

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 continu 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.

SDSS, un relevé... astronomique !

Une carte tridimensionnelle couvrant le tiers de la voûte céleste et recensant plus de 470 millions d'astres : c'est notamment ce qu'a produit SDSS, un projet emblématique des données massives en astronomie.

Céline Reylé

Chaque nuit depuis l'année 2000, le télescope de l'Observatoire d'Apache Point, dans les monts Sacramento, au Nouveau-Mexique (États-Unis), scrute le ciel. Grâce à cet instrument de 2,5 mètres de diamètre, les astronomes établissent un gigantesque relevé du ciel, nommé SDSS (*Sloan Digital Sky Survey*). Il aura fallu plus d'une décennie d'observations avec ce télescope dédié pour construire l'image du ciel la plus grande et la plus détaillée jamais obtenue (voir l'illustration).

Cette image est unique non seulement par l'étendue couverte, puisqu'elle balaie le tiers du ciel, mais aussi par sa profondeur de champ : elle recense tous les objets célestes situés dans la ligne de visée de la caméra jusqu'à des distances cosmologiques. L'accumulation de plusieurs millions d'images prises avec cette caméra a produit une image d'un térapixel (10^{15} pixels), en cinq couleurs. L'image est tellement grande que 500 000 écrans de télévision à haute définition seraient nécessaires pour la visualiser entièrement à sa résolution maximale. Elle est de surcroît complétée par des analyses spectroscopiques d'un grand nombre des astres détectés.

Les objectifs scientifiques du SDSS sont multiples et le programme comporte, dans sa phase actuelle, quatre champs

d'étude bien définis. Le volet SEGUE (*Sloan Extension for Galactic Understanding and Exploration*) étudie les étoiles de la Voie lactée pour comprendre le scénario de sa formation. Les vieilles étoiles de notre galaxie sont observées avec un spectrographe, qui décompose les rayonnements électromagnétiques reçus de l'astre visé. L'étude spectroscopique révèle ainsi sa température, sa composition chimique, sa masse, son âge, sa vitesse.

Une image de 10^{15} pixels

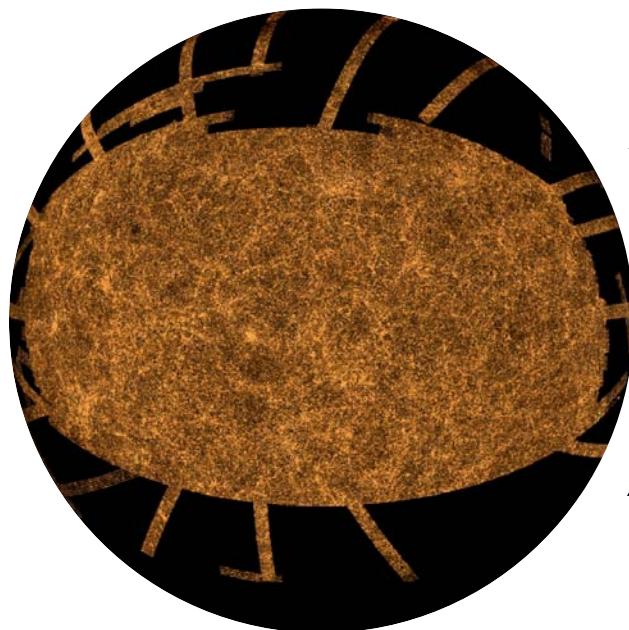
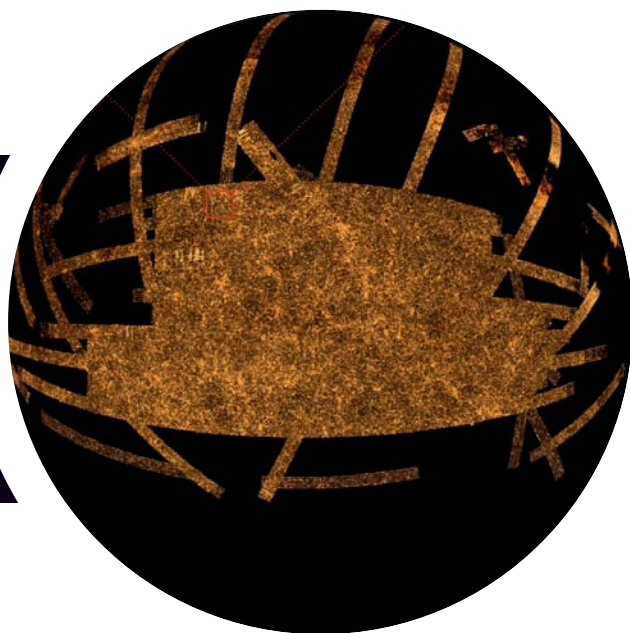
Ces informations individuelles, mises en lien d'une étoile à une autre, permettent de révéler en détail la structure spirale de la Voie lactée, de découvrir de nouvelles classes d'étoiles, d'étudier les déplacements des astres à l'intérieur de la Galaxie, etc. L'ensemble des renseignements ainsi obtenus aide à élaborer des scénarios expliquant la formation et l'évolution de la Galaxie.

Un deuxième volet de SDSS est le relevé BOSS (*Baryon Oscillation Spectroscopic Survey*), qui concerne l'Univers lointain. Son objectif est d'éclairer le mystère de l'« énergie sombre », entité de nature inconnue qui serait responsable de l'accélération de l'expansion de l'Univers, et d'aider à améliorer les modèles cosmologiques. Il utilise un

spectrographe capable de décomposer la lumière d'un millier d'astres à la fois. BOSS est chargé de mesurer les distances de plus d'un million de galaxies et quasars (noyaux de galaxies actives). Il dresse ainsi la plus grande carte tridimensionnelle des galaxies dans l'Univers. En analysant la répartition spatiale des galaxies lointaines et des quasars, on espère détecter les empreintes laissées par les ondes acoustiques (ondes de pression) qui se sont propagées dans l'Univers peu après le Big Bang.

Ces informations seront très utiles aux scientifiques qui tentent de comprendre l'évolution de l'Univers. La structure à grande échelle de l'Univers, la façon dont les galaxies sont regroupées, sont autant d'indices pouvant jouer en faveur ou en défaveur de telle ou telle théorie cosmologique.

Le relevé MARVELS (*Multi-object APO Radial Velocity Exoplanet Large-area Survey*), quant à lui, est voué à la détection d'exoplanètes. Pour ce faire, les astronomes de ce projet mesurent très précisément et de manière répétée le mouvement de va-et-vient d'environ 8 500 étoiles de type solaire et proches, mouvement oscillant dû à la présence d'un compagnon du type de Jupiter en orbite autour de l'étoile. On s'attend avec MARVELS à la découverte



d'une centaine de nouvelles exoplanètes géantes, ainsi qu'un nombre similaire de naines brunes, astres intermédiaires entre les plus petites étoiles et les planètes gazeuses.

Enfin, le quatrième volet de SDSS, l'expérience APOGEE (*Apache Point Observatory Galactic Evolution Experiment*), consiste à observer la Voie lactée dans l'infrarouge proche. Ce domaine du spectre électromagnétique permet de pénétrer à travers la poussière qui obscurcit la plupart des étoiles situées dans le disque et le bulbe galactiques.

APOGEE utilise un spectrographe à haute résolution très performant. Les astronomes mesurent grâce à lui la composition chimique et la vitesse (le long de la ligne de visée) de 100 000 étoiles géantes, ce qui fournit un aperçu sans précédent sur la structure dynamique et l'histoire chimique de la Voie lactée.

Ainsi, SDSS constitue un projet astronomique de grande ampleur, qui a déjà apporté de nombreuses découvertes. Cette immense observation du ciel, qui couvre 14 500 degrés carrés sur la voûte céleste, a répertorié 470 millions d'astres. Et parmi eux, près de deux millions de galaxies et 500 000 étoiles sont la cible d'études spectroscopiques.

La quantité d'informations produites par SDSS est énorme. Le volume total des données actuellement accessibles en ligne est de 71 téraoctets, ce qui équivaut à l'information contenue dans une bibliothèque comptant 71 millions de livres de 1 000 pages, soit bien plus qu'il n'y en a à la Bibliothèque nationale de France...

Un travail colossal est nécessaire pour traiter les données brutes provenant des instruments d'observation et en extraire les informations sur les paramètres physiques des objets astronomiques. Bien sûr, une grande partie de ce travail est effectuée par des ordinateurs. Pour BOSS, par exemple, le traitement est réalisé au Laboratoire américain Lawrence Berkeley par une grappe de 50 machines, chacune comportant huit cœurs (huit processeurs). Les capacités de stockage doivent être également à la hauteur du projet : les différents nœuds de stockage des données brutes et traitées totalisent une capacité de 250 téraoctets...

Il reste qu'une centaine de scientifiques doivent travailler des mois entiers pour préparer les données et les rendre publiques. En astronomie, on a pour tradition de rendre les données d'observation accessibles à tous. Celles du SDSS renferment sans aucun doute de nombreuses découvertes pour les années à venir. ■

CES DEUX IMAGES DE LA VOÛTE CÉLESTE

(au-dessous du plan de notre galaxie, à gauche ; au-dessus du plan galactique, à droite) ont été construites avec

les données du relevé astronomique SDSS. Chaque point lumineux est une source détectée. Les régions du ciel qui n'ont pas été observées sont en noir.

L'AUTEUR



Céline REYLÉ est astronome à l'Institut UTINAM, à Besançon, et directrice de l'Observatoire des sciences de l'Univers THETA Franche-Comté/Bourgogne. Elle participe au projet SDSS.

SUR LE WEB

Site du projet SDSS : www.sdss.org

Francis Daumas et Marion Massol

Au cœur d'un centre d'archivage

Pour stocker le déluge récent de données numériques, notamment scientifiques, de grands centres d'archivage ont été mis en place. Comment fonctionnent-ils ? Exemple avec le CINES, un centre national à Montpellier.

La plupart du temps, quand nous souhaitons conserver un document numérique, nous le copions sur un support externe, que nous stockons dans un placard. Mais sa récupération longtemps après peut se révéler compliquée. D'abord, on note rarement, ou de façon imprécise, la localisation du support et les informations pertinentes sur le contenu. Si l'on parvient à retrouver le support, encore faut-il pouvoir le lire : imaginez une photo de mariage stockée sur une vieille disquette des années 1980 ! Les équipements de lecture de ce support ont parfois disparu ou sont difficiles à trouver après tant d'années. Supposons que vous ayez trouvé le bon équipement : vous devez aussi disposer du logiciel capable de reconnaître le format du

fichier et de restituer le contenu sous une forme intelligible. Or les formats évoluent et les logiciels peuvent disparaître ou créer des fichiers qui ne seront pas lisibles par leurs futures versions. Enfin, le fichier initial ne doit pas comporter trop d'erreurs et son support doit avoir été bien conservé.

Cet exemple illustre les difficultés posées par la préservation à long terme des données : l'obsolescence et la détérioration du support, la disparition de l'outil logiciel de lecture, les évolutions du format du fichier et la perte de la localisation et de la signification de son contenu. Pour répondre à ces difficultés, des techniques d'archivage ont été développées. Nous les présenterons à travers l'exemple d'un des principaux sites d'archivage français : celui du CINES (Centre

informatique national de l'enseignement supérieur), à Montpellier, un établissement public placé sous la tutelle du ministère de l'Enseignement supérieur et de la Recherche.

Des données variées

Aujourd'hui, l'information numérique est omniprésente et protéiforme. Les données à conserver sur le long terme sont de natures variées : administratives, culturelles, scientifiques (issues d'observations ou de calculs), etc. Les livres des bibliothèques, les thèses, les articles scientifiques, les encyclopédies sont numérisés (*voir le dossier « Le savoir numérisé » dans ce numéro*). En résultent une multitude de documents à préserver, sous divers formats : textes, sons, images, vidéos...