Combiner arbres phylogénétiques et visualisation d'ensembles

Jean-Baptiste Lamy[*,**], Flora Jay[*,***]

[*]Laboratoire de Recherche en Informatique,
CNRS/Université Paris-Sud/Université Paris-Saclay, Orsay, France
flora.jay@lri.fr
[**]LIMICS, Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny,
INSERM UMRS 1142, Sorbonne Universités
jean-baptiste.lamy@univ-paris13.fr
[***]Laboratoire EcoAnthropologie et Ethnobiologie,
CNRS/MNHN/Université Paris Diderot, Paris, France

Les arbres sont très largement utilisés en phylogénie. Cependant, un arbre avec n feuilles présente les similarités entre $n-1$ sous-ensembles de feuilles, alors qu'il existe $2^n - n - 1$ sous-ensembles possibles d'au moins deux feuilles. Par exemple, pour 3 feuilles A, B et C, 4 similarités peuvent être mesurées, entre les sous-ensembles $\{A, B\}$, $\{B, C\}$, $\{A, C\}$ et $\{A, B, C\}$, mais un arbre n'en montrera que 2, *e.g.* $\{A, B\}$ et $\{A, B, C\}$ si une branche rassemble A et B. Plus n augmente, plus la vision donnée par l'arbre deviendra réductrice.

Ce problème est particulièrement important en génétique des populations (Pickrell et Pritchard, 2012), lorsque l'on étudie la diversité génétique des populations d'êtres vivants et leurs relations. Dans ce contexte, il est nécessaire de tenir compte des processus biologiques telle l'apparition de mutations dans le génome mais aussi des processus démographiques, tels que la séparation des populations (aussi appelée divergence), la variation de leur taille effective, et les migrations (mouvement d'un groupe d'individus qui quittent une population pour en rejoindre ou en créer une autre, apportant au passage leur matériel génétique). Un arbre généalogique peut représenter la composante évolutive et une partie de la composante démographique (divergence simple des populations, dérive génétique plus forte dans les populations de petite taille, etc) mais pas la composante migratoire post divergence qui induit des "flux de gènes" entre branches de l'arbre.

Afin de résoudre ce problème, nous proposons l'utilisation de visualisation d'ensembles, et notamment des boîtes arc-en-ciel (Lamy et al., 2017) et de leur variante proportionnelle (Lamy et Tsopra, 2019), et l'illustrons par une application à un sous-ensemble de données extraites du *1000 Genomes Project* (Auton et al., 2015). Une première approche consiste à visualiser les similarités comme des ensembles (sous-ensemble des populations avec les mêmes mutations).

Une seconde approche consiste à superposer arbre phylogénétique et boîtes arc-en-ciel (voir Figure 1). L'arbre (en noir) représente l'histoire démographique "prépondérante" des populations, lesquelles sont présentées en colonne (les couleurs dans les en-têtes de colonne identifient les continents). La longueur des branches de l'arbre indique le nombre de mutations depuis l'ancêtre commun. Les boîtes rectangulaires permettent de visualiser les similarités entre branches éloignées de l'arbre : par exemple la boîte bleu ciel à gauche montre une similarité

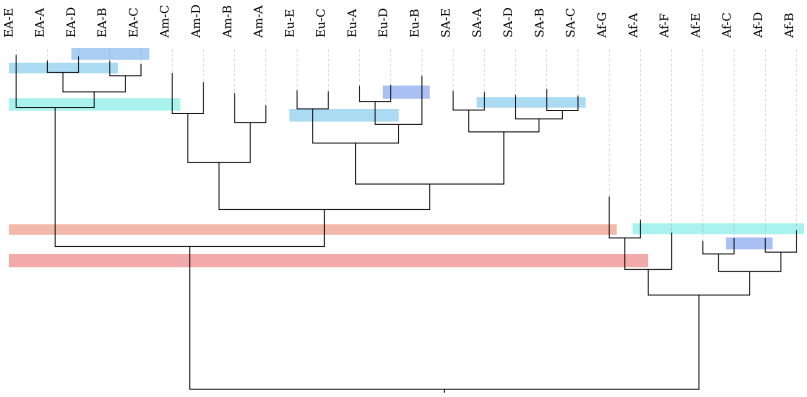

FIG. 1 – *Exemple préliminaire de combinaison arbre phylogénétique - boîtes arc-en-ciel.*

entre une population amérindienne (Am-C) et l'ancêtre des populations est-asiatique (EA), bien qu'il n'y ait pas d'ancêtre commun spécifique à l'ensemble de ces populations. Chaque boîte recouvre les branches (et sous-branches) de l'arbre correspondant aux populations qui partagent les similarités. La hauteur des boîtes est proportionnelle au nombre de mutations similaires, et la couleur indique le nombre de populations impliquées (plus la couleur chaude, plus le nombre de populations est élevé). Les boîtes correspondant à des similarités déjà visualisées par l'arbre ne sont pas affichées. Ces analyses se basent sur un sous-ensemble de marqueurs génétiques et nous demeurons prudents quant à leur interprétation.

Les difficultés rencontrées sont (1) la génération d'arbres et de boîtes arc-en-ciel ayant la même unité afin de les rendre comparables, (2) la génération de boîtes à partir de données numériques (une mutation donnée pouvant être présente chez 60% des individus d'une population, et non seulement 0 ou 100%), et (3) la complexité visuelle des représentations obtenues.

Références

Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, et et al. (2015). 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature 526*(7571), 68–74.

Lamy, J. B., H. Berthelot, C. Capron, et M. Favre (2017). Rainbow boxes : a new technique for overlapping set visualization and two applications in the biomedical domain. *Journal of Visual Language and Computing 43*, 71–82.

Lamy, J. B. et R. Tsopra (2019). RainBio : Proportional visualization of large sets in biology. *IEEE Transactions on Visualisation and Computer Graphics accepted*.

Pickrell, J. K. et J. K. Pritchard (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics 8*(11), e1002967.

Summary

In population genetics, it is important to visualize both the genetic evolution of populations and the contribution of migrations to the spreading of genetic mutations. Here, we propose an original approach combining phylogenetic tree and set visualization with rainbow boxes.