2

Quand les algorithmes font parler l'ADN

Jean-Philippe Vert, École normale supérieure, Mines ParisTech et Institut Curie, Paris

« Cracker » le code de notre génome, comprendre les facteurs externes qui régulent l'expression des gènes, déterminer les mécanismes responsables du développement de maladies génétiques : ce travail délicat et laborieux pourrait bientôt s'automatiser, grâce à de nouvelles techniques d'intelligence artificielle. De quoi faire entrer la génomique dans une nouvelle ère.

Imaginez un texte de 6 milliards de lettres, soit 600 fois plus que le nombre de caractères d'*À la recherche du temps perdu,* de Marcel Proust. Supposez maintenant que ce texte utilise un alphabet de quatre lettres (A, T, C, G) au lieu des 26 de notre alphabet latin. Ce texte, a priori indéchiffrable, c'est notre génome. Propre à chaque individu, ce dernier encode un message essentiel au bon fonctionnement de nos cellules. Certaines séquences de ce code peuvent aussi être délétères et provoquer l'apparition de maladies. La compréhension de ce texte constitue donc un Graal pour la biologie, et en particulier pour la génomique, discipline visant à comprendre la structure, la fonction, et l'évolution des génomes. Il aura fallu un demi-siècle de découvertes scientifiques et de prouesses technologiques pour réaliser le premier séquençage du génome humain, au début des années 2000. Cette entreprise, qualifiée parfois de « projet Apollo de la biologie », a ouvert la voie à l'analyse de ce texte immense.

Depuis, la technologie a progressé à une allure fulgurante, si bien que le séquençage d'un ADN humain (ou non humain) est presque devenu une opération de routine, réalisable en quelques heures pour un coût raisonnable. En parallèle, d'autres types de technologies ont été développés. D'une part, pour analyser l'épigénome, c'est-à-dire l'ensemble des modifications moléculaires qui agissent sur le fonctionnement de l'ADN sans en altérer le code. D'autre part, pour étudier le transcriptome, à savoir l'ensemble des petites molécules (ARN) produites à suite de la transcription du génome et qui jouent un rôle crucial dans la production de protéines et le fonctionnement de la cellule. Mises en commun, toutes ces données forment ce que l'on appelle un portrait moléculaire.

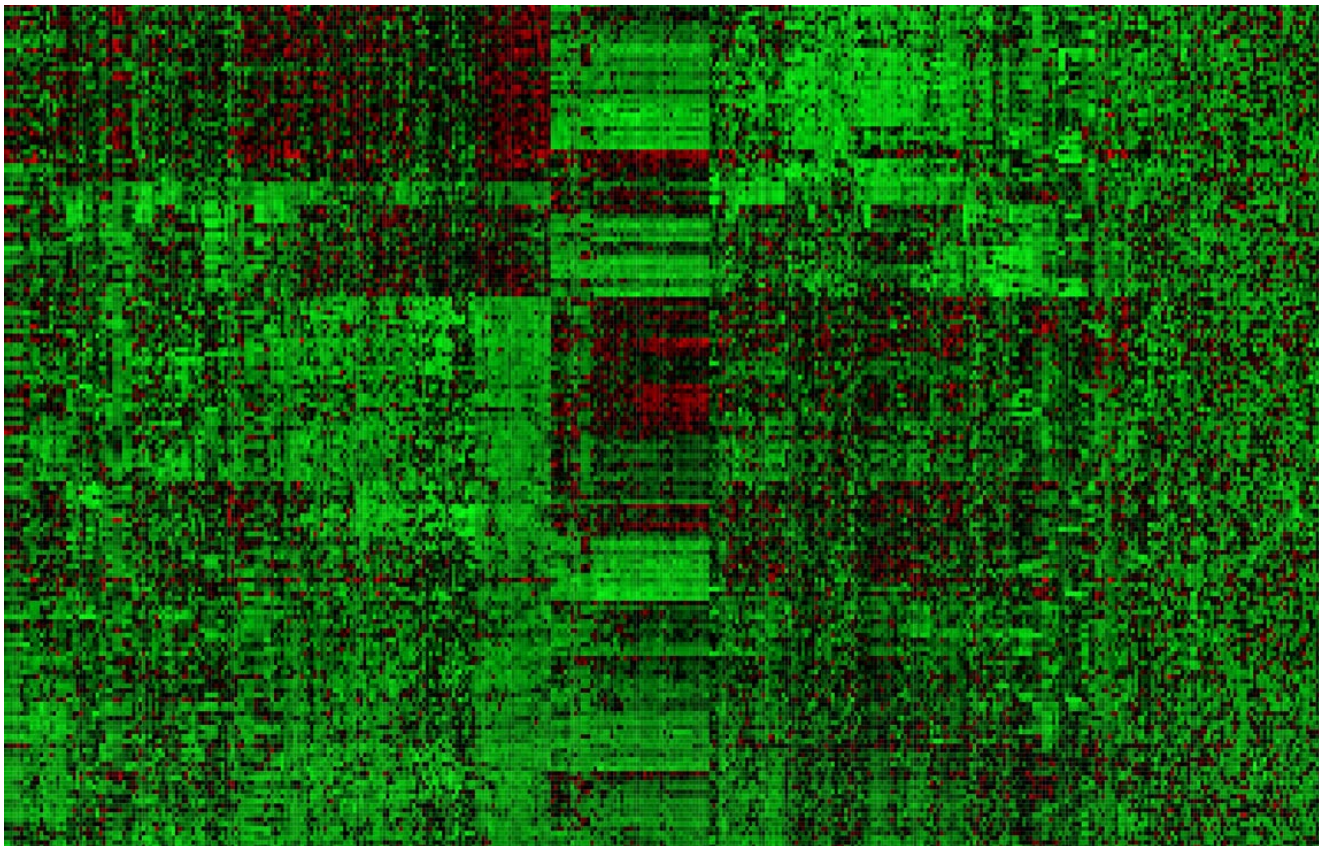
Comment l'analyser, donner du sens aux grandes quantités de données produites par ces technologies dites à haut débit ? Grâce à des techniques d'intelligence artificielle ! Et en particulier des algorithmes d'apprentissage statistique. Ces derniers « apprennent » et s'améliorent grâce à l'afflux de données. Ils parviennent ainsi à résoudre des tâches complexes, comme

DR

BIO-INFORMATICIEN
Jean-Philippe Vert est professeur au département de mathématiques et applications de l'École normale supérieure, directeur de recherche à Mines ParisTech où il dirige le centre de bio-informatique, et chef d'une équipe travaillant sur la modélisation du cancer à l'Institut Curie.

Contexte

En 1953, Rosalind Franklin, Francis Crick, James Watson et Maurice Wilkins découvrent la structure en double hélice de l'ADN, support de l'information génétique. Certains gènes ont pu être caractérisés, mais une grande partie de notre code génétique n'a pas pu être déchiffrée. L'explosion de la puissance de calcul des ordinateurs, le séquençage de l'ADN à haut débit et l'amélioration des techniques d'intelligence artificielle changent la donne.



▲ *Sur cette image, chaque ligne correspond à une tumeur du sein. Chaque colonne représente un gène plus ou moins exprimé dans cette tumeur : en vert, il l'est beaucoup ; en rouge, peu. Des algorithmes de classification non supervisée font apparaître des groupes de tumeurs aux caractéristiques génétiques proches et les classent en cinq catégories, ce qui aide les médecins à choisir un traitement adapté.*

JEAN-PHILIPPE VERT

l'annotation des données génomiques. Ce travail délicat consiste à repérer des éléments fonctionnels dans le génome : des gènes ou des séquences régulatrices de ces gènes qui remplissent une certaine fonction biologique.

Imaginez ouvrir l'un des chapitres du génome humain : une longue suite de lettres A, T, C, G, sans structure manifeste, apparaît devant vous. Comment décrypter ce langage et comprendre le message codé dans le texte ? Comment identifier les régions codant les gènes et leurs structures fines, repérer les positions de l'ADN sur lesquelles se fixent les protéines qui réguleront l'expression de ces gènes ? En suivant la démarche du biologiste, vous commenceriez sans doute par chercher des répétitions, des régularités à différentes échelles pour, peu à peu, identifier des structures cachées, inférer une sorte de grammaire.

Comparer les génomes

La force des algorithmes d'apprentissage statistique est de reproduire cette démarche de façon à traiter automatiquement les 6 milliards de lettres du génome. Une classe d'algorithmes appelés modèles graphiques est particulièrement efficace pour cela. Ils permettent en effet aux chercheurs d'inclure leurs connaissances dans une modélisation probabiliste des données, puis d'inférer des informations pertinentes en laissant l'algorithme optimiser par lui-même les paramètres du modèle sur les données réelles. Dans le cas de l'annotation de l'ADN, on utilise des modèles graphiques particuliers, baptisés chaînes de Markov cachées. Ces dernières permettent d'inférer automatiquement l'annotation du génome à partir de régularités découvertes par le modèle dans la séquence d'ADN. Ces modèles rentrent dans la catégorie des méthodes d'apprentissage dites non supervisées, car elles apprennent à annoter le génome sans qu'on leur fournisse d'informations explicites sur certaines parties du génome dont l'annotation est déjà connue. Ces modèles graphiques offrent une grande flexibilité et s'adaptent à différentes situations. Par exemple, une autre application de ces méthodes consiste à extraire des informations épigénétiques, c'est-à-dire relatives à des modifications moléculaires autour de •••

6 MILLIARDS

C'EST LE NOMBRE DE NUCLÉOTIDES que contient notre ADN. Ces nucléotides sont l'adénine (A), la thymine (T), la cytosine (C) et la guanine (G). L'ensemble forme notre code génétique.

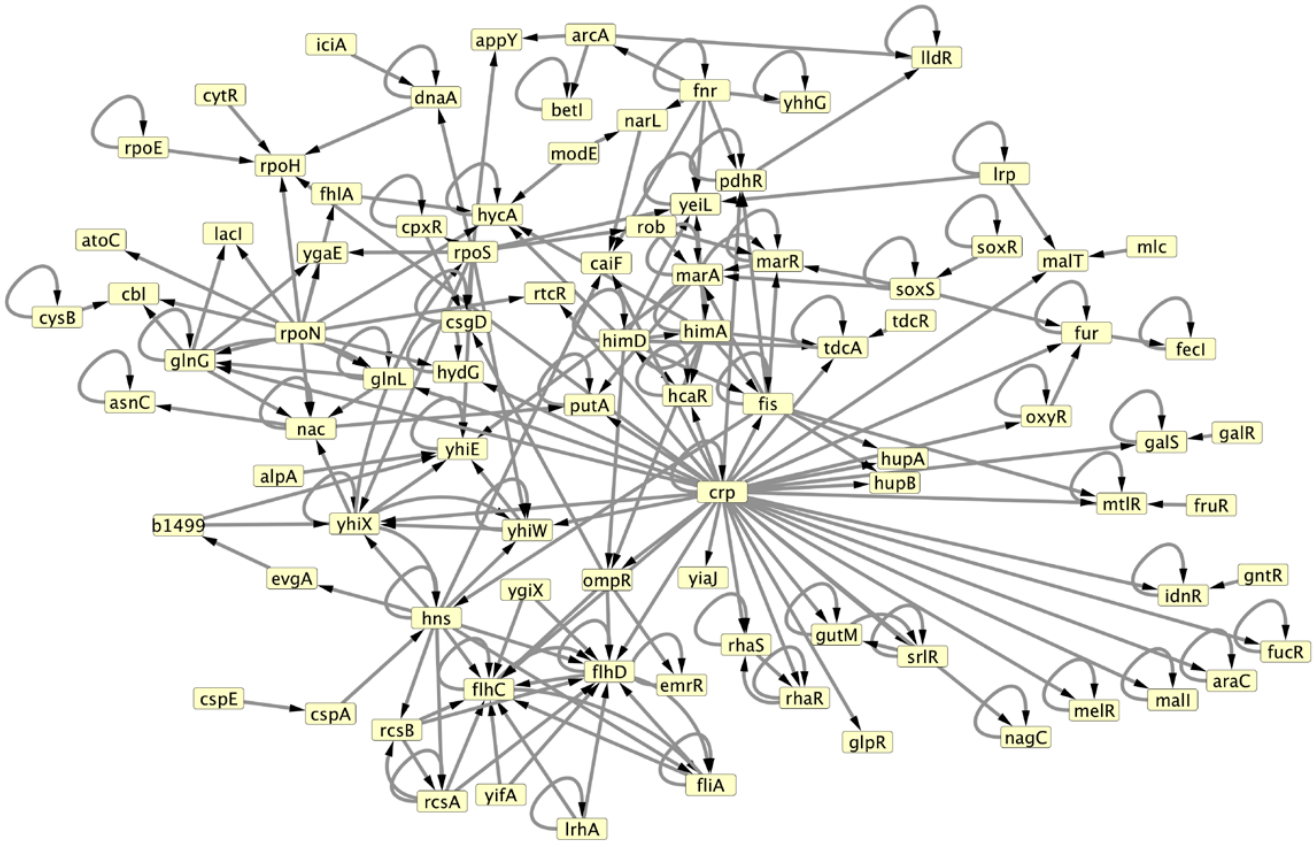
••• l'ADN. C'est ce qui a été fait dans le cadre du projet international Encode en 2012, visant à établir une annotation précise des parties fonctionnelles du génome humain à partir de portraits moléculaires mesurés dans différents types cellulaires (1).

Toutefois, le meilleur moyen de faire parler l'ADN est de comparer les génomes. En filant la métaphore littéraire, l'analyse d'un livre suffisamment long peut permettre de décrypter en partie les secrets d'un langage, en s'appuyant sur les répétitions de mots ou de structures grammaticales au sein du texte. Mais ce n'est qu'en comparant plusieurs livres que l'on peut voir émerger du sens. En effet, c'est en regroupant les mots par sujet lorsqu'ils apparaissent fréquemment ensemble que l'on voit apparaître des similarités entre certains livres en fonction de leur contenu ou de leur auteur. De la même manière, la génomique dite comparative, qui analyse les génomes en les comparant, est l'une des approches les plus puissantes pour acquérir de la connaissance à partir de données génomiques.

Traitement personnalisé

Historiquement, la génomique comparative s'est d'abord intéressée à la comparaison d'espèces, ce qui a permis de reconstruire l'arbre de la vie proposé par Darwin et d'identifier les gènes dont les fonctions sont spécifiquement associées à une famille d'espèces. Les modèles graphiques utilisés pour identifier la structure d'un génome unique peuvent d'ailleurs être étendus au traitement simultané de plusieurs génomes. Plutôt que de comparer les génomes entre plusieurs espèces, comme l'homme et la souris, on peut aussi comparer des portraits moléculaires de différents individus au sein d'une même espèce. Par cette approche, on peut établir des corrélations entre des variations observées dans un portrait moléculaire et des propriétés comme le rendement d'une plante ou le risque de développer une maladie.

Pour ce faire, la génomique comparative s'appuie essentiellement sur des modèles statistiques et des algorithmes d'apprentissage non supervisés. Le but ? Capturer les similarités et les variations entre données génomiques. Des techniques de réduction de dimension ou de classification non supervisées permettent, par exemple,

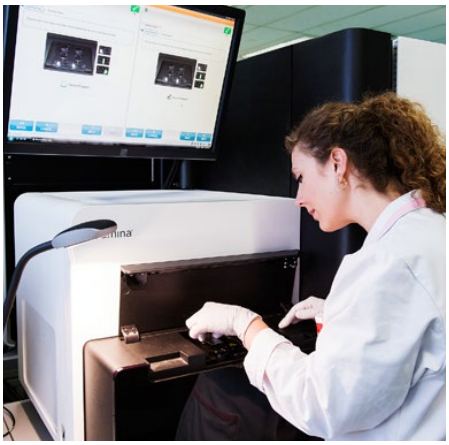


▲ *Ce réseau représente les interactions (flèches) entre des protéines dans une cellule humaine. Il a été déduit de l'analyse d'ARN messagers par un algorithme d'apprentissage. Cette approche pourrait permettre d'identifier de nouvelles cibles thérapeutiques pour certaines maladies.*

JEAN-PHILIPPE VERT

UN GRAND PROJET FRANÇAIS DE DÉCRYPTAGE DU GÉNOME

À l'instar des États-Unis, du Royaume-Uni ou de la Chine, la France a lancé en 2016 le projet France médecine génomique 2025, ambitieux programme visant à développer l'utilisation de la génomique dans les parcours de soins. Pour cela, des plateformes de séquençage seront créées sur le territoire pour, d'ici à 2020, être en mesure de séquencer 235 000 génomes par an.
Les énormes quantités de données récoltées seront traitées dans des centres de calculs et permettront d'améliorer la connaissance des maladies traitées, en particulier les maladies rares et les cancers. Cette approche aidera à personnaliser la prise en charge des patients.



▲ *Un séquenceur haut débit de l'unité de génomique métabolique du Genoscope, à Évry.*

CYRIL FRESILLON/CNRS PHOTOTHEQUE

d'identifier des sous-groupes homogènes au sein d'une population hétérogène. Ces techniques se sont invitées dans la recherche contre le cancer dès le début des années 2000, lorsqu'il a été possible d'analyser des transcriptomes complets de plusieurs centaines de tumeurs. Elles ont permis de réaliser des comparaisons qui ont révélé la grande hétérogénéité moléculaire de certains types de tumeurs. Les cancers du sein ont ainsi été divisés en cinq grandes classes en fonction de leur profil moléculaire. Selon ces classes, le pronostic et le traitement pressentis sont différents (2).

Aujourd'hui, cette classification va encore plus loin. En effet, nous sommes capables de séquencer des échantillons différents au sein d'une même tumeur, voire de séquencer des cellules uniques. Cela permet de mettre en lumière l'hétérogénéité moléculaire au sein même de la tumeur d'un patient. Grâce à des outils d'apprentissage non supervisés, comme des modèles graphiques ou des techniques de factorisation de matrice, on reconstruit ainsi l'histoire moléculaire de la tumeur à partir de ces données, et on identifie automatiquement les processus impliqués dans son apparition et sa progression. On peut par exemple déterminer si un cancer est apparu à la suite d'une exposition au soleil ou au tabac, en analysant des mutations observées dans l'ADN d'une tumeur. De façon étonnante, les techniques de factorisation de matrices utilisées pour réaliser ce genre d'expertise sont similaires à celles qui sont utilisées par les plateformes de vidéos à la demande comme Netflix pour personnaliser leurs recommandations. En génomique, ces informations précieuses peuvent aider les médecins à mieux caractériser la maladie pour un patient donné, et donc de lui apporter un traitement personnalisé. Outre ces informations d'ordre médical, certains algorithmes d'apprentissage statistique permettent d'inférer des connaissances plus fondamentales.

On parvient à identifier automatiquement les processus impliqués dans l'apparition de la tumeur

Comme toute science, la biologie accumule des connaissances en confrontant des hypothèses avec des observations. Historiquement, les hypothèses étaient formulées par les scientifiques à partir de leur intuition, et des expériences étaient réalisées pour les valider ou les invalider. La génomique, en produisant de grandes quantités de données, a quelque peu inversé ce paradigme de recherche : il est maintenant courant de commencer par générer beaucoup de données, par exemple de séquencer des centaines de génomes, puis de les analyser par des méthodes automatiques fondées sur les statistiques et l'intelligence artificielle. On fait ainsi émerger des hypothèses à partir des données.

Gènes exprimés ou non

Bien sûr, ces hypothèses doivent ensuite être validées grâce à d'autres expériences ciblées. Prenons l'exemple de la régulation de l'expression des gènes. Depuis les travaux de François Jacob, Jacques Monod et André Lwoff, qui leur valurent le prix Nobel de médecine en 1965, nous savons que chacun des 20 000 gènes codés dans notre ADN peut être exprimé ou pas – c'est-à-dire copié sous forme d'ARN messager afin de produire une protéine – en fonction de la présence ou non d'autres protéines, appelées facteurs de transcription. Ces derniers, en se fixant sur le brin d'ADN, commandent l'expression du gène cible. Mais comment identifier, pour chaque gène cible, les •••

••• facteurs de transcription qui le régulent et l'ensemble des facteurs qui influent sur l'expression des gènes ? Une solution consiste à collecter des données de transcriptomes de plusieurs centaines d'échantillons soumis à diverses conditions expérimentales, et à les comparer. Si l'on observe qu'un gène cible A est systématiquement exprimé dans les conditions expérimentales où un facteur de transcription B est également exprimé, on peut supposer que le facteur B régule A. Mais lorsque l'on a plusieurs gènes cibles et plusieurs facteurs de transcription à considérer en même temps, la situation est plus compliquée. Et c'est là que les algorithmes se révèlent très utiles.

Risque de récidive

Les réseaux bayésiens, en particulier, offrent un cadre statistique rigoureux pour inférer des interactions entre plusieurs gènes et préciser les rapports qu'entretient tel gène avec tel facteur de transcription. Les réseaux bayésiens sont des modèles graphiques particuliers qui combinent théorie des graphes (*) et statistique pour inférer des relations de causalité, comme le fait que l'expression d'un gène est régulée par un autre gène.

Depuis quelques années, d'autres méthodes fondées sur les forêts aléatoires ou la régression lasso, deux techniques populaires d'apprentissage statistique, ont aussi démontré leur intérêt pour cette tâche : elles ont obtenu les meilleures performances lors d'une compétition internationale visant à reconstruire aussi précisément que possible le réseau de régulation d'organismes bactériens et de levure (3). Cela ouvre la voie à de nombreuses applications en biotechnologie et en médecine comme l'identification de nouvelles cibles thérapeutiques.

Outre la compréhension, l'intelligence artificielle excelle dans la prédiction

Outre la compréhension de ces interactions, l'intelligence artificielle excelle dans l'art de la prédiction. Prédire le rendement d'une plante à partir de son ADN ; évaluer le risque de récidive d'un cancer, et adapter le traitement en conséquence à partir de l'expression des gènes et des mutations dans l'ADN d'une biopsie ; prédire l'efficacité d'un traitement à partir du portrait moléculaire d'un cancer…

Ces multiples tâches prédictives sont aujourd'hui essentiellement remplies par des méthodes d'apprentissage statistique supervisé. Si l'on prend l'exemple de l'évaluation des risques de récidive d'un cancer, cette approche consiste à collecter des portraits moléculaires de la tumeur sur des groupes de patients au moment du diagnostic initial, puis de suivre ces patients pendant plusieurs années. On associe un label « récidive » aux portraits moléculaires des patients victimes d'un nouveau cancer avant cinq ans, et un label « non-récidive » aux autres. Puis, à partir de ces données dites « étiquetées », on entraîne un algorithme d'apprentissage à prédire la catégorie de la tumeur (récidive et non récidive) en fonction du portrait moléculaire réalisé au moment du premier diagnostic. Dans la pratique, on combine ces données génomiques avec d'autres informations dont on dispose sur la maladie, comme la taille de la tumeur ou l'âge du patient, qui peuvent influencer le risque de récidive. Cette tâche de classification supervisée se caractérise souvent par le fait qu'on dispose pour chaque patient d'un grand nombre de données moléculaires (le niveau d'expression de 20 000 gènes, les mutations à des millions de positions dans l'ADN, etc.). En revanche, le nombre de patients inclus dans de telles expériences est souvent limité à quelques centaines.

Ce déséquilibre entre le nombre ahurissant de données par individu et celui plus modeste d'individus, est une limite problématique pour l'efficacité des algorithmes d'apprentissage. Pour pallier ce que les statisticiens appellent « la malédiction des grandes dimensions », des projets visent à collecter des données sur de grandes cohortes d'individus (lire ci-contre). En parallèle, la recherche en mathématique et en informatique pour améliorer les techniques d'apprentissage statistique en grande dimension est en pleine ébullition ! ■

(*) **La théorie des graphes** est une discipline visant à étudier des modèles abstraits de réseaux constitués de nœuds reliés entre eux par des liens. L'étude des interactions entre les gènes peut être formalisée comme un problème de graphe.

DRESSER LE PORTRAIT DES CANCERS

L'union fait la force. Le dicton sonne particulièrement vrai pour l'International Cancer Genome Consortium, réunissant 88 équipes de recherche à travers le monde. L'objectif de cette organisation ? Étudier en profondeur le génome de 25 000 tumeurs et l'ensemble des facteurs qui régulent l'expression de leurs gènes. Tous ces portraits moléculaires de cancers, obtenus entre autres grâce aux techniques d'intelligence artificielle, permettront de mieux comprendre les mécanismes par lesquels les différentes catégories de tumeurs se développent. Et donc de mieux les soigner à l'avenir.

(1) M. Hoffman *et al., Nature Methods, 18,* 473, 2012.
(2) C. Perou *et al., Nature, 406,* 747, 2001.
(3) D. Marbach *et al., Nature Methods, 9,* 796, 2012.

DAVID CARILLET/SHUTTERSTOCK