Veille citoyenne d'information sur les OGM et les semences

Prolifération des bases de données de séquences génétiques numérisées

Par Eric MEUNIER, Frédéric PRAT

Publié le 21/02/2020

D’importantes divergences existent entre les États quant au statut légal des séquences génétiques numérisées dans des bases de données. L’enjeu est de savoir si l’accès à ces séquences numérisées est soumis aux mêmes conditions que celles prévues pour les ressources génétiques : consentement préalable, engagement de partage des avantages résultant de leur utilisation, protection des connaissances traditionnelles associées et/ou interdiction de revendiquer des droits de propriété intellectuelle. Un rapport rendu fin 2018 à la Convention sur la Diversité Biologique permet de mieux comprendre le fonctionnement et les limites de ces bases de données. *Inf’OGM* en résume ici les grandes lignes.

L’institut de bio-informatique européen (EBI) héberge une base de données de séquences génétiques numérisées (voir encadré ci-dessous). Le site Internet de cette base permet d’effectuer diverses recherches : génome en tant que tel, séquences génétiques particulières, séquences de protéines, petites molécules, maladies… Outre ces recherches accessibles pour n’importe quel internaute, la base propose des outils permettant de comparer des séquences entre elles, de rechercher des séquences similaires... Certes les résultats sont libres d’accès, mais leur compréhension réservée aux spécialistes.

Un lexique prolifique

Le rapport rendu à la Convention sur la Diversité Biologique (CDB) démarre sur la question des termes utilisés. Il note que dans les discussions politiques sur l’accès aux séquences et le partage des avantages (APA), les différences de terminologie reflètent souvent des points de vue divergents sur le champ d’application du Protocole de Nagoya et des lois nationales mettant en œuvre ce partage. Les États souhaitant que les séquences numérisées soient couvertes par Nagoya parleront de ressources génétiques alors que ceux souhaitant qu’elles ne soient pas couvertes insisteront sur le terme « *numérisées* » pour les différencier des ressources physiques. Le rapport de la CDB détaille que la terminologie utilisée varie également d’un processus politique à l’autre. Le terme « *données de séquences numériques* » est utilisé dans l’étude de cadrage sur la biologie de synthèse commandée par le Traité international sur les semences (Tirpaa). La Commission pour les ressources génétiques pour l’alimentation et l’agriculture (CRGAA) de la FAO utilise de son côté le terme « *information de séquence numérique* ». Dans le cadre de la Convention des Nations unies sur le droit de la mer pour la conservation et l’utilisation durable de la diversité biologique marine des zones situées au-delà de la juridiction nationale, les termes « *ressources* in silico » et « *données de séquences numériques* » ont été utilisés. De son côté, l’Organisation mondiale de la santé (OMS) a pu utiliser le terme « *données de séquence génétique* » définies comme « *l’ordre des nucléotides trouvés dans une molécule d’ADN ou d’ARN... contenant les informations génétiques qui déterminent les caractéristiques biologiques d’un organisme ou d’un virus* ».

En 2018, la Convention sur la Diversité Biologique (CDB) recevait un rapport [1] faisant le point sur les bases de données qui rassemblent en leur sein les séquences génétiques numérisées d’organismes vivants. Focalisé sur les bases de données publiques, ce rapport permet de mieux comprendre leur fonctionnement et limites. Un rapport important pour la CDB car cette dernière encadre le fameux « *partage des avantages* » entre les détenteurs de ressources et les utilisateurs (voir encadré ci-dessous). Un partage des avantages que plusieurs gouvernements souhaitent voir concerner les ressources génétiques physiques et les séquences numérisées [2] !

Partager ou non les avantages liés à l'utilisation des séquences numérisées ?

Depuis 1992, la Convention sur la diversité biologique (CDB) détermine que les ressources génétiques sont sous la souveraineté des États qui les hébergent. Le Protocole de Nagoya, adopté en 2010 [3], assure un « *accès aux ressources génétiques et un partage juste et équitable des avantages découlant de leur utilisation* ». Pour les plantes cultivées, c'est le Traité international sur les semences (Tirpaa) qui assure ce rôle.
Si, avec ces textes internationaux, les droits des États sur les ressources sont à peu près clairs, la discussion s'est déplacée sur la nature des séquences génétiques numérisées : doit-on les assimiler aux plantes elles-mêmes, auquel cas il faut alors s'assurer d' « *un partage juste et équitable des avantages découlant de leur utilisation* » ; ou bien ces séquences sont-elles d'une autre nature, et alors elles pourraient être légalement mises en libre service sur Internet, sans aucune contrepartie ? [4]

Des lieux de stockage virtuels

La base de données de l'EBI est l'une des 1 500 bases de données publiques existantes recensées par la CDB. De telles bases sont des lieux de stockage des séquences numérisées d'ADN, d'ARN, de protéines... Certaines bases se concentrent sur une maladie, d'autres sur un organisme comme une base de données pour la banane [5], pour des vers [6] ou encore sur la bactérie *Streptococcus pneumoniae* [7]. Même l'épigénétique fait l'objet de bases de données dédiées avec l'Atlas du génome humain par exemple [8]. Pour ce qui concerne l'épigénétique végétale, il faut visiter la méta-base du consortium international d'épigénomique des plantes [9].

Mais un tel nombre de bases de données a fini par poser problème, les utilisateurs ne sachant lesquelles utiliser de préférence. La réponse apportée a donc été de créer d'autres bases de données dites méta-bases qui rassemblent dans un même endroit virtuel les données des différentes bases de données. Une collaboration internationale des bases de séquences nucléotidiques (INSDC) a ainsi été créée, qui se divise en trois : l'archive nucléotidique européenne (EMBL-EBI) hébergée au Royaume-Uni, GenBank hébergée aux États-Unis d'Amérique et la banque de données ADN hébergée au Japon. Leur objectif ? « *Capter, préserver et échanger des collections de séquences nucléotidiques et les informations associées* » explique le rapport de la CDB.

Public et privé alimentent les bases de données numériques

Historiquement, les informations stockées dans ces bases sont issues de prélèvements qui ont été faits dans la Nature, y compris dans les organismes biologiques domestiqués par des communautés humaines. De plus, les instituts tels que les jardins botaniques, muséums d'histoire naturelle, universités et autres se sont engagés dans un travail de numérisation de leur collection. Le Jardin botanique royal Kew au Royaume-Uni par exemple dispose d'environ sept millions de spécimens botaniques dans son herborarium, 50 000 spécimens botaniques dans ses jardins et 35 000 graines dans sa banque de semences. Chaque année, il reçoit environ 26 000 échantillons. L'institut se charge ensuite de vérifier que les autorisations nécessaires ont bien été collectées (et le signale si ce n'est pas le cas) puis enregistre les spécimens dans son herborarium. L'ADN est extrait et les informations génétiques enregistrées dans une base de données comme GenBank.

Ensuite, les semences sont préservées au froid. Certaines bases de données comme l'EMBL ou GenBank ont fini par atteindre aujourd'hui quelques quadrillions, c'est-à-dire "10 puissance 24" [10] nucléotides enregistrés pour environ 300 000 espèces végétales [11].

Qu'il s'agisse de bases de données publiques ou privées, les séquences enregistrées ont été obtenues par séquençage du génome d'organismes vivants. La CDB note qu'aujourd'hui « *la collecte aux champs d'échantillons physiques est une partie plus petite des stratégies de recherche commerciale* » que dans les années 80. Mais la recherche publique continue de s'intéresser à ces échantillons physiques, notamment pour ce qui concerne les espèces de microbes ou les organismes vivant dans les zones à forte diversité d'espèces, les zones d'environnement extrême ou les niches écologiques. Des programmes de science dite citoyenne, décrits par le rapport de la CDB, participent à la collecte d'échantillons de partout dans le monde « *afin de comprendre la diversité génétique et biologique* ». Le rapport note que ces programmes permettent à la recherche « *d'éviter de dépenser temps et argent pour des expéditions d'échantillonnage* » et aboutissent à des « *quantités de données massives, couvrant de vastes zones géographiques* ».

Un tel travail de collecte d'échantillons et de séquençage devrait continuer : la nature est d'une diversité constamment renouvelée et potentiellement infinie… et les appareils de séquençage, devenus portatifs, sont de plus en plus abordables financièrement. Le rapport estime en effet que bientôt, « *des individus pourront facilement, et à faible coût, séquencer des gènes à partir d'échantillons physiques n'importe où dans le monde et les envoyer* via *Internet [...] loin du lieu d'échantillonnage* ».

En termes d'utilisation des bases de données publiques, les chiffres donnent le vertige. Pour le seul projet européen EMBL-EBI, ce sont 12,6 millions de requêtes par mois qui sont enregistrées, notamment des recherches de similarités entre une séquence génétique et celles enregistrées dans les banques de données. Sans donner de chiffres précis, le rapport de la CDB souligne que « *le nombre de séquences, le nombre d'individus et le nombre d'espèces séquencés [...] augmentent. Les journaux [scientifiques] exigent que les séquences génétiques soient déposées dans des banques publiques comme condition à une publication* ». Il précise également que les « *offices de brevets peuvent également soumettre les séquences incluses dans les demandes de brevets à ces banques* ».
Dans les dernières années, ont été inclus le contexte environnemental et la localisation d'origine des échantillons séquencés, « *ce qui est important pour la science et peut contribuer aux partages des avantages* » (sic). Mais ces informations ne sont, aujourd'hui, pas toujours complètes.

À quoi servent ces bases, à quelles conditions ?

Le premier rôle d'une base de données est bien de donner accès aux données elles-mêmes. Une lapalissade, certes, mais qui a des conséquences importantes dans l'utilisation des données. Le document de la CDB détaille qu'une utilisation courante des bases de données est par exemple de chercher des régions similaires entre deux séquences dont une est dans la base et l'autre entre les mains de l'utilisateur. Pour le monde de la recherche, cela permet de comparer des séquences et de les étudier selon leur fonction, leur évolution d'un organisme à un autre ou dans le temps.
Une conséquence non négligeable selon la CDB, « *cela permet également aux chercheurs de trouver des séquences identiques dans différents organismes de manière à ne pas utiliser des séquences dont le statut légal serait incertain eu égard au partage des avantages* ». Autre utilisation : les informations numériques peuvent permettre de reconstituer l'ADN, ARN, protéines… physiques, sans avoir eu accès à l'échantillon physique lui-même.

Ainsi, certaines bases de données – mais pas toutes - imposent que toute utilisation d'une séquence génétique numérisée se fasse avec l'accord du pays où l'organisme séquencé a été collecté et qu'une utilisation commerciale fasse l'objet d'une information à la CDB. Une autre approche est celle de l'open source. Dans cette politique, les séquences numérisées sont librement accessibles par tous.
Ces quelques exemples de conditions d'accès ou d'utilisation ne reflètent néanmoins pas la majorité des bases de données. Le rapport de la CDB note ainsi que « *la majeure partie des séquences génétiques numérisées est accessible* via *des bases de données publiques qui ne requièrent pas des contributeurs ou utilisateurs de s'enregistrer ou s'identifier, d'accepter les conditions générales ou signer des accords d'utilisation* ». Une politique de libre accès souvent requise par les gouvernements finançant ces bases de données selon le rapport de la CDB.

Trois limites importantes

Cette politique de libre accès est une première limite majeure dans l'utilisation qui est faite des séquences enregistrées, car sans conditions d'utilisation particulières, aucun contrôle ne peut être fait d'éventuels droits de propriété intellectuelle déposés sur des séquences approchantes voire similaires.
Surtout qu'une seconde limite vient renforcer ce risque, celle d'une non information de l'origine des séquences enregistrées. Il s'agit cette fois d'une raison technique. Le rapport de la CDB avance en effet que, bien que certaines bases de données le fassent, il ne serait pas toujours possible de garantir *a posteriori* l'origine des séquences numérisées. À cela, une raison pratique : les séquences génétiques d'une même espèce peuvent varier d'un habitat à l'autre du fait de mutations naturelles et/ou de la domestication ; alors que des séquences d'espèces différentes et d'origines différentes peuvent être similaires et, en l'absence de traçabilité, il devient impossible de retrouver l'origine exacte de celles qui sont dans une base de données. On notera ici avec intérêt qu'un tel constat n'a jamais empêché qu'un brevet sur une séquence génétique soit délivré, ni même que son détenteur se donne les moyens d'assurer sa traçabilité afin de faire valoir ses droits.
Cette incertitude sur la fiabilité de l'origine des séquences génétiques numérisées alimente également une dernière difficulté qui est due au système en lui-même. Comme le souligne le rapport de la CDB, la frontière entre recherche publique et privée s'est fortement estompée notamment du fait des partenariats publics-privés. Une séquence génétique numérisée dans une base de données publique étant accessible à tous, il n'est « *pas toujours clair de savoir comment ce matériel sera utilisé dans le futur* ». Et de détailler que des séquences obtenues *via* une recherche publique, enregistrées dans une base de données, peuvent être utilisées commercialement, par plusieurs personnes, « *sans que le fournisseur d'origine ne soit au courant* »...

Plusieurs pistes sont avancées pour répondre aux problèmes que posent l'identification fiable de la provenance d'une séquence génétique numérisée, de son fournisseur initial et, concept récent, le calcul de la valeur financière qui peut lui être attribuée. Parmi ces pistes, et outre les conditions d'utilisation générales, le rapport de la CDB évoque la basique mise en place à la fois de données complémentaires à rattacher à une séquence et d'un identifiant unique pour les chercheurs. Mais le même rapport de la CDB souligne qu'au début du stockage des données, ces informations décrivant le lieu et le contexte environnemental dans lequel l'organisme séquencé avait été prélevé n'étaient pas renseignées. Pour certains organismes, le nom même de l'organisme pouvait ne pas être donné. Autant d'informations aujourd'hui considérées comme essentielles selon le Consortium des standards génomique fondé en 2015. Mais encore aujourd'hui, tous les enregistrements dans des bases de données ne disposeraient pas de ces informations malgré des normes établies. La CDB, donnant la parole à un gestionnaire de base, explique que « *nous dépendons de ceux*

soumettant les données. Nous ne pouvons pas tout vérifier avec une soumission toutes les six minutes en moyenne. Et nous ne communiquons pas tant que cela avec les personnes envoyant les données. Nous y travaillons et espérons que tout le monde finira par prendre ses responsabilités ». De telles informations sont en effet impératives pour pouvoir mettre en œuvre le partage des avantages mais également pour pouvoir travailler. La base de données EMBL-EBI explique ainsi que « *pour comprendre des associations entre rendement de culture et différences de flore du sol, il est utile de savoir quand et où les échantillons ont été collectés* » [12].

La Convention sur la Diversité Biologique fait donc aujourd'hui face à une question simple. Est-il possible pour les bases de données de renseigner rétroactivement l'origine des ressources génétiques à la base des séquences enregistrées ? La réponse devrait être fournie lors de réunions à venir comme celle d'octobre 2020 qui doit avoir lieu à Kunming en Chine. Rappelons cependant que pour certains acteurs, le problème est d'abord celui de l'abandon de toute brevetabilité du vivant plutôt que celui d'un illusoire partage des avantages...

---

[1] « Fact-finding and scoping study on digital sequence information on genetic resources in the context of the convention on biological diversity and the nagoya protocol », 10 janvier 2018

[2] Eric MEUNIER, Frédéric PRAT, « Internet et biopiraterie, les États ne sont pas d'accord », *Inf'OGM*, 22 janvier 2019

[3] http://www.cbd.int/abs/text/default.shtml

[4] Eric MEUNIER, « Numériser les gènes pour posséder le vivant sans partage ? », *Inf'OGM*, 9 avril 2018

[5] http://banana-genome- hub.southgreen.fr/

[6] http://www.wormbase.org/#012-34-5

[7] http://pranag.physics.iisc.ernet.in/SPGDB/

[8] http://www.genboree.org/epigenomeatlas/index.rhtml

[9] https://www.plant-epigenome.org/

[10] un quadrillion : un million puissance 4, d'où son nom, soit un 1 suivi de 24 zéros.

[11] Le nombre total des organismes vivants est inconnu, mais estimé autour de 10 à 15 millions. Un projet, le Earth bioGenome Project (EGP) a pour objectif de séquencer l'ensemble de ces génomes, voir : Eric MEUNIER, « Numériser les gènes pour posséder le vivant sans partage ? », *Inf'OGM*, 9 avril 2018 et Frédéric PRAT, « Séquencer le génome de l'ensemble des êtres vivants sur Terre », *Inf'OGM*, 27 août 2019

[12] http://www.ebi.ac.uk

Adresse de cet article : https://infogm.org/proliferation-des-bases-de-donnees-de-sequences-genetiques-numerisees/