BIO-INFORMATIQUE, LE RÊVE DE DARWIN EN GRAND FORMAT

En quelques années, le mariage de la génomique et de l'informatique a révolutionné l'étude du vivant grâce à des technologies capables de reconstruire l'évolution biologique, de la molécule à l'espèce. Explications d'**Olivier Gascuel,** pionnier de ces recherches.

BIO-INFORMATIQUE

Dans le cadre de notre partenariat avec l'Académie des sciences, les académiciens nouvellement élus fin 2019 présentent un éclairage sur leur discipline et ses enjeux scientifiques, éthiques, politiques et sociétaux, à travers leur expérience personnelle.

Charles Darwin au XIXe siècle avait fait le pari d'établir entre les espèces du monde vivant des liens de filiation, conjecturant qu'elles avaient toutes évolué à partir d'un unique ancêtre commun. Les similitudes qu'il avait observées ou, au contraire, les traits qui les différenciaient constituaient autant d'indices lui permettant d'établir leur degré de parenté.

À l'époque, le mot bio-informatique n'existait pas. Personne ne soupçonnait à quel point il allait révolutionner toute la recherche en biologie. Personne, vraiment, pas même en 1983 lorsque je suis entré au CNRS. Après un parcours en mathématiques pures, puis intelligence artificielle (IA), je me suis intéressé à la biologie moléculaire. Désirant me tourner vers des aspects plus appliqués et « utiles », je travaillais sur les séquences d'acides aminés constituant les protéines. J'utilisais des méthodes informatiques et statistiques issues de l'IA pour faire parler ces enchaînements, qui définissent la structure 3D des protéines et leur fonction au sein de la cellule et de l'organisme. De la même manière, l'ADN, qui constitue le code informatique du vivant, commençait à faire l'objet d'études systématiques. Il s'agissait de décrypter ce langage pour les virus, les bactéries, les plantes, les animaux et bien sûr pour l'homme. Mais les données étaient rares. On ne disposait alors que d'une centaine de protéines de structure connue. Aujourd'hui on en compte plus de 100 000.

C'est dans ce contexte qu'a été lancé en 1989 l'ambitieux programme international de séquençage du génome humain. Nous étions alors peu nombreux à avoir travaillé sur ces données. Mais la course mondiale au génome était ouverte, et la France avait sa carte à jouer, pour le séquençage mais surtout pour l'analyse. L'entreprise était prévue pour durer trente ans, et devait être « plus difficile que d'aller sur la Lune ». Il n'a finalement fallu que douze ans pour que soit publié le premier génome humain, en 2001. Une prouesse technique, due à l'accélération des techniques de séquençage, mais aussi au développement d'algorithmes très sophistiqués pour assembler les millions de courtes séquences d'ADN fournies par les séquenceurs en une unique séquence de plus de 3 milliards de caractères, constituant le génome humain. Un gigantesque puzzle, qui a requis des recherches bio-informatiques très avancées.

2 500 GÉNOMES HUMAINS SÉQUENCÉS EN DEUX ANS !

En 2008, nouveau défi : le projet « 1 000 génomes humains » est annoncé. 1 000 génomes, alors qu'il a fallu douze ans pour un seul ? En deux ans, ce seront finalement 2 500 génomes qui seront séquencés. Aujourd'hui, faire séquencer son propre génome coûte 500 euros, et prend quelques jours…

Ainsi, en l'espace de quelques années, la science a, avec ces moyens technologiques, réinventé la manière d'étudier le vivant. On a commencé à séquencer à large échelle virus, bactéries, espèces animales, vé-

INSTITUT PASTEUR, FRANCOIS GARDY

PROFIL

Directeur de recherche au CNRS et à l'Institut Pasteur, où il dirige le département de biologie computationnelle, membre de l'Académie des sciences, **Olivier Gascuel** est bio-informaticien. Il est à l'origine de méthodes statistiques et informatiques utilisées dans le monde entier pour reconstruire l'évolution à partir d'ADN, pour des applications en biologie, médecine et écologie.

gétales… Le rêve de Darwin en grand format, car non seulement ces données représentent le code informatique de ces espèces, mais elles permettent d'en reconstruire la parenté et l'évolution. Disposant de séquences « homologues » entre espèces, dérivées d'une même séquence ancestrale, nous avons aujourd'hui des outils qui permettent d'en reconstruire la phylogénie, c'est-à-dire leur évolution, l'histoire de leur divergence. Mes recherches portent sur ces méthodes. Elles combinent des modèles probabilistes décrivant les modifications des séquences au cours du temps, et des algorithmes pour inférer à partir de séquences contemporaines l'histoire la plus probable de leur évolution. Comme prophétisé en 1973 par Theodosius Dobzhansky, « rien en biologie ne fait sens, si ce n'est à la lumière de l'évolution ». Ainsi, en reconstruisant leur évolution, on éclaire les données de séquences si abondantes aujourd'hui. La grande majorité des méthodes d'analyse des séquences reposent aujourd'hui sur des bases évolutives.

Le premier champ d'application des méthodes que nous développons est la génomique comparative et fonctionnelle. Pour comprendre un génome, on le compare à d'autres génomes apparentés. Lorsqu'on découvre un gène, on cherche s'il en existe des homologues dans les banques de données, on en reconstruit la phylogénie et à partir de celle-ci on extrapole au gène nouvellement découvert les fonctions déjà connues de ses homologues. Cette approche est utilisée journellement par des milliers de chercheurs dans le monde.

Ces techniques sont aussi beaucoup utilisées en dehors de la recherche. Nous donnons chaque année des cours de phylogénie qui sont suivis par des acteurs des domaines de l'agroalimentaire, de la sécurité alimentaire, de la justice et de la police scientifique… La phylogénétique a par exemple permis de sauver de la peine capitale ces infirmières bulgares, qui avaient été accusées d'avoir transmis le sida à des enfants hospitalisés en Libye, en prouvant qu'ils avaient été contaminés avant l'arrivée des infirmières.

Dans mon laboratoire, nous nous intéressons plus particulièrement à l'application de ces méthodes en épidémiologie. La génomique est devenue un outil de référence pour étudier les épidémies, comme celle due au Covid-19, que nous subissons actuellement. Les recherches sur ce virus s'appuient sur le même type de méthodes et d'outils que nous avions initialement développés pour l'étude du VIH. Ceux-ci permettent de retrouver l'origine géographique d'une souche, son mode de dispersion, les groupes à risque, la dynamique de propagation, ou encore d'identifier le fameux « patient zéro ». Aujourd'hui, les politiques de santé comme le design de vaccins et de traitements s'appuient largement sur les analyses évolutives issues des séquences de pathogènes prélevées sur les patients.

En épidémiologie, ces nouvelles méthodes permettent de retrouver l'origine géographique d'une souche, son mode de dispersion, les groupes à risque ou d'identifier le fameux « patient zéro ».

LA BIODIVERSITÉ ÉCLAIRÉE

Autre champ d'application majeur : celui de la biodiversité, dans lequel la génomique tend à remplacer – et pour un coût moindre – les explorateurs partis dans de lointaines contrées recenser les espèces. L'opération prend la forme d'un carottage ou de prélèvements, dont l'analyse révèle de manière exhaustive les traces de génomes présents dans l'échantillon. La puissance de ces techniques est considérable. En passant au crible des prélèvements d'eau de rivière, il est par exemple possible d'identifier les espèces de poissons qui y vivent. Cette stratégie a été utilisée avec succès sur le Mékong pour retrouver la trace et les dernières niches d'espèces menacées. L'écologie connaît aujourd'hui un véritable changement de paradigme. L'accélération des technologies permet de lancer des programmes gigantesques, comme l'initiative « Genomes on Earth », qui a pour objectif)))

Les informations contenues dans nos génomes sont précieuses, mais indiscrètes. Elles révèlent les liens de parenté, les prédispositions aux maladies… On imagine aisément les enjeux financiers.

))) de séquencer les génomes de 1,5 million d'espèces différentes. Alors que notre époque pourrait être celle de la sixième extinction de masse des espèces vivantes, la génomique, en sondant le passé, permet de faire des projections sur l'évolution de la biodiversité, essentielles à une politique raisonnée de préservation.

DES FREINS ÉTHIQUES ET TECHNOLOGIQUES

Aujourd'hui pourtant, si les capacités de séquençage continuent de croître, le domaine fait face à plusieurs freins, notamment d'ordre éthique pour tout ce qui touche à l'humain et à la santé. Les informations contenues dans nos génomes sont précieuses… et indiscrètes. Elles révèlent tout de l'héritage génétique des personnes : leurs liens de parenté, leurs prédispositions aux maladies… On imagine aisément les enjeux financiers que de telles données, qui plus est difficiles à protéger par des verrous informatiques, peuvent représenter, pour des compagnies d'assurances, par exemple. C'est notamment ce qui limite la mise en place d'un séquençage systématique à chaque naissance. Ces problèmes ont déjà amené à restreindre les données humaines fournies aux scientifiques, qui bien souvent n'ont plus accès à l'intégralité des séquences, mais à des résumés.

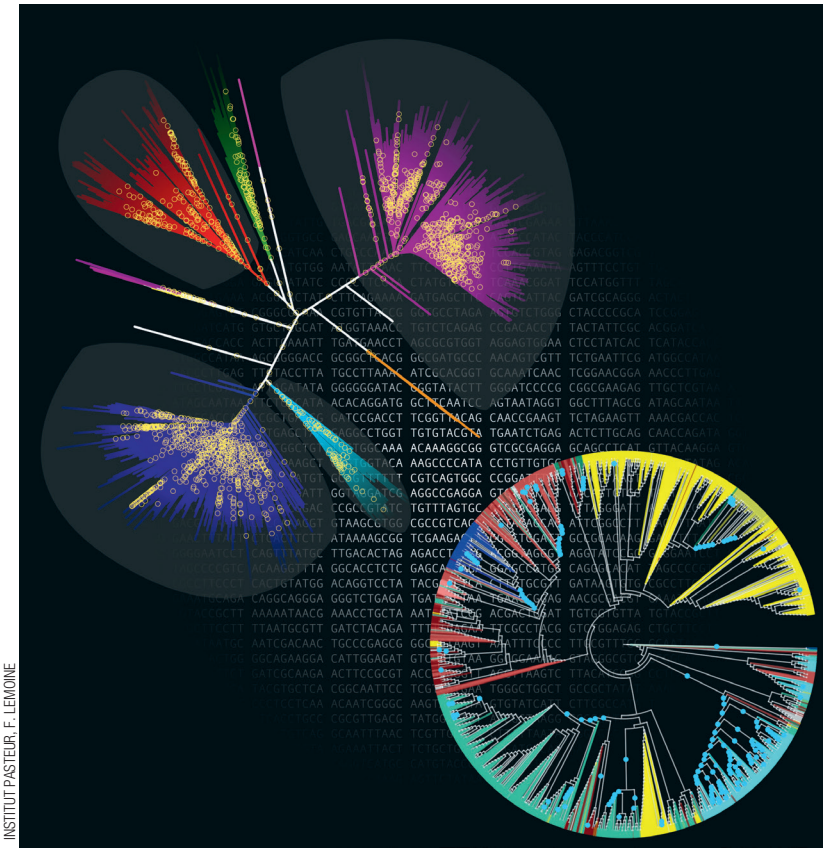
Le secteur de l'analyse, lui, est davantage touché par des obstacles technologiques. Après avoir progressé de manière exponentielle pendant de nombreuses années, les capacités de calcul des ordinateurs tendent en effet à ralentir, pour une raison très simple : la miniaturisation extrême des transistors se rapproche de l'échelle quantique à laquelle se produisent des erreurs aléatoires. Pour notre communauté de bio-informaticiens, c'est un défi important, car il nous faut développer les outils informatiques qui vont permettre d'exploiter les masses de données à très grande échelle, malgré ces limitations. Ceux dont nous disposons actuellement sont bloqués à des dimensions de données qui sont environ 100 fois inférieures à ce qu'il faudra envisager, par exemple, pour le projet « Genomes on Earth ». Ce changement d'échelle demande un effort théorique considérable : il faut repenser en profondeur les algorithmes et leurs bases mathématiques, en inventer de nouveaux, et aller vers un parallélisme massif en subdivisant les calculs en sous-tâches indépendantes. On place aussi beaucoup d'espoir dans les nouvelles méthodes d'IA, en particulier l'apprentissage profond (« Deep Learning ») qui s'est montré si efficace pour le jeu de go et d'autres secteurs comme l'analyse d'images ou la reconnaissance de la parole.

Je n'ai pas de doute que ces difficultés seront résolues. Nous vivons aujourd'hui une époque charnière, dont Darwin ne pouvait pas même rêver. Les données génomiques, alliées à la modélisation mathématique, aux algorithmes et à la puissance de calcul des ordinateurs, permettent d'explorer le passé biologique. Comme en géologie (le centre de la Terre) ou en astronomie (les confins du cosmos), on combine données et approches computationnelles pour dépasser des limites infranchissables à l'être humain.★

EN SAVOIR PLUS

Le site de l'Académie des sciences : **www.academie-sciences.fr**

Sur Interstices.info, revue multimédia de culture scientifique en ligne publiée par l'Inria, **22 focus grand public sur la bio-informatique :** https ://interstices.info/?s = bioinformatique
Dont, spécifiquement sur la phylogénie, interview d'Olivier Gascuel : https ://interstices.info/a-propos-de-la-phylogenie-moleculaire



INSTITUT PASTEUR, F. LEMOINE

Une phylogénie traduit l'évolution et les relations de parenté entre un ensemble d'organismes. En haut : un arbre phylogénique de VIH (les sous-types sont clairement séparés, par exemple, le sous-type B, majoritaire en France, est en bas). En bas à droite : une phylogénie de mammifères.