Le Big Data dans la recherche médicale

Jean-Philippe VERT

Directeur du Centre de Bio-informatique
de MINES ParisTech

Des chercheurs en sciences du vivant aux professionnels de santé, les acteurs de la recherche et de la pratique biomédicale ont de tout temps produit et conservé, souvent de manière écrite dans des dossiers médicaux ou des publications scientifiques, de la connaissance et des informations. La digitalisation rapide de ces contenus et l'explosion des volumes de données produites par de nouvelles technologies, à l'ère de la génomique à haut débit et des objets connectés, révolutionnent la recherche et la pratique biomédicale depuis quelques années. L'exploitation de «big data» biomédicales peut-elle permettre d'améliorer la rentabilité des industries de santé, de rationaliser la prise en charge des patients, d'anticiper les maladies pour mieux les traiter, voire de vaincre certaines pathologies comme le cancer ? Si les écueils scientifiques, techniques, réglementaires et éthiques ne manquent pas, les espoirs sont permis et sous-tendent d'ors et déjà de nombreuses initiatives scientifiques et industrielles.

Les «big data» de la santé

Les données utiles à la recherche biomédicale, et plus généralement à la santé, ont de nombreuses origines qu'on peut tenter de regrouper en trois grandes catégories. Premièrement, on assiste à la digitalisation massive et systématique de connaissances et d'informations traditionnellement conservées sous formes écrite, des publications scientifiques aux dossiers médicaux en passant par les résultats des 225 000 essais cliniques qui ont lieu en permanence dans le monde. Il est par exemple aujourd'hui possible d'accéder sous forme digitale à la majorité des 1,8 millions d'articles scientifiques publiés chaque année dans les 5 600 journaux indexés par la principale base de données du domaine, MEDLINE, ainsi qu'aux 24 millions d'articles plus anciens qui y sont archivés. Deuxièmement, les progrès technologiques dans les domaines de la génomique, de la protéomique ou de l'imagerie au cours des 20 dernières années ont rendu possible l'acquisition d'instruments dans les laboratoires de recherche et les hôpitaux générant d'énormes quantités de données pour décrire les caractéristiques moléculaires ou cellulaire d'échantillons biologiques. Le séquençage de l'ADN, par exemple, permet de lire le patrimoine génétique de chaque cellule codé dans les six milliards de nucléotides constituant son ADN. Depuis le séquençage du premier génome humain en 2001, le coût du séquençage a été divisé par plus de 100 000, et il est aujourd'hui possible de séquencer un individu en quelques jours pour environ 1 000 dollars (Figure 1); le séquençage est ainsi en passe de devenir un examen de routine, générant environ 100 Go de données par échantillon. Troisièmement, enfin, les informations transmises par les individus sur les réseaux sociaux, leur messagerie électronique, ou via des objets connectés comme les bracelets mesurant leur activité physique ou les lentilles de contacts mesurant la concentration de glucose dans leurs larmes, vont rapidement constituer un volume de données encore plus important que les données génomiques ou strictement médicale, utilisables pour améliorer le suivi préventif ou médical des individus. Les données biomédicale ont donc tous les aspects caractéristiques du «Big Data», notamment un très grand volume de données en croissance exponentielle (un rapport du cabinet IDC estimait le volume total des données de santé à 153 exaoctets en 2013, soit 153 milliards de gigaoctets, avec un doublement des volumes tous les

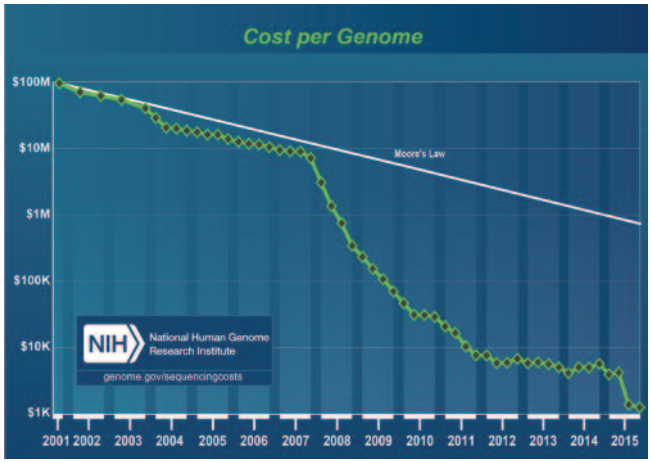

Figure 1 : Le coût pour séquencer un génome humain a été divisé par 100 000 en 15 ans, et est proche d'atteindre la barre symbolique des 1 000 dollars. Les avancées dans ce domaine sont beaucoup plus rapides que les progrès de la capacité et de la puissance de calcul des ordinateurs, décrit par la loi de Moore (source : NHGRI).

Jean-Philippe VERT

Jean-Philippe Vert est Professeur à l'École Normale Supérieure de Paris et Directeur de Recherche à MINES ParisTech, où il dirige le Centre de Bio-informatique commun avec l'Institut Curie. X-Mines et docteur en mathématiques, il s'intéresse à l'apprentissage statistique et à ses applications dans les sciences du vivant, notamment en génomique dans le cadre de la recherche contre le cancer.

deux ans) avec des contenus allant de textes en langage naturel à des images ou des séquences d'ADN, et des données de plus en plus générées par des flux continus pour des décisions devant être prises en temps réel.

Le nouveau paradigme de recherche «data-driven»

Dès les années 1990, la croissance du volume des données digitales en biologie a permis l'essor d'une nouvelle discipline, la bio-informatique, afin de développer les bases de données et les algorithmes nécessaires pour stocker et analyser ces données. Aujourd'hui, l'informatique fait partie des outils indispensables à tout chercheur en biologie, au minimum pour effectuer ses recherches bibliographiques et étudier la fonction d'un gène ou un mécanisme biologique en accédant à des bases de données spécialisées, voire de plus en plus fréquemment en utilisant des programmes dédiés pour analyser et interpréter les données qu'il ou elle aura produites dans sa démarche expérimentale. La possibilité de générer facilement et à faible coût de grandes quantités de données, comme par exemple de mesurer quantitativement et simultanément l'expression des 22 000 gènes humains dans des échantillons soumis à différentes conditions expérimentales, a d'ailleurs profondément modifié la démarche scientifique elle-même. Alors que le paradigme traditionnel de la science consiste à formuler une hypothèse puis à imaginer une expérience permettant de la réfuter ou de la vérifier, on a assisté au cours de la dernière décennie à l'émergence d'un paradigme dit «data-driven» où c'est l'analyse de larges volumes données expérimentales qui permet de formuler des hypothèses, vérifiables ensuite par de nouvelles expériences plus ciblées. Ce changement de paradigme s'est notamment traduit, depuis une quinzaine d'année, par l'apparition de nombreux projets de recherche dont le point de départ est la génération de grands volumes de données, par exemple le séquençage systématique de cohortes d'individus afin d'identifier des corrélations entre des variations génétiques et des phénotypes divers, comme le risque de développer une maladie ou la probabilité de répondre à un traitement. De la «Precision Medicine Initiative» lancée par Barack Obama en 2015, visant à collecter des données (dont le génome) d'un million d'Américains, au plan «France Médecine Génomique» remis au premier ministre Manuel Valls en 2016 et visant à développer une filière médicale et industrielle en France de médecine génomique, se basant notamment sur le séquençage de plusieurs centaines de milliers de patients par an à l'horizon 2020, on assiste à des investissements massifs visant à générer toujours plus de données. Ce nouveau paradigme donne la part belle aux méthodes de fouille de données et d'apprentissage automatique afin d'identifier des corrélations significatives au seins des données biologiques. C'est ainsi que IBM collabore étroitement avec le centre contre le cancer MD Anderson de l'Université du Texas aux États-Unis, qui traite annuellement plus de 100 000 patients, afin d'utiliser les capacités cognitives de son logiciel d'intelligence artificiel Watson pour aider les médecins à faire les meilleurs choix thérapeutiques en analysant toutes les données disponibles sur chaque patient et dans les bases de données scientifiques.

Des obstacles à surmonter

Les espoirs portés par les big data en recherche biomédicale sont donc immenses, et des investissements massifs sont en cours pour générer, collecter et analyser toujours plus de données. Les résultats concrets de cette révolution en cours tardent cependant à se manifester ; après plus de dix ans de révolution génomique, la productivité de l'industrie pharmaceutique n'a guère évolué, et l'on est loin d'avoir vaincu le cancer. Les obstacles à surmonter restent nombreux. Scientifiquement, la modélisation du vivant et la capacité prédictive des modèles mathématique mis en œuvre sont des sujets de recherche largement ouvert. Techniquement, la gestion et l'exploitation des grandes masses de données reste un défi, similaire à d'autres domaines d'applications du big data. De nombreuses questions légales se posent également lorsque l'on traite des données médicales, notamment concernant la propriété des données et la réglementation stricte entourant les autorisations de mise sur le marché. Enfin des questions éthiques se posent rapidement, concernant par exemple les tests prénataux ou l'utilisation de données personnelles. La révolution du big data dans le domaine de la médecine n'est donc pas qu'une question technique, mais nécessite l'implication et la collaboration de nombreux acteurs. ■