médecine/sciences 2012 ; 28 : 1027-8

Éditorial

Big Data en biologie

Alain Viari

« It is a very sad thing that nowadays there is so little useless information »
Oscar Wilde[1]

La publication récente des travaux du consortium ENCODE [1], ainsi que de nombreux éditoriaux récents [2, 3] ont remis en lumière le rôle et les questions posés par les grands volumes de données (*Big Data*) en biologie et en médecine.

Pourtant, la question n'est ni récente ni, loin de là, circonscrite aux sciences du vivant. L'émergence, à la fin des années 1990, des puces à ADN, la publication de la séquence du génome humain en 2001 avaient déjà, en leur temps, lancé le débat sur les questions liées à la gestion et l'analyse de ce « déluge de données » [4]. Débat qu'avivent, depuis le milieu des années 2000, les progrès considérables des dispositifs de séquençage dits de nouvelle génération (NGS), en termes de coût et de vitesse [5]. Si les premiers appareils commerciaux pouvaient produire des flux de quelques dizaines de Mo[2] par heure, les dispositifs plus récents atteignent désormais plusieurs Go (soit approximativement la taille d'un génome humain) par heure, et l'on attend de l'évolution des dispositifs en cours de développement des débits supérieurs de plusieurs ordre de grandeur.

Mais d'autres domaines scientifiques, en particulier la physique des hautes énergies, sont depuis longtemps confrontés à d'énormes volumes de données. Le flux de données produit par le *Large Hadron Collider* (LHC) du CERN (Centre européen de recherche nucléaire) est d'environ 1 To par heure (et encore ne s'agit-il là que de données pré-filtrées), soit 10 à 20 Po par an. Ces chiffres sont comparables à ceux que produisent aujourd'hui les réseaux sociaux : Facebook™ génère environ 4 Po de données par an et YouTube met chaque seconde 1 h de vidéo en ligne. Ces quelques exemples illustrent que le déluge de données n'est pas propre à la biologie mais bien un phénomène social global.

Il est donc légitime de se demander ce qui suscite l'intérêt, voire l'effroi, de la communauté biomédicale et justifie, par exemple, la création en 2012 du journal *GigaScience* [6], sous l'égide, en particulier, du *Beijing Genome Institute*, dédié à la question de la publication et de la dissémination d'études mettant en œuvre de grands volumes de données.

Il y a à cela au moins trois éléments de réponse.

Le premier est que la volumétrie, même si elle reste, en biologie et pour l'heure, en deçà de ce que l'on peut observer dans d'autres domaines, est tout de même importante et dépasse le seuil psychologique de ce qu'un chercheur peut gérer sur sa machine personnelle ou, tout simplement, faire transiter sur un réseau internet de capacité usuelle. Le second est lié à la très grande variété des données qu'il devient nécessaire d'intégrer pour produire un résultat biologique. Il s'agit d'informations moléculaires (génomique, protéomique, transcriptomique, structures 3D), cellulaires (images biologiques ou médicales) (microscopie, imagerie par résonance magnétique [IRM]), cliniques (dossier médicaux, cohortes), voire écologiques et environnementales. Ce point est plus important car les problèmes posés sont ici conceptuels et pas seulement technologiques. Enfin, le troisième élément est de nature méthodologique : les expériences *Big Data* ne se pensent et ne se gèrent pas comme les expériences traditionnelles. L'ensemble du flot de données, de l'acquisition jusqu'à la diffusion du résultat, en passant par le traitement, doit être planifié et calibré dès le départ. Ce flot engendre toujours des réductions successives et importantes des volumes de données (depuis les données brutes jusqu'au résultat biologique publié), chaque étape nécessitant des filtrages, des approximations, des « petits arrangements » avec l'incomplétude intrinsèque des informations, des contrôles statistiques. En résumé, ce flot doit être mis en œuvre par une chaîne de compétences alliant médecins, biologistes, bio-informaticiens, informaticiens et mathématiciens.

Considérons maintenant les niveaux technologiques et scientifiques concernés par ces trois éléments, afin d'identifier les enjeux et difficultés.

• Le premier niveau concerne les infrastructures matérielles (réseau, stockage, puissance de calcul). Sans vouloir en minimiser l'importance et les difficultés, c'est sans doute celui qui sera le plus facile à régler grâce à l'expérience déjà acquise par nos collègues d'autres disciplines. On pourra avantageusement s'inspirer par exemple des infrastructures hiérarchiques (de type

[1] *Saturday Review*, Novembre 1894 ; cité dans *The Economist*, février 2010, (http://www.economist.com/node/15557421).

[2] 1 bit : unité d'information (0 ou 1) ; 1 octet = 8 bits ; une base nucléique peut être codée sur 2 bits (puisqu'il y a quatre bases), mais les séquenceurs produisent également des mesures de qualité de la lecture, ce qui amène généralement à un ou deux octets par base. 1 Mo (méga-octet) = 10^6 octets ; 1 Go (giga-octet) = 10^9 octets ; 1 To (téra-octet) = 10^{12} octets ; 1 Po (péta-octet) = 10^{15} octets. Le disque dur d'un ordinateur de bureau est de l'ordre du To ; ceux d'un gros centre de calcul d'une dizaine de Po.

DOI : 10.1051/medsci/20122812001

n-tier[3]) mises en place au CERN, ou des infrastructures de type « nuage » (cloud) privées ou communautaires déjà existantes, en s'adaptant, le cas échéant, à certaines spécificités biomédicales comme le caractère décentralisé de la production (du laboratoire à l'hôpital), la nécessaire transparence d'accès et la confidentialité de certaines données. Mais les réseaux, les disques et les processeurs ne sont que le support physique de couches logicielles qui vont du système d'exploitation au système de gestion des données elles-mêmes.

• Le second niveau concerne donc la représentation et la gestion des données. Les principes importants à retenir ici sont que : (1) tout dispositif de gestion de données, quel qu'il soit, doit reposer sur un modèle (ou schéma) conceptuel de ces données, qui permet de les structurer et de les retrouver aisément et rapidement (c'est-à-dire effectuer efficacement une requête) ; et (2) tout modèle conceptuel de données dépend de l'usage que l'on souhaite faire de ces données. Cela implique qu'il n'existe pas et n'existera jamais de modèle universel pour un même type de données et, *a fortiori*, pour l'ensemble de toutes les données biologiques. Nous sommes condamnés à ce que plusieurs banques et bases de données coexistent. Le concept important pour gérer cette diversité des sources de données est celui d'interopérabilité, qui exprime la faculté qu'ont ces systèmes de gestion de données d'échanger leurs informations. Malheureusement, cette question est trop souvent réduite à sa dimension technique (logiciels différents) ou syntaxique (formats d'échange), alors que la question centrale est en réalité celle de l'interopérabilité sémantique. Comment s'assurer, par exemple, que la définition d'un gène dans la base de données A recouvre bien le même concept que celle de la base B (en particulier, lorsque ces deux bases n'ont pas été conçues pour le même usage) ? Pour une partie des données, cela se résout par la définition d'ontologies, mais également pour d'autres types d'informations (images, par exemple), le problème de la mise en correspondance reste très difficile. Si les remarques précédentes ne sont évidemment pas spécifiques au *Big Data*, les problèmes y sont néanmoins amplifiés, car les grands volumes manipulés font obstacle à un contrôle fin, à échelle humaine, de la cohérence des diverses informations. Enfin, une volumétrie importante entraîne également des questions algorithmiques spécifiques comme la définition de systèmes de compression et d'indexation dédiés, plus efficaces que les systèmes de gestion de bases de données usuels.

• Le troisième niveau, intimement lié au précédent, concerne l'analyse (incluant, entre autres, le filtrage, la fusion et la visualisation) de ces données. Là encore, des progrès algorithmiques seront nécessaires pour « passer à l'échelle ». Un point important est que ces progrès nécessiteront, au moins pour les premières étapes de filtrage, des simplifications dans les modèles biologiques sous-jacents et résulteront donc, à nouveau, d'un dialogue étroit entre expérimentalistes et modélisateurs.

Enfin, une dernière dimension qu'il convient de ne pas négliger concerne l'aspect culturel. Le phénomène *Big Data* entraîne un véritable changement de perspective du statut même de la donnée expérimentale et de la façon dont celle-ci est traitée. Dans un article récent, Cochrane *et al.* [7] s'interrogent, à juste titre, sur la nécessité d'archiver systématiquement toutes les séquences produites par les NGS. Dans le cas où le matériel biologique (ADN) analysé est reproductible ou peut être conservé en suffisamment grande quantité, ils proposent même de ne plus sauvegarder les séquences brutes, le coût de reséquençage devenant alors inférieur au coût de stockage des données informatiques. Bien entendu, il existe de nombreux cas où cette proposition est inacceptable (échantillon rare ou historique, patient unique), mais considérons-la comme un point de départ pour une réflexion plus générale sur le statut de la donnée expérimentale et de son traitement. Dans la situation actuelle, il n'est pas totalement caricatural de dire que beaucoup d'expérimentateurs commencent par se concentrer sur l'acquisition des données et ne se préoccupent que dans un second temps de les archiver, les structurer, les indexer et finalement les exploiter. Le résultat global serait sans doute meilleur en inversant ce processus : c'est-à-dire en s'interrogeant d'abord sur les informations finales recherchées, pour en déduire une structuration appropriée des données et, finalement, de dimensionner le dispositif d'acquisition et de calcul adapté (ce qui limiterait en particulier la quantité de données initiales produites peu utiles pour l'exploitation finale). Ce changement de perspective va plus loin qu'un simple dialogue entre biologistes et informaticiens, il implique une réelle co-construction des expériences et il ne s'agit pas là seulement du prix à payer pour endiguer le déluge de données mais d'une réelle opportunité scientifique en direction d'une biologie numérique. ◊

***Big Data* in biology**

[3] Un *tier* est une couche ou un niveau (*layer* en anglais). Une architecture n *tier* est une architecture matérielle hiérarchisée en *n tiers*.

LIENS D'INTÉRÊT

L'auteur déclare n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

Directeur scientifique adjoint à la direction de la recherche de l'Inria
En charge des STIC pour les sciences de la vie,
la santé et l'environnement
Inria Grenoble-Rhône-Alpes
655, avenue de l'Europe
38330 Montbonnot-Saint Martin, France.
alain.viari@inria.fr

RÉFÉRENCES

1. Skipper M, Dhand R, Campbell P. Presenting *ENCODE*. *Nature* 2012 ; 489 : 45. Voir les 5 articles et news and views publiés dans ce numéro.
2. Sagoff M. *Issues in Science and Technology*, summer 2012 ; http://www.issues.org/28.4/sagoff.html
3. The data deluge (editorial). *Nat Cell Biol* 2012 ; 14 : 775
4. Gershon D. Dealing with the data deluge. *Nature* 2002 ; 416 : 889-91.
5. Kircher M, Kelso J. High-throughput DNA sequencing: concepts and limitations. *Bioessays* 2010 ; 32 : 524-36.
6. http://www.gigasciencejournal.com
7. Cochrane G, Cook CE, Birney E. The future of DNA sequence archiving. *GigaScience* 2012 ; 1 : 2.

LECTURES RECOMMANDÉES

• Enfin, je recommande vivement la lecture du numéro consacré au *Big Data* de la revue *Ercim-News*, téléchargeable en pdf à l'url suivante : http://ercim-news.ercim.eu/en89.

TIRÉS À PART
A. Viari