Comment comparer des structures d'arbres et autres questions épineuses ?

Pierre Darlu, CNRS, INSERM U535 Epidémiologie génétique et structure des populations humaines, 94817 Villejuif

Le problème statistique de la comparaison d'arbres est un problème relativement difficile, certains disent même « problematic », « vacuous » (Felsenstein, 2004) ou « with no biological relevance » (Penny et al., 1982).

Certes, il existe des méthodes permettant de tester si des données sont plus ou moins bien « ajustées » à un arbre plutôt qu'à un autre. Mais il ne s'agit pas d'une réelle comparaison de topologies d'arbre, plutôt d'estimer (par parcimonie ou vraisemblance) l'adéquation de données à des arbres.

Certes, il existe aussi de nombreuses mesures indiquant la différence entre deux topologies (RF de Robinson et Foulds (1981), « Symetric Difference » (Hendy et Penny, 1985), , NNI (Nearest neighbour interchanges), « quartets » ou triplets résolus (Estabrook et al., 1985), « agreement subtree » ou voir A. Guénoche etc…)

Cependant, les statistiques sur ces mesures sont rares ou inexistantes. Elles se bornent le plus souvent à tester si deux arbres se ressemblent plus que deux arbres aléatoires. De simples tests de permutations permettent d'estimer la probabilité de rejeter l'hypothèse nulle Ho d'absence de congruence (lLD de Farris et al., (1994) dans le contexte de parcimonie ou tests de Shimodaira et Hasegawa (1999) dans le contexte probabiliste, par exemple) ou d'absence de corrélations entre deux matrices de distances d'arbres (test de Mantel). Mais ces tests n'indiquent pas si deux arbres sont plus proches entre eux qu'ils ne le sont d'un troisième ou d'autres arbres.

Pour contourner la difficulté, une tentative de comparaison d'arbre par des méthodes de réechantillonage sera offerte à votre critique. Elle conduit à construire un « arbre des arbres », attachant à chaque partition de cet arbre une estimation de sa robustesse. Les applications sont nombreuses, en particulier dans le domaine de la biologie évolutive où la profusion d'arbres découle naturellement de la multiplication du séquençage de nombreux gènes et espèces.

Si le temps le permet, quelques autres questions épineuses (tout au moins pour l'intervenant) pourraient être soumises brièvement à la sagacité des auditeurs, comme celle de la détection de ponctuations dans les arbres évolutifs…