m/s

médecine/sciences 2001 ; 17 : 287-9

PLÉTHORE DE DONNÉES ET PÉNURIE DE CONCEPTS ?

Jean Weissenbach

La brève histoire de la génétique est émaillée de révolutions. Certaines reposent sur des avancées conceptuelles : gène, théorie chromosomique de l'hérédité, identification de l'ADN comme matériel héréditaire, expression génique et son contrôle, etc. D'autres sont la conséquence de percées technologiques : hybrides somatiques, applications des techniques de l'ADN recombinant (clonage, séquençage, PCR, transgénèse, *knock-out*, etc.). La prochaine révolution est en marche ; elle doit découler de l'utilisation de l'information génétique à l'échelle génomique à des fins de recherche et d'applications.

A partir du moment où une partie essentielle de l'explication des caractéristiques phénotypiques était inscrite dans l'ADN, que celui-ci devenait lisible, à défaut d'être compréhensible, pouvions-nous échapper au séquençage ? Il est sans doute inutile de revenir sur l'argumentaire qui a servi à vendre le projet au cours des années 1980. Trop fut promis, et les détracteurs, qui furent nombreux à certains moments, avaient beau jeu. Aujourd'hui, personne ne conteste que ce projet avait un caractère incontournable et que c'est un succès.

Les médias ont amplement relayé la stupide guerre de communiqués entre *Celera* et le consortium public (HGP) qui a émaillé la phase de séquençage aléatoire intensive en 1999-2000. Selon certains, c'était un excellent moyen de susciter l'intérêt du public pour l'entreprise. En tout cas, on constate que, dans cette circonstance, la communication et le sensationnalisme ont largement pris le pas sur l'information, et je doute que ce que le public en a retenu puisse quelque peu refléter la réalité. On pourra aussi noter, qu'en la matière, *Celera* a montré des talents remarquables. Arriver à faire croire que le travail de séquençage était terminé, alors que l'ensemble des données dont ils disposaient ne représentait que le tiers de ce qui est nécessaire pour procéder à un assemblage satisfaisant, est un tour de force.

Mais *Celera* vient de franchir un pas supplémentaire dans le domaine de la mystification et, cette fois-ci, avec le support d'une revue scientifique de très haut niveau. Déjà, le titre sans nuances *« The Sequence of the Human Genome »* de l'article « historique » de *Science* devrait balayer toute trace de scepticisme chez d'éventuels agnostiques. Pourtant, il subsiste 170 000 trous dans l'assemblage « compartimenté » de *Celera* et la structure d'au moins un tiers des gènes est incomplète. Une lecture rapide laisse le lecteur sur l'impression que le travail de *Celera* est un succès. En fait, l'entreprise de *Celera* a été sauvée par le projet public HGP *(human genome project)*. Il était notoire que *Celera* se servirait massivement des données du HGP et, qu'en conséquence, nous ne saurions pas le fin mot sur la validité de l'approche de séquençage aléatoire global *(m/s 2000, n° 1, p. 15)*. Un examen des résultats chiffrés présentés dans l'article de *Science* montre, sans ambiguïté, que cette stratégie ne permet pas d'assembler un génome aussi complexe que le génome humain. Cette réalité est bien sûr occultée

ADRESSE

J. Weissenbach : Genoscope et Cnrs UMR-8030, 2, rue Gaston-Crémieux, 91057 Évry Cedex, France.

avec adresse. Les enjeux sont manifestement trop importants pour se laisser perturber par de vagues scrupules d'honnêteté intellectuelle. On a connu des cas de résultats fallacieux, il y en aura d'autres. Ici, rien de tel, le jeu consiste à faire croire que la stratégie de séquençage aléatoire global a permis d'assembler le génome humain (*voir* l'article de H. Roest Crollius, *p. 309 de ce numéro*). Ceci est une escroquerie à plusieurs niveaux : (1) *Celera* indique avoir introduit 2,96 équivalents génome du HGP dans son assemblage. Or, ces 2,96 ne sont pas des morceaux collectés au hasard : il s'agit de la succession de fragments parfaitement chevauchants, obtenue par découpage en « quinconce » de la séquence assemblée par le HGP. Ceci permettait donc à *Celera* de conserver l'intégralité de l'information d'assemblage du HGP, issue elle-même d'une couverture moyenne comprise entre 7 et 8 fois ; (2) *Celera* se garde bien de donner le résultat de l'assemblage de ses données prises isolément. Mais on peut en avoir une idée. Les assemblages de *Celera* couvrent au mieux la même fraction de génome que le HGP. En d'autres termes, les 10 % de séquence euchromatique qui sont absents du projet public, mais bien présents dans les 27 millions de lectures de *Celera*, n'ont pu être assemblés par leurs programmes.

A ce jour, le projet n'est pas terminé, loin s'en faut. Et pourtant beaucoup croient que le génome est séquencé alors qu'il subsiste quelque 140 000 « trous » dans la séquence du HGP. Ce caractère incomplet n'aura empêché ni le consortium public, ni *Celera* de procéder à des analyses exhaustives. Il est vrai que les 5 à 10 % manquants ne changeront pas substantiellement l'image qui émerge (*voir* l'article de R. Heilig et T. Brüls, *p. 299 de ce numéro*), mais la séquence complète, de haute qualité et soigneusement annotée, reste une nécessité incontournable pour de nombreux utilisateurs et pour la plupart des études post-génomiques.

Les utilisateurs ont à présent à se débrouiller au sein d'une jungle de données annotées par des procédures automatiques. Ces annotations sont certes loin d'être parfaites, mais elles constituent une base de travail acceptable et utile pour qui dispose du minimum de sens critique nécessaire pour séparer le bon grain de l'ivraie.

Le déroulement du programme génome aura été marqué par une implication de plus en plus forte de l'industrie privée. Les données scientifiques et leur interprétation deviennent une marchandise. Cette tendance se maintiendra-t-elle ? On peut mentionner plusieurs sources de problèmes.

1. Les données (et leur interprétation) doivent pouvoir être remises en question. Seul un libre accès public permet de les évaluer et de le faire savoir. Souvenons-nous, qu'il y a quelques mois encore, la Société Incyte annonçait 150 000 gènes humains. S'il ne s'agit pas de publicité mensongère, on peut cependant observer que des scientifiques s'exprimant pour le compte d'une société peuvent avoir des difficultés à concevoir que leurs démarches expérimentales souffrent des mêmes limitations que dans le monde académique. On ne va quand même pas désespérer Wall Street en vertu du principe de précaution !

2. Il semble aussi que les clients des sociétés de génomique commencent à se lasser de payer au prix fort. D'où l'idée de considérer qu'il existe un champ pré-compétitif qui doit rester en libre accès. Le consortium SNP *(single nucleotide polymorphisms)* public-privé auquel participent une dizaine de laboratoires pharmaceutiques et plusieurs centres publics de séquençage est l'exemple le plus marquant de cette tendance.

3. Si les données sur le génome humain et sur quelques autres génomes peuvent se vendre au prix fort, il n'en sera pas toujours de même. C'est d'ailleurs parce que des utilisateurs privés étaient prêts à payer un prix important pour la séquence du génome humain que le projet a connu cette extraordinaire accélération. Dans l'*agro-business*, la situation est déjà différente puisque chacun des géants dispose de sa propre séquence du génome du riz et qu'il existe, de surcroît, un projet public. Les génomes à faible valeur marchande, qui ne sont pas nécessairement les plus inintéressants sur le plan scientifique, ne seront financés qu'avec des fonds publics.

On peut donc se demander ce que vont devenir, à terme, ces énormes capacités de séquençage. Si les besoins académiques demeurent importants pour les années à venir, il est plus difficile de prévoir ce qu'il en sera des financements.

Il reste maintenant à tenir les promesses. Mais celles-ci ont aussi évolué avec le temps. A l'image du déroulement inattendu du séquençage (*voir* l'article de B. Jordan, *p. 290 de ce numéro*), ce que l'on compte faire aujourd'hui en post-génomique est beaucoup plus ambitieux que ce qui était imaginable au moment de la genèse du projet.

Il ne fait de doute pour personne que l'inventaire exhaustif des gènes, facilitera considérablement le clonage positionnel. Pour ce dernier, la séquence change radicalement la donne. Il reste pourtant quelque 4 000 gènes de maladies monogéniques à identifier ; cela prendra du temps. Le récent isolement d'un variant de séquence intronique de la calpaïne 10 dans le diabète de type II chez certaines populations nous autorise à imaginer que ce succès n'est qu'un infime début à l'identification d'une multitude de variants de prédisposition. Mais ce résultat indique aussi que cette voie demeure difficile.

On place aussi beaucoup d'espoirs dans l'établissement de profils de transcription exhaustifs, qu'ils soient dérivés d'hybridations à des puces à ADN ou d'expériences de type SAGE. L'intérêt diagnostique dans la classification des tumeurs a déjà été montré. Sur le plan strictement physiologique, les variations phénotypiques individuelles risquent de sérieusement compliquer les choses. Un profil de transcription est un phénotype extrêmement détaillé, où il faudra surmonter le problème des variations individuelles et se débarrasser des signaux parasites par un découpage minutieux des échantillons tissulaires.

On a souvent coutume de dire que les expériences de génomique et de post-génomique ne s'appuient pas

sur des hypothèses. Mais existe-t-il des expériences sans hypothèse sous-jacente ? Lorsqu'on recherche des ligands potentiels d'une protéine dans une expérience de double hybride, on formule l'hypothèse selon laquelle cette protéine peut interagir avec d'autres. Il s'agit d'une hypothèse, certes fruste, mais la réponse peut être hautement significative et la chance d'obtenir un résultat inattendu est bien plus grande avec un *a priori* minimaliste. Ce qui est sans doute dérangeant, c'est que la conception de telles expériences est préfabriquée : il n'y a plus qu'à s'occuper de la mise en route. Mais cela est toujours moins trivial qu'il n'y paraît.

Quand on a à appréhender la fonction de 30 000 gènes et de plusieurs centaines de milliers de protéines dérivées il faut en passer par des expériences répétitives à grande échelle ; la connaissance est à ce prix. Inversement, les données recueillies peuvent apparaître monotones, mais elle sont beaucoup plus riches et imprévisibles qu'on l'imagine.

Nous en sommes comme au lendemain des grandes découvertes, nous connaissons grossièrement l'emplacement et l'étendue des terres et des mers, le monde devient fini mais il reste à l'explorer et à l'exploiter ■

TIRÉS À PART

J. Weissenbach.