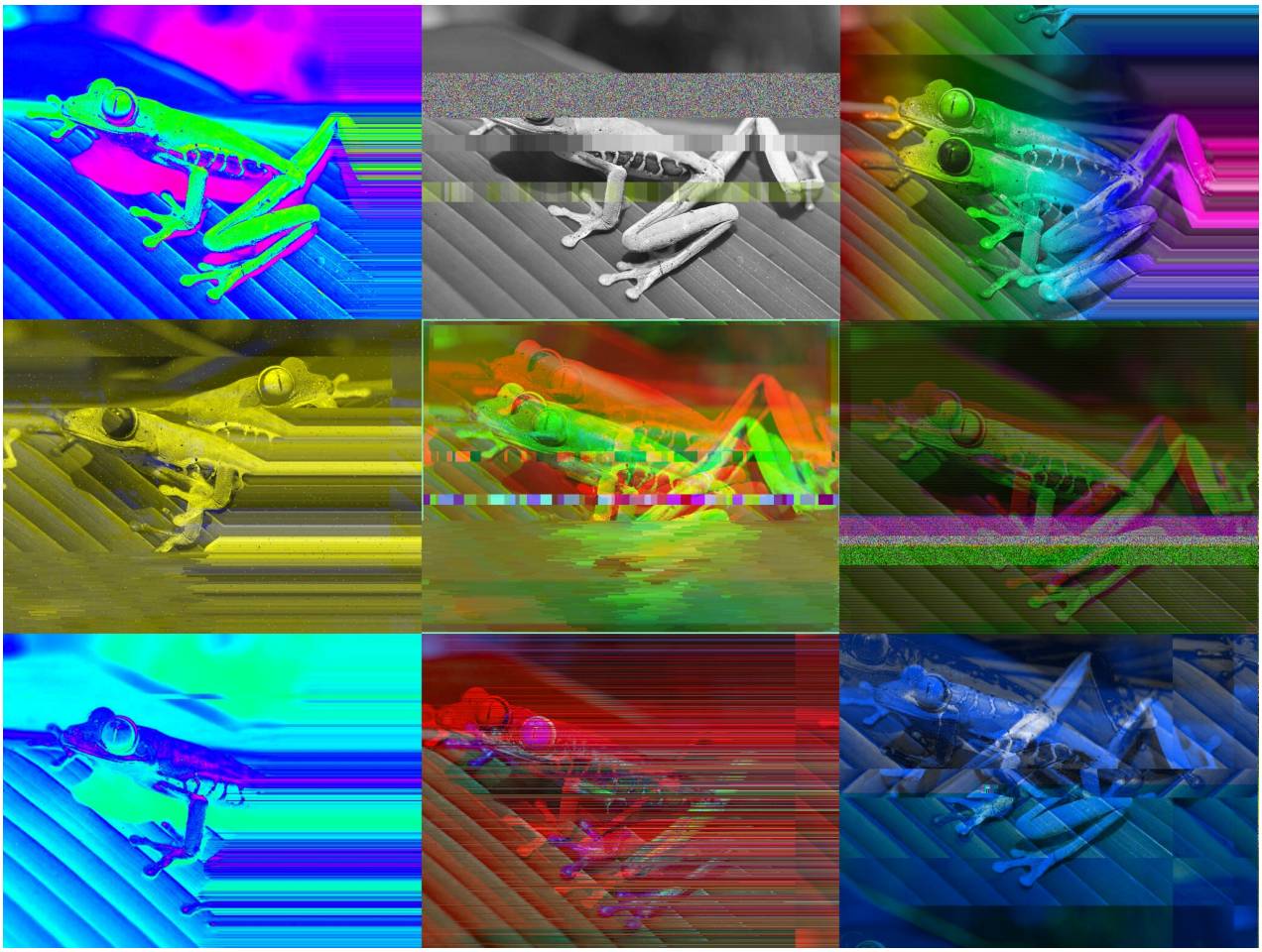
Veille citoyenne d'information sur les OGM et les semences

Artificialisation du vivant : des données numériques contaminées

Par Eric MEUNIER

Publié le 04/02/2021, modifié le 09/03/2026



La numérisation des génomes d'organismes vivants en vue d'applications biotechnologiques est le nouveau Graal des entreprises. Actuellement, certains gouvernements souhaitent exonérer les données, issues de cette numérisation, des obligations internationales de partage des avantages. Une approche qui conviendrait aux entreprises souhaitant accroître leur capacité à breveter le vivant. Cependant la maîtrise technique est loin d'être au rendez-vous. Ainsi, à l'instar de la contamination de cultures agricoles par des OGM, les bases de données des séquences

numérisées sont contaminées par des séquences qui n'ont rien à y faire ou qui ne sont pas à leur place.

Depuis plusieurs années, les génomes, protéines et autres molécules d'organismes vivants sont séquencés par des méthodes présentées comme de plus en plus rapides et économiques. Ces informations de séquences numérisées (DSI pour *Digital sequence information*) sont enregistrées dans des bases de données informatiques publiques ou privées. Comme l'a détaillé *Inf'OGM* dans son récent dossier [1], l'utilisation de ces DSI fait l'objet de débats politiques intenses au niveau international avec un enjeu majeur : les acteurs ayant les ressources et moyens pourront-ils allègrement piocher dans une biodiversité numérisée sans l'accord des États souverains et sans devoir partager les bénéfices résultants, comme cela est imposé aujourd'hui par la Convention sur la Diversité Biologique pour justement éviter toute biopiraterie ?

En attendant la réponse, il est à noter qu'à l'instar du débat sur les OGM, la numérisation des informations de séquences a toujours été présentée avec pour acquis sous-jacent une maîtrise technique absolue. Une récente étude conduite sur une base de séquence, GenBank et sur une version « *nettoyée* » de cette base, la base de données RefSeq [2] remet pourtant en cause cette prétendue maîtrise technique comme une autre étude l'avait fait voici presque vingt ans [3].

Deux millions de séquences contaminées

Les deux auteurs de cette étude se sont penchés sur une des trois bases de données de séquences majeures, GenBank. Avec les paramétrages qu'ils ont utilisés, les chercheurs ont identifié plus de deux millions de séquences de 6 795 espèces enregistrées dans Genbank comme contaminées par des séquences qui n'ont rien à y faire. Pour la base RefSeq, 114 000 séquences de 2 767 espèces sont contaminées par de telles séquences. S'ils calculent que cela concerne 0,54 % des séquences enregistrées dans Genbank ou 0,34 % des séquences enregistrées dans RefSeq, une version donc pourtant manuellement « *nettoyée* » de Genbank, les auteurs estiment que cela représente déjà une « *fraction substantielle* ». Précisions importantes, les auteurs ont simplifié le travail. Ils ont en effet considéré une séquence comme contaminante seulement dans le cas où elle est retrouvée dans les séquences d'un organisme appartenant à un autre règne du vivant. Les auteurs ont aussi fusionné deux règnes, les bactéries et les archées, rendant impossible de « *détecter des contaminations entre ces deux règnes* » du vivant. Enfin, ils n'ont pas cherché d'éventuelles contaminations par des séquences métagénomiques, c'est-à-dire d'organismes accompagnateurs et donc contaminants (microbiote…). Les résultats présentés pourraient donc bien être la partie émergée de l'iceberg.

Quant aux causes de ces contaminations, elles sont multiples selon les deux chercheurs. Cela peut être dû aussi bien aux réactifs utilisés lors des protocole de séquençage qu'à du « *matériel de laboratoire, des artefacts de traitement des échantillons ou des contaminations croisées lors des multiples tours de séquençage* ». Une des origines majeures des séquences contaminantes est nous-mêmes, *Homo sapiens*. Présent dans les laboratoires, l'humain contamine de fait les échantillons qu'il manipule. Les auteurs notent que parmi les exemples de contamination qu'ils ont trouvés figure celui de « *milliers de fragments d'ADN humains (...) retrouvés dans des génome bactérien et [dont] de nombreux ont été traduits et annotés comme protéines* ». Une situation qui fausse donc les informations enregistrées dans les bases, tant en termes d'ADN que de protéines.

Une contamination source d'erreurs potentiellement importantes

Ces contaminations relativisent le discours de maîtrise absolue des techniques de séquençage qui sous-tendent les positions actuellement mises en avant dans les négociations internationales [4].

Mais cette absence de maîtrise technique pose d'abord problème aux chercheurs eux-mêmes. Car ces séquences contaminantes, une fois enregistrées dans les bases de données, sont faussement interprétées comme appartenant à un organisme. Les auteurs de l'article soulignent ainsi que « *ces séquences contaminantes posent une variété de problèmes parmi lesquels des dénominations incorrectes de séquences dans des études de métagénomiques, des conclusions erronées de transfert horizontal ou des annotations de génome de mauvaise qualité* » [5]. Des contaminations qui pourraient donc induire les chercheurs en erreur dès le départ de leurs travaux.
Un exemple ? Les chercheurs mentionnent la présence d'au moins 4 000 nucléotides provenant d'*Escherichia coli* dans le génome du nématode *Caenorhabditis elegans*. Problème : ce nématode est l'un des organismes utilisés comme modèle en laboratoire... d'où de nombreuses erreurs potentielles.

Ces résultats relativisent de fait de nombreuses notions, qu'il s'agisse de la qualité des génomes évidemment et notamment ceux dits « *de référence* » ou encore l'annotation même des séquences. Ces annotations qui ambitionnent de lier génotypes (les séquences génétiques) et phénotypes (les caractéristiques visibles d'un organisme) pourraient, de fait, être erronées. Les deux auteurs qui ont écrit cet article proposent un nouvel algorithme nommé Conterminator qui permettrait de nettoyer les bases de données d'une partie des séquences contaminantes. Bien qu'ils aient privilégié la vitesse sur la fiabilité et l'exhaustivité, cet algorithme rappelle par sa seule existence combien les assertions de maîtrise des techniques peuvent être trompeuses. La recommandation formulée d'utiliser leur algorithme « *en routine* » sur des bases en constante augmentation témoigne d'ailleurs de l'ampleur de la tâche. À titre de comparaison, on notera que les auteurs estiment qu'un nettoyage relativement fiable de la seule base RefSeq (déjà partiellement nettoyée, rappelons-le) prendrait, avec les logiciels actuels, 30 000 années de calculs...

---

[1] Voir le dossier « Numériser le vivant pour mieux le privatiser », publié dans *Inf'OGM, le journal*, N°162.

[2] « Terminating contamination : large-scale search identifies more than 2,000,000 contaminated entries in GenBank », M. Steinegger et S. L. Salzberg, *Genome Biology*, 2020, 21:115.

[3] « Can you bank on GenBank ? », D.J. Harris, *Trends in Ecology and Evolution*, Vol.18 No.7 July 2003, p.317

[4] Voir le dossier « Numériser le vivant pour mieux le privatiser », *Op. cit.*

[5] Les auteurs entendent par là que des observations de transfert horizontal de séquences entre deux espèces pourraient ne pas en être mais simplement des contaminations des séquences d'une espèce enregistrées dans une base de données par des séquences d'une autre espèce.

Adresse de cet article : https://infogm.org/artificialisation-du-vivant-des-donnees-numeriques-contaminees/