CHRONIQUES D'UNE SÉQUENCE ANNONCÉE

1992-2002 : dix ans de programmes Génome

Ce recueil est dédié à Axel Kahn, qui m'a « lancé » dans l'écriture, à François Flori qui, non content de suivre ces chroniques depuis des années, a veillé sur la réalisation de ce livre, et surtout à Anne, dont le patient travail de relecture critique a bénéficié à beaucoup de chroniques et à tous les commentaires qui les accompagnent.

CHRONIQUES D'UNE SÉQUENCE ANNONCÉE

1992-2002 : dix ans de programmes Génome

Bertrand Jordan

EDK
EDITIONS MÉDICALES ET SCIENTIFIQUES

Outre les *Chroniques Génomiques* et de nombreux articles de vulgarisation, Bertrand Jordan a publié « Voyage autour du Génome » (Éditions John Libbey Eurotext, 1993), fondé sur les chroniques décrivant un tour du monde qui faisait le point sur le projet Génome à ses débuts. D'autres ouvrages ont suivi, les plus récents sont « Les Imposteurs de la Génétique » (Éditions du Seuil, 2000), qui traite des déformations et déviations de la Génétique, « Le chant d'amour des concombres de mer » (Éditions du Seuil, 2002), méditation marine sur la nature, l'évolution et la condition humaine, et « Les marchands de clones » (Éditions du Seuil, 2003), mise au point sur les aspects scientifiques, économiques et sociétaux du clonage animal et humain.

Le dessin de couverture provient du livre de résumés du *Genome mapping and sequencing meeting* tenu en 1989 à Cold Spring Harbor Laboratory et est dû à James Duffy. Il est reproduit avec l'autorisation de l'auteur et de « Cold Spring Harbor Laboratory Meetings and Courses Programs ».

Éditions E.D.K.
10, villa d'Orléans
75014 PARIS
Tél. : 01 53 91 06 06

ISBN : 2-84254-089-1

SOMMAIRE

PRÉFACE

DES GÈNES ET DES HOMMES

Il est rare dans l'histoire qu'un même événement eut été annoncé par trois fois, à plusieurs années d'intervalle, simultanément par les Chefs d'États et de Gouvernements des pays dominants sur le plan scientifique et technique. Tel est le cas du séquençage du génome humain dont les résultats très avancés, puis presque complets, puis enfin pratiquement complets furent présentés en grande pompe à la presse en 2000, le 14 et 15 février 2001 et, enfin, en avril 2003 afin de coïncider avec le cinquantième anniversaire de la découverte de la structure en double hélice de l'ADN. Chaque fois, le caractère exceptionnel de cette médiatisation institutionnelle internationale fut justifié par le caractère symbolique, presque métaphysique, du déchiffrage par l'homme de son programme génétique, le *« blueprint of human life »* et par les extraordinaires bienfaits que ne saurait manquer d'entraîner cette performance technique et scientifique.

Bien avant cette débauche de dithyrambes et de spéculations, Bertrand Jordan fut un observateur avisé, lucide et distancié de l'entreprise, aussi bien dans ses dimensions techniques que politiques, économiques et éthiques. À la fin des années 1980, il avait entrepris une reconversion thématique de l'équipe qu'il dirigeait au Centre d'Immunologie de Marseille-Luminy. Spécialiste des gènes codant les protéines du complexe majeur d'histocomptabilité chez l'homme, thème en parfaite cohérence avec les priorités d'un Institut d'Immunologie, Bertrand Jordan avait décidé de se consacrer plus à la génétique en elle-même. Il s'intéressait, en particulier, à la recherche par clonage positionnel de gènes de maladies génétiques à une époque où le premier grand succès en ce domaine, l'identification du gène modifié dans la myopathie de Duchenne, venait d'être publié. Un autre témoin de ce bouleversement scientifique était la revue *médecine/sciences*, créée à la suite d'une initiative conjointe franco-québécoise, en 1985. C'est au titre de membre fondateur et de Rédacteur en Chef de *médecine/sciences* que Bertrand Jordan me sollicitât afin de me présenter ses projets. Il envisageait en particulier de prendre une année sabbatique afin de pouvoir visiter, tout autour du monde, les principaux laboratoires engagés dans la « nouvelle recherche en génétique », ce que l'on allait bientôt appeler la génomique. C'est alors qu'émergea l'idée d'une collaboration régulière entre Bertrand Jordan et *médecine/sciences*, sous la forme de « chroniques génomiques » pour lesquelles j'inventais un logo évocateur, celui d'un individu stylisé arpentant à grands pas la double hélice d'ADN. Ces chroniques débutèrent avant que le tour du monde des grands laboratoires du domaine ne fût entrepris, mais servirent à accueillir, à chacune des étapes de celui-ci, les billets du Docteur Jordan. Ce voyage d'études se situait juste entre deux épisodes très significatifs qui firent entrer les études du génome dans le champ de la *« big science »*, c'est-à-dire celui d'une étude à grande échelle systématisée, optimisée et robotisée. Il s'agit du séquençage systématique des ADN complémentaires par Craig Venter, alors chercheur des NIH américains, puis de l'exploit, en 1992, de l'équipe de Daniel Cohen et Jean Weissenbach au Généthon d'Évry. Grâce à Bertrand Jordan, *médecine/sciences* devînt donc un outil privilégié de l'éveil du monde scientifique francophone à ce nouveau cours des choses. Il s'agissait en fait pour *médecine/sciences* d'une fonction logique puisque la revue avait été créée en 1985, l'année même de la publication dans *Nature* de l'identification formelle par l'équipe de Louis Kunkel, aux États-Unis, du gène muté ou délété dans la myopathie de Duchenne. C'est d'ailleurs grâce à la prise de

conscience par le directeur de l'Association Française de Lutte contre les Myopathies de l'époque, Bernard Barataud, du pouvoir de la génétique à la suite de cet exploit des américains que l'AFM s'engageait dans l'aventure du Téléthon et contribuait alors à écrire l'histoire de la génomique. L'actuel ouvrage de Bertrand Jordan commence au terme de ces prémisses, alors que se pose la question des étapes suivantes. C'est donc à un compte rendu détaillé des événements et péripéties ayant conduit au séquençage du génome humain qu'est convié le lecteur. Ce dernier pourra se familiariser avec l'évolution des stratégies expérimentales mises en œuvre, se plonger dans les arcanes de la compétition scientifique et industrielle, s'initier aux subtilités des relations complexes entre le monde scientifique et politique, prendre conscience, enfin, des incertitudes qui persistent quant à la signification réelle de ce long enchaînement d'un peu plus de trois milliards de paires de bases. Celles-ci sont en particulier illustrées par les débats actuels sur la diversité de la séquence du génome dans l'espèce humaine, et sur le nombre de gènes. Craig Venter, dirigeant l'effort de séquençage de la Société *Celera Genomics* publié le 15 février 2001 dans la revue *Science*, commenta les résultats obtenus sur un petit échantillon d'ADN appartenant à des personnes d'ethnies différentes. L'échantillon « caucasien », c'est-à-dire le sujet blanc d'origine européenne, s'avéra être le sien. On ne peut cependant en déduire que Monsieur Venter « se connaisse » pour autant mieux que le commun des mortels, marquant les limites « programmatiques » de la séquence... Quoiqu'il en soit, Craig Venter insista longuement sur le fait que la très grande homogénéité des séquences d'ADN de sujets asiatiques, africains ou européens disqualifiait les thèses racistes. En d'autres termes, la réfutation philosophique des discours inégalitaires fondant les thèses racistes pouvait maintenant laisser place à l'évidence scientifique offerte par le séquençage du génome qui devenait ainsi, vraiment, une œuvre d'une profonde dimension morale. Malheureusement, on pourrait aussi bien indiquer que l'étude du génome démontrait aussi qu'existaient en moyenne entre des individus jusqu'à 3 millions de différences... Lorsque l'on connaît les effets phénotypiques possibles d'une mutation unique, il y a bien évidemment là de quoi largement laisser place à tous les délires. D'ailleurs, nul ne conteste que les quelque 1,4 % de différences entre les génomes d'*Homo sapiens* et de *Pan troglodytes* aient des conséquences phénotypiques notables.

L'autre observation dérivée de l'analyse du génome humain et présentée comme une bouleversante surprise fut que l'on n'y pouvait détecter qu'un petit nombre de « gènes », de l'ordre de 25 000 à 30 000. Ce décompte, non encore définitif, repose avant tout sur l'identification des unités de transcription conservées entre espèces et ayant la capacité de coder des protéines. La stupeur déclenchée par cette nouvelle aussi bien que sa signification biologique méritent d'être commentées. Sur le premier plan, la métaphore d'un programme génétique vu comme un langage dont le gène constituerait les mots, et dont la signification serait ainsi largement combinatoire, conduit à n'être nullement surpris de ce qu'une régulation différente de gènes pratiquement semblables en nombre et très voisins en séquence puisse aboutir à des phénotypes extrêmement différents. Quant à l'aspect quantitatif lui-même, tout dépend à l'évidence de ce que l'on entend par « gènes ». Dans le sens des « facteurs héréditaires » de Gregor Mendel, un gène doit correspondre à tout élément transmissible du génome susceptible d'influer le phénotype. À ce titre, faut-il parler d'un ou de plusieurs gènes lorsqu'une même unité de transcription peut, par des phénomènes d'épissages alternatifs multiples ou d'utilisation de promoteurs optionnels, engendrer des transcrits et des protéines de structure et de fonction fort différentes ? Et comment considérer tous les éléments du génome qui sont transcrits, relativement conservés au cours de l'évolution mais non codants ? Si seul 1,6 % du génome des mammifères semblent avoir le potentiel de coder pour des protéines, les régions transcrites et conservées pourraient être dix fois plus étendues. Parmi elles, nul doute que figurent, en particulier, des ARN interférants, des micro-ARN ou d'autres espèces encore dont on sait qu'elles peuvent finement moduler l'expression, voire la structure du génome à différents niveaux : réarrangement génique, méthylation, trans-

cription, stabilité des transcrits, traduction... Or, chacune des unités transcriptionnelles codant pour de telles espèces d'ARN régulateurs peut avoir un effet phénotypique caractérisé et, par conséquent, mérite alors d'être appelé « gène » au sens initial du terme. Ainsi, suivant la convention à laquelle on se réfère, le génome humain, comme d'ailleurs celui des autres espèces, doit avoir de 25 000 à plusieurs centaines de milliers de « gènes ».

Ainsi apparaît-il que le séquençage du génome humain dont rend compte l'ouvrage de Bertrand Jordan n'est qu'une étape, certes importante, des progrès de la biologie. Pour reprendre la métaphore linguistique précédemment convoquée, l'on pourrait assimiler le travail ainsi réalisé à une vaste entreprise lexicographique : il va de soi que la possession d'un bon dictionnaire, constitué ici des gènes « annotés » est un outil essentiel de toute étude linguistique et littéraire, c'est-à-dire, ici, biologique. Au-delà de cette réalité, fleurissent des déclarations manifestant une incompréhension profonde de ce qu'est le génome ou reflétant des desseins divers, d'essence idéologique, commerciale ou politique. La lecture de l'ouvrage de Bertrand Jordan donnera à ses lecteurs les clés nécessaires pour, dans l'avenir, différencier, dans le discours autour des relations entre l'homme et ses gènes, ce qui revient probablement à la prudence scientifique ou bien ressort d'un discours d'une toute autre nature.

Axel Kahn

1. ENTRÉE EN MATIÈRE

AVANT-PROPOS

J'ai commencé à écrire pour médecine/sciences *en 1989, sur l'amicale sollicitation d'Axel Kahn, alors Rédacteur en chef. Il s'agissait d'un éditorial sur la « Génétique inverse », terme employé à l'époque pour la démarche qui allait, en quelques années, identifier plus d'un millier de gènes impliqués dans différentes maladies génétiques humaines. Cette collaboration est devenue régulière et intensive lorsque, en 1991, j'ai consacré une année entière à un « Tour du monde du Génome ». Il s'agissait en fait d'une enquête menée dans près d'une centaine de laboratoires, des États-Unis à l'Europe en passant par le Japon ou l'Australie, afin de voir où en était réellement ce « Programme Génome » alors à ses débuts. Les « Chroniques génomiques » rédigées au fil des visites et aussitôt publiées dans* médecine/sciences *ont donné naissance à deux livres relatant ce voyage, l'un du point de vue scientifique et organisationnel*[1]*, l'autre d'une manière plus personnelle, moins axée sur la science mais insistant sur la découverte des pays, des cultures et des individus rencontrés au cours de ce périple*[2]*. Le présent ouvrage se rapporte à une période ultérieure : il reproduit les chroniques parues dans* médecine/sciences *de 1992 à 2002. Regroupées par thèmes, elles constituent un aperçu – personnel, partiel et sans doute parfois partial – du déroulement de ce Programme Génome qui, en une décennie, est passé des premières cartes grossières et truffées d'erreurs à la séquence quasiment complète des trois milliards de lettres qui constituent notre patrimoine héréditaire. Elles illustrent les aléas de cette aventure, les illusions et parfois les déconvenues qui l'ont marquée, montrent la formidable évolution de certaines techniques et la relative stagnation d'autres. Elles font aussi une large part aux problèmes d'organisation inévitablement rencontrés au cours de cette entreprise très nouvelle pour la biologie par son ampleur et son caractère systématique. Ces textes sont reproduits tels qu'ils ont paru*[3]*, accompagnés de quelques commentaires qui les resituent dans leur contexte ou examinent si les pronostics formulés se sont ou non avérés corrects.*

J'espère que les lecteurs qui ont apprécié ces chroniques lors de leur parution les retrouveront avec plaisir, et que d'autres découvriront à travers elles les péripéties parfois oubliées de ce projet sans précédent, dont l'achèvement nous ouvre aujourd'hui des perspectives aussi passionnantes qu'à certains égards inquiétantes...

1. *Voyage autour du Génome, le tour du monde en 80 labos.* Paris : John Libbey Eurotext, 1993.
2. *Voyage au pays des gènes.* Paris : Les Belles Lettres, 1996.
3. Nous avons même conservé les conventions typographiques de l'époque de parution, y compris la référence à la revue, passée de *médecine/sciences* à *Med Sci*...

Med Sci *2000 ; 16 : 22-25*

ANNÉES GÉNOME : LES DIX GLORIEUSES

Ce premier texte n'est pas une « vraie » chronique, mais un éditorial écrit pour le numéro de médecine/sciences *de janvier 2000 consacré à « La révolution du génome ». Rédigé peu avant l'obtention du premier « brouillon » de la séquence de l'ADN humain, il situe globalement les différents éléments du Programme génome et constitue à ce titre, me semble-t-il, une bonne introduction à ce recueil.*

1988 : les Programmes Génome se mettent en place, après plusieurs années de débats lancés par le colloque de Santa Cruz 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 du *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, ils est instructif de la relire et de consulter le livre des *abstracts* correspondant. Plusieurs éléments sont frappants. Le séquençage n'occupe qu'une demie-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 déchiffrage 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 entraînèrent 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 encombrent notre ADN. À 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 20 % 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 plus tard par le précurseur Craig Venter à contre-courant de la doctrine officielle, révélée au grand jour en 1961 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 polymor-*

phisms), n'était pas encore manifeste. Il faudrait attendre l'aventure du Généthon, l'investissement massif de l'AFM 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, édifiée 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 constituaient l'objet principal du débat. 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 possiblité 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 chromosme 16 ou 19 (tâche à laquelle s'étaient attelés les laboratoires de Robert Moysis 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 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 déchiffrage 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, *Arabidopsis 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 [4].

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 bioinformatiques 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 bioinformatique très diversifiée, dont l'épanouissement a été accompagné par une fantastique accélération des performances des ordinateurs et un extraordinaire développement du réseau Internet. Bioinformatique 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 sembait 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 réétablies 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 quatre-vingts ; 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 bioinformatique 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'informatique 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.

Depuis cet éditorial (paru au tout début de l'année 2000), les séquences « brouillon » puis « finies » du génome humain et de plusieurs autres génomes ont été publiées, et les surprises n'ont pas manqué. Le nombre (probable) de gènes humains a été revu à la baisse et est passé de plus de 100 000 à 30 ou 40 000 (voir Chapitre 8). *Les phénomènes d'épissage alternatif se sont avérés bien plus fréquents que prévu, et touchent sans doute la majorité des gènes. Enfin, la découverte des ARNi, ces petits ARN transcrits mais non traduits, ouvre une nouvelle fenêtre sur les processus de régulation tout en donnant un rôle à une partie de l'ADN non codant. Le panorama effectué reste néanmoins d'actualité, et j'ai regroupé les chroniques dans un ordre qui correspond approximativement au survol effectué ci-dessus. Nous allons donc commencer par les cartes...*

This page intentionally left blank

2. L'ÈRE DES CARTOGRAPHES

Les premiers temps du programme ont été dominés par la réalisation des cartes du génome humain. On oublie un peu à quel point ces études étaient imparfaites et parcellaires dans les années 1980 : la première carte génétique générale date de 1987, et ses RFLP espacés de vingt ou quarante centimorgans la rendent peu utilisable. Quant aux cartes physiques, beaucoup pensaient lors des débuts du programme Génome qu'elles seraient impossibles à établir à l'échelle de chromosomes entiers. Ce balisage apparaissait pourtant indispensable au séquençage (objectif encore lointain compte tenu des techniques alors disponibles), mais aussi, et surtout, il était directement utile à la génétique médicale pour sa recherche de gènes impliqués dans des maladies héréditaires. De 1988 (commencement réel des études à grande échelle) jusque vers la fin des années 1990, la construction de cartes génétiques et physiques a donc constitué l'essentiel du travail de nombreux laboratoires et Genome centers. *Cette aventure, dans laquelle la France a joué un rôle de premier plan grâce au CEPH, à l'AFM et à Généthon, a connu de nombreux rebondissements dont l'écho transparaît dans les chroniques publiées à l'époque.*

Med Sci *1992 ; 8 : 1100-1101*

DES CARTES EN VOIE D'INTÉGRATION ?

Cet article, rédigé à l'issue d'un séminaire sur l'intégration des données de cartographie obtenues à divers niveaux par différentes méthodes, décrivait les premières tentatives d'unification des résultats – compromises, entre autres, par l'état primitif de l'informatique et les très faibles performances d'Internet. La véritable intégration ne devait être effectuée que beaucoup plus tard, une fois la séquence de l'ADN humain connue.

Une réunion sur « la construction de cartes chromosomiques intégrées chez l'homme » s'est tenue récemment[1] à l'initiative de Jean Frézal et avec le soutien du GREG (Groupement de Recherches et d'Études sur les Génomes).

Un séminaire « transversal »

Cette réunion regroupait une trentaine de scientifiques européens appartenant au domaine de la Génétique moléculaire humaine (ou murine) ou à celui des banques de données ; certains cumulaient d'ailleurs ces deux compétences. La confrontation devait être féconde et, pour une fois, le dialogue entre biologistes et informaticiens s'est avéré fructueux. Cette rencontre avait lieu, il est vrai, à un moment où la Génétique européenne a le vent en poupe, avec le récent succès du séquençage du chromosome III de la levure et, surtout, les remarquables résultats des équipes du Généthon en cartographie génétique et physique (voir la *Chronique génomique,* p. 1102 de ce numéro[2]). Sur le front des bases de données, les systèmes européens marquent également des points avec, entre autres, la base « ACeDB » (A *Caenorhabditis elegans* Data Base), élaborée à l'origine pour le nématode, mais maintenant adoptée pour plusieurs autres organismes. Reconnaissons aussi que nous avions affaire à des informaticiens éclairés connaissant réellement la Biologie : Newton Morton (Southampton) pour la *Location Data Base*, Charles Gautier (Lyon) pour *MultiMap*, Jean Thierry-Mieg (Montpellier) pour ACeDB, Otto Ritter (Heidelberg) pour le projet européen d'IGD *(Integrated Genome Database)*, Guy Vaisseix pour le service informatique du Généthon et, bien sûr, Jean Frézal pour Génatlas.

L'intégration, une étape indispensable

Objet du séminaire : comment aller vers des cartes intégrées. La question se pose à plusieurs niveaux. Il s'agit d'abord, dans le cadre d'une cartographie en principe homogène (carte génétique, par exemple), d'intégrer les résultats obtenus par différents groupes : c'est moins aisé qu'en apparence, car les marqueurs employés ne sont pas forcément les mêmes, et les familles sur lesquelles porte l'analyse sont parfois mal définies. La qualité des résultats peut aussi varier considérablement d'une équipe à une autre... La comparaison, à cet égard, entre la récente carte génétique NIH/CEPH combi-

1. Séminaire sur les cartes intégrées, Gif-sur-Yvette (France), 23 et 24 octobre 1992.
2. Ce renvoi correspond à la Chronique suivante, p. 23 de cet ouvrage.

nant les données de très nombreux laboratoires et la carte de microsatellites présentée par le groupe de Jean Weissenbach est instructive. La longueur totale de la première, 4 910 centimorgans (au lieu des 3 300 anciennement admis, et des 3 800 à 4 000 qui paraissent maintenant les plus vraisemblables), est presque certainement due à un nombre non négligeable d'erreurs. La carte microsatellite, beaucoup plus cohérente, présente, elle, un léger déficit (3 576 centimorgans) en raison de l'absence de marqueurs télomériques, mais est vraisemblablement plus fiable et certainement plus utile en raison de la haute informativité des marqueurs sur lesquels elle repose. Ceci pour dire que l'intégration souhaitée ne doit pas se résumer à faire la moyenne entre deux mauvaises cartes et une bonne...

L'intégration est encore plus importante, et plus délicate, lorsque des résultats obtenus par des méthodes très variées : *contigs* de YAC, cartes cytogénétiques, points de translocation, hybrides d'irradiation... doivent être combinés avec les données génétiques. Il est clair, et l'exemple du chromosome 21 le démontre, qu'une carte physique complète (fondée sur une série de YAC se recouvrant) aide puissamment à structurer et à combiner les cartes partielles préexistantes ; elle ne résout pas pour autant tous les problèmes, et impose au contraire l'emploi de méthodes rationnelles pour gérer cette intégration.

Cette exigence est d'autant plus impérieuse que le flux de données s'accélère actuellement de façon précipitée. La mise en cohérence des multiples informations obtenues s'avère problématique, même dans le cadre des fameux *chromosome-specific workshops*. Sue Povey, *chromosome editor* pour le 9, devait nous faire part de ses tribulations, relayée par Claudine Junien qui joue le même difficile rôle pour le chromosome 11. La masse d'informations provenant des travaux sur les organismes modèles doit, elle aussi, être prise en compte. Les résultats du séquençage du chromosome III de la levure – en attendant la suite, déjà engagée – peuvent éclairer sur la fonction de certains gènes, à condition que les outils nécessaires pour exploiter ces résultats soient élaborés.

Une tentative intéressante

La *Location Data Base*, développée par Newton Morton et ses collaborateurs, propose une approche logique. Plus que d'une base de données (du moins dans son état actuel), il s'agit d'un ensemble d'outils informatiques gérant la combinaison de plusieurs cartes. Les questions de préséance (en cas de contradiction entre deux cartes partielles, laquelle croire ?) restent du ressort de l'opérateur ; mais le système informatique traite les divers cas de figure d'une manière claire. La présentation de la carte composée privilégie la position, le long du chromosome, indiquée en métaphases à partir d'un télomère, et regroupe toutes les informations autour de cette échelle avec une indication de leur fiabilité. Newton Morton est, on s'en serait douté, assez critique sur la présentation des renseignements dans GDB (*Genome Data Base*, la banque de données « officielle » des *Human Gene Mapping Workshops*, implantée à Baltimore). Cette attitude était, il faut le dire, partagée par beaucoup de participants, qui reprochent à GDB son manque de convivialité et l'indigence de son gestionnaire de cartes *Map Manager*.

Les progrès de l'*Integrated Data Base*

IGD, c'est un remarquable programme du groupe de Sandor Suhoi et Otto Ritter, au DKFZ (*Deutsche Krebs Forschungs Zenter*, centre de recherches sur le cancer), à Heidelberg. Soutenus par un important contrat européen, ils ont organisé un service d'accès aux banques existantes (GDB, GBASE, OMIM...) comparable à celui qui est offert par le *Resource Centre* britannique. Mais, au-delà, leur ambition est de mettre au point un dispositif permettant l'accès à ces différentes bases à travers une interface commune,

évitant à l'utilisateur le pénible apprentissage des syntaxes particulières à chacun de ces systèmes. Ce projet a bien progressé, et la solution adoptée présente deux caractéristiques tout à fait instructives. Tout d'abord, elle emploie comme interface la base de données ACeDB – en fait, les données des autres bases sont extraites de ces dernières et chargées dans ACeDB (qui va alors, bien sûr, changer d'appellation). Cela fait partie, semble-t-il, d'un mouvement général : nombreux sont ceux qui découvrent les charmes de ce système que le projet Arabidopsis emploie maintenant sous le nom d'AAtDB *(An Arabidopsis thaliana Data Base)*. De plus, une partie des drosophilistes s'y est convertie, et le groupe de David Bentley en Angleterre, commme celui du *Lawrence Berkeley Laboratory* en Californie, y font appel pour des données sur l'homme. Enfin, c'est sous ce format que les massifs résultats des équipes du Généthon, présentées à ce séminaire par Jean Weissenbach et Daniel Cohen, vont être mis à la disposition de la communauté scientifique... Beau palmarès pour cette base qui semble décidément présenter bien des avantages, d'autant qu'elle peut aussi servir de cahier de laboratoire informatisé... c'est même sa fonction première !

Par ailleurs IGD – rejoignant en cela une tendance actuellement très nette – s'éloigne du modèle classique où la base réside exclusivement sur la machine centrale interrogée à distance par des ordinateurs réduits au rôle de « terminal stupide » *(dumb terminal)*. On constate en effet que, même avec les meilleurs réseaux, les aléas de communication subsistent, sans parler de la surcharge de la machine centrale aux heures de pointe. Parallèlement, les capacités de calcul et de stockage des micro-ordinateurs, ou des stations de travail à bas prix, se sont accrues au point qu'il devient concevable d'y implanter une version locale de la base de données, la liaison avec la machine centrale servant simplement à « rafraîchir » périodiquement cette dernière. Dans le schéma de Ritter, la base est installée localement sur une station de travail, constituant le système FRED *(Front End)*, qui communique avec le système central TED *(Target End)* à travers MIM *(MIddle Manager)* traduisant les demandes de l'utilisateur dans le langage approprié pour chaque base. Ce concept permet une réponse beaucoup plus rapide, affranchie des aléas de la connexion, tout en gardant une version à jour des données « officielles ». La base Genatlas de Jean Frézal existe maintenant, elle aussi, en une version qui peut être implantée localement sur un PC un peu « musclé ». Cette modalité très commode s'oppose à la conception centralisatrice de GDB et semble susceptible de se généraliser : grâce la baisse continue des tarifs, une station de travail sous UNIX très convenable peut maintenant être obtenue pour moins de cinquante mille francs...

Un dialogue nécessaire

De telles réunions sont, à l'évidence, utiles : celle-ci a fait évoluer les conceptions de nombreux participants et a déclenché des collaborations nouvelles. Il est à souhaiter que cette réflexion débouche sur des réalisations : le moment paraît bien choisi pour une initiative européenne dans ce domaine. La création probable de l'EBI (*European Bioinformatics Institute*, proposé par l'EMBL et, semble-t-il, sur le point d'être approuvé par la CEE) pourrait être l'occasion de concrétiser ces projets et de passer à la vitesse supérieure, au moment où l'Europe apparaît comme une source très importante de données sur les génomes...

Med Sci *1992 ; 8 : 1102-1105*

GÉNÉTHON : LA RÉUSSITE D'UN PARI

Cette chronique, parue dans le même numéro de médecine/sciences *que l'article précédent, saluait les premiers résultats de Généthon. Ce laboratoire d'un type nouveau, presque totalement financé par l'AFM, avait été créé de manière assez confidentielle, et avait souvent été présenté comme une structure de service devant aider les équipes françaises à accroître leur efficacité. En réalité, il abritait des programmes de grande ampleur, et qui commençaient alors à porter leurs fruits. La carte génétique de deuxième génération (fondée exclusivement sur des microsatellites), établie par l'équipe de Jean Weissenbach, allait rester la référence durant de nombreuses années. Quant à la carte physique, projet mené par Daniel Cohen, sa finalisation, après une première étape très prometteuse et largement médiatisée, semblait toute proche.*

Un coup d'éclat

Lors du déjà traditionnel colloque de Cold Spring Harbor à la fin du printemps de cette année, et plus récemment à Nice, au congrès *Human Genome* tenu pour le première fois hors des États-Unis, plusieurs annonces spectaculaires ont mis en relief les travaux réalisés au « Généthon ». Le palmarès est éloquent puisqu'il inclut la carte physique complète du bras long du chromosome 21 ainsi qu'une carte génétique humaine homogène et serrée fondée sur l'isolement de plus de huit cents microsatellites hautement polymorphiques et leur positionnement précis. De plus, les progrès rapides d'une entreprise de cartographie physique de l'ensemble du génome humain permettent d'espérer sa couverture presque complète vers la fin de l'année en cours... sans oublier la montée en puissance du projet « Genexpress » qui a communiqué aux banques de données plus de 2000 séquences partielles d'ADNc. Il s'agit là d'avancées de grande envergure, propulsant soudain notre pays à un rang plus glorieux que celui de lointain troisième auquel il semblait installé, et soulignant l'apport de structures et de personnalités très atypiques.

Un grand scepticisme

Un grand scepticisme régnait pourtant lorsque le projet du « Généthon » commença à s'ébruiter il y a à peine trois ans. Généthon avait pour parents deux organisations d'un modèle peu courant, le CEPH (Centre d'Étude du Polymorphisme Humain) et l'AFM (Association Française contre les Myopathies). La première, dirigée par Jean Dausset et Daniel Cohen, est bien connue pour son rôle primordial dans la gestion et la diffusion d'une collection de familles permettant de rendre cohérent l'établissement de la carte génétique, puis dans l'organisation de la collecte des données et de leur synthèse. Le CEPH s'était aussi impliqué depuis plusieurs années dans des activités de « génome lourd ». Certaines d'entre elles, comme le séquençage du complexe majeur d'histocompatibilité humain, avaient tourné court – comme d'ailleurs les autres entreprises de méga-séquençage lancées à la même époque de par le monde. D'autres semblaient marquer le pas, par exemple le programme LABIMAP dont le but affiché était de mettre au point avec la société Bertin et d'autres partenaires toute une « ligne » d'automates pour la

biologie moléculaire. Le CEPH avait, en revanche, réussi à construire dès cette époque une banque YAC de bonne qualité, dont les clones avaient une taille moyenne élevée de l'ordre de 600 à 700 kilobases, nettement supérieure à celle de Saint Louis (MO, USA). Néanmoins, pris globalement, les résultats scientifiques du CEPH pouvaient apparaître modestes en regard de sa taille, des ambitions exprimées par ses responsables et du niveau élevé de son financement.

En ce qui concerne l'AFM – suffisamment connue pour que l'on n'en retrace pas l'histoire – elle avait déjà derrière elle plusieurs téléthons réussis, mobilisant chaque année des sommes très importantes, et était devenue un élément de poids dans la recherche génétique française. Son soutien, d'abord limité aux travaux directement liés aux maladies neuromusculaires, s'était bientôt étendu à l'ensemble des maladies génétiques. Mais l'aventure du Généthon devait marquer un changement d'échelle.

Ce fut une décision presque solitaire des deux acteurs principaux, Daniel Cohen pour le CEPH et Bernard Barataud pour l'AFM. Les conseils, commissions et comités d'évaluation auraient vraisemblablement reculé devant l'ampleur du pari : rien moins que la création de toutes pièces d'un puissant laboratoire d'étude du génome, aux dimensions d'un gros institut du CNRS (plus de cent personnes), parfaitement équipé et doté d'un budget total (y compris les services) de plus de 70 millions de francs par an. Aspiration sous-jacente, sinon affichée : battre le programme génome américain sur son propre terrain, celui des cartes génétiques et physiques, tout en faisant faire un bond décisif à la recherche sur les maladies génétiques.

On comprend que le projet ait surpris. Les sommes mises en jeu paraissaient énormes par rapport aux 150 kilofrancs par an, et par chercheur, des laboratoires les mieux financés : on oubliait, bien sûr, qu'au Généthon, ces montants – certes confortables – incluent les salaires et les amortissements. La confidentialité entourant ce lancement en indisposa beaucoup ; l'ambiguïté sur les fonctions réelles de Généthon – parfois présenté essentiellement comme un laboratoire de service – entretenait la confusion ; quant à la personnalité souvent abrasive des deux personnages en cause, elle n'aidait pas à croire à des objectifs que l'on pouvait, de bonne foi, considérer comme irréalistes. N'oublions pas qu'à cette époque commençait la tragi-comédie du « GIP génome », annoncé en grande pompe à l'automne 1990 et dont la structure n'est toujours pas en place deux ans plus tard, même si un conseil scientifique a attribué en 1992 des fonds dont chacun espère constater la réalité très prochainement. Dans cet imbroglio bien français, alors que les projets américain et même anglais semblaient aller de l'avant sans accroc majeur, l'ambition de faire mieux qu'eux pouvait, à juste titre, apparaître comme déraisonnable.

Les premiers éléments connus du Généthon, et notamment le fameux atelier des Mark II, renforçaient les réserves des sceptiques. Cet appareil, seul résultat visible du programme LABIMAP, était le deuxième prototype d'une machine à *Southern blots*. D'une technologie intéressante, en particulier par sa façon d'intégrer migration et électrotransfert dans la même cuve sans manipulation supplémentaire, il semblait pourtant arriver trop tard et réaliser l'automatisation d'une méthode « obsolescente ». Les microsatellites étaient en passe de remplacer les RFLP pour la carte génétique et la PCR rendait caduque la plupart des applications des *Southern blots*. De plus, la reproduction à l'identique et à vingt exemplaires de cette machine et de tous ses accessoires (alimentation électrique, système de refroidissement, ordinateur de commande) était très dispendieuse. Bref, on pouvait légitimement craindre que cet atelier ne devienne ce que l'on appelle outre-Atlantique un « éléphant blanc », un coûteux et inutile monument à une technologie dépassée. Lors de sa présentation au colloque de Cold Spring Harbor au printemps 1991, il suscita une majorité de commentaires ironiques – il est vrai que les Américains prennent plus facilement au sérieux la cuisine de notre pays que sa technologie... Enfin, sur un plan plus organisationnel, le choix pour le Généthon d'un personnel principalement

composé de techniciens au niveau de qualification modeste était contestable : les quelques chercheurs présents, souvent à temps partiel d'ailleurs, arriveraient-ils à diriger durablement cette usine ?

Très intéressé par cette tentative – qui répondait à certaines de mes interrogations sur le retard technologique de la Biologie –, j'étais pourtant surpris devant la dimension donnée au projet, et je m'interrogeais sur son bien-fondé, d'autant qu'il absorbait une part très importante des crédits consacrés à la recherche par l'AFM. Beaucoup de scientifiques français partageaient cette réserve, accentuée par la méconnaissance des buts réels de Généthon et par une certaine envie envers les collègues qui bénéficiaient de laboratoires si somptueusement équipés. Les critiques ne manquaient donc pas, et c'est dans un climat de scepticisme, pour ne pas dire d'hostilité, que les équipes d'Évry commencèrent à fonctionner à l'automne 1990.

Un premier frémissement

L'attitude du milieu commença pourtant à se modifier à l'égard du CEPH (et par extension du Généthon) lorsqu'il devint indéniable que la banque YAC était de bonne qualité – et que, conformément aux engagements de Daniel Cohen, elle était effectivement mise à la disposition de la communauté. Pour la majeure partie des chercheurs français et étrangers (même aux États-Unis), elle se révéla d'un accès plus facile que celle du groupe de Hans Lehrach, assez comparable, ou que celle de Rakesh Anand, aux *inserts* nettement plus petits. Exploitée selon les modalités maintenant classiques du criblage PCR (diffusion de *pools* analysés par le laboratoire extérieur, les étapes finales étant effectuées ensuite au CEPH), elle joua un rôle déterminant, par exemple, dans le succès du groupe de Jean-Louis Mandel sur l'X fragile.

À Généthon même, la présence de Jean Weissenbach et son investissement méthodique dans la mise en place d'une sorte d'usine à isoler, séquencer et localiser des microsatellites témoignaient du sérieux de l'entreprise. L'important atelier de séquençage ainsi installé fut mis à profit pour identifier le gène impliqué dans le syndrome de Kallmann et, plus récemment, pour d'autres maladies génétiques. Simultanément, la batterie des Mark II rendait de grands services à plusieurs équipes en leur permettant d'effectuer rapidement les dizaines ou même centaines de « blots » indispensables à la localisation génétique de « leur » maladie : cas, par exemple, du groupe d'Arnold Munnich pour l'amyotrophie spinale. La banque de cellules implantée au Généthon engrangeait – non sans se heurter à quelques résistances de la part des cliniciens et des chercheurs dont le sentiment de propriété est très fort – les prélèvements sanguins nécessaires à l'étude de diverses maladies et en « immortalisait » une proportion notable sous forme de lignées lymphoblastoïdes. Le programme « Genexpress » de Charles Auffray semblait, lui, avoir plus de mal à démarrer, handicapé – entre autres – par les hésitations et les incohérences du CNRS ; quant au projet de cartographie physique mené par Daniel Cohen, homme orchestre et porte-parole parfois tonitruant, il restait entouré d'un certain mystère tant sur sa finalité réelle que sur son état d'avancement effectif.

La mise sur orbite

C'est à partir du printemps 1992 que le succès allait devenir évident. La construction d'une banque de « méga-YAC » (non encore publiée), dont la taille moyenne dépasse la mégabase, fut très remarquée : ces clones constituaient manifestement des réactifs aptes à permettre la poursuite d'objectifs de cartographie audacieux. L'extraction réussie d'une banque spécifique du chromosome 21 [1], première mise en œuvre aboutie d'un schéma séduisant mais sur lequel d'autres s'étaient cassé les dents, fut rapidement suivie de l'obtention d'un *contig* de YAC couvrant la totalité

de la région euchromatique de ce chromosome [2]. Au colloque de Cold Spring Harbor de mai 1992, ces résultats firent l'effet d'une bombe. Certes, le groupe américain de David Page [3] présentait une carte analogue pour le chromosome Y. Mais aucun des gros projets de cartographie de chromosomes menés dans divers centres du DOE et du NIH n'en était arrivé à ce stade, surtout pas le programme spécifique sur le 21 coordonné en principe par le *Lawrence Berkeley Genome Center*, anciennement dirigé par Charles Cantor. Signe de ce choc, la très officielle gazette *Human Genome News*, publiée conjointement par le DOE et le NIH, donne un compte rendu qui décrit en détail la carte physique du chromosome Y mais omet complètement celle du 21, mentionnant simplement la banque spécifique...

Le festival n'était pas terminé puisqu'en septembre devait paraître, dans la prestigieuse revue *Cell*, le premier article du groupe de Daniel Cohen décrivant son entreprise de cartographie générale du génome [4, 5]. Là non plus, pas d'innovation méthodologique fondamentale : c'est l'application aux YAC et à l'ensemble du génome humain de la technique des *fingerprints* mise au point par Alan Coulson et John Sulston [6]. Chaque YAC de la banque humaine totale est coupé par deux ou trois enzymes afin d'obtenir un jeu de fragments ; après migration du mélange et *Southern blot*, ces derniers sont révélés par hybridation avec une sonde correspondant à une séquence répétée de type « L1 » ou « Alu ». Les « signatures » ainsi obtenues sont alors comparées deux à deux pour déterminer lesquelles comportent suffisamment de fragments communs pour suggérer que les YAC correspondants présentent un recouvrement. C'est en somme le procédé mis en œuvre dès la fin des années quatre vingt pour cartographier le chromosome 16 au *Genome Center* de Los Alamos par recouvrement de cosmides. La force du projet CEPH/Généthon a été de partir d'une banque YAC aux *inserts* de grande taille, de disposer de la batterie des Mark II pour effectuer rapidement de très nombreux *fingerprints* dans des conditions de reproductibilité excellentes, et de s'appuyer sur un service informatique capable d'assurer les calculs nécessaires, qui requièrent une puissance informatique considérable. L'opération a été organisée sur un mode industriel – pas de place, en effet, dans une telle entreprise, pour l'approximation ou l'amateurisme – et les résultats présentés sont convaincants : les *contigs* obtenus à l'été 1992 couvrent, selon des calculs aux bases raisonnables, une bonne moitié du génome en morceaux de plus de deux mégabases, et les contrôles effectués sont rassurants sur la validité des recouvrements détectés. Tout laisse à penser que la poursuite de l'assemblage va aboutir, dans un bref délai, à la couverture de la quasi-totalité du génome (90 à 95 %) par des YAC organisés en *contigs* de grande taille, dépassant la dizaine de mégabases. Ce résultat présente une importance déterminante : la carte ainsi construite donnera un cadre de référence fiable (incarné en des clones dont il faudra, d'ailleurs, organiser la diffusion) facilitant considérablement l'approche du gène de toute maladie génétique déjà « localisée » par l'analyse dans les familles. Elle permettra aussi de structurer et d'intégrer les diverses cartes partielles (génétique, cytogénétique, à base d'hybrides d'irradiation ou de gels pulsés...) obtenues dans de nombreux laboratoires et, par là même, de les rendre beaucoup plus opérationnelles.

Le palmarès ne s'arrête pas là puisque *Nature* publie fin octobre la carte génétique du groupe de Jean Weissenbach [7], fondée sur plus de huit cents microsatellites hautement polymorphes. Elle constitue une carte de deuxième génération, beaucoup plus cohérente que celle parue tout récemment dans *Science* [8] qui rassemble les données de nombreuses équipes et fait encore un large appel aux RFLP, souvent moins utiles pour les études dans des familles de malades en raison de leur polymorphisme moins accentué. Quant à « Genexpress », il a maintenant produit un nombre de séquences partielles tout à fait respectable dont deux mille ont été directement déposées dans les banques de données sans que la moindre demande de brevet aie été envisagée. C'est là sans doute la meilleure réponse qui pouvait être apportée aux tentatives de « protection » soutenues par le NIH ou, en tout cas, par sa puissante directrice Bernadette Healy.

Changement d'ambiance

Ces succès arrivent à un moment où le programme américain hésite un peu. La démission de James Watson lui a porté un coup sévère ; la controverse sur les brevets empoisonne l'atmosphère, et les successeurs pressentis (on parle beaucoup de Francis Collins) semblent peu enclins à s'engager tant que la direction du NIH reste dans les mêmes mains (les récentes élections présidentielles peuvent entraîner des changements à ces postes « sensibles »). La percée française pose aux responsables de délicats problèmes. Il est clair qu'une réorientation de certains programmes va s'imposer, et que quelques directeurs de *Genome Centers* doivent vivre des jours difficiles. La réussite du programme européen de séquençage du chromosome III de la Levure (même s'il a coûté très cher, sept ou huit Ecus la base, soit près de dix dollars), et sa poursuite sur d'autres chromosomes contribuent à cette morosité : vu de Washington ou de San Francisco, le seul projet sérieux dans ce secteur était américain, et c'est David Botstein qui devait le mettre en œuvre... La quatrième édition du colloque *Human Genome*, précédemment tenu à San Diego et qui avait lieu pour la première fois en Europe, fut le théâtre d'un véritable festival franco-européen. Il y avait fort peu d'américains parmi les cinq cents congressistes, une cinquantaine tout au plus, une fois défalqués les exposants commerciaux. Cela tient sans doute au fait que cette réunion est moins « pointue » que le colloque annuel de Cold Spring Harbor ; mais peut-être quelques « ténors » ne tenaient-ils pas trop à se trouver confrontés aux derniers résultats obtenus sur le vieux continent...

Un « coup » médiatique

La parution de l'article de *Cell* a été dans notre pays l'occasion d'un battage médiatique qui en a indisposé certains. Il était pourtant légitime qu'un succès aussi remarquable soit souligné, logique aussi que l'AFM qui, après tout, a financé la quasi-totalité de l'opération, informe le public sur l'emploi des fonds recueillis. Il y a eu, naturellement, quelques dérapages, facilités par l'emphase de certains protagonistes et surtout par la grande ignorance de bien des « journalistes scientifiques » (j'ai reçu des appels de chroniqueurs qui confondaient carte génétique et carte physique et étaient persuadés que le Professeur Cohen avait inventé les YAC et identifié tous les gènes humains...). Ces exagérations ne peuvent décemment être reprochées aux auteurs des travaux en question ; d'ailleurs des journaux comme *Le Monde* ou même *Libération* ont, eux, rendu compte très correctement de l'événement.

Les raisons d'un succès

Cette réussite démontre qu'il faut savoir, à certaines phases, changer d'échelle et de méthode de travail. La carte physique « génome entier » sera obtenue grâce à une infrastructure technique et un environnement informatique qui n'existent, à ma connaissance, dans aucun autre laboratoire de biologie au monde. Et l'organisation très rigoureuse, quasiment industrielle, qui a été mise en place pour cette entreprise lui était indispensable. La dimension, absolument nécessaire au programme « carte physique », joue un rôle moins fondamental pour les deux autres projets Généthon. La production de microsatellites aurait pu continuer selon le mode artisanal qui avait été celui de Jean Weissenbach au départ ; elle aurait peut-être été aussi efficace en termes de rapport financement/résultats. Mais elle n'aurait pas alors abouti au coup d'éclat d'une carte complète fondée exclusivement sur des marqueurs de haute qualité, incontestablement supérieure à la carte « NIH-CEPH » publiée quelques semaines auparavant [8]. Quant au séquençage de clones d'ADN mené par Charles Auffray, il aurait sans doute pu, lui aussi, être effectué à plus petite échelle sans perte flagrante d'efficacité. Mais dans le contexte

du moment, marqué par l'hégémonie apparente du groupe de Craig Venter et l'affaire des brevets, il était important de manifester la présence d'autres équipes sur le front des ADNc. Au total, la dimension, strictement indispensable pour le premier projet, a renforcé l'impact des deux autres, tout en permettant probablement des économies d'échelle non négligeables.

Ce coup d'accélérateur brutal n'était concevable que dans le cadre très particulier de l'AFM et du CEPH, cofondateurs de Généthon (qui, tout comme eux, est une association loi de 1901). Le profil psychologique de leurs responsables respectifs a été déterminant. Leur mérite est d'avoir osé prendre de très gros risques, d'avoir cru à ce projet au point d'y jouer la crédibilité de leurs organisations respectives. Une fois décidé ce pari, qu'aucun comité d'experts (par nature prudent et conservateur) n'aurait tenté, la structure totalement privée a permis l'engagement très rapide des travaux, l'achat du matériel, le recrutement du personnel... le tout dans un délai de quelques mois. Cette rapidité fait, à juste titre, l'envie des responsables œuvrant dans le cadre institutionnel de l'Inserm ou du CNRS : ceux-là, après avoir déposé presque deux ans à l'avance une demande de création d'une unité, auront vu leur demande agréée, parfois après avoir révisé leur copie et soumis une nouvelle proposition un an plus tard, puis auront longtemps attendu leur premier poste de technicien tout en cherchant à convaincre l'administration d'effectuer quelques travaux d'aménagement indispensables... Ils savent quelles déperditions d'énergie provoquent ces structures par ailleurs si cartésiennes, et apprécient l'atout qu'a constitué le mode de fonctionnement de l'AFM en l'espèce. Mais il faut bien voir que les risques encourus, dans le cas du Généthon, étaient à la hauteur des moyens engagés... alors qu'ils sont bien minimes dans le secteur institutionnel. Quoiqu'il en soit, la percée opérée va avoir de nombreuses conséquences, parmi lesquelles, on peut l'espérer, une évolution des mentalités qui n'a que trop tardé.

Références

1. Chumakov IM, Le Gall I, Billault A, *et al*. Isolation of chromosome 21-specific yeast artificial chromosomes from a total human genome library. *Nat Genet* 1992 ; 1 : 222-5.
2. Chumakov I, Rigault P, Guillou S, *et al*. A continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* 1992 ; 359 : 380-97.
3. Foote S, Vollrath D, Hilton A, Page DC. The human Y chromosome. Overlapping DNA clones spanning the euchromatic region. *Science* 1992 ; 258 : 50-66.
4. Bellanne-Chantelot C, Lacroix B, Ougen P, *et al*. Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992 ; 70 : 1059-67.
5. Cohen D. Premier lever d'une carte physique du génome humain. *médecine/sciences* 1992 ; 8 : 881-2.
6. Coulson A, Sulston J, Brenner S, *et al*. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc Natl Acad Sci USA* 1986 ; 83 : 7821-5.
7. Weissenbach F, Gyapay G, Dib C, *et al*. A second generation linkage map of the human genome based on highly informative microsatellite loci. *Nature* 1992 ; 359 : 794-801s.
8. NIH/CEPH collaborative mapping group. A comprehensive genetic linkage map of the human genome. *Science* 1992 ; 258 : 67-80.

Med Sci *1994 ; 10 : 898-902*

CARTE PHYSIQUE DU GÉNOME HUMAIN : L'ÉTAT DES LIEUX

Le « grand succès » annoncé fin 1992 pour la carte physique du génome humain allait faire long feu. Autant la carte génétique de Généthon s'avérait complète et fiable, et rendait d'immenses services à d'innombrables équipes de Génétique médicale, autant l'espoir de baliser l'ensemble du génome humain par simple analyse des recouvrements de YAC devait être rapidement abandonné. L'intégration des marqueurs de la carte génétique (donc le recours aux résultats d'une autre méthode) allait permettre de présenter une carte physique aux ambitions plus modestes... qui devait elle-même se révéler pratiquement inutilisable en raison de très nombreuses erreurs dues au chimérisme des YAC. Tous ces problèmes, abondamment discutés à l'étranger, furent pour l'essentiel passés sous silence en France. Cette chronique tentait de faire le point sans agressivité, mais en appelant (à peu près) un chat un chat. Elle allait valoir à son auteur quelques solides inimitiés...

Septembre 1992 : un article de l'équipe de Daniel Cohen (CEPH-Généthon) dans la très cotée revue *Cell* [1] décrit une technique de *fingerprint* (empreinte) de chromosomes artificiels de levure (YAC). Appliquée à une échelle jusque-là inimaginable à la banque de « Méga-YAC » du CEPH, elle permet de déterminer les positions relatives de ces clones, et donc de construire la carte physique de l'ensemble du génome humain. La méthode, mise en œuvre à grand renfort d'appareillage sophistiqué et d'informatique, devrait selon les auteurs produire très prochainement un taux de couverture de 90 %. L'impact médiatique est considérable : on pensait jusque-là que ces cartes n'étaient réalisables que chromosome par chromosome, et qu'elles ne seraient pas établies avant plusieurs années. Aux États-Unis – jusque-là très largement en tête dans ce domaine – on accuse le coup ; tel patron de laboratoire rassemble son équipe et lui annonce qu'il va falloir changer de sujet puisque les Français ont tout raflé... Dans notre pays règne un certain triomphalisme ; des journalistes annoncent déjà que la carte est terminée, les gènes trouvés et que les applications médicales vont suivre à brève échéance.

Début mai 1994, à Cold Spring Harbor, se tient l'annuel congrès *Genome Mapping and Sequencing*, de loin la meilleure réunion sur le sujet. Quelques mois plus tôt est paru dans *Nature* un court article présentant « Une carte physique de première génération du génome humain », signé par Daniel Cohen, Ilya Chumakov et Jean Weissenbach [2]. Le balisage présenté couvre entre 50 et 80 % du génome et, malgré l'extrême concision de l'exposé, il semble bien que les moyens employés n'ont plus grand chose à voir avec la méthode vantée dans Cell un an plus tôt. À Cold Spring Harbor – comme déjà à Kobé lors de *Human Gene Mapping 93*, début novembre – l'on constate que chaque *Genome Center* poursuit la construction de la carte physique de « son » chromosome, apparemment comme si de rien n'était. Difficile de s'y retrouver dans cet imbroglio ; c'est cette situation complexe que je vais essayer d'analyser dans cette Chronique, me fondant pour l'essentiel sur les publications des uns et des autres ainsi que sur les communications ou les *posters* présentés tout récemment à Cold Spring Harbor.

Rappelons pour commencer les éléments techniques de base. Une carte physique fiable et opérationnelle doit reposer, tout le monde s'accorde maintenant à le dire, sur l'alignement de clones le long du génome. Elle représente chaque chromosome par un ensemble de segments d'ADN clonés dont les positions relatives sont connues, de préférence avec précision et certitude. Le prototype à cet égard est la carte du génome d'*Escherichia coli* publiée en 1987 par Kohara *et al.* [3] : il avait déterminé la carte de restriction de 3 400 phages contenant chacun un segment d'ADN bactérien de 15 à 20 kilobases et les avait positionnés le long des 4 700 kilobases du génome de la bactérie. Carte fiable puisqu'on peut reprendre n'importe lequel de ces clones et vérifier sa carte ainsi que ses recouvrements avec ses voisins, carte utile puisque tout chercheur intéressé par l'examen détaillé d'une région du génome situé par exemple à 37 minutes (la carte génétique d'*Escherichia coli* emploie ces unités, liées au transfert d'information entre bactéries par conjugaison) peut demander le ou les clones correspondants et les étudier tout à loisir. Les 473 clones représentant le « jeu » minimum pour couvrir ce génome sont d'ailleurs maintenant distribués par la firme *Takara Biochemicals*, sous forme d'un filtre contenant l'ADN de ces clones et appelé *The* Escherichia coli *Gene Mapping Membrane*.

Quelques téméraires s'étaient lancés dès 1987-1988 dans la construction de telles cartes pour un chromosome humain. L'entreprise semblait vouée à l'échec : un chromosome moyen mesure une centaine de mégabases alors que les cosmides – vecteurs les plus performants à l'époque – véhiculent au maximum 40 kilobases. C'étaient donc, au bas mot, dix mille cosmides qu'il fallait analyser en détail, tâche apparemment surhumaine d'autant que l'on savait dès le départ que certaines régions sont « inclonables » dans ces vecteurs. La cartographie physique du génome humain n'est devenue réellement envisageable qu'avec l'invention par Burke, Carle et Olson [4] des chromosomes artificiels de levure (YAC, *yeast artificial chromosomes*) capables de cloner des segments d'ADN humain longs de plusieurs centaines de kilobases : le nombre de « pièces » à assembler pour couvrir un chromosome est alors de l'ordre du millier, et chacun sait qu'un puzzle à mille pièces est infiniment plus facile qu'un puzzle à dix mille pièces...

Les *Genome Centers* américains, créés à partir de 1989, se sont alors donné pour objectif l'établissement de cartes, chromosome par chromosome. Parmi les stratégies employées, la plus généralement admise fut celle du *STS (sequence tagged site) content mapping* dont le principe est représenté sur la *Figure 1*. En bref, il s'agit d'abord de baliser le chromosome choisi par un millier de STS, petites séquences définies chacune par un couples d'amorces PCR et uniques dans le génome. Ces STS peuvent être obtenus en prenant des sondes déjà connues pour provenir de ce chromosome, en les séquençant partiellement, en spécifiant un couple d'amorces à partir de cette séquence – et en vérifiant que ce dernier permet l'amplification de la région choisie et d'elle seule. Les STS peuvent aussi être des « microsatellites » déjà déterminés lors de la cartographie génétique du même chromosome – ce qui offre l'avantage d'intégrer *ipso facto* la carte physique et la carte génétique puisque les balises employées sont les mêmes. Les STS servent alors au criblage par PCR d'une banque de YAC : banque spécifique du chromosome étudié si elle existe, banque générale sinon, la spécificité des STS assurant en principe l'obtention d'un YAC provenant du bon chromosome. Si la position des STS le long du chromosome est connue, cela indique du même coup celle des YAC correspondants ; dans le cas contraire, il faudra ordonner les STS soit par cartographie génétique, soit à l'aide d'un « panel » d'hybrides somatiques contenant différentes régions du chromosome. Deux YAC contenant le même STS seront forcément chevauchants, et la détermination de leur position approximative par hybridation *in situ* autorisera quelques vérifications.

Ce processus est naturellement long, complexe et coûteux. Le « tarif » qu'appliquaient les organismes comme le DOE ou le NIH était de l'ordre de dix millions de dollars (soixante millions de francs) pour un chromosome : cette somme permettait de faire fonctionner durant trois ans un groupe d'une trentaine de personnes, équipé d'un outillage performant (séquenceurs, robots), et devait aboutir à la carte physique du chromosome

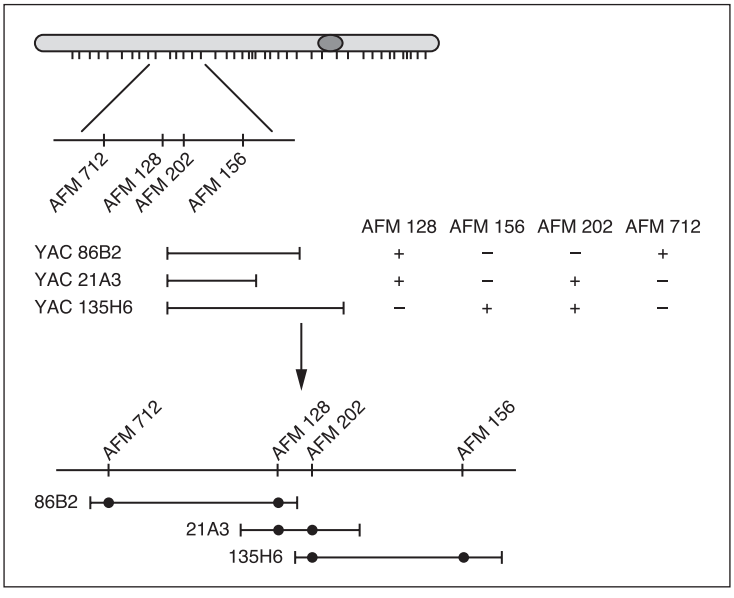

Figure 1. Construction d'une carte physique par *STS content mapping*. Les STS précédemment définis et positionnés le long du chromosome par analyse génétique (en haut) sont testés (par PCR) sur les YAC de la banque (en employant divers schémas de mélange afin de limiter le nombre de réactions à effectuer). Deux YAC qui sont positifs pour le même STS présentent *a priori* un recouvrement et l'on peut ainsi progressivement assembler des *contigs* (bas de la figure).

choisi ou, du moins, à des progrès très significatifs dans cette direction. Seuls les États-Unis semblaient s'engager dans cette course : au Royaume-Uni, l'accent était mis sur la Génétique et les ADNc, le programme Japonais sortait à peine des limbes ; quant à la France, son Programme Génome « officiel » tardait à décoller tandis que Généthon, tout nouvellement créé par l'AFM et le CEPH, paraissait centré sur la carte génétique (programme dirigé par Jean Weissenbach), les ADNc (Genexpress et Charles Auffray) et les services – l'impressionnante salle des « Mark II » avec ses vingt automates à *Southern blot* destinés aux équipes qui souhaitaient « localiser » des maladies.

En réalité, Généthon, le CEPH et Daniel Cohen préparaient une bombe. Le CEPH, au prix d'efforts prolongés et d'une remarquable persévérance, avait construit une banque de YAC contenant de très grands segments, de l'ordre de la mégabase : les fameux « MégaYacs ». Et il se livrait à l'assemblage simultané de contigs de YAC sur l'ensemble du génome humain par la technique d'empreinte mentionnée plus haut. Le principe est simple *(Figure 2)* : l'ADN d'un clone de levure contenant un YAC est coupé par une enzyme de restriction qui produit une centaine de fragments. Ces derniers sont séparés par électrophorèse, puis transférés sur un filtre de nylon, lequel est ensuite hybridé avec une séquence répétée humaine (séquence Alu par exemple) qui va révéler une ou deux dizaines de fragments. L'opération est ainsi effectuée pour chaque YAC de la banque ; la comparaison des empreintes permet de repérer ceux qui contiennent des fragments de taille identique – ce qui indique qu'ils présentent sans doute un chevauchement. Méthode largement employée, en Grande-Bretagne et aux États-Unis, pour l'assemblage de « contigs » de cosmides, elle n'avait jamais été mise en pratique avec succès pour les YAC ; elle fut appliquée à très grande échelle à Généthon. Ce n'était pas une mince affaire : il fallait répéter l'analyse sur plusieurs dizaines de milliers de YAC, dans des conditions

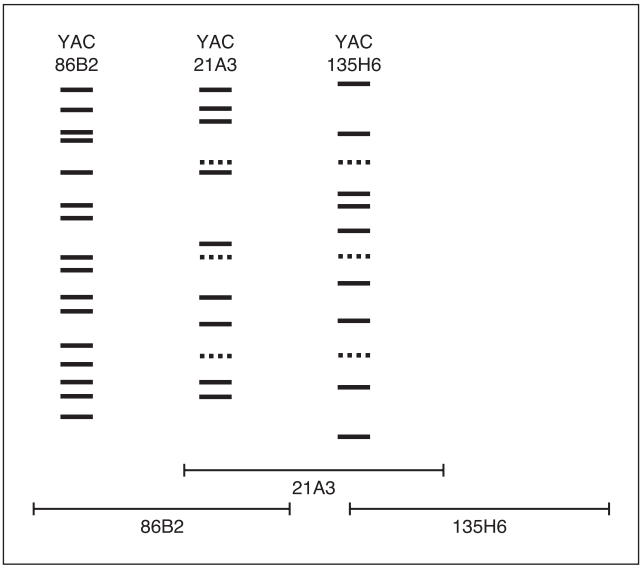

Figure 2. Analyse des recouvrements de YAC par empreinte *(fingerprint).* L'ADN de chaque clone de levure contenant un YAC est traité par une enzyme de restriction, séparé sur gel, transféré sur filtre et hybridé avec une sonde répétée humaine comme « Alu » (qui ne révèle pas les fragments d'ADN provenant des chromosomes de levure). Les positions des bandes obtenues sont comparées deux à deux ; la présence de plusieurs bandes « identiques » dans deux clones (bandes foncées entre la 1re et la 2e colonne, bandes hachurées entre la 2e et la 3e) donne une présomption de recouvrement entre les clones correspondants. Pour une discussion détaillée de cette méthode, voir [1].

autorisant une comparaison précise entre chacune de ces empreintes et chacune des cinquante mille autres. Cela posa de redoutables problèmes de standardisation des expériences, sans parler des programmes et de la puissance informatique nécessaires pour confronter cette multitude d'objets complexes. Au printemps 1992, cet assemblage semblait en très bonne voie, ce qui suscita l'article de *Cell*... et la panique des concurrents. En un ou deux ans, l'équipe de Généthon avait plus avancé que la douzaine de *Genome Centers* américains : il y avait de quoi pavoiser, et l'on ne s'en priva pas.

La carte à 90 %, annoncée pour la fin de 1992 [5] se fit pourtant attendre. Il est fréquent dans la recherche que les derniers 10 % d'un projet demandent autant d'efforts et de temps que les premiers 90 %. Ceux qui ont déterminé des séquences d'ADN ou de protéines ne me démentiront pas... Mais la lettre publiée dans *Nature* fin 1993 [2] témoigne d'un net changement de stratégie. Co-signée par Jean Weissenbach, chef d'orchestre de l'impeccable opération « carte génétique » au Généthon, elle est en fait largement fondée sur une approche du type *STS content mapping* dans laquelle les microsatellites positionnés pour la carte génétique jouent le rôle principal. Une astucieuse technique d'hybridation des produits Alu-PCR de YAC avec les autres YAC, ou avec les mêmes produits obtenus à partir d'hybrides somatiques, donne des informations complémentaires, et l'hybridation *in situ* effectuée sur quelques centaines de YAC confirme les positionnements. La méthode des *fingerprints*, qui devait à elle seule produire l'essentiel des données, est reléguée à un rang secondaire. Le résultat final est une carte dont le taux de couverture et la fiabilité sont inversement corrélés *(Figure 3)* : si l'on se limite au cas où la jonction entre deux STS est assurée par un seul YAC (niveau 1), la couverture est de 10 % ; si l'on admet entre deux STS trois YAC positionnés par Alu-PCR ou *fingerprint*

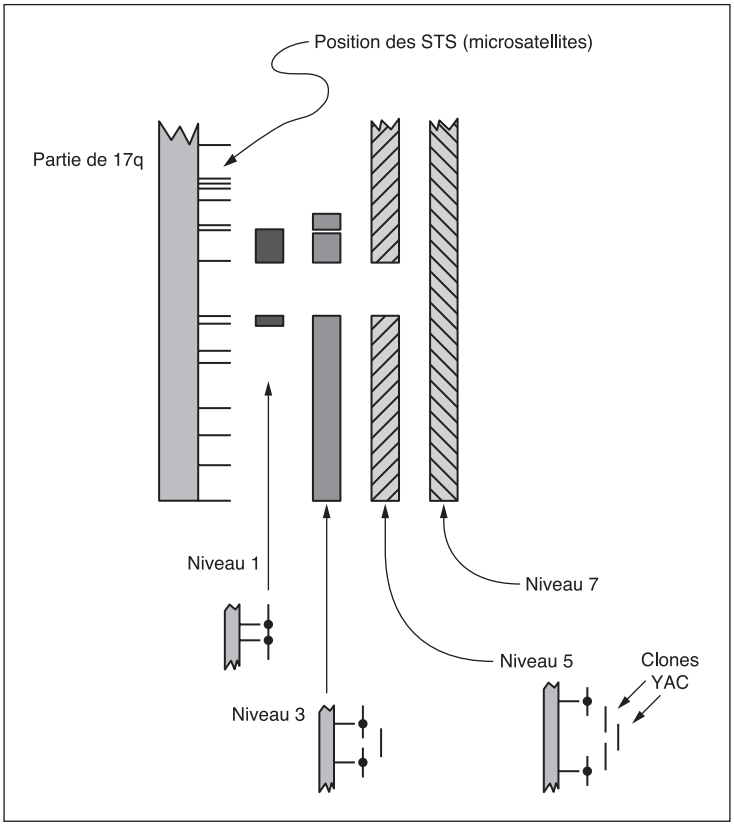

Figure 3. Zoom sur la carte physique. On voit à gauche une partie du bras long du chromosome 17, les barres horizontales repèrent la position des STS déterminée par cartographie génétique. Les quatre jeux de bandes sur la droite indiquent le taux de couverture de cette région au niveau 1 (un seul YAC joignant deux STS), 3 (trois YAC), 5 et 7. Plus le niveau est élevé, plus la couverture est complète, mais moins elle est fiable, à cause notamment des problèmes causés par le chimérisme des YAC (d'après [2]).

(niveau 3), elle monte à 30 % ; au niveau 7, elle approche 90 % – mais la probabilité d'erreurs est alors élevée.

Pourquoi tant d'incertitudes ? N'oublions pas l'échelle de l'entreprise : cinquante mille clones à étudier par diverses techniques, c'est colossal. Pensons à la facilité avec laquelle, dans nos laboratoires, l'on se trompe de clone, de marqueur de taille ou de boîte dans le congélateur... et l'on imaginera l'organisation quasi militaire qui s'impose pour manipuler mille fois plus d'objets en réduisant au minimum les risques de confusion. Mais le problème est plus profond, il tient à la nature même des YAC et aux imperfections de ce système de clonage. Les chromosomes artificiels de levure ont joué un rôle primordial, et restent encore irremplaçables (quoique BAC, PAC et autres MAC commencent à les concurrencer sérieusement[1]) ; mais ils ont le gros défaut de présenter un chimérisme

1. BAC : *bacterial artificial chromosomes* ; PAC : *pombe artificial chromosomes* ; MAC : *mammalian artificial chromosomes*. Ces différents systèmes de clonage promettent de propager de grands segments d'ADN exogène dans des bactéries, dans *Schizosaccharomyces pombe* ou même dans des cellules de mammifères – mais sont encore loin d'être pleinement opérationnels.

fréquent. Chimérisme, c'est-à-dire que le segment d'ADN inséré dans un YAC est en fait formé de deux (ou plus) fragments provenant d'autant de régions du génome. Cet inconvénient a été repéré assez vite, mais sa fréquence est longtemps restée sous-estimée. On annonçait 10, 15, peut-être 20 % de chimérisme. En réalité, les mesures les plus récentes, rapportées dans un article à paraître de Buddy Brownstein, indiquent un chimérisme moyen de 50 % – un clone sur deux contient deux segments d'origines différentes. Cela est vrai pour plusieurs banques, et même pour celle de Rakesh Anand (dite banque ICI) dont l'on a cru longtemps qu'elle était épargnée par ce fléau. Fléau, c'est bien le mot : car un clone chimérique dans un contig le fait « sauter » d'un chromosome à un autre, ou – ce qui n'est pas mieux – d'un point à un autre très éloigné sur le même chromosome. Cela d'autant plus que l'on cherche naturellement à employer les YAC les plus longs possibles pour la carte physique (moins de pièces) et que le taux de chimérisme augmente fortement avec la taille... Ainsi selon Éric Lander, lorsqu'un YAC contient deux STS, ces derniers appartiennent à deux chromosomes différents dans 60 % des cas ! On voit le problème posé aux bâtisseurs de contigs, et les risques d'erreurs qui en résultent pour les programmes d'assemblage les mieux conçus.

C'est ainsi qu'une bonne douzaine de *Genome Centers* américains poursuivent des programmes de cartographie physique. Le plus important, celui d'Éric Lander, créé fin 1992 un peu à l'image de Généthon, semble bien reprendre tout à zéro puisque – sur les YAC du CEPH – il applique une stratégie pure et dure de *STS content mapping* appuyée sur une automatisation poussée (une imposante machine permet d'effectuer 147 456 (1 536 fois 96) réactions de PCR en parallèle). Les autres centres, travaillant chacun sur un ou deux chromosomes, ont pris des options moins extrêmes. Ils emploient très largement les YAC du CEPH : la banque de méga-YAC, malgré ses imperfections (qu'elle partage avec les autres banques existantes) reste la meilleure à ce jour grâce à la taille des fragments insérés et à la couverture d'ensemble. Les laboratoires des États-Unis ont d'ailleurs fait preuve d'une grande efficacité pour la dupliquer et la distribuer dès qu'elle leur a été fournie par le CEPH. Les *Genome Centers* s'appuient également sur la carte publiée fin 1993. Peu se hasardent à lui faire confiance au-delà du niveau 3, mais elle constitue un premier « bâti » (au sens des couturières) que l'on vérifie, corrige, enrichit, et qui constitue le point de départ de cartes plus restreintes mais plus solides. Chacun s'accorde à rendre hommage au CEPH non seulement pour l'œuvre accomplie, mais aussi pour la rapidité avec laquelle les réactifs (les YAC) ainsi que toutes les informations sur la carte (contigs, chemins d'un YAC au suivant, résultats du criblage par les STS...) ont été mis à la disposition de la communauté. Paradoxalement, en raison du sous-développement informatique de la plupart des laboratoires français, ces données sont souvent plus accessibles outre-Atlantique que chez nous : leur abord (libre) implique des systèmes informatiques (de préférence une station de travail) et, en tout cas, des compétences dans l'emploi des réseaux qui sont hélas encore peu répandues en France.

Les paradoxes apparents de la situation deviennent donc compréhensibles. À l'échelle d'un génome aussi complexe que celui de l'homme, la différence entre carte génétique et carte physique tend à s'estomper : aucune des deux ne sera jamais complètement terminée, chacune demandera toujours plus de précision, de détails, et restera sujette à correction. Le placement d'un YAC devient probabiliste : il y a 90, 99 ou 99,9 chances pour cent que la position donnée soit correcte, mais nous n'aurons jamais de certitude absolue sur 50 000 YAC répartis sur trois milliards de nucléotides. Ainsi la carte physique CEPH-Généthon n'est ni « la fin de l'histoire » (ce que n'a d'ailleurs jamais prétendu Daniel Cohen) ni un « coup » raté. C'est une entreprise dont les résultats tangibles mais imparfaits servent de support aux avancées suivantes – ce qui est, au fond, la finalité de toute recherche.

Références

1. Bellanne-Chantelot C, Lacroix B, Ougen P, Billault A, Beaufils S, *et al.* Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992 ; 70 : 1059-68.
2. Cohen D, Chumakov I, Weissenbach J. A first-generation physical map of the human genome. *Nature* 1993 ; 366 : 698-701.
3. Kohara Y, Akiyama K, Isono K. The physical map of the whole *E. coli* chromosome : application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* 1987 ; 50 : 495-508.
4. Burke DT, Carle GF, Olson MV : Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 1987 ; 236 : 806-12.
5. Dorozynski A. Gene mapping the industrial way (Science in Europe section). *Science* 1992 ; 256 ; 463.

En quelques mots, l'épilogue de cette histoire : la carte physique de Généthon, en tant qu'outil opérationnel, a rapidement disparu de la scène. Elle a en revanche eu le mérite de montrer qu'il n'était pas absurde de s'attaquer à une entreprise de cette ampleur ; de plus, la banque de YAC du CEPH, malgré ses défauts (chimérisme de très nombreux clones) a été une ressource importante durant plusieurs années. La construction de la carte physique a été reprise par plusieurs Genome centers*, en particulier celui d'Éric Lander au* Whitehead Institute *(États-Unis). Elle a bénéficié de la mise au point de nouveaux vecteurs comme les BAC (*bacterial artificial chromosomes*, aux* inserts *plus petits que les YAC, mais présentant très peu de clones chimériques), et a reposé sur l'emploi systématique des STS* (sequence tagged sites) *pour repérer les recouvrements entre les clones et les positionner en s'aidant de la carte génétique. Ces efforts ont finalement abouti, vers 1997/1998, à des cartes physiques réellement utilisables, qui ont notamment servi pour le séquençage du génome humain.*

Med Sci *1994 ; 10 : 1154-1156*

FUGU STORY

Dans le contexte des programmes Génome, le Fugu est apparu, au début, comme un organisme marginal et pour tout dire un peu folklorique. Je montrais dans ce texte son intérêt principal : autoriser la comparaison de génomes proches par leur contenu génique mais de taille très différente. Une telle cartographie comparative suggère que la majeure partie de l'ADN non codant est inutile, qu'il s'agit réellement d'« ADN poubelle »... La découverte des ARNi, attribuant un rôle régulateur à certaines séquences non codantes, module un peu cette affirmation qui, à mon sens, reste néanmoins valable pour l'essentiel...

C'est en 1991, au cours d'un déjeuner à Kyoto avec Mitsui Yanagida, que j'appris le nouveau thème de recherche choisi par Sydney Brenner : un poisson, *Fugu rubripes*, jusque-là surtout connu pour la place privilégiée qu'il occupe dans la cuisine japonaise. Je pensai d'abord à une plaisanterie : Yanagida, excellent biologiste moléculaire de la Levure, est aussi un grand farceur, et nous étions justement en train de déguster de délicieuses tranches de Fugu cru... Mais l'information était bel et bien exacte.

Cette option me parut saugrenue. Certes Sydney Brenner nous a déjà joué un tour de ce genre en décidant, il y a plus de vingt ans, de se consacrer à l'étude d'un animalcule jusque-là obscur, le nématode *Caenorhabditis elegans* ; l'énorme travail accumulé depuis a porté ses fruits, et ce petit ver sera très certainement le premier organisme multicellulaire à voir son génome (100 mégabases, soit la taille d'un chromosome humain) entièrement séquencé. L'œuvre de bénédictin qu'a représenté la détermination de toutes les connexions nerveuses du nématode et l'élucidation de leur ontogenèse n'a pas encore débouché sur des découvertes renversantes, mais c'est peut-être pour dans quelques années. En tout cas le pari de *Caenorhabditis* paraissait dès le départ sensé : prendre l'eucaryote multicellulaire le plus simple possible et le soumettre aux feux croisés de l'étude génétique, de la dissection anatomique et de l'analyse moléculaire au niveau de l'ADN, c'était à l'évidence une entreprise défendable [1]. Il n'en est pas de même pour *Fugu rubripes* : ce poisson inconnu, pour lequel n'existe aucun embryon d'étude génétique (d'ailleurs quasiment inabordable pour un ensemble de raisons pratiques), a pour seul atout un génome plus compact que le nôtre. Selon une série de mesures effectuées au début des années 1960, son génome haploïde correspondrait à 0,4 picogrammes d'ADN par cellule, contre 3 picogrammes pour l'homme. Détermination ancienne, mais effectuée de manière assez précise : les globules rouges étant nucléés chez les poissons, il avait suffi de les énumérer dans un échantillon de sang puis de doser, par les bonnes vieilles méthodes colorimétriques, la quantité d'ADN contenue dans un certain volume. Des mesures complémentaires sur le sperme avaient confirmé la valeur obtenue.

À l'époque, ces chiffres n'avaient pas autrement surpris. Le Génie génétique n'était pas né, nos connaissances sur les gènes étaient fort sommaires, fondées pour l'essentiel sur quelques données obtenues chez les bactéries. La plupart d'entre nous pensaient, comme Jacques Monod, que ce qui était vrai pour *Escherichia coli* était vrai pour l'éléphant. L'on imaginait que, sauf exception, la quantité d'ADN indiquait le nombre de gènes, et personne n'était choqué à l'idée qu'un obscur petit poisson au minuscule cerveau ait huit fois moins de gènes que *Homo sapiens*. Nous n'en sommes plus là. L'intime parenté d'organismes apparemment fort différents est devenue manifeste : alors qu'à la

fin des années 1970, après cinq ans de clonage et de séquençage, l'homme et la souris apparaissaient encore comme deux mondes séparés, nous savons maintenant que pratiquement chaque gène humain a un homologue murin (et *vice versa*), et que leurs séquences codantes sont souvent les mêmes, à quelques pour cent près. Notion qui n'a d'ailleurs pas encore atteint le grand public, souvent surpris d'apprendre que la souris (sans parler des primates) est presque identique à l'homme de ce point de vue...

Tout diodon qu'il soit, le Fugu fait partie des vertébrés, et, à ce titre, nous pensons maintenant qu'il doit avoir un jeu de gènes à peu près équivalent au nôtre. Voilà donc un génome particulièrement compact, puisque – si les chiffres cités plus haut sont exacts – il comporte en tout 400 mégabases d'ADN au lieu de nos 3 000. C'était bien le motif avancé par Brenner pour le choix de ce système biologique. L'argument me semblait pourtant mince, et je soupçonne que je n'étais pas le seul. Lancer ainsi un lourd projet sur un organisme quasiment inconnu, sans autre justification que la compacité présumée de son génome, était-ce bien raisonnable ? Les données acquises sur le Fugu seraient elles vraiment transposables aux autres vertébrés, ou faudrait-il décrypter l'embryologie, le comportement et la génétique de cet organisme pour leur donner un sens ? Pour tout dire, ce programme m'apparut un peu comme la dernière lubie d'un grand esprit habitué à avoir raison contre tout le monde et voulant frapper un dernier coup avant sa retraite.

C'est à Singapour que j'eus l'occasion de voir de plus près ces travaux – et de réviser quelque peu mon jugement. Un congrès organisé par la *Asian Pacific Society of Bioscientists* y rassemblait à la mi-août 1994 quelques centaines de chercheurs provenant des quatre « petits dragons » (Singapour, Taïwan, Corée du Sud, Hong-Kong) ainsi qu'une poignée d'Australiens (mais aucun Japonais) ; ce colloque réunissait également des stars comme Robert Gallo, Michael Bishop ou Robin Weiss. Sydney Brenner était lui aussi de la partie – en tant que membre de la maison car, outre son laboratoire à Cambridge et son groupe à San Diego, ce diable d'homme dirige également une équipe de six personnes à Singapour. Dans cette « Suisse de l'Asie » au développement foudroyant (dix pour cent de croissance annuelle, et un niveau de vie maintenant supérieur à celui de la Grande-Bretagne, son ancienne métropole), la recherche biologique fondamentale était restée insignifiante jusqu'à la fondation il y a sept ans de l'*Institute of Molecular and Cellular Biology*. Somptueusement installé, mené par un directeur énergique, Chris Tan, et bénéficiant d'un budget à faire rêver un directeur des Sciences de la Vie (cent vingt millions de francs par an, salaires compris, pour deux cent trente personnes), cet Institut a réussi à attirer d'excellents chercheurs qui publient dans *Cell, EMBO J, Nature, PNAS*... et dont les travaux devaient dominer le colloque.

L'exposé que fit Sydney, et quelques discussions avec ses collaborateurs, m'ont permis de mieux mesurer les premiers résultats du travail sur le génome du Fugu et d'en apprécier les retombées possibles. Les deux équipes impliquées dans le projet (celle de Cambridge, une dizaine de personnes, et celle de Singapour ; le groupe de San Diego travaille sur d'autres thèmes) ont commencé par vérifier la taille du génome de ce poisson. La méthode a consisté tout simplement à compter les clones révélés lorsqu'on crible une banque génomique de Fugu avec des sondes déjà caractérisées provenant du même organisme et dont on sait (par une analyse en *Southern blot*) qu'elles correspondent à un gène unique. Une banale règle de trois incluant le nombre de clones examinés dans le criblage, la taille moyenne des segments insérés et le nombre de positifs trouvés indique une valeur de 384 mégabases. La précision de l'accord avec les 400 mégabases déduites des anciens dosages est sûrement fortuite, mais cela confirme bien la taille du génome de notre diodon. Pour évaluer l'effectif total des gènes, une approche astucieuse combinant séquençage partiel et exploitation des bases de données a été employée [2]. Six cents clones pris au hasard dans une banque génomique (construite à partir d'ADN fragmenté par sonication) ont subi un séquençage partiel, deux ou trois cents nucléotides. On peut alors comparer ces « étiquettes » (bien qu'il s'agisse ici d'ADN génomique et non d'ADNc) à l'ensemble des séquences géniques de mammifères contenues dans les bases de

données. Une partie d'entre elles correspond à des séquences répétées : le génome du Fugu contient des minisatellites, des microsatellites et bien sûr de multiples copies d'ADN ribosomique, mais pas d'éléments de type ALU ou L1. Pour les autres, l'on observe soit des homologies très fortes, indiquant une excellente conservation des séquences codantes, soit aucune ressemblance, signifiant que le segment de Fugu considéré n'est pas une séquence codante, ou contient un gène dont l'équivalent n'a pas encore été séquencé chez les mammifères. Finalement les 130 kb séquencées (en 600 petits morceaux) s'avèrent contenir au total environ une kilobase de séquence codante connue, soit près de un pour cent. Or l'ensemble des séquences géniques connues chez les mammifères représente trois mégabases (une fois les redondances éliminées) sur un génome de 3 000 mégabases. Si nous avions séquencé au hasard des fragments d'ADN humain selon les modalités décrites ci-dessus, nous devrions donc après comparaison retrouver dans cet échantillon un pour mille de séquences codantes connues. Chez le Fugu, la proportion de séquences géniques est dix fois plus grande, ce qui indique un nombre total de gènes équivalent puisque le génome est dix fois plus petit. Le calcul exact suggère une valeur proche de 80 % du chiffre humain, soit de 50 à 80 000 gènes.

Les hypothèses de départ sont donc validées, et le Fugu présente bel et bien un assortiment de gènes équivalent au nôtre mais codé par dix fois moins d'ADN. Les travaux menés depuis deux ou trois ans montrent comment cette compacité est atteinte. Pour les cinq ou six gènes déjà séquencés chez le Fugu, on retrouve presque toujours la même organisation que chez les mammifères : nombre d'exons et position des introns sont respectés. Mais ces derniers sont beaucoup plus petits, en général longs seulement d'une centaine de nucléotides ; de sorte qu'un gène « moyen » mesure de deux à trois kilobases, introns compris. Quant aux distances intergéniques, elles semblent osciller autour d'une valeur analogue. Du coup, le Fugu devient immédiatement utile pour l'étude de gènes humains nouvellement découverts. Ces derniers s'étendent en effet souvent sur des distances décourageantes : sans atteindre les trois millions de bases du gène de la dystrophine, nombreux sont ceux qui occupent cent ou deux cents kilobases, à raison de dizaines de petits exons perdus dans un vaste désert intronique. Le déchiffrage d'un tel ensemble est malaisé, et le raccourci du passage par l'ADNc moins utile qu'il n'y paraît : il est souvent difficile d'obtenir un clone complet. De plus les phénomènes d'épissage alternatif sont fréquents dans ces grands gènes, l'on n'est donc jamais sûr d'avoir trouvé tous les exons. Une tactique payante peut donc être d'isoler et de séquencer l'homologue Fugu du gène en question. La séquence codante ainsi obtenue sera très proche de celle de l'homme ou de la souris, et la disposition des exons généralement identique. Une fois ces dernier définis chez le poisson, il sera facile de vérifier leur présence chez l'homme et de détecter d'éventuelles différences – tout comme d'obtenir la séquence humaine grâce à un déchiffrage sélectif guidé par la connaissance de l'organisation du gène. Le gène de la Huntingtine mesure ainsi 25 kb chez le Fugu, contre plus de 300 chez l'homme ; quant à celui de la dystrophine il dépasse un peu les 100 kb – ce qui est énorme pour le diodon mais autrement plus abordable que les trois mégabases de l'homme ! Notons que jusqu'à maintenant tous les gènes cherchés dans ce système ont pu être trouvés, et que même une entité apparemment aussi spécialisée que la tétrahydrocannabinol hydroxylase (qui métabolise le produit actif du hachisch) a son gène chez le Fugu...

Un autre aspect de ce travail concerne la conservation des relations de proximité entre les gènes [3] – ce que Sydney Brenner appelle *adjacency* par répugnance à employer le terme de synténie qui tend à désigner la conservation de grandes régions. On sait qu'entre l'homme et la souris, séparés par cent quarante millions d'années dans l'évolution, les homologies entre zones chromosomiques s'étendent en général sur dix à vingt mégabases – c'est-à-dire que dans ces intervalles on retrouve les mêmes gènes, souvent dans la même disposition, comme si les chromosomes d'une espèce étaient un *patchwork* de ceux de l'autre (ou *vice-versa*). C'est environ mille cinq cents millions d'années qui nous séparent du diodon, on peut donc imaginer que les segments homologues s'étendent

sur une à deux mégabases – soit cent à deux cents kilobases en ADN de Fugu puisque ce dernier est dix fois plus compact. Cela semble effectivement être le cas, et a fait l'objet d'une communication lors du dernier *Cold Spring Harbor Genome Mapping and Sequencing Meeting*. Le nombre d'exemples étudiés est certes encore insuffisant, mais cela ouvre d'intéressantes perspectives pour la recherche de gènes impliqués dans une maladie. L'analyse génétique « localise » en général cette dernière à quelques centimorgans, quelques mégabases près ; il faut alors étudier en détail la région ainsi délimitée et faire l'inventaire des gènes qu'elle contient, pour examiner ensuite l'état fonctionnel de chacun d'eux chez les malades. Une telle zone, en l'état actuel de nos connaissances sur le génome humain, contiendra en général quelques dizaines de gènes dont seulement un ou deux connus. Leur mise en évidence demandera beaucoup d'efforts, compte tenu de la dispersion des exons dans ce grand intervalle. L'alternative maintenant offerte est d'isoler le segment correspondant chez le Fugu (une banque de YAC, avec des segments insérés allant de cent à deux cents kb, serait idéale de ce point de vue, et des efforts sont en cours pour la réaliser) : l'inventaire des séquences codantes contenues sur ce segment relativement court et maniable sera bien plus facile que chez l'homme.

L'on peut enfin spéculer – et Sydney Brenner ne s'en est pas privé, au cours d'une intervention prévue pour trente minutes et qui devait durer près d'une heure et demie – sur l'équivalence fonctionnelle entre gènes de Fugu et gènes de mammifères, et sur les façons d'en tirer parti. Imaginons en effet que l'on remplace dans une souris un gène par son homologue tiré du diodon, sous forme de clone génomique contenant les régions adjacentes 5' et 3'. Si ce gène fonctionne chez la souris (ce qui est très vraisemblable), et s'il y est correctement réglé (pari plus hasardeux mais non stupide), cela voudra dire qu'il porte toutes les séquences régulatrices en *cis* nécessaires. L'identification de ces dernières pourra alors être réalisée en comparant les séquences des régions non codantes du gène murin et de celui du poisson : on peut compter sur mille cinq cent millions d'années pour faire diverger toutes les régions non directement impliquées dans la régulation ou le codage de la protéine.

Revenons, pour terminer, sur la signification générale de cet ensemble de résultats. Il devient de plus en plus scabreux d'attribuer un rôle essentiel aux 95 % du génome humain consacrés aux introns et aux séquences intergéniques : le poisson en contient dix fois moins, ce qui ne l'empêche nullement de faire fonctionner un jeu de gènes à peu près aussi complexe que le nôtre et d'ailleurs très proche de lui. Difficile dans ces conditions d'attribuer à notre *junk DNA* une fonction cachée – à moins d'y loger la conscience ou l'âme, ce qui semble peu probable, d'autant que nous le partageons avec la vache et la souris. Non, il faut se rendre à l'évidence : notre ADN est encombré d'un magma de séquences superflues mais suffisamment peu nuisibles pour que l'évolution ne les ait pas éliminées. Cet exemple parmi beaucoup d'autres[1] nous montre que, contrairement aux apparences, la nature n'est pas parfaite. Notre plongée dans les mystères du vivant le confirme : décidément, selon l'expression profonde de Richard Dawkins, le « Grand horloger » est aveugle...

Références

1. Labouesse M. *C. elegans*, les promesses d'un petit animal intelligent : *small is beautiful*. *médecine/sciences* 1994 ; 10 : 337-41.
2. Brenner S, Elgar G, Sanford R, Macrae A, Venkatesh B, Aparicio S. Characterization of the pufferfish *(Fugu)* genome as a compact model vertebrate genome. *Nature* 1993 ; 366 : 265-8.
3. Little P. Small and perfectly formed (News and views). *Nature* 1993 ; 366 : 204-5.

1. On pourrait aussi parler des voies de transduction et de leur incroyable complexité, fruit d'un bricolage prolongé avec acharnement durant des milliards d'années...

Depuis 1994, le Fugu est resté un objet d'étude privilégié. Son génome a été séquencé aux États-Unis et en Angleterre, tandis que celui d'un de ses cousins (Tetraodon nigroviridis) *était lu dans le cadre de notre Centre National de Séquençage (CNS). La ressemblance entre les gènes humains et ceux du poisson a été confirmée, avec presque toujours le même découpage en introns et exons dans les deux espèces ; sauf exception, les gènes du Fugu sont beaucoup plus compacts que les nôtres. Il ne semble pas que les espoirs d'analyse fonctionnelle évoqués dans ma chronique aient débouché sur d'importants résultats ; en revanche, l'analyse comparative des génomes de l'homme et du Fugu (plus exactement du tétrodon étudié au CNS) a été à la base de l'estimation avancée courant 2000 par Jean Weissenbach selon laquelle le nombre de gènes humains est de l'ordre de 35 000 (voir la chronique,* Un compte d'apothicaire*, Chapitre 8). Affirmation hétérodoxe à l'époque (la valeur communément admise était de 100 000 à 150 000), mais qui tend aujourd'hui à être généralement acceptée.*

3. LE DÉTOUR PAR LES ADN COMPLÉMENTAIRES

La stratégie d'approche du génome humain par l'analyse massive et partielle de séquences d'ADN complémentaires a pris au dépourvu bien des scientifiques, moi le premier. Lors de mon tour du monde du Génome effectué en 1991, qui m'avait amené à sillonner les États-Unis durant quatre mois et à y visiter une bonne trentaine de laboratoires, j'étais passé à côté du projet que développait déjà Craig Venter. J'y ai ensuite consacré plusieurs chroniques, d'autant plus que ces travaux ont lancé la polémique sur le brevetage des séquences d'ADN et ont débouché, quelques années plus tard, sur la mise au point des puces à ADN (qui font elles aussi l'objet de plusieurs chroniques regroupées vers la fin de ce recueil, au Chapitre 7).

Med Sci *1993 ; 9 : 211-216*

LE FESTIVAL DES ADNc

Premier texte sur les EST, à un moment où l'approche était encore controversée mais où l'on fondait beaucoup d'espoirs sur la tentative française menée par Charles Auffray à Généthon.

Un vieux débat

Depuis qu'il est question de séquencer l'ADN humain, la discussion fait rage : faut-il déchiffrer l'ADN génomique « tout venant » (qui d'après les estimations généralement admises, contient seulement quelques pour cent de séquences codantes) ou s'attaquer en priorité aux seuls gènes, représentés par les clones d'ADN complémentaires préparés à partir de l'ARN messager extrait de cellules ou de tissus ? Nous avons plusieurs fois évoqué ce déjà vieux débat, et présenté dans ces colonnes, il y a quelques mois, le « deuxième souffle » du séquençage génomique [1]. L'approche par l'ADNc a, elle aussi, beaucoup de partisans et fait l'objet de travaux menés avec des moyens notables. Plusieurs publications récentes dans *Nature Genetics* – qui s'impose actuellement comme la meilleure revue semi-spécialisée du domaine – nous donnent l'occasion de faire le point. Notre tour d'horizon englobera également les premiers articles sur ce sujet parus depuis l'été 1991, ainsi que les données plus récentes glanées à l'occasion du congrès *Human Genome* à Nice, en octobre dernier, ou lors d'échanges avec les différents protagonistes de ce nouveau champ de recherches.

La construction même des banques d'ADNC – si elle est bien faite – assure que l'on va déchiffrer des séquences codantes et non des introns ou des séquences intergéniques. Aussi intéressants que soient ces derniers pour les partisans des isochores, voire les passionnés des fractales et de la « génétique numérique », leur signification biologique est obscure. Avantage, donc, aux stratégies fondées sur le séquençage d'ADNc. Mais le caractère à la fois partiel et redondant de la représentation des gènes dans toute banque d'ADNc est immédiatement venu tempérer l'enthousiasme des chercheurs. On sait en effet que chaque cellule n'exprime qu'une partie des cinquante ou cent mille gènes que renferme notre patrimoine génétique : la banque d'ADNc correspondante est donc forcément incomplète. De plus, les travaux de cinétique de réassociation de l'équipe de John Bishop, encore cités aujourd'hui bien que remontant à 1974 [2], ont clairement mis en évidence trois classes d'abondance parmi les ARN messagers d'une cellule. Les messagers très abondants comprennent un petit nombre d'espèces moléculaires dont chacune existe à des milliers ou des dizaines de milliers d'exemplaires par cellule ; ils constituent 10 à 20 % de la masse totale de l'ARN messager. Viennent ensuite les ARN moyennement abondants, quelques centaines d'espèces différentes présentes, chacune, à plusieurs centaines d'exemplaires par cellule et représentant 30 à 40 % du total. Le reste correspond aux messagers rares, qui regroupent une ou même, plusieurs dizaines de milliers d'espèces ; on en trouve une ou deux molécules par cellule, parfois même moins. Ces chiffres, obtenus par des méthodes qui paraissent aujourd'hui bien grossières, ont été pour l'essentiel confirmés par les études plus fines menées depuis le début de l'ère du clonage et restent la base de tout raisonnement.

Une approche systématique par les ADNc, visant à dresser par ce biais un inventaire des gènes humains, suppose l'analyse de très nombreux clones – pris au hasard dans une banque construite avec soin. L'optique est différente de celle du scientifique intéressé par une protéine particulière et qui en cherche le gène : il s'agit ici, au contraire, d'en examiner le plus possible. L'écueil qui menace, dès lors, le séquenceur d'ADNc est celui de retomber fréquemment sur les mêmes clones, ceux qui proviennent des messagers abondants et donc très représentés dans la banque. Son inventaire va, du coup, se limiter aux gènes les plus exprimés dans le tissu duquel il est parti – gènes qui sont souvent déjà connus : du fait même de leur forte expression, ils sont plus facilement repérables et clonables que ceux dont le messager est rare. C'est ainsi que le groupe de Peter Schimmel au MIT, qui s'était lancé dans cette aventure au début des années 1980, affichait un bilan assez décevant [3]. Ces chercheurs avaient établi une banque de muscle squelettique de lapin et séquencé plus de 150 clones tirés de cette collection : ils y trouvèrent, certes, les clones correspondant à 13 des 19 protéines musculaires connues, mais ne découvrirent aucun gène jusque-là inconnu... Cela a sans doute dissuadé beaucoup d'équipes de s'engager dans cette voie – avant que les premiers résultats du groupe de Venter ne remettent à la mode ce genre d'approche.

La fin des années 1980 : programmes nationaux, progrès de la « normalisation »

Le Programme Génome aux États-Unis, officiellement inauguré fin 1990, avait, en réalité, déjà débuté sous les auspices du DOE, du NIH et du HHMI, dès 1987-1988. Il ne comportait pas, avant 1991, de volet « ADNc », centré qu'il était sur les cartes génétiques et physiques, sur l'amélioration des technologies et sur le lancement de quelques projets de séquençage génomique : à l'époque, Craig Venter proposait le séquençage de la bande q28 du chromosome X. Les programmes nationaux annoncés ou mis en route par plusieurs « petits » pays – Grande-Bretagne, Japon ou France – faisaient, eux, une part importante à ce type de recherche. Cela traduisait sans doute le désir d'occuper un « créneau » peu investi aux États-Unis, et l'impression que ce travail pouvait être effectué sans mettre en œuvre de très gros moyens. Des méthodes nouvellement élaborées semblaient aussi rendre ce travail plus fructueux. Nous faisons ici allusion à l'égalisation ou « normalisation » des banques d'ADNc – procédé dont l'objectif est d'arriver à ce que les différentes espèces moléculaires soient présentes à des fréquences comparables dans une banque donnée.

Le principe de ces méthodes est assez simple. On part d'une banque d'ADNc non normalisée, dont on prépare en masse les segments insérés. Ce mélange est dénaturé, puis placé dans des conditions où la réassociation des brins complémentaires peut avoir lieu. Bien entendu, les espèces moléculaires abondantes, représentées à de très nombreux exemplaires dans le mélange de départ, se réassocient rapidement puisque la probabilité de rencontre des deux brins est élevée en raison de leur forte concentration. Les espèces rares, elles, se réassocient beaucoup plus lentement. De sorte que, après un certain temps de réassociation, les espèces abondantes sont majoritairement converties en ADN bicaténaire, au contraire des espèces rares. Un passage sur colonne d'hydroxyapatite (qui sépare ADN simple brin et double brin) donne alors un mélange de fragments d'ADN monocaténaires enrichi en séquences rares – ou, plus précisément, dans lequel l'abondance des séquences fréquentes a été diminuée.

La mise en œuvre de ce procédé est laborieuse. Les cycles de réassociations éliminent – c'est leur but – la majeure partie (en masse) de l'ADNc, amenant le chercheur à travailler par la suite avec des quantités d'ADN infinitésimales dont le maniement est délicat. C'est, en fait, l'inclusion d'une ou plusieurs étapes d'amplification par PCR dans les protocoles qui a permis d'aboutir à la construction de banques normalisées repré-

sentatives – publiées à peu près simultanément par une équipe japonaise [4] et par le laboratoire de Sherman Weissman à Yale (CO, USA) [5]. L'existence de ces banques – et la perspective de pouvoir appliquer les protocoles de « normalisation » ainsi définis à toute banque particulière – ont certainement rendu l'option ADNc plus attractive pour les décideurs scientifiques. Mais, curieusement, comme nous allons le voir, les travaux réalisés jusqu'ici n'ont guère fait appel à la normalisation. L'emploi de banques classiques, dans lesquelles l'abondance des clones reflète – plus ou moins – celle des molécules de messager dans la cellule de départ, s'avère très possible ; il est même préférable dans certains cas...

Les EST (ou « signatures », ou encore, « étiquettes ») entrent en scène

C'est des États-Unis qu'allait venir le premier résultat marquant. Ce fut une surprise : les travaux du groupe de Craig Venter, bien que conduits dans le cadre du NIH à Bethesda, n'étaient pas financés par le programme génome ; Venter lui-même n'était guère en odeur de sainteté auprès de Jim Watson, son directeur prestigieux – mais un peu caractériel... Et la controverse déclenchée par la demande de brevets aussitôt déposée par Venter et Reid Adler devait attirer l'attention sur ces études, bien au-delà des cercles scientifiques directement concernés. Nous n'évoquerons pas aujourd'hui cette controverse, déjà abondamment commentée.

Le premier article du groupe de Venter, paru en juin 1991, montrait l'efficacité de la méthode choisie [6]. Il s'agissait tout simplement de prendre, au hasard, des clones dans une banque d'ADNc (provenant du cerveau humain en l'occurrence), et de faire, sur chacun de ces clones, « un coup » de séquence : une seule lecture sur un séquenceur de type *Applied Biosystems*, donnant deux cents à trois cents nucléotides de séquence fiable à 97 ou 99 %. Cette information apparemment fragmentaire se révèle extrêmement utile. Comparée aux séquences enregistrées dans les bases de données, elle permet d'établir si l'on est en présence d'un gène déjà connu, ou d'une nouvelle entité. Elle autorise aussi la définition d'amorces PCR pour l'isolement du gène correspondant sous forme de cosmide ou de YAC, ainsi que la détermination du chromosome d'origine grâce à un « panel » d'hybrides monochromosomiques. L'on a ainsi obtenu un STS *(sequence tagged site)* repérant une séquence exprimée : c'est ainsi que Venter a baptisé ces entités « EST » pour *expressed sequence tags*. Enfin – du moins dans certains cas – la traduction en acides aminés du court segment d'ADN lu peut donner quelques idées sur la protéine. Ces données sont, de plus, obtenues à un coût très raisonnable : les divers laboratoires pratiquant en grand cette méthode avancent un prix de revient de l'ordre de dix dollars par étiquette. Rappelons que le séquençage complet d'un ADNc de deux mille bases, au tarif optimiste d'un dollar la base, coûte deux cent fois plus...

C'est ainsi qu'en quelques mois de travail, sans faire appel à des moyens techniques lourds, l'équipe avait séquencé plus de six cents clones pris au hasard dans une banque d'ADNc de cerveau. Trois cent trente sept de ces derniers étaient des entités jusque-là inconnues. Ce n'était pourtant que le galop d'essai puisque, un peu plus de six mois plus tard, la même équipe présentait deux mille trois cent soixante-quinze nouvelles séquences obtenues selon le même schéma opératoire [7]. On pouvait dès lors prévoir que cette connaissance – certes schématique et partielle – allait, à bref délai, pouvoir s'étendre à une fraction significative de nos cinquante ou cent mille gènes. D'autant que Craig Venter n'était pas seul à suivre cette voie...

Les « étiqueteurs » américains...

L'acteur principal, au Nouveau Monde, est naturellement Craig Venter. Comme il eut l'occasion de l'expliquer lors du *Human Genome Meeting* de Nice en octobre 1992, il est venu à l'ADNc par déception devant le peu de rendement (pour l'ADN humain) du séquençage génomique : ses articles récemment parus sur le sujet [8] soulignent en effet les difficultés de l'identification des exons, et ne montrent, après 100 kilobases de séquence, que cinq gènes dont deux déjà connus... Au contraire, un travail équivalent fournit des centaines, sinon des milliers, d'*expressed sequence tags* (EST). Sur le plan technique, il a choisi, dans un premier temps, l'emploi de banques d'ADNc commerciales : elles se sont révélées de qualité médiocre, ce qui a entraîné un changement de tactique. Qu'elles soient commerciales ou « maison », les banques proviennent d'un amorçage au hasard sur l'ARN messager, c'est-à-dire que l'endroit séquencé est placé de façon aléatoire dans la séquence – le plus souvent codante en raison de la prépondérance de cette dernière dans l'ARN messager *(m/s n° 9, vol. 8, p. 966) (Figure 1)*. Ce pari augmente la probabilité d'obtenir une information sur la structure de la protéine correspondante ; en revanche, elle peut aboutir à séquencer plusieurs morceaux du même ADNc, clonés dans différents plasmides – sans que l'on ne s'en aperçoive tant que ces séquences partielles ne se recouvrent pas. Dans une étude effectuée avec le groupe de Michael Polymeropoulos [9], une petite cinquantaine de ces EST ont été localisés sur le génome humain. Localisés est d'ailleurs un terme un peu fort, puisque, en fait, il ne s'agit que d'une « assignation » chromosomique, réalisée par réaction PCR (amorces définies à partir des séquences déjà déterminées) sur un jeu de cellules hybrides homme-souris, ou homme-hamster, contenant chacune un seul chromosome humain. L'information ainsi obtenue est par trop imprécise pour être utile en elle-même, mais elle permet d'entamer ensuite la localisation fine de chaque clone sur un *panel* spécifique du chromosome auquel il appartient.

Une autre équipe d'outre-Atlantique, celle de James Sikela [10], a publié des travaux similaires menés d'une façon un peu différente : ces auteurs ont choisi d'obtenir la séquence de l'extrémité 3' de l'ARN messager, – en amorçant la synthèse de l'ADNc par de l'oligo dT et en le clonant dans un vecteur directionnel *(Figure 1)*. Dans ce cas, les séquences portent presque toujours sur la région 3' non codante de l'ARN messager, et sont directement comparables, tant à l'intérieur du laboratoire qu'avec les autres groupes ayant pris la même option. De plus, la grande divergence de ces séquences non codantes entre l'homme et la souris, et leur fréquent polymorphisme, facilitent leur localisation ultérieure précise sur le génome. L'équipe a ainsi séquencé un peu plus de mille clones, dont plus de neuf cents semblent correspondre à des gènes jusque-là inconnus : il faut dire qu'il avait été procédé à une étape de « précriblage » pour éliminer les clones les plus abondants, parmi lesquels se retrouve la plus forte proportion de gènes déjà connus. Comme pour l'équipe de Venter, la localisation chromosomique est une étape limitante : une vingtaine seulement des clones séquencés a été localisée...

... l'approche japonaise...

Le programme Génome du ministère de l'Éducation Japonais (le « Monbusho ») incluait dès 1990 une composante ADNc notable. Je la vis en œuvre dans le laboratoire de Kenichi Matsubara au printemps 1991. C'était un petit projet, mené à l'époque par deux ou trois jeunes chercheurs et dont la conception m'avait fort intéressé ; les premiers résultats ont paru récemment dans *Nature Genetics* [11].

L'originalité de cette équipe est d'employer le séquençage, non seulement pour découvrir de nouveaux gènes, mais aussi pour évaluer le spectre d'expression d'un tissu donné. À cet effet, la banque est établie de façon à refléter le plus fidèlement possible la

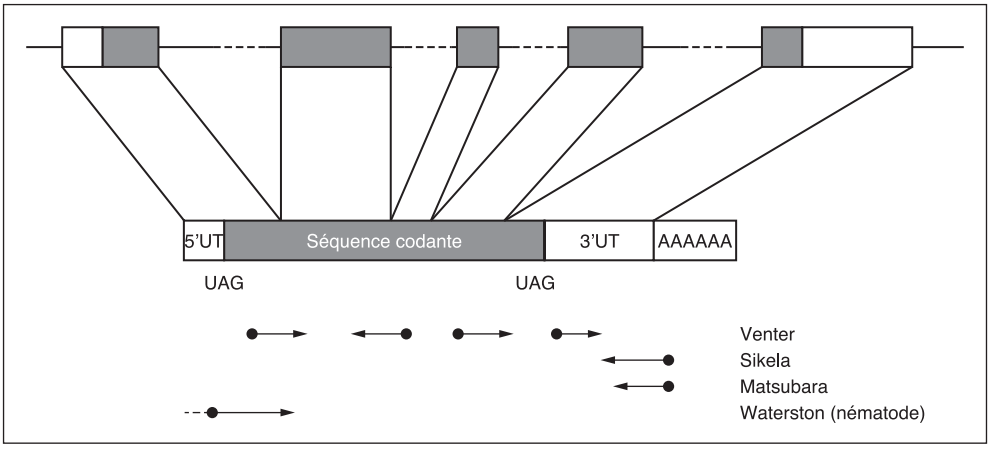

Figure 1. Les options de séquençage d'ADNc. On a figuré un gène (en haut), l'ARN messager qui en dérive et la position des séquences effectuées par différentes équipes, celles de Craig Venter [6, 7], de James Sikela [10], Kenichi Matsubara [11] et Robert Waterson [14].

composition de l'ARN messager du tissu – en clonant volontairement des segments assez courts (deux à trois cents nucléotides) à partir de l'extrémité poly A du messager. Les clones sont alors pris au hasard et séquencés selon les modalités maintenant classiques du *cycle sequencing* suivi d'un passage sur séquenceur *Applied*. L'équipe répertorie alors les séquences « redondantes » (trouvées plus d'une fois) et celles qui sont « solitaires », les compare aux banques de données, et effectue quelques contrôles pour vérifier que les valeurs de fréquence trouvées lors de l'analyse des clones reflètent bien la réalité au niveau de l'ARN messager...

Sur 982 clones ainsi séquencés, 468 appartiennent à cette catégorie des « solitaires » ; c'est là que l'on trouve les quatre cinquièmes des séquences nouvelles. Les clones redondants, eux, constituent la moitié de l'échantillon mais ne renferment que 173 espèces distinctes, dont trois très fréquentes. La majorité de ces espèces est, elle aussi, nouvelle (135 sur 173). On voit que si l'on cherche des gènes encore inconnus, il vaut mieux sélectionner des séquences peu exprimées, peu représentées dans les banques : c'était d'ailleurs assez évident *a priori*, mais les chiffres précisent ce point. Finalement, l'on retrouve les trois classes de fréquences décrites jadis par l'équipe de Bishop : trois espèces constituent à elles seules 10 % du total des clones (donc, toutes choses égales par ailleurs, 10 % de la masse de l'ARN messager cellulaire), elles représentent, dans cette cellule, la classe des messagers « très abondants ». Les 170 autres séquences redondantes forment la classe « moyennement abondante », 40 % du total, les solitaires correspondent, quant à eux, aux messagers rares. Bien entendu, la frontière entre « moyennement abondant » et « rare » est floue et arbitraire : si l'équipe avait séquencé 2 000 ou 5 000 clones au lieu de 1 000, certains des solitaires seraient devenus redondants ! Les auteurs explorent d'ailleurs cette question en prenant quelques clones solitaires et en les utilisant comme sonde sur 8 800 ou 26 400 clones de la même banque. Dans cette expérience, certains des clones testés ne sont pas retrouvés : ils sont réellement rares. D'autres sont rencontrés de une à cinq fois, ce qui permet d'estimer leur abondance.

L'examen des clones reliés à des gènes déjà connus, une petite centaine d'espèces représentée par 214 clones, permet d'aller plus loin : dans la mesure où la fonction est, dans ce cas, connue ou soupçonnée, ces données définissent une sorte de carte trans-

criptionnelle *(Figure 2)* de la cellule étudiée, qui est une lignée hépatocytaire. On voit ainsi apparaître les gènes impliqués dans la synthèse des protéines, une douzaine d'espèces constituant un quart des clones, les gènes liés à diverses autres fonctions dans le cytoplasme et les organites (une quarantaine d'espèces, un quart des clones également), et les protéines impliquées dans la sécrétion : un troisième quart des clones, contenant une quinzaine d'espèces distinctes. Le quatrième quart, quant à lui, se répartit entre protéines nucléaires, de membrane, du cytosquelette... Ce spectre, bien que partiel puisque la plupart des clones ne sont pas identifiés, donc pas rattachés à une fonction, donne néanmoins une bonne idée de ce qui se passe dans la cellule étudiée. Il n'est pas très surprenant qu'une cellule d'origine hépatique synthétise beaucoup de protéines secrétées, mais il est intéressant de constater que près de 30 % des messagers moyennement abondants sont des transcrits spécifiques du foie. La spécificité tissulaire est donc nette : les partisans du criblage différentiel ou des banques soustraites, souvent attaqués sous prétexte que tous les gènes seraient plus ou moins transcrits dans tous les tissus, en tireront quelque réconfort. Quant à la forte présence des facteurs de synthèse protéique (entre autres, le facteur d'élongation 1 alpha, qui, avec 22 clones, remporte la palme de la séquence la plus représentée), elle est liée au fait que la cellule de départ est une lignée établie, et n'a pas été retrouvée dans une banque construite directement à partir de foie lors de travaux plus récents de la même équipe.

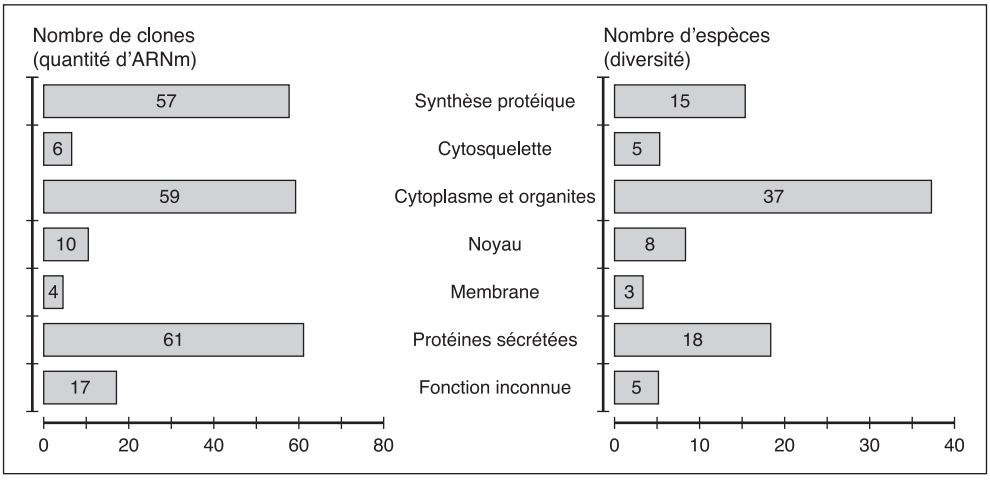

Figure 2. Spectre d'expression d'une lignée hépatique. Certains des résultats publiés par le groupe de Kenichi Matsubara [11] sont ici représentés sous forme graphique : à gauche, le nombre de clones (sur les 214 identifiés) appartenant aux diverses catégories, et à droite, le nombre d'espèces différentes pour chacune d'elles. Soulignons que, par la force des choses, cette statistique ne porte que sur les séquences correspondant à des gènes déjà identifiés, et que ce spectre est donc susceptible de se modifier fortement dans le futur.

L'idée directrice de cet exercice, dont les premiers résultats sont prometteurs, est de répéter la détermination pour chacun des 200 types cellulaires qui suffiraient à constituer notre corps selon l'ouvrage classique de Bruce Alberts, *Molecular Biology of the Cell*. Cela devrait fournir un ensemble de spectres de transcription – dont l'interprétation deviendra de plus en plus riche au fur et à mesure que plus de séquences seront connues et que les gènes correspondants auront été étudiés. Les 200 000 séquences ainsi déterminées donneront à la fois une moisson de gènes et des informations biologiques précises. Notons qu'à 10 dollars le « coup de séquence », le montant total du projet s'établit à

quelques millions de dollars, un chiffre raisonnable comparé au séquençage d'un seul chromosome humain qui, en l'état actuel des techniques, reviendrait à une ou plusieurs centaines de millions. Et un tel exercice sur les ADNc se prête aisément à un fonctionnement relativement décentralisé et collaboratif impliquant de nombreux laboratoires ; il peut ainsi jouer un rôle d'animation utile dans le monde Japonais de la recherche.

... et les autres

Les travaux rapportés ci-dessus ne rendent pas compte de l'ensemble des entreprises de séquençage systématique d'ADNc en cours de par le monde. Mentionnons le petit groupe des ADNc qui fonctionne dans le cadre du « Centre de ressources » du programme Génome britannique. Il a à son actif plus d'un millier de séquences nouvelles, bloquées pendant quelque temps par le dépôt de demandes de brevet effectuées à titre conservatoire par le *Medical Research Council* – à mon avis bien mal inspiré... Plus près de nous, le projet « Genexpress » de Charles Auffray (CNRS Villejuif et Généthon), a produit huit mille séquences dont plus de trois mille sont uniques et nouvelles, et ont été enregistrées dans les banques de données EMBL et GENBANK. Ce dépôt a été entouré d'une certaine solennité, et a donné lieu à une cérémonie, à l'UNESCO, destinée à frapper les esprits et à marquer l'opposition de ses protagonistes aux procédures de brevetage. D'autres programmes en cours, en Europe et ailleurs, permettent d'estimer que l'on dispose déjà de plus de vingt mille EST humains – nombre respectable, si on le rapproche du total estimé de nos gènes, cinquante à cent mille, et du nombre de gènes humains clonés à la fin des années 1980, moins de deux mille. Encore que ces divers chiffres ne soient pas tout à fait comparables...

Combien d'EST, combien de gènes ?

Il est instructif d'examiner le résultat des « croisements » auxquels se sont livrés certaines de ces équipes. Le groupe de Matsubara, par exemple, a comparé ses cinq cents « nouvelles » séquences aux deux mille six cents de Venter : vingt-trois seulement sont communes. Faut-il en déduire que les deux sous-ensembles (les banques d'ADNc employées) se recouvrent très peu, ou qu'elles correspondent à l'échantillonnage d'un capital de plusieurs centaines de milliers de gènes différents ? Non, car comme nous l'avons noté, les séquences de Matsubara proviennent en principe de l'extrémité 3' de l'ARN messager – alors que celles de Venter sont prises au hasard le long de la molécule *(Figure 1)*. Le recouvrement réel pourrait aller jusqu'à vingt ou trente pour cent ; la comparaison des séquences de Sikela (elles aussi en 3') avec celles du groupe japonais sera de ce point de vue révélatrice. Bref, les données n'imposent pas pour le moment une remise en cause du chiffre généralement admis de cinquante à cent mille gènes chez l'homme ; et leur accumulation indique bien qu'une fraction substantielle de cet ensemble sera prochainement « étiquetée ».

Il serait naturellement souhaitable que, pour aider aux comparaisons, les laboratoires s'accordent sur la région à séquencer. Ce n'est pas le cas actuellement ; la zone qui présenterait le maximum d'avantages est la région 5' du messager puisque, compte tenu de la faible longueur, quand elles existent, des séquences 5' non codantes, elle donnerait des informations sur la structure de la protéine *(Figure 1)*. Mais cela suppose des banques d'ADNc *full length*, dans laquelle chaque molécule de messager serait représentée par un transcrit inverse complet – ce qui est encore difficilement faisable aujourd'hui. On peut penser que les travaux d'étiquetage stimuleront le séquençage complet de clones d'ADNc, ce qui facilitera les recoupements nécessaires. En tout état de cause, l'archivage de ces séquences partielles doit être rapide afin de permettre les comparaisons entre laboratoires. Cet archivage est bel et bien réalisé pour les séquences de

Venter : le droit des brevets américains – au contraire du droit français – autorise en effet la publication des résultats avant obtention des brevets. Mais le délai entre l'obtention d'une « signature » et son archivage est encore trop long, certaines informations recueillies par les auteurs (comme les homologies avec d'autres gènes qu'ils ont décelées en interrogeant les banques de données) sont perdues dans des bases comme EMBL, et tout le système doit être amélioré pour faciliter son accès.

La plus-value de la localisation chromosomique et de la génétique

Comme nous l'avons signalé, la localisation chromosomique des « étiquettes » ne suit pas, et de loin, leur obtention. Il est en effet (relativement) aisé de charger les produits de réaction de trente clones sur un appareil *Applied Biosystems* et d'en tirer trente séquences ; en revanche, la réalisation d'un nombre équivalent de *Southern blots* ou, pire, d'hybridations *in situ* représente un travail beaucoup plus lourd. Au surplus, la sensibilité de la méthode FISH *(fluorescent* in situ *hybridization)* est encore très « limite » pour de petites sondes. Pourtant, l'information de localisation chromosomique est capitale, et à plus d'un titre. Tout d'abord, c'est elle qui transforme l'EST en un repère qui peut servir à l'intégration des cartes physiques, génétiques, et transcriptionnelles. Mais, de plus, la connaissance de la zone chromosomique d'où provient un EST, jointe à un minimum d'informations sur le tissu dans lequel il est exprimé, permet d'intéressants croisements avec les données de la Génétique clinique. N'oublions pas que l'homme est l'organisme dont la pathologie est la mieux étudiée, que trois ou quatre mille maladies génétiques sont connues, que plusieurs centaines d'entre elles sont localisées sans que pour autant le gène en cause ait été isolé. Parmi les EST nouvellement déterminés, certains vont fatalement se situer dans le segment chromosomique où « doit » – d'après l'étude génétique – se trouver le gène responsable de telle ou telle maladie. Cela ne prouve pas que l'EST corresponde à ce gène ; mais il devient à tout le moins un candidat sérieux.

À l'heure actuelle, les méthodes qui permettent de placer des ADNc sur le génome restent lourdes et lentes. L'emploi d'hybrides somatiques ne définit que le chromosome d'origine, et coûte dix fois plus cher – d'après les estimations de Craig Venter – que l'obtention de la séquence ; l'hybridation *in situ* de ces petites sondes réclame une adaptation de la technique pour presque chaque clone, contrairement à la localisation de cosmides ou de YAC qui fonctionne en routine, dans des conditions définies une fois pour toutes. L'avenir, en fait, est aux « filtres polytènes »[1], qui exploitent avec élégance les travaux de cartographie physique dans lesquels s'est récemment illustrée l'équipe de Daniel Cohen [12, 13] *(m/s n° 8, vol. 8, p. 881)*. Le concept vient du nématode, dont la carte physique est pour l'essentiel terminée. Le groupe de John Sulston a pu ainsi réaliser des filtres sur lesquels sont disposés un peu plus de 900 YAC représentant l'ensemble des cent mégabases de ce génome. Puisque la carte physique, et donc la position de chacun des YAC sont connues, ces derniers ont pu être déposés sur le filtre dans l'ordre dans lequel ils se trouvent sur la carte des six chromosomes de *C. elegans*. Ainsi, la simple hybridation d'une sonde – un ADNc par exemple – sur ce filtre va normalement révéler deux ou trois points adjacents : ce seront les deux ou trois YAC (recouvrants) qui contiennent la séquence correspondante. La position de ces points détermine le chromosome et la position de la séquence sur ce dernier, à une ou deux centaines de kilobases près. Elle indique aussi quels clones le chercheur doit demander à Sulston s'il souhaite

1. Le nom de « filtre polytène » fait référence à la drosophile. Dans cet organisme, les chromosomes polytènes des glandes salivaires ont depuis très longtemps facilité la localisation de gènes par hybridation *in situ* : ils sont en effet constitués d'environ 4 000 copies d'ADN répliqué sur place, les signaux obtenus sont donc très nets et la localisation aisée – comme avec les filtres auxquels, par analogie, John Sulston a donné leur nom.

examiner l'environnement de son ADNc ou rechercher d'autres gènes à proximité... Les projets de type EST sur le nématode font naturellement bon usage de cette possibilité. C'est ainsi que le groupe de Waterston a récemment positionné 670 ADNc : 606 n'ont posé aucun problème, les autres présentant des séquences répétées qui compliquent quelque peu l'opération [14].

Le chromosome 21 humain a été complètement cartographié, l'ensemble de notre génome est en passe de l'être : rien ne s'oppose en principe à ce que des « filtres polytènes » humains soient préparés. Le projet « Genexpress » devrait logiquement en bénéficier, et ainsi intégrer ses ADNc dans la carte physique. Cela ne sera pas immédiat, car il reste à définir la position exacte des *contigs* obtenus par l'équipe de Daniel Cohen, à établir un « jeu minimum » de YAC (inutile d'en mettre trente mille sur les filtres, cinq ou six mille devraient suffire)... et à installer la logistique de fabrication et de distribution de ces filtres qui seront à coup sûr très demandés ! Mais existe-t-il une meilleure manière de montrer l'utilité des approches génomiques lourdes, et de répondre aux critiques qui prétendent que ces travaux ne présentent pas d'intérêt biologique ?

Références

1. Jordan BR. Séquençage génomique : le deuxième souffle. *médecine/sciences* 1992 ; 8 : 854-7.
2. Bishop JO, Morton JG, Rosbach M, *et al.* Three abundance classes in HeLa cell messenger RNA. *Nature* 1974 ; 250 : 199-204.
3. Putney SD, Herlihy WC, Schimmel P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* 1983 ; 302 : 718-21.
4. Ko MSH. An « equalized cDNA library » by the reassociation of short double-stranded cDNAs. *Nucleic Acids Res* 1990 ; 18 : 5705-11.
5. Patanjali SR, Parimoo S, Weissman SM. Construction of a uniform-abundance (normalized) cDNA library. *Proc Natl Acad Sci USA* 1991 ; 88 : 1943-7.
6. Adams MD, Kelley JM, Gocayne JD, *et al.* Complementary DNA sequencing : expressed sequence tags and Human Genome project. *Science* 1991 ; 252 : 1651-6.
7. Adams MD, Dubnick M, Kerlavage A, *et al.* Sequence identification of 2,375 human brain genes. *Nature* 1992 ; 355 : 632-4.
8. Martin Gallardo A, McCombie WR, Gocayne JD, *et al.* Automated DNA sequencing and analysis of 106 kilobases from human chromsome 19q13.3. *Nat Genet* 1992 ; 1 : 34-9.
9. Polymeropoulos MH, Xiao H, Glodek A, *et al.* Chromosomal assignement of 46 brain cDNAs. *Genomics* 1992 ; 12 : 492-6.
10. Khan AS, Wilcox AS, Polymeropoulos MH, *et al.* Single pass sequencing and physical and genetic mapping of human brain cDNAS. *Nat Genet* 1992 ; 2 : 180-5.
11. Okubo K, Hori N, Matoba R, *et al.* Large-scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat Genet* 1992 ; 2 : 173-9.
12. Chumakov I, Rigault P, Guillou S, *et al.* A continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* 1992 ; 359 : 380-7.
13. Bellanne-Chantelot C, Lacroix B, Ougen P, *et al.* Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992 ; 70 : 1059-68.
14. Waterston R, Martin C, Craxton M, *et al.* A survey of expressed genes in Caenorhabditis elegans. *Nat Genet* 1992 ; 1 : 114-23.

Med Sci *1995 ; 11 : 273-276*

LA VALSE DES ÉTIQUETTES

Deux ans plus tard, l'approche par les EST (pour lesquels on essaie, en France, d'imposer le terme d'« étiquette ») a acquis droit de cité, et dbEST contient maintenant plus de deux cent mille séquences partielles d'ADNc humains. Le très joli travail de l'équipe de Bert Vogelstein (décrit ci-dessous) illustre l'apport crucial que représente la connaissance de nombreuses séquences géniques partielles, et la manière dont elle permet (en combinaison avec des données de position sur le génome) d'arriver rapidement à l'identification d'un gène important pour le cancer. La controverse sur les brevets continue, d'autant que Craig Venter est maintenant passé dans le privé, et qu'une grande entreprise pharmaceutique, Merck, adopte une stratégie tout à fait différente de publication intégrale des résultats.

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, elle-même soutenue par la puissante firme SmithKline Beecham, aurait effectué plus de 150 000 séquences partielles[1] de clones pris au hasard dans des banques d'ADNc construites à partir de divers tissus humains. Malgré les inévitables redondances (clones

1. 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 !

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 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 SmithKline 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 SmithKline 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 ; SmithKline 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 SmithKline Beecham ne contredit pas vraiment cette interprétation.

Est-ce, en fait, si scandaleux ? SmithKline 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 (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 cent 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 laissés prendre de vitesse par le privé !

Les prétentions de SmithKline 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 accumulé 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 colon. 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'une (très improbable) drogue anticancéreuse 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 de Sciences (La propriété industrielle dans le domaine du vivant, 26 et 27 janvier) devrait 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 pathologies. 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. À 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 à 10 000 F par clone[2], 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

2. 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.

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 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 70 000 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 SmithKline 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 ambigüités, 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 notre pays ne joue pas à ce stade un rôle aussi important que lors des étapes précédentes.

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 ? *Nat Genet* 1994 ; 7 : 345-6.
3. Buckholz H. *The FDA follies : an alarming look at our food and drugs in the 1980s*. New York : 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, Ownership and human genome. *Nature* 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).

L'obtention d'informations de localisation sur de très nombreux EST s'est révélée beaucoup plus difficile que je ne l'imaginais en rédigeant cette chronique. Les « filtres polytènes », si efficaces dans le cas du nématode, n'ont jamais été vraiment opérationnels pour l'homme, principalement du fait du chimérisme fréquent des YAC ; il me semble aussi que personne ne s'est sérieusement attelé à les construire, et que la volonté politique de financer de tels outils a fait défaut. Aucune des autres techniques de localisation n'a pu être transposée à grande échelle, et c'est seulement l'avènement de la séquence globale de notre génome qui a permis de placer (par simple recherche d'homologie) la plupart de ces EST. Cela dit la « carte triplement intégrée » que j'appelais de mes vœux est aujourd'hui une réalité et, de fait, l'identification du gène humain impliqué dans une affection héréditaire est maintenant effectuée en quelques mois ou même quelques semaines – pour autant que la maladie soit monogénique et que l'on ait rassemblé des familles en nombre suffisant.

Med Sci *1995 ; 11 : 1717-1719*

GÉNOME HUMAIN : L'ANNUAIRE NOUVEAU EST ARRIVÉ

Cette chronique fait écho à la publication par Craig Venter d'un grand ensemble de séquences d'ADNc accompagné d'analyses bioinformatiques, qui forme la pièce de résistance d'un numéro spécial de Nature *intitulé* The genome directory. *Elle mentionne aussi la parution d'une nouvelle mouture de la « carte physique française », très en retrait sur les perspectives annoncées en 1992 (voir* Carte physique du génome humain, l'état des lieux, *dans le Chapitre 2), et déjà un peu obsolète par rapport à la carte construite au* Whitehead Institute. *Elle évoque enfin un début d'intégration réelle de ces cartes, notamment par le positionnement d'EST sur leurs différentes versions.*

Événement scientifique, ou coup médiatique ?

La publication par la revue *Nature* d'un numéro spécial de près de quatre cents pages, le 28 septembre dernier [1], a fait l'objet d'une préparation médiatique approfondie : dossier à l'usage des journalistes développant le contenu des articles et rappelant les biographies des auteurs, conférence de presse sur invitation à Washington, et même distribution sur le réseau Internet par « ftp anonyme » des photographies de Craig Venter et de Daniel Cohen... Simple « coup de pub » visant à marquer un point dans l'intense compétition qui oppose *Nature* et *Science*, ou résultats déterminants pour l'avancée des Programmes Génome, c'est ce que nous allons tenter de démêler.

On trouve dans ce *Genome Directory* une nouvelle version de la carte physique générale, accompagnée de cartes plus détaillées portant respectivement sur les chromosomes 3, 12, 16 et 22, et enfin un article sur les ADNc. Dû au groupe de Craig Venter, ce dernier [2] occupe à lui seul 172 pages, décrit l'analyse de « 84 millions de nucléotides de séquence d'ADNc », et constitue à mon sens la nouveauté principale de ce supplément. Événement politique autant que scientifique : après des controverses acerbes qui ont défrayé la chronique depuis deux années, voici que Venter publie enfin une bonne partie de ses résultats et révèle 174 472 EST (*expressed sequence tags*, séquences partielles d'ADNc) jusque-là conservés à l'abri des regards indiscrets dans une base de données « privée »...

Les EST enfin révélés

Le dossier de presse indique que « presque 90 % des données de séquence de TIGR »[1] sont maintenant disponibles – il est en fait difficile d'évaluer exactement combien d'EST restent confidentiels. Il s'agit naturellement des plus « juteux », ceux dont la séquence risque de mener à de nouveaux inhibiteurs de coagulation ou à des cytokines

1. TIGR : *The Institute for Genome Research*, le laboratoire de Craig Venter, est très lié à *Human Genome Science*, lui-même sous contrat avec la firme SmithKline Beecham. Voir Jordan B. La valse des étiquettes. *médecine/sciences* 1995 ; 11 : 273-6.

inédites, donc à des produits commerciaux brevetables. Par ailleurs, la contribution du projet français Genexpress (qui jusqu'à fin 1994 a été le plus important producteur d'EST « publics ») n'est nulle part mentionnée dans ces commentaires... Pourtant ces séquences font partie de l'analyse présentée par Venter, de même que celles de « l'initiative Merck », vaste programme de séquençage d'ADNc conduit par Robert Waterston (Saint Louis, États-Unis) avec le soutien de cette firme. Avec, fin août, 172 388 séquences disponibles, cette entreprise ôtait beaucoup de sa valeur au « trésor de guerre » de TIGR/HGS. Son succès a sans doute joué un grand rôle dans la décision de Venter : il était temps de publier ces résultats avant qu'ils ne soient par trop dévalorisées...

Reste que l'étape est décisive : le groupe de Venter a effectué son analyse sur 174 472 EST « maison », auxquels il a ajouté 118 406 séquences supplémentaires tirées de la base spécialisée dbEST, librement accessible sur le réseau. C'est la première fois qu'une étude porte sur un ensemble aussi vaste : elle devrait permettre d'affiner l'estimation du nombre total de gènes contenus dans notre génome, tout en indiquant combien d'entre eux sont déjà étiquetés. Le calcul n'est pas simple, puisque les clones à séquencer ont été pris au hasard et que les gènes fortement exprimés sont sur-représentés dans les banques d'ADNc. Et, malheureusement, il ne suffit pas de comparer les séquences pour repérer celles qui proviennent du même gène. Le déchiffrage effectué par les différentes équipes porte tantôt sur l'extrémité 3', tantôt sur l'extrémité 5' de l'ADNc ; les séquences 5' peuvent elles-mêmes se situer en différents points selon la banque et le clone étudiés. Les erreurs de lecture compliquent encore l'analyse.

En comparant deux à deux ces presque trois cent mille petites séquences qui représentent au total 83 millions de nucléotides, Venter obtient 29 599 groupes de deux ou plusieurs EST se recouvrant totalement ou partiellement. Baptisées THC, pour *tentative human consensus sequences*, ces entités définissent en principe autant de gènes. En raison des recouvrements non détectés le nombre réel est très certainement plus faible. Les séquences « solitaires » sont, elles, au nombre de 58 384, et représentent un effectif de gènes difficile à estimer pour les raisons évoquées ci-dessus : 10 000 ?, 30 000 ? La fourchette est large. Enfin, certains gènes très peu exprimés, ou mis en œuvre seulement dans un organe spécifique à un moment précis du développement, ont pu échapper à l'étiquetage (car absents des banques utilisées) et ne sont donc pas comptabilisés. Bref, malgré leur ampleur, ces travaux ne permettent pas encore de fixer le contenu de notre patrimoine génétique. Le précédent de la levure ou du nématode, où le séquençage intégral a révélé trois ou quatre fois plus de gènes qu'attendu, incite d'ailleurs à la prudence. Quoi qu'il en soit, 10 214 seulement des séquences étudiées correspondent à des gènes connus, ce qui montre une nouvelle fois l'étendue de notre ignorance... et l'impact des informations apportées par les EST.

Venter tente aussi d'interpréter les données en termes d'expression : les EST ayant été obtenus à partir d'un grand nombre de banques d'ADNc provenant de divers tissus, leur fréquence dans chacune d'elles donne une idée du niveau d'expression du gène correspondant. C'est une démarche similaire à celle suivie au Japon par Kosaku Okubo et Kenichi Matsubara, quoique est moins rigoureuse dans la mesure où l'approche expérimentale (et les banques d'ADNc) n'ont pas été conçues dans cette optique ; les résultats ne peuvent donc être qu'indicatifs.

L'accès à ces résultats suppose la consultation de la base de données de TIGR, la *Human cDNA Database* (HCD). L'accès au niveau 1, qui concerne d'après Venter 85 % des données, est libre pour les chercheurs du secteur public sous réserve de la signature d'une décharge de responsabilité. L'accès au niveau 2, contenant les données confidentielles, reste, lui, soumis à la signature d'un accord donnant à *Human Genome Sciences* un droit de premier regard sur la valorisation des résultats obtenus. Moins satisfaisante que la disponibilité complète avec archivage dans les bases de données publiques (EMBL, GenBank), cette formule permet aux sponsors privés du projet de garder un certain

contrôle sur la diffusion des résultats (vis-à-vis de firmes concurrentes, par exemple) tout en offrant aux chercheurs des possibilités d'analyse intéressantes grâce aux logiciels développés par les informaticiens de TIGR et incorporés dans HCD.

L'article de Venter, et la mise à disposition de plus de cent mille EST nouveaux, représentent un acquis important. La cartographie de ces séquences fait l'objet d'un programme international qui devrait placer 20 000 EST sur le génome dès la mi-1996 : le clonage de gènes de maladie par l'approche des « candidats positionnels » rendra alors quasiment caduc le clonage positionnel classique. C'est ce qu'annonçait d'ailleurs, il y a près d'un an, un connaisseur en la matière nommé Francis Collins... [3].

Carte générale : affinement et consolidation

Passons maintenant aux cartes, et pour commencer à la plus globale (et la plus médiatisée), celle qui émane du CEPH (Daniel Cohen). Elle fait suite à la publication, en 1992, d'une méthode d'alignement de YAC [4] qui était censée donner très rapidement une carte couvrant 90 % du génome. C'est en réalité fin 1993 [5] qu'était publiée une « carte physique de première génération » s'appuyant largement sur les marqueurs de la carte génétique réalisée à Généthon par le groupe de Jean Weissenbach. Les « niveaux » successifs (nombre de YAC connectant deux points « sûrs ») de cette carte présentaient un taux de couverture croissant... et une fiabilité décroissante. Le présent article, lui, intitulé *A YAC contig map of the human genome* [6], indique que les « niveaux » élevés ont été abandonnées, et qu'une combinaison de techniques a été mise en œuvre pour obtenir les données les plus solides possible.

Le résultat, selon les auteurs, est un ensemble de 225 *contigs* de YAC, d'une taille moyenne de dix mégabases, couvrant 75 % du génome. Tout porte à penser que la fiabilité est effectivement au rendez-vous, que les incertitudes ont été correctement évaluées et que ces *contigs* seront une base sérieuse pour les travaux ultérieurs. Le retard enregistré par rapport à des prévisions initiales très optimistes correspond aux difficultés bien réelles rencontrées par tous les laboratoires dans la construction de cartes de grande étendue avec ces réactifs indispensables mais très imparfaits que sont les YAC. La publication de cette carte est une étape majeure dans le balisage physique de notre génome, et va encore accélérer la découverte de gènes impliqués dans des maladies en rendant immédiatement accessibles les YAC « couvrant » toute région désignée par l'analyse génétique. Elle ne constitue cependant pas un saut qualitatif comparable à la première carte génétique parue en 1987 [7] ou à la première carte physique générale publiée en 1993 [5]. De plus, une nouvelle carte physique du génome humain a été établie au *Whitehead Institute* (le *Genome Center* américain dirigé par Éric Lander) et présentée pour la première fois lors du récent « Colloque sur les séquences transcrites » tenu à l'Iles des Embiez début novembre. Construite selon une technologie analogue, utilisant largement la carte génétique de Généthon et les YAC du CEPH, elle apparaît plus fine (incorporant plus de 11 000 STS au lieu de 3 000), plus complète (couvrant environ 90 % du génome) et très fiable puisque chaque YAC est relié à son voisin par au moins deux STS. Déjà partiellement disponible sur Internet, cette nouvelle avancée relativise donc l'importance du résultat rapporté ici.

Les cartes par chromosome restent indispensables

Les quatre autres articles décrivent des cartes physiques portant chacune sur un chromosome. Ils s'appuient sur la carte générale et sur les YAC du CEPH, mais y ajoutent des marqueurs et des données provenant de multiples équipes souvent regroupées en consortiums. Ces groupes travaillent souvent depuis des années sur différentes régions d'un chromosome, y ont cherché et souvent trouvé les gènes impliqués dans les maladies

qui les intéressent, et ont rassemblé des collections d'hybrides, YAC et autres réactifs. Ils savent quelles sont les régions problématiques, celles dont les séquences répétées brouillent l'analyse, celles où de fréquentes délétions compliquent l'analyse génétique... Leur expertise détaillée, et les débats contradictoires auxquels ils se livrent garantissent une qualité de résultat nettement supérieure à celle atteinte par les grands centres qui, attaquant l'ensemble du génome, ne peuvent pas trop s'attarder sur les zones litigieuses.

Deux des cartes publiées se disent « de deuxième génération », mais cette appellation, comme le numéro de version des logiciels ou la notion de nouveau modèle automobile, repose sur des bases hautement subjectives. À mon avis, seule mérite vraiment ce nom la carte du chromosome 16 [8], coordonnée par le laboratoire de Robert Moysis (Los Alamos) et bénéficiant de la collaboration active du groupe de Grant Sutherland (Adelaide, Australie). Elle s'appuie, comme les trois autres, sur la construction d'un *contig* de YAC couvrant sans trop de lacunes l'ensemble du chromosome ; mais elle offre une résolution beaucoup plus fine grâce à un deuxième niveau reposant sur l'alignement de deux mille cosmides. Rien de très étonnant puisque le groupe de Los Alamos projetait à l'origine (avant la découverte des YAC) d'assembler un *contig* de cosmides sur ce chromosome... L'entreprise s'est avérée impossible avec ces seuls clones, trop petits (30 à 40 kilobases) par rapport aux dizaines de mégabases à baliser ; les YAC, arrivés à la rescousse, ont comblé les trous et affermi une carte qui serait, sinon, restée incomplète. Mais le travail préalable effectué permet, une fois la carte de YAC établie, de passer rapidement à ces réactifs (actuellement) irremplaçables que sont les cosmides, seuls clones de taille raisonnable que l'on puisse facilement préparer, analyser en détail ou séquencer.

Les trois autres articles [9-11] correspondent, eux, à l'affinement de cartes physiques reposant sur des YAC et à l'obtention de *contigs* fiables couvrant la quasi-totalité des chromosomes étudiés (le 3, le 2 et le 22). Cartes fiables et détaillées, directement utilisables pour la recherche de séquences transcrites par *exon-trapping* ou *cDNA capture* ; mais pour ces trois chromosomes, on est encore loin du « prêt à séquencer » *(sequence-ready map)* dont se rapproche beaucoup la carte actuelle du 16.

Des stratégies concurrentes mais interdépendantes

L'avancée des travaux sur les EST ou sur les cartes physiques présente une grande continuité, et le choix de la date de publication est assez arbitraire. La sortie fin septembre 1995 de ce *Genome Directory* doit donc beaucoup à des impératifs politiques et médiatiques ; elle autorise à tout le moins un tour d'horizon fort opportun. Bien qu'absente de cet annuaire, la carte génétique ne peut être oubliée : sa version à 5 000 marqueurs, due au projet mené par Jean Weissenbach, paraîtra bientôt. Elle a sans doute atteint le niveau de finesse nécessaire et raisonnable : ce volet est en quelque sorte clos. Le Projet Génome se recentre ainsi autour de l'amélioration des cartes physiques, du séquençage et du positionnement des EST, et de l'amorce d'un effort sérieux pour déchiffrer l'ensemble de notre ADN. Ces différents aspects ne sont pas indépendants, et leur poids respectif reste incertain. Comment équilibrer les travaux entre cartes par chromosome, qui font intervenir de nombreux groupes possédant des expertises très diverses, et carte générale, qui est plutôt l'affaire de structures importantes comme le CEPH ? La richesse du catalogue des EST, et l'obtention sans doute prochaine d'informations sur leur position dans le génome rendent-ils moins urgent le séquençage exhaustif ? Autant d'interrogations auxquelles les décideurs qui financent ces recherches devront apporter une première réponse dans les mois qui viennent – en attendant que le succès ou l'échec des travaux en cours ne change, peut-être, les données de la question.

Références

1. The Genome Directory. *Nature* 1995 ; 377 (suppl).
2. Adams MD, Kerlavage AR, Fleischmann RD, *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 1995 ; 377 (suppl) : 3-174.
3. Collins F. Positional cloning moves from traditional to perditional... *Nat Genet* 1995 ; 9 : 347-50.
4. Bellanne-Chantelot C, Lacroix B, Ougen P, Billault A, Beaufils S, *et al.* Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992 ; 70 : 1059-68.
5. Cohen D, Chumakov I, Weissenbach J. A first-generation physical map of the human genome. *Nature* 1993 ; 366 : 698-701.
6. Chumakov IM, Rigault P, Le Gall I, *et al.* A YAC contig map of the human genome. *Nature* 1995 ; 377 (suppl) : 175-298.
7. Donis-Keller H, Green P, Helms C, *et al.* A genetic linkage map of the Human Genome. *Cell* 1987 ; 51 : 319-37.
8. Doggett NA, Goodwin LA, Tesmer JG, *et al.* An integrated physical map of human chromosome 16. *Nature* 1995 ; 377 (suppl) : 335-66.
9. Gemmill RM, Chumakov I, Scott P, *et al.* A second-generation YAC contig map of human chromosome 3. *Nature* 1995 ; 377 (suppl) : 299-320.
10. Krauter K, Montgomery K, Yoon SJ, *et al.* A second-generation YAC contig map of human chromosome 12. *Nature* 1995 ; 377 (suppl) : 321-34.
11. Collins JE, Cole CG, Smink LJ, *et al.* A high-density YAC contig map of human chromosome 22. *Nature* 1995 ; 377 (suppl) : 367-96.

Med Sci *2001 ; 17 : 81-84*

ADNc : LES INCONTOURNABLES

Six ans ont passé depuis la chronique précédente. Le premier brouillon de la séquence génomique humaine est paru depuis près d'un an, et l'on pourrait penser que cela rend caduque l'approche par les EST. Il n'en est rien, car – contrairement au cas de la levure – l'interprétation de la séquence humaine est extrêmement délicate, et la définition des gènes par la seule analyse informatique s'avère quasiment impossible. La base de données publique dbEST contient maintenant des millions de séquences partielles (plus de trois millions en ce qui concerne les séquences humaines), et les EST deviennent un outil indispensable pour essayer de déterminer le nombre, la position et l'étendue des gènes humains. Les efforts redoublent pour obtenir des EST « complets », car la plupart d'entre eux correspondent encore à des séquences très partielles, généralement en 3' du gène. Par ailleurs, ces collections de clones constituent des réactifs essentiels pour la fabrication des « puces à ADN » dont le rôle ne cesse de grandir.

Au tout début des années 1990, l'on peinait à séquencer une ou deux centaines de kilobases d'ADN humain, et les résultats scientifiques de tels travaux apparaissaient assez minces par rapport aux efforts et aux fonds investis [1].

Les débuts des EST

L'option des ADNc, leur déchiffrage partiel mais massif apparurent vite comme une alternative réaliste au séquençage intégral. Lancée en franc-tireur par Craig Venter [2], très largement médiatisée par le scandale que soulevèrent les tentatives de brevets sur ces séquences partielles, l'approche des EST *(expressed sequence tags)* allait jouer un rôle central dans l'exploration de notre génome tout comme dans la mise au point de produits nouveaux. L'accumulation de ces étiquettes fut entreprise dans le secteur public, les données étant déposées au fur et à mesure dans la base *ad hoc* dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). Elle s'effectua aussi au sein de groupes privés comme *Human Genome Sciences* ou *Incyte*, qui gardaient ces résultats pour eux ou en vendaient l'accès à des industriels de la pharmacie. D'autres entreprises, notamment Merck, choisirent de collaborer avec les laboratoires académiques et de financer l'obtention de données publiques. Le nombre d'EST répertoriés augmentait rapidement, atteignant (pour dbEST) 325 000 en janvier 1996 et 550 000 un an plus tard.

Redondant par principe (puisque l'on détermine les séquences partielles de clones pris au hasard dans des banques d'ADNc), cet ensemble était analysé par des systèmes dénommés *gene index*[1] comparant toutes les séquences afin de les regrouper en *clusters*

1. Il existe aujourd'hui plusieurs de ces systèmes, en général accessibles sur le réseau. En voici une liste :
UNIGENE : http://www.ncbi.nlm.nih.gov/UniGene/index.html.
TIGR : http://www.tigr.org/tdb/tgi.shtml.
IMAGENE : http://image.llnl.gov/image/imagene/current/bin/search.
STACK : http://www.sanbi.ac.za/Dbases.html.
GENEXPRESS : http://idefix.upr420.vjf.cnrs.fr/IMAGE/Genexpress.html.

censés représenter chacun un transcrit : c'est à partir de ces données qu'avait été faite l'estimation d'environ cent mille gènes humains aujourd'hui très discutée [3]. L'utilité des EST fut encore renforcée par la localisation d'un grand nombre d'entre eux sur notre génome : effectuée massivement grâce à l'emploi des hybrides d'irradiation, celle-ci devait aboutir en 1998 au positionnement de plus de trente mille étiquettes [4]. Ces EST « pré-positionnés » devenaient ainsi une source gratuite de gènes-candidats pour les projets de génétique humaine, une fois passée l'indispensable étape de la localisation.

Le grand retour du séquençage

Pendant ce temps les techniques de séquençage s'amélioraient progressivement. Des réactifs plus stables, des « séquenceurs » plus performants, une exploitation informatique plus sophistiquée, et surtout une prise de conscience de la nécessité de planifier une opération de « Très Grand Séquençage » comme une entreprise industrielle, aboutissaient, malgré l'absence d'une révolution technique majeure, à rendre la lecture de mégabases d'ADN possible et presque abordable. Ces progrès devinrent évidents pour tous avec l'obtention en 1996 de la séquence complète de la levure. Les treize mégabases déchiffrées constituaient de loin le plus important ensemble jamais obtenu, et montraient que de tels projets étaient devenus viables ; et l'utilité de ces données était attestée par la découverte de près de trois mille gènes « nouveaux » (en plus des trois mille déjà répertoriés) chez cet organisme que l'on croyait pourtant très bien connaître après des lustres d'études biochimiques et génétiques. La démonstration était faite que seul le séquençage intégral permettait d'accéder à l'ensemble des données géniques. Parallèlement, l'avancée rapide du séquençage du nématode (cent mégabases) donnait confiance dans notre capacité à passer une étape de plus et à aborder les 3 000 mégabases de l'ADN humain.

C'est ainsi que fut lancé à partir de 1997 le déchiffrage intégral de notre ADN. Réalisé principalement en Grande-Bretagne et aux États-Unis (avec une participation minoritaire de la France, du Japon, de l'Allemagne), aiguillonné par la concurrence avec l'initiative privée de Craig Venter et de l'entreprise *Celera*, il donnait lieu en juin 2000 à l'annonce conjointe de la première séquence « brouillon » de notre génome. Annonce plus politique que scientifique, et séquence dont les critères de qualité sont assez mal définis (à l'exception des chromosomes 21 et 22 qui, eux, sont de qualité « finie ») ; il n'en reste pas moins qu'environ 90 % de notre génome est aujourd'hui déchiffré avec un taux d'erreur d'une fraction de pour cent, et que toutes ces données sont accessibles à quiconque dispose d'un accès Internet. Avancée considérable, dont on pourrait attendre une explosion de nouveaux résultats et la résolution de maintes controverses. Le moins que l'on puisse dire est que ce n'est pas toujours le cas, comme le montre l'incertitude actuelle sur le nombre de gènes humains dont l'estimation varie de moins de trente mille à plus de cent vingt mille... [3].

Les EST continuent

C'est que l'interprétation de la séquence, et notamment la détection des gènes, posent des problèmes redoutables dans le cas de l'homme (et des mammifères en général). La complexité des structures géniques, la multiplicité d'exons souvent très petits, l'existence de nombreux pseudogènes, le caractère relativement flou des signaux d'épissage et encore plus des promoteurs... tout cela rend actuellement impossible une interprétation *a priori* fiable des données humaines, même lorsqu'il s'agit de séquence « finie » : les difficultés d'annotation des chromosomes 21 et 22 le montrent très clairement [5, 6]. Du coup, loin d'être devenue caduque, l'analyse des EST et plus généralement

l'étude des ADNc continue de plus belle[2]. En témoigne notamment la progression du nombre de séquences nouvelles dans dbEST : trois millions entre 1999 et 2000 (dont un million de séquences humaines), bien plus qu'au cours des cinq premières années de cette bases de données. En fait, les informations obtenues grâce aux ADNc jouent actuellement un rôle indispensable dans l'annotation de la séquence humaine : c'est bien souvent en comparant les EST et la séquence génomique que l'on repère des exons, d'autant plus que de telles comparaisons s'accommodent assez bien de séquence de qualité « brouillon ». De plus, l'analyse des ADNc met aujourd'hui en évidence des phénomènes susceptibles de modifier profondément la compréhension de notre matériel génétique et même celle de l'évolution.

De très nombreuses séquences partielles d'ADNc continuent à être obtenues dans le cadre de projets EST, en ciblant le plus possible des tissus spécialisés dont les gènes spécifiquement exprimés n'ont pas encore été échantillonnés. De plus, un effort général est effectué pour obtenir et déchiffrer des ADNc complets *(full length)* représentant l'intégralité du transcrit. Un point général de ces travaux a été récemment effectué lors du colloque « Transcriptome 2000 » tenu à l'Institut Pasteur début novembre 2000[3]. Il a fourni la plupart des informations résumées dans cette chronique – qui n'est pas pour autant un compte-rendu de cette réunion, au cours de laquelle bien d'autres thèmes ont été abordés.

La nouvelle vague des projets EST humains a débuté dès 1997, avec le programme CGAP du *National Cancer Institute* dont l'objectif était d'explorer les tissus tumoraux en obtenant des banques d'ADNc à partir de tumeurs ou de fragments microdisséqués. La fraction de séquences « nouvelles » (non homologues à des séquences déjà contenues dans dbEST) s'est avérée importante dans ces données, et a permis d'accroître très nettement la représentativité de dbEST. Ce type d'effort est poursuivi, tant dans le cadre de CGAP que pour différents tissus spécialisés. L'équipe de Bento Soares (Université d'Iowa, États-Unis), par exemple, effectue des soustractions successives de banques d'ADNc afin d'augmenter la proportion de séquences nouvelles. Et, bien que cet article soit centré sur les travaux portant sur l'homme, n'oublions pas les nombreux projets menés sur d'autres organismes pour lesquels les EST sont souvent la principale information génomique actuellement disponible...

La quête de l'ADNc *full length*

Depuis quelque temps, l'obtention de jeux importants d'ADNc complets, et leur séquençage, sont devenus le but de nombreuses équipes. La plupart des clones d'ADNc à partir desquels des EST ont été déterminés sont en effet courts et peu représentatifs de l'ARN messager dont ils dérivent. La majeure partie des banques séquencées ont été construites par l'équipe de Bento Soares, en utilisant un amorçage sur la queue polyA de l'ARN messager, suivi de deux traitements d'égalisation pour augmenter la proportion de séquences peu exprimées (parmi lesquelles se trouve la plus forte proportion de séquences « nouvelles »). Dans ces conditions la taille moyenne des *inserts*, qui se situent tous à l'extrémité 3' du transcrit, est de l'ordre de la kilobase. Or beaucoup d'ARN messagers ont une longueur de plusieurs milliers de bases, et la région 3' non codante mesure souvent plus d'une kilobase : on pourra s'en persuader en examinant un jeu pris au hasard de grands ADNc humains sur la base de données du *Kazusa Sequencing Institute*

2. En tout état de cause les EST sont aujourd'hui le réactif presque obligatoire pour la construction de réseaux d'ADN sous forme de *macroarrays* ou *microarrays*. Je ne discute pas ici cet aspect qui est très important, même si l'on peut penser que les réseaux seront à l'avenir fondés de plus en plus sur des oligonucléotides de synthèse [7].

3. Transcriptome 2000, organisé par Charles Auffray, Bento Soares et Sumio Sugano à l'Institut Pasteur du 6 au 9 novembre 2000. Voir le site correspondant http://www.vjf.cnrs.fr/transcriptome/.

(Japon) (http://zearth.kazusa.or.jp/huge/). Les EST 3' et 5' obtenus à partir d'un tel clone peuvent donc ne contenir aucune séquence codante. Ils ne révèlent alors rien sur la nature de la protéine codée par le gène correspondant, et ne permettent aucune prédiction fonctionnelle. Pendant assez longtemps, la démarche généralement suivie a consisté à obtenir, « à l'unité », le clone d'ADNc complet à partir d'un EST jugé particulièrement intéressant en raison de son profil d'expression et/ou de sa localisation chromosomique. Cela était réalisé en criblant des banques d'ADNc spécialisées, et en pratiquant différentes manœuvres d'extension (5' RACE, par exemple) à partir du clone existant et de l'ARNm d'un tissu judicieusement choisi.

La nouvelle tendance, en cours depuis quelque temps déjà, consiste à s'efforcer d'obtenir des banques contenant une forte proportion d'ADNc complets, puis à identifier les clones correspondant à ce critère grâce à un séquençage partiel, et enfin à déchiffrer intégralement celles des séquences qui sont à la fois *full length* et nouvelles. Les méthodes employées sont variées. Certaines équipes effectuent une sélection sur la taille de l'ARNm et/ou sur celle de l'ADNc après rétrotranscription (Bento Soares, Université d'Iowa, États-Unis ; Omahu Ohara, *Kazusa DNA Research Institute*, Japon ; Stefan Wiemann, *Deutsche Krebs Forschungs Zenter*, Heidelberg, Allemagne ; Robert Strausberg, *National Cancer Institute*, Bethesda, États-Unis). D'autres utilisent la « coiffe » présente à l'extrémité 5' de l'ARN messager pour isoler les ARNm complets par un système de capture *(cap trapping)* biotine/avidine (Judy Margolin, Baylor, États-Unis ; June Kawai, *Riken*, Tsukuba, Japon) ou pour procéder à la ligation en 5' d'un oligonucléotide qui se retrouve ensuite dans l'ADNc (Sumo Sugano, Université de Tokyo, *Nedo cDNA project* (MITI, Japon). Notre Génoscope mène également un travail de ce type en utilisant des banques d'ADNc produites par l'entreprise *Lifetech*.

Les équipes effectuent ensuite une séquence en 5' et en 3' afin de déterminer si l'ADNc est complet ou presque (au minimum, présence d'un ATG dans un contexte de séquence approprié), et s'il est nouveau (en tant qu'ADNc complet), avant de procéder à sa séquence complète. Le rendement dépend beaucoup des projets et des techniques : de quelques pour cent à la moitié de clones obtenus s'avèrent être complets. Il n'est d'ailleurs pas toujours évident de le déterminer, le critère de l'ATG et de son environnement n'étant pas très restrictif ; par ailleurs la lecture révèle souvent des amorçages internes qui, lors du clonage, ont eu lieu sur des régions internes riches en A et non sur la région polyA 3' terminale. Enfin le débrouillage des différents clones d'ADNc correspondant aux épissages alternatifs *(voir plus loin)* demande beaucoup de temps. Les différents projets annoncent avoir obtenu de quelques centaines à près de 20 000 séquences de clones *full length*. Il existe certainement de nombreux doublons dans ces travaux, d'autant plus que les séquences sont transmises aux bases de données avec un certain retard, et en tous cas après la fin du séquençage complet d'une série de clones. Il n'en reste pas moins que l'ensemble de ces travaux devrait fournir rapidement la séquence complète des transcrits pour les vingt à vingt-cinq mille gènes humains les plus approchables (parce que exprimés dans des tissus accessibles).

Généralité de l'épissage alternatif

La fréquence et la complexité des phénomènes d'épissage alternatif ressortent de manière évidente à l'examen de ces résultats. Ils se présentent sous de nombreuses formes : variation de la limite des exons (extension ou raccourcissement), déplacement du site d'initiation de la transcription ou du site de polyadénylation, exons ou introns cryptiques, saut d'exons, répétition d'exons... L'obtention de séquences d'ADNc *full length* révèle ces phénomènes, dans la mesure où des clones complets s'avèrent différents après séquençage tout en partageant d'importantes zones homologues à 100 % ; la comparaison avec la séquence « brouillon » du génome joue un rôle important en permettant

de montrer que ces formes correspondent bien au même gène. Réciproquement, bien sûr, cette comparaison définit avec précision l'ensemble des exons. On parle maintenant de 50 % de gènes présentant un épissage alternatif, et le comité de nomenclature de HUGO *(Human Genome Organisation)* commence à attribuer des noms spécifiques à ces différentes formes...

Dans de nombreux cas, l'épissage alternatif fausse les résultats des *gene indexes*. Pour plus du tiers des ADNc étudiés dans le projet du DKFZ (Heidelberg, Allemagne), les séquences complètes montrent que deux, trois ou même quatre *clusters* Unigene correspondent en fait au même gène avec des exons alternatifs. Ce sont là autant de gènes « en trop » dans les dénombrements fondés sur ce type d'analyse... Plusieurs exemples montrant cinq ou six épissages alternatifs du même gène ont été présentés, sans que l'on sache si toutes les formes sont biologiquement significatives. Cette question de la pertinence biologique est évidemment capitale, et les informations à ce sujet restent fragmentaires[4]. J'aurais personnellement tendance à penser que la majorité de ces épissages ont un rôle fonctionnel, et que ces phénomènes permettent de produire 2 ou $3 \times N$ protéines à partir de N gènes (chacun donnera à N la valeur qui lui convient[5]). Ces possibilités montrent l'intérêt que peut présenter pour l'organisme une structure morcelée des gènes. Elles en constituent peut-être une justification, qui s'ajoute à l'habituel argument évolutif : la construction facilitée de protéines multifonctionnelles par assemblage de domaines protéiques ayant évolué indépendamment. N'oublions pas les modifications post-traductionnelles qui, elles, peuvent donner naissance à plusieurs entités à partir de chaque séquence d'acides aminés...

D'autres surprises ?

Un dernier élément nouveau, encore imprécis et qualitatif, ressort de ces travaux : il semble que les cas de gènes présents à plusieurs exemplaires, sur le même chromosome ou sur des chromosomes différents, soient en train de se multiplier. Ces indications résultent de la comparaison entre les ADNc complets et la séquence brouillon du génome humain ; elles sont encore fragmentaires, et peuvent dans certains cas être liées à des erreurs dans ce brouillon (séquences attribuées au mauvais chromosome ou à la mauvaise région) ; mais il se pourrait que le phénomène soit assez général et ajoute à la complexité de notre génome – tout comme à la difficulté de définir le résultat du *sweepstake* de *Cold Spring Harbor* ! (voir http://www.ensembl.org/Genesweep/).

On le voit, la saga des EST n'est pas terminée. Loin de correspondre à une étape désormais caduque dans l'analyse du génome humain, l'étude des ADNc s'avère aujourd'hui indispensable à la compréhension des nouvelles données sur notre génome ; combinée avec la séquence brouillon, elle est en train de changer l'image que nous nous faisons de notre ADN et, sans nul doute, de nous aider à comprendre comment trente mille gènes « seulement » peuvent rendre compte de la complexité de notre organisme...

4. Il semble – mais les indications sont encore fragmentaires – que les schémas d'épissage alternatifs diffèrent souvent entre l'homme et la souris. On peut en déduire que cela indique leur caractère artéfactuel... ou au contraire que cela démontre leur rôle possible dans la différenciation des espèces !

5. L'estimation faite par l'équipe du Génoscope à partir de la séquence « brouillon » couvrant près de 90 % de notre génome confirme celle publiée en juin dans *Nature Genetics* [8] et prenant en compte 42 % de cet ensemble : moins de 30 000 gènes.

Références

1. Martin-Gallardo A, McCombie WR, Gocayne JD, *et al.* Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nat Genet* 1992 ; 1 : 34-9.
2. Adams MD, Kerlavage AR, Fleischmann RD, *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 1995 ; 377 : 3-174.
3. Roest Crollius H, Jaillon O. Le nombre de gènes dans le génome humain : les paris sont ouverts. *Med Sci* 2000 ; 16 : 988-90.
4. Deloukas P, Schuler GD, Gyapay G, *et al.* A physical map of 30,000 human genes. *Science* 1998 ; 282 :744-6.
5. Dunham I, Shimizu N, Roe BA, *et al.* The DNA sequence of human chromosome 22. *Nature* 1999 ; 402 : 489-95.
6. Hattori M, Fujiyama A, Taylor TD, *et al.* The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* 2000 ; 405 : 311-9.
7. Jordan B. Jusqu'où iront les puces ? *Med Sci* 2000 ; 16 : 950-3.
8. Roest Crollius H, Jaillon O, Bernot A, *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat Genet* 2000 ; 25 : 235-8.

ADNc ou ADN génomique ?

La saga mouvementée des ADNc est intéressante à plus d'un titre. Elle indique comment des stratégies scientifiques peuvent s'opposer pour finalement se rejoindre, et aussi à quel point le « meilleur choix » dépend de multiples facteurs : taille de l'organisme, présence ou non de grands introns, niveau des connaissances préalables... Les avocats d'une focalisation du programme Génome sur les gènes (en laissant de côté introns et séquences intergéniques dont on savait dès le départ qu'ils formaient la majeure partie de l'ADN humain) n'ont pas manqué. C'était la thèse que soutenait Sydney Brenner dans les années 1980 ; la première ébauche du Programme Génome français, rédigée en 1990 par Philippe Kourilsky, centrait le projet sur l'étude des ADNc, thème prioritaire clairement affirmé. À l'inverse, la position officielle du programme nord-américain était catégorique : pas question de se disperser dans l'étude prématurée des gènes, il fallait impérativement cartographier puis séquencer l'ensemble de l'ADN humain. La percée inattendue de Craig Venter a été une surprise majeure. Devant l'efficacité de sa méthode pour révéler de nouveaux gènes, celle-ci a rapidement été adoptée par tous les programmes nationaux, avec des succès d'ailleurs très inégaux. Quelques années plus tard, la réussite du séquençage intégral de la levure, la relative facilité d'interprétation de cette séquence et les nombreux gènes ainsi mis en évidence faisaient de nouveau pencher la balance vers la lecture de l'ADN génomique. Aujourd'hui, même avec une séquence humaine fiable à 99,99 %, l'on constate que les ADNc – et de préférence des ADNc complets – sont indispensables à sa compréhension...

This page intentionally left blank

4. COORDINATION OU CONCURRENCE ?

Les questions d'organisation ont joué un grand rôle dans le déroulement des programmes Génome et, à ce titre, ont fait l'objet de plusieurs chroniques. C'était la première fois qu'une recherche biologique était menée à une telle échelle, et cela impliquait une certaine planification pour éviter les « doublons » et les gaspillages. Pour autant, et contrairement à une perception assez répandue, il n'y a pas eu d'organisation stricte ni de répartition autoritaire des tâches : c'est seulement pour le séquençage proprement dit que l'on s'en est rapproché, avec un « partage » des chromosomes à séquencer et des règles communes de diffusion des résultats. Pour la phase de cartographie et de séquençage d'ADNc, l'on a plutôt assisté au développement de programmes nationaux (États-Unis, Grande-Bretagne, France...). La Human Genome Organisation *(HUGO) a fait circuler l'information et a aidé à la coordination de ces projets sans assumer de fonction dirigeante – ne serait-ce que parce qu'elle ne disposait que de faibles ressources financières. Ce thème de l'organisation recouvre aussi les modalités de travail des* Genome centers *: le délicat équilibre entre études systématiques et recherches ciblées, la difficulté à motiver des chercheurs qualifiés pour effectuer des tâches assez répétitives, ou encore la gestion des retombées de l'étude du génome sur le quotidien des laboratoires... Les cinq textes qui suivent, rédigés de 1992 à 2001 (donc, à l'échelle de ce programme, depuis l'âge des cavernes jusqu'aux temps modernes), explorent ces différents aspects.*

Med Sci *1992 ; 8 : 476-482*

LES CONTRADICTIONS DU GÉNOME

Pour commencer, une chronique, parue au début de 1992, qui examine le mode de fonctionnement des genome centers *américains (que j'avais, pour la plupart, visités l'année précédente), analyse les difficultés qu'ils rencontrent et décrit les diverses solutions utilisées pour y faire face. On pourra noter que j'y envisage Généthon comme une sorte de « centre de ressources » : c'est ainsi qu'il était présenté à l'époque, afin sans doute de limiter les réactions négatives de la communauté scientifique française à l'égard de cette tentative inédite, qui suscitait étonnement, incrédulité et jalousies...*

Une routine abrutissante ?

C'est bien ainsi que sont souvent perçus les « programmes Génome » développés dans plusieurs pays. Beaucoup imaginent qu'ils reposent, dans des laboratoires dédiés à cet usage, sur une armée de techniciens effectuant des tâches répétitives et sans imagination sous la houlette de quelques chercheurs. Selon cette vision assez répandue il ne s'agit en somme que de répéter, presque à l'infini, les gestes courants du laboratoire de biologie moléculaire ; l'on imagine certes que quelques opérations sont prises en charge par des automates, mais l'ensemble du processus apparaît singulièrement peu créatif.

Pour ceux qui connaissent réellement les laboratoires impliqués il est clair que cette vision est dans une très large mesure fausse. Le Génie génétique n'a qu'une quinzaine d'années d'existence, et il est encore en évolution rapide. Les méthodes bougent, progressivement ou par à coups (que l'on songe à ce qu'ont changé la PCR, ou la mise au point des YAC...) et la stratégie optimale à un moment donné peut devenir très vite caduque. C'est d'ailleurs ce qui explique en grande partie la faible automatisation des laboratoires, car la mise au point d'appareils efficaces demande du temps et il est difficile de l'effectuer dans un contexte aussi changeant. Cette évolution perpétuelle a deux conséquences : la nécessité de changer son fusil d'épaule à intervalles rapprochés, afin de ne pas poursuivre une opération selon une technique maintenant obsolète et remplaçable par une autre cinq ou dix fois plus efficace ; et par ailleurs, le fait que les méthodes n'ont pas le temps de se stabiliser en une routine bien au point. Il faut donc que le personnel soit capable d'apprécier l'importance de nouveautés, de les mettre en œuvre, et également de comprendre ce qui se passe quand l'expérience ne donne pas les résultats attendus : ce que nos collègues anglo-saxons désignent du nom très expressif de *trouble-shooting*.

Il reste néanmoins vrai que l'affinement de la carte génétique humaine, la construction de *contigs* de clones s'étendant sur des dizaines de mégabases et, bien sûr, le séquençage de grandes régions d'ADN sont des entreprises comportant une part non négligeable d'un travail répétitif dont l'exécution doit être impeccable sous peine d'arriver à un rapport qualité/prix désastreux : pas question ici de tolérer l'improvisation un peu brouillonne qui est souvent rencontrée dans nos laboratoires « académiques ». Un tel manque d'organisation est admissible à la rigueur pour des travaux de petite ampleur, dans un cadre académique et au cours d'une phase de découverte de la recherche par des étudiants. Mais elle n'a pas sa place dans de grands programmes mettant en jeu des

appareillages relativement sophistiqués et – du fait même de cette automatisation – des quantités considérables de réactifs coûteux : pensons à la consommation de l'enzyme Taq polymérase dans un centre comme celui de Maynard Olson et David Schlessinger à Saint Louis (MO, USA) qui effectue près de mille réactions de PCR par jour...

Comment s'organiser ?

Les solutions apportées sont très variables et j'en ai observé diverses variétés au cours de mes visites de laboratoires à travers le monde. Certains pratiquent le « tout technicien », c'est-à-dire ont recours à du personnel d'exécution, formé à une tâche précise et pouvant l'effectuer – on l'espère – avec fiabilité et sans trop de lassitude pendant plusieurs années. C'est par exemple le cas du laboratoire de Tada-Aki Hori au Japon, qui assure la localisation par hybridation *in situ* des très nombreux cosmides isolés par l'excellent groupe de Yusuke Nakamura (Tokyo) à partir des chromosomes 3 et 11. L'équipe de Hori, installée dans un centre de recherches de la *Science and Technology Agency* (un peu l'équivalent de notre CEA et du *Department of Energy* américain) à Chiba, près de Tokyo, pratique l'*in situ* non radioactif avec un matériel classique. Les microscopes sont équipés d'appareils photo ; pas de caméra CCD et encore moins de microscope confocal. Les localisations reposent sur le travail de trois techniciennes, supervisées par un chercheur, qui effectuent plusieurs centaines de localisations par an. Les résultats sont de bonne qualité, et les opératrices paraissent contentes de leur sort. Le Généthon, dont nous reparlerons plus loin, a fait lui aussi le choix d'employer essentiellement du personnel technique, encore qu'il ait été amené à revoir à la hausse le niveau de formation demandé après quelques déboires initiaux ; enfin certains *Genome centers* aux États-Unis comme celui que dirige Rick Myers à San Francisco (chromosome 4) ou Glen Evans au *Salk Institute* (chromosome 11) fonctionnent de la même façon. Il faut noter à cet égard que le terme de technicien recouvre aux États-Unis une réalité assez différente de ce qu'elle est en France : dans cette société très mobile beaucoup sont en fait des étudiants qui jouent ce rôle pendant deux ou trois ans, le temps de faire des économies avant de reprendre leurs études et d'achever leur maîtrise ou de commencer un doctorat. Il y a donc dans cette catégorie un renouvellement constant et une motivation que la routine n'a pas eu le temps d'émousser ; quant aux difficultés de positionnement que nous connaissons souvent avec nos techniciens professionnels elles se posent autrement dans ce contexte.

D'autres laboratoires choisissent au contraire d'employer surtout des étudiants et post-doctorants. C'est alors que le caractère systématique du travail commence à poser problème. Si la tâche qui consiste à établir une carte physique complète du chromosome 7, par exemple, a une justification scientifique indéniable, elle n'en comporte pas moins une part importante de travail répétitif. C'est le cas de toute recherche en Génie génétique, où l'on finit toujours par passer beaucoup de temps à préparer des ADN, effectuer des hybridations, des réactions PCR, couler, faire migrer et interpréter des gels... mais la « rentabilité » de tels exercices semble beaucoup plus élevée lorsque l'on travaille sur une maladie génétique que dans le cadre d'une cartographie systématique. Le *post-doc* qui trime sur la recherche de gènes contenus dans un jeu de cosmides provenant du bras court du chromosome 4 peut toujours rêver qu'il va découvrir le gène de la maladie de Huntington, et qu'il aura les honneurs de revues comme *Cell, Science* ou *Nature*. En tout état de cause, pourvu qu'il isole quelques YAC, détermine leur carte et y repère grâce aux îlots CpG, aux *Zoo blots* ou à toute autre méthode quelques gènes, qu'il montre dans quel tissu ces derniers sont transcrits... il aura assemblé les éléments d'un article honorable qu'il pourra soumettre à *Genomics* ou à *Nucleic Acids Research*. Il tirera aussi de ce travail les éléments de quelques séminaires et communications dans les colloques. Il aura enfin le sentiment de maîtriser, en quelque sorte d'être propriétaire, de

ces résultats et d'avoir apporté à l'édifice de la Biologie une pierre – petite sans doute mais bien identifiable et résultant indéniablement de sa propre activité.

Pourtant on ne peut pas dire qu'un tel travail soit réellement créatif, ni au niveau conceptuel ni au niveau technique : il peut se limiter à l'application de méthodes éprouvées sans idée *a priori* et sans que l'inventivité n'entre en jeu. Ce sera pourtant une recherche qualifiée de fondamentale, ce qu'elle est peut-être dans son objet mais, à mon avis, pas dans sa démarche. La contribution ainsi apportée est limitée, entre autres, par le fait qu'elle ne s'intègre pas facilement dans une vision d'ensemble, que le degré de précision et de validité des informations obtenues n'a pas été soumis à une évaluation dont les critères soient bien définis, et que seules les données présentées dans une éventuelle publication – qui sera sans doute très abrégée vu son intérêt limité – finiront éventuellement par être consignées dans les banques de données. On peut facilement argumenter que, sauf exception, les efforts ainsi investis seront moins utiles à l'avancement des connaissances que le travail qu'aurait mené le même *post-doc* dans le cadre d'une entreprise génomique de grande ampleur comme celle qui est développée par Maynard Olson et David Schlessinger. Pourtant, dans le deuxième cas, le mieux qu'il puisse espérer est d'être signataire, avec beaucoup d'autres, d'un ou deux articles relatant quelques étapes de l'avancement de la carte d'un chromosome.

Une solution partielle consiste à organiser le travail afin que chaque chercheur ait une double activité : d'une part une participation à un projet général, peu productif dans l'immédiat mais constituant l'ossature de la stratégie du laboratoire (et justifiant souvent son financement *via* des contrats « Génome »), d'autre part un sujet personnel plus ciblé rattaché à ce thème mais susceptible de s'avérer intéressant et rentable en termes de résultats publiables. C'est une tactique qui peut être payante si ces sujets secondaires sont astucieusement choisis de façon à profiter au maximum des clones, des données acquises dans le cadre du programme principal ainsi que des méthodes mises en place pour son exécution. Mais elle porte en elle le risque d'une dérive, le danger qu'imperceptiblement le projet latéral devienne la priorité du chercheur et que l'autre activité soit délaissée : dérive bien compréhensible... L'histoire récente du tri de chromosomes dans le laboratoire de Lawrence Livermore, en Californie, en fournit un exemple. Cette équipe, avec celle de Los Alamos au Nouveau Mexique, a été parmi les premières à maîtriser la préparation de banques spécifiques à partir de chromosomes ainsi triés. Les deux centres se sont alors partagé les chromosomes humains afin d'établir, au service de la communauté, un jeu complet de librairies spécifiques pour chaque chromosome, tout en continuant à avoir une activité de recherche concernant l'amélioration des méthodes de tri et leur emploi pour étudier le polymorphisme de taille des chromosomes. Le parti choisi à Lawrence Livermore a été d'affecter une machine au tri de routine (celui qui sert à la construction de banques), une autre au tri de recherche. Bien que l'ensemble soit sous la responsabilité des mêmes scientifiques, il s'est produit progressivement un glissement qui a abouti à ce que l'essentiel de l'attention se porte sur le tri de recherche et la machine correspondante. La qualité des banques produites (et distribuées) par Lawrence Livermore n'a pas tardé à s'en ressentir ; il s'est trouvé en particulier que plusieurs préparations de chromosomes ont été contaminées par des bactéries du type *Pseudomonas*, et que les banques résultantes contenaient quelques pour cent de clones provenant de l'ADN de cette bactérie. Cette contamination était catastrophique pour certains utilisateurs : ces banques ont en effet souvent été employées pour construire des sous-banques d'îlots CpG, pour lesquelles on sélectionne par divers artifices les clones dont l'ADN contient des séquences riches en CpG, qui correspondent aux « îlots HTF », associés en général à des gènes. Or l'ADN de *Pseudomonas* contient une forte proportion de ce dinucléotide... de sorte que les banques d'« îlots HTF » obtenues à partir de ces librairies contenaient surtout des clones de *Pseudomonas*...

On voit que la dissociation entre les deux aspects du travail est très dangereuse : la solution a consisté à rénover et améliorer les machines et surtout à établir un planning

selon lequel la même machine est utilisée alternativement une semaine pour le tri de routine et une pour la recherche : ainsi s'assure-t-on d'un intérêt soutenu des chercheurs pour le parfait état du matériel qui sert pour le tri de routine. Mais le meilleur exemple d'intégration entre recherche ciblée et travail systématique, je l'ai rencontré dans le laboratoire de David Ward, pape des sondes non radioactives et apôtre de l'hybridation *in situ*. On connaît ses études sur la mise au point et la diffusion de l'hybridation *in situ* avec des sondes non radioactives : elles ont transformé une technique longue et délicate, valable seulement entre les mains de quelques spécialistes et relativement imprécise du fait de la taille des grains d'argent qui signalent la sonde... en une méthode fiable et rapide, transférable dans de nombreux laboratoires et avec laquelle un *post-doc* peut localiser une centaine de sondes en deux ou trois mois (à condition que ces sondes soient des cosmides, ce qui tend à devenir le cas général). Son laboratoire place ainsi plus de mille sondes par an : des cosmides provenant de banques spécifiques d'un chromosome, des *linking clones* employés pour la cartographie en champs pulsés..., dans le cadre d'une multitude de projets. On imagine une armée de techniciens rivés à leurs microscopes et faisant de l'*in situ* à longueur de journée : il n'en est rien, et l'organisation du laboratoire est très instructive. En fait chaque *post-doc* (il y en a une douzaine) a son propre sujet, centré sur le clonage du gène d'une maladie pour laquelle on a déjà une localisation approximative (liaison génétique, ou translocation). Le travail commence alors par l'étude de quelques dizaines de sondes (en général des cosmides) provenant de la région ou, à défaut, d'une librairie spécifique du chromosome en question ; elles sont localisées par hybridation *in situ* sur métaphases et les mieux placées servent alors à une cartographie fine sur les noyaux en interphase. On restreint ainsi la zone où peut se trouver le gène à quelques centaines de kilobases, dans lesquelles l'on dispose maintenant de plusieurs sondes ; l'approche est d'ailleurs considérablement accélérée s'il existe une translocation dans la région, ce que l'on recherche aussi par *in situ* sur des noyaux en interphase. Le *post-doc* est alors prêt à passer aux YAC et aux ADNc pour l'étape suivante de sa recherche. Avec ce système d'une redoutable efficacité, chacun fait quelques mois d'*in situ* à haute dose, plaçant une centaine de sondes dans le cadre de son projet. Les critères de validité sont clairement définis et un dispositif de vérification a été élaboré pour confirmer les localisations avant de les inscrire dans la base de données du laboratoire qui « produit » ainsi de très nombreuses localisations alors qu'il ne comprend qu'un seul technicien. Bref, une manière de travailler originale, très efficace, semblant résoudre la contradiction qui parasite beaucoup de projets Génome. Cette solution n'est possible que parce que les méthodes d'hybridation *in situ* pratiquées dans ce laboratoire sont en avance sur ce qui se fait presque partout ailleurs : cet avantage technique joue un rôle déterminant. Si les *post docs* de Dave Ward mettaient une ou deux années pour localiser leurs cent sondes, ou si tout le monde pouvait localiser cent cosmides en trois mois, cette organisation astucieuse ne tiendrait plus. L'élégante solution que nous venons d'explorer n'est donc malheureusement pas généralisable.

L'automatisation, une porte de sortie ?

L'automatisation ne pourrait-elle résoudre cette contradiction ? À l'ère des microprocesseurs et de la robotique, quoi de plus simple en apparence que de remplacer le personnel souvent surqualifié qui effectue les tâches manuelles de la recherche génomique par des automates précis, rigoureux, fiables et surtout infatigables ? Ces systèmes encore très peu répandus chez nous, je pensais à les rencontrer en grand nombre au cours de mon enquête. Aux États-Unis notamment, berceau du « Programme Génome », j'espérais beaucoup des centres du DOE. Cet organisme *(Department of Energy)*, qui avait pris l'initiative du programme au milieu des années 1980, a une tradition technologique très marquée, liée tout naturellement à son implication dans des programmes *high-tech* comme la bombe atomique et, plus récemment, la « guerre des étoiles » : on pouvait donc

s'attendre à y voir un développement massif des approches instrumentales de la biologie. Ce n'est, en fait, pas vraiment le cas. Il y a certes dans ces laboratoires (Lawrence Livermore et Lawrence Berkeley en Californie, Los Alamos au Nouveau-Mexique) plus de technologie qu'ailleurs, mais le travail à la main continue à prédominer. On y rencontre tout juste un ou deux robots Beckman *Biomek*, ce robot de pipetage qui commence à être employé pour certaines opérations de Biologie moléculaire : duplication de cultures stockées en plaques à microtitration, confection de filtres à partir de ces mêmes microplaques, réactions de séquençage... Les équipes de robotique (car il y en a) travaillent sur la construction, à partir d'éléments du commerce, de robots du même type mais plus rapides ou plus précis, ou sur la mise au point de systèmes pouvant reconnaître et prélever des colonies sur une boîte de Pétri, ou encore sur des machines à PCR capables d'accepter un très grand nombre d'échantillons ; mais la pratique quotidienne des laboratoires reste pour l'essentiel celle du travail manuel. Relative déception donc, décidément la robotique a bien du mal à pénétrer dans les laboratoires de Biologie, même dans ceux qui devraient en principe y être le plus perméables... Quelles sont les raisons d'un tel retard ? Tout d'abord le conservatisme du milieu de la recherche biologique fondamentale, où l'on hésite souvent à investir le temps indispensable à la mise au point d'une nouvelle procédure (qui serait pourtant rentable dans le long terme). Une autre raison est le manque de formation en technologie et en physique de base de la plupart des biologistes, qui rend le dialogue avec un ingénieur en robotique bien difficile... Les équipes du DOE ont pourtant une culture technique très supérieure à la moyenne ; mais même dans ces centres le fossé entre biologistes et roboticiens, comme entre biologistes et informaticiens, est profond : les cultures diffèrent, la perception des impératifs aussi. Les spécialistes veulent réaliser un programme ou une machine élégants, parfaits à leur point de vue ; le biologiste, lui, cherche plutôt la rapidité de mise à disposition d'un outil même un peu approximatif, et la facilité d'emploi.

D'autres raisons, plus fondamentales, expliquent ce retard. La principale sans doute est que les techniques du Génie génétique sont encore en évolution rapide. Compte tenu des délais qu'implique la mise au point puis l'éventuelle commercialisation d'un automate spécialisé, le risque que celui-ci soit obsolète au moment de son introduction est réel. La montée en puissance des méthodes utilisant l'amplification enzymatique (PCR) en est un bon exemple : l'amplification court-circuite toute une série d'étapes et change complètement l'ordre de grandeur de la sensibilité nécessaire dans de nombreuses applications puisque la région à examiner peut maintenant être amplifiée spécifiquement un million de fois au préalable. Du coup le marché potentiel d'une machine à réaliser des *Southern blots* se trouve considérablement rétréci puisque la PCR permet souvent d'éviter le recours à cette technique. Robotiser, c'est ainsi faire un pari risqué, surtout pour un industriel ; c'est courir le danger d'aboutir à un appareil dépassé par l'évolution des techniques ou simplement trop cher pour les laboratoires de Biologie où une machine de 500 KF est encore considérée comme du gros matériel (l'échelle n'est pas la même que pour nos collègues physiciens...).

L'automatisation ne peut donc apporter qu'une solution très partielle à notre dilemme – ce qui ne veut pas dire qu'il faille la négliger, ni qu'elle soit employée aujourd'hui au maximum de ses possibilités... Mais sa mise en œuvre s'avère délicate, en raison d'obstacles techniques, financiers et psychologiques. On peut alors penser à contourner l'obstacle en opérant une division du travail : la part systématique et bien codifiée serait assurée par quelques « Centres de Service » organisés de manière industrielle, équipés de machines et bien dotés en personnel technique ; les laboratoires auraient recours à ces centres pour les phases correspondantes de leur travail, gardant la maîtrise de leurs projets et la responsabilité de la conception et de l'exécution des parties les plus nobles mais aussi les plus délicates de leurs programmes. Il s'agirait en somme de constituer des sortes de services communs pour le Génome.

Services communs : efficacité non garantie

Les services communs ont mauvaise presse en France, et il faut reconnaître que beaucoup d'entre eux, édifiés à grands frais et dotés d'un abondant personnel, se sont révélés remarquablement peu performants. On se souvient sans doute du Centre CNRS d'élevage d'animaux de laboratoire à Orléans, qui a pendant des années fourni à nos laboratoires des souris dont le prix de revient réel (salaires et amortissements compris) était très supérieur à celui que font payer des fournisseurs commerciaux qui ne sont pourtant pas, que l'on sache, des philanthropes. On pourrait facilement multiplier les exemples. D'autres services communs fonctionnent, eux, relativement bien, mais à l'avantage quasiment exclusif d'un laboratoire : ils lui sont pratiquement intégrés (ce qui est sans doute la raison de leur relative efficacité) et n'ont de commun que le nom : ils constituent en fait une façon déguisée d'augmenter le potentiel de l'équipe en question. Au moins peut-on apprécier que dans ce cas les deniers publics ne soient pas gaspillés, même s'ils n'ont pas exactement l'usage prévu. Mais globalement la situation est loin d'être satisfaisante, et la tendance depuis plusieurs années est à la fermeture de ces services, les tâches correspondantes étant prises en charge soit par le secteur privé (une bonne solution lorsqu'on peut faire jouer la concurrence) soit par des structures intégrées aux laboratoires et bénéficiant, on l'espère, de leur dynamisme. L'échec généralement reconnu des services communs en France s'explique aussi par certaines de nos tares nationales : notre répugnance à mettre en question les structures une fois qu'elles ont été créées, une culture économique très lacunaire dans le secteur public, et l'incontournable problème du statut du personnel... car naturellement tout le monde dans ces lieux est fonctionnaire Inserm ou Cnrs, ce qui empêche le recrutement de certains hommes clefs au niveau de qualification voulu (en raison de la modicité des salaires), gêne toute adaptation en profondeur et rend très malaisée la remise en cause de la structure.

Mais ces paramètres ne sont pas les seuls, et même dans des systèmes très différents de nos organismes publics, les « Centres de Ressources » et autres services communs se heurtent à quelques difficultés. Par exemple, l'évaluation quantitative et qualitative de la demande des futurs utilisateurs n'est pas évidente, et il ne faut pas faire trop confiance aux prévisions qu'ils expriment et qui se révèlent souvent dramatiquement fausses. Les usagers savent aussi se montrer parfaitement déraisonnables, surtout si cela ne leur coûte rien : David Schlessiger, dont le laboratoire effectua généreusement (et gratuitement) une série de criblages par hybridation de sa banque YAC en 1988/1989 pour des laboratoires extérieurs en parle encore avec amertume. Il ne compte plus les cas où les sondes reçues s'avéraient inutilisables, et se souvient de ce laboratoire rappelant un an après avoir reçu les précieux clones YAC tirés (au prix de beaucoup d'efforts) de la banque, pour les réclamer de nouveau : ils n'en avaient rien fait dans l'intervalle et les avaient perdus, ou laissé se dessécher dans leur réfrigérateur... Si le groupe de Saint Louis s'est rapidement converti au criblage par PCR c'est sans doute parce que cette méthode est plus fiable - encore que Hans Lehrach (ICRF, Londres) soit d'un autre avis -, mais surtout parce qu'elle exige du laboratoire demandeur un certain effort : séquencer sa sonde, définir et synthétiser des oligonucléotides, vérifier qu'ils donnent une bande nette et unique par amplification sur de l'ADN génomique... Si le demandeur a fait ce travail, on peut supposer qu'il a réellement besoin du clone YAC qui sera éventuellement obtenu, ce qui est rassurant pour ceux qui se livreront au lourd travail du criblage. Un autre moyen de filtrer les demandes frivoles est de faire payer à leurs auteurs une somme raisonnable, mesure qui est souvent plus utile par son effet dissuasif que par la rentrée financière qu'elle occasionne. Les centres de service ont aussi du mal à obtenir des utilisateurs des informations, un *feedback* comme on dit en franglais, sur le devenir des objets qui leur ont été fournis : qualité des librairies, caractéristiques des clones... ces informations reviennent rarement au « fournisseur », auquel elles seraient pourtant précieuses, ne serait-ce que pour mieux connaître la qualité de ses produits. De ce point

de vue, le système des bibliothèques de référence de Hans Lehrach a l'avantage d'imposer un minimum de retour d'informations puisque c'est à partir des coordonnées sur le filtre du ou des clones positifs en hybridation que pourront être fournis ces derniers. En tout état de cause, il est important pour un centre de service d'avoir un cahier des charges précis et d'indiquer très nettement quels services il peut rendre, et dans quelles conditions ; sinon il devient vite parasité par des circuits parallèles. Les modes d'emploi du *HGMP Resource Centre*, largement diffusés dans le petit journal *G-nome news* représentent sans nul doute un exemple à suivre. Le centre de service doit aussi avoir une politique claire au niveau des signatures d'articles : dans quels cas se comporte-t-il en prestataire de services, quand y a-t-il par contre collaboration scientifique avec en corollaire participation aux articles ? Là aussi les règles doivent être explicites.

La situation aux États-Unis mérite aussi quelques développements. Il n'existe pas à ma connaissance dans ce pays de centres de service au sens strict, peut-être parce qu'ils sont moins nécessaires qu'en Europe en raison de la relative abondance des financements. En fait, chaque Centre d'études sur le Génome un peu important s'est rendu autonome en obtenant, par exemple, une copie de la banque YAC de Saint Louis et même d'une ou deux banques européennes. Il n'y a donc pas de système de criblage à façon (par PCR) comme celui offert en Grande-Bretagne ou, en France, par le CEPH, ce qui handicape les laboratoires qui n'ont pas de relations avec un *Genome center*. Les contrats Génome prévoient explicitement le soutien de services, souvent appelés *cores* ; ils font même l'objet au NIH d'un contrat spécial distinct des contrats de recherche. Mais ces *core services* m'ont paru en général très centrés sur le service... local, comme par exemple dans le Département de Génétique de Baylor (Tom Caskey, Houston, Texas, USA) où ils semblent surtout soutenir les (excellentes) équipes de Génétique humaine du Département. On trouve à leur tête des chercheurs qui ont fait le choix d'un travail plus routinier, moins « glorieux » que les autres équipes, mais avec par contre une situation relativement stable et un répit dans l'épuisante chasse au contrat. Rappelons en effet que ce département, comme beaucoup aux États-Unis, offre aux équipes de recherche le gîte, mais pas le couvert, qu'elles dépendent de leurs contrats pour la quasi-totalité de leurs fonds de recherche, pour leur personnel et même pour la majeure partie du salaire de leur responsable...

Comment attirer des « seniors » ?

C'est une des questions les plus difficiles à résoudre pour les centres de ressource. Les responsables du *HGMP Resource Centre* de Harrow, dans la banlieue de Londres, le centre de service principal du Programme Génome britannique la considèrent comme leur problème principal. Le *Resource Centre*, déjà mentionné à propos du programme Génome britannique [1], est un « vrai » centre de service : entendez qu'il se préoccupe essentiellement d'apporter à ses utilisateurs britanniques (et européens, on ne sait pas assez qu'il est ouvert à tout laboratoire de notre continent ou du moins de la CEE) les services qu'ils demandent. Cela comprend le criblage de banques YAC (celles de Saint Louis et de Rakesh Anand), la fourniture de sondes et d'amorces PCR, bientôt la localisation de sondes chez la souris dans le cadre du *backcross* européen, des banques d'ADNc, l'accès aux bases de données et une aide à l'emploi de systèmes informatiques... Structure de taille modérée, créée de façon très pragmatique et avec des moyens raisonnables, elle souffre de ne pouvoir attirer quelques biologistes de haut niveau. Son personnel, formé de jeunes au profil mi-chercheur mi-ingénieur avec un niveau proche du PhD, a une qualification suffisante pour effectuer les travaux courants ; mais quand quelque chose ne va pas, quand une méthode ne « marche » plus sans raison apparente ou que la mise en place d'une nouvelle technique ne réussit pas, l'absence de chercheurs chevronnés capables de faire le *troubleshooting* nécessaire se fait cruellement sentir. Jusqu'à maintenant les efforts faits pour en attirer ont échoué à cause du manque d'intérêt

de ce travail de routine pour un bon chercheur ; l'environnement peu attrayant (Harrow est une banlieue très lointaine, et il y a peu de recherche sur place : le *Resource Centre* est installé dans les locaux d'une unité du MRC fermée pour qualité scientifique insuffisante) joue sans doute aussi son rôle.

La situation actuelle du « Généthon » créé à grands frais par l'AFM à Évry avec l'expertise du CEPH montre comment les choses peuvent évoluer lorsqu'on tente de résoudre cette question en injectant de la recherche dans une structure de service. Lors de sa création en 1990 le Généthon avait en effet été présenté comme une structure où seraient rassemblés équipement lourd et personnel technique dont bénéficieraient les équipes de recherche françaises à certaines phases critiques de leurs travaux. Aux étapes particulièrement laborieuses impliquant la réalisation de centaines de *Southern blots*, un séquençage massif ou encore l'alignements de nombreux clones, ces équipes pourraient venir utiliser pendant une période limitée les équipements et le savoir-faire rassemblés à Évry avant de repartir poursuivre leurs travaux « chez elles ». Le démarrage de cet ensemble a, en fait, vite fait apparaître la pénurie de cadres ; et pour en attirer, au moins à temps partiel, dans ce bâtiment très bien équipé mais un peu perdu dans une banlieue lointaine la décision a été prise d'y implanter des projets de recherche. Plusieurs d'entre eux sont maintenant opérationnels ; le plus avancé est celui que dirige Jean Weissenbach et qui vise à définir un très grand nombre de microsatellites hautement polymorphiques afin d'apporter une contribution décisive à l'affinement de la carte génétique humaine. Parmi les autres thèmes, on doit citer celui de Daniel Cohen sur l'« extraction » de banques YAC spécifiques à partir de la banque générale du CEPH et la construction de *contigs* le long du chromosome 21, et bien sûr le projet « genexpress » de Charles Auffray (en principe cofinancé par le CNRS) qui a pour objectif un séquencage massif et partiel de clones d'ADNc selon un schéma assez analogue à celui popularisé par Craig Venter. Trois études (il y en a d'autres) innovantes, et qui ont de bonnes chances de succès ; mais ce sont des sujets de recherche, même s'ils doivent à terme déboucher sur des réactifs d'intérêt général, et l'aspect « service » primordial dans le schéma initial est un peu passé au second plan. Au lieu d'être une structure originale qui allait permettre à de nombreux laboratoires d'accéder, au moment venu, aux techniques lourdes et ainsi irriguer l'ensemble des travaux sur le Génome en France, le Généthon est devenu une sorte de deuxième CEPH, effectuant avec de très gros moyens (chacun des thèmes cités ci-dessus emploie une vingtaine de techniciens, sans parler des machines) une recherche de pointe. Il convient de tempérer cette appréciation en rappelant que Généthon a aussi des activités ouvertes sur l'extérieur, banques de cellules, séquence à façon (dans quelques cas), et confection de *Southern blots* pour différents projets ; mais même l'atelier des automates à *blots* tourne à plus de 50 % pour les besoins « internes » (CEPH et de recherches menées sur place).

Rapport qualité/prix : la bouteille à l'encre

Tout ce travail coûte cher, très cher. Les appareils valant moins de 0,5 ou 1 MF deviennent rares ; l'informatique, malgré la baisse constante des prix, absorbe des sommes élevées ; et la base de calcul du financement pour certaines structures proposées s'approche du MF par personne, somme qui inclut, il faut le dire, la plupart des salaires. On atteint là des sommets qui n'ont plus rien à voir avec les dépenses d'un laboratoire classique, estimées, pour des unités bien financées et faisant beaucoup de Biologie moléculaire et de culture de cellules, à deux ou trois fois moins. Certes une activité de service, fortement automatisée donc consommant de grandes quantités de réactifs, ouverte sur l'extérieur et « exportant » des objets onéreux à produire, une telle activité doit logiquement exiger un soutien plus lourd qu'une recherche académique classique. Mais l'évaluation précise devient très délicate, et n'est pas la moindre difficulté de ce secteur. On ne dispose pas ici de l'étalon commode du fournisseur de souris commercial qui permet-

tait d'évaluer facilement la viabilité (ou plutôt la non-viabilité) du CSEAL ; encore qu'il commence à y avoir des firmes qui font de la séquence à façon, et que si l'entreprise envisagée par le milliardaire Fréderick Bourke [2] se réalise effectivement elle pourrait fournir une référence bien utile. Mon impression, fondée sur la visite de nombreux centres et la communication de leurs budgets, est que le rapport qualité/prix est extraordinairement variable : selon la stratégie choisie, la pertinence des choix faits, la rigueur de leur mise en œuvre, la qualité de la gestion du centre... je ne serais pas étonné qu'il y ait des écarts de un à dix dans le prix de revient, par exemple, de la mégabase de contig de YAC. Évaluation malaisée, mais pourtant primordiale, car avec la nature intrinsèquement expansionniste des programmes Génome il est impératif de dépenser au mieux les sommes importantes qui leur sont allouées.

Changement d'échelle, changement de culture ?

Beaucoup des questions évoquées ci-dessus découlent des dimensions du programme Génome et de son caractère systématique. C'est en effet la première fois qu'est entrepris en Biologie un effort aussi structuré : la « guerre contre le cancer » lancée du temps de Richard Nixon, ou les recherches sur le SIDA aujourd'hui ont une ampleur comparable mais regroupent une multiplicité de travaux très divers, dont la plupart pris isolément peuvent s'inscrire dans un cadre « académique » classique. Au contraire, les tâches définies par ce projet sont bien codifiées : affiner la carte génétique jusqu'à une résolution de deux à quatre centimorgans, établir des cartes physiques de chromosomes entiers, séquencer de grandes régions d'ADN. Leur accomplissement dans de bonnes conditions impose la mise en place d'une organisation de type semi-industriel, assurant un suivi précis de l'avancement des projets et leur contrôle de qualité, ainsi qu'une gestion professionnelle sur le plan financier... toutes choses auxquelles l'on n'est pas trop habitué dans les laboratoires de recherche fondamentale. Pour autant le fonctionnement doit rester extrêmement souple et se prêter à la remise en cause des stratégies expérimentales (nous avons lourdement insisté sur le caractère évolutif des techniques), remise en cause qui doit s'étendre aux structures et aux hommes qui les animent. C'est un peu la quadrature du cercle... d'autant qu'il ne faut pas dans tout cela oublier la motivation de ceux qui exécutent ce travail. À cet égard, il faudra bien arriver à une meilleure reconnaissance des travaux systématiques, à attacher par exemple une certaine valeur à la production d'informations qui sont inscrites dans des bases de données, sans passer par une publication au sens habituel du terme. Une petite révolution culturelle, en quelque sorte...

Les programmes Génome constituent, on le voit, une sorte de banc d'essai d'une biologie différente, plus massive, plus instrumentale, plus industrielle ; et la sociologie de cette nouvelle façon de faire de la recherche n'est pas sans intérêt. Les questions qui sont posées nous amènent en effet à remettre en cause certains comportements hérités d'une ère individualiste et artisanale dont le champ est sans doute en train de se rétrécir.

Références

1. Jordan B. Grande-Bretagne : un programme Génome à dimension humaine. *médecine/sciences* 1992 ; 8 : 163-6.
2. Anderson C, Aldhous P. Genome project faces commercialization test. *Nature* 1992 ; 355 : 483-4.

Med Sci *1992 ; 8 : 720-723*

GÉNOME : LA CAVERNE D'ALI BABA... OU LE SUPPLICE DE TANTALE ?

Convaincu par mon tour du monde en 1991 que les programmes Génome avaient produit de multiples « ressources », engrangé de grandes quantités de résultats et développé de nouvelles méthodes de recherche, je cherchais par ce texte à inciter les laboratoires « ordinaires » à profiter de ces acquis. Il me paraissait en effet évident que cette jonction potentiellement très fructueuse tardait à se faire, alors même que l'accès aux informations, aux banques de données ou aux collections de clones était relativement aisé pour peu que l'on s'en donne la peine. Cette chronique est donc riche en indications pratiques et en adresses. Le lecteur notera qu'Internet commence à faire une timide apparition, et que le terme de « liaison à haute vitesse » recouvre ici une connexion à... 9 600 bauds. Rappelons qu'aujourd'hui le plus modeste des modems domestiques fonctionne à 56 K (56 000 octets/s) et qu'un branchement ADSL offre normalement un débit de 512 K.

Nombre de lecteurs de ces *Chroniques Génomiques* les parcourent avec attention (du moins je l'espère...), mais sans se sentir personnellement concernés : le monde du Génome est loin de leurs préoccupations quotidiennes, et si leur curiosité les porte à vouloir savoir ce qui s'y passe ils n'y voient pas de conséquence directe pour le déroulement quotidien de leurs travaux. Pourtant les recherches menées dans le cadre des programmes Génome, si elles impliquent souvent des grands laboratoires fortement équipés et des projets importants, produisent beaucoup d'outils et d'objets présentant un intérêt réel pour les équipes « normales ». Le but de cette chronique est de le rappeler à l'aide de quelques exemples, et de donner des indications propres à en faciliter l'accès. Les connaisseurs me pardonneront le manque d'informations nouvelles pour eux dans ce texte qui, à vrai dire, ne leur est pas destiné...

Les retombées des programmes Génome

Une équipe venant d'isoler un morceau du gène qui est l'objet de son étude a en général pour premier désir d'en disposer dans son intégralité, de préférence sous la forme d'un segment cloné contenant aussi les séquences de régulation qui gouvernent l'expression et qui peuvent parfois se trouver assez loin des séquences codantes. Plutôt que d'explorer à cet effet une banque génomique construite dans des phages ou des cosmides, il peut donc être avantageux d'isoler un chromosome artificiel de levure ou YAC *(yeast artificial chromosome)* : la taille probable du segment inséré (quelques centaines de kilobases) rend très vraisemblable l'obtention du gène complet accompagné des différents promoteurs, *enhancers*, *CAT boxes* et autres séquences pertinentes. Mais, on le sait [1], la construction de banques de YAC reste une entreprise très délicate et de longue haleine : il est hors de question de se lancer dans ce travail pour un besoin ponctuel. La réaction habituelle en ce cas pour un biologiste moléculaire est de demander un exemplaire de la banque au laboratoire qui l'a construite ; mais une banque YAC ne se transfère pas comme une librairie de phages ou de cosmides, il ne suffit pas d'envoyer

un tube contenant quelques centaines de milliers de clones que le laboratoire destinataire pourra utiliser à sa guise. Les banques de chromosomes artificiels de Levure se présentent sous forme d'un jeu de clones ordonnés dans des plaques à microtitration : une librairie représentative du génome humain (50 000 clones) occupe ainsi 500 plaques à 96 puits. La duplication et le transport d'un ensemble aussi volumineux ne sont évidemment concevables que dans un nombre très limité de situations[1]. Le chercheur pourrait, à défaut, envoyer sa sonde à un laboratoire central qui se chargerait du criblage : c'est ce qu'a fait le groupe de David Schlessinger (Saint Louis, MO, États-Unis) au début, mais il a vite croulé sous le nombre des sondes et reculé devant l'irresponsabilité des utilisateurs.

Depuis un ou deux ans l'accès à ces banques a été organisé, grâce en grande partie aux financements des programmes Génome. La modalité la plus courante consiste en un criblage par PCR à deux étapes. Dans un premier temps, le laboratoire demandeur séquence sa sonde et définit des oligonucléotides autorisant l'amplification d'un fragment spécifique. Il reçoit des détenteurs de la banque des échantillons d'ADN correspondant à des *pools*, c'est-à-dire contenant chacun l'ADN extrait d'un mélange de mille ou deux mille clones YAC, la banque entière étant ainsi représentée par quelques dizaines d'échantillons. Il effectue alors les réactions PCR sur chacun des *pools* afin d'identifier celui qui répond positivement à l'amplification (en donnant le fragment de la taille prévue) et qui doit donc contenir le clone positif recherché. Il envoie alors au laboratoire central cette information ainsi que les oligonucléotides permettant l'amplification afin que ce dernier effectue la deuxième étape du criblage. Cette procédure partage le travail et y implique le demandeur ; et la preuve est maintenant faite qu'elle constitue une façon viable de permettre à la communauté l'accès à une banque YAC. C'est en effet sous cette forme que sont exploitées plusieurs d'entre elles, en France dans le cadre du CEPH ou en Grande-Bretagne au *Resource Centre* du *Human Gene Mapping Programme* ; les financements nécessaires à la mise en place d'une telle organisation ont été fournis pour l'essentiel par des contrats « Génome » nationaux ou européens, et mettent – moyennant un certain travail ! – cet outil remarquable à la disposition de l'ensemble des laboratoires.

Comme je l'ai déjà mentionné, d'autres modalités sont envisageables, entre autres l'élégante solution des « librairies de référence » défendue par Hans Lehrach (*Imperial Cancer Research Fund*, Londres). Une banque de cosmides du chromosome X a ainsi été exploitée par de multiples équipes à l'aide de filtres à haute densité distribués par ce groupe ; pour les YAC le procédé n'est pas encore complètement opérationnel (en tant que système de distribution) ce qui est dommage car le criblage direct par hybridation présente beaucoup d'avantages pour l'utilisateur surtout si ce dernier recherche de nombreux clones... Mais en tout état de cause l'accès aux clones YAC est maintenant possible à tout laboratoire – à condition qu'il sache où s'adresser, et qu'il fasse preuve d'une certaine patience. Rappelons que les techniques d'analyse des YAC sont maintenant pour la plupart bien codifiées, à la portée de tout laboratoire de Biologie Moléculaire normalement constitué et que ces clones peuvent aussi être transférés dans des cellules et, sans doute bientôt, dans des animaux. Notons aussi que les YAC se prêtent tout à fait à la localisation chromosomique par hybridation *in situ* après marquage non radioactif : une simple amplification sur l'ADN total de la cellule de levure contenant le clone avec des amorces Alu (spécifiques de l'ADN humain) produit un mélange de sondes directement utilisable.

Autre cas de figure : une équipe vient d'isoler, chez la souris, une série de gènes apparentés par des homologies de séquence, qui pourrait représenter une nouvelle famille de récepteurs de neurotransmetteurs. Compte tenu des précédents il est conce-

1. On peut à cet égard signaler qu'il est possible d'acheter une banque YAC humaine (diffusée par la firme Clontech) ; mais le prix de cette bibliothèque sous sa forme développée (340 plaques à microtitration contenant chacune 96 clones) est en France de près de 100 000 F...

vable que ces gènes soient regroupés dans le génome, ce qui pourrait indiquer qu'ils sont issus d'une duplication récente ou qu'ils sont soumis à une régulation commune. Il est donc essentiel de savoir très vite – et à peu de frais si possible... – si cette hypothèse est tenable. L'hybridation *in situ* n'est pas la solution car les sondes dont on dispose (des ADNc de petite taille) ne donnent pas de résultats suffisamment rapides et fiables par cette technique. Mais il existe dans le cas de la souris une alternative : les systèmes de croisement interspécifiques qui, grâce à la divergence de séquence entre les ADN des deux souches, autorisent la localisation de toute sonde qui se révèle polymorphique dans ces conditions... pour autant qu'un travail préalable ait permis d'effectuer les croisements, de préparer les ADN des centaines de souris qui en résultent, et de repérer les points de recombinaison sur chacun des chromosomes à l'aide d'une étude avec un jeu de sondes de référence. Une entreprise de cette nature est en voie d'achèvement dans le cadre du *Backcross* européen qui associe les équipes de Jean-Louis Guenet à l'Institut Pasteur et le *HGMP Resource Centre* britannique ; et cela devrait aboutir avant la fin de cette année à un jeu de filtres sur lesquels il suffira d'hybrider une sonde quelconque (sous réserve qu'elle révèle un fragment différent sur les ADN des deux parents) pour déterminer son chromosome d'origine et sa position à une dizaine de centimorgans près. Il sera ainsi possible, pour reprendre l'exemple discuté, de savoir rapidement si les gènes en cause proviennent d'une même région chromosomique... sans se livrer à de nombreuses et délicates hybridations *in situ* ni se lancer dans une longue série d'analyses en champ pulsé.

À ce stade d'ailleurs peut apparaître un troisième type d'utilisation des « ressources » produites par les programmes Génome. Il s'agit cette fois de ces banques de données de plus en plus complètes qui emmagasinent et rendent accessibles les informations obtenues par les laboratoires spécialisés [2]. L'une d'elles, GBASE, est principalement consacrée à la souris, et contient un état à jour des correspondances (synténies) connues entre les chromosomes humains et murins. On sait que notre génome est étonnamment proche de celui de ce rongeur familier, que pratiquement tout gène humain existe sous une forme très proche chez la souris ; mais on constate aussi que la disposition de ces gènes le long des chromosomes est conservée sur des distances importantes. Le chromosome 11 de la souris, par exemple, contient une grande zone homologue au 17 humain, une autre correspondant au 5. Dans chacune de ces régions les gènes sont dans le même ordre chez les deux espèces. Ces synténies sont maintenant bien explorées ; de sorte que de la position d'un gène chez la souris on peut déduire presque à coup sûr sa localisation chez l'homme, surtout si l'on dispose des données les plus récentes. L'exploitation décrite ci-dessus du système de « *backcross* interspécifique » permettra donc d'obtenir non seulement l'information chez la souris, mais aussi une position très vraisemblable chez l'homme. On pourra alors se demander si des syndromes héréditaires pouvant être dus à un dysfonctionnement des récepteurs en question n'auraient pas par hasard été « localisés » (au sens de la localisation génétique, première phase de la « génétique inverse ») dans la région chromosomique humaine ainsi désignée. La consultation des données génétiques et cliniques contenues dans « OMIM » (*On-line Mendelian Inheritance in Man*, la version électronique du fameux atlas des maladies génétiques humaines de Victor McKusick) ou dans la base française Genatlas sera à même de répondre à cette interrogation... et peut-être d'indiquer ainsi une voie de recherche très féconde !

De la coupe aux lèvres...

C'est maintenant que l'on en vient au supplice de Tantale : les choses, en effet, ne sont pas aussi aisées. Pour employer les « ressources » (au sens anglo-saxon du mot) décrites ci-dessus, il faut déjà connaître leur existence : or, l'information ne circule pas toujours bien et elle n'est pas parfaitement fiable. Certains services restent parfois confi-

dentiels, et seuls les initiés peuvent, du coup, en bénéficier ; d'autres entreprises d'intérêt commun, au contraire, très largement annoncées (et quelquefois financées dans ce but par les programmes Génome nationaux ou européens), s'avèrent ne pas fonctionner concrètement. Il faut reconnaître que la technologie évolue très rapidement, et aussi que la mise en place d'une telle organisation se révèle en général nettement plus difficile que prévu... ce qui fait que les délais prévus sont rarement tenus. De plus, la tension entre recherche et service, discutée dans une précédente chronique [3] joue à plein ici ; elle aboutit parfois à des situations confuses où le scientifique « extérieur » ne sait pas ce à quoi il a droit et ce qui relève de la collaboration à négocier (avec signature d'articles à la clé...). Cela traduit un souci compréhensible de la part d'équipes qui doivent assurer à leurs membres les éléments d'une carrière ; mais on ne peut que souhaiter que les contrats « Génome » nationaux ou européens spécifient clairement les conditions dans lesquelles l'activité de service s'effectue, et que le cahier des charges soit net, précis... et public, même s'il doit être fréquemment révisé.

Tout ceci fait que de nombreux laboratoires (surtout ceux qui ne sont pas directement dans le « circuit » Génome) ne profitent pas de ces ressources considérables, alors même qu'elles pourraient considérablement accélérer leur travail.

Quelques pistes...

Nous resterons donc résolument pratiques en indiquant pour finir quelques pistes pour ceux qui voudraient, eux aussi, faire usage de ces possibilités. L'information est, on l'a compris, une denrée précieuse. Elle est abondante, mais inégalement diffusée, et n'atteint pas obligatoirement ses destinataires.

Une première mesure très simple consiste à faire en sorte de recevoir quelques petits journaux qui rassemblent une foule d'informations utiles sur les programmes Génome et sur les services qu'ils proposent. Il s'agit notamment de *Human Genome News*, le journal du Programme Génome des États-Unis (HGMIS, *Oak Ridge National Laboratory*, PO Box 2008, Oak Ridge, TN 37831-6050, USA) et de *G-nome News*, son homologue britannique (*HGMP Resource Centre*, CRC, Watford Rd, Harrow, HA1 3UJ, UK). L'abonnement à ces deux gazettes sans prétention est gratuit, et elles renferment toutes sortes de renseignements : noms (et numéros de télécopieur !) des responsables, ressources offertes, nouvelles des *Genome Centers*, annonce de colloques spécialisés...

Plus évoluées, plus complètes, mais aussi plus techniques sont les bases de données (voir pour plus de précisions le numéro « spécial séquences » de la revue *Nucleic Acids Research* [4]). Les principales en ce qui nous concerne sont GDB *(Genome Data Base)* et Genatlas pour les données génomiques hors séquence, GBASE pour la souris et les synténies, et OMIM pour les maladies génétiques. Compte tenu de la masse d'informations produites par la recherche génomique, dont la plupart ne seront jamais publiées au sens classique du mot, elles se révèlent de plus en plus indispensables. Le problème est qu'il ne s'agit pas de bases de données « simples » comme celles concernant les séquences d'ADN, qui peuvent être distribuées sur disquette ou sur disque optique. Ces dernières ont en effet une structure peu complexe (malgré le volume de leur contenu) car elles contiennent essentiellement des objets de même espèce (les séquences) pourvus de quelques annotations et exploités par un système informatique qui n'a pas à gérer de relations compliquées entre ces éléments. Au contraire, une base génomique comme *Genome Data Base* (ou Genatlas) renferme des données de nature très diverse (clones, sondes, polymorphismes, maladies, cartes génétiques, cartes physiques...) dont les interrelations sont multiples. Il faut donc les emmagasiner dans un système informatique à structure relationnelle, qui gère les relations entre les objets autant que les objets eux-mêmes. De tels systèmes ne sont pas « portables » sur le micro-ordinateur de chaque chercheur : il faudrait importer non seulement les informations, mais aussi le programme

de gestion qui les manipule et ne « tourne » que sur station de travail de type Sun. Bref, dans la pratique il faut y accéder par voie de réseau, et de réseau à haute vitesse (au moins 9 600 bauds) sinon elles sont quasiment inexploitables.

Comment faire en pratique ? Il est facile d'être enregistré comme utilisateur par ces organisations et de disposer d'un mot de passe : il suffit pour cela de remplir un formulaire (l'adresse de GDB et d'OMIM est donnée dans la référence [4], celle de GBASE est : *The Jackson Laboratory*, Bar Harbor, Maine 04609, USA, et Genatlas est bien connu des lecteurs de *médecine/sciences*, ne serait ce que par les cartes qui paraissent dans les dernières pages de chaque numéro). Mais pour employer réellement ces banques il faut pouvoir se raccorder à un réseau à grande vitesse (Internet, de préférence par une liaison par fibre optique), disposer d'un terminal (un Macintosh ou un PC peuvent suffire, mais seule une station de travail pourra exploiter pleinement l'interface graphique), et savoir s'y retrouver dans ces ensembles où la navigation est parfois complexe. Là encore, il a été créé des structures destinées à faciliter l'entrée dans ces systèmes informatiques, à diminuer les coûts de connexion et à proposer des formations spécialisées aux utilisateurs : le HGMP britannique offre un tel service, et dans le cadre d'un contrat européen le DKFZ d'Heidelberg fait de même (*European Data Resource for Human Genome Analysis*, DKFZ, Im Neuenheimer Feld 280, D6900 Heidelberg, Allemagne). On peut souhaiter que dans notre pays aussi des relais de ce genre soient établis, au niveau national ou régional ; en attendant l'accès à ces données est possible mais exige une bonne dose de patience et un environnement informatique solide. C'est d'ailleurs un domaine dans lequel les organismes de recherche devraient faire un sérieux effort, car il en va de notre compétitivité...

Sur un plan plus matériel il faut citer les nombreuses banques de cellules, de clones ou de librairies dont les services sont soit gratuits, soit peu onéreux. Tout le monde connaît l'ATCC *(American Type Culture Collection)*, mais certains ignorent qu'elle distribue aussi des banques spécifiques de chromosomes ou d'ADNc, y compris une des premières banques « égalisées », celle qui émane du laboratoire de Sherman Weissman à Yale (New Haven, USA). Une autre structure, le *Corriell Cell Repository* (401 Haddon Ave, Camden, NJ 08103, USA), diffuse un précieux jeu d'hybrides monochromosomiques : vingt-quatre lignées de cellules de hamster contenant chacune un seul chromosome humain. Plus près de nous, des banques de sondes assez complètes sont gérées par le *Resource Centre* du HGMP britannique ; notons au passage que ce centre (dont une partie du financement est européen) est ouvert à tout chercheur de la CEE. En France, le CEPH et surtout le Généthon (1, rue de l'Internationale, BP 59, 91002 Évry Cedex – Tél : 01 69 47 28 00) assurent eux aussi certains services : criblage de la banque de YAC bien sûr, mais aussi établissement de lignées lymphoblastoïdes, extractions d'ADN et préparation de *Southern blots*. Les prestations ne sont pas gratuites, mais les prix sont modérés et très inférieurs à ceux du secteur commercial : moins de 300 F pour établir une lignée lymphoblastoïde, trois à quatre cents francs pour un *Southern blot* de 18 échantillons, y compris les digestions enzymatiques. C'est donc une aide précieuse, d'autant que ces montants peuvent être inclus dans une demande de financement déposée auprès de l'Association Française contre les Myopathies...

En guise d'envoi

Cette chronique résolument concrète aura peut être déçu certains de mes lecteurs habituels parmi les spécialistes : ils n'y auront pas appris grand-chose, ils auront peut-être même relevé quelques approximations. Mais si elle a pu convaincre d'autres chercheurs d'aller voir de plus près en quoi les ressources des programmes Génome pourraient leur être profitables, si elle leur a donné quelques pistes pour commencer à y accéder... elle aura rempli son objectif !

Références

1. Jordan BR. La montée en puissance des YAC. *médecine/sciences* 1990 ; 6 : 470-2.
2. Jordan BR. Génome et informatique : condamnés à s'entendre ? *médecine/sciences* 1991 ; 7 : 726-8.
3. Jordan, BR. Les contradictions du génome. *médecine/sciences* 1992 ; 8 : 476-82.
4. Sequence supplement. *Nucleic Acids Res* 1991 ; 19 : 2221-49.

Med Sci *1994 ; 10 : 92-95*

LES HGM SE SUIVENT... ET NE SE RESSEMBLENT PAS

Écrit plus circonstanciel puisqu'il rapporte un des ateliers bisannuels appelés Human Gene Mapping Workshops – *ateliers qui ont constitué durant une dizaine d'années l'outil principal d'échange et de collationnement des informations acquises sur le génome humain et la génétique médicale. Je me souviens du premier auquel j'avais assisté, tenu à Helsinki en 1985 (HGM 8) : après trois jours de discussions animées, les volumineux documents de consensus élaborés par les différents groupes de travail (un par chromosome) avaient été tirés à deux cents exemplaires grâce à une monstrueuse photocopieuse, afin que chaque congressiste reparte avec un état des lieux de la génétique humaine. L'édition 1994 de HGM décrite ici marquait la renonciation à ce mode de traitement des informations, rendu caduc par la multiplication des acteurs, le volume des données à traiter et la montée en puissance de l'informatique. On sent en filigrane dans ce texte la nostalgie d'un monde de la génétique convivial et à taille humaine, monde en voie de disparition...*

Le dernier *Human Gene Mapping Workshop*, HGM 11, avait eu lieu à Londres en août 1991 ; à la mi-novembre 1993, HGM 93 se tenait à Kobe (Japon). HGM 93 juste après HGM 11 : ce saut soudain dans les numéros d'ordre n'est pas le fait d'une machine a voyager dans le temps. Il manifeste l'évolution d'une formule qui, après avoir eu son heure de gloire, était devenue obsolète.

Les HGM, « ateliers » biennaux depuis 1973, avaient pour but de faire le point sur l'avancement de la carte des gènes humains, chromosome par chromosome, et d'établir un document résumant les conclusions qui résultent d'un consensus – éventuellement provisoire. Ils comportaient quelques conférences et de nombreux *posters* présentant les récents travaux des participants ; mais l'essentiel se passait dans de petites salles où les chercheurs se réunissaient pour débattre des derniers résultats obtenus sur chaque chromosome. Chacun de ces mini-conclaves rassemblait quelques dizaines de cartographes du génome, sous la houlette d'un ou deux présidents, experts du chromosome en question et chargés de veiller au bon déroulement des échanges ainsi que d'en rédiger les conclusions. On a bataillé ferme, dans ces assemblées, pour savoir s'il fallait enregistrer la localisation du gène de la myopathie de Duchenne (avant qu'il ne soit cloné) entre les repères DXS41 et DXS84, bien que les données fussent un peu préliminaires et qu'une autre équipe indiquât une position légèrement différente. On y parlait aussi de nomenclature, afin d'éviter de répertorier deux fois le même gène sous des noms différents, ou au contraire de donner la même appellation à des entités distinctes ; on se demandait si les sites fragiles ou les points de translocation devaient être catalogués au même titre que les gènes... Après trois ou quatre jours de ce régime, le responsable de l'atelier suivant était élu par l'assemblée ; les présidents des comités travaillaient très tard le soir à la rédaction d'un document final de plusieurs centaines de pages qui, au prix de prodiges d'organisation, était reproduit à des centaines d'exemplaires durant la nuit et remis à chaque congresslste avant son départ.

Ce système convivial et sympathique ne devait pas résister à l'avalanche de résultats qui apparut dès la deuxième moitié de la décennie 1980, pour s'accélérer encore avec le démarrage effectif des programmes d'étude du génome humain. La création de bases de données susceptibles d'emmagasiner toutes ces informations s'avérait cruciale : la *Human Gene Mapping Library* de Yale, Genatlas à Paris en 1987, *Genome Data Base* de Baltimore à partir de 1989 s'y attelaient avec des succès divers. Mais l'enregistrement de l'ensemble des données au cours d'un atelier de quelques jours devenait acrobatique, on atteignait la limite des possibilités matérielles et surtout financières. Rassembler mille personnes, c'est presque banal s'il suffit de leur assurer gîte et couvert et de leur faire écouter, trois jours durant, quelques dizaines d'orateurs. Mais, pour *Human Gene Mapping*, il fallait de surcroît prévoir la réunion de plus de vingt comités spécialisés dans autant de salles ; il fallait surtout que chacun d'eux ait à sa disposition, en permanence, tous les moyens informatiques nécessaires pour collationner, vérifier et enregistrer dans GDB les résultats obtenus au cours des deux années précédentes. HGM 11, dernier de la série, devait ainsi coûter très cher : plus d'un million de livres sterling, soit dix millions de nos francs, pour la seule informatique sans compter les voyages et les séjours des conférenciers, présidents de comités et congressistes. Aussi fut-il décidé de tenir à l'avenir des colloques spécialisés portant chacun sur un seul chromosome, réunissant beaucoup moins de participants, et plus faciles à mettre sur pied. Décision qui devait provoquer quelques remous : les chercheurs étaient fort attachés à ce rassemblement général, occasion de prendre connaissance de l'ensemble des travaux et de revoir les collègues. Et puis tout le monde n'est pas spécialisé dans l'étude d'un seul chromosome ; de nombreux thèmes de recherche sont transversaux : leurs protagonistes doivent-ils alors assister à une quinzaine de colloques dans l'année pour se tenir au courant ? Confrontés à une grogne inattendue, les états-majors effectuèrent un repli stratégique, et l'on termina sur la notion, somme toute raisonnable, de réunions par chromosome pour l'enregistrement des données, doublées d'un grand congrès biennal permettant à tous de se rencontrer et de faire le point sans introduction de données sur les systèmes informatiques, donc sans les complications et les frais encourus à Londres.

C'est au Japon, qui depuis plusieurs années souhaitait en être le théâtre, qu'a eu lieu cette HGM « nouvelle formule » précédée d'une séance de coordination (CCM, *Chromosome Coordinating Committee*) rassemblant les présidents des réunions par chromosome. Moins couru que les précédents (environ cinq cents inscrits), hésitant un peu entre l'atelier et le symposium classique, ce colloque ne manquait pourtant pas d'intérêt à bien des égards. Comme on pouvait s'y attendre dans ce pays où règne le souci du détail, l'organisation matérielle était impeccable, l'équipement audiovisuel parfaitement adapté et le personnel nombreux et efficace, quoique pas toujours très anglophone. La participation japonaise était naturellement massive (une bonne moitié des assistants), ce qui donnait l'occasion de rencontrer de jeunes chercheurs locaux rarement présents dans les congrès internationaux. Le programme Génome, qui m'avait semblé prendre un réel départ au Japon il y a deux ans [1], confirme son décollage. Le montant des financements en 1993 est de 3,6 milliards de Yen, soit environ 180 millions de nos francs, et une augmentation de l'ordre de 20 % est prévue pour 1994. Rappelons que notre ministère de la Recherche attribue une soixantaine de MF au GREG (plus vingt-huit au CEPH), et qu'aux États-Unis la somme totale (NIH plus DOE) tend vers les deux cent millions de dollars. La liste des projets financés par le Monbusho (ministère de l'Éducation) donne une idée des lignes de force : une vingtaine portent sur la cartographie physique et le clonage positionnel, une dizaine sur l'ADNc, quinze sur des mises au point technologiques... et trente-deux sur la bioinformatique appliquée au génome. Des structures spécialisées sont en bonne voie de réalisation : mentionnons l'Institut de séquençage de Kazuza (à Chiba, près de Tokyo), financé par le gouvernement local et dont la construction devrait s'achever au printemps 1994, ou le *Human Genome Center* de l'Université de Tokyo (cartographie physique, bases de données, développements d'outils d'analyse de

séquences). Le rôle de Kenichi Matsubara, vice-président de Hugo, responsable du programme du ministère de l'Éducation et organisateur de EIGM93 semble bien établi, ce qui est de bon augure car il s'agit d'un scientifique particulièrement compétent, étranger aux coteries et d'une intégrité sans faille.

C'est d'ailleurs son équipe qui présentait certains des résultats les plus intéressants du congrès, du moins à mon sens. Il s'agit en fait de la poursuite du projet portant su l'ADN complémentaire dont j'avais vu les débuts en 1991 [1] et dont le premiers résultats ont paru l'an dernier [2, 3]. L'objectif est d'établir une carte transcriptionnelle de deux cents cellules de base du corps humain, une *body map* selon la terminologie choisie par les auteurs. Le moyen : pour chaque cellule, construire une banque d'ADNc calibrée de façon à représenter fidèlement la population des ARN messagers, et séquencer deux ou trois cents nucléotides à l'extrémité 3' de mille clones pris au hasard. La première publication du groupe [2] avait montré toutes les informations que pouvait fournir cet exercice *a priori* un peu répétitif. Il avait en effet défini une brassée de gènes « nouveaux » (jusque-là inconnus), et, surtout, donné une évaluation du taux d'expression de centaines de gènes connus. La détermination de la multiplicité des clones correspondant à des séquences déjà répertoriées donnait pour la première fois une vision d'ensemble de l'activité transcriptionnelle d'une cellule avec quelques surprises, comme par exemple le fait que le gène le plus exprimé soit celui du facteur d'élongation l alpha, retrouvé vingt-deux fois parmi les mille clones... Le but annoncé d'établir la même liste pour les 199 autres types cellulaires pouvait paraître irréaliste. Cela ne semble pas être le cas, puisqu'à Kobe le groupe présentait l'analyse des données obtenues sur vingt types cellulaires : il est devenu le centre d'une nébuleuse incluant une dizaine de laboratoires (et un industriel, Hitachi), qui appliquent cette méthode d'analyse et mettent en commun les résultats obtenus pour les différentes cellules.

Leur confrontation, grâce à une base de données qui facilite les recoupements, se révèle très féconde. On peut, par exemple, effectuer par voie informatique une sorte de « *Northern blot* synthétique » dessiné à partir de la fréquence à laquelle se retrouve une séquence dans chacune des vingt cellules étudiées ; ou, à l'aide des premières assignations chromosomiques effectuées sur deux ou trois centaines de clones, constater que ce sont les gènes à expression forte et ubiquitaire qui ont le plus souvent deux ou trois localisations chromosomiques (correspondant sans doute à des pseudogènes), au contraire des gènes spécifiques d'un tissu.

Du côté des Occidentaux, peu de nouveautés (et beaucoup d'absents, comme le programme francais Genexpress) ; seul l'informaticien de l'*Institute for Genomic Research*, le nouveau laboratoire de Craig Venter, exposait son *Expressed Gene Anatomy Database* visant à regrouper et comparer les séquences partielles obtenues par différentes équipes. Chose intéressante, l'orateur estimait le nombre total de séquences d'EST actuellement déterminées à environ 130 000, chiffre supérieur au nombre total des gènes humains (50 000 a 100 000). Même si l'on tient compte des inévitables redondances (intra- et interlaboratoires), ce chiffre indique que la moitié sans doute des gènes humains a fait l'objet d'un étiquetage – chiffre imposant bien que beaucoup de ces séquences ne soient pas encore dans les banques de données. D'une façon générale, les progrès de ces programmes montrent que l'exploration par séquencage partiel présente, dans l'état actuel des techniques et lorsqu'il est mis en œuvre de manière rigoureuse, un excellent rapport qualité/prix.

Sur le front de la technologie, le bilan japonais est plus nuancé. L'usine à séquencer de Tsukuba, cet ensemble de robots destiné à automatiser presque complètement le séquençage de l'ADN [3, 4], devait entrer en fonction courant 1992 : elle n'a, en réalité, pas vraiment démarré. Un *poster* décrivait la séquence du chromosome VI de la levure, 280 kilobases, effectuée a Tsukuba... sans passer par l'usine située dans le même laboratoire, et à l'aide de séquenceurs Applied Biosystems ; un peu plus loin, un autre *poster*

montrait, lui, quelques photos de cette installation et indiquait, avec une humilité rare, « *the performance is very poor* ». Cette tentative, qui m'avait fait forte impression en son temps, semble donc se terminer peu glorieusement. Était-elle prématurée, ou s'est-elle fourvoyée en raison d'erreurs de parcours ? En tout cas, cet échec atteste que la persistance japonaise, cette capacité souvent payante à investir dans le long terme, peut aussi avoir des effets négatifs : difficulté à renoncer à un projet, à reconnaître qu'il est raté. Le programme d'ordinateurs de « cinquième génération », qui avait tant effrayé l'Occident lors de son lancement, en est un autre exemple : malgré son fiasco (il produit des machines en principe très puissantes, mais en pratique inutilisables, dont certaines ont été fournies gratuitement à des laboratoires qui, après quelques essais, on renonce à s'en servir), il fut prolongé d'une ou deux années au-delà du terme prévu. Toujours à propos de séquençage, Shimadzu présentait un appareil proche des instruments occidentaux ; et surtout Hitachi, qui vend depuis trois ans au Japon un séquenceur analogue à l'« ALF » de Pharmacia, montrait les résultats obtenus avec un prototype utilisant l'électrophorèse sur capillaire. Technologie déjà largement discutée depuis deux ou trois ans, celle-ci promet une augmentation considérable du débit grâce à la vitesse de séparation, à la résolution qui doit permettre de lire en routine 1 000 bases (au lieu de 400 à 500 actuellement) et à la possibilité de coupler de nombreux capillaires, 96 dans le cas de la machine Hitachi, et d'effectuer ainsi autant de séquences en parallèle. La firme semble avoir résolu un des principaux problèmes techniques, celui de la détection des signaux de fluorescence avec une très haute sensibilité ; et il est instructif de voir qu'après s'être contentée, dans un premier temps, de la copie du système inventé par Wilhelm Amsorge, elle se trouve maintenant – pour autant que je puisse en juger – dans le peloton de tête de ceux qui mettent au point la prochaine génération de séquenceurs.

Quelques remarques pour conclure cette vision forcément très impressionniste d'un congrès dense et dont beaucoup de sessions se tenaient en parallèle. Le front de la bioinformatique continue à bouger, et la saga des bases de données « officielles » n'est pas terminée : Peter Pearson n'est plus directeur de *Genome Data Base*, à qui l'on reproche toujours son manque de convivialité. Le nouveau responsable de l'informatique, Ken Fasman, promet de grandes améliorations, de nouveaux outils graphiques et la prise en compte des données préliminaires. Quant à Pearson, il est maintenant charge de la restructuration d'OMIM *(On line Mendelian Inheritance in Man)*, la base de données sur les maladies génétiques tenu depuis trente ans par Victor McKusick, afin d'en faire une vraie base relationnelle : espérons qu'elle ne perdra pas dans le processus la relative facilité de consultation qui la caractérise aujourd'hui ! L'effort japonais déjà mentionné plus haut est considérable, et certaines des démonstrations présentées étaient tout à fait convaincantes. Elles témoignent, comme le système GNOME proposé par une fondation privée et qui offre une interface simple et uniforme pour toute recherche d'homologie de séquence, d'un louable souci de se mettre à l'écoute des biologistes et de développer le outils dont ils ressentent le besoin plutôt que de vouloir à tout prix faire de la « vraie » recherche informatique en perdant un peu de vue ses applications. On aimerait que cette attitude soit plus répandue, singulièrement dans notre pays... Notons encore la brillante prestation de Yusuke Nakamura, dont la réputation n'est plus à faire dans le domaine de la cancérologie : résultats importants, appuyés sur des données impeccables. Il s'agissait cette fois d'une étude très complète sur les mutations dans le gène *APC* (impliqué dans la polypose colique, et cloné par cette équipe), illustrant de façon parfaitement claire le mécanisme génétique en deux étapes de ce cancer : prédisposition héréditaire due à l'inactivation d'une copie du gène, puis inactivation de l'autre copie par une mutation somatique et apparition de tumeurs. La France n'était pas en reste grâce à l'exposé de Gilles Thomas sur la neurofibromatose de type II et le sarcome d'Ewing, avec la mise en évidence du rôle dans différents cancers d'un produit de fusion (dû a une translocation) associant l'extrémité 5' de la protéine Ewing avec le site de fixation à l'ADN d'une autre protéine – d'ou l'apparition d'un facteur de transcription anormal dont les effets sur la

régulation cellulaire sont catastrophiques. Les maladies à déterminisme génétique complexe n'étaient pas oubliées ; Mark Lathrop et William Cookson illustraient les voies d'approches et les difficultés de l'étude du déterminisme génétique de l'hypertension, du diabète ou de l'asthme, tandis que Francis Collins, nouveau directeur du programme Génome américain, montrait toute la complexité de l'« après-gène » dans le cas de la neurofibromatose de type I ; deux ans après l'isolement du gène, on n'a encore qu'une idée très partielle des fonctions possibles de la protéine correspondante.

L'éthique avait aussi sa place et, contrairement à l'habitude, les interventions la concernant n'étaient pas placées tout à la fin du colloque, au moment où la salle est clairsemée et ou chacun ne pense plus qu'à son avion. Tenue au début de la dernière journée devant un public nombreux et attentif, cette session fut marquée par les propos percutants de Nancy Wexler, la blonde et infatigable égérie des recherches sur la chorée de Huntington. À partir du récent clonage d'embryons humains et de son exploitation médiatique, elle insistait avec force sur la nécessité d'une attitude a la fois rigoureuse et offensive des scientifiques pour limiter, par exemple, l'analyse d'ADN en vue de la recherche d'une maladie ou d'une prédisposition aux seuls adultes correctement informÉs et consentants, sur leur seule demande, à l'exclusion de tout tiers (assureur, employeur ou, même, membre de la famille). Quant au représentant français, Alain Pompidou, son exposé sur les problèmes de brevetage du génome évitait la rhétorique facile et maximaliste qui a parfois prévalu dans ce domaine, posait les bonnes questions et fut, me semble-t-il, bien reçu par l'auditoire.

Au total, cette première HGM nouveau style constitue un relatif succès. Elle fait certes un peu double emploi avec l'annuelle réunion de Cold Spring Harbor *(CSH Genome Mapping and Sequencing Meeting)*, plus pointue, et les *HUGO meetings* dont le dernier se tint a Nice a l'automne 1992, plus ouverts mais aussi plus rigides sur le plan de leur organisation. Elle a perdu ce qui faisait la spécificité des HGM précédentes, l'accouchement d'un consensus sur les cartes et leur mise en mémoire, mais elle pourrait dans l'avenir être un forum ouvert où se rencontrent tous les acteurs du programme Génome, dans un cadre moins numériquement restreint (et donc moins élitiste) que Cold Spring Harbor (limité à environ deux cents participants). Il faudrait pour cela accroître le rôle des *posters* : il étaient fort bien traités au niveau de l'espace, une salle vaste et bien éclairée où ils sont restés exposés trois jours ; mais aucune tranche horaire ne leur avait été réservée, ce qui leur a fait perdre beaucoup d'efficacité comme moyen de mise en contact. La prochaine édition, HGM 95, prévue en Europe, devrait être plus fréquentée et sera sans doute le véritable test de cette nouvelle formule.

Références

1. Jordan BR. Génome au Japon : au-dela des mythes. *médecine/sciences* 1991 ; 7 : 851-3.
2. Okubo K, Hori N, Matoba R, *et al.* Large scale cDNA sequencing for analysis af quantitative and qualitative aspects of gene expression. *Nat Genet* 1992 ; 2 : 173-9.
3. Jordan BR. Le festival des ADNc. *médecine/sciences* 1993 ; 9 : 211-6.
4. Jordan BR. La robotique en biologie moléculaire : l'arlésienne ? *Biofutur* 1992 ; 108 : 22-5.

Le lecteur aura noté que je commençais à avoir quelques doutes sur la percée décisive que la plupart des observateurs attendaient du Japon dans le secteur de l'instrumentation destinée à la biologie moléculaire. On pensait généralement que Hitachi, ou d'autres firmes, allaient « démocratiser » le séquençage d'ADN en produisant des machines performantes et bon marché, des « séquenceurs personnels » que chaque chercheur tiendrait à avoir sur sa paillasse. L'industrie japonaise rééditerait ainsi le « coup »

réussi à la fin des années 1970, lorsqu'elle avait mis au point, à partir des coûteux magnétoscopes professionnels (inventés aux États-Unis), des appareils de grande diffusion rapidement plébiscités par le marché. Je discute ce thème dans une chronique sur le Japon écrite fin 1991, et dans le commentaire qui l'accompagne (Chapitre 5). *Dans un autre registre, on pourra s'étonner en me voyant évoquer un « récent clonage d'embryons humains » dans la dernière partie de cette chronique : il s'agissait en fait d'un « clonage » par division d'embryons humains très précoces (et non viables car présentant des anomalies chromosomiques), pratiqué par Jerry Hall et Robert Stillman* (George Washington University) *et qui avait fait grand bruit à l'époque.*

Med Sci *1997 ; 13 : 1176-1178*

DU PROGRAMME GÉNOME À LA « PHARMACOGÉNOMIQUE »

Trois ans plus tard, en 1997, le « monde du génome » a bien changé. L'industrie pharmaceutique a rapidement compris l'intérêt que pouvaient présenter pour elle les résultats mais aussi les modes de pensée de ce programme et les technologies qui en sont issues ; les congrès sont de plus en plus organisés par des entreprises commerciales (et non des sociétés savantes), et les industriels y ont pignon sur rue.

Human Genome Conference : on se souvient de ces congrès organisés par HUGO (surtout sa branche américaine) et l'*American Association for the Advancement of Science* (AAAS). Certains eurent lieu en Europe (à Francfort en 1991 puis à Nice en 1992), lors de la grande époque de Généthon et de ses cartes génétique et physique [1]. En mai dernier se déroulait à Monte-Carlo un colloque portant le même nom, *Human Genome Europe*... mais proposé par *Cambridge Healthtech*. Il ne s'agit plus d'une institution académique, mais d'une société tournée vers le monde industriel, et qui met sur pied de multiples réunions portant sur tous les aspects de la « génomique » qui concernent les entreprises[1].

Le congrès de Monte-Carlo n'a pas apporté de révélations scientifiques fracassantes ; il a néanmoins permis de faire un tour d'horizon des secteurs auxquels s'intéresse l'industrie[2], et de la manière dont elle les aborde. L'ambiance de ces réunions surprend un peu : les industriels, qui assuraient plus de la moitié des interventions, sont souvent avares d'informations précises, et parfois très triomphalistes sur des fondements expérimentaux ou conceptuels qui semblent minces. Le pire côtoyait le meilleur : certains ont joué le jeu, apportant des données nouvelles et des précisions qui permettaient d'appréhender les approches expérimentales employées par l'entreprise. D'autres au contraire ont fait des exposés totalement creux, affirmant à chaque diapositive la supériorité de leurs *proprietary techniques* sans jamais les dévoiler, et suscitant un agacement largement partagé. Les « académiques », en minorité parmi les orateurs et encore plus dans l'assistance, étaient chargés de faire le point sur les grands domaines de la recherche sur le génome (séquence, cartographie, bases de données...), ou rapportaient les travaux de leur groupe sur des sujets directement liés aux soucis des entreprises.

À coup sûr, ce colloque témoigne de la place qu'a pris maintenant l'étude du génome dans le monde industriel. À lire le programme ou les résumés de bien des orateurs, on pourrait même croire que la « génomique » est dominante : *« Medical genomics : from genes to products »* (Bill Haseltine, président de *Human Genome Sciences, Inc.*), *« New paradigms and opportunities for biology enabled by genomics »* (David Bailey, *Pfizer*), *« Parallel drug development with gene expression micro-arrays »*

1. Quelques titres glanés au fil de l'année : *Human Genome Project : commercial implications*, *From genes to proteins*, *Bioinformatics and genome research*...

2. Pour une description générale des activités industrielles dans ce secteur, et un autre éclairage sur ces questions, le lecteur pourra avec profit se reporter à l'article de Philippe Froguel et Catherine Smadja récemment paru dans *médecine/sciences* [2].

(Dari Shalon, *Synteni*)... Les néologismes foisonnent : *« Genomics »*, *« Pharmacogenomics »*, ou même *« Proteomics »* pour désigner l'étude « en grand » des protéines ; chacun présente sa « plate-forme technologique » et vante (souvent sans les décrire) ses *proprietary approaches* et ses gadgets aux noms alléchants : *« MAP Technology »*, *« CFLP™ »*, *« Gene-Calling™ »*...

Quelques réflexions sur le positionnement des intervenants et la nature de leur « fonds de commerce » amènent à modérer cette impression. *Cambridge Healthtech Institute* a pour seule activité l'organisation de ces colloques à l'interface génome/industrie, et tend naturellement à souligner l'importance de ce domaine de recherche pour les entreprises. Quant aux orateurs du secteur privé, quelques-uns seulement sont de « vrais » industriels de la pharmacie. Merck, Glaxo, Zeneca, Pasteur-Mérieux-Connaught... mettent effectivement au point des médicaments et s'appuient pour ce faire sur les résultats récemment acquis dans l'étude des génomes. D'autres, comme *Boehringer, Pharmacia, Perkin-Elmer Applied*... vendent des réactifs ou des appareils, et souhaitent naturellement élargir leur marché. Enfin, bon nombre de firmes *(Synteni, Spectra, Human Genome Sciences, CuraGen, Hexagen, Lexicon...)* ont pour vocation de vendre aux industriels des services leur permettant de bénéficier au mieux de la « révolution génomique ». Elles vont donc vanter, parfois jusqu'à l'exagération, toutes les perspectives ouvertes par la connaissance des génomes...

Il n'en reste pas moins qu'une mutation est en cours dans l'univers de la pharmacie. L'avancée considérable des connaissances permet aujourd'hui la définition moléculaire des « cibles » sur lesquelles devra agir un éventuel médicament ; les analogies entre séquences et entre organismes autorisent, si elles sont bien exploitées, la définition de nombreux mécanismes. L'analyse génétique de maladies multifactorielles peut être menée de manière à trouver de nouvelles indications pour des médicaments existants, ou à adapter le traitement à un groupe particulier de patients. L'industrie pharmaceutique, qui vivait depuis longtemps à l'ère de la chimie, est ainsi entrée dans celle de la biologie, et le *design* rationnel de *mechanism-based drugs* n'est plus tout à fait une chimère. Dans un monde où la mise au point d'un nouveau médicament coûte en moyenne quatre cent millions de dollars, tout apport conceptuel susceptible d'abréger le processus d'élaboration (qui s'étale actuellement sur plus de dix ans) ou d'améliorer le taux de succès constitue une avancée déterminante.

Pour donner une idée concrète des travaux d'interface représentés à ce colloque, voyons en quoi consiste l'activité de la société *Incyte*, une des plus connues de ce secteur (http://www.incyte.com). Fondée en 1991, elle emploie aujourd'hui quatre cent cinquante personnes (en comptant les filiales). Son revenu d'environ quarante millions de dollars, entièrement consacré à la recherche (au sens large), correspond à près de 15 % du budget total (salaires compris) de notre Inserm. L'objectif d'*Incyte* est de fournir aux industriels des données, des systèmes de traitement de l'information et des réactifs. En fait, l'entreprise a largement exploité l'exploration du génome par le séquençage partiel d'ADNc, obtenant au total près de deux millions d'EST[3] à partir de 350 banques grâce à ses 68 séquenceurs d'ADN. Cela représente plus du double des données « publiques » accessibles aux chercheurs académiques...

Incyte a, surtout, fait un très gros effort d'organisation de ces résultats, auxquelles s'ajoutent naturellement ceux du domaine public. Une base de données appelée *LifeSeq* contient ces séquences qui, regroupées après analyse comparative, sont copieusement commentées et accompagnées d'informations d'expression ou de localisation sur le génome. Ces dernières proviennent pour partie du secteur académique, pour partie du travail propre d'*Incyte* : l'entreprise compte positionner trente mille nouvelles étiquettes

3. Les chiffres donnés ici proviennent du rapport 1996 d'*Incyte* ou de l'exposé à Monte-Carlo de son vice-Président Jeff Seilhammer.

sur le génome, et pratique la mesure d'expression par *microchip*. L'avantage majeur de la base *LifeSeq*, c'est la multiplicité des liaisons entre informations de nature différente, la qualité de leur organisation et la finesse de l'annotation. Par exemple, une équipe d'une trentaine d'informaticiens et de biologistes a travaillé plusieurs mois pour structurer de manière logique les données fonctionnelles sur les protéines, afin de faciliter la recherche par un industriel des séquences susceptibles de correspondre à une activité enzymatique particulière... *Incyte* vend cette base de données à une vingtaine d'entreprises (parmi celles qui ont rejoint le groupe en 1996, *BASF, Johnson & Johnson, Monsanto*...), et c'est de là que proviennent pour l'essentiel ses revenus.

Le sentiment qui domine chez un chercheur « académique » lors d'une démonstration de *LifeSeq*, c'est l'envie de disposer d'un tel outil... même s'il ne contenait que les données « publiques ». L'investissement en matière grise et en informatique réalisé par Incyte est en effet très supérieur à celui que peuvent effectuer les structures officielles – même le *European Bioinformatics Institute* (EBI) ou la *National Library of Medicine* (NLM). Mais pour accéder à la base, il faut payer quelques millions de dollars... Il est question d'une éventuelle version « ouverte » de *LifeSeq*, mais le planning de sa réalisation reste flou, et son évocation me paraît surtout destinée à calmer un peu la frustration des chercheurs. Frustration qui est une des constantes d'un tel congrès : les outils actuellement développés par les industriels ou leurs fournisseurs de services sont bien proches des intérêts de la recherche fondamentale, mais restent en pratique inaccessibles aux acteurs de cette dernière.

Envisageons, par exemple, les *DNA chips* destinés à la mesure du niveau d'expression de nombreux gènes dans divers tissus. Cette technologie, qui remplacera à terme l'approche plus ancienne des membranes à haute densité hybridées avec des sondes radioactives [3, 4], n'est pas encore aussi performante que le laissent entendre ses promoteurs, mais elle représente certainement l'avenir. Elle est pour l'essentiel entre le mains d'entreprises qui ont effectué les coûteux investissements nécessaires. Les *DNA chips* se présentent en réalité sous deux versions. Ceux sur lesquels sont fixés des dizaines de milliers d'oligonucléotides, mis au point et activement promus par la société *Affymetrix* (entre autres), sont *a priori* plus adaptés au diagnostic de mutations qu'à la mesure d'expression, encore que cette dernière soit possible [5]. D'autres *microchips* (moins miniaturisés) portent une série de produits de PCR représentant quelques centaines ou milliers de gènes et sont destinés à être hybridés avec des sondes complexes fluorescentes préparées à partir de divers tissus ou cellules [6-8]. *Synteni* les baptise « GEM » (joyaux) ou *Gene Expression Modules*, et effectue, à l'aide de ces matrices portant jusqu'à dix mille gènes, des mesures d'expression pour le compte de grands groupes pharmaceutiques. Quelques industriels, comme *Glaxo Wellcome*, s'attachent, eux, à « monter » la technique dans leur laboratoire... ce que leur déconseille, naturellement, *Synteni* ! En tout cas, les GEM ne sont pas au catalogue des fournisseurs de laboratoires, et ces derniers seraient bien en peine de les utiliser pour leurs expériences, en l'absence des appareils très sophistiqués nécessaires pour « lire » le résultat.

Autre outil développé par les entreprises : des collections de mutants de souris (dix mille F1 prévues après mutagenèse chimique, *Hexagen plc*), ou même une banque de *knock-out (Omnibank, Lexicon Genetics, Inc.)*[4]. Le paradigme du clonage positionnel a fait souche, et la recherche de connexions entre marqueurs polymorphiques et maladies (ou mutants) préoccupe les industriels. N'oublions pas, dans cette liste des *proprietary resources* (informations et réactifs non diffusés), de nombreuses séquences complètes de micro-organismes. Il en existe actuellement sept dans le domaine public, mais beaucoup plus, apparemment, dans les ordinateurs de compagnies qui ont fait séquencer *in house* tel ou tel pathogène, thermophile ou auxotrophe en comptant exploiter ces données pour

4. Dans ce dernier cas, un accès partiel (150 lignées de souris) sera offert grâce à l'entreprise Merck, toujours disposée à financer des ressources « publiques ».

s'assurer le *leadership* sur leur marché. On imagine la frustration des spécialistes de la génétique bactérienne, qui pourraient mener de bien belles études comparatives... si seulement ils disposaient des mégabases de séquences éparpillées dans le secteur privé !

Cette situation met en évidence plusieurs évolutions récentes. Elle montre d'abord que les données génomiques, même celles qui paraissent très « cognitives », présentent une utilité pratique, contrairement à ce qu'affirmaient certains opposants des programmes Génome à leurs débuts[5]. Ensuite, elle illustre un brouillage de la distinction entre recherche fondamentale et appliquée. La recherche en pharmacie ne se résume plus à synthétiser à l'infini des variantes de composés connus pour les tester sur d'innombrables rats. Elle s'attache aujourd'hui aux mécanismes, donc au déterminisme génétique de maladies multifactiorielles – à condition qu'elles soient rentables, c'est-à-dire que le marché d'un futur médicament soit évalué à au moins un milliard de dollars. Pour le diabète, l'hypertension, la migraine, l'incontinence urinaire ou l'obésité, ces conditions sont remplies, et les instruments à employer pour discerner un mécanisme sont ceux de la recherche fondamentale : polymorphismes, déséquilibre de liaison, gènes candidats, bioinformatique, génomes et systèmes modèles... Disposant de moyens considérables, soumis à une concurrence féroce, les grands groupes n'hésitent pas à investir pour se doter directement ou indirectement des outils nécessaires qui, on l'a vu, sont bien souvent les mêmes que ceux des chercheurs académiques – mais ne leur sont pas accessibles par nature (car trop massifs, ou nécessitant une infrastructure très lourde) ou leur sont refusés pour des raisons de confidentialité.

La recherche industrielle dispose donc, dans certains domaines, d'une avance technologique – soit qu'elle soit la première à employer de nouvelles méthodes, soit qu'elle les applique à une échelle hors de portée des meilleurs instituts de recherche fondamentale. Il est, de plus, difficile de savoir exactement où en sont ses laboratoires en raison du secret industriel ou, parfois, de tactiques d'intoxication. Même sur des fronts apparemment fondamentaux et cognitifs, elle peut donc se trouver en concurrence directe avec des équipes de recherche académiques : les récepteurs de cytokines, les MAP-kinases... sont probablement mieux connus chez *SmithKline Beecham* ou *Schering-Plough* qu'à l'Inserm, au CNRS ou même au NIH. Certes les premiers publient rarement le résultat de leurs travaux et ne rivalisent guère pour figurer au sommaire de *Human Genetics*, *Nucleic Acids Research* ou *Immunogenetics*... mais cette situation, nouvelle me semble-t-il en biologie, n'en est pas moins fort inconfortable pour nous... Il semble en tout cas exclu de se voiler la face, et de continuer son petit bonhomme de chemin de chercheur sans tenir compte de cette nouvelle donne. Les biologistes de l'après-génome seront obligés de se tenir au courant des sujets et des stratégies du monde industriel, afin de définir les créneaux dans lesquels ils peuvent être les plus efficaces, et de mettre en place quelques partenariats. Le jeu se complique, les nouveaux joueurs n'appliquent pas les règles habituelles, il va falloir en tenir compte. En tout état de cause, si la biologie, et singulièrement la « génomique » ne sont plus seulement un exercice intellectuel mais aussi un levier en prise directe avec la réalité économique, c'est en raison même de leurs succès – et, à ce titre, nous sommes collectivement responsables de cette évolution somme toute très positive.

5. Notons que certains se sont posé la question en temps utile, comme en témoigne par exemple l'organisation d'une réunion entre scientifiques et industriels intitulée *The Genome Project and the Pharmaceutical Industry*, dès septembre 1990 (*Human Genome News*, novembre 1990).

Références

1. Jordan B. Généthon : la réussite d'un pari. *Med Sci* 1992 ; 8 : 1102-5.
2. Froguel P, Smadja C. La génomique à l'épreuve du marché boursier : faut-il avoir peur des sociétés de biotechnologie génétique ? *Med Sci* 1997 ; 13 : 843-5.
3. Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach, H. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm Genome* 1992 ; 3 : 609-19.
4. Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, Jordan BR. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* 1995 ; 29 ; 207-15.
5. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotech* 1996 ; 14 : 1675-80.
6. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995 ; 270 : 467-70.
7. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis : microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996 ; 93 : 10614-9.
8. Derisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996 ; 14 : 457-60.

Cette chronique date de six ans seulement... mais sa lecture indique à quel point le paysage a changé dans cet intervalle. On y sent le choc de la découverte du secteur de la génomique industrielle par un chercheur académique, et son désarroi devant les puissantes technologies mises en œuvre. Les microarrays*, encore à leurs débuts mais en développement rapide, lui semblent presque hors de portée des laboratoires publics, tandis qu'il enrage de ne pouvoir accéder aux riches bases de données créées par des entreprises comme* Incyte*. On note aussi que le terme de « protéomique » lui apparaît presque incongru...*

Med Sci *2001 ; 17 : 290-293*

GÉNOME : LES MÉANDRES DE LA TECHNOLOGIE

Ce texte récent jette un regard rétrospectif sur l'évolution des techniques employées pour l'étude du génome durant la décennie écoulée. Il insiste sur des aspects souvent oubliés dans la reconstruction historique a posteriori *: les espoirs déçus, les fausses pistes, mais aussi les approches imprévues qui se sont révélées fructueuses. Il apporte ainsi une conclusion adéquate à ce chapitre consacré aux questions d'organisation et de technologie.*

Lorsqu'on célèbre l'aboutissement d'un programme, la vision rétrospective de son déroulement tend toujours à gommer les incertitudes, les faux départs et les impasses pour privilégier un déroulement harmonieux et logique – mais bien souvent reconstruit *a posteriori*. La saga du Génome humain n'échappe pas à cette tendance. Il est pourtant instructif d'examiner les inflexions qui ont émaillé son histoire, ne serait-ce que pour se persuader que même un projet aussi « routinier » (d'après ses détracteurs) que le séquençage de notre génome ne s'est pas déployé de manière totalement ordonnée et prévisible. Je me limiterai ici au plan de la technologie ; on pourrait sûrement s'attacher avec profit aux aléas politico-organisationnels de ce grand projet (un peu comme l'a fait Robert Cook-Deegan pour ses premières années dans son excellent livre [1]), mais je ne me risquerai pas – du moins cette fois-ci – sur ce terrain.

Les surprises de la cartographie

Bien que ses promoteurs aient beaucoup parlé de séquençage (ne serait-ce que pour emporter l'adhésion des parlementaires nord-américains et débloquer ainsi les financements), le programme Génome, dans ses premières années, fut logiquement centré sur la construction des cartes génétiques et physiques. La cartographie génétique, fondée sur l'analyse de la transmission de marqueurs polymorphes au sein de grandes familles, allait progresser de manière relativement continue, sur la lancée du concept initial proposé par David Botstein [2]. Les premiers marqueurs employés, les RFLP *(restriction fragment length polymorphisms)* firent progressivement place aux microsatellites (séquences répétées polymorphes du type (GT)n, n étant variable). Aujourd'hui ce sont les SNP *(single nucleotide polymorphisms*, prononcé *snips*) qui tiennent le haut du pavé. Mais il s'agit toujours de détecter de petites variations dans la séquence de l'ADN, et l'étude de familles reste d'actualité. Seules la nature des marqueurs polymorphes, la méthode de détection... et l'échelle de l'analyse ont changé. Le projet mené par Jean Weissenbach au Généthon joua de ce point de vue un rôle central, en montrant l'efficacité d'une organisation centralisée et en aboutissant dès 1992 [3] à une carte génétique à base de microsatellites détaillée et fiable – qui reste encore aujourd'hui le fondement de nombreuses études de Génétique médicale.

Pour la cartographie physique, l'histoire des technologies est beaucoup plus chaotique. À la fin des années 1980, il était beaucoup question de méthodes nouvelles particulièrement adaptées à l'étude de grandes régions d'ADN : l'analyse par électrophorèse en champs pulsés, le saut le long du chromosome (dont le logo de ces chroniques est un

souvenir), et le clonage de très grands fragments d'ADN dans des chromosomes artificiels de levure ou YAC *(yeast artificial chromosomes)* [4]. La découverte d'enzymes de restriction à site rare, coupant l'ADN humain en moyenne toutes les quelques centaines de kilobases, combinée à la séparation des grands fragments obtenus par électrophorèse en champs pulsés [5], devait permettre de construire des « macro-cartes de restriction » s'étendant sur un chromosome entier, à l'image de ce qui avait été réalisé pour des virus avec des enzymes et des méthodes de séparation conventionnelles. Le saut le long du chromosome, acrobatique procédé fondé sur la circularisation de grands fragments d'ADN génomique et permettant de cloner ensemble deux fragments distants de centaines de kilobases dans le génome [6] devait compléter cette approche *top-down* (du haut vers le bas) en fournissant sondes et points de repère.

En réalité, la cartographie par électrophorèse en champs pulsés allait se révéler très aléatoire, notamment en raison des méthylations partielles de l'ADN génomique aux sites des enzymes de coupure, qui faisaient apparaître ou disparaître ces sites en fonction de l'état physiologique des cellules utilisées pour préparer l'ADN à analyser. À cela s'ajoutaient des difficultés techniques pour séparer de très grands fragments de manière reproductible. Cette méthode rendit certes de grands services pour étudier précisément des régions couvrant une ou plusieurs mégabases autour d'un gène particulièrement intéressant, mais elle échoua en tant que technique de cartographie générale. Le saut le long du chromosome, malgré son emploi par le groupe de Francis Collins dans le clonage du gène de la mucoviscidose[1] [7], s'avéra être une technique extrêmement délicate et sujette à artéfacts. Elle disparut progressivement des laboratoires – après avoir fait perdre leur latin à bien des chercheurs.

Restent les YAC. L'article de Burke, Carle et Olson en 1987[2] [8] fit l'effet d'une bombe : cloner 500 kilobases et même une mégabase d'ADN humain dans la levure, c'était réellement un résultat révolutionnaire, et qui arrivait à son heure. Dans les laboratoires du DOE *(Department of Energy)*, à Lawrence Livermore (Californie) et Los Alamos (Nouveau Mexique), des équipes dotées d'importants moyens techniques et informatiques essayaient de construire les cartes des chromosomes 16 et 19 à partir de la comparaison des cartes de restriction de milliers de cosmides. Tâche héroïque, et sans doute impossible. Le Japonais Yuji Kohara avait bien réussi, presque tout seul, à construire en 1987 une telle carte pour le génome d'*Escherichia coli* à partir d'un jeu de trois mille quatre cents segments clonés dans le phage lambda [9]. Mais *E. coli* possède en tout quatre mégabases d'ADN, alors que les chromosomes 16 et 19 en comptent chacun près d'une centaine... On pouvait logiquement penser que la représentation d'un chromosome par quelques centaines de YAC (au lieu d'une dizaine de milliers de cosmides) allait considérablement simplifier le puzzle et rendre viable cette approche *bottom-up* (du bas vers le haut).

De fait, en septembre 1992, un article du groupe de Daniel Cohen [10] paraissait dans la revue *Cell* et faisait l'effet d'une bombe. Intitulé *Mapping the whole human genome by fingerprinting yeast artificial chromosomes*, émanant du CEPH, de Généthon et de l'INRIA, il étendait l'approche du DOE à l'ensemble du génome humain à partir d'une banque de « Méga-YAC » dont les *inserts* approchaient et dépassaient parfois la mégabase. Article fortement médiatisé : le quotidien *Libération* indiquait tout bonnement que cette technique révolutionnaire allait permettre « d'achever d'ici la fin de l'année la cartographie physique des gène humains » [11]. Ces espoirs devaient être déçus. L'article beaucoup plus modeste paru un an plus tard dans *Nature* [12] et intitulé *A first-generation physical map of the human genome* s'appuyait en fait sur la carte génétique réalisée à Généthon et utilisait largement les marqueurs déjà placés par l'analyse génétique pour

1. Une rumeur invérifiable disait que même pour ce projet son rôle avait été moins important qu'il ne semble à la lecture de l'article [7].

2. On oublie trop souvent le second auteur, le Français Georges Carle.

positionner les YAC. L'approche n'avait plus grand chose à voir avec la méthode annoncée dans l'article de *Cell*. De plus, la carte présentée n'était guère utilisable dans la réalité. Une « vraie » carte physique raisonnablement détaillée et fiable ne devait apparaître que nettement plus tard, notamment grâce aux travaux de l'équipe de Lander (*Whitehead Institute*, Cambridge, États-Unis), fondés sur le repérage des STS *(sequence-tagged sites)* dans les clones mais utilisant toujours les méga-YAC du CEPH [13]. La tentative française avait au moins servi à stimuler les autres laboratoires en leur montrant qu'il était concevable d'établir de manière globale la carte physique du génome humain...

En fait, les YAC avaient trahi leurs utilisateurs par un chimérisme catastrophique : la moitié (peut-être même plus) des clones étaient formés de segments d'ADN provenant de deux régions distinctes du génome. Sans parler des clones instables dont la taille diminuait à chaque analyse, ou de ceux qui ne devenaient stables qu'après avoir subrepticement perdu la moitié de leur insert... Finalement, le mérite des YAC fut de prouver qu'il était possible de cloner de grands segments d'ADN, et d'encourager la recherche d'autres systèmes plus fiables : dès 1990, Nat Sternberg dévoilait un vecteur fondé sur le phage P1, qui permettait de propager dans des bactéries des fragments d'au moins cent kilobases [14]. Aujourd'hui, les cartes physiques « prêtes à séquencer » sont généralement construites à l'aide de BAC *(bacterial artificial chromosomes)* dont les *inserts* de deux cents kilobases sont presque toujours stables et non chimériques.

Finalement, des trois méthodes qui, en 1990, semblaient devoir aboutir aux cartes physiques, seuls subsistent aujourd'hui les BAC, héritiers des YAC. D'autres approches nouvelles, en particulier l'emploi systématique d'hybrides d'irradiation, ont joué un rôle complémentaire important. Si les « cartes intégrées » du génome sont aujourd'hui une réalité, c'est grâce à des méthodes bien différentes de celles sur lesquelles on avait parié (et investi) il y a une dizaine d'années.

La révolution du séquençage se fait attendre

Le séquençage intégral fut, dès le début, le cheval de bataille des promoteurs du Programme Génome Humain. Il semblait pourtant hors de portée des procédés existant à l'époque : technique de Maxam et Gilbert ou de Sanger, mises en œuvre à la main avec des marquages radioactifs ou, avec des réactifs fluorescents, dans les premiers « séquenceurs » développés dans le laboratoire de Lee Hood et produits par la compagnie *Applied Biosystems* dès 1987. Le débit maximum dans les meilleures conditions était estimé à une centaine de kilobases par an et par personne – trop peu pour s'attaquer aux trois milliards de bases de notre génome... Le potentiel d'amélioration de ces approches semblait pour beaucoup très limité[3].

On fondait donc beaucoup d'espoirs, notamment aux États-Unis, sur la mise au point de nouvelles méthodes. Il y avait le choix : excision puis détection par cytométrie de flux de bases individuelles (à partir d'une seule molécule d'ADN), lecture directe de la séquence par microscopie à effet tunnel ou à force atomique [15], analyse par spectrométrie de masse, ou encore séquençage par hybridation. Ces méthodes de déchiffrage « sans gel » devaient augmenter l'efficacité de plusieurs ordres de grandeur, et on considérait généralement que la majeure partie du génome humain serait déchiffrée à l'aide d'une d'elles[4]. En fait aucune de ces méthodes n'a contribué au séquençage. La première est restée bloquée par des problèmes de sensibilité (il faudrait être capable de détecter à coup sûr une seule molécule de base fluorescente). Les microscopies à champ proche, elles, ont produit beaucoup d'artéfacts interprétés de manière optimiste comme de l'ADN

3. Un document du *Department of Energy* datant de 1992 (*Primer on Molecular Genetics*, DOE, juin 1992) les qualifie de *interim sequencing technologies*.

4. Le même document indique *Third-generation gel-less sequencing technologies... are expected to be used for sequencing most of the human genome.*

jusqu'à ce qu'on se rende compte que la visualisation de cette molécule était effectivement possible dans des conditions opératoires bien précises, mais que la résolution obtenue ne permettait pas de distinguer les bases. La spectrométrie de masse commence à avoir un impact dans certaines méthodes de détection de SNP, mais pas pour le séquençage *a priori*. De même, l'hybridation à des jeux d'oligonucléotides s'avère très performante pour la recherche de mutations (donc la détection de variants par rapport à une séquence connue) mais impraticable pour la détermination de séquences inconnues.

C'est donc la méthode « conventionnelle » de Sanger qui est responsable de la quasi-totalité des trois milliards de nucléotides d'ADN humain maintenant emmagasinés dans EMBL et GenBank. Elle s'est finalement révélée de loin la plus opérationnelle, et a atteint le but requis grâce à de nombreux perfectionnements de détail, à l'automatisation de plusieurs étapes[5], à des financements conséquents et à de gros progrès tant dans une organisation de type industriel pour les opérations de mégaséquençage que dans le contrôle, l'assemblage et l'exploitation informatique des données. La surprise, ici, a été l'absence de révolution, l'échec de toutes ces nouvelles approches (je n'ai cité que les principales) dont on pouvait penser qu'au moins une se révèlerait efficace, et les potentialités d'une méthode datant de 1977 – autant dire de la préhistoire !

EST : le succès d'un franc-tireur

Le dernier exemple d'inflexion inattendue est bien sûr celui des EST *(expressed sequence tags)*. L'idée de s'intéresser préférentiellement aux gènes, donc à l'ARN messager *via* des banques d'ADNc, avait déjà été proposée plusieurs fois, notamment par Sydney Brenner. Plus près de nous, un rapport établi en 1990 par Philippe Kourilsky proposait de centrer le programme français sur le séquençage des ADNc et d'y consacrer au moins 20 % du budget[6]. Mais aux États-Unis, premier parti dans la course du génome, la doctrine officielle était de viser le séquençage intégral de l'ADN génomique humain, et de ne pas se laisser distraire par des projets latéraux. Il faut avouer que les tentatives précédentes de séquençage au hasard d'ADNc n'avaient pas été très fructueuses [16] – sans doute ces projets étaient-ils trop sous-dimensionnés pour être efficaces.

Le génie de Craig Venter fut de définir un mode opératoire à grande échelle adapté aux séquenceurs existants : prendre les clones d'ADNc au hasard dans des banques et de se contenter pour chacun d'eux d'une seule réaction de séquence donnant quelques centaines de bases avec un taux d'erreur pouvant approcher 1 % [17]. Ce mode opératoire était bien éloigné des habitudes des chercheurs, qui préfèrent choisir amoureusement quelques ADNc puis les séquencer intégralement avant de les décrire sous toutes les coutures dans un article prévu pour *Nature* – mais qui finit souvent dans une revue bien moins prestigieuse... Mais l'approche de Venter devait démontrer un rapport qualité/prix incomparable. L'approche des EST connut le succès que l'on connaît, attesté aujourd'hui par les presque sept millions de séquences contenues dans dbEST – sans parler des millions que revendiquent *Incyte* ou *Human Genome Sciences*. L'importance scientifique et économique de ces séquences partielles a été amplement démontrée, depuis le clonage *in silico* du gènes responsables de cancers héréditaires [18] jusqu'à la découverte de protéines thérapeutiques comme un facteur de croissance de kératinocytes, KGF-2 [19], rebaptisé répifermine par *Human Genome Sciences* et maintenant proche de la mise sur

5. Il y aurait aussi beaucoup à dire sur les méandres de la robotique associée aux projets Génome : les échecs ont été nombreux, même au Japon (pourtant *a priori* bien parti), et encore aujourd'hui l'atelier de séquençage entièrement robotisé (dont la mise en place fut tentée à plusieurs reprises) reste une vue de l'esprit.

6. Ce rapport (« Rapport sur le Génome Humain, élément d'un projet "Génomes" » par Philippe Kourilsky, 5 juillet 1990) demandé par le Ministre de la Recherche, Hubert Curien, et établi après consultation de nombreux chercheurs, reste très intéressant à lire aujourd'hui. Son auteur avait bien vu les enjeux et esquissé des stratégies (scientifiques et organisationnelles) tout à fait adéquates. Il est fort dommage que ses recommandations n'aient guère été suivies...

le marché comme accélérateur de cicatrisation. La séquence intégrale du génome humain, aujourd'hui disponible, ne devient interprétable que grâce à la comparaison avec les EST, comme l'expérience des chromosomes 22 et 21 l'a amplement démontré [20].

La morale de l'histoire

Sur le plan technologique (le seul que nous abordions aujourd'hui), le déroulement du Programme Génome a donc été tout sauf un long fleuve tranquille. On s'est beaucoup trompé, bien des certitudes se sont effondrées tandis que des paris osés étaient tantôt gagnés, tantôt perdus[7]. C'est la règle du jeu : la prévision ne marche vraiment bien qu'après coup ! Il faut en tirer les conséquences : de grands programmes scientifiques comme celui-ci demandent à la fois continuité et souplesse. Continuité : les « coups d'accordéon » financiers qu'a subi le projet français au fil des changements de gouvernement ont très certainement nui à l'efficacité de recherches qui demandent à être planifiées sur deux ou trois ans. Mais aussi souplesse : les inflexions technologiques peuvent rendrent très vite obsolète un programme ou des installations[8], il faut les percevoir à temps et ajuster le tir en conséquence. Et naturellement cette souplesse doit aussi caractériser les structures et les personnes. De ce point de vue, le statut d'EPST, les règles de la comptabilité publique, le code des marchés et aussi la qualité de fonctionnaire ne constituent sans doute pas le cadre le mieux adapté...

Références

1. Cook-Deegan R. *The gene wars.* New York : WW Norton and Co, 1994.
2. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 1980 ; 32 : 314-31.
3. Weissenbach J, Gyapay G, Dib C, *et al.* A second-generation linkage map of the human genome. *Nature* 1992 ; 359 : 794-801.
4. Jordan BR. Megabase methods : a quantum jump in recombinant DNA techniques. *Bioessays* 1988 ; 8 : 140-5.
5. Schwartz DC, Cantor CR. Separation of yeast chromosome sized DNAs by pulsed field gradient gel electrophoresis. *Cell* 1984 ; 37 : 67-75.
6. Collins FS, Weissman SM. Directional cloning of DNA fragments at a large distance from an initial probe : a circularization method. *Proc Natl Acad Sci USA* 1984 ; 81 : 6812-6.
7. Collins FS, Drumm ML, Cole JL, Lockwood WK, Vande Woude GF, Iannuzzi MC. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* 1987 ; 235 : 1046-9.
8. Burke DT, Carle GF, Olson MV. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 1987 ; 236 : 806-12.
9. Kohara Y, Akiyama K, Isono K. The physical map of the whole *E. coli* chromosome : application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* 1987 ; 50 : 495-508.
10. Bellanne-Chantelot C, Lacroix B, Ougen P, *et al.* Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* 1992 ; 70 : 1059-68.
11. Bensimon C. Percée décisive dans la recherche génétique. *Libération*, 18 septembre 1992.

7. Pas question de me poser en censeur sentencieux : je me suis trompé au moins aussi souvent que les autres, couvrant d'éloges les « méthodes mégabase » (gels pulsés, saut et YAC) ou exprimant mon scepticisme sur la possibilité d'obtention de la séquence humaine dans un délai raisonnable...

8. La salle des Mark II au Généthon en est un bel exemple. Ces machines automatisant la technique du *Southern blot*, mises au point (assez lentement) par l'entreprise Bertin, entrèrent en production au moment où la cartographie génétique basculait des RFLP (détectés par *Southern blot*) aux microsatellites (analysés sur gel de séquence). Cet « éléphant blanc » (2 MF par machine) fut quelque temps recyclé dans le *fingerprinting* des YAC avant d'être démantelé.

12. Cohen D, Chumakov I, Weissenbach J. A first-generation physical map of the human genome. *Nature* 1993 ; 366 : 698-701.
13. Hudson TJ, Stein LD, Gerety SS, *et al.* An STS-based map of the human genome. *Science* 1995 ; 270 : 1945-54.
14. Sternberg N. Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc Natl Acad Sci USA* 1990 ; 87 : 103-7.
15. Jordan BR. Le tunnel séquencera-t-il le génome ? *Med Sci* 1990 ; 6 ; 1007-8.
16. Putney SD, Herlihy WC, Schimmel P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* 1983 ; 302 : 718-21.
17. Adams MD, Kelley JM, Gocayne JD, *et al.* Complementary DNA sequencing : expressed sequence tags and human genome project. *Science* 1991 ; 252 : 1651-6.
18. Papadopoulos N, Nicolaides NC, Wei YF, *et al.* Mutation of a *mutL* homolog in hereditary colon cancer. *Science* 1994 ; 263 : 1625-9.
19. Jimenez PA, Rampy MA. Keratinocyte growth factor-2 accelerates wound healing in incisional wounds. *J Surg Res* 1999 ; 81 : 238-42.
20. Jordan B. ADNc : les incontournables. *Med Sci* 2001 ; 17 : 81-4.

This page intentionally left blank

5. AUX QUATRE COINS DU GLOBE

Nous allons maintenant parcourir le monde pour décrire les différents « Programmes Génome » nationaux. Je fais ici une petite entorse au principe de ce recueil, censé couvrir la période 1992-2002 : j'y gi inclus deux chroniques parues fin 1991, celles qui concernent le Japon et les États-Unis, afin que le tour d'horizon soit complet. Les textes consacrés à la France sont regroupés au chapitre suivant.

Med Sci *1991 ; 7 : 851-853*

GÉNOME AU JAPON : AU-DELÀ DES MYTHES

Première de ces chroniques : celle qui traite du Japon. Cette nation était alors au faîte de sa prospérité, avant que ne débute la crise prolongée qui la frappe depuis le milieu des années 1990. Deuxième puissance économique mondiale, elle taillait des croupières aux États-Unis dans de multiples secteurs (électronique, automobile...) : l'on avait naturellement tendance à lui prêter de grandes potentialités dans le domaine de la recherche sur le génome. L'article s'attachait à dégonfler certaines de ces illusions, tout en restant assez optimiste sur l'avenir.

Le temps des illusions

Beaucoup ont cru, vers la fin des années 1980, que le Japon s'était sérieusement engagé dans la recherche sur le Génome humain : grâce à l'ingéniosité et à l'acharnement de ses chercheurs, grâce aussi à ses incontestables compétences dans le domaine de la robotique, ce pays était – pensait-on – devenu un concurrent dangereux pour les États-Unis. Cette impression très répandue était pourtant erronée, et reposait sur une série de malentendus et d'ambiguïtés, peut-être pas toujours involontaires...

Premier malentendu, encore courant dans le grand public : le programme Génome aurait pour objet immédiat et principal un séquençage massif de l'ADN humain. Ce serait, effectivement, un « travail de Japonais » que de séquencer trois milliards de nucléotides à l'aide des techniques actuelles : dans la vision souvent déformée que l'on a de ce pays, on l'imagine bien s'attelant à une entreprise aussi colossale et abrutissante. Mais, on le sait, l'essentiel des programmes Génome est en fait consacré à l'établissement de cartes génétiques et physiques, et ce sont là, compte tenu de la constante évolution des méthodes, des activités de recherche à haut contenu intellectuel qui ne peuvent être confiées à une armée de techniciens... ou de robots !

Un deuxième malentendu, celui du programme *Human Frontiers* est venu accroître la confusion. Cet important programme international financé, au départ, presque exclusivement par les Japonais, attribue des contrats à des projets réalisés en collaboration par des équipes de plusieurs pays. On a pensé, et parfois écrit, qu'il avait pour objet l'étude du Génome et que par ce biais le Japon cherchait à prendre la direction de ces travaux. Il est vrai que la confusion était permise, car les objectif étaient présentés, comme parfois dans ce pays, d'une façon à la fois très grandiloquente et – il faut bien le dire – assez fumeuse. Mais les choses sont maintenant claires : *Human Frontiers* (dont le secrétariat est à Strasbourg, et auquel d'autres pays que le Japon commencent à contribuer) soutient des études sur le système nerveux, sur le fonctionnement du cerveau, sur certains thèmes de biologie moléculaire... mais pas sur le Génome en tant que tel.

Enfin, l'image du Japon dans ce secteur a été infléchie par les efforts d'Akiyoshi Wada, scientifique connu et très puissant dans ce pays, qui a consacré beaucoup d'énergie autour de 1985 à mettre en place des moyens de séquençage à très grande échelle. La tentative n'a pas réellement abouti à cette époque : la technologie n'était pas mûre comme on l'a vu avec le relatif échec des grands projets de séquençage lancés depuis [1]. Mais elle a donné lieu à quelques articles d'un imprudent optimisme : l'un d'eux publié par la

revue *Nature* en 1987 annonçait le séquençage à 17 cents (US) la base [2]. Cinq ans plus tard le DOE (*Department of Energy*, aux États-Unis) tente toujours d'en évaluer le coût, et les chiffres cités tournent autour de 1 à 3 dollars... Ces articles ont naturellement renforcé l'illusion que le Japon avançait très vite sur le Génome.

Réalités japonaises

J'avais pu appréhender la situation sur le terrain lors de visites de laboratoires, il y a bientôt trois ans, et les réalisations m'avaient en effet paru modestes. Cette perception était d'ailleurs en accord avec le petit nombre de travaux japonais présentés dans les colloques internationaux spécialisés comme l'annuel *Cold Spring Harbor Genome Mapping and Sequencing Meeting* et leur qualité souvent moyenne. Une étude bibliométrique récente réalisée pour la *European Science Foundation* confirme que pour l'année 1990 la « production » japonaise était nettement inférieure à celle de la France[1]... Trois causes à cela, et tout d'abord la faiblesse de la génétique médicale dans ce pays. Elle est due à des raisons culturelles : les « tares » héréditaires y sont objet de honte bien plus que chez nous, et les individus atteints sont en général assumés dans le cadre de la famille sans faire appel à des institutions ; les prélèvements sanguins (base de toute étude familiale) se heurtent aussi à de fortes réticences. L'accès aux malades est donc ardu, leur étude malaisée : or, la génétique médicale a été le substrat du développement de la génétique moléculaire humaine, puis des travaux sur le Génome...

Autre raison de ce retard, la relative faiblesse de la recherche fondamentale : le Japon a d'abord développé (avec le succès que l'on connaît) les travaux appliqués, parfois très appliqués ; et ce n'est que plus tard que les laboratoires fondamentaux ont bénéficié de budgets importants. Or, quoi qu'en pensent certains, le Génome fait partie de ce secteur (pensons aux gels pulsés, aux YAC, aux librairies de saut, à l'*in situ* non radioactif...), il ne se réduit pas à l'application de méthodes déjà éprouvées : on ne peut pas multiplier du jour au lendemain les laboratoires, et encore moins les compétences, nécessaires à un développement de ces études.

Troisième explication, plus circonstancielle cette fois : la multiplicité (comme en France d'ailleurs) d'agences gouvernementales s'occupant de recherche, et les difficultés de concertation entre leurs programmes, difficultés auprès desquelles les brouilles périodiques entre l'Inserm et le CNRS font figure de querelle d'amoureux... Ces luttes entre ministères expliquent qu'il y ait à l'heure actuelle non pas un mais quatre ou cinq programmes Génome (même si leur coordination s'améliore, comme on le verra plus loin), et qu'il soit particulièrement délicat de s'y retrouver dans un paysage aussi complexe.

Le démarrage

Mais au Japon comme ailleurs, les situations ne sont pas figées ; et une enquête sur place de près d'un mois, comportant des discussions avec les principaux responsables et la visite de leurs laboratoires, fait apparaître une très nette montée en régime des recherches dans ce secteur. Cela est en grande partie dû au travail d'organisation de Kenichi Matsubara, scientifique respecté et fin diplomate, qui dirige le programme du ministère de l'Éducation Nationale (Monbusho) mais joue aussi un rôle dans la coordination avec les autres projets : ceux de la *Science and Technology Agency* (STA), du ministère de la Santé, et de l'Agriculture. L'ensemble de ces programmes reçoit un financement annuel de l'ordre de deux milliards de Yens, soit une petite centaine de MF (millions de francs) : soit à peu près autant en valeur absolue que ce qui est prévu en

1. *Report on Genome Research 1991, European Science Foundation*, Strasbourg, France, 1991.

France, et donc sensiblement moins par rapport à la puissance économique du pays ; cela n'en reste pas moins un investissement notable[2].

Le programme du Monbusho (une vingtaine de MF) vise dans un premier temps à développer, systématiser et coordonner des études actuellement en cours : banques d'ADNc complet dans des vecteurs d'expression (H. Okayama, à Kyoto), isolement de sondes et cartographie génétique des chromosomes 3 et 11 (Y. Nakamura, Tokyo), mise en place de systèmes informatiques et accès aux banques de données (M. Kanehisa, Kyoto puis Tokyo), ou même séquençage de *S. pombe* (M. Yanagida, Kyoto) et cartographie physique de *S. cerevisiae* (K. Isono, Kobe), pour n'en citer que quelques-uns. Le fait important, et relativement nouveau, est qu'il ne s'agit plus là de grands programmes futuristes et un peu abstraits mais de travaux tangibles, déjà menés à un bon niveau international par des équipes compétentes ; l'objectif de K. Matsubara est de leur donner les moyens de se développer tout en les coordonnant dans une optique « Génome ». Le programme n'est pas encore centré sur un chromosome ou une région donnée. L'approche est donc très pragmatique, et on commence à en voir les résultats dans plusieurs des laboratoires cités ci-dessus.

L'autre programme important, plus important même que celui du Monbusho, est géré par Yoshi Ikawa pour la STA : près de quarante MF par an. Il finance d'une part des recherches dans deux « laboratoires propres » : le *Riken Life Science Center* (Tsukuba) et l'*Institute of Radiation Biology* à Chiba (l'un et l'autre près de Tokyo), et d'autre part des équipes universitaires par le biais de contrats qui représentent souvent pour elles un apport précieux. Comme son nom l'indique, la STA privilégie les aspects technologiques : c'est sous ses auspices qu'a été construite la première bonne banque YAC japonaise, celle que T. Imai (maintenant dans le laboratoire de Y. Nakamura à Tokyo) a commencée à Saint Louis (MO, USA) puis continuée au *Riken* à Tsukuba ; c'est également au *Riken* qu'est poursuivi l'effort d'automatisation du séquençage d'ADN (suite du programme Wada) dont je dirai quelques mots par la suite. Le laboratoire de Chiba (à ne pas confondre avec le futur institut de séquençage, financé, lui, par les autorités locales et des industriels et dont la revue *Nature* s'est fait l'écho il y a quelque temps[3] se consacre principalement à l'hybridation *in situ* : il a ainsi positionné plusieurs centaines de cosmides sur les chromosomes 3, 11 et 21, ceux sur lesquels la STA concentre son effort. Et, comme mentionné précédemment, les contrats distribués sont un apport vital pour les laboratoires universitaires qui reçoivent souvent par cette voie plus d'argent que du Monbusho. Certains programmes de la STA apparaissaient peu convaincants par le passé, car les réalisations concrètes n'étaient pas à la hauteur des ambitions affichées : des progrès notables ont eu lieu, la formation des chercheurs a porté ses fruits, et des travaux de bonne qualité sont maintenant en cours.

Restent encore les programmes du ministère de la Santé (une dizaine de MF, pour des études ciblées sur certaines maladies génétiques) et du ministère de l'Agriculture, avec notamment un projet « Génome du riz » et un financement équivalent... ainsi que les efforts du MITI[4] pour motiver les industriels. Il s'agit là, autant que je sache, d'intentions non encore traduites dans les faits : à suivre donc, mais sans incidence significative dans l'immédiat.

2. Compte tenu de la diversité des règles comptables, les montants ne sont que très approximativement comparables d'un pays à un autre, et ne donnent que des ordres de grandeur.

3. *Nature (News)* 1991 ; 349 : 640.

4. MITI : ministère du Commerce International et de l'Industrie, fortement soutenu par les industriels japonais et impliqué dans des programmes comme *Human Frontiers* ou dans la création d'instituts de pointe, par exemple le *Protein Engineering Research Institute* (PERI) à Osaka.

Compter avec le Japon

Le retour dans les mêmes laboratoires après un intervalle de deux ou trois ans s'avère ainsi extrêmement instructif : le changement d'ambiance, en ce qui concerne les recherches sur le Génome, est très sensible. Plusieurs de ces équipes, dans le secteur des YAC et de leur exploitation, de la cartographie de chromosomes, de l'analyse d'ADNc... se situent maintenant à un bon niveau international ; les programmes du Monbusho et de la STA, et la personnalité de K. Matsubara ne paraissent pas discutés ; et l'on a le sentiment que l'implication du Japon dans ce domaine est sérieuse et durable.

Les changements sont particulièrement nets en ce qui concerne l'instrumentation. On s'attend certes à ce que les Japonais soient performants dans ce secteur : ils l'ont montré à propos de produits comme les ordinateurs ou les magnétoscopes, et il y a à cela des raisons structurelles : intégration de grands groupes industriels comme *Mitsubishi* qui disposent de compétences internes pour l'ensemble des technologies, tendance « culturelle » à privilégier l'investissement à long terme, choix général de l'option robotique (par opposition à l'importation de main d'œuvre). Mais comme j'ai déjà eu l'occasion de le dire [3], les premiers résultats en ce qui concerne l'instrumentation associée au Génome s'étaient avérés décevants : relatif échec de la tentative d'A. Wada, ou commercialisation par *Seiko Instruments* d'appareils automatisant une technique périmée, le séquençage d'ADN par la technique de Maxam et Gilbert...

Ces balbutiements appartiennent au passé ; des appareils performants sont maintenant fabriqués, et mis sur le marché au Japon, parfois à l'étranger : robot *Seiko* très compact pour effectuer les réactions de séquence (méthode Sanger, bien sûr), système à *Imaging Plate* de Fuji qui remplace l'autoradiographie dans nombre de cas, robot *Kubota* pour les préparations d'ADN qui peut traiter 144 échantillons sans intervention... Et Eichi Soeda m'a montré au *Riken* (Tsukuba) une étonnante « usine à séquencer l'ADN » (rapidement mentionnée dans un récent numéro de *Nature*[5], sorte de chaîne associant une douzaine de robots qui prennent chacun en charge une opération : prélèvement des colonies bactériennes, préparation de l'ADN, réactions de séquence, coulage des plaques... jusqu'au démoulage des gels pour les jeter en fin de chaîne. Cet ensemble comprend non seulement les robots, mais aussi les « convoyeurs » assurant le transfert des échantillons d'une machine à la suivante. Il reste naturellement à voir comment cela fonctionne, quelle est la fiabilité et si la production est réellement à la hauteur des chiffres impressionnants annoncés : environ cent mille nucléotides de séquence brute par 24 heures. Il s'agit en tous cas, à ma connaissance, d'un système et d'une tentative uniques au monde, et c'est en fait une des retrombées du programme lancé par Wada il y a plus de six ans : on travaille dans le long terme au Japon !

Il faut donc s'attendre à une percée de l'instrumentation japonaise, d'autant que les industriels de ce pays appliquent une stratégie qui leur a bien réussi par le passé : mettre à la portée d'un large public des appareils précédemment produits en petit nombre et à des prix élevés. Hitachi, par exemple, qui commercialise actuellement un séquenceur d'ADN comparable à la machine LKB/Pharmacia, serait très avancé dans la mise au point d'un « séquenceur personnel » beaucoup moins cher et destiné à une large diffusion : le marché potentiel d'une telle machine, qui concernerait la plupart des laboratoires de Biologie moléculaire, serait sans nul doute considérable. Ces différents développements techniques, les laboratoires Japonais en seront évidemment les premiers bénéficiaires...

Si mon analyse est exacte, le Japon va donc devenir dans les années qui viennent un acteur de premier plan dans le domaine du Génome : fait nouveau, dont il faudra tenir compte, et qui devrait nous inciter à plus d'efforts de collaboration. Pour des raisons à la fois géographiques et culturelles, ces relations sont plus complexes à mettre en place

5. *Nature (News)* 1991 ; 351 : 593.

qu'avec la Grande-Bretagne ou les États-Unis ; mais il semble exister un certain nombre d'opportunités qu'il serait judicieux de saisir... avant que d'autres ne le fassent à notre place.

Références

1. Jordan BR. Les heurs et malheurs du séquençage d'ADN à grande échelle. *médecine/sciences* 1991 ; 7 : 612-3.
2. Wada A. Automated high-speed DNA sequencing. *Nature* 1987 ; 325 : 771-2.
3. Jordan BR. Les sigles et les gros sous. *médecine/sciences* 1990 ; 6 : 288-90.

Les pronostics relativement optimistes qui concluent ce texte n'ont pas été confirmés : depuis dix ans, on a toujours l'impression que le programme Génome japonais est sur le point de démarrer réellement. À mon avis, une des causes de cet état de fait est que les schémas culturels nationaux se prêtent mal à une recherche technologique multidisciplinaire et aventureuse (comme l'ont été les approches développées par Craig Venter, John Sulston ou Leroy Hood). Au Japon, les équipes fonctionnent au consensus, les scientifiques marginaux sont mal acceptés et les connexions entre chercheurs de discipline différentes sont encore plus difficiles qu'ailleurs. De surcroît, les luttes entre organismes et ministères rivaux, un instant calmées sous la houlette de Kenichi Matsubara, ont repris de plus belle. À ce jour, on ne peut pas dire que le Japon ait fourni une contribution majeure au programme Génome humain, comparable à celle de la Grande-Bretagne, voire de la France (dont le PIB ne représente pourtant que le tiers du celui du Pays du Soleil Levant). Et même sur le plan de l'instrumentation, l'industrie japonaise n'a pas effectué la percée attendue : les séquenceurs d'ADN, les automates de laboratoires ou les scanners *de* microarrays *restent pour l'essentiel l'apanage des firmes nord-américaines.*

Med Sci *1991 ; 7 : 957-959*

USA : UN PROGRAMME GÉNOME SOLIDEMENT INSTALLÉ

Datant lui aussi de fin 1991, un tableau plutôt positif mais finalement assez exact du programme Génome aux États-Unis, qui se termine sur une évocation des atermoiements français.

Plantons le décor

Le programme Génome aux États-Unis est bien vivant ; il est maintenant enraciné dans le tissu scientifique du pays et l'on n'imagine pas qu'il puisse être abandonné, ou même sérieusement ralenti, ce qui n'était pas le cas il y a un ou deux ans. Sa direction bicéphale est assurée par le DOE *(Department of Energy)* et le NIH *(National Institutes of Health)*, le premier étant à peu près l'équivalent de notre CEA, l'autre de l'Inserm ; la *National Science Foundation*, qui correspond au CNRS, est peu engagée dans la recherche biologique. Les laboratoires qui effectuent les travaux vont de l'équipe universitaire classique, très focalisée sur des problèmes biologiques précis (l'isolement du gène impliqué dans une maladie, en général) à l'« usine à cartographier » qui se consacre à l'établissement de cartes physiques ou génétiques, utilise beaucoup de machines et fonctionne selon un *planning* de type industriel. Mais situons d'abord le terrain des uns et des autres.

Le DOE dispose d'un budget (annuel) de l'ordre de 60 millions de dollars qui alimente trois centres principaux : Lawrence Livermore (Anthony Carrano, chromosome 19) ; Los Alamos (Bob Moysis, chromosome 16) et Lawrence Berkeley (ancien responsable Charles Cantor, chromosome 21), ainsi que des équipes plus restreintes localisées dans ses autres laboratoires : Argonne, Brookhaven, Oak Ridge. Enfin, un quart des crédits alimente des groupes extérieurs comme celui de George Church (Harvard, MA, USA) ou même étrangers comme celui de Grant Sutherland (Adelaide, Australie) qui collabore avec le centre de Los Alamos pour la carte génétique et cytogénétique du chromosome 16.

Le NIH, lui, finance (à l'intérieur d'un budget annuel de 108 millions de dollars) une dizaine de centres : Rick Myers (UCSF, San Francisco, CA : chromosome 4, en démarrage) ; David Schlessinger (Washington University, St-Louis, MO : un centre qui « tourne » déjà pour la carte physique par YAC du X et du 7) ; Glen Evans (Salk, San Diego, CA pour le 11) ; Ray White et Ray Gesteland (Université de l'Utah, Salt Lake City pour la carte génétique des 16, 17 et 5) ; Tom Caskey (Houston, TX pour des régions du X et du 17) ; Francis Collins (Université du Michigan, Ann Arbor, MI pour la recherche de gènes impliqués dans des maladies), et, tout récemment, Beverly Emanuel (Université de Pennsylvanie, PA) pour la cartographie du chromosome 22). De plus Éric Lander (MIT, Boston, MA) est financé pour la cartographie du Génome de la souris et David Botstein et Ron Davis (Stanford, CA) pour un projet exploratoire de séquençage de la levure. Chacun de ces laboratoires emploie au total trente à quarante personnes et reçoit un budget de l'ordre de 10 millions de dollars (salaires compris) pour quatre à cinq ans.

La sociologie des *Genome centers*

Seuls certains de ces centres sont à l'heure actuelle vraiment opérationnels ; les autres sont en cours d'installation, leurs contrats ayant été accordés fin 1990 ou début 1991. Mais on voit déjà les différences : à Saint Louis, toute une organisation très structurée est mise en place pour effectuer systématiquement et de bout en bout la cartographie physique (par alignement de YAC en *contigs*) des chromosomes X et 7. David Schlessinger et ses collaborateurs ont la volonté de faire les choses à fond et de manière homogène sur l'ensemble de ces chromosomes, et la ferme intention de ne pas dériver sur l'étude d'une région particulière, aussi attractive puisse-t-elle être. À Houston, Tom Caskey, au contraire, a explicitement obtenu le contrat Génome pour renforcer les « services » (criblage de YAC, cytogénétique, séquence, informatique...) d'un ensemble d'équipes qui chacune poursuit avec vigueur et compétence l'isolement du gène impliqué dans une ou deux maladies.

Les différences entre DOE et NIH sont également marquées. Le DOE, plus « professionnel », plus technologique, avec un personnel plus stable et un climat de compétition un peu moins intense, fournit un cadre favorable au fonctionnement de Genome Centers voués à l'étude complète d'un chromosome : c'est le cas de Lawrence Livermore et de Los Alamos. Peu de post-docs et encore moins d'étudiants dans ces équipes, mais beaucoup de technologie, des programmes à long terme et une ambiance souvent moins tendue que dans les laboratoires « académiques » ; de bons résultats aussi, d'autant plus que les YAC sont arrivés à point nommé pour compléter une stratégie de cartographie physique engagée à l'époque des cosmides et qui commençait à atteindre ses limites. Les *Genome centers* du NIH, quant à eux, ont malgré leur diversité le caractère commun d'être proches du monde universitaire, avec un flux de personnel important, des étudiants, des *post-docs*, du mouvement, de la vie... mais aussi du mal à assumer l'aspect systématique de l'entreprise, à maîtriser l'envie d'explorer les pistes ponctuelles mais prometteuses qui apparaissent au cours d'un travail systématique, et qui peuvent amener à le délaisser. C'est une contradiction difficile à gérer mais importante : la raison d'être du programme Génome est justement de promouvoir une approche globale. Dans le système « académique » classique, ce type de travail n'est sans doute pas suffisamment reconnu...

Quelques enseignements : les cartes sont réalisables...

La cartographie génétique a beaucoup progressé, et une carte comportant en moyenne un marqueur polymorphique tous les centimorgans constitue maintenant un objectif réaliste [1] et d'ailleurs presque atteint pour certaines régions [2]. Mais les progrès les plus sensibles concernent l'obtention de cartes physiques complètes, fondées sur du matériel cloné (cosmides et surtout YAC) et couvrant l'ensemble d'un chromosome. Celles-ci sont bien avancées pour le 7, le 11, le 16, l'X, le 19 ; elles seront sûrement obtenues rapidement pour les petits chromosomes comme le 21 ou le 22 ; puis, dans un deuxième temps, pour d'autres. La division du travail par chromosomes n'est pas artificielle dans ce cas, car il y a effectivement beaucoup d'avantages d'échelle à étudier l'ensemble d'un chromosome donné, et l'aboutissement est une collection de clones extrêmement utile pour toute étude portant sur une région particulière. Notons que la carte physique complète d'un chromosome moyen représente plusieurs années de travail pour un laboratoire d'une trentaine de personnes, et une dépense d'une dizaine ou une vingtaine de millions de dollars en l'état actuel des techniques. Dans toute ces entreprises, les YAC ont pris aux cosmides une belle « part du marché ». La banque générale établie à Saint Louis a maintenant été reproduite à une dizaine d'exemplaires et mise en œuvre dans chaque *Genome center* ; l'effort des équipes porte sur l'amélioration des méthodes de criblage (criblage tout PCR ou sur *pools* de clones séparés en champs pulsés) et d'ana-

lyse (clonage d'extrémité par *vector PCR*, par recombinaison homologue ou par circularisation, obtention de sondes internes par Alu-PCR ou Line-PCR... [3] plus que sur la construction de nouvelles librairies ; sur ce point, l'Europe semble un peu en avance avec les banques établies par Rakesh Anand [4], le CEPH [5] et Tony Monaco [6]. Cela n'empêche pas d'autres équipes de travailler au développement de nouveaux vecteurs, en essayant en particulier d'assurer la propagation de grands segments d'ADN exogène dans des bactéries, plus faciles à manipuler que la levure [7].

... L'ADNc a le vent en poupe...

Il y a toujours eu de bonnes raisons de vouloir privilégier l'étude de l'ADNc par rapport au séquençage d'ADN génomique « tout venant » : accès direct aux gènes, séquence « utile », travail débouchant sur la fonction... mais il existait un sérieux obstacle technique : la représentation très inégale des différents ARNm dans toute banque d'ADNc. Cette difficulté semble maintenant résolue, puisqu'au moins deux laboratoires ont démontré que l'on pouvait « normaliser » ou « égaliser » ces banques (c'est-à-dire faire en sorte que les différentes séquences y soient à peu près également représentées). L'analyse systématique des ADNc, qui était déjà le point fort des programmes Génome britannique, français et japonais, devient très à la mode, et le DOE a lancé récemment un appel d'offres sur ce thème. Il s'avère d'ailleurs qu'un séquençage systématique peut fournir des informations très instructives même si la banque ADNc que l'on emploie n'est pas normalisée [8].

... L'informatique est très importante !

Un *Genome center* typique comporte une équipe d'informatique de quatre à huit personnes, beaucoup plus que ce que l'on trouve au niveau de la plus grosse de nos unités Inserm ou CNRS ; la *Genome Data Base* (GDB) installée par Peter Pearson à Baltimore regroupe une trentaine d'informaticiens et de gestionnaires, avec un budget de plusieurs millions de dollars par an ; et la station de travail est en passe de remplacer le MAC haut de gamme ou le PC/AT dans les laboratoires et sur le bureau des chercheurs. J'ai exposé récemment les raisons de cette débauche d'informatique [9] ; je répéterai simplement ici que l'amélioration des techniques et leur début d'automatisation aboutissent à multiplier les résultats, les clones, les données ; que certaines de ces méthodes font elles-mêmes de plus en plus appel à l'informatique pour la saisie directe des données et leur manipulation ultérieure ; que l'archivage de ces résultats et leur diffusion à la communauté scientifique selon des modalités précises et après une procédure de validation à définir revêtent une grande importance [10] Il y a donc beaucoup de tâches à assumer, et les moyens débloqués, aux États-Unis, sont à la hauteur – ce qui ne veut pas dire que tout aille pour le mieux : le fameux fossé entre informaticiens et biologistes est encore trop réel et la communication entre ces deux cultures très différentes ne se fait pas sans heurts. Mais dans l'ensemble les choses avancent, les carnets de laboratoire informatisés commencent à fonctionner, les bases de données privées, semi-privées et publiques se mettent en place ; l'accès à ces banques est assurément une des conditions d'un travail efficace en Génétique humaine.

Et la France ? *(bis)* [11]

À la fin de ce bilan somme toute largement positif se pose inévitablement la question : qu'en est-il en France ? Cela d'autant plus que notre pays, qui occupe actuellement un honorable troisième rang mondial dans le domaine, vit une situation complexe. D'un côté des laboratoires « académiques » corsetés dans un cadre institutionnel (Inserm ou

CNRS) très rigide, avec une gestion de personnel extrêmement lourde et une marge de manœuvre limitée ; de l'autre des structures très « débridées » (sur le plan des moyens et de la gestion), le CEPH et le Généthon ; au milieu un « GIP génome » encore en devenir, tiraillé entre ses différents partenaires et qui a bien du mal à se constituer... Il y a beaucoup à dire, trop pour le faire ici : ce sera l'objet d'une prochaine chronique !

Références

1. Jordan BR. Les yeux plus gros que le ventre ? *médecine/sciences* 1990 ; 6 : 576-9.
2. Petersen MB, Slaugenhaupt SA, Lewis JG, Warren AC, Chakravarti A, Antonarakis SE. A genetic linkage map of 27 markers on human chromosome 21. *Genomics* 1991 ; 9 : 407-19.
3. Jordan BR. La montée en puissance des YAC. *médecine/sciences* 1990 ; 6 : 470-2.
4. Anand R, Riley JH, Smith JC, Markham AF. A 3,5 genome equivalent multi access YAC library : construction, characterisation, screening and storage. *Nucleic Acids Res* 1990 ; 18 : 1951-6.
5. Albertsen HM, Abderrahim H, Cann HM, Dausset J, Le Paslier D, Cohen D. Construction and characterization of a yeast artificial chromosome library containing seven haploid human genome equivalents. *Proc Natl Acad Sci USA* 1990 ; 87 : 4256-60.
6. Larin Z, Monaco AP, Lehrach H. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc Natl Acad Sci USA* 1991 ; 88 : 4123-7.
7. Jordan BR. Des vecteurs de clonage à la pelle. *médecine/sciences* 1991 ; 7 : 503-4.
8. Adams P, *et al.* Complementary DNA sequencing : expressed sequence tags and Human Genome Project. *Science* 1991 ; 252 : 1651-6.
9. Jordan BR. Informatique et biologie : condamnés à s'entendre ? *médecine/sciences* 1991 (sous presse).
10. Aldhous P. Human genome databases at the crossroads. *Nature* 1991 ; 352 : 94.
11. Jordan BR. Programme Génome : et la France ? *médecine/sciences* 1990 ; 6 : 807-9.
12. Jordan BR. Génome au Japon : au-delà des mythes. *médecine/sciences* 1991 ; 7 : 851-3.

Med Sci *1992 ; 8 : 163-166*

GRANDE-BRETAGNE : UN PROGRAMME GÉNOME À DIMENSION HUMAINE

Le projet britannique, lui, donnait une forte impression de sérieux et de pragmatisme, et allait en effet produire d'excellents résultats.

La Grande-Bretagne, un terrain privilégié

La Grande-Bretagne bénéficie d'une forte tradition de génétique humaine. La qualité de ses chercheurs et cliniciens est connue ; son système de santé étatisé a considérablement facilité les travaux sur l'épidémiologie, et par ailleurs ce pays est le berceau de la Biologie moléculaire. C'est là en effet qu'a été mis au point le séquençage des protéines, qu'a été découverte la structure de l'ADN ; et les méthodes modernes d'étude des gènes et de leur régulation ont leur origine au Royaume-Uni au moins autant qu'aux États-Unis. Les relations des Grands-Bretons avec leurs cousins d'outre-Atlantique sont intenses et privilégiées, mais – malgré un certain exode des cerveaux souvent déploré – n'ont pas abouti à une situation de type colonial : si beaucoup de chercheurs partent aux États-Unis, beaucoup aussi reviennent ensuite dans leur pays d'origine, et les liens établis sont à l'avantage des deux parties. En dépit d'un financement modeste, la recherche biologique britannique tire fort bien son épingle du jeu et présente indiscutablement un rapport qualité/prix excellent. Il est donc particulièrement indiqué d'examiner sa situation dans le domaine de la recherche sur le Génome et de comparer ses réalisations à celles du programme développé aux États-Unis.

Le *Human Gene Mapping Project* : modeste mais pragmatique

C'est au début de 1989 que la Grande-Bretagne a officiellement annoncé son programme Génome, intitulé *Human Genome Mapping Project* (HGMP). Un montant de 11 millions de livres (110 MF) était alloué pour trois ans au *Medical Research Council* (MRC) à cet effet, cette période devant être suivie d'une augmentation annuelle de 45 MF au budget du MRC pour soutenir le même thème mais sans qu'il constitue un chapitre à part. Les priorités annoncées concernaient la cartographie génétique et physique de « régions génomiques intéressantes », le séquençage d'ADNc, les organismes modèles (souris et nématode) et les évolutions méthodologiques. Ce programme devait être mis en œuvre par la création d'un centre de ressources *(Resource Centre)* destiné à aider les équipes en gérant des collections de sondes, les librairies de YAC *(yeast artificial chromosomes)*, et, par ailleurs, par un ensemble de contrats aux laboratoires existants, le *Directed Programme of Research*.

Le programme avait été élaboré après des discussions serrées. Le MRC (incarné sur ce thème par Sydney Brenner), et le puissant *Imperial Cancer Research Fund*, dirigé par Walter Bodmer et déjà bien impliqué dans les travaux sur le Génome s'en étaient, semble-t-il, disputé le *leadership*. Quoiqu'il en soit la question fut tranchée par le ministre de l'Éducation et de la Science qui confia cette entreprise au MRC ; Tony Vickers, un

scientifique investi depuis plusieurs années dans l'administration de la recherche, en fut nommé directeur et se consacra à sa mise en place. Les premiers contrats furent attribués dans le courant de l'année 1989, le *Resource Centre* commença à fonctionner en 1990, et le « petit journal » du programme britannique, appelé *G-String* puis *G-nome News*, en est déjà à son neuvième numéro. On souhaiterait que chez nous les déclarations ministérielles soient suivies d'effets aussi immédiats...

Le *Resource Centre* : une structure au service des laboratoires

C'est une des originalités du programme Génome britannique que de comporter un *Resource Centre*, un laboratoire central de service destiné à faciliter le travail des équipes : la finalité est analogue à celle du « Généthon » créé par l'Association Française contre les Myopathies, mais les installations sont nettement plus modestes. Ce centre, installé dans la banlieue ouest de Londres, regroupe une vingtaine de personnes et bénéficie d'un financement, salaires compris, d'environ un million de livres par an (10 MF) : sa taille et ses ressources sont donc comparables à celles d'un *Genome Center* aux États-Unis, même si l'objectif est un peu différent. Ses cinq services s'occupent d'informatique, d'ADNc, de YAC, de sondes et amorces, et enfin de la cartographie du génome de la souris. Les prestations sont offertes gratuitement aux 550 usagers déjà inscrits, dont on attend essentiellement un retour d'informations. Le personnel est jeune ; le directeur pour la Biologie, Ross Sibson, a travaillé quelques années dans l'industrie et en particulier chez Amersham. Tony Vickers, qui dirige l'ensemble du programme britannique, suit de près les activités du Centre.

Le groupe d'informatique poursuit des buts multiples : formation des utilisateurs, mise en réseau de l'ensemble des moyens de calcul, connexion aux réseaux internationaux, et mise au point de logiciels spécialisés. Il est très orienté vers l'usager, et l'on note une agréable absence de sectarisme au niveau des choix « politiques » : le responsable, F. Rysavy souhaite par exemple offrir l'accès à Genatlas et considère que c'est le taux de consultation qui doit décider de l'avenir d'une base de données (plutôt qu'une décision bureaucratique *a priori*). Le groupe des ADNc s'est attaché au séquençage partiel d'un grand nombre de ces entités, selon la tactique de la « signature » (3 ou 400 bases de séquence fiable à 99 % sur chaque clone) popularisée aux États-Unis par Craig Venter [1] mais proposée en fait par Sydney Brenner depuis plusieurs années. La première année de fonctionnement de ce service de taille modeste a vu le traitement de 6 000 clones qui ont fourni 1 500 séquences dont la moitié sont nouvelles (jusque-là inconnues). L'équipe des YAC offre un service de criblage pour la librairie de Saint Louis ainsi que celle de Rakesh Anand *(ICI pharmaceuticals)*[1] [2] ; le service des « sondes et amorces » gère une collection de sondes et la fourniture (gratuite) d'amorces ; le Centre, enfin, est partenaire dans l'entreprise européenne de cartographie du génome de la souris par la méthode des croisements interspécifiques et doit courant 1992 offrir un service de localisation de sondes par cette méthode.

En somme, le *Resource Centre* constitue une structure modeste, mais qui semble solide et surtout directement en prise avec les utilisateurs et leurs demandes. Il est dommage peut-être qu'elle ne se soit pas (pour le moment) plus impliquée dans des approches ambitieuses du type de celle développée par Hans Lehrach *(voir plus loin)* ; c'est sans doute la rançon de son souci de répondre aux demandes immédiates des équipes, sans leur imposer une stratégie *a priori*.

1. Cette banque semble toujours détenir le ruban bleu de la plus faible proportion de clones chimériques : moins de 5 %.

Un tissu de recherche dense et vivace

Il est difficile dans une chronique nécéssairement brève de décrire une réalité aussi riche que celle de la recherche en Grande-Bretagne dans ce secteur ; les quelques cas évoqués ci-dessous le sont à titre d'exemple, pour montrer les spécificités de la génétique moléculaire dans ce pays.

Hans Lehrach (ICRF, Londres) et les bibliothèques de référence

ICRF *(Imperial Cancer Research Fund)* est une fondation privée qui collecte des sommes considérables : environ 400 MF par an, plus donc que l'AFM ou l'ARC en France, et qui emploie environ 1 700 personnes dont un bon millier directement impliquées dans la recherche. Son laboratoire principal, installé *41 Lincoln's Inn Fields* à Londres, regroupe une cinquantaine d'équipes dont plusieurs sont concernées par le génome. Le groupe de Hans Lehrach est celui qui se rattache le plus à notre sujet. C'est un chercheur qui a beaucoup réfléchi aux stratégies expérimentales, en a évalué de façon aussi quantitative que possible les performances et la fiabilité, et a développé un ensemble de méthodes originales, très différentes de celles mises en œuvre ailleurs et en particulier aux États-Unis. Il parie en effet sur les « bibliothèques de référence » : des banques d'ADN réalisées dans différents vecteurs (cosmides ou YAC), ordonnées ensuite en plaque à microtitration puis exploitées grâce à des filtres à haute densité contenant 10 ou 20 000 colonies disposées régulièrement à l'aide d'un robot. Des jeux de filtres identiques sont distribués aux laboratoires extérieurs qui peuvent les hybrider avec les sondes de leur choix, repérer la ou les colonies positives et renvoyer l'information au laboratoire de Londres qui leur fournit alors les clones correspondants. Cette manière de procéder présente le double avantage de faire réaliser l'étape la plus délicate, l'hybridation, par le laboratoire extérieur (particulièrement motivé par l'obtention du clone), et de permettre une accumulation toute naturelle des informations au laboratoire central. Les données ainsi receuillies sont complétées par l'hybridation des mêmes filtres avec des sondes complexes (*pools* de clones, ADNc total...) ou peu spécifiques (oligonucléotides) de façon à récupérer le plus grand nombre de données possible par expérience [3]. Si l'on compare cette méthode à celle des STS *(sequence tagged sites)* généralement pratiquée aux États-Unis, on ne peut qu'être frappé par ses attraits : la stratégie des librairies de référence remplace les STS par les clones de la librairie, beaucoup plus directement utilisables, tout en assurant une mise en mémoire naturelle des données obtenues au fur et à mesure ; elle est utile aux autres laboratoires tout au long de sa mise au point et non uniquement une fois qu'elle a abouti à une carte complète. L'approche est remarquablement efficace et sans faille – si toutes les expériences envisagées marchent réellement en routine. C'est un peu sur ce terrain que l'on attend Hans Lehrach, car les techniques employées sont délicates, et certains ne croient pas à leur fiabilité dans le cadre d'un emploi quotidient. Mais l'équipe a marqué plusieurs points importants ces derniers temps : par exemple la démonstration de la validité de la méthode de *fingerprinting* de toute une série de clones par hybridation en filtre à haute densité avec un jeu d'oligonucléotides [4]. Le groupe, une petite trentaine de personnes, effectue un travail considérable avec un financement total d'environ 7,5 MF par an : moins qu'un *Genome center* aux États-Unis, plus évidemment qu'une unité Inserm de cette taille en France...

Les cosmides à l'anglaise : les nématodes...

On connaît l'histoire du nématode *(Caenorhabditis elegans)* sur lequel Sydney Brenner s'est beaucoup penché par le passé. Ce n'est pas en principe notre tasse de thé puisque ces *Chroniques Génomiques* concernent essentiellement le génome humain...

mais cet organisme dont le génome mesure environ 100 mégabases, soit une taille comparable à celle d'un chromosome humain moyen, constitue un remarquable banc d'essai pour les méthodes, et c'est à ce titre que nous en parlons. Alan Coulson et John Sulston se sont attelés depuis 1984 à l'établissement de sa carte physique complète ; ils n'ont pas choisi la facilité puisqu'au lieu de se contenter d'une carte de restriction ils se sont engagés dans l'analyse systématique des recouvrements de cosmides afin d'établir un *contig* (série de cosmides contigus) couvrant entièrement le génome.

La carte physique est en fait pratiquement terminée [5]. L'assemblage des cosmides a été effectué par la méthode des *fingerprints*, mise au point par Sulston ; les « trous » inévitables ont été ensuite comblés par des YAC isolés d'une librairie construite par Bob Waterston (St-Louis, MO, USA). 17 000 cosmides au total ont été étudiés et leurs recouvrements analysés par un système informatique mis au point par Sulston. Les YAC, eux, ont servi de sonde pour la librairie de cosmides, et réciproquement, afin d'établir des ponts pour relier les *contigs*. Tout ceci a été fait, pour l'essentiel, à la main : de l'artisanat donc, comparé aux ressources et au personnel des centres du DOE, mais un artisanat efficace, très bien organisé, et qui atteint ses objectifs...

Il faut dire que la communauté du nématode semble autrement plus coopérative que celle de la génétique humaine, et que les interactions dans les deux sens ont été nombreuses, confiantes et fréquentes. Le laboratoire se mobilise actuellement sur le début du séquençage du génome de *C. elegans*, avec des objectifs ambitieux mais réalistes : trois mégabases de séquence au centre du chromosome 3 (région riche en gènes) en trois ans, avec le groupe de Bob Waterston à Saint Louis (MO, USA). Ce programme de méga-séquençage de deuxième génération paraît avoir de fortes chances de réussir contrairement à certaines tentatives précédentes [6].

... et le chromosome 11

Dans la même mouvance technique, mais en un lieu différent et, cette fois, sur du matériel humain, l'équipe de Peter Little (*Imperial College of Science, Technology and Medicine*, Londres) effectue une analyse systématique du bras court du chromosome 11. Les cosmides provenant d'une librairie réalisée à partir d'un hybride somatique sont analysés selon une procédure très semblable à celle de Coulson, mais quelques étapes sont réalisées de façon semi-automatique à l'aide de l'ubiquitaire robot Beckman Biomek. L'efficacité paraît comparable à celle des entreprises analogues menées aux États-Unis à Lawrence Livermore (CA) et Los Alamos (NM) en ce qui concerne la détection des recouvrements et la croissance des *contigs* ; dans les deux cas, il est clair qu'il faut passer par les YAC pour achever la carte. Mais la modestie des moyens mis en œuvre ici est remarquable : trois ou quatre personnes et 1,5 MF par an ; une fois encore le rapport qualité/prix de la recherche britannique paraît excellent.

Édimbourg prépare l'après-gène

Édimbourg, c'est (entre autres) une importante unité du MRC *(MRC Human Genetics Unit)*, où l'on retrouve des noms connus : Nick Hastie, Howard Cooke, David Porteous, Robin Allshire... On se souvient peut-être aussi qu'ici travaillait Ed Southern (avant d'émigrer à Oxford), dans une équipe qui portait à la fin des années soixante-dix le nom de *Mammalian Genome Unit* : sans doute le premier laboratoire à comporter dans son titre le mot « Génome », qui a depuis fait fortune !

L'Unité regroupe environ 200 personnes et son orientation générale fait une large part au modèle souris ainsi qu'au ciblage génique et généralement aux approches fonctionnelles ; mais elle reste branchée sur la clinique et est également tout à fait au point

en ce qui concerne les technologies moléculaires les plus récentes. On y a développé entre autres des vecteurs de ciblage génique très performants, une banque de YAC du chromosome 11, ainsi qu'une série de chromosomes X « raccourcis » par intégration d'un télomère dont on imagine sans peine l'utilité pour de nombreux laboratoires. L'Unité comporte aussi un étonnant « conservatoire », le *MRC Human Genetic Registry* ; près de 5 000 personnes enregistrées, sur lesquelles plus de 1 000 médecins ou administratifs fournissent annuellement des informations. Deux dames y tiennent (à la main et sur support papier) de grands arbres généalogiques regroupant parfois plus de cent personnes et comportant toute sorte de détails sur chacune d'elles. C'est un outil de travail certainement inappréciable, géré avec un sens pratique tout britannique ; mais qu'en dirait chez nous la redoutable Commission Informatique et Libertés ?

Il s'agit là sans doute d'une des meilleures unités du MRC. Moins connue qu'un laboratoire comme celui d'ICRF à Londres, ou que le *Laboratory of Molecular Biology* de Cambridge, elle mérite l'attention ; n'oublions pas qu'à Édimbourg se trouve aussi Adrian Bird, revenu après quelques années très productives à l'*Institute of Molecular Pathology* à Vienne, et dont les travaux sur les îlots CpG font autorité...

« *Rule Britannia !* » ?

Règne sans partage, comme à la plus belle époque de l'Empire Britannique ? On peut du moins légitimement conclure que la Grande-Bretagne a fort bien réussi le démarrage de son programme Génome : malgré un budget de recherche modeste, dans des laboratoires souvent équipés de façon spartiate et où, visiblement, on fait attention au moindre microlitre d'enzyme, ce sont fréquemment des travaux excellents qui sont réalisés. Intelligence (la densité de penseurs au mètre carré est remarquable), pragmatisme, originalité (une caractéristique personnelle très prisée outre-Manche), ce cocktail s'avère d'une redoutable efficacité. Et cet exemple tout proche gagne à être médité : il est sans doute plus facile à transposer chez nous que les mastodontes du DOE, et aucune fatalité ne condamne notre pays à rester – comme c'est le cas actuellement – nettement derrière nos collègues d'outre-Manche...

Références

1. Adams MD, Kelley JM, Gocayne JD, *et al.* Complementary DNA sequencing : expressed sequence tags and Human Genome project. *Science* 1991 ; 252 : 1651-6.
2. Anand R, Riley JH, Butler R, *et al.* A 3,5 genome equivalent multi access YAC library : construction, characterisation, screening and storage. *Nucleic Acids Res* 1990 ; 18 : 1951-6.
3. Lehrach H. Hybridization fingerprinting in genome mapping and sequencing. *Genome Analysis* 1990 ; 1 : 39-81.
4. Craig AG, Nizetic D, Hoheisel JD, Zehetner G, Lehrach H. Ordering of cosmid clones covering the herpes simplex virus type I (HSV-I) genome : a test case for fingerprinting by hybridisation. *Nucleic Acids Res* 1990 ; 18 : 2653-60.
5. Coulson A, Kozono Y, Lutterbach B, *et al.* YACs and the *C. elegans* genome. *Bioessays* 1991 ; 18 : 413-7.
6. Jordan BR. Les heurs et malheurs du séquençage à grande échelle. *médecine/sciences* 1991 ; 7 : 612-3.

Les impressions positives rapportées dans cette chronique ont été dans l'ensemble confirmées. Le HGMP Resource Centre *a été l'un des rares centres de ressources à être la fois opérationnel et réellement au service des utilisateurs. Il a beaucoup fait pour diffuser les méthodes génomiques dans l'ensemble du tissu de recherche britannique. Notons en passant qu'ici encore je présente à tort le Généthon comme une structure de*

service... Le pronostic sur le succès du mégaséquençage lancé sous la houlette de John Sulston s'est révélé exact : ce travail a donné naissance au Sanger Centre *qui fut, durant plusieurs années, le plus productif de centres de séquençage mondiaux. Seule l'entreprise de Hans Lehrach, malgré son incontestable séduction intellectuelle et les moyens importants qui lui ont été consacrés à ICRF et, plus tard, dans le cadre du programme Génome allemand, n'a pas eu le retentissement attendu et a finalement assez peu contribué à l'établissement des cartes du Génome. Le programme Génome britannique a de plus bénéficié du soutien de la puissante* Wellcome Trust*, qui a pris de belle manière le relais du* Medical Research Council *lorsque ce dernier a commencé à diminuer ses financements : cela a assuré au programme britannique une continuité qui a beaucoup manqué en France...*

Med Sci *1993 ; 9 : 460-464*

OÙ EN EST LE PROGRAMME GÉNOME RUSSE ?

Dans cette chronique, je rapportais les ambitions et les moyens non négligeables (à ses débuts) d'un programme Génome soviétique assez mal connu en Occident. Il avait démarré en fanfare à l'époque de Gorbatchev, bénéficiant de financements en dollars américains, et avait richement équipé en matériel moderne certains laboratoires. Mais, en 1993, date de mon séjour, le programme Génome russe et, plus encore, la recherche biologique étaient déjà en piteux état. Pourtant, quelques équipes survivaient, et l'on pouvait espérer une remise en route progressive du système avec l'aide de collaborations bien choisies.

La renaissance de la génétique en URSS

Après des débuts prometteurs à l'époque de Vavilov ou de Koltsov, la génétique en Union soviétique avait quasiment disparu durant l'ère de Lyssenko – cette phase d'obscurantisme devait se prolonger même sous Kroutchev, jusque vers le milieu des années 1960. La génétique et la biologie moléculaire étaient toutes deux très mal vues durant cette longue période au cours de laquelle – selon une interprétation bien grossière du marxisme – la part de l'inné, du génétique devait à tout prix être niée. Seuls comptaient l'acquis, le culturel : c'est ainsi que l'on espérait faire du passé table rase pour construire la société socialiste qui devait accoucher de « l'homme nouveau ». Nous savons comment ce rêve s'est transformé en cauchemar...

La renaissance devait être lente et progressive. Un des signes avant-coureurs, en 1959, fut la fondation à Moscou d'un *Institute of Radiation Biology*, qui allait devenir plus tard l'Institut de biologie moléculaire. Son directeur, Victor Engelhardt, venait de Sibérie, et semble avoir été un homme de caractère. Il se battit avec succès pour obtenir la reconnaissance d'une science alors idéologiquement suspecte. Il devait même réussir à faire soutenir une thèse à un de ses anciens élèves, Alexandre Bayev, qui n'avait pourtant pas fini de purger une longue peine de prison (dix-sept ans derrière les barreaux). Il avait eu la malchance de suivre des cours de philosophie politique avec un proche de Boukharine, avant la disgrâce, puis la liquidation de ce dernier. Encore Bayev eut-il bien de la chance puisque la plupart de ses condisciples furent, eux, fusillés pour le même crime...

Ce même Bayev entama, à 50 ans passés, une brillante carrière scientifique ; il fut, en même temps que Holley aux États-Unis et Sanger en Grande-Bretagne, un des premiers à aborder le séquençage de l'ARN. Il travaillait, comme ses collègues américains et britanniques, sur les ARN de transfert, les seules molécules assez petites – avec l'ARN ribosomique 5S – pour être accessibles à des méthodes très laborieuses. Respecté par ses collègues, autant, sans doute, pour ses qualités scientifiques qu'en raison de son histoire personnelle, il devenait bientôt membre de la puissante Académie des Sciences de l'URSS. C'est lui qui, en 1988, convainquit Gorbatchev de lancer un programme Génome humain fort bien doté en roubles et même en dollars. Son élève Andrei Mirzabekov, nouveau directeur de l'Institut de biologie moléculaire après la disparition d'Engelhardt, devenait vice-président de HUGO, et un bureau de cette organisation était prévu

à Moscou. Bref, le programme Génome humain de l'URSS débutait sous les meilleurs auspices, à l'époque de la détente, de la *glasnost* et de la *perestroïka*, et l'on pouvait attendre beaucoup de la créativité enfin libérée des scientifiques soviétiques.

L'environnement russe en 1993

Trois ans plus tard, tout a bien changé. L'URSS a disparu, la fantomatique CEI n'a pas d'existence réelle, et la Russie, coupée de son ancien empire et dépouillée de son statut de grande puissance, s'enfonce dans une crise dont l'on ne perçoit pas l'issue. Ce n'est pas le chaos, et cette société marche encore à peu près, du moins pour autant que puisse en juger un étranger y passant quelques jours en ce début d'année 1993 : les services publics fonctionnent, les bâtiments sont chauffés, les boutiques ne sont pas vides et les trains partent à l'heure. Certes, les trous dans la chaussée sont nombreux, comme les mendiants dans les rues et le métro, mais ce n'est pas Calcutta ni même Alger. Pourtant, les distorsions sont évidentes et l'adaptation à l'économie de marché semble loin d'être réalisée. La structure des prix paraît aberrante à un Occidental : une place de concert coûte le prix d'une demi-orange, la traversée de Moscou en taxi équivaut au salaire mensuel d'un chercheur « senior ». Si l'on traduit les prix en devises (à la mi-janvier, 500 roubles s'échangent contre un dollar), on arrive a des chiffres complètement surréalistes : rémunération minimum à 20 francs, 100 francs par mois pour un scientifique « senior ». Une inflation galopante (20 % par semaine...) ne facilite pas les rééquilibrages indispensables, et l'instabilité politique est extrême ; personne ne se hasarde à faire de pronostics sur la durée de vie du système politique actuel, même exprimée en mois. Dans une telle situation, il faut avoir la foi chevillée au corps pour privilégier le long terme et parier sur l'avenir...

Dans ce contexte difficile, la recherche ne constitue apparemment pas une priorité. Le statut des chercheurs, il y a peu de temps élite privilégiée de la société – du moins pour ceux qui n'étaient pas trop contestataires –, a beaucoup décliné ; leurs salaires ne suivent pas, même de loin, la courbe vertigineuse des prix. Ils sont atteints aussi dans leurs moyens de travail : l'approvisionnement des laboratoires en réactifs devient très difficile lorsque ceux-ci – c'est souvent le cas – proviennent de l'Ouest et doivent être payés en devises. Dans beaucoup d'instituts, l'absence de quelques produits (enzymes de restriction, molécules radioactives, marqueurs fluorescents...) ou de pièces de rechange indispensables aboutit au chômage technique – d'autant plus que les scientifiques, pour simplement survivre, sont amenés à consacrer une bonne part de leur temps à se procurer des denrées de base ou à exercer une autre activité pour gonfler un peu leur maigre budget.

Pas étonnant que, dans ces conditions, les frontières étant maintenant ouvertes, l'émigration devienne une tentation très forte. De nombreux scientifiques ont ainsi quitté la Russie pour les États-Unis ou l'Europe. Partis parfois pour quelques mois, ils trouvent à l'Ouest des conditions de travail et de vie très supérieures – malgré quelques désillusions – à celles de leur pays. Lorsque arrive le moment prévu pour le retour, beaucoup décident de rester dans leur lieu d'accueil, au moins pour quelques années. On peut ainsi craindre la disparition de pans entiers de la recherche russe.

Les ambitions du programme Génome russe

Le programme Génome humain russe existe pourtant, avec sa structure, ses objectifs et des crédits non négligeables – même s'il est malaisé d'apprécier à quoi ils correspondent concrètement aujourd'hui. Lancé en 1989, il a bénéficié dès le départ de fonds considérables : 25 millions de roubles – censés a l'époque valoir autant de dollars US – et 5 millions de dollars en devises pour les achats à l'étranger. Cela représentait

un effort comparable à celui consenti par l'État en France, l'année dernière – pour un produit national brut qui est à peu près du même ordre. Programme structuré en sept chapitres dont les intitulés rejoignent ceux que l'on trouve ailleurs (cartographie génétique, cartographie physique, séquençage, informatique, applications cliniques...), il était géré par autant de comités d'experts. L'ensemble de l'organisation était basé à l'Institut de biologie moléculaire de Moscou ; 300 contrats furent ainsi attribués à travers l'URSS dans un premier temps, à travers la Russie par la suite.

À sa tête, l'on retrouve l'académicien Bayev – toujours vert à 89 ans, malgré les épreuves qu'il a subies, et plusieurs responsables dont beaucoup appartiennent à cet institut. Citons notamment Andrei Mirzabekov, promoteur du séquençage d'ADN par hybridation, ainsi que des personnages honorablement connus à l'Ouest comme Lev Kisselev, Alexandre Zelenin ou encore Eugene Sverdlov de l'Institut de génétique moléculaire de Moscou. Le programme couvre une vaste gamme de sujets, de la génétique clinique aux nouvelles méthodes de séquençage, mais est en principe centré sur quatre chromosomes (3, 5, 13 et 19) ; il fait état de nombreuses collaborations avec les laboratoires américains, principalement ceux du DOE *(Department of Energy)*. Est-ce parce que la structure un peu bureaucratique de ce dernier organisme convient mieux aux habitudes russes que celle, plus ouverte et plus compétitive, du NIH ? Toujours est-il que Charles Cantor, homme du DOE, est plus souvent cité que Jim Watson, et qu'on se réfère à ses stratégies expérimentales plus qu'à celles prônées par Francis Collins, Éric Lander ou Ray White.

Le budget effectif du programme, en 1992, a été de 130 millions de roubles. Il est extrêmement délicat d'évaluer ce que représente cette somme. Si l'on se base sur le prix de vente d'une machine à PCR russe, 50 000 roubles pour un appareil qui, sous cette forme, vaut à peu près 20 000 francs en France, la somme représente environ 50 millions de nos francs. Mais si on la convertit au taux actuel du change (un franc égale 100 roubles), on arrive à peine à plus d'un million... Selon l'étalon choisi, on a ainsi un montant très décent (compte tenu de la situation du pays) ou une misère : seul un examen de la réalité des laboratoires est susceptible de nous éclairer.

Qu'en est-il sur le terrain ?

Une dizaine de jours en janvier 1993, employés à la visite de cinq laboratoires à Moscou et de trois autres à Saint-Petersbourg, m'ont permis une première approche de cette réalité.

L'institut que dirige Mirzabekov à Moscou, appelé indifféremment Institut de biologie moléculaire ou Institut Engelhardt, du nom de son ancien directeur, est indubitablement le laboratoire le plus actif parmi ceux que j'ai vus. Il occupe un immense bâtiment de style « classico-stalinien » avec colonnades, statues et bas-reliefs, construit à l'origine pour abriter un Institut de géologie et de recherches minières. À l'initiative de Khrouchtchev, qui pensait qu'un laboratoire de recherches minières devait se trouver à proximité des mines et non au centre de Moscou, le bâtiment fut attribué aux biologistes. Visiblement, son architecte n'avait aucune idée de ce qu'est la recherche en géologie ou en biologie : c'est une suite de grandes pièces au sol parqueté, aux murs de pierre épais, dans lesquelles il a fallu rajouter à grand peine fluides, conduits et alimentations électriques. L'ensemble est vieillot et assez dégradé, mais il y règne une activité certaine, même si les 500 personnes annoncées sur l'organigramme ne sont pas toutes présentes en même temps.

L'équipe d'Andrei Mirzabekov se consacre à l'étude de la structure de la chromatine, ainsi qu'à la mise au point du séquençage de l'ADN par hybridation, technique dont il est l'un des apôtres avec l'ex-Yougoslave Radomir Crkvenjakov (maintenant à Argonne, aux États-Unis) et Edwin Southern, à Oxford. Le principe est commun aux trois équipes :

il s'agit de fabriquer une matrice sur laquelle sont fixés des milliers d'oligonucleotides de séquence connue, puis d'hybrider sur cette dernière l'ADN à séquencer, préalablement marqué par un réactif fluorescent. À partir de l'image qui en résulte, comportant un point positif pour chaque oligonucleotide qui a retenu un fragment de la sonde parce qu'il lui est complémentaire, des algorithmes permettent de déduire la séquence.

Cette méthode se heurte aux différences de stabilité des hybrides selon la séquence de l'oligonucléotide : ceux qui sont riches en A et en T forment des doubles brins dont la température de fusion est basse, au contraire des séquences comportant beaucoup de G et de C. Or, il n'est évidemment pas question de traiter différentes zones de la matrice à des températures différentes... Le procédé employé par Mirzabekov est astucieux : il fixe ses oligonucléotides non à la surface d'une lame, mais dans de tout petits blocs d'acrylamide. La réaction d'hybridation a ainsi lieu à l'intérieur de l'acrylamide, donc en présence d'une importante concentration locale d'oligonucléotide. Mirzabekov a montré que, dans ces conditions, la température de fusion dépend beaucoup de la concentration de l'oligonucleotide. Dès lors, il peut « ajuster » la température de fusion dans chaque bloc en y déposant une quantité d'oligonucleotide calculée en fonction de sa séquence, de façon à obtenir une température de fusion identique pour tous les hybrides formés sur la matrice.

L'autre point critique du séquençage par hybridation est qu'il devient réellement praticable seulement si l'on peut en miniaturiser l'élément central : une lame comportant tous les octanucléotides possibles, capable, d'après les simulations, de séquencer directement un fragment d'ADN long de cent bases, doit en porter 65 536 ! Il est impératif de les disposer sur une surface de petites dimensions, donc de réduire la taille de chaque élément. Pour ce faire, Mirzabekov s'est assuré la collaboration d'un ancien laboratoire militaire spécialisé dans la mécanique de précision et les dispositifs optoélectroniques ; deux machines réalisées par cette équipe permettent de fabriquer des *oligonucleotide chips* comportant des blocs de 80 micromètres de côté, espacés de 20 micromètres (ce qui correspond à une densité de 10 000 éléments par centimètre carré), et de les « charger » individuellement en oligonucléotides. Naturellement, le résultat de l'hybridation se lit à l'aide d'un microscope couplé à une caméra et à un ordinateur. Mirzabekov espère, par la suite, se rapprocher des techniques employées en informatique afin de poursuivre la miniaturisation. Cela n'est pas forcement chimérique si l'on pense que les microprocesseurs actuellement installés dans les machines de bureau un peu performantes comportent, sur un ou deux centimètres carrés, plus d'un million de transistors...

Dans le même institut, plusieurs équipes travaillent à établir la carte physique du chromosome 3. L'approche choisie – dans laquelle il me semble reconnaître l'influence de Charles Cantor et de Cassandra Smith – repose sur les concepts de *linking* et de *jumping*, très à la mode à la fin des années 1980 – mais qui n'ont pas tenu toutes leurs promesses depuis, leur mise en œuvre se révélant un peu acrobatique. Le projet part d'un hybride somatique contenant le chromosome 3, à partir duquel sont établies une banque de *linking clones* – contenant chacun un site de coupure par l'enzyme de restriction à site rare Not I et les séquences adjacentes – et une banque de *jumping clones* dont chacun comporte les deux extrémités d'un grand fragment génomique produit par la même enzyme. Un séquençage systématique de tous ces clones permet alors de les assembler afin de constituer la carte physique du chromosome. La taille des fragments Not I est évaluée par des expériences complémentaires faisant appel à des analyses par électrophorèse en champs pulsés. Le séquençage de ces différentes entités est effectué pour partie sur place, dans le groupe de séquençage de l'institut dirigé par Vladimir Zakhariev, pour partie à Novosibirsk où se trouve une importante équipe réalisant du séquençage à façon.

La méthylation fréquente, mais incomplète et variable, des sites NotI dans l'ADN génomique est un des points faibles de cette stratégie. Elle empêche leur coupure par

l'enzyme, d'où des « trous » dans la carte – en tout état de cause bien moins opérationnelle et utile à l'ensemble de la communauté qu'une carte appuyée sur des fragments clonés (un *contig* de YAC, par exemple). Quoi qu'il en soit, le projet mené en collaboration avec le groupe de George Klein au *Karolinska Institute*, à Stockholm, est assez avancé ; un gros travail préalable a été effectué par un chercheur maintenant émigré, Eugene Zabarovky, sur la construction de vecteurs bien adaptés à l'établissement de telles banques (la série des vecteurs lambda SKN dont l'on dit ici le plus grand bien). Reste que cette entreprise doit logiquement subir, elle aussi, les contrecoups des récents résultats du Généthon, qui la rendent sans doute quelque peu caduque ; notons que, malgré la place d'Ilya Chumakov (originaire de l'*Institut Engelhardt* et même du groupe de Mirzabekov) dans ce travail, on ne semble pas en avoir mesuré ici toutes les retombées...

L'on trouve aussi à l'*Institut Engelhardt* une équipe d'informatique spécialiste du génome. Elle semble s'être principalement attachée à développer plusieurs bases de données tournant sur micro-ordinateur PC. Il y a en effet abondance de PC dans les laboratoires que j'ai visités, mais pas de *Macintosh* ni, sauf très rares exceptions, de stations de travail type SUN ou *Silicon Graphics*. Par ailleurs, les communications téléphoniques avec l'extérieur sont très difficiles, et il est impraticable pour le moment d'interroger à distance des bases de données comme GDB, OMIM, GBase... Les informaticiens russes se sont donc consacrés à la récupération des informations et à leur collationnement dans des logiciels adaptés au PC et ne réclamant qu'une capacité de stockage limitée grâce à d'astucieux algorithmes de compression. Les systèmes que j'ai pu essayer m'ont paru commodes et conviviaux ; ils témoignent d'un savoir-faire certain, mais souffrent d'une mise à jour effectuée une ou deux fois par an seulement, et n'apportent rien d'essentiellement nouveau par rapport à des *packages* commercialises dans nos pays. La créativité de leurs auteurs serait mieux employée à bâtir des systèmes originaux sur des machines modernes (le PC à processeur 286, au bâti massif et aux disquettes de 5''1/2, est roi ici), plutôt qu'à « réinventer la roue ».

Parlons un peu, pour terminer, du chromosome 19, un des quatre choisis par le programme russe. Certains travaux sont réalisés à l'Institut Engelhardt, dans le groupe d'Alexandre Zeleniev, mais le centre de gravité se situe plutôt autour d'Eugene Sverdlov. Ce dernier est simultanément responsable d'un département au *Shemiakin Institute of Bioorganic Chemistry*, et directeur de l'*Institute of Molecular Genetics*, dans la banlieue Nord de Moscou, à côté d'un institut de physique nucléaire (l'Institut Kurchatov) dont il faisait autrefois partie, ce qui le rendait inaccessible aux visiteurs pour cause de sécurité nationale. Le « Shemiakin » est le seul « beau » bâtiment de recherche qu'il m'ait été donné de voir en Russie : marbres, sculptures, plantes vertes, halls immenses et somptueux, il a été construit au début des années 1970 pour un universitaire puissant et très probablement atteint de mégalomanie avancée. Il n'est que de parcourir le bureau directorial dont a hérité l'actuel responsable, Vladimir Ivanov, pour en prendre conscience : avec ses antichambres et son secrétariat, l'ensemble occupe bien deux ou trois cents mètres carrés où tout n'est que boiseries, cuir fauve, verre épais et tube chromé. Les mauvaises langues disent que cet institut a absorbé la majeure partie de la dotation en devises de l'ensemble de la recherche civile soviétique, et l'appellent « le Taj Mahal de Shemiakin ». C'est en tout cas une digne relique de « l'ère de la stagnation », durant laquelle les hiérarques cacochymes du Kremlin multipliaient les constructions inutiles et somptueuses...

Pour en revenir au chromosome 19 – qui, rappelons le, est aussi le thème principal du *Genome Center* d'Anthony Carrano à Lawrence Berkeley, en Californie –, l'accent semble mis ici principalement sur les séquences exprimées. Plusieurs tactiques sont utilisées par les chercheurs installés au Shemiakin pour obtenir des banques d'ADNc spécifiques de ce chromosome. L'une part de l'ADN nucléaire d'un hybride somatique sur lequel un amorçage PCR employant les séquences Alu (présentes dans les introns, et donc dans le produit de transcription primaire qu'est l'ARN hétérogène nucléaire) devrait

sélectionner des séquences exprimées humaines. L'autre fait appel aux méthodes de soustraction entre la cellule hybride et une lignée de hamster pour sélectionner les clones correspondant aux séquences présentes uniquement dans le premier – en éliminant, naturellement, les séquences très conservées entre les deux espèces.

À l'*Institute of Molecular Genetics*, on se consacre plutôt aux approches génomiques, en particulier à la construction d'une banque de YAC spécifique du chromosome 19. Cela est réalisé à partir du même hybride somatique, mais avec une assez astucieuse méthode de tri « génétique » des clones humains, par recombinaison homologue dans la levure avec un plasmide portant une séquence Alu (donc susceptible de recombiner uniquement avec de l'ADN humain) et un marqueur métabolique permettant de sélectionner les levures qui l'ont acquis. Dans ces différents travaux – qui mettent en œuvre des schémas expérimentaux déjà publiés à l'Ouest –, les résultats quantitatifs (nombre de clones, taille...) paraissent convaincants, mais, curieusement, le contrôle de l'origine des clones (la vérification du fait qu'ils proviennent effectivement du chromosome 19) n'a été effectué que très ponctuellement – de sorte qu'il plane une certaine incertitude sur la qualité réelle des bibliothèques élaborées. Le laboratoire a des contacts et même une collaboration établie avec celui de Carrano – mais, cette fois encore, il ne semble pas avoir tiré toutes les conséquences des résultats récemment obtenus, notamment en France.

Des points communs

Les différents instituts visités, à Moscou comme à Saint-Petersbourg, sont tous bien, et même très bien, équipés, mieux, à vrai dire, que les meilleurs laboratoires publics de notre pays. Le matériel est abondant, récent, généralement d'origine occidentale. Le groupe de séquençage de l'Institut de biologie moléculaire de Moscou, par exemple, dispose de deux robots Biomek et de trois séquenceurs (un LKB/Pharmacia, deux *Applied Biosystems*). Partout des micro-ordinateurs PC, des ultracentrifugeuses Beckman, des systèmes d'analyse par HPLC... Visiblement les débuts du programme Génome, lorsque l'URSS existait encore et disposait de devises, ont permis un grand nombre d'achats. Les appareils sont récents, et la majorité paraissent être en état de fonctionnement ; seule l'informatique commence à dater un peu. Le cadre, en revanche, n'est pas à la hauteur et, à l'exception du *Shemiakin Institute* déjà cité, les locaux sont vieillots, dégradés et mal entretenus – ce qui est peu compréhensible, vu l'abondant personnel de service qui y est attaché.

En revanche, en ce qui concerne les réactifs, la situation est réellement critique. Les devises sont très rares, les prix occidentaux extraordinairement élevés par rapport aux crédits et aux tarifs locaux ; et du fait de la dislocation de l'URSS, des fournisseurs qui faisaient partie de cet ensemble (les Pays baltes pour les enzymes de restriction, l'Ukraine pour les éléments Peltier indispensables aux machines à PCR) exigent maintenant des devises – quand ils n'ont pas arrêté toute production pour cause de manque de matières premières ou de guerre civile. Certes, les Russes sont habitués à se « débrouiller » par leurs propres moyens, ils ont remis en route une production de nucléotides marqués au voisinage de Moscou et des laboratoires préparent eux-mêmes les réactifs nécessaires à la synthèse d'oligonucléotides ou à certaines techniques de « sondes froides »... mais la perte d'efficacité est très sensible.

Le départ à l'Ouest des chercheurs qualifiés est l'autre grand problème. La gravité de cet exode varie selon les lieux. Au dire des Russes, ce sont les meilleurs instituts qui ont perdu le plus de monde, puisqu'ils avaient le personnel le plus compétent. C'est souvent vrai, mais pas toujours : l'Institut Engelhardt semble avoir bien réussi à tirer son épingle du jeu. Ce n'est pas un désert, à la différence de certains laboratoires où l'on ne rencontre que de rares personnes, la plupart des chercheurs « senior » étant à

l'étranger... « en stage », dit-on pudiquement. Andrei Mirzabekov, lui, a appréhendé le danger, et a allumé des contre-feux assez efficaces. Il a dissous les départements traditionnels pour créer un nombre important de petits groupes, plus d'une trentaine, donnant ainsi une position de responsabilité à de nombreux jeunes scientifiques. Il les a fortement encouragés à demander des contrats russes ou étrangers – démarche inhabituelle ici, mais facilitée par le fait que le programme génome est géré sur place, ce qui ne diminue sûrement pas les chances des demandeurs de la maison. Les chercheurs peuvent aller jusqu'à tripler leur rémunération en puisant sur les contrats, chose bien nécessaire, vu le niveau du salaire officiel... Du coup, même si une cinquantaine de chercheurs de l'Institut Engelhardt sont en séjour à l'étranger, relativement peu s'y sont fixés jusqu'ici.

Au *Sherniakin Institute*, au contraire, malgré le cadre splendide et l'équipement surabondant (il y a même des stations de travail SUN, chose exceptionnelle dans ce pays), les « seniors », dans leur quasi-totalité, sont partis pour de bon. C'est certes l'occasion d'une promotion éclair pour des jeunes qui, à 30 ans, se retrouvent à la tête d'un département, mais il n'est pas certain que l'efficacité soit au rendez-vous. Dans d'autres équipes, c'est l'impression de vide qui domine ; et j'ignore quelle peut être la situation dans les laboratoires moins cotés : les instituts rattachés à l'Académie des sciences (ceux que j'ai visités) représentent le « dessus du panier » de la recherche russe.

La qualité intellectuelle des interlocuteurs ne fait guère de doute. Ils semblent avoir une large culture scientifique, font preuve de curiosité et développent des approches expérimentales souvent empreintes d'originalité. Ils ne se laissent pas arrêter par les difficultés, et l'on sent qu'ils ont derrière eux une tradition de « débrouillardise » due à des conditions de travail ardues, sans aide possible de l'extérieur. Cette habitude de « compter sur ses propres forces » a parfois des conséquences négatives, notamment cette tendance à réinventer la roue, déjà vue à propos d'informatique et qui se retrouve dans d'autres secteurs ; s'y ajoute une sous-estimation de l'importance des contacts informels et une certaine méconnaissance de la « littérature grise » : seule l'information officielle, celle qui paraît dans les revues scientifiques, est prise en compte. Néanmoins, les scientifiques russes donnent dans l'ensemble le sentiment d'être plus cultivés et créatifs que beaucoup de leurs homologues américains (ou français). Mais cette créativité, hier corsetée dans un système très autoritaire, doit aujourd'hui s'exprimer dans un contexte complètement chaotique. La hiérarchie reste pesante et la distribution des contrats a du mal à devenir une réelle compétition ouverte...

De quoi peut être fait l'avenir ?

Il est facile d'échafauder des scénarios-catastrophe – malheureusement assez plausibles. On peut imaginer, en l'absence d'une amélioration de la situation économique du pays, la poursuite du *brain drain* vers les États-Unis et l'Europe ; on peut prévoir l'obsolescence rapide du matériel encore moderne qui équipe aujourd'hui les laboratoires, et sa mise hors d'usage – faute de pièces de rechange – à la suite de pannes triviales. L'un des deux seuls micro-ordinateurs *Macintosh* vus en Russie équipait un séquenceur d'ADN de marque *Applied Biosystems* ; il était bloqué en raison d'un problème de logiciel probablement élémentaire (mais les informaticiens russes ne connaissent pas les *Mac*). Le représentant à Moscou d'*Apple Computers* réclamait 40 dollars (en devises) pour venir faire un diagnostic, une somme quasiment inaccessible pour le laboratoire en question... On peut craindre aussi que la mise sur pied d'une industrie nationale des réactifs ne s'avère extrêmement aléatoire dans un environnement aussi perturbé, où l'inflation galopante et les distorsions aiguës d'une économie en transition fournissent à ceux qui ont quelques capitaux mille occasions de trafics bien plus fructueux que la préparation d'enzymes de restriction. C'est ainsi que pourraient disparaître, en un an ou deux, des pans entiers de la recherche biomédicale, ne laissant que des instituts

aux trois quarts déserts, coquilles vides où subsisterait peut-être l'apparence, mais non la réalité de la recherche. Ce serait un désastre, car la reconstitution ultérieure d'une communauté de recherche viable prendrait de nombreuses années.

La Russie n'est est pas encore là. Il existe actuellement plusieurs instituts confortablement équipés, dont les chercheurs sont compétents et capables de réaliser des travaux de bonne qualité. Il ne leur manque pour cela que quelques moyens, en quantité pour ainsi dire catalytique : de quoi acheter les réactifs que l'on ne peut trouver ou fabriquer sur place, de quoi aussi donner aux scientifiques les suppléments de salaire (quelques centaines de francs par mois...) leur permettant de vivre sans exercer en parallèle d'autres activités. Les montants en cause sont ridicules : le directeur d'un bâtiment de trois cents personnes les évaluait à 20 000 dollars, 100 000 francs par an pour l'ensemble de son institut. Il semble donc urgent de mettre en place rapidement des collaborations dans lesquelles les laboratoires russes pourront apporter leur savoir-faire. Celui-ci couvre toute la biologie moléculaire « classique », le séquençage, l'analyse de protéines en gel à deux dimensions, la production et la caractérisation de protéines recombinantes, sans oublier les projets d'allure plus « génomique » comme celui de Mirzabekov. En contrepartie, des soutiens financiers modestes aideraient ces laboratoires à continuer de fonctionner et à garder leurs chercheurs. La toute récente annonce par l'Inserm de « contrats de coopération Est/Ouest » va dans le bon sens ; il serait souhaitable que cette initiative soit reprise par d'autres, car les huit contrats prévus sont loin de couvrir l'ensemble des possibilités.

*Hélas, c'est plutôt le scénario-catastrophe qui s'est réalisé. Dans les soubresauts politiques et économiques qui n'ont pas fini de secouer la Russie, la recherche est vraiment passée au énième plan. Ayant participé, en 1994/1995, à un essai de collaboration de longue durée avec l'Institut Engelhardt, j'ai pu directement constater les décourageantes difficultés matérielles et administratives impliquées par de telles tentatives. Et, autant que j'ai pu en juger lors d'un récent voyage, fin 2002, la situation n'a pas, depuis, évolué de manière positive. Les équipements qui m'avaient fait bonne impression en 1993 n'ont pas été renouvelés, et les laboratoires semblent moins actifs qu'à l'époque. Quelques équipes surnagent, grâce à des contrats de recherches provenant des États-Unis ou d'Europe, ou à des chercheurs qui font la navette entre leur pays et l'Occident. Parmi ces derniers, Andrei Mirzabekov, qui a passé plusieurs années à l'*Argonne national laboratory *(États-Unis) pour développer ses* microarrays *à blocs d'acrylamide, est maintenant revenu à l'Institut Engelhardt où il a créé une* start-up *qui tente d'exploiter ces procédés. En dépit de quelques groupes encore actifs, le tableau d'ensemble est assez navrant, et la Russie sera sans doute durablement absente des recherches génomiques et post-génomiques.*

Med Sci *1995 ; 11 : 1162-1164*

ALLEMAGNE : ENFIN UN PROGRAMME GÉNOME HUMAIN

Cette chronique présente le programme Génome humain allemand, dont j'avais à l'époque une vision assez précise puisque je faisais partie du comité chargé de son évaluation. Comme on le constatera à la lecture, j'étais plutôt optimiste sur son déroulement.

Cinq ou six ans après les États-Unis, la Grande-Bretagne, le Japon et la France, voici que l'Allemagne annonce le lancement d'un grand programme d'étude du génome humain. Mieux vaut tard que jamais, dira-t-on... Il est vrai que ce pays est un cas à part. L'opinion publique y est très soupçonneuse à l'égard de la science, et plus encore du génie génétique. La connotation des termes, *gentechnik* ou *gentechnologie*, qui le désignent est même franchement négative : *Gentechnologie, der Gipfel der Ausbeutung alles lebendigen*, « Génie Génétique, sommet de l'asservissement de tout le vivant », proclament les banderoles des activistes. Si, de plus, il s'agit de l'homme, la méfiance devient extrême. Attitude compréhensible, compte tenu des terribles souvenirs de l'ère nazie : les scientifiques allemands des années 1930 ont largement contribué à la justification des programmes d'« hygiène raciale » du III^{e} Reich. Après 1945, la plupart d'entre eux sont restés les piliers de la génétique officielle, et il a fallu attendre 1988 pour que les rapports entre science et national-socialisme soient dénoncés par Benno Muller-Hill dans son livre *La science meurtrière* [1]. Du fait de ce passé chargé, des thèmes comme l'analyse des composantes héréditaires du comportement sont quasiment tabou, et les études génétiques les plus innocentes sont vues avec suspicion. L'Allemagne compte pourtant de très bons laboratoires dans ce domaine, mais aucun n'a encore engagé les travaux systématiques qui caractérisent une approche proprement génomique.

Le vent a pourtant commencé à tourner. Les excès de certains groupes anti-*gentechnik*, allant jusqu'à l'attentat à la bombe, ont sans doute lassé l'opinion ; les chercheurs et les décideurs industriels mettent en cause une législation tâtillonne qui a contraint plusieurs entreprises à transférer leurs activités de Génie génétique à l'étranger [2]. Sur le front scientifique, la *Deutsche Forschungs Gemeinschaft* (DFG), principale agence gouvernementale de financement de la recherche, avait lancé en 1987 un (modeste) programme intitulé « Analyse du génome humain par les méthodes de la biologie moléculaire ». Soutenue à partir de 1990 par le ministère de l'Éducation, de la Science, de la Recherche et de la Technologie (BMBF en initiales allemandes), cette initiative a porté ses fruits en aidant des équipes d'excellente qualité. Plus récemment, la puissante Société Max Planck (dont les fonds proviennent pour moitié du gouvernement fédéral, l'autre moitié étant fournie par les *Länder*) a joué un rôle décisif en choisissant les deux nouveaux directeurs de l'Institut Max Planck de Génétique Moléculaire (Berlin). Elle a en effet nommé Hans Lehrach, génomiste de grand renom et inventeur de nombreuses techniques ingénieuses qui prennent le contre-pied de certains paradigmes admis aux États-Unis, et Hans-Hilger Ropers, généticien humain de bonne envergure. De ce remue-ménage émerge finalement un programme national intitulé *Human Genome Research*. Annoncé fin juin par le Dr Rütgers, Ministre de l'éducation et de la recherche,

c'est un projet déjà structuré, dans lequel on retrouve sans trop de surprise nombre des idées-force qu'a soutenues Hans Lehrach depuis bientôt dix ans : l'emploi systématique de banques sous forme de filtres à haute densité, le recours à l'hybridation plutôt qu'à la PCR, et l'intégration des résultats dans une base de données centrale.

Ce programme arrive très tard, alors que plusieurs pays ont déjà beaucoup investi dans le génome. Ce n'est pas forcément un inconvénient : le projet allemand peut s'appuyer sur une analyse approfondie de l'état de l'art, prendre en compte l'ensemble des connaissances déjà acquises, et choisir un créneau prometteur sans être tenu par la nécessité de terminer des travaux en cours afin de les rentabiliser. Situation inverse, par exemple, de celle des centres de Lawrence Livermore et Los Alamos aux États-Unis, qui se sont impliqués très tôt, dès 1988, dans la cartographie physique de chromosomes entiers avec les outils de l'époque, les cosmides. Ils se sont vite trouvés confrontés à des avancées technologiques (YAC, cartographie par microsatellites et *sequence tagged sites*) qui ont dans une large mesure rendu leur travail caduc. En outre, un programme entamé aujourd'hui peut bénéficier de l'étude des différents modèles organisationnels mis en place (les *Genome Centers* aux États-Unis, la constitution d'un centre de ressources en Grande-Bretagne, l'expérience de Généthon...), analyser leurs avantages et leurs limites et en tirer des leçons utiles pour sa propre structuration.

À première vue, les choses se présentent plutôt bien. Le projet engage conjointement le BMBF et la DFG, avec l'appui de la société Max Planck et de l'industrie. Il est inscrit dans la durée, huit ans, avec un soutien de 50 à 60 millions de Deutschmarks par an, près de deux cent millions de nos francs. Ses concepteurs ont donc compris – et surtout réussi à faire admettre aux politiques – que dans ce secteur on ne pouvait pas travailler au coup par coup et qu'une planification à moyen terme était indispensable. Il est aussi résolument ancré dans la communauté scientifique internationale : les demandes doivent être rédigées en anglais, et les Allemands sont minoritaires dans le comité scientifique chargé de veiller à leur évaluation. Visiblement on a saisi que dans un « petit » pays (tout est relatif...) il fallait éviter une consanguinité excessive...

Qu'en est-il du contenu du projet ? Son objectif est clairement formulé : « Identification et caractérisation systématiques de la structure, de la fonction et de la régulation des gènes humains, en particulier ceux qui ont une importance médicale ». Ce centrage affirmé sur les gènes constitue un choix logique en 1995. La carte génétique est, sinon terminée, du moins arrivée à la limite de ce qui apparaît utile et raisonnable ; la carte physique, encore imparfaite, est en voie d'affinement rapide : s'embarquer aujourd'hui dans ce domaine ne serait pas raisonnable. Les gènes, leur expression, leur fonction constituent en revanche le nouveau front sur lequel il convient de se porter. Un volet très important du projet s'appuie sur les concepts défendus par Hans Lehrach : production massive et distribution à grande échelle de « ressources », et collation, centralisation des données obtenues grâce à leur emploi. Précisons tout de suite que les ressources en question sont essentiellement des banques (génomiques ou d'ADNc), diffusées sous forme de filtres à haute densité portant chacun des dizaines de milliers de clones régulièrement disposés grâce à des robots *ad hoc*.

On se souvient que Hans Lehrach s'est fait le champion de ce type d'approche, qui avait été initiée par son équipe à l'*Imperial Cancer Research Fund* (Londres) et a déjà rendu de fiers services à quantité de chercheurs. Elle a l'avantage de donner accès aux banques sous une forme très commode : le criblage ne demande qu'une simple hybridation sur un filtre prêt à l'emploi, et son dépouillement est grandement facilité par la disposition régulière des clones. De plus ce système facilite la centralisation de l'information : pour récupérer les clones détectés sur le filtre par une sonde, l'utilisateur en communique les coordonnées au laboratoire central, qui peut du coup emmagasiner le renseignement dans sa base de données. Les indications obtenues en divers lieux s'additionnent donc tout naturellement, et l'on obtient « gratuitement » des résultats comme la

liaison physique entre des sondes employées dans deux équipes différentes – du seul fait qu'elles ont, par exemple, révélé le même YAC sur un filtre à haute densité. Ce système réclame néanmoins une expertise et une infrastructure importantes pour assurer la construction et la conservation des banques et surtout pour la production à grande échelle de filtres de bonne qualité. Une informatique performante s'impose également pour gérer toutes les données et les mettre à la disposition de la communauté. Le projet comporte donc la mise en place d'un centre de ressources chargé de ces opérations, disposant de plusieurs dizaines de personnes, de matériel lourd et de financements conséquents. Sa mission principale consiste à mettre banques et filtres à la disposition de l'ensemble des partenaires. Il doit de plus établir une liaison étroite avec quelques centres de recherche importants qui, effectuant des travaux à grande échelle grâce aux outils fournis, collaboreront avec le centre pour l'archivage des résultats.

Mais l'appel d'offres qui vient d'être publié est ouvert, et donne la possibilité de créer des centres de recherche (ou, plus ponctuellement, de financer des équipes) dans les champs du « très grand séquençage », de l'informatique appliquée au génome ou d'études systématiques sur la fonction de nombreux gènes. Avec les montants prévus, et compte tenu du fait que l'infrastructure existant en Allemagne est déjà bien fournie, on peut effectivement envisager d'alimenter simultanément un centre de ressources, un programme en réseau assez centralisé et des projets explorant d'autres voies. Une telle ouverture est indispensable. La méfiance envers le génome n'a pas disparu en Allemagne, la communauté scientifique elle-même n'y échappe pas totalement, et un plan centré sur une approche unique – émanant, qui plus est, d'un étranger récemment implanté (Hans est autrichien) – serait difficilement accepté. Le projet initial, très « Lehrarchien », a donné lieu à d'intenses discussions dont la revue *Science* s'est fait récemment l'écho [3] ; elles ont abouti à un intitulé plus ouvert mais qui garde une bonne cohérence.

Sauf anicroche imprévue, le programme Génome Humain allemand semble donc bien parti. En regard, la situation en France apparaît confuse, et notre propre programme prend des allures d'ectoplasme. Durant quelques années, l'implication massive de l'AFM, avec la création de Généthon, a permis de grands progrès qui nous ont placés dans le peloton de tête : la dernière version de la carte génétique (cinq mille marqueurs, moins d'un centimorgan de distance moyenne entre les repères) va être publiée très prochainement, et la carte physique CEPH/Généthon, malgré ses limites, a joué un rôle de premier plan. Mais dorénavant l'AFM se tourne plus nettement vers la thérapie – évolution parfaitement logique – et réduit par conséquent son investissement dans la génétique. Le secteur public n'a pas vraiment pris le relais, bien que les effets d'annonce n'aient pas manqué. Le « Programme Génome français », révélé en grande pompe à l'automne 1990 ne s'est concrétisé qu'en 1993. Le Groupement de Recherches et d'Études sur les Génomes (GREG) a alors fonctionné durant deux années à un niveau budgétaire raisonnable (une bonne soixantaine de millions annuels) avant de voir ses attributions et son budget brutalement réduits – dans une relative confusion – début 1995. Les « Actions concertées coordonnées » lancées par le ministère ont repris certains de ses champs d'action, avec un flou notable, des moyens mal définis et un évident recouvrement des thématiques. À l'heure actuelle, les règles du jeu ne sont pas établies de manière claire, et les demandeurs comme les évaluateurs s'interrogent. Le nouveau secrétariat d'État à la Recherche sera-t-il à même de redresser la barre ? Dans une période d'austérité budgétaire peu favorable aux dépenses « improductives », on peut craindre que les bonnes intentions affichées ne se traduisent pas dans les faits. Compte tenu des avancées très significatives récemment réalisées dans notre pays, ce serait fort regrettable...

Références

1. Müller-Hill B. *Murderous science*. New York : Oxford University Press,1988.
2. Abbott A. Germany will ease requirements of gene technology laws in bow to researchers. *Nature* 1993 ; 360 : 286.
3. Kahn P. Germany warily maps genome project. *Science* 1995 ; 268 : 1556-8.

En fait, le programme Génome humain allemand ne semble pas avoir été un grand succès et, à ma connaissance, deux difficultés majeures ont compromis son efficacité. D'une part, la réaction du milieu scientifique allemand contre la relative main-mise de Hans Lehrach sur son organisation – déjà brièvement évoquée dans la chronique – s'est poursuivie et a abouti à un certain éparpillement des financements et des projets. D'autre part, Lehrach lui-même n'a pas utilisé à bon escient les financements considérables dont il disposait, tant pour les structures de recherche que pour celles de service. Leur montée en régime trop rapide et insuffisamment structurée s'est apparentée à une fuite en avant et a fortement nui à leur efficacité. En revanche, les actions menées à la même époque pour développer la recherche industrielle en biotechnologie et génomique – à peine mentionnées dans mon texte – ont effectivement permis de nombreuses créations d'entreprises et ont nettement amélioré la situation sur ce plan.

Med Sci *2001 ; 17 : 760-761*

GÉNOME : QUAND LA CHINE S'ÉVEILLERA...

J'avais fait, en 1987, une tournée assez approfondie des instituts chinois, et en avais retiré une impression catastrophique : locaux dégradés et mal tenus, appareils sommeillant sous des housses, et activité très faible : visitant durant quinze jours des laboratoires de biologie moléculaire, j'avais une seule fois vu un chercheur manipulant à sa paillasse... Revenant en 2001 et voyant des installations modernes, des scientifiques dynamiques et affairés, j'ai naturellement été très sensible aux changements intervenus. Mon enthousiasme devant cette nouvelle réalité est bien perceptible dans le texte qui suit.

La Chine, malgré l'existence depuis 1994 d'un Programme Génome national, a joué jusqu'ici un rôle mineur dans l'ensemble des projets internationaux touchant au Génome humain. Il semble qu'elle ait actuellement la volonté, et les moyens, de prendre une place nettement plus importante, c'est du moins l'impression retirée d'un récent voyage (mars/avril 2001) dans ce pays[1] où j'ai pu visiter des centres de séquençage importants... et constater *de visu* qu'ils fonctionnaient réellement (*voir* [1] pour des informations générales).

La Chine, c'est une banalité de le rappeler, est un immense pays en plein développement – un développement chaotique et inégal. L'aisance évidente d'une partie notable de la population côtoie un dénuement flagrant ; une dictature fossilisée régit une économie de marché dynamique et sans complexes... Dans ces villes en perpétuel chantier où des immeubles d'un luxe insolent côtoient des ruelles presque sordides, des laboratoires ultramodernes ont surgi. Ils ont été créés de toutes pièces à l'écart des centres universitaires traditionnels encore très dégradés et dont les conditions de travail sont souvent d'un autre âge. C'est ainsi qu'à Pékin (Beijing), le *Beijing Genomics Institute* emploie plus de trois cents personnes dans un bâtiment neuf, au sein d'une zone industrielle et technologique très éloignée du centre-ville mais proche de l'aéroport.

Cet Institut [2] est essentiellement consacré au séquençage, et dispose pour ce faire de trente-deux séquenceurs modernes, des appareils *Megabace* à 96 capillaires[2] de la firme *Molecular Dynamics*. Cette machine (rivale du séquenceur *Perkin-Elmer 3700*) peut effectuer 6 *runs* par 24 heures en déterminant à chaque fois 96 séquences de 500 nucléotides environ. La production totale (théorique) de cet atelier est donc d'environ 20 000 séquences de 500 nucléotides par 24 heures. Lors de ma visite, 31 machines sur

1. L'un des objectifs de ce voyage était la participation à un atelier national UNESCO sur les aspects éthiques de la génétique et de la biotechnologie – sujet « chaud » en raison de l'adoption de lois à tendance eugénique il y a quelques années.

Cet atelier (voir http://south.genomics.org.cn/unsceo/index.htm) a permis de constater l'existence d'un vif débat sur ces sujets en Chine ; il fera peut-être l'objet d'une prochaine chronique...

2. Les séquenceurs à capillaires présentent, par rapport aux machines classiques effectuant l'électrophorèse sur plaque, l'avantage d'éliminer les opérations manuelles (coulage des gels et chargement des échantillons), et – grâce au meilleur refroidissement des capillaires – permettent l'emploi de tensions plus élevées, donc des runs plus brefs. Les machines *Molecular Dynamics Megabace* et *Perkin-Elmer 3700* sont à capillaires ; la génération précédente des systèmes à plaque est représentée notamment par les appareils de type *Perkin-Elmer 377* et *LiCor*.

32 étaient en fonctionnement effectif, ce qui suggère que la production réelle est proche de ce chiffre – l'installation tourne 24 heures sur 24, les opératrices fournissant 40 heures de travail en quatre jours suivis de trois jours de repos. L'alimentation de ces séquenceurs – sous-clonages, production des ADN, réactions de séquence – mobilise une centaine de personnes, qui travaillent pour l'essentiel manuellement, sans robots. L'assemblage des séquences est l'affaire d'une centaine d'informaticiens et bio informaticiens, dotés de nombreuses stations de travail et d'un super-ordinateur chinois (monté à partir de composants d'origine nord-américaine), un *Dawning 3000*, donné pour une vitesse de 400 gigaflops (400 milliards d'opérations par seconde). L'ensemble a donc une capacité supérieure à notre Génoscope, qui bénéficie d'une centaine de machines *LiCor* nettement moins performantes (électrophorèse sur plaque, permettant des lectures plus longues mais avec un débit beaucoup plus faible) et d'un personnel moins nombreux mais, il est vrai, doté d'une robotique nettement plus développée.

Le *Beijing Genomics Institute* est un centre public (ou du moins *not for profit*, il y a des distinctions dont la subtilité m'échappe un peu), soutenu par la puissante Académie des Sciences chinoise (CAS, *Chinese Academy of Science*) et dirigé par un énergique chercheur chinois, Huanming Yang, qui a effectué plusieurs séjours post-doctoraux en Europe et aux États-Unis, et est actuellement secrétaire général du programme Génome chinois. Le personnel de l'Institut est surtout formé de jeunes techniciens diplômés en chimie, biologie ou informatique après deux ou trois années d'études universitaires, avec quelques étudiants en thèse et très peu (cinq ou six) de chercheurs de niveau post-doctoral. L'absence de cadres est une constante de ces laboratoires, liée au développement récent de ces recherches et à la perte de toute une génération d'étudiants en raison de la Révolution Culturelle (pratiquement aucun enseignement durant six ou sept ans...). D'importants efforts d'animation scientifique (séminaires, voyages) sont semble-t-il entrepris pour pallier l'isolement géographique du centre.

Les grands projets menés dans cette structure sont au nombre de trois. Tout d'abord, le séquençage du génome du riz, un hybride très productif utilisé en Chine et baptisé *Super-rice* ainsi que ses deux variétés parentales, *indica* et *japonica* [3]. Ensuite, un projet de grande envergure (budget total de l'ordre de 500 millions de francs) sur le génome du porc [4], en collaboration avec un consortium danois alliant secteur public et privé. Ce programme comporte à la fois le séquençage de grandes collections de clones d'ADNc et celui de l'ADN génomique proprement dit. Enfin, une contribution au séquençage du génome humain (la Chine s'est engagée pour 1 % de la séquence totale) et, au-delà, des travaux sur la diversité humaine grâce au déchiffrage de régions sélectionnées chez les 57 « nationalités » que comporte ce pays. Il reste de la place pour de « petits » projets (quelques mégabases). Ces derniers pourraient fournir à des équipes françaises de fructueuses occasions de collaboration... à la condition de « renvoyer l'ascenseur » vers nos collègues chinois en effectuant quelques séjours sur place, en participant à l'animation scientifique du centre et à la formation de chercheurs chinois, et enfin en aidant à la mise en place des approches de la Génomique moderne, encore assez embryonnaires ici.

Un deuxième grand centre de séquençage, le *Hangzhou Genomics Institute* [5], est situé à Hangzhou, agréable ville moyenne (moyenne pour la Chine, elle compte tout de même près de deux millions d'habitants), à moins de deux cents kilomètres de Shanghai. Mis sur pied en grande partie grâce à des subventions et des prêts de la municipalité (plus de deux cents millions de francs), c'est en fait une annexe du centre de Beijing. Ce laboratoire dispose de trente-huit séquenceurs *Megabace* (qui là encore étaient presque tous en marche lors de ma visite) et d'un ordinateur du même type que dans la capitale, mais reçoit les échantillons prêts au chargement de Beijing et se borne au séquençage proprement dit (30 personnes) et à l'assemblage des séquences (30 personnes également). Les thèmes sont les mêmes que dans la « maison-mère », et l'animation scientifique sûrement plus limitée.

La dernière structure visitée est le *Chinese National Human Genome Center at Shanghai*, situé lui aussi dans un parc scientifique et technologique dans le nouveau Shanghai (zone de Pudong). Ce centre, d'une centaine de personnes, est plus diversifié que les deux précédents : séquençage (avec un parc plus modeste d'une dizaine de *Perkin-Elmer 377*, cinq *Megabace*, trois *Perkin-Elmer 3700*), mais aussi des groupes de génomique et de génétique humaine (en collaboration). Là aussi, le manque de cadres est flagrant, et l'anglais est très peu pratiqué en dehors de quelques responsables. Ce centre est dirigé par le Professeur Chen Zhu, francophone et francophile. Son potentiel est intéressant en raison de son équipement et de sa liaison avec les équipes universitaires de Shanghai ; il m'a néanmoins semblé moins actif que les deux précédents. Une équipe d'un des instituts de la CAS à Shanghai *(Institute of Biochemistry and Cellular Biology)* collabore avec lui et réalise notamment des *microarrays* avec des résultats de bonne qualité.

Ce panorama impressionniste ne prétend pas être exhaustif ; il existe certainement d'autres structures travaillant sur le génome dans une optique à grande échelle – bien que les deux centres de séquençage visités soient probablement les plus productifs. Mais il suffit à montrer que la contribution de la Chine ne peut plus être négligée dans ce domaine. Contrairement à ce qui a pu être le cas dans le passé, les centres annoncés existent réellement, les machines sont bien là, et pas sous des housses : elles fonctionnent souvent à plein régime. Il s'ouvre peut-être des opportunités de collaboration intéressante pour certaines équipes françaises. Cela d'autant plus que plusieurs des responsables ont effectué des séjours prolongés en France (c'est vrai pour Huanming Yang comme pour Chen Zhu) et sont restés plutôt francophiles – ou en tous cas très intéressés à contrebalancer la toute-puissance des États-Unis[3] par des collaborations avec les laboratoires de notre pays...

Références

1. http://www.chgc.sh.cn.
2. http://www.genomics.org.cn/ (avec version anglaise).
3. Li H. China to sequence hybrid rice genome. *Science* 2000 ; 288 : 1331.
4. Li H. China, Denmark team up to tackle the pig. *Science* 2000 ; 290 : 913-4a.
5. http://www.south.genomics.org.cn/ (principalement en chinois).
6. Li H. Genomics. Money and machines fuel China's push in sequencing. *Science* 2000 ; 288 : 795-8.

3. Le comportement de Craig Venter et de l'entreprise Celera vis-à-vis des chercheurs et des institutions chinoises n'a pas été au-dessus de tout reproche. Celera a notamment facturé près de trois cent mille dollars le séquençage d'un virus de trois cent kilobases tout en se réservant une partie de la propriété industrielle correspondante [4]... Le choix quasi exclusif fait, en Chine, en faveur de *Molecular Dynamics* (rival de Perkin-Elmer et donc de Celera) pour l'équipement des centres de séquençage n'est probablement pas sans rapport avec ces démêlés...

This page intentionally left blank

6. HÉSITATIONS HEXAGONALES

La situation française est souvent évoquée au fil de ces chroniques, mais en général de manière allusive et rapide. On trouvera ci-après deux textes spécifiquement consacrés à l'organisation du programme Génome dans notre pays. Le premier, écrit en 1992, présente l'état des lieux et les différents acteurs publics et privés. Il évoque ensuite la difficile mise en route du « Programme Génome français » annoncé dès l'automne 1990 (voir le Flash *page suivante) par Hubert Curien mais qui n'allait se concrétiser qu'en 1993 avec le GREG – lequel ne vécut qu'à peine trois ans. Le deuxième est consacré à la création de ce qui est aujourd'hui le Génoscope (Centre National de Séquençage, à Évry).*

Med Sci *1990 ; 6 : 908*

FLASH

« Flash » de dernière minute, rédigé dans la précipitation pour le numéro de novembre 1990 de la revue en raison de l'annonce du « Programme Génome Français ». Le moins que l'on puisse dire, c'est que la mise en place de ce programme n'a pas bénéficié du même sentiment d'urgence...

Le programme français « Génome humain » auquel je faisais allusion dans ma dernière Chronique Génomique *(m/s n° 8, vol. 6, p. 807)* a été officiellement annoncé par le Ministre de la recherche et de la technologie, Hubert Curien, le 17 octobre dernier. Rappelons que l'étude de ce programme avait été confiée à Philippe Lazar (Directeur de l'Inserm), et le projet établi par Philippe Kourilsky après consultation d'un certain nombre de scientifiques impliqués dans ce domaine.

Quelles sont les grandes lignes de ce « Programme National Génome Humain » ? Sur le plan scientifique, il privilégie le créneau de l'étude des régions codantes, c'est-à-dire du séquençage des ADNc (et non de l'ADN génomique « tout venant »). C'est une option raisonnable mais moins évidente qu'elle n'en a l'air : elle pose un certain nombre de problèmes méthodologiques qui sont justement évoqués dans la chronique de ce mois. Le séquençage exhaustif de petits génomes, ainsi que les développements informatiques, sont les deux autres points forts du programme, à côté d'un soutien à l'informatisation, à la distribution de matériel biologique et à la formation du personnel. Ce GIP devrait disposer d'environ 50 MF en 1991 et 100 MF en 1992 : cela représente un investissement tout à fait notable. Le début de mise en place devrait se situer dans les semaines à venir, mais à l'heure actuelle le Président du GIP (dont la personnalité aura une grande importance) n'est pas désigné.

À l'heure où le programme européen sur le Génome démarre enfin pour de bon (le premier appel d'offres vient de sortir), ces décisions sont très opportunes. J'espère pour ma part que les moyens annoncés sont réellement des moyens nouveaux, et que les problèmes de personnel auxquels j'ai déjà souvent fait allusion seront traités de façon adéquate dans le nouveau cadre. En tous cas, bonne chance au Programme National Génome Humain !

Med Sci *1992 ; 8 : 264-267*

GÉNOME FRANÇAIS : DE GRANDES ESPÉRANCES...

On trouvera ci-dessous l'analyse de la situation française que j'annonçais à la fin de la chronique consacrée en 1991 aux États-Unis (voir Chapitre 5*). Terrain miné, les enjeux scientifiques, politiques et personnels étaient élevés et les oppositions au sein du milieu de la recherche assez vives. La déclaration tout à fait officielle rapportée dans le* flash *ci-dessus n'avait été suivie d'aucun acte concret, et le retard français semblait se creuser. Les succès des travaux – jusque-là confidentiels – menés au Généthon allaient pour un temps mettre notre pays en pointe dans le monde du Génome ; mais on verrait plus tard que cela ne remplaçait pas un programme complet et structuré comme celui de la Grande-Bretagne.*

Une troisième place honorable

L'examen de critères quantitatifs permet d'affirmer que notre pays se situe à la troisième place mondiale pour les recherches sur le Génome humain. La *Figure 1* montre quelques-uns de ces indicateurs pour quatre nations (États-Unis, Japon, Grande Bretagne, France), en regard de leur produit national brut. Que l'on envisage, comme l'a fait la Fondation Européenne de la Science [1], le nombre de publications touchant à la cartographie des gènes chez l'homme, ou la proportion de nos compatriotes cooptés à HUGO avant que certains *lobbies* ne s'organisent, ou encore que l'on se livre à une analyse des communications présentées au congrès annuel du Génome à Cold Spring Harbor, la conclusion sera la même. Nous pouvons en tirer un certain réconfort puisque notre recherche se trouve ainsi placée devant celle du Japon ; d'un autre côté le fait que notre production soit au moins deux fois plus faible que celle de la Grande-Bretagne incite à une grande modestie... En tout état de cause la France fait partie des « poids lourds » dans ce secteur, ce qui lui donne à la fois des droits et des responsabilités.

Sur le plan qualitatif l'évaluation est naturellement plus délicate. On peut noter quelques réussites récentes dans de très médiatiques « courses au gène » : localisation de l'amyotrophie spinale par l'équipe d'Arnold Munnich [2], clonage de la région de l'X fragile et analyse de sa variation par celle de Jean-Louis Mandel [3], et isolement du gène impliqué dans le syndrome de Kallman par le groupe de Christine Petit et Jean Weissenbach [4]. On remarquera cependant que dans chacun de ces cas la compétition était très sévère, et que les publications françaises sont intervenues en même temps (à très peu de chose près, dans un sens ou dans l'autre) que celles d'une ou deux équipes étrangères travaillant sur le même sujet. De façon plus « structurelle » la place de notre pays sur le plan international est due pour une bonne part à l'existence du CEPH et à la position centrale qu'il a su prendre dans l'établissement de la carte génétique humaine. Nous reviendrons plus loin sur cette organisation qui présente nombre de caractéristiques inhabituelles ; rappelons simplement que Jean Dausset a su dès 1980 prévoir que l'étude génétique de l'homme serait grandement facilitée par la constitution d'un « jeu » de familles dont l'ADN pourrait être fourni à de nombreux laboratoires et qui permettrait l'obtention d'un jeu cohérent de données. De fait, le CEPH a joué et continue de jouer un

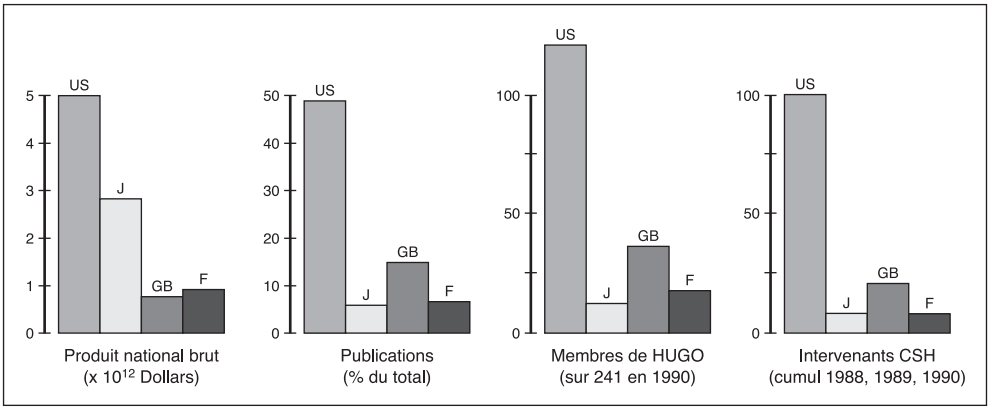

Figure 1. Quelques éléments de comparaison entre quatre grands pays. Premier panneau, à gauche, leur produit national brut en milliers de milliards de dollars ; puis leur production scientifique estimée par leur part des articles touchant au Génome humain (étude réalisée pour la Fondation Européenne de la Science). Le troisième panneau montre le nombre de membres de l'organisation HUGO en mars 1990. Les campagnes de cooptation suivantes ont été marquées par l'apparition de lobbys efficaces dans certains pays comme l'Allemagne (plus de dix élus en une campagne), alors que d'autres pays s'en désintéressaient (un seul candidat français) ; enfin le quatrième panneau montre le cumul des intervenants en session plénière au colloque annuel *Genome mapping and sequencing* de Cold Spring Harbor. On y remarque un certain écrasement des non-Américains...

rôle primordial dans l'affinement de la carte génétique humaine, tout en s'étant investi dans plusieurs entreprises « génomiques », en association avec une autre structure *ad hoc*, le Généthon.

Une remarque me paraît s'imposer à ce stade : le tissu scientifique français a été peu innovant sur le plan méthodologique (même s'il a su être créatif au niveau des structures). Si l'on passe en revues les grandes et petites révolutions conceptuelles et technologiques qui ont permis les progrès dans l'étude du Génome, on verra que leur origine est presque toujours à l'étranger. Les concepts nouveaux – comme la cartographie génétique à l'aide des polymorphismes de l'ADN [5], le développement d'approches comme la Génétique inverse – sont apparus outre-Atlantique. Les outils nouveaux de la cartographie physique, ceux qui ont rendu possible des cartes couvrant des chromosomes entiers, ont été eux aussi mis au point à l'étranger : gels pulsés [6], YAC [7], hybridation *in situ* non radioactive, *exon trapping*, proviennent pour l'essentiel des États-Unis ; l'approche systématique par *contigs* de cosmides [8], les hybrides d'irradiation, de Grande-Bretagne ; quant au renouveau des méthodes de microdissection son origine se situe pour l'essentiel en Allemagne [9]. La grande stabilité qui caractérise les structures de recherche françaises n'a donc pas été mise à profit pour faire des investissements à long terme dans les développements méthodologiques, et c'est à mon avis dommage. Cet état de fait tient sans doute à la faible estime dont est l'objet la technologie dans notre pays et dans nos instances d'évaluation. Quelles qu'en soient les raisons, nos laboratoires se sont limité à appliquer, parfois de façon très performante, les techniques mises au point ailleurs, mais n'ont que très rarement eu un rôle innovant de ce point de vue.

Un secteur associatif puissant

Contrairement à ce que l'on pense souvent, la recherche biologique française n'est pas le domaine réservé du « tout État ». Certes la place des laboratoires publics, Inserm, CNRS, avec leurs forces et leurs faiblesses, est très importante. Mais l'initiative privée joue, et depuis longtemps, un rôle notable. Cela est particulièrement vrai pour le cancer avec les fondations comme la Ligue nationale française contre le cancer (LNFCC) ou l'Association de recherche sur le cancer (ARC) qui mobilisent des moyens considérables. Pour le génome, l'entité « incontournable » est l'Association française contre les myopathies. L'AFM existe depuis de nombreuses années, mais elle a été transformée à partir du milieu des années quatre-vingts par un président très dynamique et convaincu de l'importance de la recherche médicale. L'association a eu recours à des modes de financement jusque-là inhabituels dans notre pays, en particulier le fameux Téléthon, et a drainé par ce moyen depuis 1987 des sommes impressionnantes : entre 200 et 300 millions de francs (MF) chaque année. Une partie de ce montant finance diverses aides aux malades, mais la fraction consacrée à la recherche est significative : entre 100 et 200 MF. Pour prendre la mesure de ce que représente une telle somme, il faut la mettre en regard du budget d'un organisme public, et analyser les différents postes que comporte ce dernier. Le secteur « Sciences de la Vie » du CNRS, pour prendre cet exemple, dispose d'un budget annuel total de l'ordre de 2 000 MF. Le chiffre est impressionnant, mais il faut savoir que les dépenses de personnel (fonctionnaire, donc inamovible) représentent plus de 75 % de ce total. Les 25 % restants, soit 500 MF, couvrent l'ensemble des dépenses de fonctionnement, d'équipement, d'investissement immobilier... Comme il n'est guère possible de restreindre considérablement les dotations des laboratoires d'une année sur l'autre, cela signifie que la marge réellement disponible pour des actions nouvelles et volontaristes est de quelques dizaines de millions de francs pour l'ensemble des disciplines couvertes par ce secteur (les chiffres pour l'Inserm sont similaires, avec un peu moins de dépenses de personnel). Les 100 MF injectés par l'AFM dans la recherche sur les maladies génétiques (en sus du fonctionnement du « Généthon » mentionné plus loin) pèsent donc très lourd en comparaison, et donnent à cette association un impact considérable.

Cet effet est d'autant plus sensible que les objectifs et l'état d'esprit des représentants d'une association de malades diffèrent de ceux d'un responsable d'organisme. À travers les 250 ou 300 MF du Téléthon, c'est une demande sociale qui s'exprime, et qui réclame que l'on fasse, vite, quelque chose pour lutter contre ces maladies et obtenir des résultats concrets. Il est bon pour les chercheurs d'être confrontés à cette saine impatience ; mais des distorsions sont possibles, et pas toujours faciles à éviter.

Des modes d'organisation antinomiques

Le monde de la recherche génomique française combine deux extrêmes. D'un côté des laboratoires « publics », Inserm ou CNRS en général ; il y en a peut-être une vingtaine qui comptent au niveau international. Ces structures sont emprisonnées dans le corset du service public : financements modestes, décisions lentes. Il faut par exemple compter seize mois au minimum entre le dépôt d'une demande de création d'unité Inserm toute ficelée, et son éventuelle création, sans aucune garantie sur les moyens en personnel et crédits qui lui seront accordés. Ces entités ont surtout fort peu d'autonomie (le laboratoire n'a pas la personnalité morale, et le directeur ne peut signer aucun contrat) et quasiment aucun moyen de gérer avec efficacité leur personnel. Difficile dans ces conditions de construire des structures dynamiques, évolutives, capables de réagir « au quart de tour », de mettre de gros moyens sur un thème chaud et de réviser leurs priorités dès que nécessaire... On peut certes arriver à un fonctionnement acceptable, mais au prix de beaucoup d'énergie et de quelques tours de passe-passe. Autant d'efforts perdus pour la

recherche : la créativité dépensée pour tenter d'embaucher un informaticien au prix du marché grâce à des cumuls souvent illégaux aurait mieux été employée à affiner une stratégie expérimentale. L'intervention de l'AFM, qui distribue assez largement contrats et bourses, met de l'huile dans les rouages (parfois même beaucoup d'huile) mais ne suffit pas à compenser les défauts du système.

De l'autre côté de la barrière, les structures très atypiques que sont le CEPH et le Généthon. Le CEPH, fondé au départ avec les fonds personnels de Jean Dausset (cela mérite d'être signalé), est toujours une association loi de 1901. Il est néanmoins financé en grande partie sur des fonds publics, avec une subvention du ministère de la Recherche d'une vingtaine de millions de francs ; un montant à peu près équivalent est fourni par d'autres sources dont l'AFM, la CEE... Ses activités sont multiples, et il n'est pas toujours facile de faire la part de ce qui est du « service » (le criblage de la banque de YAC, la fourniture de ADN des familles...) et l'activité de recherche proprement dite, d'autant plus que le CEPH est lui-même imbriqué avec le Généthon (voir plus bas). Mais globalement on peut noter que cette structure d'une cinquantaine de personnes est largement financée, et que son fonctionnement s'apparente plus à celui d'une entreprise que d'un laboratoire public, surtout en ce qui concerne le point crucial de la gestion du personnel. C'est à mon avis, on l'aura compris, une bonne chose ; et de ce point de vue le CEPH s'apparente à un de ces *Genome Centers* établis aux États-Unis. En revanche, il me semble que ces derniers, bien financés eux aussi (une cinquantaine de MF par an, salaires compris) sont soumis à une évaluation plus rigoureuse, à un planning de résultats plus quantifié que ne l'est le CEPH.

Le Généthon, lui, est un résultat de la « saine impatience » mentionnée plus haut. Au départ, le choc ressenti par Bernard Barataud, président de l'AFM, en constatant que des étapes cruciales mais très laborieuses du processus qui mène au gène d'une maladie étaient réalisées, pour l'essentiel, de façon manuelle. Diagnostic partagé par Daniel Cohen (CEPH), et qui amena à la conception d'un centre spécialisé équipé de matériel lourd et susceptible d'être utilisé à tour de rôle par des équipes de recherche pour accélérer leur travail à certaine phases critiques. C'était le concept du Généthon, installé depuis maintenant plus d'un an à Évry, dans la banlieue sud de Paris. Important investissement puisque l'AFM y consacre cette année un budget de 74 MF (l'équipement est amorti sur trois ans) et y a recruté 130 personnes. L'idée est d'y rassembler un ensemble d'automates, par exemple vingt exemplaires de la machine à *Southern blots* concue dans le cadre du programme LABIMAP[1] que viennent utiliser les équipes extérieures : plus d'une dizaine y ont eu effectivement recours. Le Généthon abrite aussi des projets comme celui de Charles Auffray sur le séquençage massif d'ADNc, de Daniel Cohen sur la cartographie physique à l'aide de YAC ou de Jean Weissenbach sur l'isolement de très nombreux « microsatellites » destinés a faire progresser la cartographie génétique.

Le monde du Génome français est donc, on le voit, très divers. Il est même un peu schizophrénique ou en tous cas conflictuel, partagé qu'il est entre des laboratoires opérant dans un cadre très contraignant (bien plus qu'un *Genome Center*, pour reprendre ce point de comparaison) et d'autres assez débridés et disposant de moyens très supérieurs. C'est un terreau fertile pour le développement de jalousies inévitables entre équipes où les dotations (et même les salaires !) diffèrent considérablement. Ajoutons à cela que ce monde comporte plusieurs fortes personnalités dont la modération n'est pas la vertu première et qui ont une nette tendance à prendre toute critique pour une attaque personnelle, et l'on comprendra que le secteur n'est pas de tout repos et qu'il y règne une ambiance peu favorable à la sérénité.

1. Programme européen, dans le cadre « Euréka », associant le CEPH, l'entreprise Bertin, la société Amersham... pour la mise au point d'un série d'automates de biologie moléculaire.

Faut-il un programme Génome ?

Puisque ce dernier n'a pas encore réellement démarré, on peut légitimement se poser la question. Après tout, nous sommes à la troisième place, les bonnes équipes trouvent sans trop de peine des contrats (grâce surtout à l'AFM)... pourquoi ne pas continuer comme cela ? À mon avis – cela n'étonnera sans doute personne – un tel programme est indispensable, et vite. L'AFM, qui donne actuellement un ballon d'oxygène à ces recherches, a clairement annoncé son intention de financer la Génétique pendant trois ans (déjà bien entamés) mais de passer ensuite au soutien des approches thérapeutiques (thérapie génique ou autres). C'est de toutes façons un soutien aléatoire : qui peut assurer que l'intérêt du public se maintiendra, que le Téléthon continuera à « marcher », qu'une autre cause ne sollicitera pas le public avec succès ? Il faut donc organiser le relais, et donner aux laboratoires publics la souplesse nécessaire (en finances et en personnel) pour répondre aux défis du Génome dans le cadre d'un programme planifié sur quelques années. Il faut aussi assurer l'avenir du CEPH (et du Généthon) en les faisant rentrer – dans une certaine mesure – dans un cadre plus « normal » mais, surtout, sans perdre les éléments de souplesse indispensables. Un programme Génome français clairement annoncé, avec des structures transparentes, des moyens d'action et un financement assuré sur quelques années pourrait utilement restructurer ce monde dont j'ai décrit les contradictions, tout en impulsant les actions spécifiques indispensables (informatique, banques...). Il permettrait aussi qu'il y ait – enfin ! – un représentant de la France au niveau international. D'importantes négociations sont actuellement engagées : sur les banques de données et particulièrement celle de Baltimore, la Genome Data Bank, sur la brevetabilité des séquences... La participation de la France est désirée, attendue dans ces négociations ; mais dans le flou actuel personne ne peut réellement parler en son nom, et les occasions d'infléchir le cours de événements se perdent... Il y a donc urgence.

Le difficile accouchement du GIP Génome

Commençons par un peu d'histoire... Au printemps 1990 (deux ans déjà) le ministre de la Recherche chargeait Philippe Lazar, directeur de l'Inserm, de lui faire des propositions pour un programme Génome français. À l'époque les États-Unis avaient investi dans ce secteur de façon significative (plus d'une centaine de millions de dollars par an), le Japon annonçait ses intentions, et la Grande Bretagne avait effectivement démarré un an plus tôt son *Human Gene Mapping Programme*. Il y avait donc urgence... À son tour, Philippe Lazar chargeait Philippe Kourilsky (Institut Pasteur) de rédiger un projet, ce que faisait ce dernier après une large consultation des scientifiques engagés dans ces recherches. Les propositions de Philippe Kourilsky recommandaient la mise sur pied d'un programme bénéficiant de crédits nouveaux de l'ordre de 100 MF par an, privilégiant l'étude des gènes (donc des ADNc) tout en soutenant informatique, organismes modèles et développements technologiques. Sur le plan organisationnel, il proposait la mise en place d'un « groupement d'intérêt public » (GIP) permettant d'associer secteur public et privé et insistait sur la nécessité d'une création rapide de ce GIP avec un directeur au profil de « manager » ainsi que sur la nécessité d'une gestion souple du personnel. À l'automne 1990, le 17 octobre exactement, une conférence de presse du ministre annonçait le programme Génome français, selon des modalités assez proches du rapport de Philippe Kourilsky (à part la gestion du personnel) et avec un financement nouveau de 50 MF pour 1991 et 100 pour 1992 [10]. Les choses semblaient donc bien parties.

La suite des événements allait démentir ces espérances. En mars 1991 (donc près de six mois après l'annonce ministérielle), Jacques Hanoune était chargé d'une mission exploratoire pour examiner comment ce GIP pourrait être mis sur pied. Des mois de négociations difficiles s'ensuivirent, au cours desquelles il apparut que les différents par-

tenaires potentiels du GIP avaient des idées très différentes sur ce qu'il devrait être et sur la répartition des sièges au conseil d'administration, que certains ne voyaient pas l'intérêt d'un GIP et qu'au niveau même du ministère l'appui à ce projet n'était pas unanime... Au mois d'août, lors de la onzième *Human Gene Mapping Workshop* à Londres, les choses n'avaient apparemment pas avancé. Quelques scientifiques français manifestèrent une inquiétude qui fut largement répercutée par la presse. Peu après, le ministère s'attacha à répartir, dans des délais très courts, des contrats de recherche pour un montant total de plusieurs dizaines de millions de francs. Compte tenu de ces délais la répartition fut faite dans des conditions acrobatiques et discutables (sans appel d'offres), suscitant des tensions supplémentaires dans un milieu qui n'en avait pas besoin. Mais ce fut ensuite au tour des heureux élus (ceux à qui l'on avait distribué de l'argent) d'être mécontents, car ces crédits qu'on leur avait octroyés n'existaient pas réellement ; plus précisément c'étaient des « autorisations de programme » et non des « crédits de paiement » (les initiés comprendront) et pour les transformer en espèces sonnantes et trébuchantes il fallait ponctionner l'Inserm ou le CNRS... évidemment ravis du procédé...

Il semble maintenant que cette question soit en voie de règlement pour au moins une partie de ces crédits. Quant au programme Génome, il pourrait sous peu bénéficier d'un nouveau chargé de mission, un célèbre biologiste moléculaire de la levure qui vient de quitter la direction de son laboratoire de Gif-sur Yvette. Espérons que cette fois la tentative aboutira : il est grand temps...

Références

1. Academia Europaea. *Research on the Human Genome in Europe and Its Relation to Activities Elsewhere in the World*. Strasbourg : Academia Europaea, 1991.
2. Melki J, Abdelhak SS, Sheth P, *et al.* Gene for chronic proximal sminal muscular atrophies maps to chromosome 5q. *Nature* 1990 ; 344 : 767-81.
3. Heitz D, Rousseau F, Devys D *et al.* Isolation of sequences that span the Fragile X and identification of a Fragile X-related CpG island. *Science* 1991 ; 251 : 1236-9.
4. 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.
5. Botstein D, White RL, Skolnick M, Davis RW. Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *Am J Hum Genet* 1980 ; 32 : 314-31.
6. Schwartz DC, Cantor CR. Separation of yeast chromosome-sized DNAs by pulsed field gel electrophoresis. *Cell* 1984 ; 37 : 67-75.
7. Burke DT, Carle GF, Olson MF. Cloning of large segments of exogenous DNA into yeast artificial-chromosome vectors. *Science* 1987 ; 236 : 806-8.
8. Coulson A, Sulston J, Brenner S, Karn J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc Nat Acad Sci USA* 1986 ; 83 : 7821-5.
9. Ludecke HJ, Senger G, Claussen U, Horsthemke B. Cloning defined regions of the human genome by microdissection of banded chromosomes and enzymatic amplification. *Nature* 1989 ; 338 : 348-50.
10. Jordan BR. Le programme français Génome humain. *médecine/sciences* 1990 ; 9 : 908

Le « célèbre biologiste moléculaire » auquel je faisais allusion à la fin de cette chronique est bien sûr Piotr Slonimski. Il allait en effet devenir directeur du GIP GREG (Groupement de Recherches et d'Études sur les Génomes) finalement créé en 1993, étranglé en 1995 et fermé en 1996...

Med Sci *1996 ; 12 : 237-239*

TRÈS GRAND SÉQUENÇAGE : TROMPE-L'ŒIL POLITIQUE, OU NÉCESSITÉ SCIENTIFIQUE ?

Nous en venons maintenant à notre actuel Génoscope, que l'on baptisait alors « Centre de Très Grand Séquençage » (après le « Train à Grande Vitesse » et la « Très Grande Bibliothèque »). Projet à la gestation difficile, mais qui a abouti à une structure correctement dimensionnée et relativement performante, il était à l'époque l'enjeu de luttes de pouvoir entre le ministère, les organismes de recherche et l'AFM.

Une période critique pour les travaux « génomiques »

La conjoncture pour les recherches sur les génomes est décidément morose dans l'hexagone. L'Association Française contre les Myopathies recentre son action vers les thérapies (ce que nul ne songe à lui reprocher) et réduit son soutien à la Génétique ; du côté des organismes de recherche, les dépenses sont sévèrement encadrées, et les dotations aux laboratoires Inserm et CNRS diminuent en francs courants. Le GREG (Groupement de Recherches et d'Études sur les Génomes), chargé « d'animer et coordonner les actions scientifiques et les programmes de recherches *(sur les génomes, NDLR)* menés en France », a perdu les trois quarts de ses crédits, et ignore s'il aura un budget en 1996. Quant au ministère de l'Éducation et de la Recherche, ses « Actions Concertées Coordonnées-Sciences du Vivant » (ACC-SV) ont été mises en place à la hâte au printemps dernier. Leurs quatorze comités ont examiné plus de 1 500 dossiers, mais les financements décidés n'étaient toujours pas débloqués fin Novembre, et l'opération ne sera pas renouvelée l'année prochaine... De surcroît, les projets retenus relèvent plus de l'analyse fonctionnelle des gènes et de la génétique humaine que du Génome proprement dit[1]. En fait, il est devenu patent qu'il n'existe plus de Programme Génome français. Et les discussions qui se poursuivent au ministère sur le lancement d'un centre de « Très Grand Séquençage » peuvent sembler surréalistes : n'est-il pas incongru, dans ce contexte, d'envisager un investissement aussi lourd qu'un centre de Très Grand Séquençage (TGS) ?

Le séquençage redevient d'actualité

Le déchiffrage des génomes présente pourtant un intérêt majeur, et les projets qui débutent un peu partout donnent à réfléchir. Après les illusions des débuts, après une longue période de rodage et de déboires, le succès enregistré pour la levure, puis pour le nématode, a redonné confiance aux promoteurs du séquençage génomique et démontré son utilité. Il a montré que la lecture de centaines de kilobases, et même de mégabases, était réalisable sans attendre la mise au point des nouvelles techniques (spectrométrie de masse, effet tunnel...) dont on nous rebat les oreilles depuis bientôt dix ans. Les dividendes, pour ces organismes dont le génome est très compact, ont été d'une ampleur

1. Les intitulés des projets retenus et le montant des financements sont publiés dans le numéro d'octobre de la Lettre du GREG.

inattendue : la découverte de centaines de gènes jusque-là inconnus va donner du travail à des générations de thésards. Sur cette lancée, il redevient concevable de s'attaquer au génome humain. L'analyse des EST commence à montrer ses limites [1] et souligne la nécessité d'un séquençage exhaustif, même si le prépondérance du *Junk DNA* dans notre matériel génétique rend l'opération plus hasardeuse et moins rentable.

C'est bien ainsi que l'entendent le NIH *(National Institutes of Health)*, qui a lancé aux États-Unis un appel d'offres en ce sens, et surtout le *Wellcome Trust* (Grande-Bretagne), qui vient d'annoncer sa décision de financer le déchiffrage de cinq cents mégabases par le Centre Sanger au cours des cinq prochaines années. Le chiffre, un sixième de notre génome, laisse rêveur, tout comme le « tarif » annoncé, dix pence (quatre-vingts centimes) par base... Ce montant, très inférieur au dollar encore communément admis, serait rendu possible par des économies d'échelle ainsi que par un parti-pris de lecture rapide visant une fiabilité à 99,9 % seulement. Le tout nouveau Programme Génome allemand [2] fera lui aussi, une part au séquençage massif ; il n'est donc pas étonnant que l'on s'interroge dans notre pays. Le GREG avait organisé vers la fin 1994, une table ronde pour discuter de l'opportunité d'un programme de Très Grand Séquençage ; le ministère a, lui, réuni au début de cette année un groupe de travail, qui a rendu, peu avant les élections présidentielles, un rapport favorable au lancement d'une initiative de ce type sous réserve que des moyens significatifs lui soient alloués. Le flambeau a été repris par la secrétaire d'État à la Recherche du premier gouvernement Juppé, puis par son remplaçant à ce poste après le remaniement de novembre. Il semble donc sérieusement envisagé de lancer en 1996 une entreprise de cette nature, et de participer à un effort international dont les retombées scientifiques et industrielles seront sans nul doute considérables.

Une carte qui ne peut être jouée à moitié

Je me demande pourtant si l'on mesure bien, parmi ceux qui nous gouvernent, ce que représente un projet significatif dans ce domaine. Un centre de Très Grand Séquençage, pour être crédible par rapport aux programmes étrangers, doit produire au moins dix mégabases de séquence terminée par an : un ou deux génomes bactériens, ou une fraction non négligeable d'un chromosome humain. À un dollar par base, cela représente cinquante millions de francs par an pour les frais de fonctionnement et les salaires. Le chiffre serait beaucoup plus faible si l'on appliquait le barème du *Wellcome Trust*, mais ce dernier est-il réaliste, surtout pour un centre qui débute ? Notons que pour le projet européen sur la Levure, les laboratoires contractants ont été rétribués au taux de quinze francs par base... Un tel centre impose aussi des investissements notables : il doit regrouper une vingtaine de séquenceurs, autant de robots pour la préparation des échantillons, et enfin une informatique performante, soit un montant total de l'ordre de cinquante millions. Pour mettre en œuvre ce matériel et en tirer les mégabases attendues, il faut au moins une trentaine de personnes : des scientifiques, des informaticiens et de nombreux techniciens, le tout coordonné par quelques excellents organisateurs. L'on arrive ainsi à un effort annuel compris entre cinquante et cent millions de francs, à poursuivre durant plusieurs années. Le *Wellcome Trust* s'est engagé pour un montant de cinquante millions de livres sterling, quatre cent millions de francs : nos gouvernants veulent-ils, et peuvent-ils, faire un effort comparable ?

Deux autres points méritent d'être abordés lorsqu'on envisage un Très Grand Séquençage « à la française » : le personnel, et les compétences. Il semble exclu de faire fonctionner une structure de ce type avec du personnel technique fourni par le CNRS, l'Inserm ou l'Enseignement supérieur. Ces organismes subissent actuellement une diminution du nombre de postes, et il est peu probable qu'ils puissent en mettre des dizaines à la disposition de TGS ! Le savoir-faire spécialisé requis ne se trouvera pas obligatoire-

ment parmi le vivier des organismes, et le recrutement rapide indispensable au démarrage de l'opération tout comme sa durée de vie *a priori* limitée sont peu compatibles avec le statut de fonctionnaire. L'expérience négative des centres de service comme le CSEAL d'Orléans souligne les difficultés d'une telle approche. Il faudrait donc que les ministères concernés acceptent de créer une fondation, et la financent pour qu'elle emploie du personnel sous contrat à durée déterminée – modalité qui semble bien difficile à envisager, même pour le gouvernement actuel...

Les compétences, enfin : l'histoire récente du grand séquençage est émaillée d'échecs, et bien des laboratoires qui s'y sont engagés il y a cinq ou six ans ont arrêté les frais après avoir atteint des résultats très inférieurs à leurs objectifs. Seuls quelques-uns, dont celui de John Sulston, ont transformé l'essai. Il est instructif d'observer la montée en puissance du projet Nématode et sa parfaite conformité avec les prévisions : cent kilobases de séquence finie en 1991, quatre cents en 1992, une mégabase en 1993... le débit est maintenant d'une mégabase par mois pour chacun des deux centres (Saint Louis et Centre Sanger). Or il faut reconnaître qu'en France, aucune équipe n'a encore publié cent kilobases de séquence finie d'un seul tenant. Cela ne saurait tarder – mais nous sommes encore loin du Très Grand Séquençage, et il semble utopique de faire ce pari sans prévoir une étape intermédiaire.

Le Très Grand Séquençage plaît aux politiques

Pourquoi alors tant de déclarations, tant de bruit autour de ce projet ? D'abord, parce qu'il revêt effectivement une importance stratégique : il serait grave que la France soit absente de cette aventure, alors qu'elle avait pris (grâce à l'AFM) une place de premier plan dans la construction des cartes du génome. Être réduit au rôle d'utilisateur de séquences déterminées par d'autres, c'est accepter une place au fond de la salle, et prendre le risque de manquer les avancées de la recherche, les développements instrumentaux qu'elle impulse et les retombées industrielles qu'elle engendre. Mais le Très Grand Séquençage est aussi un objectif facilement lisible – au contraire des notions complexes de cartographie génétique ou de *contigs* de YAC – et donc séduisant pour les politiques. Rappelons nous qu'aux États-Unis le Programme Génome avait été « vendu » au Congrès sous cet aspect, à la fin des années 1980 – alors que l'idée de séquencer intégralement notre ADN était pour le moins prématurée. Aux État-Unis, le coup a réussi. Il a permis de lancer le programme, quitte à modifier considérablement son contenu une fois le principe acquis. Dans notre pays, la manœuvre semble malheureusement avoir échoué : le Très Grand Séquençage est envisagé à budget constant – et même décroissant puisque le montant de la ligne Sciences de la Vie au ministère (sur lequel serait pris le TGS) sera plus faible en 1996 qu'en 1995. Pour engager l'opération il faudrait alors supprimer tout financement spécifique des autres travaux sur les génomes, et fermer le GREG...

On évoque parfois le concours de structures industrielles ou caritatives. Certes, l'industrie s'intéresse au séquençage de l'ADN humain ou de certains pathogènes ; certes, l'AFM soutient le séquençage massif de régions promotrices réalisé par l'entreprise Genset à Évry, et pourrait éventuellement s'impliquer dans un nouveau projet mené au sein de cette entreprise. Il me semble pourtant que le Très Grand Séquençage, surtout lorsqu'il s'agit d'ADN humain, a vocation à être effectué dans un cadre majoritairement public : ceci suppose logiquement que son financement le soit aussi.

Un choix à faire sans ambiguïté

L'alternative est donc claire. Soit nous arrivons à faire comprendre aux pouvoirs publics que l'engagement effectif d'une activité de séquençage à très grande échelle est un enjeu scientifique et industriel majeur, et à les convaincre de dégager les moyens

supplémentaires indispensables, soit au minimum cinquante millions de francs par an pendant plusieurs années. Il faudra alors prévoir une montée en régime programmée et contrôlée, qui pourrait passer par la création de deux ou trois centres de « Grand Séquençage » devant chacun produire une mégabase par an ; le plus performant d'entre eux serait ensuite retenu pour passer à la vitesse supérieure. Cette tactique permettrait également d'explorer diverses options quant à l'ADN à séquencer, question que nous n'avons pas traitée ici : un « modèle » comme *Arabidopsis thaliana*, une bactérie, une région du génome humain ? Si en revanche l'action doit se limiter à réclamer trois techniciens à l'Inserm, quatre au CNRS, à acheter cinq séquenceurs et à inaugurer en grande pompe un centre qui n'aura de Très Grand Séquençage que le nom... mieux vaut s'abstenir. En tout état de cause, la décision prise devrait inclure un soutien régulier à la « vraie » recherche génomique (à ne pas confondre avec la Génétique Médicale), géré par le GREG (après tout créé pour cela) ou, éventuellement, par une autre structure, mais en évitant les incessants changements de cap et les doublons dont nous avons eu un bel exemple cette année...

Références

1. Jordan B. Génome humain : l'annuaire nouveau est arrivé. *médecine/sciences* 1995 ; 11 : 1717-9.
2. Jordan B. Allemagne : enfin un programme Génome humain. *médecine/sciences* 1995 ; 11 : 1162-4.

La belle et triste histoire du « Programme Génome Français »

Le Centre National de Séquençage (CNS ou Génoscope) a été effectivement créé en 1997 avec un statut de Groupement d'Intérêt Public (GIP) et un financement récurrent de l'ordre de 100 millions de francs par an, ce qui lui a permis de tenir une place honorable dans la compétition internationale (bien qu'à l'échelle admise en 2003 il ne s'agisse que d'un « petit » centre de séquençage).

En revanche, la politique générale du Génome en France a subi bien des vicissitudes, les chroniques consacrées aux autres pays y ont parfois fait allusion. J'en ai fait une analyse assez détaillée dans un ouvrage paru en 1996 [1]. En résumé, le Groupement d'Études et de Recherches sur les Génomes (GREG), GIP finalement créé en 1993 (en principe pour 6 ans) afin d'incarner le programme Génome français, n'a véritablement fonctionné que deux ans. Il fut privé de ses moyens puis dissous par le nouveau ministre de la Recherche, François Fillon, à la faveur du changement de majorité politique intervenu fin 1995. De fantomatiques « Actions Coordonnées et Concertées » lancées alors par le ministère de la Recherche ne l'ont pas vraiment remplacé. Au même moment, l'AFM, qui avait fortement financé les recherches sur le Génome à Généthon mais aussi dans de nombreuses équipes à travers la France, arrêtait progressivement ce soutien pour se consacrer aux approches thérapeutiques et tout particulièrement à la thérapie génique.

Le choix du lieu d'implantation du Génoscope fut l'occasion de batailles acharnées entres les tenants de Gif-sur-Yvette, d'Évry ou de la Faculté des Saint-Pères à Paris. Finalement l'AFM, qui tenait à développer l'environnement du Généthon à Évry et à en faire une « Génopole » (ville du Génome), devait gagner la partie et obtenir l'installation du Centre à Évry. Du coup, avec le CNS, Généthon et l'entreprise Genset qu'avait entre-temps rejointe Daniel Cohen accompagné d'une bonne partie des cadres du CEPH,

cette ville nouvelle concentrait tout un pan de la Génomique française. Cet ensemble allait en 1998 être officiellement structuré en tant que Génopole, attirer d'autres laboratoires et devenir un centre important pour la création ou l'implantation d'entreprises de biotechnologie.

En 1999, conscient du retard français (5 000 emplois en Génomique industrielle pour la France contre 30 000 pour la seule Californie), le ministère dirigé par Claude Allègre lançait une vague de créations de « Génopoles-bis » qui visait à généraliser ce modèle de synergie entre recherche, enseignement et industrie dans le secteur du Génome. Aujourd'hui, ces Génopoles, à Lille, Strasbourg, Lyon, Marseille, Montpellier, Toulouse ou Nantes, ont deux ou trois ans d'existence. Elle ont fédéré – à des degrés divers – les laboratoires existants, leur ont permis d'accéder aux plateaux techniques indispensables pour faire de la vraie génomique et ont suscité l'émergence de nombreuses start-up. *Mais ces acquis – soulignés par une récente expertise scientifique confiée à l'Organisation Européenne de Biologie Moléculaire (EMBO), pourtant très sceptique au départ – sont à leur tour menacés. Près d'un an et demi après l'installation du gouvernement Raffarin, celui-ci ne semble pas avoir défini concrètement sa position vis-à-vis de la Génomique et des Génopoles. Certes, de bonnes paroles ont été prononcées, mais les actes (et les crédits de paiement) ne suivent pas : sur le plan budgétaire, le « réseau des Génopoles » est en* stand-by *depuis le printemps 2002.... On retrouve ici de manière criante cette absence souvent critiquée de continuité dans la politique scientifique française.*

Référence

1. Jordan B. La belle et triste histoire du « Programme Génome Français. In : *Génétique et Génome, la fin de l'innocence*. Paris : Flammarion, 1996.

This page intentionally left blank

7. L'IRRUPTION DES PUCES

La technologie des réseaux ou arrays *est très probablement la retombée méthodologique la plus importante des programmes Génome. Ces projets, imposant l'étude simultanée et, si possible, quantitative de nombreux objets, ont poussé à l'introduction de méthodes « parallélisables » qui ont progressivement remplacé les approches linéaires (un gène, une question, une variable à la fois) régnant jusque-là en biologie moléculaire. L'exemple* princeps *de cette mutation est le remplacement du* Northern blot *(mesure qualitative du niveau d'expression d'un gène dans quelques tissus, au prix d'une manipulation délicate et s'étendant sur plusieurs jours) par la puce à ADN qui autorise une évaluation semi-quantitative des niveaux d'expression de milliers de gènes en 24 heures, parfois moins. Les chroniques qui suivent présentent l'évolution de ces méthodes, de 1998 (elles étaient alors très peu pratiquées en France) à 2002.*

Med Sci *1998 ; 14 : 1097-1102*

VOYAGE AU PAYS DES PUCES

Première chronique sur les puces à ADN dans médecine/sciences*, à une époque où, en France, cette technologie était considérée à la fois comme inaccessible et toute-puissante. Ce texte, rédigé à la suite d'un colloque spécialisé tenu à San Francisco, faisait le point sur ce secteur et en montrait les potentialités tout en tentant de dissiper certaines illusions...*

Les « puces à ADN » continuent à faire l'objet d'une attention soutenue de la part des industriels et des médias. Chacun vante son approche : *microarrays* sur lame de verre pour *Synteni*, *Molecular Dynamics* et bien d'autres, centaines de milliers, bientôt millions d'oligonucléotides que la firme *Affymetrix* fait pousser *in situ* par des techniques proches de celles employées pour la fabrication des microprocesseurs, sans oublier diverses versions des bonnes vieilles membranes Nylon qui ont sans doute encore un bel avenir devant elles à en juger par le nombre de firmes qui les commercialisent sous des formes variées. Dans ce secteur où l'industrie est présente en force, la communication commerciale prend parfois le pas sur l'information scientifique et, malgré le paragraphe légal de rigueur sur les *forward-looking statements*[1], il est souvent difficile de savoir où en sont effectivement ces différentes approches. Je vais faire le point, en me limitant à la mesure du niveau d'expression de gènes et en m'appuyant sur les informations glanées lors d'un récent colloque[2].

Pourquoi donc s'intéresser autant à ces mesures d'expression ? Deux raisons s'imposent. Savoir à quel niveau s'exprime un gène dans différents tissus, différentes situations physiologiques ou pathologiques, c'est indubitablement faire un pas vers la compréhension de sa fonction. Information limitée, l'expression ne dit pas tout, et ses indications sont parfois ambigües ou trompeuses : nous connaissons tous des gènes à expression ubiquitaire, et qui pourtant jouent un rôle crucial dans un seul tissu[3]. Elle n'en reste pas moins précieuse. La deuxième raison est technique : ces données sont à l'heure actuelle les seules que nous soyons capables d'obtenir pour de grands ensembles de gènes, les seules qui puissent être rassemblées à l'échelle des dizaines de milliers d'entités révélées par le séquençage de banques d'ADNc ou de génomes entiers. Après le génome, le transcriptome[4] !

Le monde industriel a accumulé dans les coffres-forts virtuels de ses *proprietary databases* des millions de séquences partielles d'ADNc, déchiffrées par des entreprises

1. Ce paragraphe se cache à la fin des communiqués de presse triomphalistes et en module considérablement le sens – en résumé, il explique que ce qui est annoncé peut constituer une extrapolation et ne doit pas être pris à la lettre !

2. Congrès IBC *Biochip Technologies*, San Francisco (22 au 24 juin 1998).

3. Songeons par exemple à l'enzyme adénosine désaminase, exprimée dans la plupart des tissus mais critique pour le bon fonctionnement des lymphocytes et dont l'absence est responsable d'un déficit immunitaire sévère sans autre symptôme majeur.

4. Terme au sens assez flou, qui désigne la connaissance du patron d'expression de l'ensemble des gènes d'un organisme. Mais tout dépend de ce que l'on entend par profil d'expression : suffit-il de quelques tissus de base, ou faut-il avoir étudié les deux cents types cellulaires répertoriés dans l'organisme ? Et quid des variations d'expression dans les situations pathologiques ? Le transcriptome est donc une notion assez mal définie.

comme *Incyte* ou *Human Genome Sciences*, qui s'ajoutent au million d'EST humains disponibles dans dbEST [1] ; il a compris (et quelques *success-stories* l'ont démontré aux incrédules) que la valeur de cette information pouvait être décuplée par la connaissance des patrons d'expression. L'étude conjointe des séquences, avec toutes les ressources d'une bioinformatique de pointe, et des tissus dans lesquels s'expriment les gènes correspondants, ainsi que la comparaison de ces profils entre tissus normal et pathologique, aboutissent à la définition de nouvelles protéines thérapeutiques dont certaines sont déjà en phase d'essai clinique [2].

Recherche cibles désespérément...

Plus généralement, ces travaux révèlent de nouvelles « cibles » *(targets)*, protéines jouant un rôle clef dans divers processus physiologiques ou pathologiques (inflammation, coagulation, auto-immunité...). Le but de l'industrie pharmaceutique consiste alors à synthétiser des molécules pouvant agir sur ces entités, et ainsi moduler le processus en cause ; elle dispose pour cela des ressources de la chimie combinatoire, capable de produire en peu de temps des centaines de milliers de composés, et de techniques de criblage sophistiquées *(high-throughput screening assays)* qui peuvent en quelques semaines déterminer quelles molécules interagissent avec la cible et sont donc susceptibles d'avoir une action thérapeutique. Notons à cet égard que le nombre total de cibles contre lesquelles sont dirigés les médicaments existants est évalué à environ quatre cents seulement ; les analystes du domaine estiment que la Génomique en a déjà révélé un nombre équivalent, et que leur nombre total sera de plusieurs milliers d'ici deux ou trois années. On comprend que l'industrie s'intéresse de très près à l'analyse à grande échelle des niveaux d'expression qui, avec la connaissance des séquences et la bioinformatique, joue un rôle central dans la mise en évidence de ces entités.

Sur le plan technique, on assiste à la domination des méthodes fondées sur l'hybridation de sondes complexes avec de grands jeux de segments d'ADN fixés sur un support. Le *differential display* [3] et ses variantes ont, malgré leur difficulté expérimentale, leurs artefacts et leurs faux positifs, rendu de grands services dans les laboratoires. Ces systèmes, bien adaptés à la recherche d'un ou deux gènes fortement modulés lors du passage d'une situation à une autre, sont à ce titre pertinents pour maint projet de recherche académique ; mais leur débit est trop faible, leur fiabilité trop aléatoire pour des études à grande échelle. L'ingénieuse technique *SAGE* [4], réduisant chaque gène exprimé à une information minimum, un *tag* d'une quinzaine de nucléotides, a tenté de nombreuses équipes ; les difficultés de mise en œuvre ont pour le moment empêché sa généralisation[5]. Son impact dans le domaine reste aujourd'hui marginal.

Les techniques en lice

Macroarrays, *microarrays* et « puces à ADN » : toutes ces méthodes reposent sur l'emploi d'une sonde complexe préparée à partir de l'ensemble des ARN messagers d'un tissu et hybridée avec un jeu ordonné *(array)* de milliers de cibles représentant chacune un gène différent. La grande force de cette approche – comme le souligne depuis bientôt dix ans Hans Lehrach (Berlin, Allemagne) [5] – est leur fondamental parallélisme (au sens informatique du terme). Chaque hybridation donne en effet un renseignement sur chaque gène représenté dans le jeu, même s'il se résume à « non exprimé à un niveau détectable dans ce tissu ». Ces informations sont cumulatives, puisque chaque niveau d'hybridation

5. Dans des mains expertes, elle donne néanmoins de bons résultats ; une entreprise, Genzyme (www.genzyme.com/sage), se consacre à sa diffusion et propose aux industriels la réalisation de « profils digitaux d'expression » à façon.

mesuré est relié à une cible, une séquence, un gène précis. De plus, rien ne s'oppose à ce que les collections de gènes mis en jeu s'approchent du nombre total de gènes humains – certaines des expériences actuelles n'en sont pas loin.

Les *macroarrays*

Voyons donc où en sont ces techniques. La plus ancienne (et qui n'a sans doute pas dit son dernier mot) est celle des « membranes à haute densité », qui tendent aujourd'hui à prendre le nom de *macroarrays* par référence à leur version miniaturisée, le *microarray*. Elle utilise des clones d'ADNc ou leurs produits de PCR disposés régulièrement sur des membranes de Nylon, hybridés avec des sondes complexes radioactives produites par rétrotranscription d'une préparation d'ARN messager ou total. L'hybridation, conduite en conditions d'excès de cible par rapport à la sonde, aboutit à la fixation sur chaque segment d'ADN d'une quantité de l'espèce correspondante de la sonde proportionnelle à son abondance dans le mélange d'ARN de départ ; l'ensemble des données est acquis de manière quantitative, généralement par un système à écran phosphore, et mesure ainsi le niveau d'expression de chacun des gènes représentés – plus précisément l'abondance relative de chacun des ARN messagers dans le mélange de départ.

Des robots aptes à la confection de *macroarrays* sont disponibles dans le commerce, les systèmes de détection sont en place dans de nombreux laboratoires, et cette méthode est à la portée de beaucoup d'équipes, d'autant que plusieurs fabricants commercialisent des membranes prêtes à l'emploi portant quelques centaines ou quelques milliers de gènes [6]. Sous sa forme la plus répandue, cette technique utilise des membranes d'assez grandes dimensions, avec une densité de quelques dizaines de dépôts par cm^2 et un espacement entre les *spots* supérieur à un millimètre, compatible avec une détection par des sondes marquées au ^{32}P ou, mieux, au ^{33}P. Une évolution vers la miniaturisation est en cours, avec un espacement réduit à deux ou trois cents microns et une détection colorimétrique [7] ou radioactive (avec des systèmes à haute résolution). L'emploi de sondes fluorescentes est interdit par l'autofluorescence des supports existants. Cette méthode a pour elle sa flexibilité, l'excellente sensibilité absolue de la détection radioactive et sa large gamme dynamique, ainsi que la capacité élevée des membranes Nylon qui permet la fixation de fortes quantités d'ADN-cible et abaisse ainsi le seuil de détection. Elle est employée par de nombreuses équipes « académiques » [8-11] et par plusieurs industriels, notamment *Hyseq* [12].

Les *microarrays*

Les méthodes plus récentes sont celles des *microarrays* et des puces à oligonucléotides ou *oligo-chips* (les puces à ADN proprement dites). Un *microarray* comporte quelques milliers de gènes représentés par des produits de PCR qu'a déposés un robot *ad hoc* sur une lame de verre traitée ; l'hybridation est effectuée avec une sonde complexe obtenue par rétrotranscription du mélange d'ARN en présence de nucléotides substitués autorisant une détection ultérieure (directe ou non) par fluorescence. L'acquisition des résultats est généralement effectuée par un système de balayage laser équipé d'une optique confocale, les signaux étant mesurés par un photomultiplicateur. Ce système a été mis au point, dans le secteur académique, par l'équipe de Pat Brown à Stanford qui a publié plusieurs articles très convaincants [13-16] et donne sur son site [17] des renseignements détaillés – y compris un manuel permettant en principe de construire son propre système ! L'approche a également été employée par plusieurs groupes industriels, notamment *Synteni* (entreprise fondée par un collaborateur de Pat Brown), et un système complet devrait bientôt être commercialisé par l'industriel *Molecular Dynamics* à un prix de l'ordre de 1,5 MF. Les *microarrays* ont l'avantage de la compacité : on peut envisager

des jeux comportant cinquante ou cent mille cibles. De plus, le double ou éventuellement le triple marquage sont possibles, ce qui permet des comparaisons directes entre sondes différentes.

Les *oligo-chips*

Quant aux *oligo-chips*, ils ont été développés au départ dans une autre perspective, celle de l'encore mythique « séquençage par hybridation » [18-20]. Des milliers d'oligonucléotides de séquence connue sont greffés sur une petite surface de verre ou de silicium ; la lecture de la puce, après hybridation avec un ADN marqué par fluorescence, donne des informations sur sa séquence. Les *chips* vendus par *Affymetrix* [21] sont destinés à des applications de recherche de mutations (VIH, p53, cytochrome C). Il s'agit là d'un « quasi-séquençage » dans lequel la puce porte un jeu d'oligonucléotides représentant la séquence « officielle » du gène considéré : l'hybridation avec une sonde provenant d'un prélèvement biologique révèle alors les éventuelles différences entre la séquence présente chez le patient et la séquence habituelle. Mais *Affymetrix* s'intéresse beaucoup à la mesure de niveaux d'expression, et a produit dans ce but des *expression chips* dans lesquels chaque gène est représenté par un jeu d'oligonucléotides répartis le long de sa séquence codante ; la viabilité de cette approche a été démontrée [22, 23]. Bien que cette façon d'évaluer des niveaux d'expression puisse paraître inutilement compliquée, elle bénéficie de l'impressionnant potentiel de miniaturisation de la technique : *Affymetrix* fabrique actuellement des puces d'un peu plus d'un centimètre de côté portant 320 000 oligonucléotides différents, susceptibles de mesurer le niveau d'expression de milliers de gènes, et affirme pouvoir passer bientôt à des puces portant des millions d'éléments. Comme les méthodes de fabrication des *oligo-chips* suivent de près celles de l'industrie des microprocesseurs dont on connaît les prodiges récents en termes de miniaturisation, ces affirmations sont à prendre au sérieux.

Problèmes et limites de ces méthodes

Quelques précisions sont néanmoins nécessaires pour apprécier correctement l'état de l'art. Mentionnons d'abord un problème commun aux deux premières méthodes, celui de l'identification des cibles ou, plus précisément, de leur vérification. *Macroarrays* et *microarrays* se fondent en général, pour constituer leurs jeux de gènes, sur les séquences complètes et partielles déjà publiées, et représentent chaque gène par un clone d'ADNc (en fait, son produit de PCR) issu de la collection rassemblée par le consortium IMAGE [1], dans laquelle chaque clone est rattaché à au moins une séquence dans dbEST. Malheureusement, cet assortiment est entaché de multiples erreurs et 10 % à 20 % des clones IMAGE ne correspondent pas à leur séquence[6]... Pour bien faire, chaque clone destiné à un *array* doit donc être séquencé à nouveau, ce qui alourdit considérablement les projets. Les *oligo-chips*, eux, échappent à cet écueil ; mais ils ne peuvent naturellement inclure que des entités déjà séquencées, au contraire des micro- et macro-*arrays*.

Un deuxième point est celui de la sensibilité et de la masse d'échantillon nécessaire pour préparer une sonde complexe. Ces deux paramètres sont interdépendants : la détec-

6. C'est-à-dire que le clone IMAGE obtenu d'un distributeur n'est pas celui qui a été séquencé pour dbEST : erreur d'identification de clone ou de rattachement de fichier qui souligne la difficulté à gérer rigoureusement des collections aussi étendues (un million et demi de clones actuellement, toutes espèces confondues...). Cela met sans doute aussi en évidence un préjugé nord-américain lié à la stratégie des STS, selon lequel seule l'information de séquence est essentielle, l'entité « clone » restant secondaire. Lors du lancement du programme d'EST pour la souris (qui a aujourd'hui produit plus de 300 000 séquences partielles), il n'était même pas prévu de conserver les clones et c'est seulement devant le tollé des utilisateurs que cette phase fut incluse – sans doute à contre-cœur, et peut-être pas avec tout le soin nécessaire. Notons aussi l'annonce récente de la contamination de nombreux clones IMAGE par un bactériophage...

tion est d'autant plus sensible (en termes de détection d'espèces rares) que la concentration de sonde est élevée, puisque les signaux lui sont proportionnels (c'est même le principe de la mesure). Et les quantités mises en jeu à ce niveau sont comparables pour les trois méthodes, de l'ordre du microgramme d'ARN messager. Quantité élevée, qui interdit par exemple d'effectuer une analyse à partir d'une biopsie ou de quelques dizaines de milliers de cellules obtenues par tri cellulaire, et qu'il est urgent de réduire. Regardons de plus près ce résultat à première vue paradoxal, puisque les concentrations de la sonde complexe diffèrent considérablement, les volumes d'hybridation se comptant en millilitres d'un côté *(macroarrays)*, en microlitres de l'autre *(microarrays, oligo-chips)*. Un calcul à partir des données publiées [11, 16, 23] montre pourtant que le seuil de détection (le nombre minimum de molécules d'une espèce spécifique qui doit être présent dans la sonde afin d'obtenir un signal détectable) est du même ordre, soit quelques dizaines de millions (Bertucci, communication personnelle). Cela découle du fait que la quantité de cible fixée sur Nylon est beaucoup plus importante que sur lame de verre[7], ainsi que de la meilleure sensibilité intrinsèque de la détection radioactive. De plus, l'agitation du milieu d'hybridation dans lequel baigne la membrane évite tout épuisement local de la sonde, contrairement à ce qui se passe pour les *microarrays* souvent hybridés sous une lamelle de microscope avec un volume total d'une dizaine de microlitres. Naturellement, l'amplification de la sonde pratiqué par *Affymetrix* [22, 23] recule cette limite ; mais, en l'absence de résultats publiés prouvant qu'elle ne fausse pas les valeurs d'abondance relative mesurées, cette méthode reste contestable.

Venons-en maintenant à la validation des mesures. Il ne suffit pas de collecter des dizaines de milliers de niveaux d'expression. Il convient aussi de garantir la validité des chiffres obtenus, l'absence de divers artefacts qui peuvent les affecter, et de connaître les limites de confiance des différentiels observés. On estime aujourd'hui que, dans des expériences bien conduites (quelle que soit l'approche), une variation d'un facteur deux pour le niveau d'expression d'un gène donné est significative. La comparaison entre une large série de sondes, de tissus ou de conditions demande plus de soins et de références, soit absolues, soit relatives pour aboutir à des données fiables.

Reste enfin l'interprétation détaillée des résultats. La masse d'informations qui résulte de ces expériences est considérable : le niveau d'expression de milliers de gènes est mesuré dans plusieurs tissus et/ou sous diverses conditions. Examiner ces données, les appréhender, les comprendre, n'est pas un problème trivial. De nombreux développements sont encore nécessaires au niveau des logiciels d'analyse et de représentation, afin d'extraire de ces chiffres le maximum de sens – et aussi de les rendre disponibles à la communauté des biologistes.

Les grandes tendances

Après ces remarques critiques, qui veulent donner de la réalité une vision plus nuancée que les communiqués parfois triomphalistes d'entreprises en compétition féroce sur un marché porteur, reste à appréhender l'avenir de technologies qui continuent d'évoluer rapidement et font l'objet d'investissements considérables. Plusieurs tendances étaient perceptibles lors du colloque déjà cité. Je les regrouperai en trois thèmes : début de « sectorisation » du marché, accent sur l'exploitation des données, émergence possible de méthodes révolutionnaires.

7. Le signal observé est proportionnel non seulement à la concentration de la sonde, mais aussi à la quantité de cible : la probabilité d'hybridation pour chaque molécule de cible est indépendante du nombre de ces molécules, donc les signaux s'ajoutent à la seule condition que les interactions entre ces molécules soient négligeables, ce qui est le cas tout au moins sur support Nylon où les molécules d'ADN occupent moins de 1 % du volume du dépôt.

Sectorisation du marché : expression barbare pour indiquer que, du moins pour les industriels concernés, il semble se profiler deux champs d'application, celui des puces complexes présentant un très grand nombre d'éléments et pouvant, par exemple, mesurer l'expression de dix mille gènes simultanément, et celui de systèmes plus simples visant quelques centaines d'entités. Le fait que la firme *Affymetrix* détienne des brevets lui assurant (d'après ses avocats) l'exclusivité d'*arrays* comportant mille éléments ou plus joue sans doute un rôle ; mais il est vrai aussi que beaucoup d'applications, notamment dans le domaine clinique, peuvent se contenter de suivre une ou deux centaines de gènes, et que le prix des puces complexes (de l'ordre du millier de dollars)[8] est à cet égard dissuasif. Des entreprises comme *Genometrix* [24] font cette analyse et s'équipent de manière à produire massivement des puces simples et à les commercialiser pour quelques dizaines de dollars l'unité, tout en proposant des systèmes de lecture ne faisant pas appel à une optique confocale, donc relativement abordables.

Accent sur l'exploitation des résultats : après l'apparition de systèmes pouvant mesurer des milliers de niveaux d'expression par expérience, les équipes qui ont commencé à les utiliser ont été rapidement submergées par l'abondance des données et ont dû créer des méthodes d'analyse permettant de les traiter et d'aboutir – si possible – à des conclusions biologiques. Le *National Cancer Institute* (NCI au NIH) a lancé un projet sur la pharmacologie moléculaire du cancer [25] dans le cadre duquel l'expression de dix mille gènes (mesurée à l'aide de *microarrays* fabriqués par l'entreprise *Synteni*) est maintenant suivie dans soixante lignées représentant les principaux types de cancer. Cette masse de données (six cents mille valeurs...) est corrélée avec la sensibilité de ces différentes lignées à un large jeu de molécules. Des analyses de corrélation sophistiquées permettent alors de déceler, par exemple, les cas où la forte expression d'un groupe de gènes dans certaines lignées est associée avec la sensibilité à une certaine drogue. Les gènes en question deviennent alors des cibles potentielles pour ce composé, pouvant éclairer l'étude de son mode d'action...

Un des plus beaux exposés de cette conférence fut celui de Michael Byrne (de l'entreprise *Genetics Institute*, en collaboration avec plusieurs équipes académiques). Il s'agit cette fois de travaux effectués grâce à des *oligo-chips* produits par *Affymetrix*, avec qui *Genetics Institute* collabore depuis trois ans, et qui touchent des thèmes d'intérêt immunologique et médical. Une illustration, l'étude de l'activation de cellules T/NK dans le diabète, menée avec le jeu de quatre chips *Affymetrix* qui permet de suivre l'expression de 6 800 gènes humains. Le matériel de départ est privilégié puisqu'il s'agit de lignées de cellules T/NK, établies à partir de deux jumeaux monozygotes dont l'un seulement est diabétique. Les profils d'expression ont été mesurés, dans les deux cas, avant et après stimulation des cellules par un anticorps anti-CD3. Chez la personne non diabétique, une augmentation d'expression très nette est observée pour plusieurs dizaines de gènes – ce qui, au passage, donne pour la première fois une vision exhaustive du processus à ce niveau et met en évidence plusieurs effets nouveaux – ; chez son jumeau, la plupart de ces effets sont absents ou fortement réduits. Dans cette situation très bien définie (puisque, naturellement, le *background* génétique est identique), on a donc maintenant une douzaine de candidats, parmi les 6 800 gènes humains connus, à une implication dans le mécanisme pathologique.

J'évoquerai enfin un dernier exemple d'application large des profils d'expression. Celui-ci déclenche, du moins chez le chercheur « académique », un fort sentiment de frustration, lié au caractère secret des (très intéressantes) données obtenues. L'entreprise *Synteni* a été rachetée par la très performante *Biotech Incyte* [26], qui a fait sa spécialité de la détermination d'EST et de leur inclusion dans des bases de données perfectionnées – accessible aux industriels moyennant une redevance d'une dizaine de millions de dollars

8. Notons que toutes ces puces sont présentées comme à usage unique, ce qui est naturellement dans l'intérêt de leurs fabricants...

[27]. *Synteni* mesure actuellement sur une série de quatre *microarrays* le profil d'expression de quarante mille gènes humains[9] dans divers tissus normaux et pathologiques, en conditions de *stress*, après traitement par des drogues... et incorpore ces résultats dans une *UniGEM Expression Database* reliée aux bases de données d'*Incyte*, et dont l'accès sera vendu – certainement très cher – aux compagnies intéressées. On touche là un point déjà souligné [27] : ces informations de nature fondamentale, potentiellement utiles pour toute la Biologie, sont obtenues dans le secteur privé... et n'en sortiront pas. Décidément, la quête de nouvelles cibles amène les laboratoires industriels à entrer en concurrence directe avec la recherche fondamentale académique... mais avec de tout autres moyens !

Nouveautés et perspectives

Émergence de techniques révolutionnaires enfin : la plus spectaculaire à mon sens était présentée par une équipe de la *Tuft University*. Il s'agit de la réalisation de *microarrays* sur fibres optiques. D'un diamètre total inférieur au millimètre, la fibre comprend dix mille filaments individuels ; à l'extrémité de chacun, une cavité de quelques microns dans laquelle se trouve une microsphère conjuguée à un segment d'ADN. Lorsqu'on trempe l'extrémité du faisceau dans quelques microlitres d'une solution contenant un mélange de fragments d'ADN fluorescents, l'hybridation a lieu et la fluorescence correspondante peut être excitée par une illumination laser et mesurée individuellement pour chaque microfibre. Les résultats montrés pour cette méthode en cours de développement et dont le principe a déjà été publié fin 1996 [28] sont assez convaincants et en démontrent la faisabilité, même si de nombreux détails pratiques restent à résoudre. En extrapolant un peu, il est possible d'imaginer des fibres encore plus fines autorisant des mesures... à l'intérieur d'une cellule ! Science-fiction certes, mais qui illustre le bouillonnement d'idées et la qualité de l'innovation dans ce secteur. Signalons aussi de nombreux travaux sur la détection de l'hybridation par des méthodes électriques (et non par la mesure de fluorescence), susceptibles d'accroître considérablement la vitesse d'acquisition, de lever les derniers obstacles à une miniaturisation encore plus poussée et d'aboutir à des puces intégrant les segments d'ADN et le dispositif de mesure.

Tirons maintenant quelques leçons de ce périple au pays des *macroarrays*, *microarrays* et puces à ADN. Il est clair qu'aujourd'hui la mesure du profil d'expression de dizaines de milliers de gènes dans des dizaines de conditions différentes est techniquement possible, que son utilité est scientifiquement et industriellement incontestable, et que les outils d'analyse permettant d'en déduire des cascades de signalisation ou de métabolisme commencent à apparaître. Il est évident aussi que la majorité des projets menés le sont dans un cadre et avec des financements industriels. Leurs résultats ne seront pas généralement disponibles, ce qui est fort dommage pour nous tous... Notons tout de même les quelques travaux (comme ceux du NCI) destinés à être rendus publics, et souhaitons que des financements (ainsi que des équipes compétentes et motivées) soient trouvés pour qu'un ensemble cohérent de résultats aboutisse dans des bases de données publiques. Dans un registre plus optimiste, il est permis d'espérer que l'amélioration de ces méthodes débouche sur une accessibilité accrue. Celle-ci est aujourd'hui réduite, peu d'appareils sont disponibles et leur prix est encore prohibitif, mais on peut compter sur une intense concurrence entre les industriels qui, à l'instar de *Genometrix*, s'attachent à mettre au point des versions *light* et abordables de cette technologie.

Restera à intégrer concrètement ce versant de l'expérimentation biologique parmi les approches que nous pratiquons au laboratoire. Cela suppose une double stratégie.

9. Cet ensemble comporte la petite dizaine de milliers de gènes actuellement « connus » (dont la séquence complète, le produit et souvent la fonction ont été déterminées), et environ trente mille *clusters* constitués, comme dans Unigene (http://www.ncbi.nlm.nih.gov/UniGene/index.html), par regroupement informatique des séquences d'EST obtenues par Synteni.

D'une part, un investissement massif dans la bioinformatique, de manière à faire le meilleur usage des données qui seront rendues publiques ; d'autre part, la mise en place d'équipements bien dimensionnés, permettant de réaliser des *arrays* de quelques milliers de gènes adaptés aux besoins de nos équipes, et de les exploiter de manière efficace. Inutile à cet égard de vouloir faire la course avec *Affymetrix* ou *Synteni* ; pas question pourtant de se priver d'employer cette technologie, d'autant qu'elle va devenir de plus en plus accessible. Et, bien sûr, il faudra veiller à éviter de se placer en compétition directe avec la recherche industrielle et chercher, au contraire, une complémentarité fondée sur la connaissance intime de différents systèmes biologiques...

Références

1. http://www.ncbi.nlm.nih.gov/dbEST/index.html.
2. Marshall A. HGS launches « first » genomics product in clinic. *Nat Biotechnol* 1998 ; 16 : 129.
3. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992 ; 257 : 967-71.
4. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995 ; 270 : 484-7.
5. Lennon GG, Lehrach H. Hybridization analyses of arrayed cDNA libraries. *Trends Genet* 1991 ; 7 : 314-7.
6. Clontech « Atlas array » : http://www.clontech.com/clontech/Catalog/Hybridization/Atlas.htmlG.
   Genome Systems « Gene discovery array » : http://www.genomesystems.com/GDA/.
   Research Genetics « GeneFilters » : http://www.resgen.com/.
7. Chen JJW, Wu R, Yang PC, *et al.* Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* 1998 (sous presse).
8. Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach H. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm Genome* 1992 ; 3 : 609-61.
9. Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, Jordan BR. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* 1995 ; 29 : 207-15.
10. Zhao N, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis : a novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995 ; 156 : 207-13.
11. Pietu G, Alibert O, Guichard V, Lamy B, Bois F, Leroy E, Mariage-Samson R, Houlgatte R, Soularue P, Auffray C. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996 ; 6 : 492-503.
12. http://www.hyseq.com/.
13. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995 ; 270 : 467-70.
14. Derisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996 ; 14 : 457-60.
15. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis : microarray-based expression monitoring of 1 000 genes. *Proc Natl Acad Sci USA* 1996 ; 93 : 10614-9.
16. Derisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997 ; 278 : 680-6.
17. http://cmgm.Stanford.EDU/pbrown/.
18. Khrapko KR, Lysov YuP, Khorlin AA, Ivanov IB, Yershov GM, Vasilenko SK, Florentiev VL, Mirzabekov AD. A method for DNA sequencing by hybridization with oligonucleotide matrix. *DNA Seq* 1991 ; 1 : 375-88.
19. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991 ; 251 : 767-73.
20. Southern EM, Maskos U, Elder JK. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides : evaluation using experimental models. *Genomics* 1992 ; 13 : 1008-17.

21. http://www.affymetrix.com/.
22. Lockhart DJ, Dong H, Byrne MC, *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996 ; 14 : 1675-80.
23. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart, DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* 1997 ; 15 : 1359-67.
24. http://www.genometrix.com/.
25. Weinstein JN, Myers TG, O'Connor PM, *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997 ; 275 : 343-9.
26. http://www.incyte.com/.
27. Jordan B. Du programme Génome à la « pharmacogénomique ». *Med Sci* 1997 ; 10 : 1176-7.
28. Ferguson JA, Boles TC, Adams CP, Walt DR. A fiber-optic DNA biosensor microarray for the analysis of gene expression. *Nat Biotechnol* 1996 ; 14 : 1681-4.

Med Sci *2000 ; 16 : 950-953*

JUSQU'OÙ IRONT LES PUCES ?

Été 2000 : la technologie des puces à ADN a maintenant pris pied en France, de nombreux laboratoires sont équipés (bien que pas toujours opérationnels), et d'autres questions se posent : comment traiter les résultats, comment « faire du sens », obtenir des informations biologiquement significatives à partir de l'avalanche de données ainsi obtenues ? Et comment va évoluer la technologie, dans laquelle de ses variantes est-il le plus judicieux d'investir ? Tels sont les points sur lesquels cette chronique essaie d'éclairer le lecteur.

Les mesures d'expression grâce à l'emploi de réseaux d'ADN continuent à avoir le vent en poupe. Des articles majeurs montrent l'apport de ces méthodes dans des questions de recherche fondamentale [1] ou clinique [2]. Les colloques spécialisés fleurissent, tandis que l'offre commerciale d'*oligo-chips* et de *microarrays* prêts à l'emploi commence à s'élargir. De nombreux laboratoires français sont maintenant équipés – quoique pas toujours opérationnels, la mise en œuvre de ces méthodes s'avérant en général plus laborieuse que prévu.

Un rapide survol de la situation actuelle

Les différentes approches en présence sont maintenant bien connues [3, 4]. Rappelons-les brièvement avant de passer à l'exercice de prospective scientifico-technique qui est l'objet de cette chronique. Dans tous les cas la mesure d'expression repose sur l'hybridation entre un mélange d'ADNc marqués préparé à partir des ARN messagers de l'échantillon biologique, et un ensemble de segments d'ADN de séquence connue (appelés tantôt « cibles », tantôt « sondes » selon les auteurs) fixé sur un support. Après incubation et lavage le taux d'hybridation est quantifié pour chaque élément du réseau. Il est proportionnel à l'abondance relative de l'espèce d'ARNm concernée dans l'échantillon de départ, et donne donc une mesure du niveau d'expression du gène correspondant.

La première famille de méthodes repose sur l'emploi de produits de PCR obtenus à partir de jeux de clones d'ADNc représentant chacun un gène. Elle se décline d'abord en *macroarrays* (ou « filtres à haute densité »), réalisés sur des membranes de nylon d'une centaine de cm^2 de surface ; dans ce cas l'échantillon est presque toujours marqué de manière radioactive. Cette approche, la plus ancienne, reste assez performante surtout pour des jeux de gènes pas trop nombreux et pour des échantillons peu abondants, et offre l'avantage d'être compatible avec le matériel déjà présent dans les laboratoires. Les *microarrays*, version miniaturisée de cette approche, sont généralement réalisés sur verre ou sur silicium, et peuvent porter une dizaine de milliers d'éléments sur quelques centimètres carrés ; l'échantillon est dans ce cas marqué directement ou indirectement par fluorescence. Ces réseaux miniature peuvent aussi être réalisés sur nylon et révélés par colorimétrie, ou par radioactivité avec un lecteur à haute résolution. Dans ce dernier cas leur sensibilité permet la mesure à partir de très petits échantillons, ce qui est un avantage déterminant dans certaines situations [5]. *Macroarrays* et *microarrays* peuvent être confectionnés au laboratoire (pour autant que celui-ci dispose des robots *ad hoc* et

de l'expertise nécessaire pour les employer) ou achetés (fort cher) à des fabricants pour le moment peu nombreux. Notons qu'à part les réseaux sur support Nylon ces *microarrays* sont à usage unique.

Le deuxième type d'approche, proposé pour le moment presque exclusivement par la firme Affymetrix, emploie des centaines de milliers d'oligonucléotides synthétisés directement sur la « puce » pour effectuer le même travail. Le relatif manque de spécificité des 20- ou 25-mères utilisés est compensé par une redondance élevée, chaque gène étant représenté par une vingtaine d'oligonucléotides auquel s'ajoutent autant de témoins présentant une différence de séquence et permettant l'évaluation du bruit de fond. La fabrication de ces *oligo-chips* réclame une technicité et des investissements hors de portés des laboratoires de recherche : la méthode photo lithographique employée requiert, pour une puce comportant des 20-mères, quatre-vingts (20×4) masques permettant l'illumination sélective des différentes régions de la puce au cours des étapes de synthèse. Les *oligo-chips* sont donc obligatoirement achetés, à un prix unitaire de l'ordre de 5 000 à 10 000 francs – et l'expérimentateur s'engage à ne même pas essayer de les réutiliser...

Chacune de ces méthodes a ses avantages et ses inconvénients, en termes de performances, de coût, de commodité... Mais la situation est loin d'être figée, et des évolutions en cours permettent de penser que les procédures employées dans deux ou trois ans seront assez différentes de celles d'aujourd'hui. Aussi dans cette chronique vais-je me risquer – le jeu est dangereux... – à indiquer les directions dans lesquelles il me semble que l'analyse du transcriptome va s'engager dans un futur relativement proche.

Vers des puces « génome entier » ?

Avec l'obtention de la séquence humaine, le désir de disposer de réseaux permettant d'analyser simultanément l'expression de l'ensemble des gènes humains (ou murins) va logiquement s'exacerber, la disponibilité tant de la séquence que de grands jeux de clones facilitant en principe une telle réalisation (déjà effective dans le cas des 6 400 gènes de la levure). Il faut pourtant souligner que la technologie actuelle n'a pas encore atteint le degré de miniaturisation permettant de représenter l'ensemble des gènes humains[1] sur une surface inférieure à celle d'une lame de microscope. Les *microarrays* les plus denses comportent actuellement de l'ordre de 2 000 dépôts par cm^2, soit environ 20 000 sur la surface utile d'une lame de microscope – sans compter les duplicats et les contrôles. Il s'agira donc plutôt d'un jeu de quatre ou cinq lames représentant l'ensemble des gènes humains – ce qui est déjà une performance remarquable. En fait, l'augmentation de cette densité ne poserait pas de problème insurmontable du point de vue de l'analyse puisque la résolution des systèmes confocaux employés pour la lecture atteint 10 ou même 5 micromètres ; mais il paraît difficile de réaliser en routine des dépôts d'ADN dont le diamètre soit inférieur à 100 micromètres[2], amenant en pratique à un « pas » de 200 micromètres qui correspond à la densité mentionnée plus haut. Des changements dans les systèmes de dépôt, et la mise au point de nouveaux types de surfaces, pourraient accroître cette densité ; il semble pourtant que la complexité et le coût de la préparation de dizaines de milliers de produits de PCR risquent en tout état de cause de freiner la diffusion de *microarrays* « génome entier ».

La situation n'est pas très différente actuellement pour les puces à oligonucléotides, mais elle est sans doute plus susceptible d'évolution. Les puces aujourd'hui commercialisées par Affymetrix portent environ 300 000 oligonucléotides. Néanmoins, la faible lon-

1. Ce nombre est très discuté, puisque les estimations actuelles vont de moins de 30 000 à 120 000. Je raisonne ici sur une valeur d'environ 50 000.

2. Un dépôt d'un tel diamètre correspond à environ un nanolitre de solution. Les têtes des imprimantes à jet d'encre sont capables de projeter des gouttelettes bien plus petites, quelques picolitres ; mais une solution d'ADN n'a pas la fluidité des encres spéciales mises au point pour cette application.

gueur de ceux-ci (conséquence du médiocre rendement des réactions photochimiques employée) impose d'en consacrer trente à quarante (contrôles compris) pour mesurer l'expression d'un seul gène, ce qui ramène le nombre de gènes analysables à moins de dix mille. Pour aller plus loin, il faudrait accroître la densité des puces et/ou réduire le nombre d'oligonucléotides par gène. Un meilleur choix des séquences de ces derniers, grâce à des progrès dans les algorithmes employés, peut permettre d'en réduire l'effectif – mais cette réduction ne sera que modérée afin de ne pas affecter la qualité de la mesure. L'augmentation du nombre d'éléments par puce est possible : les « plots » des puces Affymetrix sont actuellement de 24 par 24 micromètres, alors que l'industrie des microprocesseurs est descendue bien au-dessous du micromètre. Des puces comportant un ou plusieurs millions d'éléments sont donc possibles, mais c'est la lecture optique qui va alors atteindre ses limites en résolution et en sensibilité. De nouvelles techniques de synthèse *in situ (voir plus loin)* permettront sans doute de dépasser ces obstacles.

En tout état de cause, et au moins dans le futur immédiat, d'éventuels *microarrays* ou *oligo-chips* « génome entier » (humain ou murin) seront certainement très onéreux. De plus les problèmes d'acquisition, de stockage et d'interprétation des résultats seront importants compte tenu de la masse de données. De ce fait, l'emploi de réseaux d'ADN spécialisés représentant des jeux limités mais pertinents de gènes devrait rester tout à fait attractif.

La victoire prévisible des oligonucléotides

Les nouveaux venus au monde des *microarrays*, une fois qu'ils ont réuni les robots et les systèmes de lecture indispensables, découvrent souvent à leurs dépens la difficulté et le coût de la constitution de collections de milliers de clones d'ADNc validés, tout comme celui qu'entraîne la production de produits de PCR en qualité et quantité suffisantes. Divers problèmes (les erreurs dans la collection IMAGE, la contamination de nombreux clones par des bactériophages, sans parler des questions de propriété industrielle) compliquent encore la tâche. Sans doute pour les même raisons, il existe encore peu de *microarrays* commercialement disponibles. De ce côté, le statut un peu ambigu des clones IMAGE[3] tout comme la menace de procès de la part d'Affymetrix jouent sûrement un rôle de frein. En tout état de cause, les *microarrays* sont aujourd'hui chers (au moins 10 000 F pour 2 000 gènes), et il est probable qu'ils le restent : les économies d'échelle que l'on peut attendre de la fabrication de milliers de *microarrays* sont modestes, puisque les coûts, tant pour l'obtention des produits de PCR que pour l'opération de dépôt, sont en gros proportionnels au nombre de réseaux produits.

Les puces à oligonucléotides ne souffrent pas des mêmes problèmes. Fondées sur la connaissance de la séquence (dont le rythme d'accumulation ne cesse de s'accélérer), elles éliminent tout recours aux collections de clones d'ADNc. De plus, les économies d'échelle dans la fabrication peuvent être considérables. La firme Affymetrix pratique, à l'aide de machines automatiques, une série de 80 (20×4) cycles d'addition de bases afin de construire des jeux d'oligonucléotides (20-mères) sur des « gaufrettes » *(wafers)* de verre à partir desquelles sont ensuite découpées les puces individuelles. Les premières fabrications permettaient ainsi l'obtention, en quelques heures, de 49 *oligo-chips* à partir d'un *wafer* ; aujourd'hui, un jeu de 400 petites puces à très haute densité – portant chacun autant, ou même plus, d'oligonucléotides différents que les anciennes – est produit durant

3. Les clones IMAGE sont librement accessibles et distribués à tout laboratoire moyennant une simple participation aux frais de stockage et d'envoi. Mais leur utilisateur s'engage en principe à rendre public tout résultat obtenu grâce à leur utilisation, ce qui rend par exemple difficile pour une entreprise la vente à des industriels de réseaux à façon réalisés à partir de clones IMAGE, ces firmes ne souhaitant généralement pas publier leurs résultats. D'autres collections de clones présentent également des restrictions d'usage qui rendent leur emploi pour la fabrication de *microarrays* commerciaux difficile.

la même période à partir d'une gaufrette de même taille. Les économies d'échelle correspondantes n'apparaissent pas de manière évidente dans le catalogue de la compagnie, mais cette situation changera sûrement si la concurrence s'intensifie.

De fait, plusieurs laboratoires et entreprises mettent actuellement au point la production d'oligonucléotides *in situ* par des approches différentes, reposant par exemple sur l'addition très rapide de réactifs aux différents sites de la puce. Ces additions sont réalisées grâce à des systèmes proches des têtes d'imprimante à jet d'encre. De telles procédures utilisent la chimie « classique » de synthèse d'ADN, dont l'excellent rendement permet d'envisager la synthèse de grands oligonucléotides, 50 ou même 100-mères (Edwin Southern, communication personnelle). La spécificité d'hybridation est alors telle que le niveau d'expression d'un gène peut être mesuré avec un tout petit nombre de plots (ou même un seul) : du coup une puce de dimensions raisonnables pourra représenter un grand nombre de gènes. De surcroît, cette approche est beaucoup plus flexible que celle d'Affymetrix. La fabrication d'une puce « à façon » requiert simplement une programmation différente de la tête d'impression, et non la confection d'une série de 80 masques destinés à cibler la synthèse par voie photochimique.

Le développement de ces technologies ne dépend pas seulement de facteurs scientifiques et techniques. La question des brevets est déjà très « chaude » dans ce domaine. La firme Affymetrix en détient une multitude[4] qui, selon son interprétation, lui permettent de s'opposer à la vente de pratiquement tout type de réseau d'ADN. Ces brevets sont néanmoins attaqués, notamment par Edwin Southern qui le premier décrivait l'emploi de « multitudes d'oligonucléotides attachés à des lames de verre » lors d'une communication au symposium de Cold Spring Harbor en 1989 [6]. Ce conflit – déjà pris en main par une pléiade d'avocats – sera, espérons-le, tranché d'une manière qui encourage la concurrence...

Il est en tous cas très probable que les puces à ADN à base d'oligonucléotides, fondées uniquement sur la connaissance de la séquence, seront de plus en plus utilisées dans le futur. Cela sera certainement le cas pour les réseaux « standard », dont l'exemple actuel est le jeu complet de gènes de la levure. Le règne des oligonucléotides pourra s'étendre au-delà, pour les puces spécialisées, à condition que les nouvelles méthodes de fabrication flexibles tiennent leurs promesses, facilitent la confection de réseaux sur mesure et en réduisent le coût.

De l'artisanat à l'industrie

Les puces à oligonucléotides discutées dans le paragraphe précédent, nécessitant outillage et expertise spécialisés, ne seront pas confectionnées par les laboratoires de recherche mais achetées auprès d'industriels. Même pour les *microarrays* fondés sur l'emploi de clones d'ADNc et de produits de PCR, une tendance vers l'emploi de produits commerciaux se fera sentir. Il n'est pas rentable pour des équipes ou même des instituts de recherche d'investir des ressources importantes pour construire des réseaux standard : une fois encore, le cas de la Levure est un bon exemple. Une telle tâche peut être assurée de manière plus efficace par l'industrie ou, dans certains cas, par des centres de ressources publics. Ce n'est pas pour autant que la fabrication de *microarrays* va disparaître des laboratoires : des réseaux à façon permettant de tester des jeux limités mais choisis de gènes resteront nécessaires dans beaucoup de situations expérimentales, et la flexibilité maximale sera atteinte en les réalisant au laboratoire. Il existera sûrement d'autres possibilités : certains fabricants disposant d'une technologie flexible pourront offrir la fabrication de jeux *ad hoc*, d'autres vendront des jeux de produits de PCR « prêts

4. Une interrogation du site du *US Patent and Trademark Office* (http://www.uspto.gov) sur les mots clés Affymetrix et oligonucleotide fait apparaître pas moins de 58 brevets...

à déposer ». La situation finale sera sans doute complexe, les jeux standard étant réalisés industriellement tandis que les réseaux spécialisés seront produits dans le cadre de différents arrangements entre le monde académique et l'industrie. Dans ce contexte, il est à l'évidence vital de standardiser les formats et les systèmes de détection afin que chaque réseau produit par l'industrie n'impose pas l'emploi de son système de lecture « propriétaire » coûtant au bas mot un demi-million de francs...

Du réseau isolé à la « puce-laboratoire »

La technologie des puces biologiques n'est pas limitée aux réseaux d'ADN. L'intégration de diverses fonctionnalités sur des *chips* dont les dimensions se mesurent en centimètres est bien engagée : ces systèmes peuvent gérer des liquides, effectuer des filtrations, des réactions de PCR et même réaliser des électrophorèses capillaires [7]. Leur mise au point est fortement poussée par les besoins des industries pharmaceutiques qui doivent aujourd'hui tester littéralement des millions de composés issus de la chimie combinatoire *(high throughput screening)*. Il leur faut effectuer ces essais très rapidement, de manière massivement parallèle et en utilisant le volume minimum de réactifs afin de limiter les coûts – d'où l'intérêt évident de la miniaturisation. À terme, et au moins pour les applications cliniques et industrielles, les mesures d'expression – portant probablement sur un nombre limité de gènes – auront vocation a être intégrées dans de tels systèmes. C'est par exemple la forme sous laquelle cette approche pénètrera dans les services d'oncologie clinique – si du moins les indications montrant l'intérêt clinique de telles données se confirment [2].

Du marquage préalable à la détection électrique

Le marquage fluorescent de l'échantillon est relativement complexe, interfère avec l'hybridation et impose des systèmes de détection très sensibles et coûteux ; l'emploi de la radioactivité pose problème dans de nombreux environnements et sa résolution reste limitée même avec les détecteurs les plus performants (et les plus chers). Il est souhaitable de pouvoir détecter l'hybridation, et d'en quantifier le degré, à l'aide d'une autre méthode. De préférence, ceci devrait passer par la mesure d'un signal électrique et, dans l'idéal, n'imposerait aucune modification de l'échantillon avant l'expérience. De nombreux groupes déploient des efforts dans cette direction [8, 9]. Les méthodes envisagées vont de la détection d'un changement subtil des propriétés électriques après hybridation jusqu'à des approches très « exotiques » comme l'emploi de microbalances « pesant » la masse supplémentaire du matériel hybridé, ou la détermination du nombre de molécules double-brin (donc hybridées) par microscopie à force atomique. La faisabilité de certaines de ces méthodes a été démontrée ; reste à savoir si elles pourront atteindre la sensibilité requise et le débit nécessaire. Il est probable que leurs premiers emplois se situeront dans des applications comme la détection de bactéries ou de mutations, pour lesquelles une réponse qualitative peut suffire ; elles seront éventuellement étendues ensuite à la mesure d'expression pour laquelle une quantification précise est requise.

Amélioration des méthodes d'analyse et centralisation de données normalisées

Les aspects informatiques et bioinformatiques sont très importants et n'avaient pas été suffisamment pris en compte au début de la révolution des réseaux d'ADN. Même aujourd'hui, la validation et l'analyse des données d'expression restent relativement grossières [10]. En outre, la plus grande partie des résultats reste indisponible en dehors du

laboratoire d'obtention (parfois même en son sein...), et les jeux de données présentés par certaines équipes sur leurs sites Web ne sont pas directement comparables entre eux faute d'un format commun. De grands efforts sont faits actuellement pour améliorer la situation en mettant au point des logiciels d'analyse plus sophistiqués. Ces derniers comportent à la fois des analyses statistiques, de corrélation et de *clustering*, et des liens directs vers les informations constamment actualisées disponibles sur le réseau. De surcroît, un travail important est en cours pour définir un format de données standard qui devrait permettre d'archiver les données d'expression et de les rendre disponibles à tous – à la manière dont ont été traitées les données de séquences. Il faut néanmoins avoir conscience que le problème des résultats d'expression est bien plus complexe, compte tenu des différentes méthodologies employées et de la qualité assez variable – et délicate à évaluer – de données obtenues. Malgré ces difficultés, les progrès vont certainement être rapides. Il devrait devenir possible de tirer plus d'informations des résultats d'expression *(data mining)* mais aussi d'obtenir un profil d'expression assez complet pour tout gène d'intérêt à partir d'une interrogation de quelques sites sur le réseau.

Le futur des mesures d'expression

On peut, sans risque, prévoir que la mesure d'expression va rester d'actualité. Certes, d'autres méthodes susceptibles de donner une information fonctionnelle vont être améliorées, rendues plus rapides et applicables à un plus grand nombre d'objets ; études d'interaction de protéines, protéomique en général, inactivation de gènes dans différents modèles animaux. Mais l'utilité de la mesure d'expression à grande échelle, facilitée par la disponibilité générale des informations de séquence et poussée par l'amélioration de la technologie des réseaux d'ADN, va certainement rester une approche majeure en biologie durant plusieurs années. Elle va sans doute aussi se banaliser dans un certain nombre d'applications cliniques ou industrielles, jouant dans ce cas un rôle tantôt complémentaire tantôt alternatif par rapport à l'identification par l'ADN ou la recherche de mutations.

Références

1. Iyer VR, Eisen MB, Ross DT, *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* 1999 ; 283 : 83-7.
2. Alizadeh AA, Eisen MB, Davis RE, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000 ; 403 : 503-11.
3. Jordan BR. Voyage au pays des puces. *Med Sci* 1998 ; 10 : 1097-102.
4. Granjeaud S, Bertucci F, Jordan BR. Expression profiling : DNA arrays in many guises. *BioEssays* 1999 ; 21 : 781-90.
5. Bertucci F, Bernard K, Loriod B, *et al.* Sensitivity issues in DNA array-based expression measurements : performance of Nylon microarrays for small samples. *Hum Mol Genet* 1999 ; 8 : 1715-22.
6. Southern EM, Maskos U. Synthesis of oligonucleotides tethered to a glass surface : applications in the analysis of nucleic acid sequences. *Genome Mapping and Sequencing Meeting*. New York : Cold Spring Harbor University Press, 1989 : 136.
7. Talary MS, Burt JP, Pethig R. Future trends in diagnosis using laboratory-on-a-chip technologies. *Parasitology* 1998 ; 117 (suppl) : S191-203.
8. Souteyrand E, Cloarec JP, Martin JR, *et al.* Direct detection of the hybridiation of synthetic homo-oligomer DNA sequences by field effect. *J Phys Chem B* 1997 ; 101 : 2980-5.
9. Wang J, Jiang A, Mukherjee B. New label-free DNA recognition based on doped nucleic-acid probes within conducting polymer films. *Anal Chim Acta* 1999 ; 402 : 7-12.
10. Ermolaeva O, Rastogi M, Pruitt KD, *et al.* Data management and analysis for gene expression arrays. *Nat Genet* 1998 ; 20 : 19-23.

Med Sci *2001 ; 17 : 893-896*

PUCES À ADN : LES BREVETS CONTRE LE PROGRÈS ?

*Ce titre un peu provocateur exprimait la révolte du scientifique devant la façon dont l'obsession de la « propriété industrielle » peut bloquer l'évolution d'un secteur et, en effet, ralentir le progrès. En ce qui concerne les puces à ADN, il s'agissait bien sûr d'*Affymetrix*, de ses innombrables brevets (126 à ce jour...) et de ses tentatives pour affirmer ses droits sur tout type de réseau. J'y étais (et suis toujours) particulièrement sensible, ayant été proche de Hans Lehrach qui a développé dès le milieu des années 1980 les premiers systèmes parallèles d'analyse de l'ADN (les « filtres à haute densité »), ayant assisté en 1989 à la présentation de Ed Southern sur ses* « oligonucleotides tethered to a glass surface »*... et ayant moi-même, dans mon laboratoire, élaboré dès 1992 la technologie des réseaux d'ADN pour la mesure d'expression.*

Encore les brevets...

Les brevets – surtout lorsqu'ils touchent au vivant – n'ont pas bonne presse : la plupart de nos concitoyens sont choqués lorsqu'ils découvrent qu'il est possible, en toute légalité, de breveter la séquence d'un gène humain, à partir du moment où celle-ci était jusque-là inconnue et où une hypothèse sur la fonction de la protéine correspondante est émise. Mais, avancent leurs défenseurs, ces brevets sont indispensables pour faire avancer la recherche : les investissements considérables que consentent les industriels pour leurs investigations ne peuvent être justifiés que s'ils gardent, pour un temps limité, un droit exclusif sur l'application commerciale de leurs découvertes. Ces arguments ne sont pas sans valeur, et le système des brevets n'est sûrement pas aussi noir que le disent certains de ses détracteurs. Pourtant, il dérape parfois de manière évidente, et en vient à inhiber tant la recherche que le jeu normal du marché. C'est, me semble-t-il, ce qui se passe actuellement dans le domaine très à la mode des puces à ADN. L'exemple est intéressant, car il montre comment cette institution peut devenir contre-productive, et cela d'un strict point de vue économique, sans faire intervenir le moindre jugement éthique.

Un peu d'histoire

Les « puces à ADN » sont aujourd'hui à la pointe de l'actualité : *macroarrays*, *microarrays*, puces à oligonucléotides, ces systèmes constituent un des espoirs de la génomique. Ils permettent en effet d'analyser sur des milliers de gènes l'effet d'un changement d'état physiologique, d'une infection ou d'un médicament. De fait, ils constituent à ce jour la seule méthode capable d'apporter une information de type fonctionnel à un rythme comparable à celui de l'obtention des séquences d'ADN. Ces puces sont à la mode depuis deux ou trois années, depuis qu'elles sont employées de manière relativement massive, et que les analystes financiers puis la grande presse ont découvert leur existence. En fait, elles remontent à la fin des années quatre-vingt, avec les « filtres à haute densité » popularisés par l'équipe de Hans Lehrach (qui travaillait alors aux laboratoires

de l'*Imperial Cancer Research Fund*, à Londres). Ces membranes de Nylon mesurant 22 par 22 cm et portant 9 216 clones d'ADN (plasmides ou cosmides) [1, 2] permettaient déjà d'accéder aux banques d'ADN (qui étaient alors une ressource rare) d'une manière très efficace. Rappelons que le procédé courant à l'époque comportait l'étalement de bactéries sur des dizaines de filtres disposés dans autant de boîtes de Pétri, suivi d'un laborieux traitement (fixation, hybridation, lavage, exposition...), le tout pour (peut-être) récupérer deux ou trois clones « positifs » avant de recommencer toute l'opération lors du prochain *screening*... Dès cette époque, l'idée d'utiliser ce format pour analyser simultanément le niveau d'expression de milliers de gènes était dans l'air [3], mais les premiers résultats devaient attendre le début des années 1990 [4].

À la même époque, l'ingénieux Edwin Southern – immortel inventeur du *Southern blot* qui, par analogie, a donné son nom au *Northern blot* puis au *Western blot* – faisait une communication au deuxième colloque *Genome Mapping and Sequencing* tenu à *Cold Spring Harbor* en mai 1989 [5, 6], dans laquelle il décrivait la fabrication de réseaux d'oligonucléotides fixés sur lame de verre, et montrait les premiers résultats obtenus lorsqu'on les hybridait avec des segments d'ADN. À l'époque, l'accent était sur le séquençage plus que sur la mesure d'expression : l'on espérait que le « séquençage par hybridation » allait bientôt détrôner la méthode de Sanger, espoir qui devait être déçu par la suite [7].

Pendant ce temps-là, Steve Fodor, alors inconnu de la communauté des biologistes, tentait vainement de mettre au point un procédé de synthèse des protéines sur un support solide à l'aide de réactions photochimiques. Ce travail n'ayant pas abouti, il se reconvertissait dans la synthèse d'ADN, et parvenait bientôt à développer une technique permettant de synthétiser *in situ* des oligonucléotides sur une lame de verre, comme l'avait fait Southern auparavant. Sa méthode était différente : elle gouvernait les réactions par des illuminations successives de différentes régions de la lame assurant ainsi l'activation de réactifs photosensibles. Cela autorisait la synthèse de nombreux oligonucléotides différents sur la même lame grâce à l'emploi de masques aux dimensions précises [8]. Notons que le rendement de synthèse est modeste, ce qui limite la longueur de ces segments à une vingtaine de bases. Cette méthode, proche à beaucoup d'égards de celles employées pour la fabrication des microprocesseurs qui font vivre nos ordinateurs, allait bénéficier pleinement des connaissances accumulées et de l'infrastructure mise en place par l'industrie de la microélectronique. Elle débouchait rapidement sur des « puces à oligonucléotides » comportant, dans une surface de l'ordre du centimètre carré, plusieurs dizaines de milliers de plots portant chacun quelques millions d'exemplaires d'un oligonucléotide de séquence définie à l'avance. Ces puces étaient au départ conçues pour effectuer un « quasi-séquençage », en fait une détection de mutation sur des gènes déjà connus : la puce « p53 », par exemple, comporte un ensemble de séquences correspondant aux exons de ce gène et permet, par la simple hybridation d'un produit de PCR obtenu à partir du sang d'un malade et marqué par fluorescence, de savoir si l'on est en présence de la séquence « standard », ou s'il existe des mutations, et lesquelles. Le procédé serait un peu plus tard adapté à la mesure d'expression.

Le début des mesures d'expression à grande échelle

Durant la première moitié des années quatre-vingt-dix, plusieurs équipes allaient progressivement mettre au point la mesure d'expression simultanée de milliers de gènes. Elles utilisaient essentiellement des produits de PCR provenant de clones d'ADNc, déposés sur des membranes de nylon et hybridés avec des échantillons radioactifs [9-11]. Les premiers résultats de l'équipe de Patrick Brown à Stanford [12], avec des ADN fixés sur des lames de verre et un marquage par fluorescence, constituaient, eux, le début d'une approche se prêtant mieux à la miniaturisation, et plus attrayante pour les industriels ou les cliniciens du fait de l'élimination de la radioactivité.

En 1996, l'entreprise *Affymetrix* fondée par Steve Fodor revendiquait, avec des résultats assez convaincants, la possibilité pour ses puces à oligonucléotides d'effectuer également des mesures d'expression [13], et en faisait bientôt son principal cheval de bataille. Parallèlement, une *start-up*, *Synteni*, fondée par un collaborateur de Patrick Brown, l'initiateur des réseaux d'ADNc sur lame de verre, proposait des services de mesure d'expression effectués sur ses *microarrays* comportant 5 000 puis bientôt 10 000 gènes. Les laboratoires et les industriels prenaient progressivement conscience de la puissance de ces techniques, les fabricants de robots et de systèmes optiques proposaient des appareils adaptés, bref, le mouvement était lancé. Après une période où il se publia beaucoup plus d'articles de revue sur les nouvelles méthodes d'expression que de résultats effectifs [14], les vrais travaux commençaient à apparaître ; il est clair aujourd'hui que cette approche, lorsqu'elle est bien menée, est très efficace et génératrice de résultats importants en recherche fondamentale et appliquée.

Un foisonnement de méthodes et d'initiatives

En cette année 2001, les méthodologies mises en œuvre sont très variées. Les *macroarrays*, membranes de plusieurs dizaines de centimètres carrés employées avec des sondes radioactives, gardent une bonne part de marché même si ce format est moins « tendance » que les systèmes miniaturisés ; d'ailleurs les fabricants qui les commercialisent ces membranes ont une curieuse tendance à les baptiser *microarrays*[1]. Les véritables *microarrays*, préparés sur lame de verre à partir de clones d'ANDc avec révélation fluorescente, sont largement employés, vendus prêts à l'emploi par plusieurs firmes ou réalisés en interne par les laboratoires qui se sont équipées des robots *ad hoc*[2]. Les *oligo-chips* d'*Affymetrix* existent en de nombreuses versions (homme, souris, drosophile, rat...) permettant à chaque fois l'analyse du niveau d'expression d'une dizaine de milliers de gènes et, malgré leur prix élevé (plus de mille euros l'unité), se vendent apparemment bien. De nouvelles techniques apparaissent, notamment des *microarrays* portant des oligonucléotides de grande taille (60 à 80 nucléotides, au lieu des 25 qui constituent la limite du procédé photochimique d'*Affymetrix*) qui promettent une compacité bien meilleure[3] ainsi que des mesures probablement plus fiables [15, 16]. Des approches plus « exotiques » se développent, puces actives de *Nanogen* ou systèmes totalement intégrés comme le *Geniom DNA processor* de *FeBIT* (http://www.febit.com), un appareil qui assure à la fois la fabrication du réseau, son hybridation avec l'échantillon et la lecture du résultat... Bref, un sain bouillonnement d'idées et d'initiatives comme peut en produire une économie de marché où toute bonne idée trouve à se financer et peut prendre une part de marché, pour autant qu'elle apporte un plus en regard de la technologie existante.

Un terrain miné

En fait, la situation est loin d'être aussi saine. Elle risque même d'être bloquée par des arguments juridiques qui cachent mal la volonté d'une entreprise de se réserver à tout prix ce marché prometteur, estimé à trois milliards de dollars pour l'année 2004. Cela manifeste aussi l'absurdité d'un système dans lequel les brevets sont à la fois tout puissants et accordés avec une légèreté confondante.

---

1. Pour éviter d'être moi aussi accusé de jouer sur les mots, je dirai qu'au-dessus de 1 mm d'écartement entre les centres des plots (100 plots au cm^2 ou moins) on est en présence d'un *macroarray* ; au-dessous de 500 µm (au moins 400 plots au cm^2), d'un *microarray*. Entre les deux, il y a effectivement ambiguïté.

2. On peut aussi réaliser des *microarrays* sur nylon, et c'est même une modalité particulièrement efficace du point de vue de la sensibilité [18].

3. Dans ce cas, il suffit d'un ou deux oligonucléotides pour mesurer le niveau d'expression d'un gène, au lieu de trente à quarante pour les 20-mères d'Affymetrix. Un array de 40 000 plots peut donc suivre 20 à 40 000 gènes.

On sait que toute *start-up* qui se respecte doit, pour boucler avec succès son tour de table et convaincre les « capitaux-risqueurs » de parier sur son avenir, détenir de la « propriété industrielle », autrement dit des brevets accordés, ou tout au moins déposés, qui lui assurent une certaine exclusivité sur un procédé ou un marché. L'entreprise *Affymetrix*, dès sa fondation, a joué ce jeu à fond et détient actuellement, aux États-Unis, plus de cent brevets. Ces derniers, que l'on peut consulter sur le site de l'Office Fédéral des Brevets (http://www.uspto.gov/patft/index.html), couvrent un champ très large, bien trop large en fait. À en croire *Affymetrix*, ils lui assurent l'exclusivité de tout réseau d'ADN comportant plus de 400 éléments au centimètre carré – autant dire, tous les *microarrays* existants[4]. Ces brevets, qui font bien sûr l'objet de contestations, sont le support d'attaques d'*Affymetrix* envers des entreprises accusées de les avoir enfreints. C'est ainsi qu'un procès est en cours depuis plusieurs années entre *Affymetrix* et *Incyte*, qui a racheté *Synteni* et repris à son compte les mesures d'expression à façon et, depuis peu, la vente directe de *microarrays*. Un autre procès, cette fois à l'initiative de la firme britannique *Oxford Gene Technology*, qui exploite les procédés de Southern, semble actuellement tourner à l'avantage d'*Affymetrix*. Il faut dire que cette entreprise, dont la capitalisation boursière représente aujourd'hui plus de deux milliards de dollars, a les moyens de payer de nombreux avocats, et détient un pouvoir d'intimidation considérable...

Pouvoir d'intimidation qui vient de pousser *Operon*, une filiale de *Quiagen*, à retirer ses *microarrays* du marché américain. Cette société, une de celles justement qui développe la voie intéressante que représentent les *microarrays* à base d'oligonucléotides longs, prenait déjà soin de ne commercialiser que des réseaux peu denses, comportant moins de 400 éléments par cm^2, afin justement d'éviter d'encourir les foudres d'*Affymetrix*[5]. Or une nouvelle interprétation des brevets suggère qu'ils pourraient couvrir aussi tout *array* comportant plus de 60 éléments par cm^2 (une densité de *macroarray* très moyen) : moyennant quoi *Operon* a décidé de cesser toute vente aux États-Unis [17].

On en arrive là à une situation ubuesque, où une société qui a certes développé un procédé très intéressant, mais n'a inventé ni les réseaux d'ADN ni la mesure d'expression à grande échelle, est apparemment en mesure de bloquer le système. Appuyée sur sa puissance financière et sur des brevets accordés de façon bien légère, elle semble capable d'empêcher le développement de procédés qui pourraient s'avérer plus efficaces que le sien. Les investissements nécessaires pour mettre au point, puis amener sur le marché de nouveaux types de puces à ADN sont importants, et les industriels (ou « capital-risqueurs ») hésiteront à les engager si la commercialisation de leurs produits est à la merci d'un concurrent puissant, pourvu d'un large assortiment de brevets et assisté d'avocats agressifs. Tout cela rappelle naturellement l'histoire de la PCR et des brevets détenus par Roche – qui ont ralenti pour un temps les exploitations commerciales de ce procédé, sans pourtant avoir de conséquences dramatiques pour la recherche : le brevet acquis par Roche n'empêchait pas l'emploi de la PCR dans les laboratoires académiques, et l'entreprise a su assez rapidement accorder des licences – certes onéreuses – plutôt que de bloquer le marché. Dans le cas des puces à ADN, c'est la mise au point de nouveaux procédés qui nécessite un marché ouvert, afin que les firmes y affectent les moyens nécessaires : elle peut donc être considérablement ralentie ou même stoppée dans une telle conjoncture.

4. Aux dernières nouvelles, ce serait même tout réseau dépassant la densité de 60 plots par centimètre carré, donc même les filtres à haute densité, qui existaient pourtant à l'époque où Fodor était encore (au sens figuré) en culottes courtes...

5. On pourrait penser que cela fait d'Operon un concurrent peu dangereux, face aux centaines de milliers d'oligonucléotides par cm^2 que propose Affymetrix. Mais il faut quarante 20-mères à Affymetrix là où un seul 70-mère suffit à Operon : le combat n'est donc pas si inégal...

La *biotech* malade des brevets

Cette situation est symptomatique d'une dérive actuelle de la biotechnologie, secteur où les brevets prennent une importance démesurée et où l'existence d'un peu de propriété industrielle (aussi fragiles que soient les revendications de brevets souvent encore en cours d'examen) compte souvent plus que la solidité scientifique du projet ou que la compétence des acteurs de la nouvelle entreprise. Dérive qui prend toute son ampleur aux États-Unis, où une importante population d'avocats pousse aux procès tous azimuts tandis que certains juges rendent des décisions aberrantes (comme ces trois milliards de dollars accordés récemment à un fumeur contre la firme *Phillip Morris*), décisions qui encouragent d'autres procès tout aussi déraisonnables. Elle montre aussi les risques d'abus de position dominante dans un secteur en plein développement. Certes, le procédé *Affymetrix* est efficace, l'entreprise a été la première à proposer à la vente des puces prêtes à l'emploi permettant la mesure d'expression pour des milliers de gènes, et son potentiel d'évolution vers une miniaturisation accrue reste important. Faut-il pour autant la laisser libre d'étouffer le développement d'autres approches ?

À cet égard, un programme lancé depuis la fin de l'année dernière par notre ministère de la Recherche pose, à la réflexion, quelques questions. Au départ, les intentions sont excellentes : puisque la technologie d'*Affymetrix* représente – personne ne le conteste – une des voies majeures pour l'analyse du transcriptome, le ministère souhaite qu'elle soit sérieusement testée en France et finance l'installation des machines correspondantes dans deux centres, à Strasbourg et à Paris. Comme le débit de ces systèmes est important, il importe de permettre à d'autres équipes françaises de les utiliser, ce qui pose le problème du prix des puces : même après le rabais « académique » consenti par *Affymetrix*, il reste très élevé, près de 1 000 euros par puce (à usage unique). Le ministère réserve donc des crédits à cet effet et procède à un appel d'offres afin que les meilleurs projets se voient subventionner le coût des puces à hauteur de 75 % du tarif académique, les 25 % restant à la charge du laboratoire. Encore une fois, tout cela découle des meilleures intentions ; mais cela revient tout de même à subventionner, avec l'argent public, l'entreprise *Affymetrix* afin de rendre ses produits plus attrayants pour nos laboratoires... Il est possible que cela assure à la firme entreprise une clientèle qui, ayant goûté à sa version de la technologie, aura tendance (à moins que les résultats ne soient catastrophiques) à poursuivre dans cette voie. Qu'en pensent les laboratoires, comme le LETI du CEA, qui tentent de mettre au point des puces de deuxième génération et vont avoir à affronter *Affymetrix* sur le marché (et peut-être devant les tribunaux) ?

Les chances d'une évolution positive

Il faut espérer que les concurrents d'*Affymetrix* ne se décourageront pas et poursuivront la mise au point et la commercialisation de nouveaux systèmes susceptibles d'apporter un plus aux laboratoires ou aux services hospitaliers. Heureusement, certains d'entre eux disposent à la fois de ressources importantes, du fait de leur dimension, et d'une technologie originale. *Corning*, le géant du verre (capitalisation boursière : 20 milliards de dollars), a mis au point un ingénieux procédé d'impression de *microarrays* qui lui permet de produire dix réseaux de 10 000 plots par minute ; *Motorola*, firme connue dans le secteur de la microélectronique (35 milliards) s'appuie, elle, sur un procédé original de puce à oligonucléotides. D'autres challengers, comme *Agilent* (ex *Hewlett-Packard*), *Incyte* ou même *Clontech,* ne sont dépourvus ni d'approches originales ni de moyens.

Il serait bon aussi que les offices de brevets cessent d'accepter des revendications exagérément larges et – balayons devant notre porte – que les brevets européens et français soient sérieusement examinés de ce point de vue. Tout le monde (ou presque...) aurait à gagner à un libre développement de ces méthodes qui jouent un rôle si important en Génomique...

Références

1. Monaco AP, Nizetic D, Craig A, *et al.* Human genome linking with cosmids and yeast artificial chromosomes. *Genome mapping and sequencing meeting*. New York : Cold Spring Harbor University Press, 1989 : 90.
2. Craig AG, Nizetic D, Hoheisel JD, Zehetner G, Lehrach H. Ordering of cosmid clones covering the herpes simplex virus type I (HSV-I) genome : a test case for fingerprinting by hybridisation. *Nucleic Acids Res* 1990 ; 18 : 2653-60.
3. Lennon GG, Lehrach H. Hybridization analyses of arrayed cDNA libraries. *Trends Genet* 1991 ; 7 : 314-7.
4. Gress TM, Hoheisel JD, Lennon GG, Zehetner G, Lehrach H. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm Genome* 1992 ; 3 : 609-61.
5. Southern EM, Maskos U. Synthesis of oligonucleotides tethered to a glass surface : applications in the analysis of nucleic acid sequences. *Genome mapping and sequencing meeting*. New York : Cold Spring Harbor University Press, 1989 : 136.
6. Southern EM, Maskos U, Elder JK. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides : evaluation using experimental models. *Genomics* 1992 ; 13 : 1008-17.
7. Jordan BR. Chroniques génomiques. Génome : les méandres de la technologie. *Med Sci* 2001 ; 17 ; 290-3.
8. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991 ; 251 : 767-73.
9. Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, Jordan BR. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* 1995 ; 29 : 207-15.
10. Zhao N, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis : a novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995 ; 156 : 207-13.
11. Pietu G, Alibert O, Guichard V, *et al.* Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996 ; 6 : 492-503.
12. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995 ; 270 : 467-70.
13. Lockhart DJ, Dong H, Byrne MC, *et al.* Expression monitoring by hybridization to high-density oligonucleotide Arrays. *Nat Biotechnol* 1996 ; 14 : 1675-80.
14. The chipping forecast. *Nat Genet* 1999 ; 21 (suppl 1).
15. Jordan BR. Chroniques génomiques. Jusqu'où iront les puces ? *Med Sci* 2000 ; 16 ; 950-63.
16. Lee P, Hudson TJ. La puce ADN en médecine et en science. *Med Sci* 2000 ; 16 : 43-9.
17. Jones MM. Operon discontinues sale of arrays in North America due to concerns about Affy patents. *BioArray News* 2001 ; 1 : 1-9.
18. Bertucci F, Bernard K, Loriod B, *et al.* Sensitivity issues in DNA array-based expression measurements : performance of Nylon microarrays for small samples. *Hum Mol Genet* 1999 ; 8 : 1715-22.

Med Sci *2002 ; 18 : 287-301*

PUCES-ACTUALITÉS

Pour terminer, une chronique plus récente qui insiste beaucoup sur les contrôles de qualité dans les mesures d'expression (thème un peu oublié dans l'enthousiasme des débuts) tout en présentant quelques nouveautés et en mentionnant le relatif piétinement des « puces à protéines ».

Un congrès spécialisé fournit souvent l'occasion de faire le point sur un domaine de recherche. Voici donc quelques informations tirées d'une récente réunion à Zurich, le colloque *Lab-on-a-Chip and Microarrays for Post-Genome Applications*, organisé pour la quatrième fois en janvier 2002 par le *Cambridge Healthtech Institute*. Cette raison sociale d'apparence académique recouvre, on le sait, une entreprise commerciale dont l'activité essentielle consiste à tenir chaque année une cinquantaine de colloques spécialisés. Leurs titres souvent évocateurs *(From proteins to profits, Effective Drug Discovery, Smarter Lead Optimization...)*, tout comme les tarifs d'inscription (1 000 à 2 000 €) pour les non-académiques, non compris le logement dans les hôtels de luxe où ils ont systématiquement lieu) montrent bien qu'ils s'adressent à un public composé en grande partie d'industriels. Ils ne sont pas pour autant dénués d'intérêt, au contraire... J'ai choisi de concentrer cette chronique sur trois domaines : la qualité des données obtenues, qui devient (enfin !) une préoccupation générale, la description rapide de quelques nouveautés prometteuses, et enfin l'état actuel plutôt décevant des *protein arrays* ou, pour parler français, des réseaux de protéines. Ce faisant, je passerai sous silence d'autres thèmes tout aussi intéressants et donnerai de ce domaine une vision partielle et personnelle, mais permettant au moins d'éclairer certaines tendances actuelles.

Qualité, mon doux souci...

Dans le célèbre livre de Robert Pirsig *Traité du Zen et de l'entretien des motocyclettes* [1], le narrateur donne un jour à ses étudiants la dissertation suivante : « Qu'est ce que la qualité ? ». Tout le monde se retrouve fort embarrassé pour répondre à cette question, et il ressort finalement des discussions que chacun sait reconnaître la qualité – mais que personne n'arrive à définir précisément en quoi elle consiste. Pour la mesure d'expression par réseaux d'ADN, les éléments fournis ont généralement été très limités : une ou deux belles images, accompagnées de chiffres censés quantifier la sensibilité du système et le plus faible différentiel d'expression mesurable. La sensibilité était souvent exprimée par l'abondance minimum détectable pour une espèce donnée d'ARN messager au sein du mélange extrait de la cellule ou du tissu analysés. Une valeur « magique » était d'ailleurs souvent avancée : 1/300 000, ce qui correspond approximativement à une molécule d'ARN messager par cellule de mammifère. Cette « sensibilité » dépend en fait directement de la quantité d'échantillon utilisée (plus on « charge » l'expérience en ARN, plus une espèce rare est détectable), dont la valeur précise était souvent cachée au fin fond des *Materials and Methods*. Quant au plus faible différentiel mesurable, il était généralement évalué à un facteur 2 sur toute la gamme de mesure, alors que la précision

dépend en fait très fortement de l'intensité mesurée, les fortes valeurs étant naturellement les plus précises en valeur relative. Tout cela n'était pas terriblement sérieux...

Aujourd'hui les préoccupations concernant la qualité des mesures sont beaucoup plus présentes. La nouveauté des *microarrays* s'est émoussée, la pratique de nombreux laboratoires a montré la puissance mais aussi le coût de cette technologie... et la facilité avec laquelle l'on peut se noyer dans une mer de données de qualité souvent douteuse. Les revues deviennent plus exigeantes, et les efforts engagés par plusieurs organismes pour organiser des bases de données d'expression les amènent tout naturellement à définir des normes de documentation et de qualité. Ces soucis étaient illustrés lors du colloque de Zurich par plusieurs présentations, notamment une étude détaillée décrite par Sophie Wildsmith (*GlaxoSmithKline*, Royaume-Uni) qui, dans le cadre de travaux de toxicologie sur le rat, montrait l'importance de la répétition des mesures. Une analyse de la variance globale entre expériences montre que la mise en évidence d'une différence significative entre les profils d'expression d'animaux témoins et traités (il s'agit de travaux de toxicologie sur le rat) réclame six répétitions de chaque hybridation. Je serais curieux de savoir combien des nombreux articles publiés dans ce domaine au cours des deux ou trois dernières années ont effectué autant de contrôles... Les avantages et inconvénients des différentes méthodes utilisables pour normaliser les données d'une expérience à une autre, la mesure de la déviation standard et de sa variation selon le domaine d'intensité considéré ont fait aussi l'objet de nombreuses présentations et discussions.

La qualité doit naturellement être présente dès le départ : qualité des ADN et de leur dépôt sur le réseau. On se souvient des difficultés rencontrées avec la collection de clones d'ADNc IMAGE, ressource très utile mais comportant souvent un nombre d'erreurs inacceptable [2], même dans les jeux de clones *sequence-verified* vendus par certains fournisseurs. On sait aussi que l'entreprise *Affymetrix* a dû l'année dernière reconnaître que de nombreuses séquences présentes sur l'une de ses puces « souris » avaient été lues à l'envers dans les bases de données. Mais même lorsqu'elles sont correctement conçues, les puces *Affymetrix* (et les autres) peuvent présenter des défauts, comme le montrait une intéressante étude présentée par Andreas Hohn, de l'entreprise *Gene Data* (Bâle, Suisse). *Gene data*, comme de nombreuses autres firmes, commercialise un ensemble de logiciels destinés aux utilisateurs de *microarrays*. Cette offre comporte des programmes capables d'analyser les images d'hybridation et de repérer les défauts qu'elles peuvent présenter : imperfections locales dues à un accident de fabrication, rayures, poussières, zones ayant mal hybridé à cause de la présence d'une bulle dans la solution...

Afin de montrer les possibilités de ces systèmes, Hohn a repris l'ensemble des données d'un article publié l'an dernier sur l'expression génique chez la drosophile [3] et pour lequel les données primaires (donc les images) étaient accessibles. L'analyse ainsi effectuée sur 14 hybridations montre que trois sont franchement mauvaises, sept présentent des défauts qui peuvent néanmoins être corrigés (grâce aux logiciels de *Gene Data*), le reste étant de qualité correcte. Les défauts graves sont pour l'essentiel inhérents aux puces, donc à leur fabrication ou à leur manipulation avant emploi (mais les puces *Affymetrix* étant scellées dans une cartouche, les fausses manœuvres possibles pour l'utilisateur sont assez limitées)... En somme, il ne fait pas faire une confiance aveugle, même aux fabricants qui ont le plus d'expérience. Apparemment, il ne faut pas non plus s'obstiner à essayer d'utiliser toutes les expériences, même celles qui sont un peu douteuses : l'étude montre qu'en éliminant les images défectueuses, le rendement (le nombre de gènes dont l'expression s'avère différentielle de manière fiable) augmente très nettement.

Dans ce registre de la qualité, un mot des efforts de normalisation et d'archivage des données d'expression menés notamment à l'*European Bioinformatics Institute* (EBI) à Hinxton (Royaume-Uni). L'objectif (poursuivi actuellement par plusieurs institutions, et même par au moins une entreprise) est d'emmagasiner les résultats dans des bases de données accessibles à tous (ou moyennant finances pour les projets privés), un peu

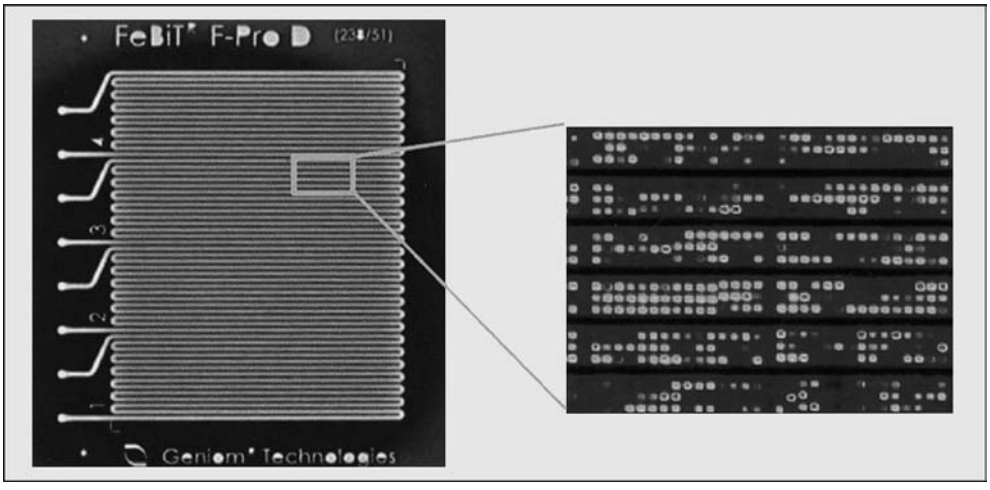

Figure 1. Le Geniom 1. Cette figure, reproduite avec l'aimable autorisation de FeBit, montre l'élément central de la machine Geniom 1, le *DNA processor*. C'est une plaque transparente (à gauche) comportant un ensemble de quatre canaux indépendants ; la synthèse d'oligonucléotides se fait directement sur la paroi plate du canal par voie photochimique, l'illumination étant assurée par un système comportant 40 000 micromiroirs commandés individuellement, et les réactifs étant apportés *via* les canaux. L'ensemble dure environ six heures pour la synthèse de 25-mères. L'hybridation est ensuite effectuée dans les mêmes canaux, la solution *ad hoc* circulant à l'intérieur de ceux-ci ; la lecture du résultat d'hybridation se fait grâce à une caméra CCD, toujours dans le *DNA processor* (image de droite). Dans le modèle actuel le nombre total d'éléments dans les quatre canaux est de 32 000 (taille individuelle 32 μm) ; la longueur des oligonucléotides peut atteindre 60 bases.

comme cela a été fait à partir de la fin des années soixante-dix pour les séquences d'ADN. Naturellement, c'est beaucoup plus compliqué : pour être significatives, les données d'expression doivent préciser le type de plate-forme, les conditions de mesure, la nature exacte de la sonde et les détails de la méthode de marquage... afin qu'un jour peut-être il soit possible de consulter de telles bases pour en tirer, à partir de résultats archivés par des dizaines de laboratoires différents, le profil d'expression d'un gène dans un très grand nombre de tissus et de situations. La mise en place d'une base de données spécifique, appelée MGED *(Microarray Gene Expression Database)* est donc une tâche ardue, à laquelle s'est attelée une équipe d'une quinzaine de personnes assistée par plusieurs groupes de travail [4]. Chemin faisant, il s'est avéré nécessaire de définir les informations qui devaient obligatoirement être fournies pour permettre l'archivage d'une expérience. Cela a été formalisé en une sorte de protocole appelé MIAME *(Minimum Information About a Microarray Experiment)* définissant les renseignements à fournir, qui devraient idéalement devenir obligatoires pour toute publication [5]. Mon impression devant ces efforts est mitigée, et, pour tout dire, je crains que ce système n'entre jamais réellement en vigueur. L'exemple concernant MIAME, donné à Zurich par Helen Parkinson (EBI), comportait 11 pages de texte pour une série de cinq hybridations... Peut-être aurait-il été plus judicieux de mettre tout de suite en place un système d'archivage, même très imparfait, afin d'habituer les laboratoires à cette procédure, et de l'améliorer progressivement ensuite ? La mise au point d'une version *light* de MIAME, appelée *MIAME express*, montre que l'on est conscient du problème à l'EBI. En tous cas, l'idéal d'une base de données d'expression archivant les résultats de nombreux laboratoires et réellement utilisable de manière transversale semble – malheureusement – encore bien lointain...

Puces à protéines : l'expectative

Le secteur des *protein arrays* occupait une place importante dans le programme de ce colloque : plus d'un tiers des présentations y étaient consacrées. Ayant assisté dans un passé récent à des conférences annonçant la construction de réseaux comportant des milliers de protéines, potentiellement utilisables pour étudier les interactions de systèmes complexes, j'étais fort intéressé par les nouveaux développements dans ce secteur. Je devais rester sur ma faim... Il existe bien des puces susceptibles de comporter dix mille plots, réalisées par microfabrication et comportant des surfaces spécialement mises au point pour fixer les protéines dans de bonnes conditions. David Wilson (*Zyomix*, États-Unis) en montrait des exemples – pour préciser ensuite que, ne disposant pas de dix mille protéines à placer sur ces réseaux, l'équipe s'était limitée à des réalisations beaucoup plus modestes (six unités identiques de 25 plots sur une puce). En fait, comme ceci est fort bien exposé dans un récent article de *Nature* [6], la fabrication de telles puces se heurte à au moins trois difficultés : disposer de nombreuses protéines pures à déposer sur le réseau, maîtriser leur interaction avec les entités qu'elles sont censées fixer, et avoir une méthode de détection sensible et quantitative. Ces problèmes sont résolus pour les réseaux d'ADN : l'amplification par PCR (ou la synthèse d'oligonucléotides) fournissent facilement l'ADN, l'interaction est assurée par l'hybridation, réaction bien connue et aux paramètres bien maîtrisés, et le marquage en masse de l'échantillon par différentes méthodes donne de bons résultats. Ils restent par contre très présents pour les réseaux de protéines...

L'essentiel des réalisations présentées à Zurich concernait donc des puces de faible complexité (de trois à vingt cinq plots), sur lesquels sont disposés des anticorps bien caractérisés destinés à capter une cible connue à l'avance et dont la présence est détectée ensuite par un deuxième anticorps apportant le système de détection – l'habituel *sandwich assay* des dosages immunologiques. C'est, en somme, une transposition du système ELISA sur un support de type réseau, ce qui permet de multiplier le nombre de paramètres mesurés et de diminuer le coût de chaque dosage. Parmi les systèmes décrits, un réseau d'anticorps monoclonaux pour détecter des pesticides dans l'eau (Claus Christensen, Université de Copenhague, Danemark), une autre puce pour suivre les protéines caractéristiques de l'arthrite rhumatoïde (Sara Mangialaio, *Novartis Pharma*, Bâle, Suisse), ou encore (en renversant le système) un réseau constitué d'antigènes viraux afin de détecter les anticorps correspondants dans le sérum humain (sérodiagnostic) (Tito Bacarese-Hamilton, *Imperial College*, Royaume-Uni). Rien de grandiose donc (il manquait quelques ténors à ce colloque), mais des applications qui peuvent trouver leur place si le conservatisme des laboratoires d'analyse et la puissance des concurrents dont la technologie ELISA est bien installée ne les réduisent pas au rôle de figurantes. Ce serait dommage, car ces approches préparent la voie à des puces à protéines plus complexes dont le rôle, malgré les difficultés rencontrées, sera sûrement important dans l'avenir.

Quelques nouveautés

Un colloque *Cambrige Healthtech* serait décevant s'il ne comportait pas quelques innovations technologiques propres à faire rêver ceux que ce domaine inspire. Pour ma part, trois items (dont deux au moins ne sont pas tout à fait nouveaux) répondaient à ce critère. Pour commencer, le système de PamGene [7], firme hollandaise représentée par son vice-président Alan Chan. Il s'agit d'un réseau de faible complexité destiné aux applications cliniques, et réalisé sur un support très particulier, une sorte d'éponge d'alumine comportant d'innombrables canaux verticaux. L'ADN est fixé dans toute l'épaisseur de ce support, et la goutte de solution d'hybridation déposée sur la puce la traverse de manière répétée sous l'influence d'une pression de gaz variable (la tension superficielle évite son arrachement du support). L'hybridation (avec une sonde fluorescente) est suivie

en direct au microscope (le support mouillé est transparent) et est très rapide : une dizaine de minutes pour une identification bactérienne par hybridation d'un produit de PCR avec des oligonucléotides fixés sur le réseau. C'est me semble-t-il un bon exemple de puce à ADN adaptée aux applications cliniques pour lesquelles le nombre de plots nécessaire est faible, mais où les résultats doivent être acquis rapidement – et, de préférence, par des méthodes qui ne dépaysent pas trop le laboratoire d'analyse. Son application à des mesures d'expression est en cours de mise au point, mais va sans doute se heurter à des problèmes de sensibilité. Il existe au moins une autre entreprise développant une technologie analogue, *MetriGenix* (États-Unis), mais elle n'était pas présente à Zurich.

Autre nouveauté dans le domaine de l'analyse : le système des *Nanobarcodes* présenté par Michael Natan de *SurroMed* (États-Unis) [8]. Il s'agit de minuscules barreaux métalliques longs de quelques micromètres et constitués de deux ou trois métaux (argent, or, cuivre). Ils sont fabriqués par synthèse électrochimique sur une sorte de moule (microscopique), ce qui permet d'alterner les couches déposées et d'obtenir des barreaux comportant jusqu'à une dizaine de bandes de différents métaux et d'épaisseur variable. On peut alors fixer sur chaque type de barreau un réactif (protéine ou ADN), puis les mélanger pour analyser un échantillon complexe, la lecture du résultat se faisant au microscope (automatisé) ou par un système de détection en flux. Cette technologie, qui montre l'alliance de la microfabrication et de la biologie, peut avoir des applications dans des domaines très variés.

Terminons cette revue des nouveautés (il y en avait d'autres) par un appareil plus conventionnel mais néanmoins assez étonnant, le « Geniom 1 » de l'entreprise allemande *FeBit* (Mannheim, Allemagne) [9] *(Figure 1)*. J'y avais fait une rapide allusion dans une précédente chronique [10], mais ai pu constater cette fois la réalité de cette machine et les progrès accomplis depuis une année. Il s'agit d'un appareil d'un seul tenant, qui est capable de synthétiser des puces à oligonucléotides, puis d'effectuer l'hybridation et la lecture du résultat – tout le matériel nécessaire tenant à l'intérieur d'un cube dont le côté mesure un peu moins d'un mètre ! L'élément central est une sorte de cartouche transparente à l'intérieur de laquelle la synthèse est réalisée par voie photochimique (en utilisant des réactifs différents de ceux d'Affymetrix, et un multimiroir programmé de *Texas Instruments*). La même cartouche sert à l'hybridation, et les résultats sont lus ensuite par un système optique à base de caméra CCD. Cela semble trop beau pour être vrai... Pourtant les machines existent réellement, une série de dix prototypes a été réalisée (l'un d'eux est en service au centre du cancer de Heidelberg) et la véritable mise sur le marché de l'appareil est prévue pour la fin de l'année. L'énorme avantage de ce système est bien sûr sa flexibilité, permettant de modifier le réseau au jour le jour en fonction des résultats des expériences. Les données montrées, encore préliminaires, étaient assez convaincantes, et on ne peut que souhaiter à cette petite entreprise (60 salariés, fondée en 1998) de poursuivre avec succès dans cette voie très innovante – qui risque néanmoins de se heurter à la concurrence de puces prêtes à l'emploi et produites en grande série à des prix plus doux que les tarifs actuels...

Pour conclure...

Bien d'autres aspects des *microarrays* étaient présentés lors de ce colloque, notamment leur emploi pour analyser l'expression génique dans différents types de cancer. Ce genre d'étude, activement poursuivi par de nombreuses équipes et entreprises, donne des résultats prometteurs en termes de classification des cancers. Il fournit des informations dont on peut penser qu'elles permettront d'ajuster le traitement administré à la nature précise de chaque cancer et donc d'améliorer à la fois le pronostic et le confort du malade. Il y a donc là une importante possibilité d'application clinique dans un avenir proche, et ce thème mériterait à lui seul une chronique ou même un article de fond. Je

n'ai pas non plus traité autrement qu'en passant la très importante question de la bioinformatique associée aux mesures d'expression, que ce soit pour acquérir et vérifier les mesures, pour les analyser par différents modes de regroupement (pour ne pas dire *clustering*), et pour essayer d'en tirer du sens du point de vue biologique. Elle faisait d'ailleurs l'objet d'un deuxième colloque *Cambridge Healthtech* et mériterait, elle aussi, une chronique entière...

Ce qui apparaît en tous cas de manière évidente, c'est que la logique des *arrays* fait tache d'huile dans la biotechnologie actuelle. Logique dont les éléments essentiels sont l'emploi de réseaux (d'ADN, de protéines, ou même de cellules comme dans le cas des *phenotype arrays* présentés par Barry Bochner, *Biolog*, États-Unis) autorisant un grand nombre de mesures simultanées, et la miniaturisation qui réduit les besoins en échantillons et réactifs, diminuant ainsi les coûts. Le mode de réalisation de ces réseaux semblait, lors de ce colloque, avoir atteint une certaine stabilité. C'est sans doute trompeur, car de nombreuses firmes parmi les plus innovantes n'étaient pas représentées et l'on peut s'attendre à de nouvelles surprises durant les mois qui viennent ; mais le principe même des réseaux est, lui, bien installé et va certainement continuer à étendre son règne sur la biologie à grande échelle comme sur de nombreuses applications cliniques.

Post-scriptum

Quelques nouvelles sur le front des brevets et de la concurrence, pour donner une suite à ma récente diatribe sur le sujet [10]. *Affymetrix* a fait la paix avec *Oxford Gene Technologies*, l'entreprise liée à Edwin Southern que je considère comme le vrai inventeur des puces à oligonucléotides, il lui en a coûté près de vingt millions de dollars. La paix a aussi été conclue avec l'entreprise *Incyte* (les deux compagnies s'étaient réciproquement attaquées depuis plusieurs années), qui s'est retirée en octobre 2001 du marché des *microarrays* tout comme *Corning*, autre concurrent potentiellement dangereux. Motorola, Agilent et un certain nombre de *start-ups* (certaines mentionnées dans cette chronique) restent dans le domaine, mais *Affymetrix* (qui emploie maintenant plus de 900 personnes) est clairement la force dominante. Pour ceux qui souhaiteraient suivre de près ce paysage changeant, une revue spécialisée qui semble bien informée mais est malheureusement fort chère, *BioArray News* [11], et un site tendant à rassembler l'ensemble des informations, qui était payant mais va devenir gratuit, BioChipNet [12].

Références

1. Pirsig RM. *Zen and the art of motorcycle maintenance*. Londres : Bodley Head, 1974. Édition française la plus récente : Collection *Points*. Paris : Seuil, 1998.
2. Halgren RG, Fielden MR, Fong CJ, Zacharewski TR. Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res* 2001 ; 29 : 582-8.
3. De Gregorio E, Spellman PT, Rubin GM, Lemaitre B. Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. *Proc Natl Acad Sci USA* 2001 ; 98 : 12590-5.
4. www.ebi.ac.uk/microarray/.
5. Brazma A, *et al.* (30 auteurs). Minimum information about a microarray experiment (MIAME) : toward standards for microarray data. *Nat Genet* 2001 ; 29 : 365-71.
6. Alison Abbott Betting on tomorrow's chips. *Nature* 2002 ; 415 : 112-4.
7. www.pamgene.com.
8. www.surromed.com.
9. www.febit.com.
10. Jordan BR. Chroniques génomiques. Puces à ADN, les brevets contre le progrès. *Med Sci* 2001 ; 17 : 893-6.
11. www.BioArrayNews.com.
12. www.bmi-heidelberg.com/BioChipNet/.

Puces « recherche », puces « cliniques » et « labopuces »

La saga des puces continue de plus belle. Cette façon « massivement parallèle » d'effectuer des mesures biologiques répond en effet à un impératif actuel : l'approche globale des phénomènes biologiques. Elle permet aussi, grâce à sa miniaturisation, de très importantes économies de réactifs : c'est vital pour l'industrie pharmaceutique qui peut être amenée, au cours de la mise au point d'un médicament, à effectuer des centaines de milliers ou même des millions d'essais. Selon que chacun de ces tests consomme quelques microlitres, ou au contraire un millilitre de réactif, le coût sera sans commune mesure, et un programme de « criblage à haut débit » qui serait impraticable s'il était effectué dans de classiques plaques à 96 puits peut devenir abordable lorsqu'il est pratiqué à l'aide de puces ad hoc.

Au niveau de la technologie, et pour les puces complexes (capables de mesurer l'expression de dizaines de milliers de gènes), l'entreprise Affymetrix reste en position dominante, détenant plus de la moitié du marché (ce qui ne l'empêche pas de continuer, aujourd'hui encore, à perdre de l'argent). Aucun de ses concurrents ne paraît en mesure de lui disputer ce leadership *dans l'immédiat. Cette firme semble en revanche assez mal placée pour aborder un secteur en pleine expansion, celui des puces à usage clinique, destinées à affiner un diagnostic ou à prédire la réponse à un médicament. Les besoins sont assez différents de ceux de la recherche. Il s'agit cette fois de mesurer, si possible très rapidement et avec un appareillage d'utilisation aisée, le niveau d'expression de quelques dizaines ou centaines de gènes dans un prélèvement, une biopsie ou un fragment de tumeur. L'information souhaitée (« y a-t-il lieu de procéder à une chimiothérapie post-opératoire ? ») doit alors découler d'une interprétation quasi automatique des données obtenues. Les puces très complexes et les machines sophistiquées d'Affymetrix sont surdimensionnées pour ce type d'application, alors que la technologie de PamGene, par exemple, s'avère bien adaptée. D'autres sociétés ont mis au point des systèmes présentant des performances similaires. Ce secteur clinique intéresse beaucoup d'entreprises : si ces tests se généralisaient, le marché correspondant serait de très grande ampleur. Par ailleurs, les firmes misant sur des approches nouvelles poursuivent leur route avec des succès variables. Corning tout comme Motorola ont abandonné leurs projets « puces » pourtant bien avancés, jugeant sans doute que la rentabilité était trop aléatoire ; au contraire, FeBit a réussi un nouveau « tour de table » qui lui a apporté une trentaine de millions d'euros, et commercialise son appareil dès juillet 2003.*

Un mot pour terminer sur les « labopuces » (Lab-on-a-chip) *dont, me semble-t-il, on parle moins en ce moment. Certes, l'idée d'intégrer l'ensemble des fonctionnalités nécessaires à l'exécution d'un test biologique sur une puce de verre, de plastique ou de silicium est extrêmement séduisante. Cette labopuce pourrait, par exemple, concentrer l'échantillon biologique, lyser les cellules, extraire l'acide nucléique puis effectuer une hybridation et lire le résultat. Mais sa réalisation pratique, à l'échelle industrielle et à un coût acceptable, demande de multiples travaux et des mises au point qui ne sont visiblement pas terminés. La labopuce reste une solution d'avenir, mais il faudra attendre encore un peu avant de la voir envahir les laboratoires...*

This page intentionally left blank

8. UNE SÉQUENCE TANT ATTENDUE...

Nous en venons donc à la séquence proprement dite. Elle est souvent évoquée dans ces chroniques mais, finalement, peu d'entre elles lui sont explicitement dédiées. Cela résulte d'un choix conscient de ma part : le séquençage était en général largement traité dans la presse scientifique ou de grande vulgarisation. J'ai donc eu tendance à me focaliser sur les autres aspects du programme Génome qui me semblaient tout aussi importants. Voici néanmoins trois chroniques sur la séquence. Les deux premières datent du début, le temps des illusions, des désillusions puis du véritable démarrage ; la dernière relate le vif débat sur le nombre de gènes humains déclenché par la publication de la première « séquence brouillon », courant 2000...

Med Sci *1991 ; 7 : 612-613*

LES HEURS ET MALHEURS DU SÉQUENCAGE D'ADN À GRANDE ÉCHELLE

J'entame de nouveau un chapitre par une chronique un peu antérieure à la période (1992-2002) couverte par ce recueil. Elle est pourtant à sa place ici : elle reflète bien les difficiles commencements du séquençage à grande échelle que l'on croyait pourtant, en 1990, à portée de main.

Un grand optimisme

À la fin des années 1980, l'on envisageait avec beaucoup d'optimisme le séquençage de grandes régions d'ADN : les « séquenceurs » commercialisés par Applied Biosystems, puis annoncés par Du Pont et Pharmacia/LKB devaient permettre un grand bond en avant de la productivité tout en évitant l'utilisation d'isotopes radioactifs. Les systèmes robotiques allaient rapidement automatiser la plupart des procédures annexes, et des moyens informatiques sophistiqués devaient faciliter l'assemblage, la vérification des séquences et, dans un deuxième temps, leur interprétation. On envisageait donc sans trop d'inquiétude des programmes visant à établir des séquences sur une ou même plusieurs mégabases (1 mégabase : 1 000 kilobases, un million de nucléotides) : souvenons-nous des projets de séquençage du complexe majeur d'histocompatibilité (deux mégabases), de la bande q28 du chromosome X (dix mégabases) ou de l'ensemble des locus codant pour les gènes du récepteur T chez la souris et chez l'homme.

La stratégie adoptée

Rappelons la stratégie qui avait été adoptée, avec quelques variantes, par tous les groupes concernés. L'unité de séquençage était l'ADN cloné dans un phage (15 à 20 kb) ou dans un cosmide (40 kb). Le fractionnement de ce grand segment d'ADN en sous clones était fait selon une tactique de *shotgun*, avec fragmentation aléatoire de l'ADN par sonication suivie du sous-clonage des fragments (il existe des méthodes plus systématiques, nous en reparlerons, mais elles ont toutes paru trop lourdes pour être appliquées à grande échelle). Suivaient alors la préparation des matrices simple brin, les réactions de séquence avec des amorces ou des précurseurs fluorescents, et l'analyse des produits de réaction par un « séquenceur » automatique, en général de marque Applied. La capacité de ces machines, de l'ordre de cinq à dix mille nucléotides de séquence brute par jour (24 pistes, 400 nucléotides lus par piste), devait permettre une « production » supérieure au million de nucléotides « bruts » par an et par machine et donc, avec plusieurs machines, plusieurs centaines de kilobases de séquence confirmés par an (notons qu'il faut cinq à dix kb de séquence brute pour une de séquence confirmée). Les résultats donnés par ces machines devaient enfin être repris par un système informatique effectuant l'assemblage de toutes ces petites séquences.

On déchante...

Aujourd'hui, début 1991, on est loin du compte ; la plus longue séquence jamais établie reste celle du virus EBV, faite à la main en Angleterre, et les entreprises de mégaséquençage ne rencontrent plus le même enthousiasme. Bien peu de projets ont produit plus d'une centaine de kilobases de séquence confirmée d'un seul tenant, et les « séquenceurs fous » sont moins triomphalistes que par le passé. Ils se sont heurtés, en fait, à toute une série de problèmes techniques et organisationnels liés au saut quantitatif qu'ils tentaient d'effectuer et aux difficultés entraînées par une automatisation partielle. Sur le plan technique d'abord les machines à séquencer, qui n'automatisent en fait qu'une des étapes de la détermination de séquence (la séparation des fragments d'ADN sur gel et leur détection) se sont généralement avérées très exigeantes sur la quantité et la qualité des mélanges réactionnels qu'on leur donne à analyser. Là où la méthode manuelle s'accommode de petites quantités de préparations d'ADN peu purifiées, la machine exige une charge plus importante d'un échantillon quasiment parfait, et si l'on voit autant de « séquenceurs » inutilisés dans les laboratoires, c'est en grande partie à cause de ce problème. En dehors des questions de mise au point, il y a à cela une raison fondamentale. Dans la méthode classique la détection finale est effectuée par un détecteur de grande taille (le film à rayons X) auquel on laisse tout son temps, des dizaines d'heures en général, pour enregistrer les données globalement sur toute la surface du gel. La machine, au contraire doit détecter les fragments « au vol », en cours de migration et pendant un temps très court, ce qui impose une sensibilité extrême et laisse peu de marge pour compenser d'éventuelles imperfections de l'échantillon. La contrainte sur la qualitè de l'ADN, commune à des degrés divers aux différentes variantes de ces machines, est particulièrement mal venue ici : on aurait aimé automatiser différentes étapes préalables comme la préparation d'ADN à partir des clones. Or, les méthodes de préparation qui donnent de l'ADN très pur font presque toujours intervenir une étape de centrifugation qui est particulièrement difficile à automatiser : les techniques utilisant des filtrations sont plus facilement automatisables mais donnent en général un ADN de moins bonne qualité. On touche là à un problème général, celui des goulets d'étranglement : une procédure complexe ne peut pas aller plus vite que la plus lente de ses étapes ; et il ne sert pas à grand-chose d'automatiser une partie du processus si l'autre reste manuelle et lente. L'expérience de ceux qui ont mis en place des laboratoires de mégaséquençage montre que sitôt un goulet d'étranglement éliminé, un autre apparaît... de sorte que la mise en œuvre du projet ressemble à une course d'obstacles.

La technique n'est pas seule en cause

Une autre difficulté, de nature différente celle-là, tient à l'interférence entre développement et production. L'objectif est l'obtention de données de séquences : il faut donc simultanément organiser leur établissement selon un certain processus et chercher en permanence à améliorer ce même processus. Il y a un conflit potentiel entre ces deux fonctions, surtout quand ce sont les mêmes personnes qui l'assument, et sur les mêmes appareils ; il suffit de quelques essais par semaine pour bloquer la moitié des machines et réduire de beaucoup la production. Inversement si l'on ne fait pas de développement on risque d'investir beaucoup de temps et d'argent de façon inefficace avec des techniques en train de se périmer. L'équilibre n'est pas facile à trouver ! Un autre problème organisationnel est celui du personnel : le travail effectué dans un tel contexte n'est pas particulièrement passionnant (la situation est très différente de celle du chercheur qui séquence « son » gène) : à quel type de personne faut-il le confier ? Deux écoles à ce sujet : les uns pensent qu'il faut plutôt choisir des personnes d'un niveau d'études moyen que l'on formera « sur le tas » et qui accepteront de faire durablement ce travail répétitif. D'autres au contraire préfèrent viser haut en acceptant le fait que la personne engagée

ne fera ce travail qu'un ou deux ans et quittera ensuite le laboratoire pour passer à un emploi plus motivant. Ils considèrent que les inconvénients de cette instabilité seront compensés par la rapidité d'apprentissage et les capacités plus larges de ces employés. La deuxième option m'a paru majoritaire dans les laboratoires aux États-Unis, ce qui est sans doute lié (entre autres) à une tradition de mobilité d'emploi qui n'existe pas en France.

On repart sur des bases plus raisonnables

Les laboratoires de mégaséquençage ont maintenant tiré la leçon des difficultés rencontrées et affichent des buts plus modestes ; ils ont aussi pris conscience de la nécessité d'une organisation de type industriel. Cette organisation impose de séparer production et développement, avec d'un côté une unité de production, et de l'autre des chercheurs ou ingénieurs qui perfectionnent les méthodes sans interférer au jour le jour avec cette unité. Une plus grande attention est portée à l'entretien et au dépannage des machines avec les moyens locaux, sans avoir à faire appel au constructeur, et surtout on se livre à une étude détaillée de toutes les étapes du processus pour évaluer de façon réaliste la production à chaque étape, rationaliser l'ensemble et faire porter l'effort maximum sur les étapes les plus limitantes. Toutes ces considérations ne seraient pas déplacées dans une usine d'automobiles ; et ceci marque bien que le passage à cette échelle entraîne un changement qualitatif, qu'il ne s'agit pas simplement de faire cent fois plus, mais aussi de le faire autrement...

Et les « petits » laboratoires

À côté de ce secteur du mégaséquençage, il existe de très nombreuses équipes qui ont de temps en temps besoin de déterminer la séquence de quelques kilobases ou dizaines de kilobases d'ADN. Ce besoin dépasse d'ailleurs largement la communauté de la Génétique humaine et concerne peu ou prou tous les biologistes moléculaires. Pour le moment ce secteur est resté fidèle aux méthodes traditionnelles de séquençage, et les machines actuelles sont à la fois trop chères, trop productives et trop « perturbantes » (par les modifications en amont et en aval qu'elles entraînent) pour s'y imposer. Il y a pourtant un potentiel d'innovations important dans ce secteur, et des possibilités commerciales auxquels les constructeurs sont sensibles : car il y aura peut-être quelques dizaines de centres de mégaséquençage, mais il existe des milliers de laboratoires faisant un peu de séquence et intéressés à ce que cela devienne plus facile, plus rapide et moins cher... On peut citer, en vrac, parmi les innovations qui s'imposeront peut-être dans ce domaine, la détection directe de la radioactivité, ou au moins le remplacement des films à rayons X par des *imaging plates* que l'on expose (comme un film) pendant quelques heures au maximum pour les « développer » ensuite en quelques minutes dans un appareil qui utilise un faisceau laser pour leur lecture. Dans les deux cas on gagne en vitesse, et surtout l'image est directement enregistrée dans une mémoire d'ordinateur. Autre option, le remplacement des isotopes radioactifs par les systèmes de marquage par photoluminescence qui pourrait enfin arriver à s'imposer dans les équipes de recherche.

L'électrophorèse capillaire qui a déjà fait un « tabac » dans les laboratoires de biochimie peut être,elle, la base d'appareils de séquence plus modestes en production et en prix que les machines Applied ou Du Pont... On voit que la gamme des innovations possibles reste large, ce qui me permet de terminer cette chronique sur une note optimiste !

Remerciements

Je remercie Leroy Hood, Ben Koop et Tim Hunkapiller (*Caltech*, Pasadena, CA), Richard Gibbs (*Institute of Medical Genetics*, Houston, TX) et Paul Silverman (*Beckman Instruments*, Fullerton, CA) pour les discussions qui ont largement nourri cet article.

Med Sci *1992 ; 8 : 854-857*

SÉQUENÇAGE GÉNOMIQUE : LE DEUXIÈME SOUFFLE

Un an plus tard, l'on commence à reprendre courage : le succès du séquençage d'un chromosome de levure montre à la fois la faisabilité et l'intérêt de la lecture de quelques mégabases d'ADN, même si l'on est encore bien loin de se sentir prêt à aborder les trois millards de bases de notre génome. Au-delà de la Levure, le séquençage de l'ADN du nématode révèle déjà les qualités d'organisation et le réalisme du tandem Sulston/Waterston – sans doute le seul, à cette époque, à avoir produit les kilobases puis les mégabases promises dans le délai annoncé et au coût prévu. On notera au passage que – me conformant en cela à l'opinion majoritaire – j'imaginais que le génome humain pourrait bien contenir plusieurs centaines de milliers de gènes...

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 ultra-sensible 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 séquençer, 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 de 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 Martinsried, 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 (vérifiée) avaient été déterminé, dont 25 seulement avec des machines, tout le reste à la main ; la comparaison des 35 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], révèla 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 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 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 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 nombreus 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 impressioniste 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 des 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 kilobases 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...

Références

1. Jordan B. 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 B. 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. *Genomics* 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 chromsome 19q13.3. *Nat Genet* 1992 ; 1 : 34-9.
9. Jordan B. 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 J-P, 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. *Nat Genet* 1992 ; 1 : 348-53.

UN COMPTE D'APOTHICAIRE

La chronique reproduite ci-après a connu un destin assez particulier : elle n'a jamais été publiée ! À la suite d'erreurs de transmission et d'un concours de circonstances, elle s'avérait faire double emploi avec un texte sur le même sujet rédigé par l'équipe du Génoscope pour le même numéro de novembre 2000 de médecine/sciences *(Roest Crollius H, Jaillon O. Le nombre de gènes dans le génome humain : les paris sont ouverts.* Med Sci *2000 ; 16 : 988-90). Alors qu'elle était déjà à l'état d'épreuves, nous avons donc décidé de la retirer (une version largement modifiée devait paraître dans la revue* Biofutur *en décembre 2000). On trouvera ci-dessous le texte qui aurait dû figurer dans* médecine/sciences*, et qui fait le point sur une des grandes surprises de l'année 2000, année de la « séquence brouillon de l'ADN humain » : le faible nombre de gènes trouvés par rapport aux 100 000 (ou plus) attendus.*

Un débat paradoxal

La controverse qui est apparue au printemps dernier à propos du nombre des gènes humains [1] doit laisser rêveurs nombre de nos lecteurs. Comment ? Après dix ans de programmes Génome, treize Téléthons, des cartes génétiques et physiques annoncées à grand fracas, alors qu'officiellement notre génome a été séquencé à plus de 90 %... non seulement on ne connaît pas le nombre exact de gènes qui y sont inscrits, mais de surcroît les estimations s'étalent de moins de 30 000 à plus de 120 000 ! L'étonnement provient aussi de la récente baisse des évaluations. Alors que l'opinion générale semblait tendre vers une valeur proche de 100 000, deux équipes au moins annoncent maintenant des nombres beaucoup plus faibles. Pourtant le séquençage a jusqu'ici plutôt mis en évidence des gènes supplémentaires, comme par exemple pour la levure. En réalité la controverse couvait depuis plusieurs années ; elle prend aujourd'hui un tour aigu en raison de l'obtention d'une séquence quasi complète, du fait aussi de l'importance des intérêts commerciaux en jeu.

Une détermination délicate

Si ce débat a lieu, c'est bien sûr parce que la définition des gènes à partir de la séquence est plus complexe pour le génome humain que pour la levure ou le nématode. Chez *Saccharomyces cerevisiae*, l'ADN est codant à plus de 50 %, la plupart des gènes sont d'un seul tenant et la prédiction des parties codantes à partir de la séquence s'avère rapide et fiable. La quasi-totalité des ORF (*open reading frames*, cadres de lecture ouverts) mis en évidence par le seul séquençage ont été confirmés ensuite de manière expérimentale. Pour l'ADN humain, la dispersion systématique des séquences codantes (qui ne représentent en tout que quelques pour cent de l'ADN) en une multiplicité d'exons parfois très courts et séparés par de longs introns rend leur détection aléatoire. Ce problème est mis en évidence par l'analyse de régions séquencées par ailleurs très bien connues, notamment celle du premier chromosome humain à être entièrement déchiffré, le chromosome 22 [2]. De fait, un tiers environ des exons prédits par les meilleurs programmes

d'analyse n'ont pas d'existence réelle (faux positifs) et près d'un tiers des « vrais » exons ne sont pas prédits (faux négatifs)[1]. Pourtant il s'agit là d'études portant sur des données de séquence « finie », de très bonne qualité (taux d'erreur inférieur à 0,02 %) [2]. Même dans ces conditions, un gène qui comporte plusieurs dizaines d'exons sera presque toujours incorrectement prédit. Et comment décidera-t-on si la trentaine de région codantes détectées dans une région donnée provient d'un seul gène, ou de plusieurs – d'autant que l'on sait très mal reconnaître les promoteurs ? L'utilisation de séquence « brouillon » ne fait que compliquer l'analyse. Les « signaux » (séquences caractéristiques de début et de fin d'exon par exemple) déjà très flous deviennent fort difficiles à repérer lorsque le taux d'erreur atteint 0,1 ou même 1 %.

Il faut donc obligatoirement (du moins dans l'état actuel de la science) utiliser d'autres informations pour repérer et compter les gènes. La méthode la plus employée depuis plusieurs années repose sur l'emploi des séquences partielles d'ADNc accumulées par millions depuis le début des années 1990, les EST *(expressed sequence tags)*. La comparaison d'un EST avec une séquence génomique conduit en effet au repérage de zones homologues qui normalement correspondent à des gènes. La méthode s'accommode bien de données de type « brouillon », une similitude à 97 ou 98 % étant en général aussi révélatrice qu'à 100 %. La détection des gènes dans les séquences d'ADN humain fait donc un large appel à ces données, qui ont été déterminantes pour le succès de nombreux travaux récents en génétique médicale. Leur emploi pour évaluer globalement le nombre de nos gènes pose néanmoins quelques problèmes.

Les *clusters* d'*Unigene*

Nous disposons aujourd'hui dans les bases de données publiques (dbEST) de plus de deux millions de séquences partielles d'ADNc humains. Plusieurs industriels, notamment *Human Genome Sciences* et *Incyte*, affirment pour leur part en détenir un nombre encore plus élevé. Bien entendu, notre génome ne comporte pas des millions de gènes : la collection présente une forte redondance, due au fait que ces séquences ont été obtenues à partir de clones pris au hasard dans de nombreuses banques d'ADNc établies à partir de divers tissus. Mais il est possible de comparer toutes ces séquences entre elles (et d'utiliser également les autres informations contenues dans les bases de données) pour regrouper celles qui *a priori* proviennent du même gène. C'est le but de la construction de *gene indexes*, qui résultent de l'analyse des séquences par un ensemble de logiciels de regroupement. Le plus connu est *Unigene* [3], le système « officiel », dont les résultats constamment tenus à jour sont disponibles sur le site du *National Center for Biotechnology Information* (http://www.ncbi.nlm.nih.gov/). La version 119 (août 2000) comporte 89 985 *clusters*, regroupements d'EST présentant une homologie de séquence sur une partie de leur longueur ou « rattachés » les uns aux autres par comparaison avec une autre séquence. On pourrait donc considérer que l'ensemble de ces données indique l'existence d'au moins 90 000 gènes dans notre génome. Reste à évaluer ce qui manque dans dbEST : bien que ces deux millions de séquences aient été déterminées à partir de clones issus de très nombreuses banques d'ADNc réalisées à partir de tous les tissus imaginables (ou du moins accessibles), il peut toujours exister des gènes très peu exprimés, ou exprimés uniquement dans un tissu bien particulier à un instant précis du développement embryonnaire, et qui échapperaient ainsi à ce comptage. Une manière simple d'estimer cette correction est de considérer l'ensemble des gènes qui ont été identifiés dans le cadre de la recherche sur les maladies génétiques, et de voir quelle fraction de ces derniers est représentée dans dbEST. Cette évaluation donne un chiffre de 90 % :

1. Je résume ici une assez longue discussion qui utilise non seulement les données sur le chromosome 22 mais aussi celles de plusieurs régions très bien connues utilisées pour étalonner les logiciels de prédiction. Voir pour plus de détails l'article de Dunham *et al.* [2].

elle est critiquable à différents égards, mais on peut sans doute retenir son ordre de grandeur et considérer que la grande majorité de nos gènes sont représentés dans dbEST. L'on arrive ainsi au chiffre souvent cité d'environ 100 000 gènes dans notre génome. C'est à partir d'une analyse de ce type que l'équipe de Craig Venter à TIGR avait proposé en 1995 une fourchette de 60 à 70 000 gènes [4] et que diverses estimations ont été réalisées depuis – la dernière, celle de l'équipe de John Quackenbush (toujours à TIGR) aboutissant à une valeur de 120 000 [5]. Nous reviendrons sur cette étude toute récente après avoir indiqué la nature des problèmes que pose ce genre de calcul et donné une idée du fondement des nouvelles évaluations.

Les problèmes de dbEST et d'*Unigene*

La fiabilité des *gene indexes* est en effet discutable, pour des raisons qui tiennent à la fois au contenu de dbEST (ou d'autres bases de données similaires) et aux méthodes de regroupement utilisées. Les courtes séquences qui constituent les EST (300 à 500 nucléotides en général) proviennent de l'extrémité 5' ou 3' de clones pris au hasard dans des banques réalisées à partir de différents tissus. Il s'y glisse à l'occasion des contaminants : fragments d'ADN génomique, introns non épissés, ou même traces d'ADN bactérien. Or ces banques ont généralement été égalisées, c'est-à-dire qu'elles ont subi un traitement (à base de dénaturation et de ré-association contrôlées de l'ADN avant clonage) visant à réduire la fréquence des séquences les plus abondantes afin d'augmenter la chance de trouver de « nouveaux » gènes par séquençage au hasard. Ce traitement, qui améliore en effet la richesse des banques, augmente aussi la proportion des séquences provenant d'évènements artefactuels rares comme ceux que nous venons d'évoquer. De plus, beaucoup de gènes présentent des phénomènes d'épissage alternatif produisant plusieurs transcrits à partir d'un même ensemble d'exons. Là aussi, l'égalisation des banques va ramener à une abondance équivalente le transcrit majeur et le produit très rare d'un épissage aberrant, pas nécessairement significatif du point de vue biologique. Deux EST obtenus à partir de ces transcrits peuvent parfaitement ne pas contenir de séquence commune et donc apparaître comme les produits de deux gènes différents.

Vient ensuite la méthode utilisée pour regrouper ces séquences afin de construire les *clusters*. Elle va tenter de tenir compte des problèmes que je viens de mentionner en effectuant une filtration préalable des séquences (pour écarter par exemple celles qui ont une composition en bases par trop anormale pour de l'ADNc humain), puis va effectuer l'assemblage le plus précis possible. Mais, compte tenu du taux d'erreur des séquences contenues dans dbEST (au moins 1 %, souvent plus), la qualité du regroupement réalisé reste problématique. Et la plupart des artefacts évoqués plus haut tendent à augmenter le nombre de *clusters* et donc le nombre apparent de gènes. Notons par exemple que sur les 90 000 *clusters* d'Unigene, près de 30 000 sont définis par un seul EST... Il est assez peu vraisemblable qu'un « vrai » gène ait été échantillonné une seule fois en deux millions de séquences, et probable que la majorité de ces EST (et donc des *clusters*[2] correspondants) appartiennent à la catégorie des artéfacts. Mais, bien sûr, il est impossible d'exclure qu'un gène, exprimé à très bas niveau dans l'embryon, ait pu effectivement donner naissance à un seul EST issu d'une banque construite à partir d'un fœtus précoce, et il est délicat de purement et simplement éliminer ces 30 000 *clusters* du seul fait qu'ils contiennent une seule séquence.

On comprend donc que l'analyse et le regroupement des EST présentent de multiples difficultés. Il existe d'ailleurs toute une série d'autres *gene indexes*, construits par différentes équipes en essayant de résoudre les problèmes dont nous avons résumé ci-

2. *Clusters* un peu particuliers, puisqu'ils ne contiennent qu'une seule séquence...

dessus la nature. On conçoit aussi que la déduction du nombre de gènes humains à partir de ces données soit plus complexe que l'approche naïve que nous avons évoquée plus haut. Les résultats récents de séquençage offrent-ils une voie alternative ?

L'apport du séquençage complet de chromosomes

Le séquençage complet du chromosome 22 a été publié à la fin de 1999 [2], et l'analyse détaillée de sa séquence par toutes les méthodes disponibles a amené ses auteurs à avancer un nombre total de 679 gènes (dont 134 pseudogènes). Puisque ce chromosome représente environ 1,1 % de notre ADN, une simple règle de trois indiquerait environ 61 000 gènes pour l'ensemble de notre génome. Si l'on tient compte d'autres données indiquant que ce chromosome est un peu plus riche en gènes que la moyenne, d'un facteur évalué à 1,38, le nombre descend à 45 000. Ces chiffres restent éminemment contestables, malgré la qualité de la séquence obtenue (taux d'erreur inférieur à 0,02 %). L'estimation à partir d'un échantillon représentant à peine plus de 1 % de notre génome est aléatoire, la précision des facteurs utilisés pour l'extrapolation est limitée, et le chiffre même de 679 gènes est sujet à révision après des études plus approfondies. La mise en évidence des gènes dans la séquence a d'ailleurs largement fait appel aux données de dbEST, il ne s'agit donc pas d'un comptage réellement indépendant. La séquence suivante, celle du chromosome 21 publiée en mai 2000 [6], indique 225 gènes pour une région correspondant à 1 % du génome, chiffre inférieur de moitié à celui du chromosome 22 et illustrant bien les incertitudes de ce genre de comptage. Au fur et à mesure que « sortent » les séquences complètes des différents chromosomes humains, la représentativité de l'échantillon va certes s'améliorer, mais il restera encore un doute sur le nombre de gènes annoncé pour chacun d'eux et sur les erreurs possibles, par excès ou par défaut, selon les méthodes utilisées. Pour vraiment évaluer le nombre total de gènes contenus dans notre ADN, il faut non seulement disposer d'une fraction réellement significative de la séquence du génome, mais aussi savoir l'exploiter de manière purement informatique pour y détecter les gènes. On voit que nous sommes loin du compte...

Exofish entre en scène

On se souvient du Fugu, ce poisson japonais au génome très compact (400 mégabases pour un jeu de gènes comparable au nôtre) dont Sydney Brenner avait préconisé l'étude au début des années 1990 [7, 8]. Notre Génoscope avait dès le début choisi de s'intéresser à ce modèle, prenant pour objet d'étude un cousin non toxique du Fugu, *Tetraodon nigroviridis*. Ayant obtenu plus de cent mégabases de séquence sur ce génome, l'équipe du Génoscope a cherché à l'employer pour détecter des gènes dans le génome humain. On peut en effet penser (et les étalonnages préliminaires réalisés sur des gènes humains connus le montrent) que les seules régions conservées entre ces deux ADN séparés par quatre cents millions d'années du point de vue de l'évolution sont les zones codantes. Diverses mises au point ont abouti à la méthode baptisée *Exofish*, qui définit à partir de la comparaison de séquence des régions conservées dans l'évolution baptisées *Ecores*, et déduisent du nombre d'*Ecores* le nombre de gènes. Sans rentrer dans le détail des calculs, décrits dans un article bien documenté paru en juin 2000 dans *Nature Genetics* [9], disons qu'ils paraissent convaincants, que l'application de cette méthode prédit environ 600 gènes sur le chomosome 22, et que les différents tests effectués donnent des résultats raisonnables. Bien entendu, la méthode n'exige pas que l'on dispose de l'ensemble de la séquence de l'un ou l'autre organisme.

L'application de cette procédure aux 42 % du génome humain contenus dans les bases de données publiques en décembre 1999 indique un peu moins de 12 000 gènes, soit pour l'ensemble du génome presque 28 000. En « tirant » au maximum les différents

paramètres vers une estimation plus haute, les auteurs arrivent à 34 000. En tout état de cause, ce chiffre est donc très bas par rapport à ce qui était généralement admis. Il semble pourtant avoir été établi de manière sérieuse... et il est en accord avec une autre analyse récente, pourtant fondée, elle, sur l'utilisation des EST.

Une autre manière d'employer les EST

Dans le même numéro de *Nature Genetics* [10], Brent Ewing et Phil Green (connu notamment pour ses performants algorithmes d'assemblage de séquence) ont appliqué aux données de séquence humaine une méthode déjà employée pour le nématode. En bref, il s'agit de mesurer le recouvrement entre deux jeux incomplets de séquences géniques (effectifs n1, n2) dont l'un au moins ne doit pas présenter de biais. Il est aisé de comprendre que le nombre m2 de recouvrements dépendra du nombre de gènes : plus ce dernier est élevé, moins les recouvrements entre les deux ensembles seront fréquents. En fait, avec des hypothèses raisonnables, on trouve que le nombre total de gènes est égal à $n1 \times n2/m2$. L'important, naturellement, est de bien choisir les jeux de séquences. Pour le premier, les auteurs ont pris soit les 679 gènes définis sur le chromosome 22, soit une série de 7 600 gènes obtenus en regroupant les séquences complètes d'ARNm contenues dans *Genbank*. On peut argumenter que ces deux collections sont assez proches d'un échantillonnage au hasard. Le deuxième jeu est construit à partir de dbEST en ne retenant (avec des critères de qualité assez stricts, y compris un réexamen des données de séquence brute) que les *clusters* contenant l'extrémité 3' de l'ARNm : les auteurs en définissent un peu plus de 43 000 à partir d'un million d'EST[3]. On détermine alors la fraction de séquences en commun entre le jeu de référence et celui tiré de dbEST, et l'on en déduit une estimation du nombre total de gènes. Les deux calculs aboutissent pratiquement au même chiffre, soit environ 34 000 gènes ! Cette utilisation intelligente de données partielles, assortie d'un regard très critique sur la qualité des séquences contenues dans dbEST, aboutit donc à un résultat qui conforte celui du Génoscope.

L'unanimité n'est pas atteinte : TIGR contre-attaque

Cette série d'articles dans *Nature Genetics* se termine néanmoins sur une estimation haute, celle de l'équipe de John Quackenbush à TIGR *(The Institute for Genomic Research)* [5]. Il s'agit cette fois de la construction d'un *gene index* censé être de très haute qualité, et de son emploi pour prédire le nombre total de gènes contenus dans notre ADN. Les auteurs ont tenté de résoudre les différents problèmes mentionnés précédemment. Ils sont partis de 1,6 millions d'EST tirés de dbEST, en ont éliminé près de cent mille suspectés d'être des contaminants, puis ont assemblé ces séquences entre elles (en utilisant leurs propres programmes et des critères stricts) et avec un jeu de 54 000 séquences complètes ou incomplètes de transcrits humains principalement obtenues à partir de la base de données *Genbank*. Ils ont éliminé tous les *clusters* contenant un seul EST (il y en avait plus de 300 000[4] et arrivent finalement à un chiffre de 75 000 *clusters*. Ils constatent que seulement 55 % des gènes connus (annotés dans Genbank) sont contenus dans ces *clusters*, et en déduisent donc que le nombre total de gènes est de 75 000/0,55 soit 136 000. En introduisant une correction pour tenter de tenir compte de l'épissage alternatif, leur estimation descend à 110 000. Leur calcul appliqué au chromosome 22 y prédit près de 1 800 gènes (au lieu des 679 trouvés) et l'extrapolation au génome entier arrive cette fois à 120 000.

3. Mais ils ne considèrent pas que ces 43 000 *clusters* représentent autant de gènes, *voir plus loin*.
4. Les chiffres donnés ici (nombre d'EST, et plus loin fraction de gènes représentés dans les *clusters*) sont différents de ceux que j'ai indiqués pour *Unigene* et correspondent vraisemblablement à des critères d'assemblage de séquence plus restrictifs.

Ces résultats sont donc en contradiction avec les deux articles discutés précédemment, et en revanche en accord avec les valeurs indiquées par les acteurs industriels de ce domaine : depuis longtemps tant *Human Genome Sciences* que *Incyte*, annoncent des chiffres largement supérieurs à 100 000. Des esprits malveillants pourraient y voir une manière pour ces entreprises de valoriser leur « trésor de guerre » vis-à-vis de leurs clients : la publicité d'*Incyte*, reproduite sur la *Figure 1*, montre qu'un tel soupçon n'est pas totalement absurde. Elle figurait en bonne place dans le numéro du 18 mai 2000 de la revue *Nature*, et annonce 60 000 gènes « *not available anywhere else* »... en s'appuyant sur une estimation de 120 000 gènes humains fondée sur l'analyse de la base de données de la firme. Évidemment, si le nombre total de gènes tombe à 30 ou 40 000, cet argument de vente devient beaucoup moins séduisant ! Plus sérieusement, il n'est pas certain que les auteurs de ce dernier article aient réellement pu éliminer tous les écueils de la construction d'un *gene index* malgré leurs efforts, et notamment qu'ils aient traité correctement la question de l'épissage alternatif. La correction de 20 % environ qu'ils introduisent pour en tenir compte semble bien faible en regard des données récentes indiquant la fréquence du phénomène [11, 12], d'autant plus que l'égalisation des banques tout comme la construction des *clusters* tendent à majorer sa contribution au nombre apparent de gènes.

Un pari et des conséquences

Cette question du nombre des gènes humain suscite un grand intérêt dans le milieu des génomistes, avec un vif débat lors de la dernière réunion de *Cold Spring Harbor*, et a suscité l'organisation d'un *sweepstake* dont les règles sont accessibles sur Internet, tout comme l'état actuel des paris. On voit sur la *Figure 2* (partie centrale de la page *Gene Sweepstake*) que l'éventail est large, de moins de 30 000 à plus de 300 000 (le petit nombre de parieurs est dû au fait qu'il faut être physiquement présent à Cold Spring Harbor pour s'inscrire). Résultat (et désignation du gagnant) à Cold Spring Harbor en 2003... Je n'ai pas encore enregistré mon pari, mais il se situera, on l'a deviné, dans la fourchette basse de l'estimation.

Plus sérieusement, ce nombre présente une grande importance pour notre compréhension de la biologie et de l'évolution. Si réellement nous n'avons que 30 ou 40 000 gènes, à peine deux fois plus que le nématode avec ses 959 cellules, sa physiologie et son comportement certes fascinants mais néanmoins fort simples par rapport aux nôtres, cela indique que la complexité de l'organisme découle essentiellement de la régulation des gènes, de leurs interactions et de celles de leurs produits – et pas avant tout de leur nombre. Cela nous éloigne de certaines visions simplistes, implicites dans la présentation de l'ADN comme le maître-plan *(blueprint)* de l'organisme, et montre quelle variété de structures et de fonctions peut découler d'un nombre d'éléments génétiques relativement restreint mais dont la combinatoire atteint une très grande complexité. Et sur le plan pratique, cela souligne l'importance de toutes les études fonctionnelles, de cet après-génome dont la plupart des outils restent à développer sinon à inventer.

D'autres conséquences d'une révision à la baisse sont à attendre. En ce qui concerne les mécanismes de l'évolution, il était assez courant ces derniers temps d'évoquer un processus de tétraploïdisation menant des 17 000 gènes du nématode ou de la drosophile aux 80 000 humains. Cette séduisante hypothèse d'une double duplication générale (suivie de la divergence des séquences permettant l'apparition de nouvelles fonctions) devient peu vraisemblable. Sans doute sera-t-on aussi conduit à une réévaluation du rôle biologique de l'épissage alternatif, dont les vicissitudes des *gene indexes* nous révèlent la fréquence insoupçonnée. L'incertitude actuelle souligne aussi, naturellement, la nécessité de progrès dans les méthodes d'annotation de génomes complexes. Il semble que les progrès soient assez lents dans ce domaine, les performances des logiciels de

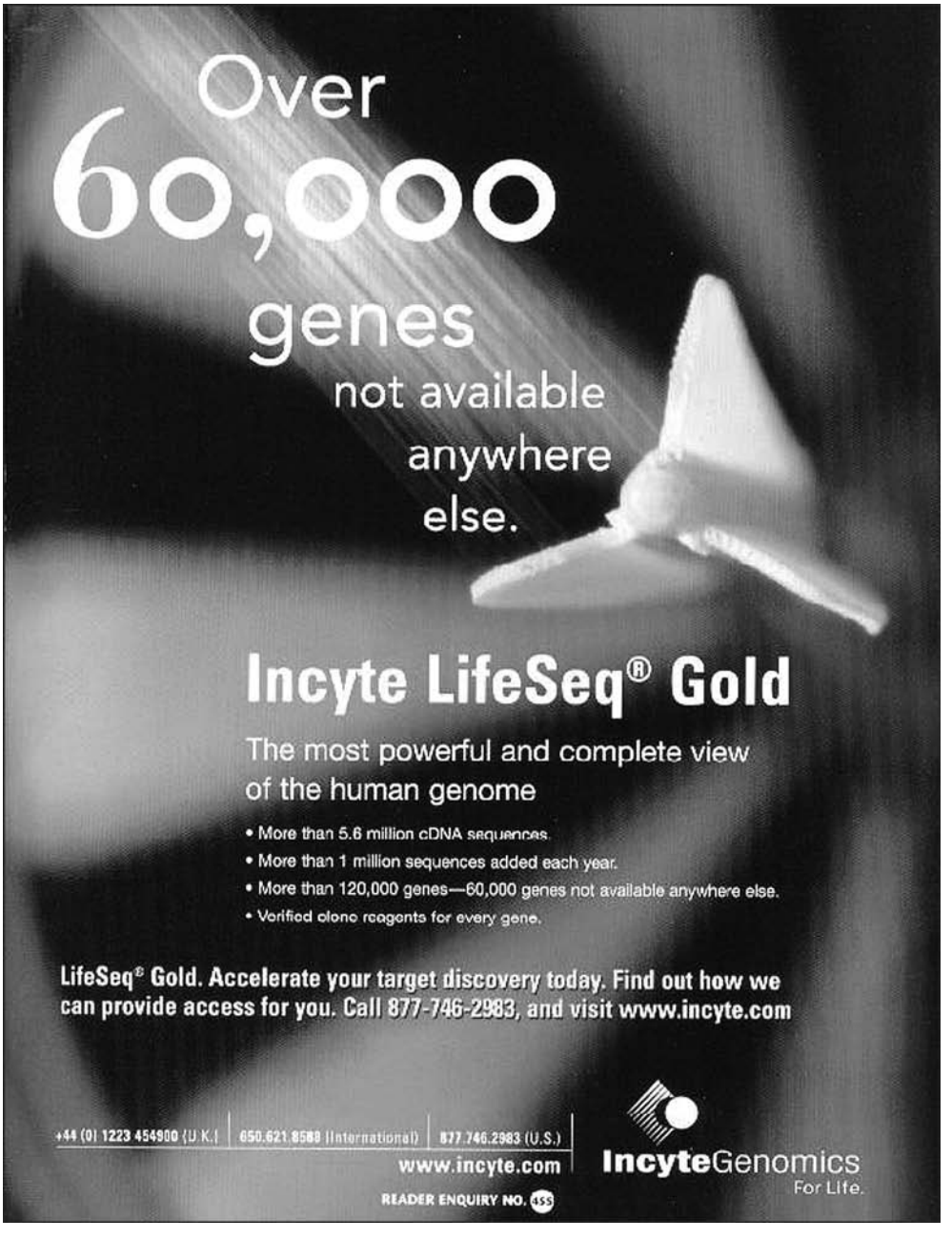

Figure 1. Publicité de la firme *Incyte*, parue dans la revue *Nature* (18 mai 2000).

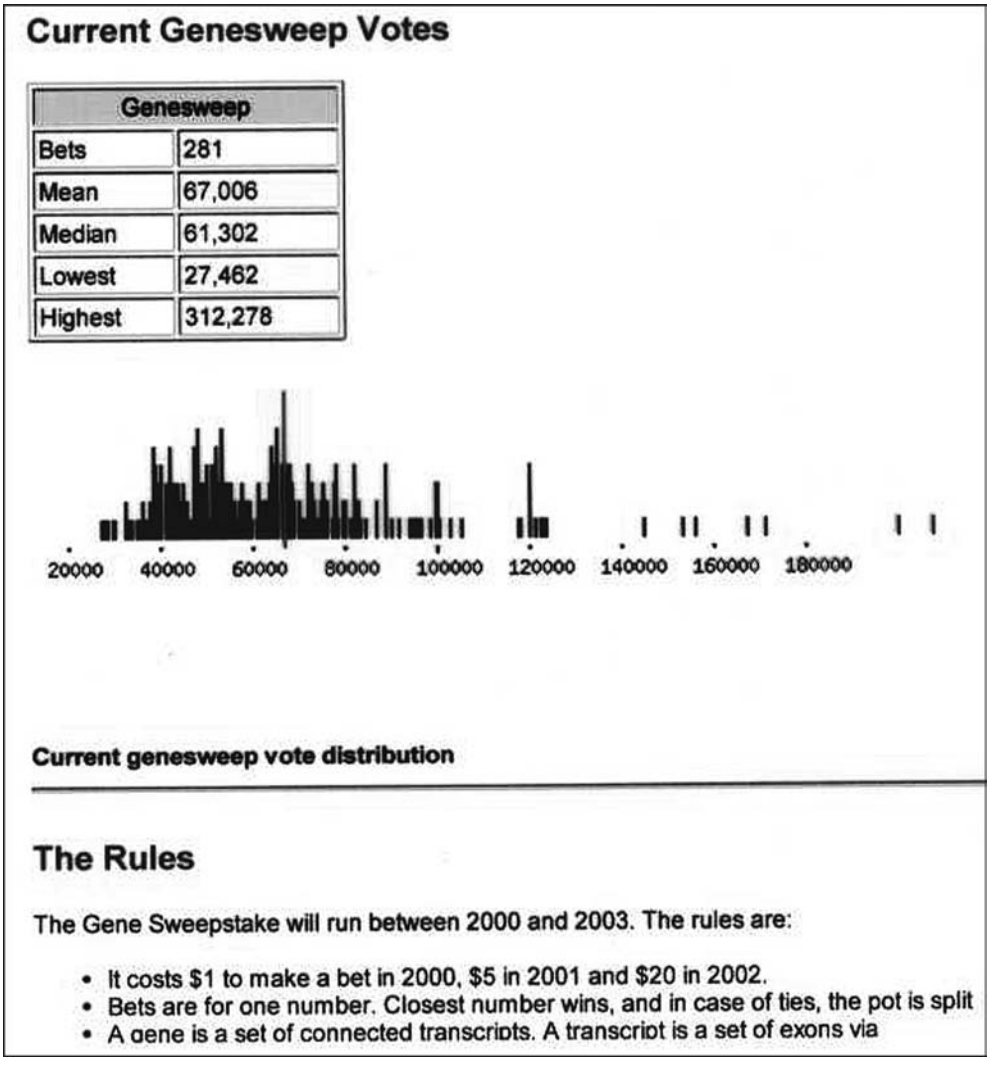

Figure 2. Le pari de *Cold Spring Harbor* (page *Gene Sweepstake*, sur le serveur *Ensembl*).

détection de gènes à partir de la séquence d'ADN n'ayant pas fait d'énormes progrès ces dernières années. La disponibilité de grandes quantités de données issues des génomes complexes va certainement stimuler la recherche et peut-être amener à mettre l'accent sur d'autres approches (comme le montre l'article du groupe du Génoscope). Les mathématiciens de plus en plus nombreux qui s'intéressent à l'interprétation du génome devraient trouver là matière à de fructueux travaux.

Références

1. Pennisi E. Human Genome Project. And the gene number is... ? *Science* 2000 ; 288 : 1146-7.
2. Dunham I, Shimizu N, Roe BA, *et al.* The DNA sequence of human chromosome 22. *Nature* 1999 ; 402 (6761) : 489-95.
3. Boguski MA, Schuler GD. ESTablishing a human transcript map. *Nat Genet* 1995 ; 10 : 369-71.
4. Adams MD, Kerlavage AR, Fleischmann RD, *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 1995 ; 377 : 3-174.
5. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120 000 genes. *Nat Genet* 2000 ; 25 : 239-40.
6. Hattori M, Fujiyama A, Taylor TD, *et al.* The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* 2000 ; 405 : 311-9.
7. Jordan B R. Fugu Story. *Med Sci* 1994 ; 10 : 1154-6.
8. Elgar G, Sandford R, Aparicio S, Macrae A, Venkatesh B, Brenner S. Small is beautiful : comparative genomics with the pufferfish (Fugu rubripes). *Trends Genet* 1996 ; 12 : 145-50.
9. Roest Crollius H, Jaillon O, Bernot A, *et al.* Estimate of human gene number provided by genome-wide analysis using tetraodon nigroviridis DNA sequence. *Nat Genet* 2000 ; 25 : 235-8.
10. Ewing B, Green P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* 2000 ; 25 : 232-4.
11. Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. *Genome Res* 1999 ; 12 : 1288-93.
12. Bretta D, Hankea J, Lehmanna G, *et al.* EST comparison indicates 38 % of human mRNAs contain possible alternative splice forms. *FEBS Lett* 2000 ; 474 : 83-6.

Le pari de Cold Spring Harbor *a trouvé son épilogue au printemps 2003. En fait, on ne connaît toujours pas le nombre exact des gènes humains, quoique la plupart des scientifiques s'accordent sur une valeur de l'ordre de 30 000. Mais les termes du pari lancé en 2000 stipulaient que le vainqueur devait être annoncé en 2003 ; ceci fut donc fait, en se fondant sur les données du site* Ensembl *(www.ensembl.org), projet commun à l'EMBL (Laboratoire européen de biologie moléculaire) et à l'EBI* (European Bioinformatics Institute) *et qui constitue sans doute la meilleure analyse actuelle du génome humain.* Ensembl *répertorie 24 847 gènes humains, et c'est donc Lee Rowen (*Institute for Systems Biology*, Seattle, États-Unis) qui a emporté le « pot » constitué par les 461 parieurs : il avait prévu 25 947 gènes.*

9. EN GUISE DE CONCLUSION...

Ces chroniques émanent d'un observateur jouissant d'une relative indépendance. Durant la période où elles ont été rédigées, je n'étais ni directeur d'institut, ni responsable de département au CNRS ou à l'Inserm, mais simple chef d'équipe dans un bon laboratoire (le Centre d'Immunologie Inserm/CNRS de Marseille-Luminy) bénéficiant d'un financement très correct. Ayant effectué en 1991 une sabbatique-enquête d'une année qui me donnait une vision d'ensemble du monde du Génome, participant à plusieurs comités nationaux et étrangers intervenant dans ce domaine, je n'avais pourtant pas de responsabilité institutionnelle lourde. Je pouvais donc m'exprimer librement, développer une vision personnelle (d'ailleurs parfois erronée) sans non-dit ni langue de bois, d'autant que la revue médecine/sciences *n'a jamais exercé aucune censure sur la forme ou sur le fond de mes chroniques. L'histoire ainsi racontée à travers ces textes est avant tout sincère, malgré les antagonismes très vifs qui régnaient dans le milieu, les oppositions entre AFM, GREG, ministère, organismes de recherche et les personnalités parfois explosives de certains protagonistes. Je ne suis pas peu fier, à cet égard, du fait que mon texte sur le Très Grand Séquençage, paru dans* médecine/sciences*, ait été repris tel quel et* in extenso *tant dans le bulletin des Sciences de la Vie du CNRS que dans le Journal de l'AFM.*

Pour autant, l'histoire ainsi racontée est incomplète : par exemple, elle ne décrit pas la compétition entre le programme public et l'entreprise Celera *pour l'obtention du premier brouillon de la séquence humaine. D'une manière générale, elle témoigne d'une prise de conscience un peu tardive du rôle des entreprises en recherche génomique. Je pense pourtant que ce recueil, malgré ses défauts et ses manques, aura permis au lecteur de revivre une période passionnante, et peut-être même d'en tirer quelques leçons... Ces dix années ont vu la recherche biologique passer (au moins en partie) de l'artisanat à la* Big science*, tandis que faisaient irruption la technologie et l'informatique, et que les questions posées changeaient de nature. Ce basculement déjà très avancé va certainement se poursuivre ; il n'est pourtant pas dit qu'il doive s'étendre à l'ensemble de cette science. Au contraire même, certaines disciplines abusivement délaissées au profit de la toute-puissante biologie moléculaire, comme la*

physiologie ou la systématique, devraient être réinvesties par les chercheurs et, bien sûr, par les organismes qui les financent[1]*. Mais la connaissance de l'architecture et du texte de notre génome, et de celui de bien d'autres organismes, est un acquis irréversible dont il faudra tenir compte dans tous les secteurs de cette science du Vivant qui va, j'en suis convaincu, dominer le III*e *millénaire.*

1. On est conscient aujourd'hui – pour ne citer que cet exemple – des difficultés rencontrées pour analyser en détail le phénotype parfois fort discret de la souris mutante résultant de l'inactivation dirigée d'un gène.


Achevé d'imprimer par Corlet, Imprimeur, S.A.
14110 Condé-sur-Noireau
N° d'Imprimeur : 72563 - Dépôt légal : octobre 2003
Imprimé en France

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank