



Années Génome : les dix glorieuses

Bertrand Jordan

B. Jordan: Centre d'immunologie Inserm/Cnrs de Marseille-Luminy, case 906, 13288 Marseille Cedex 09, France.

► Dix ans après le démarrage effectif des programmes Génome, deux années après le véritable début du séquençage de notre ADN, nous jetons un regard sur cette intéressante période. Malentendus du début, focalisation sur les cartes, apparition inattendue de l'approche des EST, importance croissante des organismes modèles : cette décennie est riche en rebondissements et en évolutions majeures, notamment l'irruption de l'industrie dans le monde du Génome... au moment où « notre » séquence est en vue. ◀

1 988: les programmes Génome se mettent en place, après plusieurs années de débats lancés par le colloque de Santa Cruz (États-Unis) en mai 1985. La communauté scientifique, dans sa grande majorité, est sceptique. Des biologistes de renom (David Botstein, Robert Weinberg, entre autres) expriment même une opposition farouche: certains se convertiront bientôt. L'image publique du projet, véhiculée par les lobbyistes qui ont réussi à convaincre le congrès américain d'investir des sommes qui paraissent colossales (deux cent millions de dollars « qui seraient tellement mieux employés à faire de la vraie biologie»), apparaît très décalée par rapport à la réalité. Pour le journaliste ou l'homme de la rue (si tant est qu'il en ait entendu parler), c'est le « Grand programme du séquençage du Génome », projet international centré essentiellement sur la séquence. Et l'on croit que cette entreprise est strictement coordonnée par une sorte de comité central du Génome dont on attribuera bientôt le rôle à HUGO, la *Human Genome Organisation*, fondée cette année-là.

Les faits sont tout autres: plusieurs programmes nationaux démarrent effectivement, mais dans une grande dispersion accompagnée d'une vive concurrence: deux programmes aux États-Unis, ceux du *National Institutes of Health* et des *Department of Energy*, quatre ou cinq au Japon, un en Grande-Bretagne... Et surtout, le contenu effectif de ces travaux porte

pour l'essentiel sur la construction des cartes (carte génétique, carte physique), préalable incontournable au séquençage global, qui semble une entreprise chimérique ou en tout cas fort lointaine. C'est ce recentrage sur la construction des cartes, objectif plus réaliste et, par ailleurs, directement utile à la génétique médicale, qui entraînera le ralliement d'une partie de la communauté ou du moins une baisse sensible de son hostilité.

Un colloque fondateur

Le premier colloque *Genome Mapping and Sequencing* tenu à *Cold Spring Harbor* (New York, USA) au printemps 1988 reflète ces tendances. J'en rapportais quelques éléments dans une *Chronique génomique* avant la lettre (le début officiel de cette série se situant en 1989) intitulée *Chronique de Cold Spring Harbor* [1]. Un peu plus de dix ans après, il est instructif de la relire et de consulter le livre des *abstracts* correspondant. Plusieurs éléments sont frappants. Le séquençage n'occupe qu'une demi-journée, la matinée du dimanche (lorsque beaucoup de participants se préoccupent surtout de rejoindre l'aéroport Kennedy à l'aide d'incertaines limousines aux horaires aléatoires), et est vu sous l'angle des premières machines (aujourd'hui si courantes) et de quelques tentatives aux ambitions à la fois modestes dans l'absolu et démesurées par rapport aux réalisations de l'époque. Déchiffrer une ou deux mégabases dans une

région du génome particulièrement intéressante: ces tentatives (comme celle, à l'époque, du CEPH – Centre d'étude du polymorphisme humain – qui se proposait de séquencer les deux ou trois mégabases du complexe majeur d'histocompatibilité, ou celle du laboratoire de Fred Blattner qui, avec le groupe japonais de Katsumi Isono à Kobé, se lançait très logiquement dans le déchiffrement des quatre mégabases du génome d'*Escherichia coli*) vont échouer lamentablement. Les problèmes d'organisation, de gestion des flux, de contrôle de qualité, les efforts nécessaires à l'identification et à l'élimination des goulots d'étranglement dans une entreprise à l'échelle industrielle avaient été largement sous-estimés. Ils entraînent l'arrêt de ces projets après une ou deux années au cours desquelles les séquences obtenues représentaient un cinquième ou un dixième des prévisions, tandis que les coûts, eux, dépassaient allègrement le mythique « un dollar par base » que l'on avait pourtant espéré réduire rapidement à un chiffre dix fois plus faible [2].

Le séquençage intégral du génome humain semblait reporté aux calendes grecques, et beaucoup (moi-même compris) doutaient fort qu'il soit effectué un jour – compte tenu de l'énormité de la tâche, des difficultés induites par la multiplicité des séquences répétées, et du maigre intérêt des 95 % de séquences non codantes qui encombrant notre ADN. A moins qu'une technique révolutionnaire ne multiplie subitement par dix ou cent la vitesse d'obtention des données tout en réduisant considérablement les coûts? Aucune des approches sur lesquelles reposaient à l'époque nos espoirs (lecture directe par microscopie en champs proches, analyse à très haute sensibilité de bases excisées d'une molécule d'ADN, ou même séquençage par hybridation) ne s'est aujourd'hui encore révélée réellement opérationnelle, elles étaient d'ailleurs pratiquement absentes de cette première édition du colloque. Si aujourd'hui le séquençage est réellement engagé, avec déjà près de 30 % de nos trois mille mégabases déchiffrées au cours des deux dernières années, c'est à l'aide de la

méthode publiée dès 1977 par Fred Sanger. Méthode rendue plus rapide, plus puissante, grâce à une multitude de petites améliorations, à une automatisation encore partielle, et surtout à une organisation de type industriel, quasiment militaire: les grandes « usines à séquencer » regroupent des dizaines ou même des centaines de machines, un personnel technique spécialisé très encadré et une informatique performante, le tout dans le cadre de projets très définis et structurés de manière rigoureuse.

Une autre absente de taille en 1988: l'approche des EST (*expressed sequence tags*) (séquences partielles de milliers de clones d'ADNc), initiée deux ans plus tard par le précurseur Craig Venter à contre-courant de la doctrine officielle, révélée au grand jour en 1991 par le « scandale » des brevets sur ces entités, et qui aujourd'hui, avec trois millions d'EST dans le domaine public et bien plus dans les bases de données industrielles privées, occupe une place centrale dans la plupart des programmes de Génomique. Bel exemple des bouleversements que peut introduire une idée originale, et de l'impossibilité de planifier strictement une entreprise pourtant aussi routinière et stupide (d'après ses détracteurs) que l'établissement de l'anatomie du génome humain...

Les cartes avant la séquence

Le colloque de 1988 était donc majoritairement consacré aux cartes. La première carte génétique de l'ensemble du génome humain, coordonnée par Helen Donis-Keller et qui fit la couverture de *Cell* en 1987, n'avait pas apparemment suscité beaucoup d'émules: seuls deux ou trois *abstracts* font état d'efforts dans ce domaine: enrichissement de la carte générale, ou cartographie plus précise d'un ou deux chromosomes individuels. Son deuxième souffle, fondé sur la découverte et l'utilisation massive des microsatellites, plus fréquents, plus polymorphes et se prêtant mieux à une exploitation semi-automatisée que les incommodes RFLP (*restriction fragment length poly-*

morphisms), n'était pas encore manifeste. Il faudrait attendre l'aventure du Généthon, l'investissement massif de l'AFM (Association française contre les myopathies) grâce à l'impulsion de Bernard Barataud, l'audace de Daniel Cohen et la rigueur de Jean Weissenbach pour qu'apparaisse, en 1992, la première carte de deuxième génération qui devait, entre autres, multiplier instantanément par dix le rythme de la localisation de maladies génétiques. La carte génétique de troisième génération, éditée sur l'exploitation intégrale du polymorphisme humain, est aujourd'hui bien engagée: la découverte et la détection des SNP (*single nucleotide polymorphisms*), pour laquelle plusieurs techniques d'analyse performantes s'affrontent encore, a déjà permis de positionner plusieurs dizaines de milliers de balises et promet l'obtention de la « carte à cent mille marqueurs » indispensable pour l'élucidation du déterminisme des maladies multigéniques et pour de nombreuses approches de « pharmacogénétique » [3]. Comme pour les EST, secteur public, entreprises industrielles, et aussi consortiums public-privé contribuent à l'accumulation de ces données non sans concurrence, duplication d'efforts et coups médiatiques dirigés plus vers les investisseurs ou la Bourse que vers la communauté scientifique.

Les cartes physiques, en revanche, étaient au centre du débat en 1988 et en constituaient l'objet principal. Deux approches s'opposaient: les cartes « par le haut » (*top down*) et « par le bas » (*bottom up*). Par le haut: l'invention en 1984 par David Schwartz et Charles Cantor de la séparation de très grands fragments d'ADN par électrophorèse en champs pulsés et la découverte des enzymes de restriction à site rare ouvraient, en principe, la possibilité de construire des cartes de restriction de chromosomes entiers. En fait, ces cartes, peu fiables en raison de leur sensibilité aux polymorphismes et à la méthylation de l'ADN, devaient s'avérer très délicates à construire et encore plus à exploiter, et leur contribution au balisage de notre génome a été modeste, sans commune mesure avec les efforts déployés. L'approche « par le bas » était la bonne: la constitution

de collections de clones contigus, des *contigs*, couvrant de grandes régions et si possible des chromosomes entiers, pouvait rendre accessible à une expérimentation détaillée tout point du génome (élément crucial pour la génétique médicale) et préparer la voie au séquençage. Mais caractériser et ordonner cinq ou dix mille cosmides pour « couvrir » le chromosome 16 ou 19 (tâche à laquelle s'étaient attelés les laboratoires de Robert Moyses et Anthony Carrano sous l'égide du *Department of Energy*) n'était pas une mince affaire, et beaucoup doutaient que ce soit possible. De nouveaux systèmes de clonage permettant de propager de grands fragments d'ADN allaient faciliter la tâche.

L'arrivée des chromosomes artificiels de levure, les YAC (*yeast artificial chromosomes*) (publiés en 1987 par David Burke, Georges Carle et Maynard Olson, mais dont la mise en place effective dans les laboratoires concernés devait prendre beaucoup de temps), et plus tard celle des P1 et des BAC (*bacterial artificial chromosomes*), vecteurs de clonage bactériens plus maniables et plus fidèles, allait sauver les cartes physiques, aujourd'hui établies de manière à peu près sûre, sur l'ensemble du génome après quelques tentatives très médiatisées mais peu opérationnelles. Le séquençage impose encore l'affinement et la fiabilisation de ces cartes (obtention de cartes « prêtes à séquencer », *sequence-ready*), généralement effectué au sein des grands centres de séquençage qui ont pris en charge le déchiffrement de tel ou tel chromosome – à moins de faire l'impasse sur cet aspect et de tenter, comme le fait Craig Venter, le séquençage en *shotgun* intégral (on séquence au hasard, on assemble ensuite), approche qui a montré sa puissance pour les petits génomes, et qui, pour les plus grands (drosophile, homme, souris), s'appuiera sans doute (sans trop le dire) sur des éléments de cartographie et de séquence déjà présents dans les bases de données publiques...

Aujourd'hui plus personne (ou presque) ne doute de l'importance – de l'intérêt – du caractère opérationnel des données sur les génomes. Les génomes : les organismes-modèle, bactéries, levure, nématode, *Arabi-*

dopsis thaliana, drosophile et j'en passe, ont pris une place qui n'était pas évidente il y a dix ans.

Place psychologique : le succès du Projet levure, pourtant éparpillé entre une cinquantaine de laboratoires européens et objet à son début des sarcasmes de nos collègues d'outre-Atlantique, a redonné confiance aux séquenceurs en montrant qu'il était effectivement possible de déchiffrer intégralement plus d'une dizaine de mégabases.

Place scientifique surtout, dans la mesure où la parenté des gènes d'organismes très éloignés dans l'échelle de l'évolution s'est révélée forte (grâce aussi à des techniques bio-informatiques de comparaison de plus en plus performantes), allant parfois jusqu'à une interchangeabilité fonctionnelle dont on n'osait pas rêver.

Il n'est pratiquement plus de projet de recherche biologique qui ne fasse appel à un moment ou à un autre aux données génomiques accessibles grâce aux bases de données disponibles sur Internet ; beaucoup sont même largement fondés sur leur exploitation à l'aide d'une bio-informatique très diversifiée, dont l'épanouissement a été accompagné par une fantastique accélération des performances des ordinateurs et un extraordinaire développement du réseau Internet. Bio-informatique presque absente en 1988 (aucune communication sur ce sujet, et à peine trois *posters* sur une quarantaine), et qui joue aujourd'hui un rôle central dans toute la génomique.

Les industriels s'intéressent de près au génome

L'irruption du secteur privé est sans doute le changement le plus fondamental apparu au cours de cette décennie du Génome. Certes, l'industrie n'était pas totalement absente à *Cold Spring Harbor* en avril 1988 : n'oublions pas que le travail d'Helen Donis-Keller était effectué dans le cadre de la firme *Collaborative Genetics*. Mais cela semblait presque une bizarrerie à l'époque, et n'a d'ailleurs pas porté bonheur à la firme en question. L'ambiance du colloque était encore

très universitaire. Aujourd'hui l'industrie détient sans doute plus de données génomiques (séquences de bactéries et de pathogènes végétaux, EST d'homme, de souris, de rat, collections de SNP...) que les bases publiques. Certaines de ces informations sont rendues disponibles au bout d'un certain délai, après que les brevets aient été déposés ; d'autres restent confidentielles et sont parfois rétablies par la recherche académique.

L'industrie pharmaceutique (mais n'oublions pas non plus les firmes agro-alimentaires) s'est lourdement investie dans l'exploitation (et parfois l'acquisition) de données sur les génomes : effort compréhensible en regard de l'accélération attendue dans la découverte de nouveaux médicaments pour lesquels les dépenses de recherche et développement sont de l'ordre de cinq cents millions de dollars par molécule. Que représentent quelques millions investis ou payés à un sous-traitant s'ils permettent de hâter ce processus, de mieux choisir les cibles et donc de réduire le coût total ? L'investissement peut être direct, en développant les programmes de recherche génomique *in house* ; il est le plus souvent indirect, prenant la forme de contrats avec des *start-ups* spécialisées dans tel ou tel aspect de la génomique : Incyte et beaucoup d'autres aux États-Unis, Genset chez nous. Ces importantes mises de fonds démontrent (s'il en était encore besoin) l'intérêt des données obtenues sur les génomes et la perspicacité de ceux qui ont lancé ces projets dans les années 1980 ; elles posent aussi de redoutables problèmes aux équipes académiques qui se retrouvent parfois en concurrence directe avec un département de recherche industriel sans disposer des mêmes ressources, notamment en bio-informatique ou pour la mesure en grand de l'expression des gènes, deux approches centrales en génomique.

Les années « biologie »

Les années qui viennent, on peut l'affirmer sans risque, seront celles de la biologie. Le seul phénomène scientifico-industriel d'importance comparable est celui de l'informa-

tique et du réseau Internet. Encore s'agit-il là avant tout d'impact sur la société, sur son mode de fonctionnement, sur le positionnement des individus. La génomique, tout en ayant des conséquences sociétales du même ordre de grandeur, constitue de plus un enjeu de première dimension au niveau cognitif: la compréhension du vivant, appuyée maintenant sur une connaissance précise et détaillée de sa base génétique, va se développer dans des directions dont nous pouvons imaginer les grandes lignes (développement, neurobiologie, réseaux d'interactions) mais dont les révélations sont par nature imprévisibles. Les années 1980 et 1990 ont été une époque fabuleuse pour la biologie: attendons-nous, dans les années 2000, à d'autres progrès, d'autres

surprises... et bien sûr aussi à d'autres problèmes ! ■

RÉFÉRENCES

1. Jordan B. Cartographie et séquence du génome humain: chronique de *Cold Spring Harbor Laboratory*. *Med Sci* 1988; 4: 448-50.
2. Wada A. Automated high-speed DNA sequencing. *Nature* 1987; 325: 771-2
3. Jordan B. Du Programme Génome à la « pharmacogénomique ». *Med Sci* 1997; 13: 1176-8.
4. Jordan B. Séquençage génomique: le deuxième souffle. *Med Sci* 1992; 8: 854-7.

MS2000

Summary

Genome, the ten-year itch

Genome projects really started ten years ago, and all-out sequencing of our DNA is just two years old: now is a good time to assess this very interesting period. It has been quite eventful; misunderstandings at the start, strong subsequent focus on maps, unplanned appearance of the very fruitful EST approach, increasing importance of model organisms... This decade has been rich in turning points and significant changes, such as the recent major impact of Industry in the Genome world... just as « our » sequence is around the corner.

Cours de Biologie Moléculaire de la Cellule

Enseignement pratique

13 mars - 14 avril 2000

Ce cours, conjointement organisé par l'Institut Pasteur et l'Institut Curie, se déroulera du 13 mars au 14 avril 2000 à plein temps, à l'Institut Pasteur à Paris. Il est destiné à des chercheurs du secteur public et privé, ayant une formation des facultés de sciences, de médecine, de pharmacie ou des écoles vétérinaires. Les candidats doivent avoir une bonne connaissance, niveau maîtrise, en biologie moléculaire. Les techniques de base de biologie moléculaire ne seront pas enseignées (exemple : clonage, séquençage de gènes, etc.). Ce cours donne lieu à un diplôme de l'Institut Pasteur suite à un examen qui se déroulera à la fin du mois d'avril.

Le thème central de ce cours concerne l'étude de la cellule eucaryote. Cet enseignement est très orienté vers l'initiation expérimentale, et fera une large place aux nouvelles techniques ainsi qu'à la démarche scientifique actuelle pour l'étude des fonctions cellulaires. Les travaux pratiques seront accompagnés de conférences théoriques sur les thèmes suivants :

- Organisation fonctionnelle de la cellule : compartiments membranaires, cytosquelette, polarité cellulaire
- Les routages intracellulaires : transport des protéines membranaires et sécrétées, endocytose des macromolécules
- Les contacts et la communication entre cellules
- La différenciation cellulaire
- La signalisation et la transduction des messagers cellulaires
- Le cycle cellulaire.

Les techniques mises en œuvre seront celles de l'analyse génétique, la transfection et l'expression de gènes clonés, la culture cellulaire, la reconstitution *in vitro* des fonctions cellulaires, la visualisation des constituants cellulaires y compris par les techniques les plus récentes de microscopie confocale et d'imagerie.

Avec la participation de : Ch. Babinet, M. Bornens, P. Bregestovski, R. Bruzzone, P. Cossart, A. Dautry-Varsat, F. Dautry, M. Dubois-Dalq, S. Dufour, E. Fabre, B. Goud, B. Hoflack, C. Hopkins, L. Johannes, E. Karsenti, D. Louvard, P. Mangeat, U. Nehrbass, J.-C. Olivo, J. Pouyssegur, J.-P. Thiéry et M. Weiss. Les cours théoriques seront assurés par des enseignants français et européens.

Responsables du cours : A. Dautry-Varsat et D. Louvard.

Renseignements et Inscriptions

Mme Banisso, Secrétariat des Enseignements et des Stages
Institut Pasteur, 28, rue du Dr-Roux, 75724 Paris Cedex 15, France.
Tél. : 01 45 68 81 41 ou 01 40 61 33 62 - Fax : 01 40 61 30 46
E-mail : rbanisso@pasteur.fr