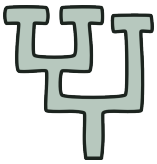
PhyPipe: an automated pipeline for phylogenetic reconstruction from multilocus sequences



Nicolás D. Franco-Sierra[1], Mateo Gómez-Zuluaga[2], Juan F. Díaz-Nieto[1], Javier C. Alvarez[1]

[1]Grupo CIBIOP, Departamento de Ciencias Biológicas, Universidad EAFIT. Medellín, Colombia. Email: nfranco@eafit.edu.co
[2]Centro de Computación Científica APOLO, Dirección de Informática, Universidad EAFIT. Medellín, Colombia.

Website: https://gitlab.com/cibiop/phypipe/wikis/home
Repository: https://gitlab.com/cibiop/phypipe

Phylogenetic reconstruction based on multiple loci of DNA sequences is nowadays a common task in many fields of biological research such as: ecology, epidemiology, evolution, and taxonomy, among many others. This process involves several methodological steps for going from a DNA sequence to a final output tree (i.e., sequence alignment, testing substitution models, concatenation of loci, phylogenetic reconstruction). Phylogenetic analyses require the use of different bioinformatic tools that demand human supervision in intermediate steps, in particular, it is necessary to edit input files with the output of previous analyses. A disconnected approach between those steps could lead to a time consuming and error-prone task, specially when multiple runs are required (e.g., same dataset and different run parameters or different datasets with or without varying the parameters). Thus, this seems to be a common problem in bioinformatic routines. Previous efforts have developed workflow management tools that can handle routines using different phylogenetic systematic software; nonetheless, none of such routines are particularly developed to tackle the problem of developing partitioned phylogenetic analyses. Our approach is particularly useful for reducing both error in intermediate steps and overall time in the phylogenetic reconstruction, in addition to facilitate the multilocus reconstruction when major or minor changes are needed to the initial dataset or running parameters.

In this work we present Phypipe, an automated pipeline written in a programming language for bioinformatic pipelines named bpipe [1], that aims to facilitate and speed-up phylogenetic reconstructions using multilocus concatenated analyses. Phypipe is intended to be used with multiple-locus data. The program receives as input multiple FASTA files (one per locus) containing DNA sequence data and it also supports previously aligned sequences. DNA data can be either from coding or non-coding regions from organelle or nuclear genomes.

The workflow consists of four major stages using the following dependencies:

- An alignment step (MAFFT and RevTrans for coding sequences)
- Concatenation of all alignment files in a single file (homescript in Python)
- Best partition scheme search (partitionfinder)
- Phylogenetic reconstruction using Bayesian Inference (BI) or Maximun Likelihood (ML). (MrBayes for BI, and Garli or RAxML for ML)

The routine implements wrapper scripts in Python for reading each stage's output file and creating the corresponding configuration file of subsequent steps. Often, these kind of analyses are performed by wet lab biologists or researchers with little computational experience, consequently an automated approach would be helpful to obtain results in a short time, avoiding the software manipulation and learning curve. This software is also useful for researchers in molecular systematics or bioinformaticians with previous experience in the included tools. Although detailed knowledge of the included software is not mandatory to run Phypipe, the potential of this pipeline would be maximized if the researcher is familiar with each independent program. Thus, researchers can modify software parameters before running the complete pipeline and obtain results while saving time.

[1]Sadedin, S. P., Pope, B., Oshlack, A. (2012) Bpipe: A tool for running and managing bioinformatics pipelines. *Bioinformatics* 28(11) 1525-6. doi:10.1093/ bioinformatics/bts167