

# Plus d'octets au service de l'ADN

Des bioinformaticiens brésiliens créent des outils pour étudier des génomes

Marcos Pivetta

PUBLIÉ EN FÉVRIER 2013

Il y a encore 10 ans, il n'y avait pratiquement pas de génomes complets à analyser. On manque aujourd'hui de programmes et de main d'œuvre spécialisée pour faire face à la quantité de séquences d'ADN déjà enregistrées sur des bases publiques de données et qui sortent quotidiennement de la nouvelle génération de séquenceurs. Ces machines qui sont extrêmement rapides déterminent les paires de base du matériel génétique, appelées lettres chimiques, à un prix des milliers de fois inférieur à celui pratiqué au début des années 2000, quand l'aventure qui a menée au séquençage du premier génome humain a pris fin. Afin de relever les défis dans ce domaine, le mathématicien João Meidanis, fondateur associé de l'entreprise Scylla Bioinformatique et professeur à l'Université Publique de Campinas (Unicamp), a investi dans une ligne de recherche visant à créer des méthodes plus simples et plus efficaces pour comparer deux ou plusieurs génomes.

Aux côtés de Pedro Feijão, son ancien élève de doctorat, il a formulé en 2009 la base théorique

d'une méthode appelée single-cut-or-join (SCJ) pour comparer des génomes entiers. L'année dernière, il a testé cette méthode dans la pratique sur les génomes de certains organismes comme les plantes et les bactéries. «Avec notre méthode, nous pouvons comparer facilement deux ou plusieurs génomes sans augmenter de manière exponentielle le nombre de calculs, comme c'est le cas dans d'autres méthodes», affirme João Meidanis. «Nous pouvons ainsi construire des arbres généalogiques et voir quels sont les génomes les plus proches et les plus éloignés du point de vue évolutif. Le mathématicien a été l'un des coordonnateurs en bioinformatique du projet qui a séquencé en 2000, le génome de la bactérie *Xylella fastidiosa*, responsable de la maladie amarelinho (petit jaune) dans les orangeaies. Ce travail a permis à la recherche brésilienne de faire la une de la revue scientifique *Nature*.

Pour confronter tout le matériel génétique d'une espèce à celui d'une autre, les chercheurs doivent recourir à des simplifications. La principale est de considérer que les gènes présents dans les génomes

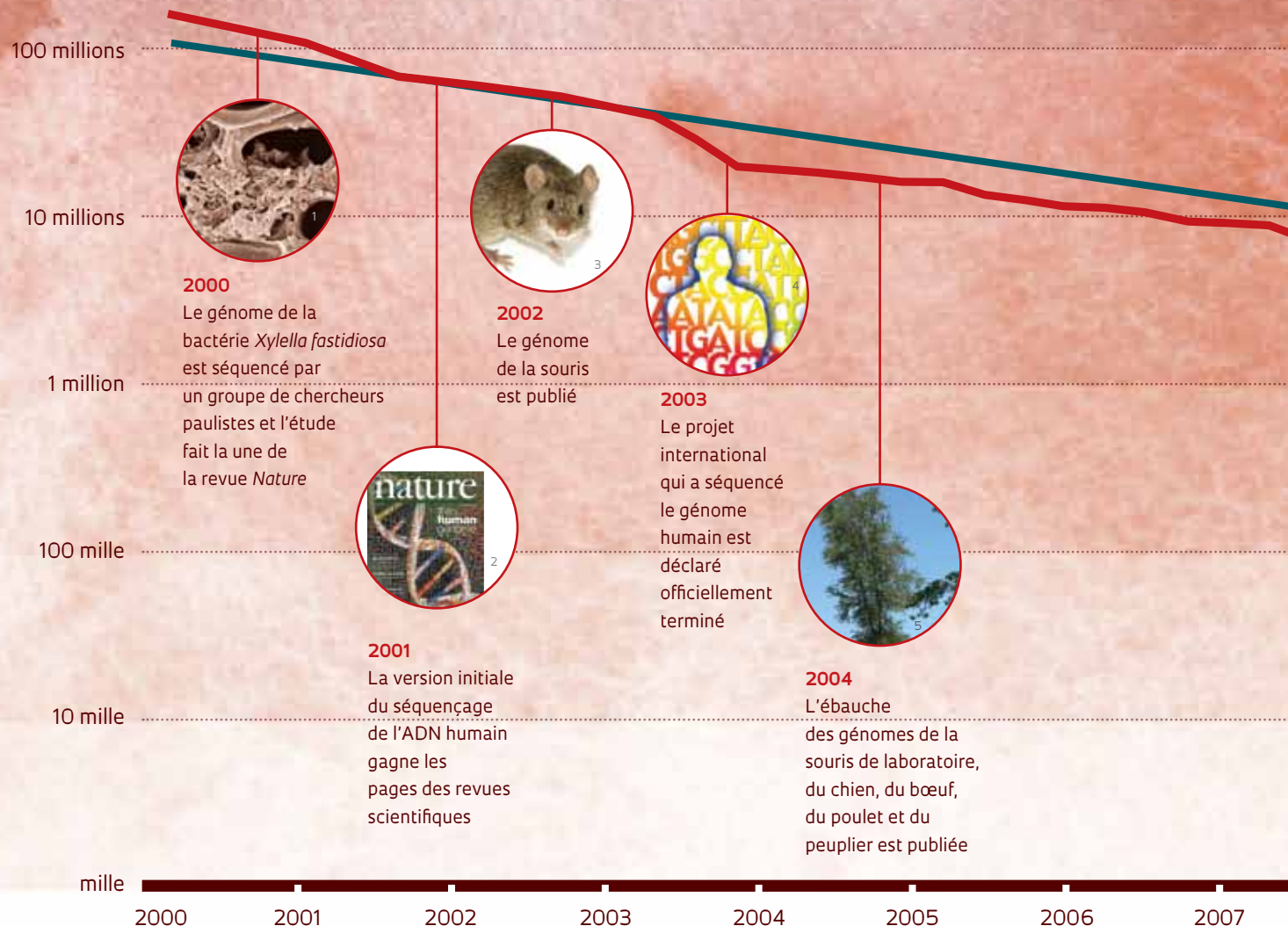


0 39 233 09660 273 650 96

# Le renversement de l'ADN

En une décennie, le coût de séquençage du génome humain est passé de 100 millions dollars US à moins de 10 mille dollars US

\* axe et en escale logarithmique (en dollars US)



comparés sont exactement les mêmes, bien qu'ils soient ordonnés de manière différente dans la séquence spécifique de chaque organisme. En partant de ce raisonnement, les méthodes pour comparer les génomes comptabilisent le nombre de réarrangements nécessaires pour qu'un génome se transforme en un autre. Ces réarrangements sont dus au déplacement de grands segments d'ADN qui ont lieu au fil du temps dans la séquence originale. Ainsi, plus le nombre de réarrangements qui séparent deux génomes sera faible, plus ils seront proches dans l'arbre évolutif.

Dans leur méthode, João Meidanis et Pedro Feijão ont formulé une définition alternative du concept de point de rupture (breakpoint), paramètre important pour trouver des réarrangements dans une séquence et ainsi calculer la proximité de

deux génomes. Le point de rupture est l'endroit où il y a une interruption d'un long segment conservé dans les génomes qui sont en train d'être comparés.

L'année dernière, ils ont encore redéfini une autre méthode de comparaison de génomes plus élaborée que le SCJ. Tout d'abord proposée en 2000, cette deuxième méthode permettait de confronter uniquement des génomes circulaires. Une fois perfectionnée, cette méthode est devenue également utile pour comparer le matériel génétique de chromosomes linéaires. «C'était une des limitations de la méthode originale», commente Pedro Feijão, aujourd'hui employé chez Scylla. La nouvelle méthode basée sur ce que les mathématiciens appellent formalisme algébrique par adjacences n'a pas encore été testée sur des génomes réels. Pour l'instant elle n'existe que de manière théorique.

## LOI DE MOORE

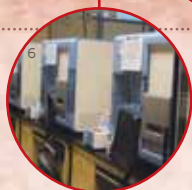
Jusque dans les années 2000, la baisse du prix du séquençage a suivi la théorie de la réduction du coût technologique de la loi de Moore, pour ensuite chuter fortement



2011 Des scientifiques chinois de l'Institut BGI annoncent un projet pour séquencer 3 millions de génomes de plantes, d'animaux, d'êtres humains et de microorganismes



2012 Le génome complet de 1 092 personnes originaires de différents continents est publié dans la revue *Nature*



2008 Les machines de deuxième ou de dernière génération arrivent sur le marché et font chuter le prix du séquençage des génomes

SOURCE: NHGRI GENOME SEQUENCING PROGRAM

2008 2009 2010 2011 2012

## MÉTAGÉNOMIQUE

João Meidanis n'est évidemment pas le seul à sentir les effets de cette nouvelle réalité dans son domaine d'intervention. De retour au Brésil depuis 2011, après avoir passé huit ans au Virginia Bioinformatics Institute, aux États-Unis, João Carlos Setubal, actuellement professeur titulaire de l'Institut de Chimie de l'Université de São Paulo (USP), a constaté que les demandes pour monter des services et mener des recherches dans son domaine ont augmenté et se sont sophistiquées ces derniers temps. Depuis qu'il est revenu à São Paulo, il a par exemple déjà reçu 16 propositions pour collaborer à des initiatives menées par d'autres chercheurs. «Les séquenceurs de dernière génération produisent une quantité astronomique de données génomiques, protéomiques et sur le métabolisme

des organismes», déclare João Carlos Setubal, qui a également été l'un des coordonnateurs de bioinformatique du projet sur Xylella. «Actuellement, grâce à la réduction des prix de la technologie, tout projet de recherche peut séquencer le génome d'un organisme avec un minimum de financement».

La métagénomique qui étudie le microbiote d'une niche écologique est un nouveau domaine qui est apparu ces dix dernières années pour les biologistes et les bioinformaticiens. Le principal projet de João Carlos Setubal est un programme thématique de la FAPESP en métagénomique sur les microorganismes présents au zoo de São Paulo. Dans cette approche, au lieu d'isoler et de cultiver les microorganismes pour extraire de manière séparée l'ADN de chaque espèce, le chercheur retire directement un échantillon de l'environnement à étudier. Dans cet échantillon, l'ADN de différentes espèces est «mêlé» et il incombe au bioinformaticien de trouver des méthodes pour séparer et caractériser le matériel génétique de chaque espèce. «Nous sommes en train d'étudier trois microbiomes au zoo qui sont, le compostage réalisé par les employés du parc, l'eau des étangs et les excréments des singes hurlleurs», déclare Setubal.

La métagénomique est également une manière simple de découvrir des organismes inconnus dans un habitat spécifique. L'équipe d'Ana Tereza Ribeiro de Vasconcelos, coordonnatrice du centre de bioinformatique du Laboratoire National d'Informatique Scientifique (LNCC), à Petrópolis, a participé à la découverte de bactéries magnétiques trouvées dans la lagune d'Araruama, sur le littoral de Rio de Janeiro, l'une des plus salines du monde. L'une des bactéries trouvées est la *Candidatus magnetoglobus multicellularis*. La bactérie identifiée par Ulysses Lins, de l'Université Fédérale de Rio de Janeiro (UFRJ), est difficile à isoler de son environnement et à conserver dans un milieu de culture. «Nous sommes aujourd'hui engagés sur une dizaine de projets en matière de métagénomique», déclare Ana Tereza, qui possède trois séquenceurs dans son laboratoire et qui compte sur une équipe d'environ 25 personnes.

L'échelle de temps et les financements engagés dans des projets consacrés à l'analyse de l'ADN ont radicalement changé au cours de la dernière décennie. Durant les premières années de l'ère génomique, seuls les grands projets osaient s'aventurer dans ce nouveau domaine. En avril 2003, quand le consortium public International qui avait séquencé pour la première fois un génome humain est arrivé à son terme, cette méga-initiative avait consommé 13 ans de travail, fait appel à des centaines de scientifiques d'au moins 18 pays (y compris le Brésil) et avait coûté 12 millions de dollars US. Dans une moindre proportion mais même ainsi grandiose, le séquençage de la Xylella aura coûté 12 millions US\$ à la FAPESP

en impliquant 192 chercheurs durant trois ans.

Actuellement, le séquençage de génomes est devenu une activité de 10 à 20 mille fois plus avancée sur le plan financier qu'il y a une décennie, selon les données du National Human Genome Research Institute (NHGRI), aux États-Unis. L'arrivée en masse sur le marché début 2008 des séquenceurs de deuxième génération qui utilisent une technologie différente de celle des premières machines de type Sanger, ont fait chuter les prix du séquençage à un rythme accéléré qui dépasse de loin les gains de performance prévus par la Loi de Moore dans l'informatique. Aujourd'hui, en deux ou trois jours, pour quelques milliers de dollars, il est possible de déterminer les 3 milliards de lettres chimiques de l'ADN d'une personne. «La bioinformatique est un nouvel outil, une loupe qui nous permet de mieux comprendre le phénomène biologique, qui n'a pas changé, mais qui peut être vu sous un autre angle», déclare Gonçalo Pereira, de l'Institut de Biologie (IB) de l'Unicamp.

Cependant, séquencer est une chose, et extraire de l'information utile sur les milliards de données fournies quotidiennement par les ordinateurs dans les mains des scientifiques en est une autre bien plus complexe. «Le séquençage en soi est actuellement bon marché, c'est devenu une commodity, mais l'analyse des données est chère», déclare le scientifique informaticien João Paulo Kitajima, de Mendelics, entreprise récemment créée qui travaille

sur des diagnostics génomiques personnalisés. «La demande en bioinformatique a augmenté de manière exponentielle et il y a un gap entre l'offre et la demande de spécialiste au Brésil et à l'étranger».

Il est difficile d'estimer avec précision la taille de la communauté de bioinformaticiens dans le pays. Selon Guilherme Oliveira, président de l'Association Brésilienne de Bioinformatique et de Biologie Informatique (AB3C), environ 300 personnes, comme des professeurs, des élèves et des chercheurs, conservent un lien avec l'entité. «Auparavant le bioinformaticien était autodidacte», déclare Oliveira, coordonnateur du centre de bioinformatique de la Fiocruz de Minas Gerais. «Aujourd'hui une bonne partie de ces personnes sortent d'un troisième cycle et il y a un spécialiste du secteur dans chaque état. Ce qui est nouveau, c'est que des entreprises travaillent maintenant dans ce domaine». De grandes universités brésiliennes comme l'USP, l'UFRJ, l'Université fédérale de Minas Gerais (UFMG), et la Fiocruz ont des programmes de troisième cycle spécifiques en bioinformatique. D'autres universités travaillent le thème comme une ligne de recherche dans le

## La réduction des coûts de séquençage a permis aux techniques génomiques de devenir financièrement accessibles

## La Chine possède le plus grand centre de séquençage du monde

En moins de 15 ans, un centre chinois de bioinformatique est passé de la condition de petit associé, du consortium de petit associé, du consortium international qui a cartographié le premier génome humain, à celle de la plus grande puissance mondiale en termes de séquençage d'ADN. Fondé en 1999, Beijing Genomics Institute (BGI, signe en anglais) est aujourd'hui propriétaire de 180 séquenceurs, la plupart de dernière génération et qui peuvent produire quotidiennement 6 téraoctets de données, ce qui équivaut aux génomes complets de 2 mille individus. Ils ont 4 mille employés et des filiales aux États-Unis, en Europe et au Japon. Les performances des chinois, qui travaillent sur une échelle gigantesque, créent les conditions nécessaires pour que le prix du séquençage du génome humain passe d'ici peu à 1 000 dollars US et leur offrent la possibilité de participer à des projets de pointe, qui vont bien au

delà du séquençage génétique d'un symbole national, le panda géant, prouesse atteinte il y a trois ans.

En 2010, par exemple, l'Institut BGI a séquencé le premier génome complet d'un ancêtre de l'homme en séquençant l'ADN d'un esquimau vieux de 4 mille ans. En 2012 il a fourni l'ADN de 100 chinois dans un projet international qui étudie le génome d'environ mille individus provenant de différentes régions de la planète.

L'année dernière le centre s'est également proposé de séquencer durant les prochaines années 3 millions de génomes d'êtres humains, de plantes, d'animaux et de microorganismes.

La politique chinoise est agressive à tous les niveaux, non seulement dans le domaine scientifique, mais également sur le plan commercial. Outre le fait de vendre ses services en bioinformatique,

l'Institut BGI essaie de se garantir un accès aux avancées les plus récentes du secteur. En début d'année, le centre asiatique a reçu le feu vert des américains pour acheter pour une somme de 177 millions de dollars US une compagnie californienne, Complete Genomics, qui a développé une nouvelle technologie de séquençage et dont les résultats seraient plus précis que ceux obtenus avec les méthodes actuelles utilisées dans le monde entier.



Panda géant: l'un des génomes que les chinois du Beijing Genomics Institute (BGI) ont séquencé entièrement



Les eaux hyper-salines de la lagune d'Araruama, à Rio de Janeiro: des études métagénomiques y trouvent des bactéries magnétiques

cadre du troisième cycle et dans un domaine plus large comme la biologie ou l'informatique.

Le travail de séquençage et d'analyse du génome du *Schistosoma mansoni*, parasite responsable de l'esquistossomose, est le projet de plus grande visibilité sur lequel ait travaillé le centre de bioinformatique de l'unité de Minas Gerais de la Fiocruz, ces dernières années. Toutefois, les six séquenceurs et les 15 spécialistes en bioinformatique du secteur dirigé par Oliveira ont participé à environ 60 autres projets distincts qui incluent des études sur le génome du cancer, sur les agents infectieux, sur les races de bovins et des travaux en métagénomique. Aujourd'hui, le centre analyse et gère encore des données pour le Réseau de Recherche d'Identification Moléculaire de la Biodiversité Brésilienne (BR-BoL), coordonné par Cláudio Oliveira, de l'Institut de Biosciences de l'Université Publique Pauliste (Unesp) de Botucatu, qui cataloguera 120 mille exemplaires de 24 mille espèces en quatre ans. Le BR-BoL est le bras brésilien du projet international Barcode of Life (Code-barres de la Vie), destiné à identifier des espèces à travers la caractérisation de leur ADN.

La bioinformatique s'est répandue dans le pays et a atteint des centres très éloignés des grandes capitales du Sud-Est. Artur Silva mène des recherches dans ce domaine et collabore avec des groupes de São Paulo à l'université Fédérale de l'état du Pará (UFPA). Depuis le mois de mai dernier, Sandro de Souza, qui durant des années a dirigé ce secteur à l'Institut Ludwig de Recherche sur le Cancer de São Paulo, se trouve à l'Institut du Cerveau de l'Université Fédérale de l'état du Rio Grande do Norte (UFRN). Il n'a pas de séquenceur à ses côtés pour l'aider dans la capitale Potiguar [N.T. : la ville de Natal], mais cela ne l'inquiète pas du tout. «Il est même possible de faire du séquençage dans le nuage internet si on veut», affirme Sandro de Souza. «Je commence mes travaux en neurosciences sans aucun problème».

En effet, Sandro de Souza a également accès à toutes les machines de l'Institut Ludwig qui, après la fermeture de l'Institut dans la capitale pauliste, ont

été transférées à la Faculté de Médecine de Ribeirão Preto (FMRP) de l'USP. C'est là, l'année dernière, qu'a commencé à fonctionner le Centre de Médecine Génomique. «Les techniques génomiques et bioinformatiques révolutionneront la pratique médicale comme c'est déjà le cas pour l'imagerie médicale», déclare Wilson Araújo da Silva Júnior, l'un des responsables du nouveau centre de la FMRP.

Pour que les services de séquençage et d'analyse d'ADN et d'ARN deviennent plus accessibles, l'Unicamp inaugurera, le premier mars, le Laboratoire Central de technologies de Haute Performance (LacTAD), qui travaillera dans les domaines de la génomique, la protéomique, la biologie cellulaire et la bioinformatique proprement dite. Parmi les équipements du laboratoire, il y a deux séquenceurs de nouvelle génération de l'entreprise Illumina qui en quelques jours sont capables de séquencer complètement un ADN humain, et un troisième équipement pour séquencer des régions spécifiques des génomes. Normalement, les machines du centre fonctionnent depuis leur arrivée l'année dernière à l'université et leur installation dans différentes unités. À partir du mois prochain, elles commenceront à fonctionner dans le bâtiment de 2 000 mètres carrés construit pour le LacTAD.

«Nous pensons qu'il y a une demande réprimée pour ce type de service et la bioinformatique est devenue un goulot d'étranglement pour de nombreuses recherches dans les domaines biologiques», déclare le chimiste Ronaldo Pilli, Directeur de la Recherche de l'Unicamp, responsable du projet du nouveau laboratoire. «Nous suivons la tendance mondiale en offrant ce type de service de manière centralisée. Ainsi, il est plus facile d'acheter, d'exploiter et de maintenir les équipements à jour». Les équipements du LacTAD ont coûté environ 5,5 millions de reais et ont été acquis à travers le programme Multiutilisateurs de la FAPESP. Le bâtiment a été financé par l'université pour un montant de 4 millions de reais.

Le LacTAD fournira ses services aux chercheurs de l'Unicamp ainsi qu'à d'autres universités et aux entreprises. Sur le site du laboratoire il y a un formulaire destiné aux chercheurs qui veulent coter le montant de leurs services. «Nous ferons des travaux qui pourront coûter de 100 reais à 100 mille reais» déclare Ronaldo Pilli. C'est la démocratisation de la bioinformatique. ■

## Projets

1. Études de la diversité microbienne dans le Parc Zoologique de São Paulo – n° 11/50870-6; **Modalités** Programme Biota – Projet Thématique; **Coordonnateurs** João Setubal – USP; **Investissement** 1 711 698,25 reais (FAPESP);

2. EMU: Laboratoire Central de Technologie de haute Performance – n° 09/54129-9; **Modalités** Programme d'Équipements Multiutilisateurs; **Coordonnateurs** Fernando Ferreira Costa – Unicamp; **Investissement** 6 034 431,00 reais (FAPESP).