

AUX RACINES DE L'ÉVOLUTION

LE LOGICIEL QUI ÉLAGUE L'ARBRE DE LA VIE

MALGRÉ UNE MASSE D'INFORMATIONS GÉNÉTIQUES SANS CESSER GRANDISSANTE, IL N'EST PAS TOUJOURS AISÉ DE CLARIFIER LES LIENS DE PARENTÉ EXISTANT ENTRE LES ESPÈCES VIVANTES. UNE MÉTHODE RÉCENTE FAIT LE TRI ET PERMET DE TRANCHER DANS CERTAINS CAS NOTABLES DE DÉSACCORD ENTRE MORPHOLOGISTES ET GÉNÉTICIENS

L'arbre de la vie illustre les liens de parenté qui existent entre toutes les formes de vie sur Terre. Depuis toujours les biologistes tentent de placer aux bons endroits les branches de ce vénérable végétal en se basant sur la comparaison de caractères morphologiques puis, surtout, génétiques. Pourtant, malgré les quantités astronomiques de données génétiques à disposition et le développement d'ordinateurs de plus en plus puissants pour les traiter, les modèles mathématiques qui dessinent ces arbres phylogénétiques commettent parfois encore des erreurs. C'est dans le but de remédier à ce type de défaillances que Juan Montoya-Burgos et Carlos Rivera-Rivera, respectivement directeur de laboratoire et doctorant au Département de génétique et évolution (Faculté des sciences), ont mis au point un nouveau logiciel. Publié le 23 février dans la version en ligne de *Molecular Biology and Evolution* – il illustre également la couverture de l'édition imprimée du mois de juin –, leur outil, baptisé LS³, permet de résoudre un biais spécifique appelé «l'attraction des longues branches», une erreur qui peut avoir une fâcheuse tendance à s'amplifier lorsqu'on cherche à l'atténuer à grand renfort de nouvelles données. Explications.

Les longues branches s'attirent Pour réaliser leurs arbres phylogénétiques, les biologistes disposent de plusieurs modèles. Ces derniers analysent les différences apparaissant entre les séquences génétiques des espèces étudiées et calculent les liens de parenté les plus probables.

Ces modèles sont basés sur le principe que toutes les espèces prises en compte évoluent à des vitesses proches.

«Le biais d'attraction des longues branches (LBA) est un artefact – une erreur – produit par les modèles lorsque ceux-ci comparent des espèces dont les vitesses d'évolution sont très différentes les unes des autres, explique Juan Montoya-Burgos. Certaines espèces accumulent en effet des mutations dans leur patrimoine génétique plus rapidement que d'autres (lire ci-contre). Dans ce cas, les valeurs

D'UN SEUL COUP, LES LAPINS ET LES RONGEURS, POURTANT PROCHES DU POINT DE VUE MORPHOLOGIQUE, SE SONT VUS SÉPARÉS LES UNS DES AUTRES

moyennes des paramètres qu'utilise le modèle s'éloignent trop de la réalité, ce qui génère un résultat erroné.»

Ces lignées à forte vitesse d'évolution ont la particularité d'apparaître sur les arbres phylogénétiques avec une barre (branche) plus longue que la moyenne. Les modèles, qui interprètent mal ces vitesses d'évolution hétérogènes, ont alors tendance à produire des arbres dans lesquels ces longues branches sont artificiellement rapprochées les unes des autres – d'où le nom du biais.



JUAN MONTOYA-BURGOS

POISSON-CHAT DU GENRE «CLARIAS», DE LA FAMILLE DES «CLARIIDAE» ET DE L'ORDRE DES SILURIFORMES.

A une époque où les machines à séquencer l'ADN produisent des millions de nouvelles séquences chaque mois, on pourrait penser qu'il suffit d'augmenter la quantité de données pour atténuer ce biais, voire le faire disparaître. Le problème, c'est qu'il s'agit souvent d'une erreur systématique. Dans un tel cas, multiplier les données génétiques fournies au modèle ne fait qu'augmenter son amplitude.

Le LBA s'est notamment manifesté lors des premières tentatives visant à classer les mammifères placentaires à l'aide de données génétiques. D'un seul coup, les lagomorphes (lapins, lièvres...) et les rongeurs (souris, écureuils, cochons d'Inde...), pourtant proches du point de vue morphologique, se sont vus séparés les uns des autres, les seconds étant placés sur le même embranchement que les primates (humains, singes...) et tous les laurasiathériens (hérissons, chauves-souris, chiens, chevaux, baleines, vaches, cochons...).

Autre exemple: les modèles d'évolution, sous l'influence du LBA, ont dans un premier temps estimé que les arthropodes (insectes, arachnides, crustacés...) avaient une parenté commune plus récente avec les deutérostomes (de l'étoile de mer à l'être humain) qu'avec les nématodes (vers ronds). En réalité, il est actuellement admis que dans ce trio, ce sont les arthropodes et nématodes qui sont les cousins les plus proches.

L'arbre à poissons-chats «J'ai moi aussi été confronté à l'artefact LBA, raconte Juan Montoya-Burgos. Cela remonte à une quinzaine d'années, lorsque je travaillais à ma thèse sur les poissons-chats (siluriformes), un ordre qui regroupe, à lui seul, le quart des poissons d'eau douce du monde et qui doit son nom aux barbillons que ses membres portent sur le nez.»

Divisés en trois lignées bien distinctes, les Loricarioidei (L), les Siluroidei (S) et les Diplomystidés (D), les poissons-chats représentent un cas d'école pour la classification des espèces. Sur la base de considérations morphologiques, les biologistes estiment que L et S partagent un ancêtre commun (SL) qui lui-même possède un autre ancêtre commun plus ancien avec D. Cette classification (LS-D) est solide du point de vue morphologique et, pour Juan Montoya-Burgos, difficile à mettre en doute.

Sa surprise est donc totale lorsque, dans le cadre de sa thèse, il soumet à l'analyse deux gènes communs à tous les poissons-chats et dont il a décrypté les séquences pour un grand nombre d'espèces. Pour le premier, l'analyse phylogénétique produit les mêmes relations trouvées par les morphologistes (LS-D). Pour le second, en revanche, elle aboutit d'abord sur un ancêtre commun entre S et D puis sur un autre, plus ancien, avec L (SD-L), en contradiction flagrante avec le résultat précédent.

« D'emblée, je n'ai pas cru à l'information délivrée par le second gène, affirme Juan Montoya-Burgos. J'ai tout de suite pensé à un artefact. D'ailleurs, on remarque sur l'arbre que certaines branches sont plus longues que d'autres et qu'elles s'attirent. Mais les moyens théoriques et informatiques de l'époque étant insuffisants, j'étais incapable de le prouver. Je n'ai donc pas publié ce résultat en me disant que je réglerais ce problème plus tard. »

Cependant, en 2006, une équipe américaine publie dans *Molecular Phylogenetics and Evolution* l'analyse de deux autres gènes communs aux poissons-chats et obtient le même résultat que celui que Juan Montoya-Burgos a précisément écarté. Contre l'intuition du biologiste genevois, la solution fournie par la biologie moléculaire (SD-L) est alors considérée comme le meilleur reflet de la réalité, au détriment de celle des morphologistes (LS-D).

Plusieurs années plus tard, toujours pas convaincu par cette nouvelle classification, Juan Montoya-Burgos décide enfin de se pencher sérieusement sur le problème du LBA. Partant du constat qu'il est très difficile d'améliorer les modèles pour qu'ils correspondent mieux à la réalité, lui et Carlos Rivera-Rivera choisissent de faire le contraire, soit d'adapter les données au modèle. La démarche est provocante, mais elle se base sur des critères très précis et solides. Ainsi est né LS³.

Enlever des données « En résumé, il s'agit d'une méthode objective permettant de trier les données de manière à ne conserver que celles qui sont adaptées aux modèles, explique Juan Montoya-Burgos. Concrètement, le logiciel élimine les portions de séquences d'ADN – et non pas l'ensemble de l'information pour une espèce comme cela s'est fait avant – qui ont évolué trop vite. Il commence par ôter celle qui est la plus problématique et répète l'analyse. Si le biais est toujours présent, il enlève une deuxième portion et ainsi de suite, jusqu'à ce que les vitesses d'évolution des séquences restantes soient homogènes. »

Pour s'assurer que le logiciel fonctionne bien, il a été confronté à des simulations, c'est-à-dire à des ensembles de données conçus de telle manière qu'ils génèrent à coup sûr des artefacts. Il a ensuite dû résoudre deux problèmes réels dont les solutions sont connues, en l'occurrence ceux précités des lapins et des vers ronds. A chaque fois, LS³ s'en est sorti haut la main en élaguant judicieusement les portions de séquences évoluant trop vite, afin de faire ressortir les branchements sous-jacents.

CE N'EST PAS SEULEMENT LES SÉQUENCES QUI ONT ÉVOLUÉ TROP VITE QUI POSENT PROBLÈME MAIS AUSSI CELLES QUI CHANGENT TROP LENTEMENT

L'étape suivante consiste désormais à vérifier que le logiciel est capable de trancher dans des cas non encore résolus, comme celui des poissons-chats. « Curieusement, cela n'a pas fonctionné tout de suite, admet Juan Montoya-Burgos. Le logiciel a enlevé des données tant et plus qu'il n'en restait plus assez à la fin pour faire tourner le modèle de manière fiable. C'est alors que nous avons compris que ce n'est pas seulement les séquences qui ont évolué trop vite qui posent problème mais aussi celles qui changent trop lentement. Nous avons donc perfectionné notre logiciel, dont la nouvelle version s'appelle LS4 [non encore publié, ndlr]. Grâce à elle, nous avons enfin pu résoudre l'arbre des poissons-chats. »

Et donner raison à qui ?

Aux morphologistes.

Anton Voss

ARBRE PHYLOGÉNÉTIQUE
COMPORTANT DES ESPÈCES D'EUKARYOTES (ORGANISMES COMPOSÉS DE CELLULES POSSÉDANT UN NOYAU).

LUCA (« LAST UNIVERSAL COMMON ANCESTOR ») EST LE DERNIER ANCÊTRE COMMUN UNIVERSEL DE TOUTES LES FORMES DE VIE SUR TERRE, Y COMPRIS LES ARCHAE ET DES BACTÉRIES QUI SONT ABSENTES DE CET ARBRE.

L'ÉVOLUTION EST LE RÉSULTAT DE MUTATIONS GÉNÉTIQUES ALÉATOIRES CONSERVÉES SOIT PAR LE HASARD (MUTATIONS NEUTRES), SOIT PAR LA SÉLECTION NATURELLE.

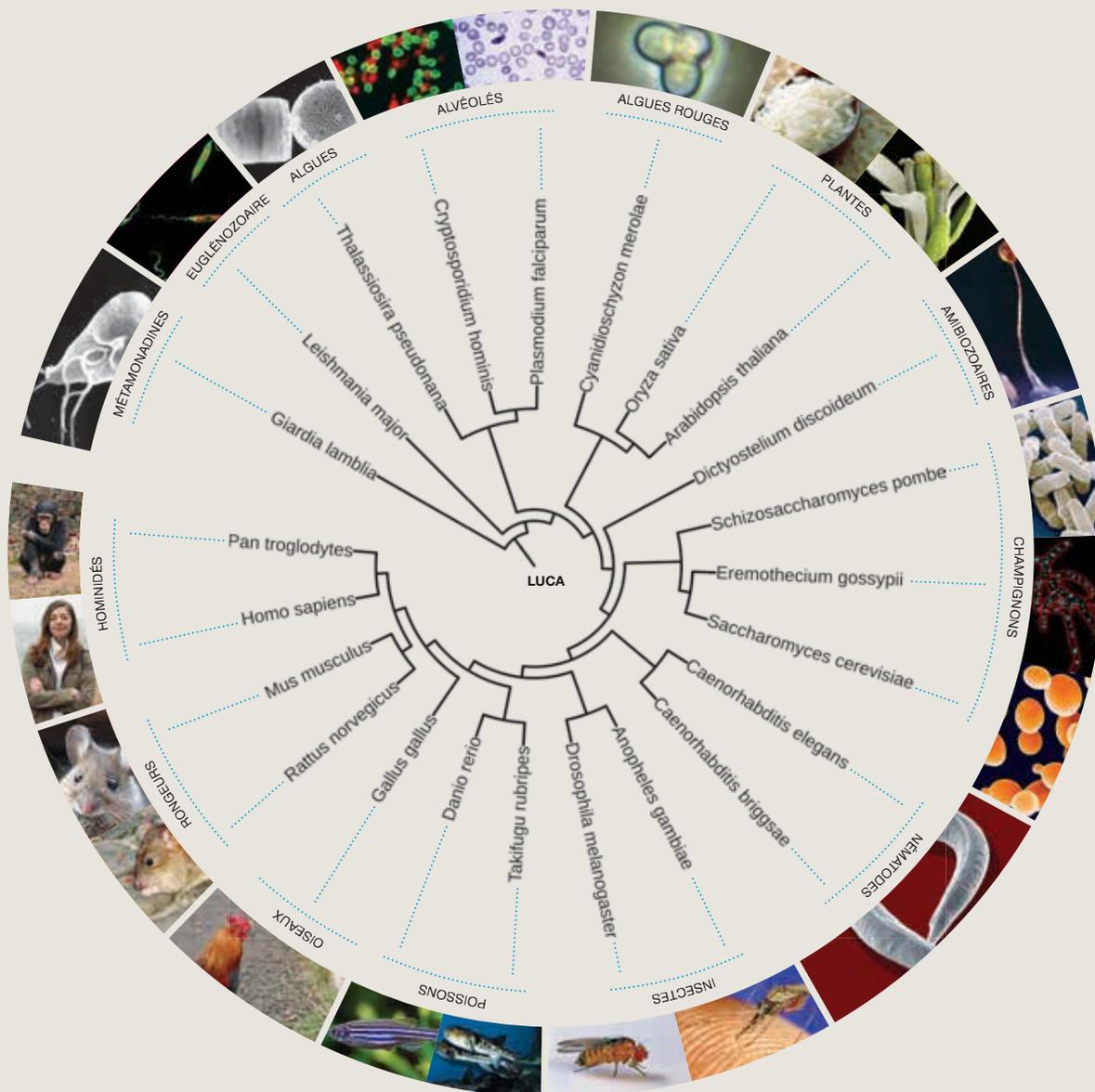
LA SOURCE DE CES MUTATIONS SE TROUVE EN GÉNÉRAL DANS DES RATÉS QUI SURVIENNENT RÉGULIÈREMENT DANS LES CELLULES REPRODUCTRICES LORS DE LA RÉPLICATION DE LEUR ADN. CERTAINES DISPARAISSENT PARCE QUE LA PROGÉNITURE N'EST PAS VIABLE, D'AUTRES SONT CONSERVÉES, CAR ELLES SONT NEUTRES OU BÉNÉFIQUES.

PLUS LE TEMPS SÉPARANT DEUX GÉNÉRATIONS EST COURT (C'EST SOUVENT LE CAS CHEZ LES ESPÈCES DE PETITES TAILLES MAIS PAS TOUJOURS), PLUS L'ACCUMULATION DE MUTATIONS EST IMPORTANTE.

PLUS LA TAILLE DE POPULATION (TAILLE EFFICACE) D'UNE ESPÈCE ANIMALE EST PETITE, MOINS LA SÉLECTION NATURELLE EST VIVE, PERMETTANT LA AUSSI À UN PLUS GRAND NOMBRE D'ERREURS DE SUBSISTER ET DE CIRCULER DANS LE PATRIMOINE GÉNÉTIQUE DE LA POPULATION.

LES GÈNES TRÈS IMPORTANTS POUR LA SURVIE NE CHANGENT PRESQUE PAS, LA MOINDRE MUTATION POUVANT ÊTRE FATALE À L'INDIVIDU.

D'AUTRES FACTEURS PEUVENT ÉGALEMENT ENTRER EN LIGNE DE COMPTE COMME LA DUPLICATION, UN PROCESSUS QUI VOIT UN DOUBLEMENT D'UNE PORTION DU CHROMOSOME OU DU CHROMOSOME ENTIER (VOIRE MÊME DE TOUT LE GÉNOME) ET LORS DUQUEL LA VITESSE D'ÉVOLUTION EST LA AUSSI MOMENTANÉMENT AUGMENTÉE.



DANS LA RAMURE DE L'ARBRE DE LA VIE

De l'arbre de la vie, on ne connaît en somme que la cime. Les innombrables extrémités de brindilles qui composent cette vaste canopée représentent les espèces telles qu'elles existent aujourd'hui. Plonger en dessous, là où les branches se rassemblent de proche en proche, revient à remonter le cours de l'évolution et à identifier les liens de parenté entre les différentes formes de vie.

Ce type de représentation est notamment popularisé par Charles Darwin dans son *Origine des*

espèces au moyen de la sélection naturelle en 1859. Le savant britannique voit dans la compétition entre les rameaux bourgeonnants la même rivalité qui oppose les espèces pour assurer leur survie. Celles qui parviennent à la lumière représentent les espèces qui ont réussi à s'adapter. Les branches mortes symbolisent, quant à elles, les lignées éteintes dont il ne reste, au mieux, que des fossiles.

Aujourd'hui, grâce à la génétique et à la puissance des ordinateurs, il est possible de dessiner des

arbres phylogénétiques de grande envergure, englobant les animaux, les plantes mais aussi les bactéries et les archaebactéries (voire même les virus). Ces arbres complexes ne sont pas exempts d'erreurs (lire ci-dessus), mais ils permettent de fournir une image de plus en plus précise de l'histoire de l'évolution. Sur ces représentations, la canopée symbolise les espèces vivantes. A la racine, là où toutes les branches finissent par se rejoindre, se trouve, théoriquement du moins, l'organisme qui a été le dernier ancêtre

commun à toutes les formes de vie actuelles : LUCA (*Last Universal Common Ancestor*). Il ne s'agit pas de la première forme de vie apparue sur Terre puisque LUCA descend lui-même d'autres ancêtres. Les chercheurs tentent actuellement de reconstruire quelques-unes de ses propriétés en se basant sur les points communs à toutes les formes de vie connues (présence d'ADN, d'ARN, fabrication de protéines, division cellulaire, etc.).