m/s DOSSIER

médecine/sciences 1992 ; 8 : 1100-1

Des cartes en voie d'intégration ?

Bertrand R. Jordan

Une réunion sur « la construction de cartes chromosomiques intégrées chez l'homme » s'est tenue récemment* 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).* 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 « Multi-Map », Jean Thierry-Mieg (Montpellier) pour ACeDB, Otto Ritter (Heidelberg) pour le projet européen d'IGD (*Integrated Genome Database*), Guy Vasseix pour le service informatique du Généthon et, bien sûr, Jean Frézal pour Genatlas.

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 combinant 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, 4910 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 décifit (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. Cela 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.

* *Séminaire sur les cartes intégrées, Gif-sur-Yvette, 23 et 24 octobre 1992.*

Une tentative intéssante

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 *Ressource 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, comme 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, représentées 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 local 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 à la baisse continue des tarifs, une station de travail sous Unix très convenable peut maintenant être obtenue pour moins de 50 000 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 souhaitable 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... ■

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

TIRÉS A PART

B.R. Jordan.