Towards alife-generated benchmarks for phylogeny

Marco Foley[1], Jonathan Rouzaud-Cornabas[1,2], Eric Tannier[1,3], and Guillaume Beslon*[1,2]

[1]BEAGLE – INRIA, Institut National des Sciences Appliquées [INSA] - Lyon – France
[2]Laboratoire d'Informatique en Image et Systèmes d'information – Université Lumière - Lyon 2, Ecole Centrale de Lyon, Université Claude Bernard Lyon 1, Centre National de la Recherche Scientifique : UMR5205, Institut National des Sciences Appliquées de Lyon – France
[3]Laboratoire de Biométrie et Biologie Evolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

Résumé

Inspired by the double-blind principle that governs testing in science, we propose a new way to test methods for molecular evolution with computer simulations. Here, two teams (the INRIA Beagle Team, specialized in computational evolution, and the CNRS/LBBE Le Cocon Team, specialized in phylogeny) worked concurrently, Beagle producing evolutionary simulations – without information about the analysis tools – while Le Cocon tested phylogenomic tools on the simulated data without information on the way they have been generated.
Blind sequence generation: The Beagle team adapted its Aevol platform to allow for the simulation of 4-bases sequences (while the original platform uses binary sequences). This allows analyzing the simulated genomes with on-the-shelf phylogenomic tools. Using this new version, we let a population evolve for 800.000 generations. Then, we simulated a random branching process and simulated evolution along the branches up to generation 1.000.000. This results in 40 different populations that evolved for the same duration in the same conditions but that diverged in their past at random times. We extracted the genome of the 40 best final organism and sent them to Le Cocon for "double-blind" analysis.

Blind phylogenomic reconstruction: A first attempt to align the 40 sequences with MAFFT – handling, as most alignment softwares, local mutations (substitution, InDels) – gave no satisfying results, which convinced the inference team that it was necessary to account for rearrangements (duplication, inversion, translocation). We then used the Mauve sequence aligner, which segments the genomes into aligned pieces. The aligned pieces, scattered across all initial genomes, were concatenated to produce 40 aligned virtual sequences, which are each rearranged segments of the initial sequences. This alignment was given as input to IQtree with a "model test" option to let the program choose the inference model, resulting in an inferred tree. Importantly, none of the tools used integrate knowledge about the simulation software and the simulations have been produced without a priori knowledge about the tools operated by the inference team.
Comparison of the inferred tree with the ground-truth showed that its shape matched almost exactly, with three differences that correspond to the lower branch supports of the inferred tree. As far as we know, this is the first time an artificial life simulation software produced

*Intervenant