5]. Techniquement, le problème n'est pas simple. Il est relativement facile d'assigner les EST à un chromosome donné, grâce à une série de réactions PCR sur les ADN d'un jeu d'hybrides somatiques : des centaines et même des milliers d'assignations ont pu ainsi être effectuées par différents laboratoires. Cela coûte déjà 1 000 ou 2 000 F par EST (plus de dix fois le prix de la séquence qui a défini ce dernier), et surtout cela n'apporte pas grand-chose. Savoir qu'un gène se trouve sur le chromosome 3 ne permet pas d'établir une corrélation avec une maladie génétique. Il est nécessaire d'aller plus loin, jusqu'à la localisation, avec si possible une précision de l'ordre de la mégabase. A ce niveau de précision, l'hybridation *in situ* reste la méthode de choix, mais elle s'applique malaisément à des sondes courtes comme les ADNc, surtout dans une opération systématique où les clones à positionner se comptent par milliers. Dans les conditions techniques actuelles, l'obtention d'un clone de grande taille (YAC ou P1), puis son positionnement par FISH, représentent la voie la plus sûre – mais cela coûte de 5 000 à 10 000 F par clone*, ce qui est prohibitif vu le nombre d'entités à traiter...

La solution réside sans doute dans l'exploitation de la carte physique et des segments clonés qui la sous-tendent. On peut l'envisager directement, avec un schéma du type des « filtres polytènes » qui a si bien réussi à la communauté du nématode. Cela revient, une fois la carte physique du génome établie, à déposer sur des filtres à haute densité les YAC dans l'ordre dans lequel ils sont positionnés sur les chromosomes. Une simple hybridation du segment d'ADN à localiser sur un tel filtre révèle alors deux ou trois « spots » positifs adjacents, indiquant à la fois le chromosome et la position sur ce dernier à quelques centaines de kilobases près. Il est aussi concevable de s'inspirer des techniques de *physical trapping* pour reconnaître et donc placer d'un coup toutes les séquences codantes que contient un YAC donné. Une autre alternative est l'emploi de « panels » d'hybrides d'irradiation. Quelle que soit la méthode qui se révélera finalement la plus efficace, il est extrêmement probable que la localisation d'EST va prochainement « décoller » et être pratiquée à une grande échelle.

Imaginons donc qu'existe un catalogue des 70 000 gènes humains. Leur séquence, le plus souvent partielle, permet une première approche de la fonction associée ainsi

** Les coûts donnés sont forcément approximatifs, ils tentent néanmoins de tenir compte de tous les facteurs (y compris personnel et infrastructure). Leur estimation s'appuie sur l'expérience de Généthon et de différents Genome Centers ainsi que sur les prix (naturellement supérieurs) pratiqués par des firmes privées offrant ces services.*

que la recherche d'homologies avec les gènes d'organismes comme *Saccharomyces cerevisiae*, *Escherichia coli* ou *Drosophila melanogaster*. Imaginons de plus que la position de chacun de ces 70000 EST ait été établie à une ou deux mégabases près. Il devient alors possible de croiser dans tous les sens une multitude de données : connaissance des fonctions chez les procaryotes ou les eucaryotes primitifs, motifs de séquence, catalogue des maladies génétiques humaines et des mutants de souris... pour en tirer des éclairages nouveaux et des résultats très significatifs. Le gène *hMLH1*, pour en revenir à lui, aurait pu être identifié sans aucune manipulation, par simple interrogation des bases de données. Les études de mutation auraient constitué la seule expérimentation nécessaire. Un troisième type d'informations, portant cette fois sur l'expression, apporterait un « plus » supplémentaire : une idée du niveau auquel chaque EST est transcrit dans une série de tissus orienterait les interprétations de manière décisive. Mais, encore une fois, le libre accès à l'ensemble de ces données apparaît comme une condition *sine qua non* pour que toutes leurs potentialités soient réalisées.

C'est bien ainsi que l'entendent de nombreux scientifiques, et même une firme, la société Merck. Celle-ci se propose tout simplement de financer l'obtention d'un nombre important d'EST destinés à être placés dans le domaine public. Philanthropie ? Certes non. Pour des raisons évidentes, Merck ne souhaite pas laisser à Smith Kline Beecham la disponibilité exclusive des EST ni, surtout, l'accès privilégié aux résultats biologiques qui en découlent. En finançant pour ce travail des laboratoires honorablement connus, et en rendant largement accessibles les résultats, elle compte selon les termes de ses porte-parole « optimiser la possibilité que cette information serve à améliorer la santé humaine » [6] – et qu'en dérivent des médicaments que Merck sera bien placé (quoique sans droits exclusifs) pour développer. Rappelons que les quelques millions de dollars mis en jeu ne sont pas exorbitants pour de grands groupes pharmaceutiques. Ceux-ci ont récemment accepté, aux États-Unis, de s'engager à payer aux femmes qui avaient reçu des implants mammaires défectueux des indemnités s'élevant au total à plusieurs milliards de dollars... Une firme comme Merck peut donc, moyennant un investissement modeste, se forger une excellente image de marque dans le milieu scientifique (gage de bonnes collaborations futures) et enlever à un rival l'exclusivité d'un secteur qui pourrait se révéler lucratif dans l'avenir. Elle devrait d'ailleurs être rejointe par d'autres partenaires, organismes, fondations ou autres entreprises, soucieux de profiter de cette occasion de remettre les EST dans le domaine public. Cette initiative arrive-t-elle après la bataille ? Ce n'est pas certain, les laboratoires qui doivent effectuer le séquencage (comme celui de Bob Waterston) sont d'une efficacité redoutable, et ils peuvent tirer parti de l'expérience des autres pour cataloguer les EST en un temps record.

Qu'en est-il de la participation française à ce projet ? L'hypothèse d'une harmonieuse imbrication entre les trois projets de Généthon pour aboutir à une carte intégrée au triple niveau génétique, physique, et transcriptionnel ne s'est qu'en partie réalisée. La carte génétique a effectivement joué un rôle primordial dans l'établissement de la carte physique [7]. Mais l'affinement de cette dernière est maintenant poursuivi au CEPH, et son emploi pour positionner les étiquettes déterminées à Évry ne semble pas être un objectif prioritaire. En tout état de cause, les choses seront moins simples que pour le nématode, en raison du chimérisme qui touche environ un YAC sur deux. Les incertitudes qui en résultent imposent, soit une assignation chromosomique préalable pour lever les ambiguïtés, soit des schémas astucieux comme celui que met au point (avec les YAC du CEPH) Donald Moir, de Collaborative Research. Le capital important que constituent les étiquettes caractérisées par le projet Genexpress – déjà assignées pour un grand nombre d'entre elles [8] – devrait pourtant être valorisé. Comme j'ai essayé de le montrer, c'est la carte triplement intégrée qui va rendre la connaissance du génome réellement opérationnelle : il serait fort dommage que la France ne joue pas à ce stade un rôle aussi important que lors des étapes précédentes ■

Bertrand R. Jordan

Directeur de recherche au Cnrs, responsable du groupe « Structure du Génome et fonctions immunitaires », CIML, Inserm/Cnrs, case 906, 13288 Marseille Cedex 9, France.

RÉFÉRENCES

1. Brenner S. Loose ends. *Curr Biol* 1994 ; 4 : 384.

2. Fields C, Adams MD, White O, Venter JC. How many genes in the human genome? *Nature Genet* 1994 ; 7 : 345-6.

3. Buckholz H. The FDA follies: an alarming look at our food and drugs in the 1980s. *Basic Books*, 1994.

4. Papadopoulos N, Nicolaides NC, Wei YF, *et al.* Mutation of a *mutL* homolog in hereditary colon cancer. *Science* 1994 ; 263 : 1625-9.

5. Dickson D. « Gene map » plan highlights dispute over public vs private interests. *Nature* 1994 ; 371 : 365-6.

6. Williamson AR, Elliston KO. *Nature* (correspondence) 1994 ; 372 : 10.

7. Jordan B. Carte physique du génome humain : l'état des lieux. *médecine/sciences* 1994 ; 10 : 898-902.

8. Auffray C, Behar C, Bois F, *et al.* Intégration au niveau moléculaire de l'analyse du génome humain et de son expression : signatures de séquence et d'hybridation de clones d'ADNc du muscle squelettique et du cerveau et assignation chromosomique des gènes correspondants. *CR Acad Sci Paris Ser III* 1995 (sous presse).

TIRÉS À PART

B.R. Jordan.