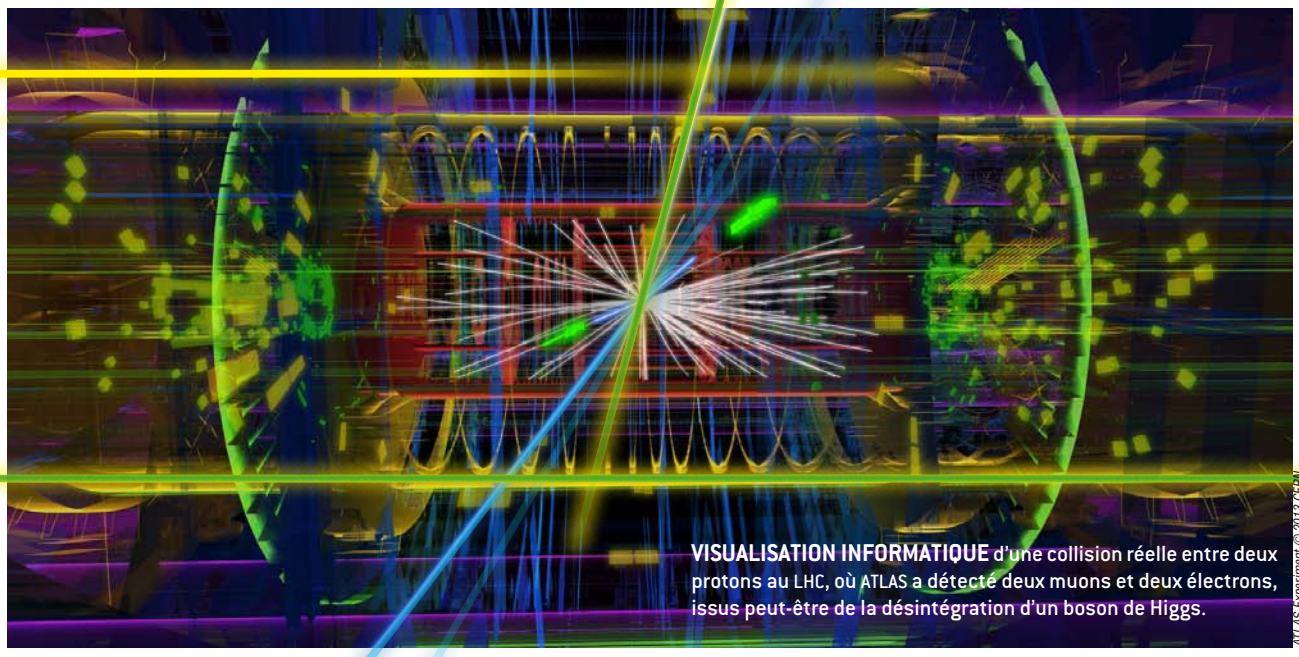

**VISUALISATION INFORMATIQUE** d'une collision réelle entre deux protons au LHC, où ATLAS a détecté deux muons et deux électrons, issus peut-être de la désintégration d'un boson de Higgs.

ATLAS Experiment © 2013 CERN

ans de prise de données (2011 et 2012) où des millions de particules ont été produites (dont environ 500 000 bosons de Higgs, en théorie), ATLAS n'a détecté que quelques centaines de bosons de Higgs se désintégrant en deux photons (contre quelques milliers si le détecteur avait été parfait).

Les flux de données que transmettent les détecteurs sont considérables. Les collisions engendrent en effet des milliers de particules secondaires, le passage d'une particule dans un détecteur se traduisant par des impulsions électroniques à partir desquelles les physiciens peuvent déterminer sa trajectoire et son identité. Les millions d'impulsions électroniques créées par chaque collision sont alors numérisées par des circuits intégrés aux détecteurs, avant d'être transmises par fibres optiques aux ordinateurs d'acquisition de données. Elles y sont regroupées sous formes de fichiers binaires de « données brutes ».

Chaque collision constitue, au niveau de chacune des quatre expériences, un événement dont la taille en données brutes est de l'ordre du mégaoctet. Un tri intelligent est effectué en moins d'une seconde pour ne garder que les événements potentiellement intéressants. Réalisé grâce aux circuits électroniques en amont puis à des grappes d'ordinateurs et des algorithmes spécialisés, ce tri réduit le taux d'acquisition à environ 200 événements intéressants par seconde. Ces données brutes sont enfin enregistrées à un taux équivalent à deux CD par seconde pour chacune des quatre expériences, sachant que la capacité d'un CD est de 700 mégaoctets.

## Un gigaoctet par seconde

Les données brutes sont traitées afin d'extraire les informations « physiques » (les caractéristiques de chaque trajectoire, l'impulsion et l'identité de chaque particule détectée), puis de les analyser. Ces traitements réduisent la taille de l'information d'un facteur deux à dix. Ils sont en général répétés plusieurs fois par an sur l'ensemble des données acquises depuis le démarrage, pour tenir compte du développement continu des algorithmes de traitement et d'étalonnage du détecteur.

Chacun des quatre complexes de détection enregistre en un an $10^{10}$ collisions, chiffre comparable au nombre d'étoiles de la Voie lactée. Le traitement des données ainsi accumulées est un défi, en raison tant des flux (de l'ordre du gigaoctet par seconde) que des volumes (plusieurs dizaines de pétaoctets chaque année, un pétaoctet étant égal à 1 000 téraoctets, soit $10^{15}$ octets).

À tout instant, quelques milliers de chercheurs dans le monde sollicitent des ressources de calcul et de stockage afin d'analyser ces données. Pour faire face à ce défi, une « grille de calcul » a été mise en œuvre.

Cette grille met en commun les ressources de calcul et de stockage de nombreuses machines indépendantes, chaque processeur traitant un seul événement. Les événements étant indépendants les uns des autres, peu de données sont à transférer entre les processeurs. Le principal intérêt de ce système est qu'il suffit d'ajouter de nouvelles entités pour augmenter d'autant les capacités de calcul de la grille, laquelle peut traiter des pétaoctets de données. La grille du LHC s'appuie ainsi sur près de 200 centres de calcul répartis dans le monde entier. Ce réseau est hiérarchisé en quatre niveaux, selon la nature et la qualité des services fournis, connectés par des liaisons à haut débit.

Pour traiter l'ensemble des données enregistrées par les expériences du LHC, les besoins en capacité de calcul atteints en 2012 équivalaient à 500 000 ordinateurs de bureau et plus de 200 pétaoctets d'espace de stockage sur disques. Un gigantisme informatique à la mesure du gigantisme (27 kilomètres de diamètre...) du LHC lui-même ! ■

# Génomes, protéomes et transcriptomes à foison

Guy Perrière

**Quelques grandes banques de données centralisent les résultats du séquençage de l'ADN, des ARN et des protéines des organismes vivants. Mais l'avalanche de séquences fait émerger un autre type d'organisation.**

L'étude moderne du monde vivant et de son histoire repose en bonne partie sur l'analyse bio-informatique de données moléculaires provenant des différents organismes. Il s'agit en particulier du matériel génétique – le génome, dont le support est l'ADN –, des ARN transcrits par le génome, ou « transcriptome », et des protéines exprimées – le « protéome ».

La possibilité d'accéder rapidement, par le biais de banques de données, aux séquences de nucléotides (pour les génomes et les transcriptomes) ou d'acides aminés (pour les protéomes) est l'une des raisons des succès qu'a rencontrés la bio-informatique au cours des 30 dernières années. Or l'arrivée à partir de 2005 de nouvelles techniques de séquençage à très haut débit, puis leur démocratisation vers 2010, ont révolutionné ce domaine.

En effet, si les instituts en charge de la maintenance des banques ont pu gérer pendant trois décennies un flux sans cesse croissant de données, ce n'est désormais plus le cas. Confrontés à un déluge sans précédent de séquences, ces organismes doivent désormais s'adapter et, ici comme dans bien d'autres domaines des sciences et de la technologie, l'avenir appartient aux structures en réseau et non plus à des grands systèmes centralisés.

L'histoire des banques de données de séquences prend ses racines en 1965 avec la publication par la physico-chimiste américaine Margaret Dayhoff du premier volume de son *Atlas of Protein Sequence and Structure*. Entre 1965 et 1978, ces livres, qui compilent les séquences de quelques dizaines de protéines, constitueront une source d'inspiration pour toutes les futures collections informatisées. C'est ainsi que *GenBank*, la toute première banque de séquences nucléotidiques d'ambition internationale, a été mise en place en 1979 au Laboratoire américain de Los Alamos. Elle sera rapidement suivie par la création de celle de l'EMBL (le Laboratoire européen de biologie moléculaire) en 1981, puis de la *DNA Data Bank of Japan* (DDBJ) en 1984.

Si la DDBJ est gérée depuis sa création par le *National Institute of Genetics* à Mishima, *GenBank* et la banque de l'EMBL ont connu plusieurs localisations différentes. Aujourd'hui, *GenBank* est maintenue par le NCBI *(National Center for Biotechnology Information)* à Bethesda. La base de l'EMBL l'est par l'EBI *(European Bioinformatics Institute)* à Hinxton, en Grande-Bretagne, qui dispose d'une capacité de stockage d'environ 18 pétaoctets ($18 \times 10^{15}$ octets).

Bien qu'à leurs débuts, le contenu et la taille de ces trois banques étaient relativement différents, une collaboration internationale s'est peu à peu établie et, depuis une vingtaine d'années, leur contenu est virtuellement identique : une séquence soumise à n'importe lequel des trois centres est transmise aux deux autres dans un délai de 24 heures au plus.

## Les grandes banques ne sont plus exhaustives

La caractéristique principale de ces banques est qu'elles permettent d'accéder librement à la quasi-totalité des séquences biologiques obtenues par des laboratoires publics ainsi que par un grand nombre de laboratoires privés. C'est cet accès non limité qui a contribué aux avancées importantes obtenues par la bio-informatique au cours de son existence en tant que discipline scientifique. En effet, cette disponibilité immédiate et exhaustive des données a énormément facilité la reproductibilité des résultats, un point essentiel en recherche scientifique.

Pendant près de 30 ans, le volume des données soumises à ces trois collections a crû de façon exponentielle, avec un temps de doublement moyen de l'ordre de 18 mois. La période 2000-2010 a même vu ce temps de doublement diminuer, avec

le séquençage de nombreux génomes ou transcriptomes complets, dont le premier génome humain en 2001.

Cependant, depuis 2010, un changement de tendance assez surprenant s'est produit : ce temps de doublement s'allonge, alors même que les méthodes de séquençage à très haut débit se sont généralisées et qu'il n'y a jamais eu autant de séquences produites dans les laboratoires. Une première explication à ce phénomène inattendu tient au fait que les centres responsables de la maintenance des banques de séquences sont de moins en moins à même de supporter la charge financière que représentent l'achat continuel de capacités de stockage supplémentaires ainsi que le développement et la maintenance des infrastructures associées.

Un autre problème est que les volumes de données sont désormais tels qu'il n'est plus possible de les transmettre en un temps raisonnable aux centres *via* le réseau. Une solution de plus en plus adoptée, pour qui veut voir ses séquences figurer dans les banques, est d'expédier un disque dur sur lequel sont sauvegardées les séquences en question, le transfert s'effectuant ensuite sur place ! Il s'agit là d'une sorte de retour aux pratiques antérieures à Internet : jusque vers la fin des années 1980, c'est par envoi postal de supports physiques (à l'époque, des bandes magnétiques ou des disquettes) que l'on transférait les séquences.

Enfin, la question de l'accès aux séquences « brutes », non assemblées et non annotées (c'est-à-dire telles qu'elles sortent des machines de séquençage), est également un problème d'importance. Du fait de la quantité de séquences de ce type, les centres ne proposent plus un accès direct aux entrées individuelles, mais à des archives compressées pouvant en contenir un très grand nombre. La survie de ces archives est fréquemment remise en question, ce service ayant déjà été supprimé une fois au NCBI, puis rétabli.

Dans ce contexte, de plus en plus de séquences ne sont tout simplement pas envoyées aux centres de saisie et leur mise à disposition à la communauté se fait par l'intermédiaire de banques de données locales, mises en place dans le cadre de projets limités. Une conséquence de cela est qu'il existe désormais, outre les trois collections généralistes mentionnées, une multitude de banques spécialisées, que celles-ci soient dédiées à un organisme spécifique ou à une problématique biologique particulière. La revue *Nucleic Acids Research* publie chaque année un numéro spécial consacré aux principales banques disponibles dans le monde. Dans son édition de janvier 2013, elle en recensait pas moins de 178. Cependant, ce catalogue est très loin d'être exhaustif et le nombre de banques spécialisées existantes est bien plus élevé.

## Structuration en réseau et moyens mutualisés

Du fait qu'une quantité croissante de séquences ne sont plus envoyées aux centres de saisie, les trois collections généralistes ne peuvent plus être considérées comme complètes. Or cette perte d'exhaustivité a d'ores et déjà des répercussions sur la facilité de la reproduction des résultats qui était l'apanage de la bio-informatique. Dans ce contexte, une tendance lourde est à la structuration en réseau et à la mutualisation des moyens.

En France, depuis 2005, la bio-informatique s'est structurée sous la forme d'un réseau national (le ReNaBi) regroupant six centres régionaux, chaque centre fédérant un nombre variable de plates-formes et d'équipes de recherche. Suite à l'obtention en 2011 d'un financement dans le cadre des « Investissements d'avenir » du gouvernement, le ReNaBi est en train d'évoluer vers une nouvelle structure qui prendra le nom d'IFB (Institut français de bio-informatique). L'IFB sera constitué d'une part d'une plate-forme centrale et, d'autre part, des six centres ReNaBi originaux. Ce financement à long terme permettra d'accroître sensiblement les capacités de calcul et de stockage des données de séquençage mises à la disposition de la communauté des chercheurs français.

Au niveau européen, l'EBI a lancé, début 2007, l'initiative ELIXIR. Elle vise à fédérer les grands centres de bio-informatique (nationaux ou régionaux) dans un réseau européen, chaque nœud ayant une spécificité thématique. Tout comme dans le cas français, une telle structuration permettrait effectivement de répartir la charge, aussi bien en quantités de données à gérer qu'en termes d'infrastructures, pour ce qui est du stockage ou de l'archivage des séquences. Dans le cas d'une participation de la France à cette initiative, c'est l'IFB qui deviendrait le nœud national ELIXIR. ■

## Quelques chiffres

**6 887 génomes complets publiquement accessibles (en septembre 2013).**

**670 milliards de nucléotides (324 millions de séquences) dans la version de septembre 2013 de la banque de l'EMBL, soit environ 223 fois le génome humain.**

**54 milliards de nucléotides supplémentaires entre juin et septembre 2013, soit plus que tout ce qui a été introduit dans la banque de l'EMBL entre juin 1982 et septembre 2004.**

**3 milliards de séquences de 100 nucléotides de long produites au cours d'une seule expérimentation sur une machine *Illumina HiSeq 2500*.**


### L'AUTEUR

Guy PERRIÈRE est directeur de recherche du CNRS à l'UMR 5558, à Lyon. Il dirige le Pôle Rhône-Alpes de bio-informatique (PRABI) et est représentant français dans le réseau européen EMBnet.


### BIBLIOGRAPHIE


M. A. Quail *et al.*, **A tale of three next generation sequencing platforms : comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers**, *BMC Genomics*, vol. 13, article 341, 2012.

Y. Kodama *et al.*, **The Sequence Read Archive : explosive growth of sequencing data**, *Nucleic Acids Res.*, vol. 40, pp. D54-D56, 2012.

G. Cochrane *et al.*, **The future of DNA sequence archiving**, *GigaScience*, vol. 1, article 2, 2012.