SPLACE: a tool to SPLit, Align and ConcatenatE genes for phylogenetic inference

Renato Renison Moreira Oliveira, Santelmo Vasconcelos, Guilherme Oliveira

Universidade Federal de Minas Gerais

Abstract

The production of phylogenetic trees containing multiple genes is best accomplished by concatenating all aligned genes into a supermatrix instead of generating a tree for each gene and then inferring the phylogeny by the consensus of all trees, i.e. a supertree. The advent of NGS technologies made it easier and cheaper to obtain multiple gene information from a large number of organisms of interest, generating a more robust supermatrix. The supermatrix, then, can be used in the phylogenetic reconstructions to generate a species tree. Many studies have used the supermatrix strategy to infer the phylogeny among species, such as when analyzing genomes from prokaryotic organisms and from eukaryotic organelle genomes (mitogenomes and plastomes). Building a supermatrix can be very time-consuming, especially if there is a large number of genes from many organisms to use in the analysis. Some published tools, such as SequenceMatrix, TaxMan, ScaFoS, TNT, and Phyutility, aim to concatenate gene files, but they require aligned gene files, what can be a problem with a large number of genes, as we mentioned. Here we present SPLACE, a tool to SPLit, Align and ConcatenatE the genes from all the species of interest to generate a supermatrix file, and consequently, a phylogenetic tree. To generate the supermatrix of n organisms, SPLACE will need n fasta files, each one containing all the g genes from a particular organism. First, SPLACE splits the genes from an organism, gathering the genes that have the same name from the n organisms into a single fasta file, therefore generating g new fasta files, each one containing the same gene from different organisms. Then, SPLACE aligns each one of the g fasta files using the MAFFT aligner, generating g' new fasta files. Finally, the genes in the g' fasta files that came from the same organism are concatenated into a single sequence, generating a single fasta file with the supermatrix containing n sequences, each representing one of the n organisms. Phylogeny is then reconstructed using the supermatrix fasta file. We used SPLACE to build a phylogenetic tree for all plant species with complete nuclear genome deposited on NCBI, using its respective chloroplast genes. The supermatrix was then submitted to CIPRES portal to generate a maximum likelihood phylogenetic tree using RAxML, with a bootstrap of 1000 replicates. The resulting phylogenetic tree showed the proper proximity among the plant species that belong to the same family.

Funding: 372439/2019-5, CNPq