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.