m/s

Chroniques génomiques

par Bertrand JORDAN

Génome et informatique : condamnés à s'entendre ?

Besoins importants, moyens à la hauteur

Après quatre mois d'enquête aux États-Unis à travers une vingtaine de laboratoires, une donnée apparaît de façon évidente : l'importance de l'informatique. Ce n'est pas vraiment une surprise, mais l'effort fait dans le cadre du programme Génome est impressionnant : 10 à 20 % des crédits sont consacrés à cette discipline, ce qui est considérable. Cela signifie en pratique qu'un « Genome Center » typique comporte une équipe d'informatique de quatre à huit personnes (sur une quarantaine de chercheurs et de techniciens en tout), beaucoup plus que ce que l'on trouve dans la plus grosse de nos unités Inserm ou Cnrs ; que la Genome Data Bank (GDB) mise en place par Peter Pearson à Baltimore (MD, USA) dispose de très gros moyens ; et que la station de travail est en passe de remplacer le Mac haut de gamme ou le PC/AT dans les laboratoires ou les bureaux des chercheurs qui travaillent sur le génome dans ce pays.

Impératifs multiples : saisie et interprétation des données...

Les méthodes utilisées dans ce type de recherche impliquent de plus en plus souvent un ordinateur qui met directement les données en mémoire : analyse d'images de microscopie confocale ou évaluation de signaux en *in situ* fluorescent saisis par une caméra CCD, commande d'appareils à champs pulsés selon des cycles complexes, ou encore exposition d'autoradiogrammes sur *imaging plate* analysé ensuite par faisceau laser. La manipulation ultérieure des informations ainsi obtenues fait alors tout naturellement appel à l'informatique : au lieu de superposer à la main et à l'œil une photo agrandie d'un gel coloré au bromure d'éthydium et l'autoradiographie d'une hybridation du même gel après transfert, on fera coïncider électroniquement les images *directement enregistrées* de ces deux étapes de l'expérience... Ce qui permettra au passage de les redresser, de les amener exactement à la même échelle et d'appliquer des critères objectifs de coïncidence de bandes, tout en archivant directement le résultat de la comparaison. C'est un exemple, il y en a beaucoup d'autres !

... Cahiers de laboratoire informatisés...

L'amélioration des techniques, leur début d'automatisation aboutissent inévitablement (c'est après tout leur but !) à multiplier les informations, les clones, les données : il devient impossible de les comptabiliser et les archiver à la main. Les méthodes manuelles ne permettent en effet de gérer qu'un nombre limité de ces objets, surtout dans un laboratoire de recherche aux thèmes fluides, au personnel changeant et dans lequel tout le monde est très occupé : qui n'a pas vécu l'énervante recherche du clone qu'avait isolé un post-doc (parti depuis) il y a un an ou deux ? D'autant que l'affaire se poursuit une fois le clone retrouvé : il reste encore à mettre la main sur sa carte de restriction et sur le « bout de séquence » qu'avait fait un étudiant pour définir des amorces PCR... Multiplier par dix ou cent le nombre de segments d'ADN isolés et caractérisés sans perfectionner considérablement leur stockage et l'archivage des informations qui les concernent, c'est courir à la catastrophe. Il y a donc un besoin évident de « cahiers de laboratoire informatisés », d'un système informatique commode où les informations soient consignées au fur et à mesure par chaque membre de l'équipe après vérification selon un processus bien défini : le but est de pouvoir retrouver non seulement le clone que l'on cherche, mais aussi toutes les informations qui ont été un jour ou l'autre obtenues à son sujet.

Un exemple : Lawrence Livermore

J'ai pu voir un exemple assez évolué de ce type de cahier-base de données à Lawrence Livermore (CA, USA), dans le Genome Center où est étudié le chromosome 19 sous la direction de Tony Carrano. L'analyse est fondée principalement sur l'étude d'une librairie de cosmides provenant de ce chromosome, cosmides que l'on s'efforce d'assembler en *contigs* (série de clones présentant des recouvrements). Pour ce faire, la « signature » de chaque cosmide (une série de fragments de restriction le caractérisant) est acquise directement sur un système dérivé du séquenceur d'ADN Applied Biosystems ; la comparaison des signatures (8 000 environ à ce jour) permet au système de déduire quels cosmides sont susceptibles de constituer un *contig*. Ce dernier est alors répertorié dans le système informatique et peut être montré sur l'écran avec toutes ses caractéristiques — y compris son degré de fiabilité, correspondant à l'étendue du recouvrement observé et figuré par une couleur sur l'écran.

Les *contigs* sont ensuite vérifiés par une autre méthode : hybridation *in situ* des deux cosmides extrêmes, qui doivent (compte tenu de la résolution de cette technique) apparaître pratiquement confondus sur le chromosome ; l'exploitation d'une librairie de YAC *(yeast artificial chromosomes)*, enfin, permet de relier les *contigs* entre eux. Toutes ces informations, et l'« histoire » de chaque *contig*, sont conservées en mémoire et sont accessibles : ce qui autorise un retour sur l'évolution des résultats, l'application de critères changeants mais uniformes et l'affinement de la stratégie utilisée. Ce système remarquable, œuvre d'un groupe dirigé par Elliott Branscome, est un bon exemple d'interaction entre informatique et biologie ; ce n'est sans doute pas par hasard qu'il a été réalisé dans un laboratoire du DOE (*Department of Energy*, équivalent, aux États-Unis, de notre CEA), où le niveau technologique est bon et le personnel suffisamment stable pour que la symbiose recherchée puisse avoir lieu. A l'opposé, j'ai vu dans un laboratoire dont je tairai le nom une superbe base de données... utilisée uniquement par l'informaticien et le patron, mais que les post-docs, préoccupés avant tout de la recherche du gène de « leur » maladie et peu concernés par la stratégie d'ensemble du groupe, ignoraient superbement !

... Bases de données semi-privées ou semi-publiques...

Mais ces résultats, ces clones ne sont pas réservés au laboratoire qui les a obtenus : ils doivent être diffusés à des collaborateurs extérieurs, puis (le plus vite possible, on l'espère...) mis à la disposition de la communauté scientifique selon des modalités précises et après une procédure de validation rigoureusement définie. Cela nous amène à la question des banques de données, de leur structure, de leur accès et de leurs relations. Leur structure doit être très souple : les « objets » à classer étaient hier des plasmides, ce sont aujourd'hui des cosmides, des YAC ou des STS *(sequence tagged sites)*, couple de séquences oligonucléotidiques définissant par PCR un point unique du génome (voir la Chronique sur l'« OPA des STS », *m/s n° 2, vol. 7, p. 175*), demain ce seront peut-être des clones P1 ou des points d'échange de chromatides sœurs... La nature des relations entre ces différentes données peut aussi changer, d'où l'intérêt général pour les structures informatiques de type relationnel comme le système « Sybase » qui sont bien adaptées à la gestion de ce genre de situation. Les laboratoires importants développent ainsi leur propre dispositif, faisant en général appel à une ou plusieurs stations de travail, quoique le Mac et le logiciel « Hypercard » gardent encore quelques fidèles. Ces systèmes, ou du moins les plus évolués d'entre eux, gèrent plusieurs sous-ensembles de données : les résultats propres au laboratoire, non encore vérifiés ou hautement concurrentiels ; ceux qui proviennent de, ou sont accessibles à, des collaborateurs extérieurs ; et enfin les données auxquelles tout un chacun peut accéder et qui sont, ou seront, transférées vers les bases de données publiques. La gestion de ces divers niveaux de confidentialité n'est pas simple, et la communication entre les différentes banques (avec les contrôles afférents) pose d'intéressants problèmes informatiques, organisationnels et même politiques.

... Banques générales...

L'aboutissement de ce processus est l'obtention d'un ensemble de données solidement établies (résultant d'un consensus) et pouvant servir de base aux travaux suivants. C'était, on se le rappelle, l'objectif des *Human Gene Mapping Workshops* qui se tenaient tous les deux ou trois ans depuis le début des années 70 ; les résultats relatifs à chaque chromosome y faisaient l'objet de mises au point et de débats parfois passionnés. De ce conclave sortait un livre (de plus en plus épais au fil des années) rassemblant la totalité des informations, des sondes, des polymorphismes, des cartes génétiques et physiques. Cette sorte de bible servait alors de référence jusqu'à la conférence suivante. La nécessité d'un relais informatique est rapidement apparue en raison de la masse sans cesse grandissante de données ; à l'heure actuelle, cette fonction est principalement assurée par la « Genome Data Base » (GDB) installée à John Hopkins University (Baltimore, MD, USA) sous la responsabilité de Peter Pearson et avec le soutien financier du *Howard Hughes Medical Institute* dans un premier temps, du NIH *(National Institutes of Health)* et du DOE *(Department of Energy)* dans un deuxième. La GDB regroupe, l'ensemble des résultats concernant le génome, à l'exception des séquences d'ADN* ; cela va des sondes aux sites fragiles en passant par les cartes, les données sur la souris et l'atlas des maladies génétiques humaines établi par Victor McKusick, sans oublier le nom et le numéro de fax de la personne à qui l'on peut demander tel ou tel clone... Banque récente, bien structurée, bénéficiant de moyens considérables, elle est très complète et extrêmement utile, même si son emploi demande un certain apprentissage. Elle est en principe accessible sans conditions, en mode lecture *(read only)* ; l'introduction de données, elle, est soigneusement contrôlée par un processus complexe faisant intervenir les *chromosome editors* désignés lors des *Human Gene Mapping Workshops*. Des « nœuds » secondaires doivent faciliter la consultation de GDB en réduisant les coûts de communication : l'un d'eux est déjà installé en Angleterre (Northwick Park, Londres), le démarrage d'un deuxième à Heidelberg (RFA) dans le cadre du centre de recherches sur le cancer (DKFZ) est imminent et des négociations sont en cours avec d'autres pays dont le Japon...

... Et enjeux politiques !

Mais GDB n'est pas seule au monde. Dans notre pays la base « Genatlas » a été créée à l'initiative de Jean Frézal à partir de 1987 *(voir m/s n° 6, vol. 7, p. 595)* ; d'autres existent ailleurs, notamment au Japon. La concurrence est sans doute souhaitable : on peut légitimement penser qu'il n'est pas bon de laisser le monopole du stockage de l'information génétique à un organisme implanté dans

* La question des banques de données de séquence mérite, elle aussi, une discussion approfondie ; ce sera — peut-être — pour une autre fois...

un pays déjà largement dominant dans ce secteur de recherche, et où quelques responsables ont fait preuve de tendances impérialistes non négligeables. GDB est en effet totalement financé par les États-Unis à l'heure actuelle, et les *chromosome editors* (ceux qui contrôlent l'entrée des données dans la banque) sont, pour la plupart, anglo-saxons. Dans un domaine très proche, l'existence de deux banques majeures pour la séquence d'ADN (EMBL, européenne, et GenBank aux États-Unis) a été bénéfique en assurant une certaine émulation, et un sain pluralisme. Le problème est que nous n'en sommes pas là pour les données de cartographie : les moyens mis en œuvre par GDB (une trentaine de personnes, plusieurs millions de dollars de budget annuel) sont supérieurs d'un bon ordre de grandeur à ceux dont disposent ses « concurrents », et la connexion directe avec les *Human Gene Mapping Workshops*, dont GDB est la banque de données officielle, verrouille en quelque sorte le système. Il y a là de difficiles décisions politiques à prendre : faut-il tenter de mettre en place une alternative à GDB (avec les investissements correspondants, une ou deux dizaines de millions de francs par an), faut-il plutôt chercher à s'associer tant au financement *qu'à la gestion* de cette banque ? C'est la deuxième voie qu'a recommandée récemment la Fondation européenne pour la science (ESF)... En tout état de cause, il paraît essentiel que la communauté scientifique française puisse avoir accès dans de bonnes conditions à l'ensemble des données existantes, et donc en particulier à GDB.

Difficile fusion de deux cultures

On voit qu'une discussion de l'informatique peut nous emmener sur des terrains inattendus... Nous terminerons cette chronique en évoquant une autre question qui, pour « extra-scientifique » qu'elle soit, n'en présente pas moins une grande importance. Il s'agit de la collaboration entre biologistes et informaticiens : même aux États-Unis, où d'importants moyens ont été débloqués pour mener des projets communs, cela ne se passe pas toujours très bien. Il s'agit en effet d'effectuer une jonction, mieux, une fusion entre deux communautés dont la formation, le langage, la « culture » pourrait-on dire sont très différents. Les motifs d'incompréhension ne manquent pas, et la communication demande un sérieux effort : effort de la part des biologistes pour bien conceptualiser et expliciter leurs besoins, pour acquérir un minimum de connaissances en informatique (*computer literacy* comme disent nos collègues anglo-saxons), pour accepter aussi que tout logiciel un peu complexe exige une période de mise au point ; effort également de la part des informaticiens pour privilégier la recherche de solutions rapides (rapidement opérationnelles) et efficaces même si elles ne sont pas optimales : *quick and dirty*, comme l'on dit ici, aux dépens de l'élégance informatique. Inutile de dire que cette symbiose n'est pas réalisée partout ; et, curieusement, le fait que le responsable d'un groupe d'informatique soit ou non biologiste de formation ne semble pas — du moins dans les cas dont j'ai eu connaissance — jouer un rôle déterminant.

Conclusion (provisoire, car tout change vite...)

Au terme de cette discussion, qui reflète le point de vue d'un usager plus que celui d'un spécialiste (que les professionnels de l'informatique ou des bases de données me pardonnent mes approximations...), j'espère avoir pu faire ressentir au lecteur l'importance de la question, et sa complexité. Il ne servirait assurément à rien d'accumuler les informations sur le génome si nous n'étions pas capables de les archiver, d'y accéder et, si possible, d'y comprendre quelque chose ; mais les façons d'assurer ces tâches, et même le détail des objectifs, restent en grande partie à définir, et l'enjeu politique n'est pas loin : le savoir et le pouvoir ont, on le voit, des relations très intimes... ■

Bertrand R. Jordan

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