Archivage sur l'ADN des données numériques de l'agronomie

FICHE *QUESTIONS SUR…* n° 06.06.Q01

Mots clés : mégadonnée numérique - *big data* - archivage - ADN - polymère - agronomie

L'information est le moteur de la croissance socio-économique de la civilisation depuis ses débuts. Aujourd'hui, les activités agronomiques dépendent de manière croissante de l'accumulation et de l'usage massif de données numériques. Mais, selon les prévisions, le volume de ces dernières va atteindre des niveaux tels que leur stockage, leur archivage et leur traitement dans les centres actuels dédiés va rapidement atteindre ses limites. Sans parler de l'impact environnemental.

Face à cette situation, cette fiche explore une alternative prometteuse : l'archivage des données au moyen de molécules codantes comme l'ADN. Un chantier de 20 ans !

L'AMPLEUR DU DÉFI

Où en sommes-nous ?

L'agronomie a vécu durant les dernières décennies une véritable transition numérique qui s'est manifestée à toutes les étapes du cycle de vie des végétaux et des animaux d'élevage : sélection génétique, pilotage des cultures ou du cheptel, prévision de récoltes et process de transformation… Cette transition technologique s'appuie principalement sur l'usage de modélisations mathématiques et de très grandes quantités de données maintenant disponibles grâce aux nombreux capteurs installés, aux capacités de l'imagerie, notamment satellitaire, et aux bases de données existantes.

L'application de nouvelles méthodes numériques optimisant le cycle de vie et la production agronomiques a ainsi permis de développer de nombreux outils d'analyse et d'aide à la décision. Toutes ces évolutions ont généré une augmentation exponentielle de la quantité de données numériques à stocker et à exploiter.

Bien entendu, l'agriculture n'est pas le seul secteur concerné par l'explosion du volume des mégadonnées numériques (*big data*) : ces dernières incluent également nos connexions familiales, amicales et professionnelles, nos livres, vidéos et photos, nos données médicales, celles de la recherche scientifique, de l'industrie, des services (banques, assurances…), des transports, etc.

Résultat à ce jour

L'ensemble des données accumulées par l'humanité atteint des niveaux astronomiques. Cette gigantesque "sphère globale des données", comme on l'appelle parfois, comprend autant de caractères (une lettre, un chiffre, un symbole…) qu'il y a d'étoiles dans l'univers observable, soit quelques dizaines de milliers de milliards de milliards (cf. *Figure 1*) ! Elle est en outre en perpétuelle croissance, avec un doublement tous les 2 à 3 ans environ, soit un facteur de 100 à 1 000 en 20 ans.

Figure 1 : Ciel étoilé illustrant l'abondance des mégadonnées accumulées par l'humanité.

Il en découle que ces données sont de moins en moins conservées sur un terminal informatique local et de plus en plus dans des centres de données qui fonctionnent au sein de réseaux mondiaux de transmission. Ceci inclut le "cloud" ou "nuage" qui n'en est qu'une modalité plus virtuelle et automatisée offrant des ressources à la demande.

Les trop lourds centres de stockage actuel de données

Typiquement, un gros centre de données moderne contient un exaoctet (un milliard de milliards de caractères), un million de serveurs (des ordinateurs sans clavier ni écran), des disques durs, (ou mémoires statiques) et des bandes magnétiques.

Chacune de ces structures, abritée dans un immense hangar bâti sur un terrain de plusieurs centaines de milliers de m^2, consomme plus d'électricité qu'une ville de 100 000 habitants, dont environ 40 % pour le refroidissement des serveurs (d'où l'intérêt de positionner ces centres dans des régions froides du globe). Sans compter des milliers de tonnes de métaux chers, de plastiques issus du pétrole, des terres rares (ressources coûteuses à produire) et un investissement de quelques milliards d'euros pour une durée de vie de 20 ans. Autre inconvénient : les technologies de stockage utilisées par ces centres sont rapidement frappées d'obsolescence, tant au plan du format, du dispositif de lecture/écriture que du support, lequel nécessite d'effectuer des copies tous les 5 à 7 ans pour garantir l'intégrité des données.

Le marché global des dispositifs de stockage de données sur disques durs (ou statiques) et bandes s'élevait en 2020 à 57 milliards de dollars, en augmentation de 2 à 3 % par an. Il faut bien avoir à l'esprit qu'il existe plusieurs millions de centres de données sur la planète, quelle que soit leur taille, en incluant ceux des compagnies ; et que cette multitude de centres est reliée au reste du monde par des réseaux de connexion, qui sont eux-mêmes de gros consommateurs de ressources. Si chaque centre avait une capacité d'un exaoctet, leur nombre pourrait être réduit à environ 50 000, mais cela ferait encore beaucoup.

Cette gigantesque infrastructure couvre actuellement un millionième des terres émergées de la planète et consomme 2 à 4 % de l'électricité produite chaque année dans les pays avancés. Pour donner une image, si tous ces centres formaient un seul pays, celui-ci serait le cinquième plus gros consommateur d'électricité au monde, entre l'Inde et le Japon.

Sachant qu'en 2040 nous aurons, d'après les prévisions, 100 à 1 000 fois plus de données à stocker, il est clair que le modèle actuel de conservation deviendra largement insuffisant, tout en étant insupportable sur le plan environnemental.

DEUX PISTES POUR RELEVER LE DÉFI

Les réponses à ce redoutable défi multiforme sont de plusieurs ordres. Tout en reconnaissant l'importance de la limitation et de la compression des données et de l'éducation des usagers, nous nous limiterons ici à seulement deux pistes.

1. Distinguer les données méritant d'être conservées

Bien sûr, toutes les données agronomiques ne sont pas d'égale importance. Schématiquement, il convient d'en distinguer trois types.

- Les données d'usage instantané et d'intérêt nul dans le futur, comme celles provenant des capteurs d'un engin agricole, qui ont déterminé une réaction immédiate et unique.
- Les données intermédiaires, par exemple l'évolution temporelle de la production d'un animal de rente.
- Celles méritant un archivage de longue durée, parfois nommé "stockage froid", par exemple des données pédoclimatiques ou biogéographiques, des bases de données des semences, des ravageurs, des animaux et de leur génétique. C'est ce type de données qui sera discuté dans la suite de cette fiche.

2. Changer radicalement le modèle des centres de données

Nous avons vu plus haut que les centres dédiés au stockage et à l'archivage des données reposent sur des solutions admirables sur le plan technique, mais n'offrant plus de marges suffisantes d'optimisation pour faire face à l'avalanche prévisible des données dans un futur proche. En outre, leur impact environnemental est déjà problématique. Un chantier de 20 ans se dresse devant nous pour révolutionner ces technologies.

L'Académie des technologies s'est penchée sur cette question durant deux années et a publié un rapport en 2020. Après avoir passé en revue et écarté plusieurs approches ne présentant pas les caractéristiques

requises, son groupe de travail s'est focalisé sur les molécules porteuses d'information, telles que l'ADN (Acide DésoxyriboNucléique, support du patrimoine héréditaire) ou d'autres polymères[1] très prometteurs.

Comment coder et stocker avec l'ADN ?

Potentiellement, l'ADN – utilisé ici comme agent chimique en dehors du vivant – offre des densités de stockage informationnel dix millions de fois supérieures à celles des mémoires traditionnelles : l'actuelle *sphère globale des données* tiendrait dans une fourgonnette ! L'ADN présente en outre l'immense avantage d'être stable à température ordinaire durant plusieurs millénaires, sans consommation énergétique[2] ; il peut aussi être aisément multiplié ou détruit à volonté. En outre, l'obsolescence du support ADN ne se produira pas tant que l'homme disposera des technologies nécessaires à l'écriture et à la lecture de l'ADN, qui font partie intégrante de la médecine moderne.

Pour archiver et retrouver des données dans l'ADN, il convient d'enchaîner 5 étapes (cf. *Figure 2*) : coder le fichier de données binaires dans l'alphabet de l'ADN qui possède quatre lettres (A, C, T et G) ; puis écrire, stocker, lire l'ADN ; et enfin décoder l'information lue.

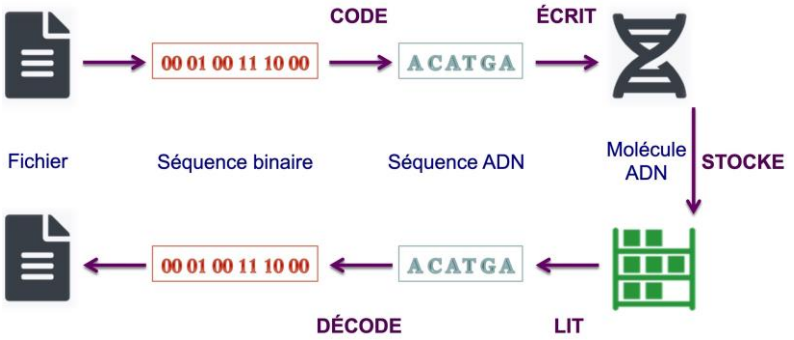

Figure 2 : Étapes du processus de stockage des mégadonnées numériques sur l'ADN. Ici sont représentés pour exemple 12 bits successifs extraits du fichier numérique. Ces 12 bits sont codés sous la forme de 6 nucléotides qui sont écrits en succession dans une molécule d'ADN. Cet ADN est ensuite stocké. Puis il est lu, et la séquence de nucléotides ainsi obtenue est décodée pour reconstituer le fichier numérique d'origine. *Crédit : François Képès.*

Un prototype réalisant ces opérations fonctionne depuis mars 2019 chez Microsoft aux États-Unis. Notons que la preuve de principe de cette approche d'archivage moléculaire des données a été apportée ; cependant, sa viabilité économique ne sera atteinte qu'en réduisant ses coûts et en améliorant sa vitesse d'un facteur 1 000 pour la lecture et de 100 millions pour l'écriture.

Est-ce réaliste ?

Le défi peut sembler énorme. Ce serait oublier la célérité des progrès des technologies de l'ADN, dont les performances doublent tous les 6 mois (un facteur 1 000 en 5 ans), à comparer à 2 ans dans les domaines du stockage électronique et informatique.

Dans le futur proche, le handicap principal de l'ADN résidera donc dans la lenteur des procédés de lecture et surtout d'écriture. Son usage se cantonnera donc initialement au stockage froid et à l'archivage de données nécessitant d'être conservées longtemps (la troisième catégorie distinguée ci-dessus), un domaine où ses avantages sont évidents, en compétition ou en complémentarité avec l'actuelle solution, qui repose, comme on l'a vu, sur la bande magnétique (cf. *Figure 3*).

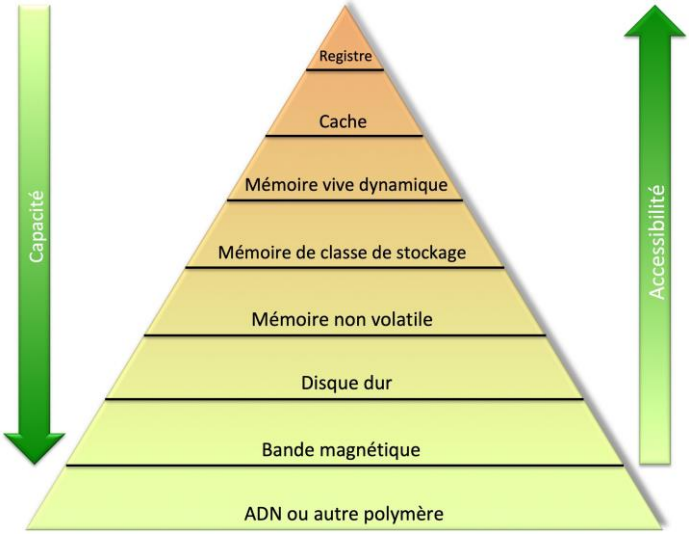

Figure 3 : Pyramide des types de mémoires dans les systèmes informatiques. En bas, l'usage hypothétique de l'ADN ou d'un autre polymère.
Crédit : François Képès et Carlo Reita

[1] Un polymère est une longue molécule constituée de nombreuses sous-unités répétées, appelées monomères. Pour stocker de l'information, il faut un polymère constitué d'au moins deux sortes de monomères, représentant le '0' et le '1'. L'ADN par exemple a quatre monomères différents : 'A', 'C', 'T', 'G'.

[2] Mais les opérations sur l'ADN ont un coût. Le IARPA (*Intelligence Advanced Research Projects Activity*, USA) estime que la consommation énergétique globale serait diminuée d'un facteur 1 000 en usant de l'ADN, comparée à l'approche conventionnelle https://www.iarpa.gov/index.php/research-programs/mist

 Fiche consultable sur le site internet www.academie-agriculture.fr onglet "***Publications***" puis "***Table des matières des documents de l'Encyclopédie***".

D'autres pistes ?

Enfin, plusieurs lignes de recherche sont parties du constat que l'ADN n'est pas nécessairement le polymère "numérique" le plus performant hors de la cellule : soit que son alphabet soit trop limité (quatre lettres), soit que sa physico-chimie ne soit pas optimale. Ce constat donne lieu à des approches alternatives s'éloignant plus ou moins de l'ADN, pour envisager d'autres hétéropolymères ou copolymères linéaires présentant des avantages théoriques. Lorsque leurs performances de lecture, d'écriture et d'édition rejoindront celles de l'ADN, ce qui pourrait prendre moins d'une décennie, ces polymères très prometteurs feront probablement irruption sur le marché de l'archivage de l'information numérique.

François KÉPÈS, membre de l'Académie d'Agriculture de France janvier 2022

Ce qu'il faut retenir :

Au-delà des approches théoriques, empiriques ou éducatives visant à limiter la quantité faramineuse d'informations générées par l'humanité, l'archivage moléculaire des informations constitue un enjeu majeur et stratégique à horizon proche.

Face aux limites physiques qu'atteignent les centres de données, la technologie moléculaire d'archivage des mégadonnées a le potentiel de devenir économiquement viable entre 2025 et 2040, progressant de marchés de niche dans les 5-10 ans vers des marchés plus globaux dans les 10-20 ans.

Les marchés de niche se cantonneront initialement au stockage froid et à l'archivage de données nécessitant d'être conservées longtemps avec peu ou pas de modifications.

Pour en savoir plus :

- Académie des Technologies : Rapport "*Archiver les mégadonnées au-delà de 2040 : la piste de l'ADN*" (groupe de travail animé par François KEPES), 2020 https://www.academie-technologies.fr/publications/archiver-les-megadonnees-au-dela-de-2040-la-piste-de-ladn/
- Académie d'Agriculture de France et Académie des Technologies : Rapport commun "*L'agriculture face à ses défis techniques*" (sous la direction de Bernard LE BUANEC), Presse des mines, 2019
- Jean-Charles HOURCADE, Erich SPITZ, Franck LALOË : *Longévité de l'information numérique.* Académie des Technologies & Académie des Sciences, EDP Sciences, 2010
- Howard COLQUHOUN, Jean-François LUTZ : *Information-containing macromolecules.* Nature Chemistry 6:455-456, 2014