Modeling molecular evolution with fast adaptive divergence at speciation

Marc Manceau^{1,2}, Hélène Morlon², Amaury Lambert¹, ¹ Collège de France, 11 place Marcelin-Berthelot, Paris 5ème ² Ecole Normale Supérieure, 46 rue d'Ulm, Paris 5ème



Motivation

- There is a long-standing debate as to whether molecular evolution proceeds in a punctuated versus gradual way at the macro scale.
- However, models of molecular evolution assume that sequences evolve gradually. Statistical tests that have been performed [1] are not modelbased and might be subject to phylogenetic reconstruction artifacts.
- The fanciest relaxed clock models assume that the substitution rate changes through time [2].
- Recent studies suggest that there is signal in biological data supporting both "auto-correlated" and "non-auto-correlated" rates of molecular evolution [3].



How could we model the "non-auto-correlated part" in a more process-based fashion ?

In particular, this signal could be better captured by punctuated events at the macro scale, capturing episodes of fast disruptive selection at speciation.



We want to compute the likelihood of the reconstructed tree with spikes in order to speed up simulations and allow inference from biological data.

■ On the reconstructed tree,

- u probability of spike per speciation
- κ probability of substitution at spike per nucleotide

On a lineage extending from t_0 to t_1 , the number of spikes is Poisson distributed with parameter

Let u(t) be the probability that the descendance of a species extant at time t be extinct before the present.

The spike rate on a lineage of the reconstructed tree is $2b\nu u(t)$

$$\int_{t_1}^{t_0} 2b\nu u(s)ds = 2b\nu(t_0 - t_1) - 2\nu \ln \frac{1 - \frac{b}{d}e^{(b-d)t_0}}{1 - \frac{b}{d}e^{(b-d)t_1}}$$

Perspectives

■ Testing possible Phylogeny dating and spike positioning : sampling from inferences : $\mathbb{P}(S, T \mid A)$

Parameter inference by maximum likelihood : trying to recover (b,d,ν,κ)

 $\mathbb{P}(\mathcal{A} \mid \mathcal{T}, \mathcal{S}) \qquad \qquad \mathbb{P}(\mathcal{S} \mid \mathcal{T}) \qquad \qquad \mathbb{P}(\mathcal{T})$

 $\mathbb{P}(\mathcal{S},\mathcal{T}\mid\mathcal{A}) \propto \mathbb{P}(\mathcal{A}\mid\mathcal{T},\mathcal{S})\mathbb{P}(\mathcal{S}\mid\mathcal{T})\mathbb{P}(\mathcal{T})$

 \blacksquare Model selection : Testing the support for spikes with a LRT testing if $\nu \neq 0$

Looking specifically for loci supporting fast positive selection at speciation events.

Comparing the fit of the model to other molecular clocks. Trying to nest distinct models in one in order to perform model selection.

References

[1] Pagel M., C. Venditti, and A. Meade. 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. Science 314 :119–121.

[2] Lepage T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Mol. Bio. and Evol. 24 :2669–2680.

[3] Lartillot N., M.J. Phillips, F. Ronquist. 2016. A mixed relaxed clock model. Phil. Trans. R. Soc. B 371: 20150132