m/s 2000

*médecine/sciences 2000 ; 16 : 10-6*

# Séquence des génomes : le feu d'artifice

**Jean Weissenbach**
**Marcel Salanoubat**

J. Weissenbach, M. Salanoubat : Genoscope, 2, rue Gaston-Crémieux, 91057 Évry Cedex, France.

▶ La publication de la séquence complète de deux génomes bactériens en 1995 marque l'entrée dans l'ère du séquençage à grande échelle. Depuis lors, les données de séquence s'accumulent à un rythme de plus en plus impressionnant, alors que la communauté scientifique n'est plus en mesure de procéder à leur analyse immédiate. Il y a un an, la séquence d'un premier génome de métazoaire était pratiquement achevée. Les séquences d'autres génomes d'eucaryotes multicellulaires (drosophile et arabette) sont très avancées. Durant l'année en cours, une version préliminaire de la séquence de l'essentiel du génome humain devrait devenir disponible. Elle sera suivie d'une version plus complète ainsi que d'une autre séquence du génome humain obtenue en parallèle par une entreprise privée qui utilise une stratégie très différente. Les capacités de séquençage constituées pour le génome humain pourront ensuite être mobilisées pour séquencer, en l'espace de quelques années, un certain nombre de grands génomes d'espèces modèles et d'espèces d'intérêt agronomique. ◀

Avec le basculement dans l'ère de l'après-génome la biologie vit une nouvelle révolution. Dans son utilisation première d'outil de la génétique – et alliée aux nouveaux outils informatiques – la séquence d'un génome est d'une puissance totalement inédite. Les séquences des premiers génomes complets, conjointement à d'autres avancées techniques, ont apporté beaucoup plus et, notamment, la possibilité d'appréhender pour la première fois une cellule, un organisme dans sa globalité, ce qu'on qualifie aujourd'hui de génomique fonctionnelle. Des travaux majeurs, comme les études d'associations génétiques à l'échelle du génome, l'analyse du transcriptome, sont aujourd'hui possibles, mais ne peuvent être envisagés qu'à partir de séquences. Dans ce qui suit, nous discuterons brièvement quelques événements clés et tendances des cinq dernières années et essaierons de faire un point sur la situation présente et une projection sur le futur immédiat en matière de séquençage de génomes.

## Une ère nouvelle de la biologie

Le chemin parcouru depuis une précédente *synthèse* de *médecine/sciences* sur ce sujet [1] est impressionnant *(figure 1)*. En 1995, aucune séquence complète n'était connue en dehors de celle des génomes viraux. Aujourd'hui, la séquence de plus d'une vingtaine de génomes procaryotes, d'une taille allant de un à quatre millions de paires de bases, est disponible publiquement, ainsi que celle de deux génomes eucaryotes. Un nombre encore plus important de séquences de génomes entiers est en chantier. Entre les premiers génomes viraux en 1977-1978 et les premiers génomes bactériens en 1995, la taille des réalisations a augmenté d'un facteur 1 000 en 20 ans. Avec la séquence du génome humain d'ici 2002-2003, un autre facteur 1 000 aura été gagné en moins de 10 ans. Cette accélération vertigineuse résulte en premier lieu d'une augmentation considérable des moyens financiers consacrés au séquençage. Elle n'est pas la conséquence d'une révolution méthodologique mais celle d'améliorations des techniques de séquençage, de leur automatisation plus poussée, d'une organisation spécifique concentrant les moyens au sein de grosses structures, les centres de séquençage et, enfin, d'un recours massif à des outils informatiques capables de traiter et de canaliser les flots de données brutes. Le séquençage des génomes est, de plus en plus, une activité industrielle où chaque étape doit être finement contrôlée. Tout biologiste qui visite un centre de séquençage pour la première fois est frappé par l'accumulation de machines, de terminaux d'ordinateurs, d'automates et, ergonomie oblige, par la faible densité en personnel.
Des moyens importants sont donc aujourd'hui consacrés au séquençage ; ils sont aussi l'objet de remises en cause périodiques. Face à des interrogations récurrentes sur l'intérêt de ces programmes coûteux, n'oublions pas ce qu'était la vie d'un biologiste moléculaire il y a dix à vingt ans, séquençant et alignant quasi manuellement

plusieurs milliers de nucléotides. En raison de la forte diminution des coûts dans les grands centres (de 4 à 7 centimes la base brute et de 1 à 2 F la base finie), il est aujourd'hui moins cher de séquencer à faible taux de couverture (1 à 3 x, *voir glossaire*) un génome bactérien inconnu pour y rechercher quelques fonctions, que de les isoler par des approches de génétique et de clonage moléculaire. De nombreux exemples récents de clonage positionnel montrent que le séquençage complet devient aussi la méthode de choix pour la recherche d'un gène de mammifère dans un intervalle de l'ordre du mégabase. Un pays comme la France consacre annuellement de l'ordre de 100 MF au séquençage massif et aux travaux périphériques (cartographie, analyse des données), c'est-à-dire à peu près ce que doivent coûter 120 à 150 chercheurs du secteur public. Il s'agit d'un effort significatif mais non démesuré et qui reste modeste en comparaison de celui des États-Unis ou du Royaume-Uni.

## Les génomes de micro-organismes

Les annonces successives de la séquence complète du génome d'*Haemophilus influenzae* [2] et de *Mycoplasma genitalium* [3] par une stratégie de séquençage aléatoire global *(voir glossaire, p. 15)*, sans cartographie préalable, furent une surprise considérable et permirent à Craig Venter et au laboratoire TIGR *(The Institute for Genomic Research)* de s'affirmer comme les *leaders* du nouveau domaine hautement technologique du séquençage massif. La génomique des micro-organismes est alors entrée dans une période de croissance exponentielle, avec plus de 25 génomes complets séquencés à ce jour (http://www.tigr.org), la plupart selon cette même stratégie de séquençage aléatoire globale. Bien que cela ait déjà été prévu par l'examen des séquences des chromosomes de levure connus, l'analyse de ces génomes complets a d'abord révélé un nombre considérable de gènes de fonction totalement ignorée, sans aucun homologue connu dans d'autres génomes. L'observation de nouveaux gènes dans les génomes récemment séquencés ne semble en outre guère se tarir et l'on peut se demander combien de nouvelles protéines totalement inédites restent à découvrir dans le monde des procaryotes.

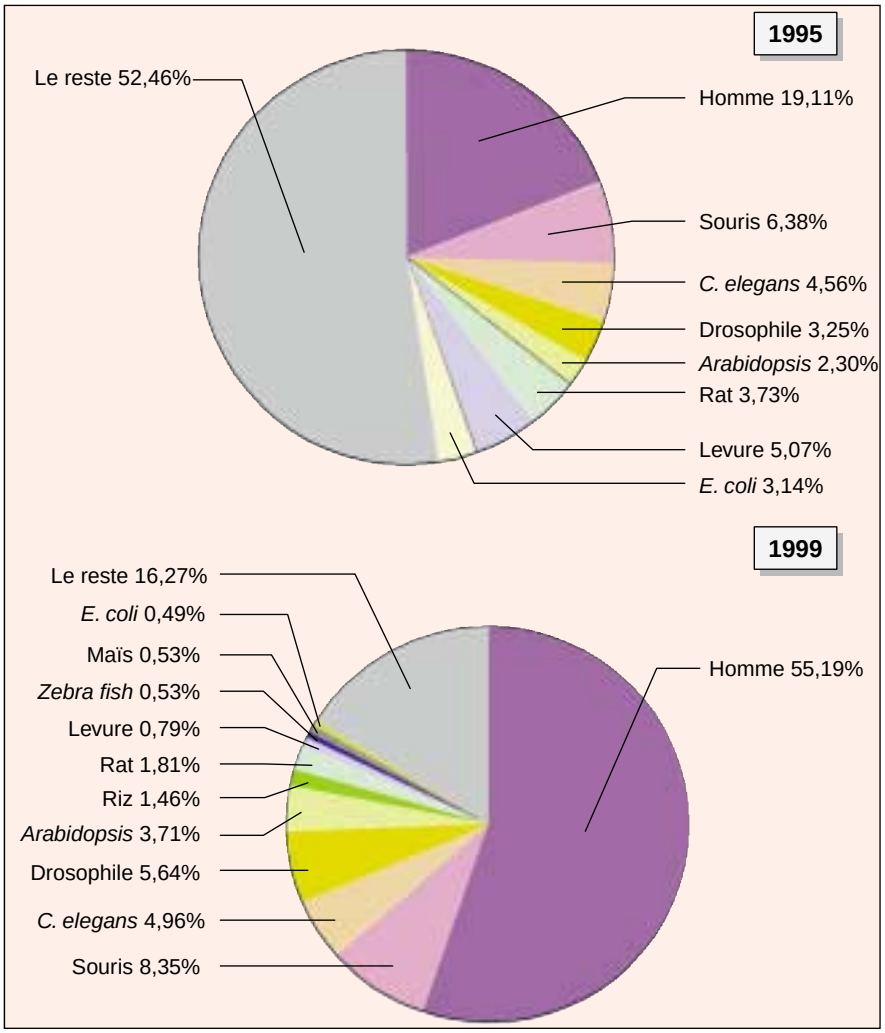


Figure 1. ***Évolution de la quantité de données relatives présentes dans les bases de données EMBL/GENBANK entre 1995 et 1999.*** *Les données de 1995 correspondent à celles publiées dans* médecine/sciences *[1]. Celles de 1999 correspondent à la* release 114 *accessible au site ftp://ncbi.nlm.gov/genbank/gbrel.txt. Cette* release *comprend les séquences génomiques et les EST.*

L'examen plus approfondi de ces génomes nous enseigne aussi qu'un nombre important de ces gènes inconnus codent en fait pour des enzymes catalysant des étapes du métabolisme primaire que nous ignorons ou des étapes connues qui sont catalysées dans d'autres espèces par des protéines sans homologie. D'autres gènes inconnus sont responsables de la synthèse de métabolites secondaires, également à découvrir, notamment en tant que sources potentielles de nouveaux antibiotiques

Le séquençage d'un génome bactérien devient un effort modeste à côté des projets portant sur les génomes de plantes et d'animaux. A côté des quelque 25 génomes publiquement disponibles, il en existe d'autres dans des bases de données privées et plus d'une cinquantaine de projets portant en majorité, mais pas exclusivement, sur des pathogènes, notamment des micro-organismes parasites eucaryotes, sont en cours. Devant les limitations de nos moyens de lutte actuels contre les espèces devenues résistantes à la quasi-totalité des antibiotiques en usage, l'exploration des génomes de pathogènes à la recherche de nouvelles cibles, de nouvelles voies métaboliques sensibles, est aujourd'hui une des pistes majeures pour la mise au point de nouvelles molécules anti-infectieuses.

L'analyse comparée des génomes de procaryotes a également permis d'observer que le processus de transfert horizontal entre micro-organismes ne se limite pas aux gènes de résistance aux antibiotiques. Ces transferts ont été nombreux au cours de l'évolution qu'ils ont probablement façonnée de manière importante [4]. L'étendue de cette diversité biologique a également été revue considérablement à la hausse ces dernières années. En effet, à la suite des travaux de Norman Pace et de ses collègues [5], on s'est aperçu que plus de 90 % des flores microbiennes de notre environnement sont constituées d'espèces (bactéries et *archaea*) non cultivables. Les résultats de l'ensemble de ces travaux s'appuient sur l'analyse comparative des gènes des ADN ribosomiques et, principalement, des gènes de l'ADN ribosomique 16S, que l'on peut amplifier par PCR *(polymerase chain reaction)* à partir d'ADN extrait de la flore microbienne de milieux naturels. Ils font apparaître en 1999 un domaine bactérien constitué de plus d'une trentaine de phylums [6] alors qu'à peine une douzaine était connue à la fin de la décennie précédente [7]. Un tiers de ces nouveaux phylums est exclusivement constitué d'espèces non cultivables, parfois très abondantes dans la nature, qui sont uniquement connues par leur ARN 16S. Ce monde ignoré est aujourd'hui accessible par les séquences d'ADN. Il est essentiel à notre environnement et nous concerne aussi directement puisque notre propre flore microbienne est, elle aussi, constituée dans sa très grande majorité d'espèces non cultivables [8]. Aujourd'hui encore, certains phylums du domaine des bactéries n'ont pas de représentant dont le génome soit en cours de séquençage, mais, au vu de la facilité avec laquelle un génome bactérien peut être séquencé, ces lacunes pourraient être rapidement comblées.

## Les génomes d'eucaryotes supérieurs

### Plantes : *Arabidopsis*, une mauvaise herbe utile

Le premier végétal supérieur dont le génome complet va être disponible au cours de cette année est celui d'une mauvaise herbe de la famille des crucifères, *Arabidopsis thaliana* (arabette). Cette angiosperme est utilisée comme plante dicotylédone modèle par des centaines de laboratoires. Les 135 mégabases (Mb) du génome d'*Arabidopsis* contiennent de l'ordre de 20 000 gènes [9]. En 1996, l'AGI *(Arabidopsis Genome Initiative)* a été créée à l'échelle internationale pour en organiser le séquençage [10] appuyé sur un ensemble de données génériques essentielles pour la conduite du projet. Ces données et celles disponibles précédemment ont permis de coordonner le séquençage de cette plante de manière exemplaire entre les partenaires.

Le séquençage de grands génomes, tant animaux que végétaux, repose classiquement sur la construction d'une suite ininterrompue de clones de chromosomes artificiels bactériens (BAC, *bacterial artificial chromosome*) ordonnés et dont les extrémités se recouvrent légèrement (une autre stratégie, proposée par Craig Venter, est celle du séquençage aléatoire global ; *voir glossaire, p. 15*). Les BAC sont ensuite séquencés indépendamment les uns des autres et la séquence complète du génome est obtenue par la réassociation de leurs séquences grâce aux régions en léger recouvrement et aux informations de localisation physique. Ces informations sont constituées de cartes physiques construites sur la base de données d'empreintes de restriction ou de contenu en STS *(sequence tagged sites)* des clones de BAC. Par ailleurs, en parallèle au séquençage, les séquences d'extrémités de BAC, selon le procédé des STC *(sequence tag connector)*, permettent de déterminer les clones ayant un chevauchement minimal [11]. Pour le génome d'*Arabidopsis*, la stratégie d'identification des BAC à séquencer combine les deux approches décrites ci-dessus. Chaque BAC, ainsi déterminé, est sous-cloné sous forme de petits fragments de plusieurs milliers de paires de bases. Le séquençage proprement dit commence alors par une phase de séquençage au hasard des sous-clones de ces BAC de façon à obtenir une couverture de 8 à 10 (séquençage aléatoire). Ces séquences sont assemblées à l'aide de programmes informatiques. A ce stade, la séquence est constituée de quelques régions continues sur plusieurs dizaines de kilobases (kb) (3 à 10 en général) par BAC. Vient ensuite la phase de finition, au cours de laquelle les lacunes sont comblées, les séquences de mauvaise qualité vérifiées et la qualité de l'ensemble contrôlée.

Cette stratégie permet d'obtenir rapidement (avant finition) l'essentiel des données pour l'inventaire des gènes. C'est la raison pour laquelle les séquences sont mises dans le domaine public dès les premiers essais d'assemblage. Un des enjeux majeurs demeure néanmoins la qualité. Il est crucial que la séquence soit complète et que la qualité soit aussi élevée que possible (les grands centres de séquençage se sont engagés à limiter à $10^{-4}$ leur taux d'erreur). La séquence devant être à la base d'autres travaux, une séquence de mauvaise qualité serait la source de fausses pistes, nécessiterait de coûteuses corrections, etc. C'est pour cette raison que beaucoup de soin est apporté à la finition, dont le coût est aussi élevé que la phase de séquençage aléatoire, qui pourtant donne 95 % des données.

Au 15 octobre 1999, la séquence du chromosome 2 et du chromosome 4 d'*Arabidopsis* a été déterminée dans sa quasi-intégralité [12, 13]. Concernant l'ensemble de ce génome, 52,5 % ont été séquencés et annotés, 26,2 % ont été séquencés et sont en cours d'annotation et 5,3 % sont sous forme de données préliminaires.

Bien que la totalité de la séquence ne soit pas encore disponible, un certain nombre de caractéristiques de ce génome peuvent déjà être mises en évidence, comme une densité en gènes élevée et quasi uniforme (un gène tous les 5 kb environ) et la présence de régions dupliquées sur différents chromosomes. Malgré les programmes extensifs de séquençage d'EST *(expressed sequence tags)* qui ont été entrepris chez cette plante, seuls 40 % à 50 % des gènes prédits possèdent un EST correspondant connu, ce qui souligne la nécessité du séquençage complet pour identifier l'ensemble des gènes d'un végétal supérieur. De plus, la moitié des gènes prédits n'ont pas de fonction connue.

La quantité de ressources engendrées fait d'*Arabidopsis* un système expéri-

mental de choix pour déterminer la « fonction » des gènes chez les végétaux [14]. Le défi auquel la communauté scientifique « végétale » est confrontée est de transposer ces connaissances aux plantes d'intérêt agronomique majeur, notamment les céréales. Cette transposition peut être effectuée sur la base de l'identité des séquences et de la conservation de l'ordre des gènes entre *Arabidopsis* et les céréales dont la divergence est estimée entre 130 et 200 millions d'années. Malheureusement, la conservation de synténie entre *Arabidopsis* et les céréales est probablement insuffisante pour être utile et il n'est jamais sûr de pouvoir distinguer entre orthologues et paralogues sans recours à la comparaison des séquences complètes. Or, cette distinction est essentielle pour assigner, à partir de la fonction décrite chez *Arabidopsis*, un rôle potentiel au gène de céréale correspondant, qui restera cependant à confirmer expérimentalement.

C'est pour cette raison qu'il est apparu nécessaire de séquencer le génome d'une céréale. Le choix s'est porté sur le riz en raison de son importance agronomique (la production est supérieure à 500 millions de tonnes par an) et de la taille relativement réduite de son génome (430 Mb) comparée à celle d'autres céréales comme le maïs et le blé, respectivement 6 et 40 fois plus grande. Le projet de séquençage du génome du riz est en cours d'organisation. Cette organisation se calque sur le modèle de l'AGI avec le choix d'une lignée de riz commune et la constitution de données génériques. Le consortium de séquençage qui regroupe le Japon, la Chine, les États-Unis, Taiwan, Singapour, la Thaïlande, l'Inde et la France devrait achever le séquençage de ce génome pour l'an 2004.

## Animaux

Beaucoup a déjà été dit sur la séquence du génome du nématode *(m/s 1999, n° 5, p. 695)*. Le génome de la drosophile ne devrait pas tarder à être connu *(voir plus loin)*.

La séquence du génome de la souris sera l'outil le plus utile pour l'interprétation de la séquence du génome humain en s'appuyant en particulier sur des comparaisons de séquence. Ce séquençage a connu une importante accélération depuis 1997 (près de 25 Mb à ce jour). Il devrait être mis en chantier à grande échelle dans la foulée du génome humain et profiter des importantes capacités de séquençage constituées en 1999. Le NIH *(National Institutes of Health)* envisage de financer un programme de séquençage complet à partir de l'an 2000, en commençant, à hauteur de 20 millions de dollars, par un premier programme destiné à préparer le terrain et acquérir les données génériques nécessaires au projet.

## Génome humain

### Sortie de la cartographie

Si le début des années 1990 a été marqué par une très intense activité dans le domaine de la cartographie, celle-ci s'est fortement ralentie à partir du milieu de la décennie, mais elle a aussi considérablement évolué. Cette évolution est pour partie la conséquence de l'amélioration des techniques de séquençage qui permettent aujourd'hui d'obtenir, aisément et massivement, des données de séquence servant à définir des réactifs de cartographie (STS, STC).

Avec près de 10 000 marqueurs microsatellites *(voir glossaire, p. 15)*, la densité de la carte génétique de deuxième génération est largement suffisante pour localiser aisément un gène de maladie monogénique avec une précision de 1 à 2 millions de paires de bases. Une carte génétique de troisième génération, constituée de polymorphismes nucléotidiques simples (SNP, *single nucleotide polymorphisms*), est en cours de constitution. Bien que nettement moins informatifs que les microsatellites, les SNP sont très abondants. Ils sont distribués au hasard, à une fréquence moyenne d'environ une variation pour 1 000 bases entre deux chromosomes homologues, et se prêtent à des techniques de criblage intensif sur des puces à ADN [15]. Leur moindre « informativité » peut donc être largement compensée par la quantité. De plus, pour remplir sa fonction, qui est de mettre en évidence les associations de SNP avec des gènes de maladies communes, cette carte doit reposer sur une densité très élevée de marqueurs [16]. Devant l'ampleur de la tâche, un consortium international, constitué de 10 des plus grands groupes de l'industrie pharmaceutique et de plusieurs centres de séquençage a décidé de coordonner un projet visant à identifier et à cartographier 300 000 SNP collectés de manière aléatoire. La cartographie de ces SNP pourra se faire sur la version préliminaire de la séquence du génome humain actuellement en cours. Cet outil, dont les données seront publiques, pourra ensuite être utilisé pour des travaux de recherche de gènes de prédisposition à des maladies communes.

La carte physique, fondée sur des ensembles ordonnés et chevauchants de chromosomes artificiels de levure (YAC, *yeast artificial chromosome*) dépasse les 90 % de couverture du génome. Malheureusement, ces YAC présentent de très nombreux réarrangements et ne peuvent donc être retenus comme matériel de départ pour la séquence. Les collections de grands fragments d'ADN génomique sont à présent établies à partir de banques de BAC, beaucoup moins remaniés que les YAC. Il existe actuellement deux banques de BAC, correspondant à une couverture de 20 à 25 « équivalents-génome » chacune, mais les cartes de préséquençage ne sont disponibles que pour une fraction (30 % à 40 %) du génome. Ces cartes se sont faites au fur et à mesure des progrès du séquençage [17]. En raison des changements récents de stratégie *(voir plus loin)*, la progression de la carte de BAC se fera pratiquement en parallèle de la progression de la séquence d'une manière très semblable à celle décrite ci-dessus à propos des plantes.

On dispose aujourd'hui, grâce aux programmes de séquençage massif de banques d'ADNc, de plus de 1 600 000 séquences d'EST correspondant à environ 60 000 gènes [18]. Un réseau international de laboratoires américains et européens a procédé à la cartographie de ces EST à l'aide des collections d'hybrides d'irradiation *(voir glossaire, p. 15)*, particulièrement utiles pour intégrer les gènes aux cartes existantes [19, 20]. Un total d'environ 30 000 gènes a ainsi été cartographié à l'aide de ces hybrides d'irradiation sur une carte

constituée de 1 000 intervalles délimités par des marqueurs de la carte génétique [21]. La carte de ces produits d'expression devrait en particulier faciliter la recherche et l'identification de gènes de maladies. Mais les approches s'appuyant sur la localisation des collections de séquences exprimées ne permettent pas d'aboutir à un inventaire complet des gènes. Seule la séquence génomique devrait, à terme, permettre de dresser un inventaire complet des gènes.

### Accélération du séquençage

Les premiers grands centres de séquençage publics et privés ont commencé à apparaître vers 1993-1994 (*Sanger Centre,* Université de Washington, TIGR, HGS, Incyte). John Sulston et Bob Waterston, respectivement directeurs du *Sanger Centre* et du *Genome Sequencing Center* de l'Université de Washington à St Louis (USA), ont proposé dès 1994 un programme de séquençage de l'ensemble du génome humain, s'étalant sur une dizaine d'années, réparti entre un petit nombre de centres. Vers 1994-1995, ces mêmes centres ont mis en route des projets importants (plusieurs mégabases) de séquençage du génome humain alors que moins du quart du programme de séquençage du nématode était réalisé. Fin 1995, le *Wellcome Trust* (la fondation caritative qui finance le *Sanger Centre* et qui n'a qu'un lien historique avec Glaxo-Wellcome) annonça le financement du séquençage d'un sixième du génome humain au prix d'1 F la base alors que ce n'est qu'à présent qu'on a atteint des coûts aussi bas. Il s'agissait donc à l'époque d'un pari, et c'est sans doute une des raisons pour lesquelles le financement a été si lent à s'établir de façon massive aux États-Unis. Ce programme public a néanmoins débuté avec une forte coordination entre les Américains et les Britanniques. Le rôle *leader* des Britanniques dans le soutien et les initiatives sur le projet est à souligner. On peut aussi sérieusement s'interroger sur la manière dont les États-Unis auraient laissé le libre accès aux données de séquence sur le génome humain sans cette participation massive du *Wellcome Trust*. D'autres pays se sont progressivement joints au projet de séquençage, il s'agit notamment du Japon (3 % à 5 %), de l'Allemagne (2 % à 3 %), de la France (2 % à 3 %) et, tout récemment, de la Chine (1 %). Le Centre National de Séquençage-Génoscope ne fonctionne que depuis 1998 et ce n'est que depuis 1999 qu'il contribue au projet génome humain en séquençant la plus grande partie du chromosome 14.

### Coup de pied dans la fourmilière

En mai 1998, Craig Venter et le groupe Perkin-Elmer annonçaient leur intention de séquencer le génome humain par une approche de séquençage aléatoire de l'ensemble du génome. Une nouvelle entreprise, Celera-Genomics, a été constituée à cette fin. Elle est équipée de 300 exemplaires d'un nouveau séquenceur multicapillaire (ABI3700) capable de produire 5 à 8 fois plus de données brutes que les instruments précédents, pour un coût d'exploitation légèrement réduit, en raison d'une beaucoup plus grande automatisation et d'une moindre consommation de réactifs. Ce projet devrait être réalisé pour la fin de 2001 dans cette nouvelle entreprise, et financé par des fonds privés. Pour attirer les investisseurs, la rentabilité financière de l'ensemble sera assurée par la constitution d'un portefeuille de quelques centaines de brevets portant sur des utilisations de gènes identifiés au cours du déroulement du projet. L'entreprise a débuté par un projet pilote à la mesure : le génome complet de la drosophile (qui est en cours d'assemblage). Celera-Genomics produira en un an (à partir de l'automne 1999) l'ensemble des données brutes nécessaires : 30 milliards de bases (soit une couverture de 10 x sous forme de 70 millions de lectures de fragments). Le temps restant serait consacré à la cartographie et à l'assemblage du très grand nombre de fragments obtenus. De manière à identifier un grand nombre de SNP, il est prévu de séquencer en parallèle le génome de plusieurs individus. Craig Venter a aussi accompagné le lancement en fanfare de l'opération Celera de violentes critiques, grandement injustifiées, du projet public.
A côté des chances de succès très discutées de l'opération Celera, ce projet fait surtout peser le risque de privatiser une grande partie des données de séquence du génome humain. A la suite de nombreuses discussions au sein de la communauté des « usagers » académiques, consécutives à l'annonce du projet de Celera, les acteurs principaux du projet public (*Wellcome Trust*, NIH, DOE-*department of energy*) ont reconsidéré leur stratégie et ont, en premier lieu, procédé à une augmentation substantielle des budgets des centres de séquençage publics (*Sanger Centre,* Royaume-Uni, MIT/Whitehead, USA, *St Louis University,* USA, etc.). Paradoxalement, alors que les attaques de Craig Venter visaient clairement à disqualifier et à provoquer l'abandon du projet public, c'est un fort accroissement du financement qui s'est produit : plus de 220 millions de dollars ont été dégagés pour l'obtention du *working draft*.
Il a été décidé que les centres de séquençage publics procéderont, de manière prioritaire, à la phase de séquençage aléatoire à faible couverture (5X) de BAC qui seront ordonnés parallèlement au séquençage, aboutissant ainsi pour le printemps 2000 à une version brute et non achevée de la séquence *(working draft).* Cela devrait permettre de disposer de données partielles mais suffisantes pour des projets de recherche de gènes responsables de maladies dans des régions données. C'est dans ce domaine, qu'à ce jour, l'absence de séquence pose les problèmes les plus criants. Cette version brute aurait donc pour intérêt majeur de satisfaire une partie des besoins immédiats de nombreux utilisateurs. Cependant, il est à craindre aussi que d'avoir fourni 80 % à 90 % des données ne démotive une partie des producteurs. Il y a donc un risque de rester avec une séquence inachevée du génome humain.

### Deux séquences pour le prix d'une ?

Ainsi, le *working draft* est en cours de réalisation. A ce jour, 44 % de la séquence du génome est disponible publiquement dont 16 % sous forme de séquence achevée (dont celle du chromosome 22 [22]) ou pratiquement achevée et 28 % sous forme de version préliminaire *(figure 2)*. Une version plus achevée à partir d'une couverture en séquence brute de 10 x devrait être dis-

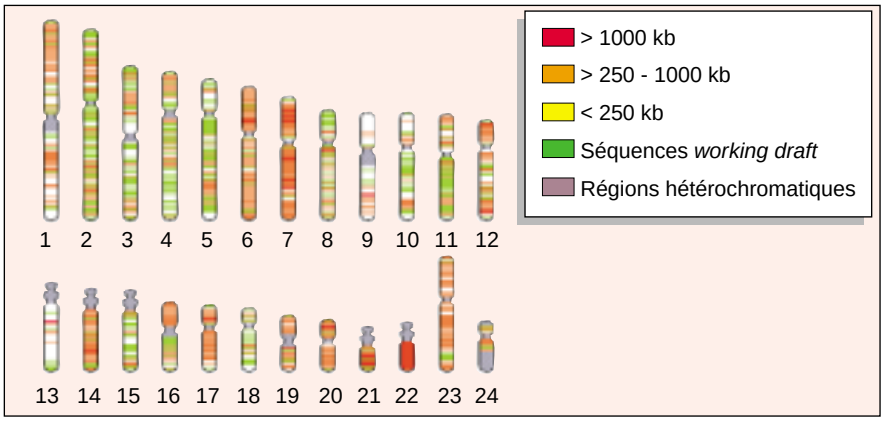


Figure 2. ***Progrès accomplis dans le séquençage du génome humain.*** *La coloration en rouge et orange indique la localisation des régions déjà séquencées et leur degré de continuité. La couleur verte correspond à des séquences de qualité* working draft *[23].*

ponible vers le début de 2001. La séquence complète est toujours prévue pour 2003. Notons que les bénéfices réalisés par Perkin-Elmer avec les ventes massives de réactifs et d'appareillage financés par le projet public sont recyclés dans l'opération Celera.

Il sera intéressant de voir la teneur du communiqué de victoire dans lequel Craig Venter annoncera qu'il a réussi à établir la séquence du génome de la drosophile, sachant que la séquence, actuellement déposée par Celera dans les bases de données, est constituée de fragments dont la taille moyenne est de 25 kb. Il a aussi indiqué que la séquence du génome humain déterminée par Celera sera accessible publiquement, mais nous ne savons ni dans quel délai ni sous quelle forme. Il s'agira sans doute d'une séquence consensus déduite d'un assemblage, mais nous ignorerons probablement la manière dont celui-ci a été établi. Selon ce qui fut annoncé, la séquence du génome humain de Celera devrait être constituée d'un ensemble qui ne dépasserait pas 5 000 fragments, soit environ 100 à 400 par chromosome. De surcroît, s'il est fait massivement usage du *working draft* pour l'assemblage de la séquence de Celera, nous n'aurons pas la démonstration de la faisabilité de l'approche aléatoire globale. Des discussions entre Celera et le NIH, sur une éventuelle fusion des projets, ont été entamées.

## Conclusions

Il n'est pas possible ici de traiter du problème de l'interprétation des séquences qui demeure, comme chacun sait, le prochain défi à relever. D'une manière générale, l'interprétation est avant tout le problème de l'utilisateur. Le producteur de données de séquence ne peut assurer qu'un niveau d'interprétation informatique fruste. Celle-ci repose soit sur des comparaisons avec des séquences existantes, soit fait appel à des programmes de prédiction capables de reconnaître des éléments caractéristiques dans les séquences. Mais les données utilisées pour les comparaisons sont en augmentation constante et les programmes de prédiction évoluent avec le temps. Il est donc préférable de faire cette interprétation informatique au moment du besoin. Quoi qu'il en soit, cette interprétation doit être considérée comme une base qui doit permettre de déterminer la façon la plus efficace de réaliser le travail expérimental pour parvenir à l'identification et à la détermination de la structure et de la fonction des gènes ■

### * GLOSSAIRE *

***Stratégie de séquençage aléatoire :*** *tout processus de séquençage génomique commence par une phase aléatoire où les sous-clones correspondant soit à un BAC, soit à un génome complet sont séquencés à leurs deux extrémités sur une longueur de 500 à 900 paires de bases. Le terme aléatoire indique que l'assemblage de ces séquences ne nécessite aucune information sur la position relative des différentes séquences ainsi déterminées. L'assemblage est fondé sur les régions communes contenues dans les séquences obtenues aléatoirement. Il est à noter que de telles régions communes ne peuvent se trouver que si une paire de base déterminée a été séquencée plusieurs fois, ce qui explique la couverture utilisée (6 à 8) pour la phase de séquençage aléatoire. Jusqu'à présent le séquençage aléatoire a été utilisé avec succès pour des régions d'une taille comprise entre 100 kb (un BAC) et quelques mégabases (génomes bactériens par exemple). Classiquement, pour des génomes de plus grande taille, le séquençage s'effectue BAC à BAC (voir* Arabidopsis *et génome humain projet public). Craig Venter, dans son approche de séquençage aléatoire global, fait l'hypothèse d'une extension de cette approche à des génomes de grande taille.*

***Couverture n x :*** *n représente le nombre moyen de fois qu'une base est présente dans une banque de clones ou a été lue au cours d'une phase de séquençage aléatoire. Synonymes : génomes équivalents, profondeur.*

***Hybrides d'irradiation :*** *lignées cellulaires, ayant intégré des fragments de chromosomes humains et provenant de la fusion de cellules de rongeur avec des cellules humaines préalablement irradiées aux rayons X. La taille des fragments chromosomiques intégrés dépend de la dose de rayons X auxquelles les cellules humaines ont été soumises.*

***Microsatellites :*** *courtes répétitions en tandem de séquences très simples (le motif de base est composé de di-, tri-, tétranucléotides, etc.) répétées 10 à 30 fois, et disséminées en de nombreux emplacements distincts dans le génome. Un nombre différent de répétitions du motif de base est fréquemment observé sur les deux allèles d'un individu ou entre les individus.*

## RÉFÉRENCES

1. Weissenbach J. Le génome humain entre médecine et science. *Med Sci* 1995 ; 11 : 317-23.
2. Fleischmann RD, Adams MD, White O, *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995 ; 269 : 496-512.
3. Fraser CM, Gocayne JD, White O, *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science* 1995 ; 270 : 397-403.
4. Koonin EV, Tatusov RL, Galperin MY. Beyond complete genomes : from sequence to structure and function. *Curr Opin Struct Biol* 1998 ; 8 : 355-63.

## RÉFÉRENCES


5. Pace NR. A molecular view of microbial diversity and the biosphere. *Science* 1997 ; 276 : 734-40.
6. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* 1998 ; 180 : 4765-74.
7. Woese CR. Bacterial evolution. *Microbiol Rev* 1987 ; 51 : 221-71.
8. Suau A, Bonnet R, Sutren M, *et al.* Direct analysis of genes encoding 16S rDNA form complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* 1999 ; 65 : 4799-807.
9. Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M. *Arabidopsis thaliana* : a model plant for genome analysis. *Science* 1998 ; 282 : 662-82.
10. Bevan M, Ecker J, Theologis S, *et al.* Objective : the complete sequence of a plant genome. *Plant Cell* 1997 ; 9 : 476-8.
11. Venter JC, Smith HO, Hood L. A new strategy for genome sequencing. *Nature* 1996 ; 381 : 364-6.
12. Lin X, Kaul S, Rounsley S, *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 1999 ; 402 : 761-8.
13. The european union *Arabidopsis* genome consortium and the Cold Spring Harbor, Washington University in St Louis and PE Biosystems *Arabidopsis* sequencing consortium. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 1999 ; 402 : 761-8.
14. Bouchez D, Hofte H. Functional genomics in plants. *Plant Physiol* 1998 ; 118 : 725-32.
15. Wang DG, Fan JB, Siao CJ, *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998 ; 280 : 1077-82.
16. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 1999 ; 22 : 139-44.
17. Sanger Center, Washington University Genome Sequencing Center. Toward a complete human genome sequence. *Genome Res* 1998 ; 8 : 1097-108.
18. Hillier LD, Lennon G, Becker M, *et al.* Generation and analysis of 280 000 human expressed sequence tags. *Genome Res* 1996 ; 6 : 807-28.
19. Gyapay G, Schmitt K, Fizames C, *et al.* A radiation hybrid map of the human genome. *Hum Mol Genet* 1996 ; 5 : 339-46.
20. Hudson TJ, Stein LD, Gerety SS, *et al.* An STS-based map of the human genome. *Science* 1995 ; 270 : 1945-54.
21. Deloukas P, Schuler GD, Gyapay G, *et al.* A physical map of 30 000 human genes. *Science* 1998 ; 282 : 744-6.
22. Dunham I, Shimizu N, Roe BA, *et al.* The DNA sequence of human chromosome *Nature* 1999 ; 402 : 489-95.
23. Jang W, Chen HC, Sicotte H, Schuler GD. Making effective use of human genomic sequence data. *Trends Genet* 1999 ; 15 : 284-6.


## TIRÉS À PART

J. Weissenbach.


m/s 2000

## Summary

### Genome sequencing : the Big Bang

The publication of the complete sequence of two bacterial genomes in 1995 marked the beginning of the era of large-scale sequencing. Since then, sequence data has been accumulating at more and more impressive rhythms, such that the scientific community can no longer keep pace by immediate analysis of the results. One year ago the sequence of the first metazoan genome was practically completed. Sequencing of other multicellular eukaryotic genomes *(Drosophila* and *Arabidopsis)* is at an advanced stage. This year, a « working draft » of the major part of the human genome sequence should become available. This will be followed by a more complete version, and by another human genome sequence produced by a private company using a quite different strategy. The sequence capacities that have been established for the human genome can later be mobilized for the sequencing, in the space of a few years, of several large genomes of model organisms and species of agronomic interest.