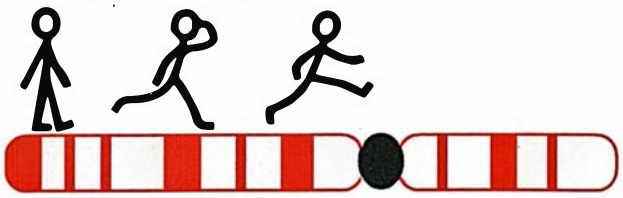
m/s

Chroniques génomiques



par Bertrand JORDAN

Séquençage génomique : le deuxième souffle

Nous avons évoqué, l'année dernière, les espoirs déçus des chercheurs qui avaient cru, dès 1987/1988 , pouvoir se lancer dans le séquençage de grandes régions génomiques [1] : ils s'étaient heurtés à de multiples obstacles les amenant à réduire considérablement leurs prétentions et même, dans certains cas, à abandonner leur tentative. Rappelons qu'à l'époque, on attendait la séquence complète d'*Escherichia coli* (quatre ou cinq mégabases) et de *Salmonella typhimurium*, sans parler de l'ensemble du complexe majeur d'histocompatibilité humain (trois mégabases) : or, aucun de ces programmes n'a abouti à ce jour. Mais le vent semble avoir tourné, et l'on a vu récemment les premiers résultats de plusieurs grands projets de séquençage. Ils prouvent que de telles entreprises sont possibles et montrent, de plus, que leurs retombées au niveau de la connaissance sont substantielles.

Une technologie peu évolutive

On pensait généralement, dans les années 1980, que les techniques mises en œuvre pour déterminer la séquence de l'ADN allaient subir une mutation. Le séquençage multiplex de George Church [2] devait décupler l'efficacité de la voie traditionnelle employant la radioactivité, et des démarches plus novatrices étaient étudiées. On comptait sur la microscopie à effet tunnel, sur l'identification des nucléotides par spectrométrie de masse ou la détection de bases isolées par fluorométrie ultrasensible pour gagner plusieurs ordres de grandeur en vitesse, tout en réduisant considérablement les frais [3]. Plus récemment, le séquençage par hybridation est apparu comme une alternative intéressante [4] ; mais pour le moment, aucune de ces approches n'a fait ses preuves. Le déchiffrage de l'ADN fait toujours appel, pour l'essentiel, au procédé mis au point dès la fin des années 1970 par Alan Coulson et Fred Sanger — amélioré à divers égards, mais dont le principe reste identique — : produire, par réaction enzymatique à partir de l'ADN à séquencer, une série de fragments marqués ayant tous la même origine et des extrémités différentes mais spécifiques d'un nucléotide donné, puis séparer ces fragments sur un gel très résolutif et déduire la séquence de la position des bandes observées.

Des perfectionnements de détail

Des améliorations ont certes été apportées. La facilité d'emploi de la technique a été fortement accrue par la découverte d'enzymes plus spécifiques et moins sensibles à la structure secondaire de l'ADN, par de multiples *kits* prêts à l'usage, et par l'utilisation de la PCR en amont du séquençage proprement dit. Et les « machines à séquencer », qui prennent en charge la séparation des fragments et la détection des bandes, ont maintenant droit de cité. La firme *Applied Biosystems*, première arrivée sur le marché, a vendu plus de 800 appareils de par le monde, le Suédois *LKB/Pharmacia*, son rival le plus sérieux, a mis en service une centaine d'appareils. La synthèse d'oligonucléotides a fait d'énormes progrès, et ces réactifs, qui permettent le séquençage à partir d'un point choisi du génome, sont maintenant à la portée de tout laboratoire. L'informatique associée à la détermination de séquence a, elle aussi, bien progressé depuis les logiciels mis au point par Robert Staden à la fin des années 1970. L'un des principaux avantages des machines est qu'elles assurent l'entrée directe des informations de séquence dans une mémoire d'ordinateur : dans la méthode manuelle, la lecture des autoradiographies, puis la saisie au clavier des résultats, sont une des principales sources d'erreurs.

Un nouveau réalisme

Mais le redémarrage des grands programmes de séquençage génomique est surtout dû, à mon avis, à un changement d'attitude. On attendait une révolution technique : c'est une évolution culturelle qui a eu lieu. Les responsables ont compris la nécessité d'une organisation rigoureuse, particulièrement bien décrite dans un récent article du laboratoire de Leroy Hood [5]. Ils ont aussi cessé de s'illusionner sur les coûts, et ont accepté de prévoir un prix de revient nettement supérieur au chiffre mythique d'un dollar US la base, du moins dans la phase du démarrage. Un projet en bonne voie, celui du séquençage de l'ADN du nématode par les équipes de John Sulston et de Bob Waterston [6], et le déchiffrage complet du chromosome III de la levure par un consortium européen [7] illustreront ces propos ; nous n'oublierons pas d'évoquer le récent séquençage

d'une centaine de kilobases d'ADN génomique humain par l'équipe de Craig Venter [8] qui, on le voit, ne se cantonne pas à l'ADNc...

La séquence du nématode : ce n'est qu'un début...

J'ai déjà exprimé ici [9] tout le bien que je pensais de l'équipe britannique qui, après avoir cartographié les 100 mégabases du génome du nématode, entame son séquençage. Les résultats initiaux, publiés en mars de cette année, portent sur 121 kilobases, soit trois cosmides provenant du milieu du chromosome III (le nématode en a six). La stratégie employée commence par un séquençage au hasard *(shotgun)* avec deux machines *Applied Biosystems*, très appropriées à cette étape grâce à leur grand débit. Dans un deuxième temps, la séquence est terminée de façon dirigée, c'est-à-dire en utilisant des amorces de séquençage synthétisées à partir des séquences déjà connues, afin de combler les trous subsistants et de vérifier les régions litigieuses. L'avantage de cette approche est sa souplesse : on peut modifier la part relative des deux étapes en fonction des enseignements de l'expérience. Le début est modeste, un peu plus d'une centaine de kilobases, mais ce n'était qu'une mise en train : les prévisions sont de 800 kilobases pour 1992 et 2000 en 1993. Le prix de revient annoncé, pour le laboratoire tel qu'il tourne aujourd'hui, est de l'ordre de un dollar US la base, bien qu'il ait été nettement plus élevé pour la séquence présentée en raison des mises au point effectuées en cours de route.

Que « dit » cette séquence, au delà de son aspect de banc d'essai technologique et organisationnel ? Une analyse informatique indique qu'environ 33 de ces 121 kilobases sont codantes. Il n'est pas simple de déterminer le nombre de gènes auxquels cela correspond, compte tenu de la présence d'introns chez le nématode : si les programmes d'analyse savent maintenant à peu près trouver les exons dans une séquence, ils ont beaucoup de mal à savoir où commencent et finissent les gènes, à distinguer intron et séquence intergénique. On peut néanmoins estimer que ces trois cosmides contiennent une trentaine de gènes. Nombre élevé, très supérieur aux prévisions : si l'on extrapole à l'ensemble du génome du nématode — en tenant compte de ce que l'on sait déjà sur les différences de densité génique selon les régions — l'on arrive à plus de 15 000 gènes au total. Cela fait beaucoup pour ce minuscule invertébré qui comporte en tout 959 cellules...

Le chromosome III de la levure ou un consortium réussi

C'est à la levure que devait revenir l'honneur d'être le premier organisme à voir un de ses chromosomes entièrement séquencé. L'agencement du travail était inhabituel, puisque les trois cent et quelques kilobases à déchiffrer furent réparties entre 35 équipes, chacune en prenant en charge une dizaine avec un financement de deux Écus (soit actuellement trois dollars US) par base séquencée. Commencé début 1989, ce projet était entouré d'un grand scepticisme : beaucoup pensaient que les niveaux de compétence des divers laboratoires étaient trop hétérogènes, que leur coordination serait impossible, et que la tentative n'aboutirait pas dans les délais. Parcourant les États-Unis au printemps 1991, alors que cette séquence était déjà presque terminée, j'entendis beaucoup parler des plans américains de séquençage de la levure, encore à l'état d'esquisse ; mais jamais personne, et surtout pas David Botstein, chaud partisan de cette entreprise, ne mentionnait le programme européen. Toujours est-il que la plupart des équipes remplirent leur contrat, que le centre coordinateur situé à Martinaried, en Allemagne, assembla les séquences, et que le travail était pour l'essentiel terminé à la mi-1991. Un total de 385 kilobases de séquence effectuées en double (par des laboratoires différents) permit d'évaluer le taux d'erreurs à environ 0,4 % : une erreur toutes les 2 000 bases, performance tout à fait honorable.

Cette séquence, qui fit l'objet d'un article publié en mai 1992 par la revue *Nature* [7], révéla un très grand nombre de nouveaux gènes : près de 150, sur un chromosome très étudié où l'on s'attendait à en trouver une cinquantaine. Dans le cas de la levure, où les introns sont rarissimes, la reconnaissance des gènes d'après la séquence est facile, et le chiffre d'un gène toutes les deux kilobases ainsi obtenu est fiable — d'autant plus qu'il a été aisé de montrer que dans leur quasi-totalité, ces derniers sont effectivement transcrits. Quelques-uns de ces « nouveaux » gènes présentent une similitude avec des séquences déjà répertoriées dans les bases de données, provenant de l'homme, de la drosophile, du xénope ou même d'*Arabidopsis*. Le « jeu minimum » de gènes requis pour assurer les fonctions vitales de la levure augmente donc en flèche, et les connaisseurs parlent maintenant de 6 000 à 7 000 gènes pour ce très modeste eucaryote inférieur... D'autres éléments intéressants commencent à apparaître, comme la confirmation d'une forte corrélation entre densité en gènes et fréquence de recombinaison : les zones du chromosome III où l'on trouve le plus de gènes par kilobase, sont aussi celles où la fréquence de recombinaison (toujours par kilobase) est la plus élevée. Cela va dans le sens de théories comme celle de Roeder [10] qui associent les deux phénomènes. Finalement, la somme de 2,65 millions d'Écus, montant total de ce programme, ne paraît pas exhorbitante au vu des résultats. Il semble toutefois évident que l'on ne continuera pas à faire du mégaséquençage en assemblant des consortiums de dizaines de laboratoires travaillant à la main et qu'il va fatalement falloir passer par une certaine concentration pour réduire les frais. C'est d'ailleurs ce qui se met en place pour la phase suivante, portant sur les chromosomes II et XI, représentant un total d'une bonne mégabase et demie.

Craig Venter ne séquence pas que l'ADNc

L'équipe de Craig Venter à Bethesda (MA, USA) est surtout connue par l'efficacité dont elle a fait preuve dans le séquençage massif et partiel de clones d'ADNc pris au hasard dans une banque... et par sa décision très contestable de chercher à breveter ces séquences. Cependant, les participants au colloque sur la cartographie

et la séquence du génome à Cold Spring Harbor en 1988 se rappellent peut-être que ce même Craig Venter y avait présenté un grand projet de séquençage génomique : le déchiffrage de la bande q28 du chromosome X — soit la bagatelle d'une dizaine de mégabases. Ce projet n'obtint pas les financements demandés, et ne fut donc pas réellement entamé ; mais il n'est pas surprenant de voir ce groupe continuer à s'intéresser à l'ADN génomique. Son récent article dans le premier numéro de la nouvelle revue *Nature Genetics* [8] rapporte le séquençage d'une zone du chromosome 19 humain autour du *locus* ERCC1, qui contient un gène impliqué dans la réparation de l'ADN. Ce chromosome avait été choisi par le laboratoire d'Anthony Carrano (Lawrence Livermore, un des *Genome Centers* du *Department of Energy*) pour en établir la carte physique par recouvrement de cosmides. Trois de ces derniers ont été séquencés par le groupe de Craig Venter, selon des modalités maintenant classiques : tactique *shotgun* et appareils *Applied Biosystems*.

L'assemblage des multiples petites séquences a été compliqué par la présence de très nombreux éléments répétés et a nécessité des expériences complémentaires de cartographie fine par enzyme de restriction ; et le repérage des gènes n'a pas été sans aléas. C'est que l'on analyse ici de l'ADN humain, que les gènes comportent de nombreux introns et sont assez dispersés : selon les propres estimations des auteurs, les programmes informatiques employés ont un taux de « faux négatifs » de l'ordre de 60 %, autrement dit, ils « ratent » les exons plus d'une fois sur deux... Cinq gènes ont été trouvés dans cette zone de 116 kilobases : il s'agit du gène *ERCC1* déjà caractérisé, d'un proto-oncogène *fosB*, d'un autre dont le produit ressemble à une phosphatase et de deux gènes nouveaux, ne présentant aucune homologie avec les séquences connues. La moisson est moins riche que pour levure et nématode ; elle indiquerait néanmoins de 75 à 150 000 gènes au total chez l'homme. Ce travail montre que l'on peut effectivement séquencer des centaines et, sans doute, des milliers de kilobases d'ADN humain, mais que la tâche est en même temps plus ardue et moins riche d'informations que pour des organismes simples dont le génome est plus compact.

Quelques conclusions

Tout épilogue ne peut être que provisoire : si demain une technique (révolutionnaire ou « évolutionnaire ») rend le séquençage dix fois plus rapide, ou en ramène le coût à 0,50 F la base, tout changera. C'est parfaitement envisageable puisqu'après tout, bien d'autres techniques ont fait des progrès comparables dans le passé récent. En attendant, on peut déjà tirer quelques enseignements valables dans le cadre technique d'aujourd'hui.

• *Le séquençage d'une mégabase est aujourd'hui faisable.* On croyait, à tort, que c'était le cas il y a quatre ou cinq ans ; mais c'est maintenant une réalité démontrée. Il n'est pas chimérique de prévoir un budget d'un mégadollar par mégabase, à condition que la région à séquencer soit déjà entièrement clonée dans des cosmides. L'interprétation informatique des séquences ne semble plus poser de problèmes insurmontables, et la détection des exons par les programmes récents est relativement satisfaisante, bien qu'il reste encore délicat de définir les limites des gènes dans le cas d'ADN humain.

• *Le nombre (estimé) de gènes ne cesse de croître.* Chaque région entièrement séquencée a révélé beaucoup plus de gènes qu'attendu. La majorité d'entre eux semble avoir échappé à l'analyse génétique, parce que leur inactivation n'a pas de conséquences apparentes pour l'organisme — du moins dans les conditions du laboratoire. Cela veut-il dire pour autant que ces gènes sont inutiles ? Qu'en est-il pour l'homme, et quelle validité accorder au chiffre généralement admis de 50 ou 100 000 gènes lorsqu'on découvre que le nématode, avec ses 959 cellules, en a sans doute 15 000 ?

• D'un point de vue plus « utilitaire », l'on peut aujourd'hui s'interroger sur *l'emploi du séquençage dans la recherche du gène d'une maladie*. Après localisation génétique, on se trouve généralement en face d'une zone de quelques centimorgans, quelques milliers de kilobases, qui doit contenir le « gène morbide ». Les différentes méthodes actuelles visant à faire le catalogue des « gènes candidats » présents dans cette région sont relativement hasardeuses et, en tout état de cause, non exhaustives. Pourquoi ne pas séquencer toute la région afin de faire une fois pour toutes le catalogue des exons qu'elle contient ? Cette approche a été mise en œuvre par le groupe de Christine Petit avec l'aide du Généthon pour découvrir le gène du syndrome de Kallmann [11] ; il est vrai que dans ce cas la zone avait été restreinte à moins de 100 kilobases, mais il n'est pas forcément chimérique de penser à multiplier ce chiffre par dix. Le procédé a le mérite d'être systématique et est, par exemple, sérieusement envisagé par des responsables de l'AFM.

• *Séquençage génomique et séquençage d'ADNc sont-ils opposés ou complémentaires ?* Le séquençage de l'ADNc, selon le schéma popularisé par Craig Venter, donne à peu de frais une information très partielle, susceptible d'établir assez rapidement un catalogue de tous les gènes transcrits dans le tissu à partir duquel a été construite la banque ; mais cette tactique très efficace ne fournit ni la séquence complète (indispensable aux prédictions de structure et, *a fortiori*, de fonction), ni la localisation du gène, et risque de fournir une vision très impressionniste du génome. Le séquençage génomique a les avantages inverses — moyennant un coût très élevé. Trop élevé, dans les conditions d'aujourd'hui, pour battre le rapport qualité/prix du séquençage de l'ADNc si l'on s'adresse à notre grand génome ; en revanche, pour la levure et le nématode, l'approche génomique est sans doute la plus rentable. Il suffirait que le séquençage devienne plus abordable — à la suite par exemple d'innovations permettant la lecture de 2 000 bases par échantillon au lieu de 400 actuelles — pour qu'il en soit de même chez l'homme.

Post scriptum : nous noterons encore, pour être plus complets, deux articles récents rapportant de grandes séquences : l'un, très attendu, provient de l'équipe de Fred Blattner [12] et porte sur un peu moins de 100 kilo-

bases dans le génome d'*Escherichia coli* ; le deuxième [13] concerne une centaine de kilobases au voisinage du *locus* de la chorée de Huntington et émane du décidément très prolifique Craig Venter... ■

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

RÉFÉRENCES

1. Jordan BR. Les heurs et malheurs du séquençage à grande échelle. *médecine/sciences* 1991 ; 7 : 612-3.
2. Church GM, Kieffer-Higgins S. Multiplex DNA sequencing. *Science* 1988 ; 24 : 185-8.
3. Jordan BR. Le tunnel séquencera-t-il le génome ? *médecine/sciences* 1990 : 6 : 1007-9.
4. Cantor CR, Mirzabekov A, Southern E. Report on the sequencing by hybridization workshop (special feature, meeting report). *Genomics* 1992 ; 13 : 1378-83.
5. Wilson RK, Koop BF, Chen C, Halloran N, Sciammis R, Hood L. Nucleotide sequence analysis of 95 kb near the 3' end of the murine T cell receptor alpha/delta chain locus : strategy and methodology. *Geromics* 1992 ; 13 : 1198-208.
6. Sulston J, Du Z, Thomas K, *et al.* The C. elegans genome sequencing project : a beginning. *Nature* 1992 ; 336 : 37-41.
7. Oliver SG, Van der Aart QJM, Agostoni-Carbone ML, *et al.* The complete DNA sequence of yeast chromosome III. *Nature* 1992 ; 357 : 38-46.
8. Martin Gallardo A, McCombie WR, Gocayne JD, *et al.* Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nature Genetics* 1992 ; 1 : 34-9.
9. Jordan BR. Grande-Bretagne : un programme Génome à dimension humaine. *médecine/sciences* 1992 ; 8 : 163-6.
10. Keil RL, Roeder GS. *Cis*-acting, recombination-stimulating activity in a fragment of the ribosomal DNA of *S. cerevisiae*. *Cell* 1984 ; 39 : 377-86.
11. Legouis R, Hardelin JP, Levilliers J, *et al.* The candidate gene for the X-linked Kallmann syndrome encodes a protein related to adhesion molecules. *Cell* 1991 ; 67 : 423-35.
12. Daniels DL, Plunkett G, Burland V, *et al.* Analysis of the Escherichia coli genome : DNA sequence of the region from 84.5 to 86.5 minutes. *Science* 1992 ; 257 : 771-8.
13. McCombie WR, Martin-Gallardo A, Gocayne JD, *et al.* Expressed genes. Alu repeats and polymorphisms in cosmids sequenced from chromosome 4p16.3. *Nature Genetics* 1992 ; 1 : 348-53.