

forme à partir d'ions positifs qui sont extraits d'une des électrodes. Quand la chaîne métallique relie les deux électrodes, le courant d'électrons peut facilement circuler et la résistance est faible. Le bit d'information vaut alors 1. En inversant le champ électrique, les atomes redeviennent des ions et migrent vers l'électrode : la chaîne se brise, la résistance augmente et le bit d'information passe de 1 à 0. Le processus est réversible.

Les progrès sont si rapides dans ce domaine qu'il est difficile de donner les caractéristiques de ces mémoires. Le temps d'écriture varie de la nanoseconde à la milliseconde suivant les dispositifs. Par ailleurs, les techniques de fabrication, surtout pour les mémoires à oxydes, sont relativement courantes, et leur coût de production est très compétitif. En revanche, les phénomènes microscopiques mis en jeu restent à mieux comprendre.

Tous les dispositifs présentés n'en sont pas au même point de maturité et leur présence sur le marché peut prendre

encore quelques années. Si les disques durs commencent déjà à disparaître dans les appareils nomades, les mémoires dynamiques pourraient n'être remplacées que d'ici trois à cinq ans.

Un horizon encore flou

La demande de mémoires toujours plus efficaces va continuer de croître. Cela encouragera les universités, les centres de recherche et les constructeurs à explorer les pistes évoquées, ainsi que de nombreuses autres. La concurrence étant forte et favorable à l'inventivité, l'évolution peut être rapide, et il est difficile de voir clairement quels composants s'imposeront dans cinq, dix ou vingt ans. La tendance qui se dégage est l'abandon de l'idée d'une technologie prédominante et universelle qui remplacera toutes les mémoires existantes. La solution optimale est de faire cohabiter plusieurs techniques et d'améliorer les performances de chacune. Jusqu'où ? On l'ignore. ■

■ BIBLIOGRAPHIE

H.-S. Wong *et al.*, **Metal oxide RRAM**, *Proceedings of the IEEE*, vol. 100, n° 6, pp. 1951-1979, 2012.

T. Kawahara *et al.*, **Spin-transfer torque RAM technology : Review and prospects**, *Microelectronics Reliability*, vol. 52, pp. 613-627, 2012.

M. Hilbert et P. Lopez, **The world's technological capacity to store, communicate and compute information**, *Science*, vol. 332, pp. 60-65, 2011.

G. W. Burr *et al.*, **Phase-change memory technology**, *J. Vac. Sci. Technol. B*, vol. 28, n° 2, pp. 223-262, 2010.



france culture

LA MARCHÉ DES SCIENCES

Découvertes, inventions, aventures savantes au fil de l'Histoire

Aurélien Luneau
14h-15h / chaque jeudi

franceculture.fr

SCIENCE

Stocker de l'information dans l'ADN

Christophe Dessimoz

Grâce à l'avancée des techniques de la génomique, il devient possible d'utiliser l'ADN comme support numérique pour archiver des données sur le long terme.

Deux piscines olympiques: tel est le volume qu'il faudrait pour contenir l'ensemble des données produites dans le monde en 2012 (environ 3×10^{21} octets), si on les stockait dans les toutes récentes clefs USB de un téraoctet (10^{12} octets). Au fil des ans, ce volume croît de façon exponentielle et, avec lui, le besoin de supports numériques fiables pour archiver l'ensemble des informations sur le long terme. De plus, la durée de vie des supports actuels étant limitée (pas plus de quelques dizaines d'années), la pérennisation des données nécessite une maintenance constante et lourde, fondée sur des copies multiples et régulières sur divers supports (voir l'article de Francis Daumas et Marion Massol dans ce numéro).

Pourtant, il existe depuis plus de trois milliards d'années un support numérique universel, compact et stable: l'ADN, la molécule porteuse de l'information génétique des organismes vivants. Codé sous la forme d'une succession de quatre molécules – les nucléotides A, T, G, C –, le programme de la vie est numérique: les nucléotides codent de l'information, au même titre que les bits (0 ou 1) de l'informatique. L'ADN est de plus très stable – on a retrouvé, dans des carottes de glace, des gènes datant de 450 000 à 800 000 ans – et dense: si l'on codait sur de l'ADN toute l'information produite en 2012, elle tiendrait dans le coffre d'une voiture...

En outre, le fait que l'ADN soit universel limite les risques d'obsolescence. Tant

qu'il y aura de la vie sur Terre et un intérêt pour la comprendre, la technique pour lire l'ADN existera. Aussi les progrès récents de la génomique nous ont-ils conduits, à l'Institut européen de bio-informatique, à étudier la possibilité d'utiliser l'ADN comme support d'archives numériques.

Tous les sonnets de Shakespeare sur ADN

Peu après la caractérisation de l'ADN, publiée en 1953 par James Watson et Francis Crick, l'idée d'utiliser cette molécule comme support d'information a été envisagée. En 1964, un physicien et officier radio russe, Mikhail Neïman, voit dans les systèmes et processus d'information biophysiques un moyen de miniaturiser les dispositifs de stockage et de traitement de l'information. L'idée a été mise en œuvre à la fin des années 1990, avec le développement du génie génétique. Plusieurs équipes ont su encoder de l'information sur l'ADN, mais les informations étaient courtes (un message secret annonçant le débarquement du 6 juin 1945, une comptine anglaise avec paroles, musique et petite image, la Déclaration des droits de l'homme...). Surtout, les manipulations génétiques étaient laborieuses et *ad hoc*.

Ces dernières années, cependant, les progrès des techniques de séquençage ont entraîné un tournant important. En 2010, le biologiste américain Craig Venter, de

L'AUTEUR



Christophe DESSIMOZ est professeur de bio-informatique au Collège universitaire

de Londres. Il est aussi affilié à l'Institut européen de bio-informatique (EMBL-EBI). Article écrit en collaboration avec Marie-Neige Cordonnier.

BIBLIOGRAPHIE

N. Goldman *et al.*, Towards practical, high-capacity, low-maintenance information storage in synthesized DNA, *Nature*, vol. 494, pp. 77-80, 2013.

G. M. Church *et al.*, Next-generation digital information storage in DNA, *Science*, vol. 337, pp. 1628-1629, 2012.

Entretien avec Ch. Dessimoz à écouter sur: bit.ly/pls_433_dessimoz

l'Institut du même nom, et son équipe ont synthétisé un génome complet d'un million de paires de bases (ou nucléotides) à partir du code du génome d'une bactérie, qu'ils ont introduit dans une autre espèce de bactérie dont ils avaient ôté le génome. Si l'opération était à nouveau *ad hoc* et fort coûteuse (estimée à plusieurs dizaines de millions d'euros), elle montrait qu'il était désormais possible de coder une grande quantité d'information sur de l'ADN en un temps raisonnable.

Depuis, le coût du séquençage a baissé de façon drastique, ce qui ouvre encore de nouvelles perspectives. Encoder de l'information sur l'ADN n'est pas compliqué. Il suffit de passer d'un système binaire (0, 1) à un système codé en base 4 en convertissant chaque couple binaire – 00, 01, 10, 11 – en un des quatre nucléotides A, T, G, C. Toutefois, deux aspects fondamentaux sont à prendre en compte. D'une part, coder une longue séquence coûte cher. Nous avons donc pris le parti de synthétiser des fragments courts (150 paires de bases), meilleur marché.

D'autre part, il s'agit de limiter au maximum les risques d'erreurs. Les erreurs de lecture des séquences se concentrent là où se suivent plusieurs nucléotides identiques. Pour éviter ces répétitions, nous avons converti les données binaires en base 3 et non en base 4, et utilisé le quatrième nucléotide comme une sorte d'opérateur de permutation. Le code en base 3 est composé de 0, 1, 2, auxquels on fait correspondre un nucléotide A, T, G ou C. Toute répétition est évitée en modifiant la correspondance chiffres-nucléotides après chaque chiffre codé. Par exemple, si l'on veut coder 22 et que le premier 2 est codé par A, le second 2 sera codé par une nouvelle correspondance ne faisant intervenir que les trois autres nucléotides (T, G ou C).

Enfin, pour limiter encore les erreurs, nous avons synthétisé des fragments d'ADN dont les séquences se chevauchent, de façon que chaque nucléotide soit couvert par quatre fragments différents, et nous avons ajouté, au début et à la fin de chaque fragment, des informations indiquant sa position dans la séquence complète.

Nous avons testé notre méthode sur cinq fichiers différents : un fichier texte en code ASCII comportant tous les sonnets de Shakespeare, le PDF d'un article scientifique, un extrait sonore du

Où stocker l'ADN synthétisé ?

Un endroit frais, sombre et sec suffirait à conserver de l'ADN lyophilisé pendant plusieurs milliers d'années. Une telle banque existe déjà : inaugurée en 2008, la Réserve mondiale de semences du Svalbard, une immense enceinte réfrigérée creusée dans l'île norvégienne du Spitzberg pour préserver la diversité des semences, n'a pas de personnel permanent sur place. Une cave à vin conviendrait aussi...



discours de Martin Luther King (format mp3), une photo (format JPEG 2000) et le texte (ASCII) du code qui nous sert à convertir les données binaires en base 3, pour montrer que l'on peut encoder le codage lui-même...

En deux jours, les données ont été encodées sur 150 000 brins d'ADN. L'ADN a ensuite été lyophilisé et envoyé par avion en Allemagne, où il a été réhydraté et séquencé en deux semaines. Nous avons retrouvé l'intégralité des données, sauf celles contenues dans deux fragments de 25 nucléotides. Ces deux fragments appartenaient à une région répétitive, ce qui nous a permis de les reconstituer manuellement et d'atteindre 100 pour cent de réussite (contre 99,999 pour cent sans ces deux fragments). Les deux fragments avaient chacun la particularité d'avoir des séquences leur permettant de se replier sur eux-même, ce qui a empêché leur séquençage.

Un intérêt pour le stockage à long terme

Nous avons ainsi réussi à stocker de l'information sur de l'ADN et à la restituer selon un procédé déclinable pour toute donnée numérique. Bien sûr, notre technique a ses limites. Avant tout, même s'ils seront sans doute réduits à quelques heures à l'avenir, les temps d'accès, d'écriture et de lecture sont moins bons que ceux d'une clef USB. L'ADN ne peut constituer un support compétitif que pour l'archivage de données à long terme. De plus, le volume de données testé était assez petit – 739 kilooctets, l'équivalent d'une disquette –, mais en théorie, il est possible de passer à une échelle beaucoup plus grande. De même, le procédé reste cher – 7 000 euros par mégaoctet –, mais ce coût diminuera vite si les tendances actuelles se poursuivent.

La technique d'encodage peut encore être améliorée et optimisée. Notamment, le passage à des données plus importantes nécessitera une automatisation et une miniaturisation plus poussées. D'ici quelques années, pour des utilisations sur le long terme, il pourrait devenir plus économique d'investir dans un support certes plus cher à l'achat, mais avec très peu de frais de maintenance, l'ADN, que d'opter pour un support bon marché tel que les bandes magnétiques, mais qui doit être relu tous les cinq ans. ■

L'ADN

mémoire numérique du vivant

Vincent Daubin, Simon Penel et Éric Tannier

L'ADN, un code à quatre lettres, est une source d'information sur l'histoire du monde vivant. En comparant les génomes actuels, on reconstitue les gènes d'espèces disparues il y a plusieurs centaines de millions d'années.

L'ESSENTIEL

- Par les mécanismes de l'hérédité, les génomes actuels recèlent des informations sur l'évolution du vivant depuis plusieurs milliards d'années.
- L'information est brouillée par le jeu des mutations, de la sélection naturelle et du transfert de gènes entre espèces.
- Néanmoins, en modélisant l'histoire des gènes, les phylogénéticiens ont accès à des informations sur des espèces dont on n'a plus trace aujourd'hui.



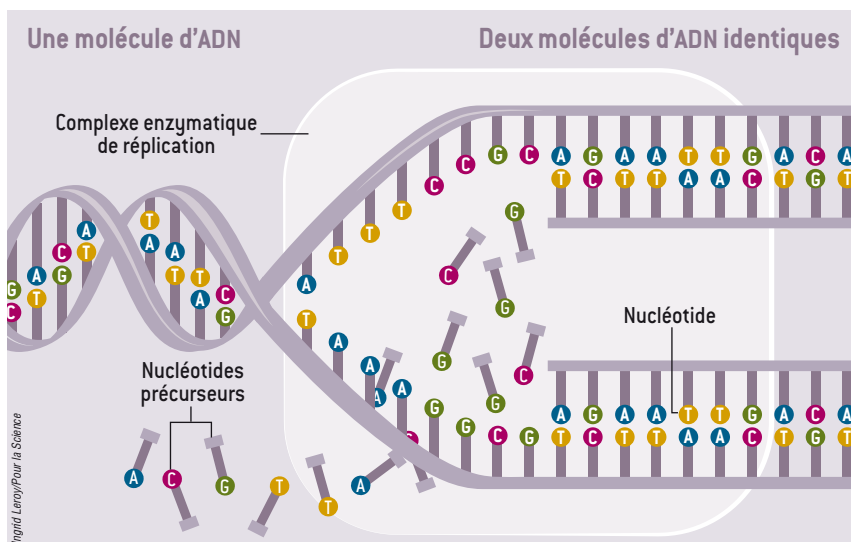
Depuis plus de trois milliards d'années, des informations numériques écrites en langage ADN sont copiées, plus ou moins fidèlement, puis interprétées, conduisant à la formation de nouvelles générations d'organismes. À ce titre, la théorie de l'évolution est une théorie de l'information, analogue à celle qu'on utilise en mathématiques depuis sa formalisation en 1948 par le mathématicien américain Claude Shannon : un ensemble de lois décrivant comment un objet – un texte dans le cas de la théorie mathématique, un organisme dans le cas de la reproduction – peut être reconstitué à partir d'une information sur cet objet, en général codée de façon plus compacte que l'objet lui-même. La biodiversité actuelle résulte de ce processus de duplication – ou réplication – avec erreurs des données génétiques héritées d'une multitude d'ancêtres.

Dans les génomes des organismes actuels figurent les traces de ces milliards d'années de transmission d'informations codées génétiquement. Aussi, chaque lignée ayant enregistré des événements différents depuis sa séparation des autres, il est possible, en comparant ces génomes, de reconstruire l'histoire du vivant. Depuis 50 ans, les phylogénéticiens, qui étudient les relations évolutives entre les organismes, comparent les séquences ADN des espèces actuelles. Avec les progrès du séquençage génomique, ces recherches ont changé d'échelle. Elles permettent aujourd'hui de comprendre à la fois les mécanismes de l'évolution et l'origine des espèces. Quelles informations anciennes peut-on espérer décrypter dans les génomes actuels ? Quelles sont les méthodes employées ? Quelles sont les limites de cette approche ? L'analogie avec la théorie de l'information apporte plusieurs réponses.

L'hérédité : une théorie de l'information

Si, en semant des graines de tomate, on obtient des tomates et non des courgettes, c'est parce que dans les graines figure une information qui dicte la forme de la plante. La nature de cette information, celle qui est transmise des parents à leur progéniture dans tout le monde vivant lors de la reproduction, a été longuement débattue. Selon la théorie de la préformation, par exemple, dont on trouve des traces dès l'Antiquité, chaque graine contenait

© Shutterstock/Tatiana53



1. LE MÉCANISME qui confère sa pérennité à l'information codée dans le génome est la réplication de l'ADN, orchestrée par un complexe protéique, le complexe enzymatique de réplication. Les deux brins de la molécule d'ADN sont séparés et, sur chaque brin, sont recrutés les nucléotides complémentaires de la séquence lue par le complexe (A et C sont respectivement les complémentaires de T et G). On obtient ainsi deux nouvelles molécules d'ADN identiques.

l'intégralité du futur organisme, déjà formé, mais en plus petit – une sorte de modèle réduit de l'organisme. Mais pour la théorie de l'épigenèse, sa concurrente depuis l'Antiquité jusqu'au XVIII^e siècle, les graines constituaient un plan de l'organisme, que le développement allait réaliser selon des lois d'organisation qui restaient à découvrir. L'épigenèse était ainsi déjà presque une théorie de l'information au sens de Shannon.

Au début du XX^e siècle, la génétique a donné à l'hérédité son caractère numérique: le gène – qui n'était pas encore identifié comme un segment d'ADN – était un objet abstrait que l'on pouvait observer dans un nombre fini d'états disjoints, les allèles, codant les caractéristiques héréditaires des organismes. La découverte, en 1953, de la structure moléculaire du support de l'information héréditaire a clarifié le concept de gène: le plan des organismes contenu dans les graines est un code numérique porté par une molécule, l'ADN.

Cette molécule est un assemblage linéaire de petites molécules – des nucléotides – de quatre types: adénine (A), guanine (G), cytosine (C) et thymine (T). Cet assemblage porte une information numérique: comme les mots dans une écriture alphabétique ou les séquences de 0 et de 1 qui codent les données et les programmes dans les ordinateurs, l'ADN forme un texte dont les mots sont les gènes, écrits dans

un alphabet de quatre lettres A, G, C et T, correspondant aux quatre nucléotides. Ce texte code le plan de l'organisme tel que le prévoyait l'épigenèse.

La réplication, un excellent moyen de conservation

Portée par l'ADN, une information numérique est donc transmise, de génération en génération, lors de la reproduction. Cette information, que l'on nomme le génome, est contenue dans toutes les cellules de tous les êtres vivants. Elle se propage par la réplication de l'ADN, qui le rend pérenne (voir la figure 1). En effet, le support lui-même, comme les multiples supports de l'information numérique en informatique (carte perforée, cassette, cartouche, disquette, disque, mémoire flash), n'est pas conservé indéfiniment. L'ADN d'une cellule qui meurt se dégrade vite, notamment parce qu'il est assimilé par d'autres organismes.

Dans des conditions exceptionnelles de conservation, on peut retrouver une partie de l'information qu'il contient après quelques milliers d'années, difficilement après quelques dizaines de milliers d'années. Ce temps de conservation peut paraître long à l'échelle de nos vies, mais la réplication, elle, permet une préservation de l'information à une tout autre échelle: depuis plus de trois milliards d'années elle permet la copie des génomes et, en quelque sorte, la conservation de cette information dans les organismes vivants.

Grâce à l'ADN et à sa réplication, l'information contenue dans les génomes se propage, se transmet, se copie, se pérennise. Les organismes vivants, en se reproduisant, permettent cette transmission. En informatique aussi, la copie et la transmission de l'information sont un moyen de conservation des données pour faire face aux défaillances possibles des supports. Néanmoins, toute copie, toute transmission *via* un support, provoque des erreurs. Aucun support ni aucun canal de transmission n'est totalement sûr.

Le codage et la transmission de l'information numérique afin de minimiser ces erreurs sont l'objet de la théorie des codes. Le principe est d'introduire de façon méticuleuse et parcimonieuse une « redondance » dans l'information, afin de détecter

Lexique

Aufhebung : Ce terme allemand cher à Hegel désigne, en biologie, la dualité de la mémoire génétique du monde vivant, à la fois préservée et peu à peu détruite par les mutations.

Dérive génétique : Évolution d'une population au hasard de mutations sans effet sur la reproduction, et donc non soumises à une pression de sélection.

Épigenèse : Théorie selon laquelle les graines constituent un plan de l'organisme et non une copie complète miniature de l'organisme parent.

Épigénétique : Ensemble des modifications de l'expression des gènes qui ne sont pas codées dans le génome.

Phylogénie : Généalogie du vivant fondée sur l'étude des caractères anatomiques et moléculaires des organismes vivants.

et de corriger les erreurs. De même, la réplication de l'ADN ne produit jamais des copies parfaites de l'information. La structure de l'ADN et des mécanismes de réparation permettent de corriger une partie des erreurs, sur le principe de la redondance. L'ADN est constitué de deux brins enroulés en une double hélice. Indispensable à la réplication, cette structure en double brin est aussi une façon de conserver l'information en deux copies identiques. Chaque nucléotide a un complémentaire, auquel il est apparié dans la double hélice : adénine est complémentaire de thymine, et guanine de cytosine. En conséquence, l'information est présente deux fois – sur une séquence de nucléotides et sur son complémentaire.

Lors de la réplication, il arrive que les mécanismes dérapent et que dans le brin copié, un nucléotide soit perdu ou inséré, ou pris pour un autre. Il en résulte un « mésappariement », c'est-à-dire la mise en correspondance d'un nucléotide avec un autre qui n'est plus son complémentaire. La plupart du temps, les erreurs ainsi détectées sont réparées en rétablissant le nucléotide complémentaire.

Mais le mécanisme de correction a ses limites, et les génomes répliqués contiennent toujours des différences par rapport à leurs ancêtres. D'abord parce que la fidélité a un coût pour l'organisme, en énergie et en temps passé à fabriquer la génération suivante. Ces erreurs constituent ce que l'on nomme les mutations. Il en existe de multiples possibles, de toutes les formes

et de toutes les tailles : remplacement d'un nucléotide par un autre, insertion ou perte d'un nucléotide ou d'un segment de nucléotides, duplication ou déplacement d'un segment, duplication du génome tout entier. Au moins une centaine de mutations séparent chez l'homme les génomes des parents et des enfants, qui contiennent environ trois milliards de nucléotides. La seconde raison pour laquelle les systèmes vivants ne sont pas parfaitement fidèles, c'est parce que dans ces erreurs figurent parfois les clés de leur survie.

Une information qui évolue est-elle préservée ?

Aufhebung, en allemand, signifie à la fois préservation et destruction. C'est un de ces mots à double sens que le philosophe Georg Wilhelm Friedrich Hegel (1770-1831) affectionnait pour évoquer la dialectique, et c'est le mot qu'ont repris, en 1965, le biologiste austro-américain Emile Zuckerkandl et le chimiste américain Linus Pauling pour décrire la conservation de la mémoire du monde vivant par les molécules : préservation infinie en théorie grâce à la réplication, mais aussi destruction progressive par les mutations.

La plupart des mutations dans les génomes ne se propagent pas aux générations suivantes. D'une part, la sélection naturelle fait un tri. Si une mutation modifie l'efficacité de la reproduction dans un

■ LES AUTEURS



Vincent DAUBIN est directeur de recherche CNRS au Laboratoire Biométrie et biologie évolutive (UMR CNRS 5558) de l'Université Claude Bernard Lyon I.



Simon PENEL est ingénieur d'étude CNRS dans le même laboratoire.



Éric TANNIER est chargé de recherche de l'Inria dans le même laboratoire.

Ils sont tous trois membres du projet ANCESTROME qui vise à reconstruire les caractéristiques des organismes disparus en analysant leurs génomes avec des méthodes phylogénétiques (ancestrume.univ-lyon1.fr).

L'ADN et la métaphore informatique, comparaisons et raisons

Voir l'ADN comme une mémoire numérique pousse à l'analogie entre le monde vivant et les ordinateurs. L'ADN est le support d'une information numérique – le génome –, tout comme un disque dur est le support du logiciel ou des données. La métaphore peut être poussée loin, puisque les génomes sont des chaînes de caractères, à l'instar des données informatisées, suites de 0 et de 1. Le génome peut être vu comme un programme contenant les instructions nécessaires à la fabrication et à la reproduction d'un organisme, et le code génétique, qui permet la traduction des séquences génomiques en pro-

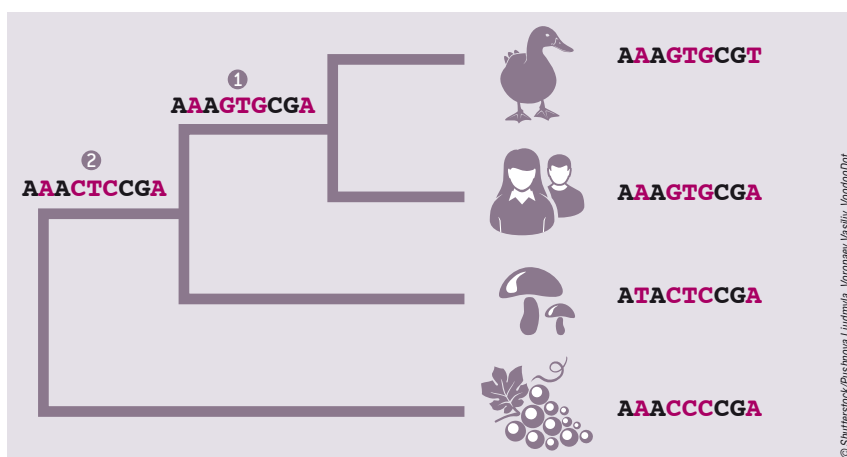
téines, est une sorte de moyen d'interpréter ces instructions.

On peut aller encore plus loin. L'évolution a été qualifiée de bricolage par le biologiste François Jacob. Or la conception d'un programme informatique pourrait souvent, sans mettre en cause le professionnalisme des programmeurs, supporter cette qualification : un logiciel évolue toujours au fil du temps, chaque programmeur reprenant des morceaux de code de ses prédécesseurs en les amendant ou en intégrant des modules extérieurs, tels des transferts horizontaux. L'évolution des données numériques est ainsi comparable à celle des génomes.

Le génome a toutefois un statut particulier. Il constitue une sorte de programme autonome, puisqu'il code aussi la fabrication de la machine qui sera capable d'exécuter le programme et de se dupliquer. Là, on dépasse les possibilités des ordinateurs et des programmes informatiques (ce qui sera peut-être démenti sous peu par les imprimantes 3D autorépliquatives *RepRap*). De plus, le génome ne porte pas tout le matériel héréditaire : une part importante d'information est transmise lors de la reproduction, en dehors de la séquence, par « épigénétique ». Faute de bon modèle d'évolution intégrant ces caractères, la métaphore informa-

tique s'arrête à l'information numérique, c'est-à-dire au génome.

On a les métaphores de son temps. Au XVIII^e siècle, les machines les plus perfectionnées étaient les automates, et la métaphore a beaucoup été utilisée pour décrire de façon mécaniste le fonctionnement du corps humain. On a aussi beaucoup comparé l'ADN à un langage, à une carte ou à un atlas. Aujourd'hui, la métaphore informatique est facilitée, d'une part, par l'utilisation constante des ordinateurs pour stocker les génomes et, d'autre part, par l'idée qu'il serait possible d'utiliser l'ADN pour le stockage de données informatiques.



© Shutterstock/Pushnova Ludmila, Voropaev Vasily, VoodooDot

2. SUR CET ARBRE PHYLOGÉNÉTIQUE SIMPLIFIÉ, on peut reconstruire, à partir des séquences d'un même gène repérées chez le canard, l'humain, le cèpe et le raisin, les séquences du même gène de l'ancêtre commun de l'humain et du canard d'une part (1), et de l'ancêtre des animaux et des champignons d'autre part (2). On suppose que les ressemblances sont dues à un héritage commun plutôt qu'à des événements indépendants. Les séquences les plus proches sont celles du canard et de l'humain (une seule mutation). La séquence du même gène chez leur ancêtre commun sera donc AAAGTTCGA ou AAAGTTCGT. Les séquences du cèpe et du raisin portant un A à la fin, la première option est sans doute la bonne. La séquence de l'ancêtre commun du cèpe, du canard et de l'humain est celle qui minimise le nombre de mutations entre cet ancêtre et le dernier parent dans l'arbre, le raisin, soit AAACCCGA.

BIBLIOGRAPHIE

G. J. Szöllösi *et al.*, Lateral gene transfer from the dead, *Systematic Biology*, vol. 62, pp. 386-397, 2013.

J.-G. Dumas *et al.*, Théorie des codes, Dunod, 2013.

G. J. Szöllösi *et al.*, Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations, *PNAS*, vol. 109, pp. 17513-17518, 2012.

S. Abby *et al.*, Lateral gene transfer as a support for the tree of life, *PNAS*, vol. 109, pp. 4962-4967, 2012.

B. Boussau *et al.*, Parallel adaptations to high temperatures in the Archaean eon, *Nature*, vol. 456, pp. 942-945, 2008.

E. Zuckerkandl et L. Pauling, Molecules as documents of evolutionary history, *Journal of Theoretical Biology*, vol. 8, pp. 357-366, 1965.

sens ou dans l'autre, alors sa fréquence augmente ou diminue dans une population au fil des générations. D'autre part, il n'est pas systématique qu'une mutation induise un avantage ou un inconvénient pour la reproduction. Dans certains génomes, par exemple ceux des mammifères ou de certaines plantes qui sont de longs textes verbeux, la plupart des mutations sont sans effet immédiat. Dans ce cas, la fréquence des génomes qui portent une mutation suit une marche aléatoire sans pression de l'environnement. C'est le principe de la dérive génétique.

Que cette variation soit influencée ou non par la sélection naturelle, la fréquence finit toujours par atteindre 0 ou 1. Si elle atteint 0, la mutation a disparu de la population ; il n'en reste pas de trace dans les génomes. Si elle atteint 1, tous les génomes de la population sont porteurs de la mutation. Dès lors, il ne reste pas de trace dans les génomes de leur état ancestral, avant la mutation.

Avec un nombre réaliste de mutations par génération (une centaine), la dérive remplace la plupart des lettres du texte génomique en quelques millions de générations. Ainsi, on ne peut retrouver des traces que de génomes ayant au plus quelques millions d'années – ce qui représente une

partie infime de l'histoire du monde vivant. En revanche, la sélection permet une meilleure conservation de l'information : on retrouve dans les génomes actuels des traces de l'origine du vivant, il y a presque quatre milliards d'années.

Pour reconstituer l'information ancienne, on compare plusieurs génomes actuels dans un cadre phylogénétique. Par exemple, prenons les génomes d'une variété de vigne, d'un cèpe, d'un canard et d'un humain (voir la figure 2). Les ressemblances entre séquences permettent de reconstituer les relations de parenté entre ces espèces, ce qui se traduit par un arbre phylogénétique, c'est-à-dire une généalogie des espèces. Cette généalogie implique que le canard et l'humain ont un ancêtre commun qui leur est exclusif et qui descend lui-même de l'ancêtre qu'ils partagent avec le cèpe. On identifie alors des morceaux (des gènes) suffisamment similaires pour que l'on suppose qu'ils descendent d'une même séquence ancestrale : on trouve ATACTCCGA dans la vigne, AAACCCCGA chez le cèpe, AAAGTTCGA chez le canard et AAAGTTCGT chez l'humain.

À partir de l'analyse de ces séquences actuelles d'un même gène, on reconstitue les séquences du gène chez les ancêtres communs des espèces considérées en supposant que les ressemblances entre espèces sont plus fréquemment dues à leur héritage commun qu'à des événements indépendants qui auraient donné le même résultat : dans notre exemple, cette hypothèse se traduit par un scénario unique où chacune des positions variables (en rouge) n'a changé qu'une seule fois au cours de l'évolution des espèces.

En multipliant le nombre de gènes et de génomes comparés (aujourd'hui on peut avoir accès aux séquences de plusieurs milliers de génomes, chacun contenant des milliers de gènes, voir l'article de G. Perrière dans ce numéro) et avec l'aide de modèles plus réalistes prenant en compte nos connaissances des processus biologiques, on peut ainsi estimer des séquences de génomes ancestraux et décrire l'histoire des mutations.

En pratique, on sait aujourd'hui reconstituer les séquences quasi complètes d'ancêtres récents, par exemple le génome de la bactérie qui fut probablement responsable des grandes épidémies de peste du Moyen Âge, à l'aide du séquençage de ses descendantes, des

souches de *Yersinia pestis* contemporaines. Pour des ancêtres plus éloignés, tel celui des mammifères qui vivait il y a une centaine de millions d'années, une partie de l'information a été perdue par dérive. Cependant, pour les parties du génome contraintes par la sélection, les gènes en particulier, la reconstruction reste tout à fait possible. On peut même connaître le nombre de chromosomes et la façon dont les gènes étaient organisés dans ces génomes. À partir de ces données, on commence à comprendre les modes de vie des espèces disparues (température optimale de croissance, profondeur préférée d'anciens mammifères marins, utilisation de l'oxygène par des bactéries anciennes, alimentation, taille ou longévité de mammifères anciens, histoire des migrations et mélanges humains...), un peu comme si l'on retrouvait le code d'un vieux programme et que l'on reconstituait son fonctionnement à partir de ses versions nouvelles (voir la figure 3).

Cependant, l'application de ces méthodes à l'ensemble du monde vivant

se heurte à une difficulté : nombre d'organismes ont la capacité d'intégrer dans leur génome des gènes de leurs contemporains, proches apparentés ou non. Ces transferts horizontaux violent les hypothèses de reconstruction des génomes ancestraux, qui supposent que l'ancêtre commun des séquences appartient à l'ancêtre commun des espèces.

Le transfert de gène : obstacle ou indice ?

Il est apparu que chaque gène, quelle que soit sa fonction, pouvait, à un moment ou un autre de son histoire, avoir été transféré. À long terme, aucun gène n'a donc la même histoire que les organismes qui le portent. Ce constat semblait remettre en cause la possibilité théorique de reconstruire l'évolution du vivant à partir des gènes.

Ainsi, le gène était trop « individualiste » pour refléter l'histoire des organismes. N'avait-il pour autant rien à en raconter ? De fait, lorsqu'on observe l'ensemble des

gènes et, pour peu que l'on sache interpréter les singularités que révèle chaque histoire de gènes, on s'aperçoit que les transferts horizontaux contiennent une information inédite sur l'histoire du vivant. D'une part, ils peuvent venir à l'appui d'une hypothèse phylogénétique : si un ancêtre a reçu un gène par transfert, ce gène doit être observé chez ses descendants. Ainsi, la présence du gène peut être un caractère partagé permettant de reconnaître les organismes qui sont proches.

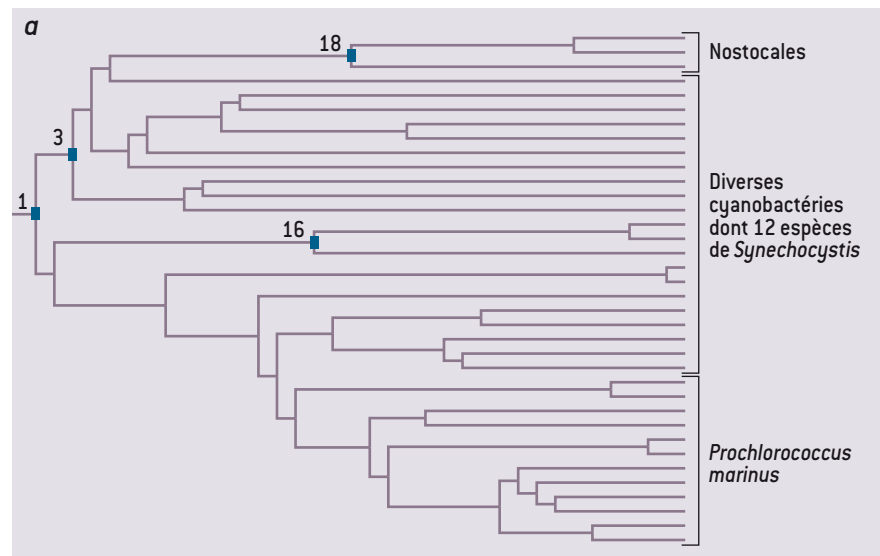
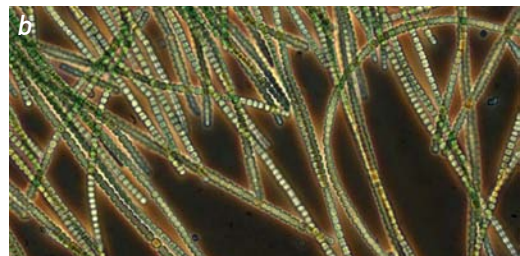
D'autre part, un échange de gènes entre deux organismes ne peut avoir lieu que s'ils sont contemporains. Le transfert horizontal contient donc une information sur les âges relatifs des lignées qui échangent des gènes : on ignore à quelle époque exactement vivaient les organismes anciens, mais, avec des relations de descendance et des échanges entre espèces contemporaines, on peut espérer établir une chronologie de l'histoire de la vie. Or les transferts ont lieu principalement chez les organismes microscopiques (bactéries, virus, champignons unicellulaires, autres

TRANSFERT DE GÈNES ET DATATION

Par définition, le transfert de gènes entre deux espèces n'est possible que si les deux espèces sont contemporaines. Ce processus, qui transparaît dans le génome des espèces actuelles, donne aux phylogénéticiens de précieuses informations sur les âges relatifs d'espèces anciennes dont les traces fossiles sont rares. En 2012, à partir de 36 génomes d'espèces différentes, les auteurs ont ainsi reconstitué la phylogénie chronologique des cyanobactéries, les premiers organismes à avoir fait de la photosynthèse et produit de l'oxygène en grande quantité (a). Les plantes ne récupèrent cette faculté en intégrant des cyanobactéries au fonctionnement de leurs cellules.

Les nœuds de l'arbre sont positionnés de gauche à droite selon leur chronologie. Le nœud 1, le plus ancien, est parfois daté à 3,8 milliards d'années. Les plantes ne figurent pas sur cet arbre, mais on voit la naissance au nœud 18 du clade des cyanobactéries multicellulaires, les nostocales, qui forment des filaments verts de cellules différenciées (b). Comme ce clade a reçu, par transfert, des gènes d'autres mor-

ceaux de la phylogénie, on estime que cette forme de multicellularité est relativement récente. De même, la présence de nombreux transferts du clade du nœud 3 vers le clade du nœud 16 lui donne une origine relativement récente, alors que les hypothèses habituelles plaçaient cette diversification à la racine des cyanobactéries.



Adapté de G. J. Szallasi et al., PNAS, vol. 109, pp. 17513-17518, 2012

organismes unicellulaires tels que l'amibe ou la paramécie). Ces organismes ne laissent presque aucune trace fossile, ce qui rend la datation des trois quarts de l'histoire de la vie inaccessible aux méthodes classiques. Le transfert fournit ainsi l'outil de datation dont la petite taille de ces organismes nous privait.

Le transfert recèle encore d'autres traces inexploitées de l'histoire de la vie. Parmi les êtres vivants anciens, certains ont laissé des descendants, d'autres non. Il est raisonnable de penser que ceux qui n'ont laissé aucun descendant sont bien plus nombreux. Pourtant, on reconstitue l'histoire à partir des génomes actuels – les descendants de ceux qui, par définition, ont laissé des descendants ! On n'a ainsi accès qu'à une faible partie de la biodiversité ancienne.

Et si l'on détecte un échange de gènes entre deux organismes anciens, c'est à partir des descendants de celui qui a reçu le gène. En revanche, il n'y a aucune raison pour que celui qui l'a donné, lui, fasse partie de la minorité qui a eu des descendants. Statistiquement, il est probable qu'on ait ainsi accès à un gène d'une espèce ancienne sans descendant, qui n'aura donc pas transmis ses gènes en dehors de ce transfert. On est donc face à une biodiversité beaucoup plus importante que celle que nous offrirait la seule transmission d'ascendant à descendant.

Les gènes ne sont pas les marqueurs dont on avait rêvé pour reconstruire l'histoire de la vie. Ils racontent parfois une autre histoire, la leur. Mais si on sait modéliser la complexité de ces histoires, ils permettent de trouver, dans les génomes,

des informations dont on pensait avoir perdu la trace. Il reste sans doute encore beaucoup à découvrir, dans nos cellules, sur notre passé.

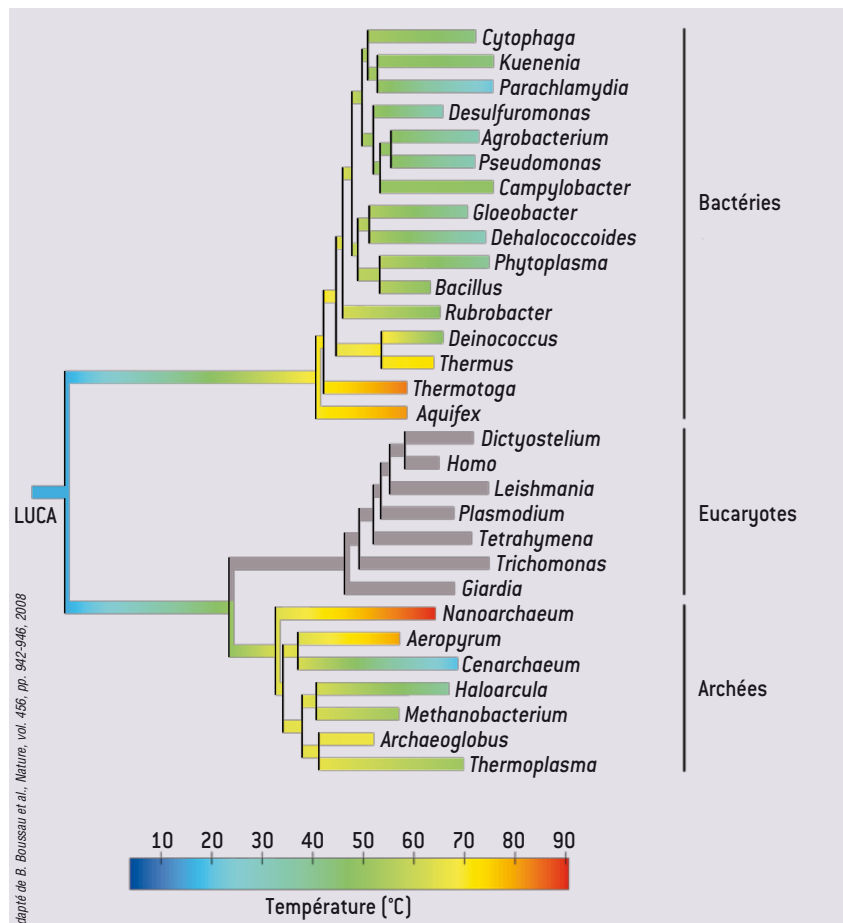
D'autres supports que l'ADN ?

En informatique, les supports de l'information numérique ont beaucoup évolué depuis les cartes perforées. Ils sont encore très divers aujourd'hui. Dans le monde vivant, il existe également plusieurs types de supports, mais c'est la double hélice d'ADN qui s'est établie comme le support universel chez les organismes cellulaires. Pourtant, ce support n'aurait pas été le premier.

L'ARN, toujours utilisé aujourd'hui par certains virus et conservé comme intermédiaire pour la lecture de l'information par les cellules, pourrait bien l'avoir précédé. Il avait l'avantage d'être capable de catalyser des réactions enzymatiques, offrant la possibilité d'organismes primitifs assez simples, dont le support d'information génétique est en même temps actif biochimiquement. Mais cette activité est également un facteur d'instabilité, et les organismes de ce monde à ARN auraient finalement adopté un média dédié pour préserver leur précieuse information génétique.

Cette transition est comparable à l'invention de l'imprimerie qui, plus rapide pour copier l'information écrite et moins sujette à l'erreur, a remplacé le scribe. Ou, en informatique, à l'abandon des cartes perforées qui stockaient l'information numérique lisible par les ordinateurs, pour des supports électromagnétiques.

Aujourd'hui, l'ensemble du vivant a adopté l'ADN comme média de stockage et de transmission de l'information numérique. Même l'informatique envisage d'effectuer cette transition : l'ADN pourrait remplacer les supports magnétiques pour stocker l'information numérique et être utilisé par les ordinateurs à la place des disques durs (voir l'article de Ch. Desimoz dans ce numéro). Cette innovation pourrait avoir des conséquences sur le monde vivant lui-même : pour des organismes habitués à récupérer l'ADN dans l'environnement par transfert horizontal, les grandes quantités d'ADN présentes dans des machines constitueraient un substrat pour inventer de nouveaux gènes. L'informatique elle-même changerait alors le cours de l'évolution ! ■



3. RECONSTITUTION DE LA TEMPÉRATURE OPTIMALE DE CROISSANCE de l'ancêtre commun à tout le vivant (LUCA) à partir de l'analyse phylogénétique d'organismes actuels (65 eucaryotes, 60 archées et 331 bactéries). En se fondant sur l'observation selon laquelle chez les procaryotes (archées et bactéries), la température optimale de croissance est corrélée à la proportion de nucléotides C et G dans le génome, des chercheurs du laboratoire des auteurs ont décrit une évolution de la température en deux temps : LUCA vivait à environ 20 °C, puis la tolérance thermique de ses descendants a augmenté, avant de diminuer pour les espèces les plus récentes.

COLLECTION **SCIENCE** LES PETITS THÉMATIQUES

Spécial T.I.P.E. 2013-2014

Retrouvez tous les articles de référence
parus dans *Pour la Science* sur
le thème **Transfert-Échanges**

Déclinés en biologie, physique, informatique,
archéologie et environnement, ces cinq
feuiltables numériques vous aideront
pour préparer votre T.I.P.E. !



3,99€
35 p.
environ

Des dizaines d'autres sujets à découvrir sur
www.pourlascience.fr/petitsthematiques

La mémoire biologique est-elle codée ?

Franck Chaillan

La mémoire biologique repose sur des mécanismes beaucoup plus élaborés que le codage binaire des mémoires numériques. Elle est efficace, pertinente, adaptée au contexte et capable d'oubli. Mais, malgré la plasticité cérébrale, sa capacité de stockage est probablement limitée.

Aujourd'hui, c'est mon anniversaire. Ce matin, je suis venu au laboratoire en moto, en passant par le bord de mer, puis en empruntant l'autoroute. Quand je conduis, dès que j'aperçois la moindre petite lumière rouge à l'arrière d'une voiture, je ralentis pour éviter tout freinage d'urgence. En revanche, je ne prête plus attention à la haute tour qui domine le port. Pour dépasser rapidement la voiture qui précède, je regarde dans mes rétroviseurs, je coupe les gaz, j'embraye, je passe de la 6^e à la 5^e vitesse, je relâche l'embrayage, j'accélère et je me penche sur la gauche pour déboîter, tout cela automatiquement. Enfin, arrivé au laboratoire, je fais une recherche bibliographique, je rassemble mes connaissances et réfléchis à la façon de les organiser pour rédiger cet article sur la mémoire.

Pourquoi vous ai-je raconté cela ? Si ce court récit ne vous a guère intéressé (et c'est tout à fait normal !), il illustre la place que prend la mémoire dans notre quotidien et ses innombrables formes. En quelques lignes, j'ai fait appel à la mémoire épisodique, celle qui me concerne, en évoquant mon anniversaire et pour rédiger cette introduction ; la mémoire spatiale en décrivant mon trajet ; la mémoire non

déclarative qui m'aide à ne pas faire attention aux événements ou objets habituels (la tour) ; la mémoire non déclarative procédurale, celle que j'utilise pour conduire, dépasser, embrayer, passer les vitesses sans que j'aie à y réfléchir ; la mémoire déclarative sémantique lors de ma recherche bibliographique et ma mémoire de travail pour organiser mes idées.

Les prémices de la plasticité

La mémoire biologique, qui permet le stockage des informations, est multiple. Elle est étroitement liée à l'apprentissage, c'est-à-dire à l'acquisition d'informations qui permettent à un individu d'adapter son comportement à l'environnement. Dès 1890, le psychophysiologiste américain William James proposa l'existence d'une mémoire primaire qui maintient temporairement une information dans le cours de la pensée, et d'une mémoire secondaire qui stocke cette information à long terme.

En 1911, le psychologue américain Edward Thorndike émit l'idée que la plasticité du système nerveux permet la création des traces mnésiques (ou souvenirs), lesquelles sont associées à des

comportements. Aujourd'hui, le rôle de la plasticité cérébrale dans la formation, le stockage, le rappel et la reconstruction des souvenirs, leur éventuelle évolution et l'oubli est avéré. Nous examinerons ici en quoi consistent la plasticité cérébrale et son rôle dans la mémoire. Puis nous comparerons les performances de la mémoire biologique et des mémoires numériques, et nous nous interrogerons sur la façon dont les secondes peuvent perturber la première.

Commençons par rappeler que la mémoire est à court terme ou à long terme. La mémoire de travail est dite à court terme, car elle maintient actives durant quelques secondes ou minutes des informations nécessaires à un raisonnement ou à une opération mentale. Dès que l'opération est terminée, les données présentes dans la mémoire de travail sont oubliées ou stockées dans une mémoire à long terme. Cette dernière correspond au stockage des souvenirs. Elle se subdivise en une mémoire déclarative et une mémoire non déclarative. La mémoire sémantique, celle des connaissances que l'on a sur le monde, et la mémoire épisodique, celle des événements qui nous sont propres, sont des mémoires déclaratives. La mémoire procédurale, celle



L'ESSENTIEL

- La mémoire biologique repose sur une chaîne d'acteurs, des plus élémentaires (diverses molécules) aux plus complexes, de vastes réseaux de neurones activés simultanément.
- La plasticité cérébrale est la clé de la mémoire biologique.
- La mémoire numérique n'est que la somme de toutes les connaissances humaines. Elle ne crée pas – pour l'instant – de savoir nouveau, contrairement à la mémoire biologique.

des apprentissages moteurs, des automatismes, la mémoire des apprentissages non associatifs (ne plus réagir à un stimulus, par exemple la sonnerie de l'horloge), ou encore le conditionnement classique (de Pavlov), sont des mémoires non déclaratives. Et cela n'est qu'un panorama superficiel de « l'anatomie de la mémoire ». Cette rapide description macroscopique posée, que s'est-il passé quand j'ai appris à conduire une moto et quand, aujourd'hui, je conduis ? Pour le comprendre, nous devons nous placer au niveau des neurones, les supports de nos pensées et de nos souvenirs. Rappelons d'abord comment fonctionne un neurone et comment il communique avec ses voisins.

Tout est une question de communication entre neurones

Un neurone se compose de dendrites qui captent les informations, d'un corps cellulaire qui les intègre et d'un axone qui les transmet vers d'autres neurones. Les informations qu'il véhicule circulent des dendrites vers les terminaisons des axones. Ces informations sont des signaux électriques résultant des différences de charges électriques régnant de part et d'autre de la membrane neuronale. En effet, la répartition des ions négatifs (chlore, Cl^- , et protéines chargées négativement) et celle des ions positifs (sodium, Na^+ , potassium, K^+ et calcium, Ca^{++}) diffèrent entre l'intérieur et l'extérieur des neurones.

Le milieu intracellulaire est riche en ions potassium et en protéines, les ions sodium et chlore étant plutôt concentrés à l'extérieur. L'intérieur d'un neurone contient davantage de charges négatives que l'extérieur. Il en résulte une différence de potentiel de l'ordre de -65 millivolts, valeur qui varie en fonction du type de neurones et de l'espèce animale. Cette différence de potentiel est nommée potentiel de repos. Quand une telle différence de potentiel règne de part et d'autre de la membrane neuronale, le neurone est dit polarisé. Il est inactif et ne véhicule aucune information, aucun potentiel d'action.

Mais cette différence de potentiel varie quand des ions traversent la membrane du neurone (y entrent ou en sortent) par différents types de canaux (voir l'encadré pages 112 et 113). Prenons l'exemple de la propagation d'une information entre deux