Détecter les effets de la sélection naturelle sur l'ADN par intelligence artificielle

Publié le 02.04.19 |Par Frantz Depaulis

Des approches d'intelligence artificielle présentent de nouvelles applications pour détecter l'impact de la sélection positive sur le génome. Elles démarrent par une phase supervisée sur simulations, pour être ensuite appliquées sur des données naturelles. La résolution est progressive suivant un processus dit d'apprentissage profond. Les premiers résultats obtenus sur des données humaines semblent très prometteurs.

La plupart de la variation de l'ADN à l'intérieur des espèces est probablement due à des mutations dites *neutres* au sens où l'évolution de leur fréquence allélique est essentiellement aléatoire sous l'effet de la dérive génétique. Les rares mutations avantageuses et celles délétères modifient quant à elles le phénotype, qui est la cible de la sélection, positive dans le premier cas, négative dans le second. Les mutations *avantageuses*, en conférant un avantage aux individus qui les portent, représentent la principale source de l'adaptation évolutive. Leur évolution (sélection positive) diffère significativement de celle des mutations neutres. Cependant, trouver les traces de la sélection positive le long du génome est une tâche ardue qui nécessite de séquencer et d'analyser l'ADN d'un grand nombre d'individus.

L'évolution de mutations neutres dans une population peut néanmoins être affectée par une mutation avantageuse voisine via un effet d'entraînement ou *auto-stop sélectif*. Ces mutations neutres sont en effet transmises avec leur voisine et montrent alors un apparent effet sélectif. L'impact sélectif est ainsi étendu le long de l'ADN. C'est la présence d'une portion d'ADN particulièrement peu variable qui permet de détecter l'existence d'un phénomène de sélection. Cette portion d'ADN est cependant restreinte car, au fur et à mesure que l'on s'éloigne de la zone sélectionnée, les processus de recombinaison génétique tendent à défaire l'association entre la mutation avantageuse et les mutations neutres les plus éloignées. Ainsi, la portée de l'auto-stop, quoique étendue par rapport au site cible de la sélection, reste limitée à une région. La localisation de telles régions permet d'identifier des allèles qui pourraient avoir été sélectionnés positivement. Un exemple sélectif type est celui de la capacité à digérer le lactose (un sucre du lait) à l'âge adulte chez l'homme. Différentes mutations du gène de la lactase (*LCT*), apparues indépendamment, sont responsables de ce phénotype. La fréquence de ces allèles, faible avant l'avènement de l'agriculture et de l'élevage au Néolithique, a par la suite rapidement augmenté dans certaines populations (lire à ce sujet *Les conséquences évolutives de changements culturels : l'exemple de la révolution néolithique*). Les allèles situés à proximité de ces mutations ont donc également été sélectionnés (Itan et coll., 2009).

Les méthodes d'analyse statistiques initialement développées pour repérer des indices de sélection positive comparaient la variabilité génétique observée sur un locus avec la variabilité attendue dans le cas d'un modèle neutre (locus soumis à des mutations neutres uniquement). Ce modèle reposait sur l'hypothèse que l'évolution de la fréquence d'une mutation neutre est uniquement due au hasard. Il faisait également de fortes hypothèses sur la démographie de la population considérée (population homogène, d'effectif constant...). De ce fait des résultats significatifs pouvaient aussi bien être dus à la sélection qu'à des effets démographiques.

Une nouvelle approche interdisciplinaire utilisée pour ces analyses de détection de la sélection repose sur l'intelligence artificielle qui vise à imiter le fonctionnement du cerveau humain. Plus précisément, les analyses les plus utilisées reposent sur les méthodes dites d'*apprentissage profond* (*deep learning*) avec un affinage progressif de la classification entre régions de l'ADN sélectionnées ou non, voire entre différents régimes sélectifs (Shrider et Kern, 2017). Elles impliquent un apprentissage dit supervisé. Le programme apprend d'abord sur des jeux de données témoins où la réponse exacte est connue. Ces jeux de données sont ici obtenus par simulations de données virtuelles. Le programme ainsi entraîné est ensuite appliqué sur les données réelles, au préalable inconnues du programme, pour détecter les régions de l'ADN soumises à sélection.

Un des principaux arguments en faveur de ce type de méthodes est qu'elles font des hypothèses moins fortes sur le mode d'évolution des populations que les méthodes statistiques initialement utilisées. L'étape préliminaire de supervision reste néanmoins dépendante d'un scénario démographique généralement inféré à partir des données. Si ce scénario ne correspond pas à la réalité, les résultats pourraient s'avérer incorrects. Une autre critique tient au caractère de boîte noire de l'approche : on ne sait pas précisément ce qui constitue le signal supposé sélectif. Il est alors difficile d'avoir un regard critique sur les résultats et d'imaginer des hypothèses alternatives. Des recherches actuelles visent cependant à caractériser ce signal.

L'application de ces méthodes d'analyse sur des jeux de données humains a permis de retrouver des résultats connus lors d'études préalables – dont la sélection positive de certains variants du gène de la lactase – et de proposer de nouvelles zones de l'ADN qui pourraient avoir été sélectionnées positivement. Pas moins de 20 000 d'entre elles ont ainsi été récemment détectées lors d'une étude portant sur l'ADN d'environ 2 500 individus (Shrider et Kern, 2017). S'il reste à caractériser précisément les fonctions de ces zones, certaines catégories fonctionnelles apparaissent déjà surreprésentées. C'est le cas de gènes impliqués dans des cancers, le développement du système nerveux central, les interactions avec des virus, l'immunité et la reproduction. Plus généralement sur ces dernières données, les résultats suggèrent que la plupart des mutations sont préalablement neutres mais que, ponctuellement, certaines peuvent devenir avantageuses lorsque l'environnement change (par exemple lors de changements de l'aire de répartition de la population ou de l'espèce considérée). Une seconde étude sur une population africaine faisait ressortir une sélection positive de certains allèles de gènes impliqués dans le métabolisme (Sugden et coll., 2018). L'interprétation des chercheurs était que ces allèles permettaient un stockage accru des réserves de graisse en prévision de périodes de famine.

En résumé, ces résultats suggèrent que l'intelligence artificielle pourrait être plus efficace pour détecter les zones de l'ADN soumises à sélection positive que les méthodes statistiques initialement développées.

Références

- Itan, Y., Powell, A., Beaumont, M. A., Burger, J., & Thomas, M. G. (2009). The origins of lactase persistence in Europe. PLoS computational biology, 5(8), e1000491.
- Schrider, D. R., & Kern, A. D. (2017). Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Molecular biology and evolution, 34(8), 1863-1877.
- Sugden, L. A., Atkinson, E. G., Fischer, A. P., Rong, S., Henn, B. M., & Ramachandran, S. (2018). Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nature communications, 9, 703. doi:10.1038/s41467-018-03100-7

CRÉDITS

AUTEUR(S)/AUTRICE(S)

Frantz Depaulis

Spécialisé en génétique des populations moléculaire, il développe et applique des méthodes d'analyses de données. Membre du CNRS, il travaille à l'École normale supérieure (Paris).

MISE EN LIGNE

Pascal Combemorel

Agrégé de SVT, il est le responsable éditorial du site Planet-Vie depuis septembre 2016.