The Journal of Open Source Software

Phylen: automatic phylogenetic reconstruction using the EggNOG database

Ignacio Ferrés[1] **and Gregorio Iraola**[1]

1 Unidad de Bioinformática, Institut Pasteur de Montevideo, Uruguay

DOI: 10.21105/joss.00593

Software

- Review
- Repository
- Archive

Submitted: 19 February 2018
Published: 04 May 2018

Summary

High-throughput sequencing is dramatically increasing the amount of genetic data available from all domains of life, but particularly from bacteria. The smaller size of bacterial genomes allows to sequence large collections of strains, mainly from species that deserve interest for their importance as human or farm animal pathogens. Phylogenetic analysis has become a standard tool to understand the evolutionary history, epidemiology and virulence of these bacteria, and the availability of genomic information has allowed to move from single-gene (e.g. using the 16S rRNA gene) to multilocus or core genome trees that bring us closer to a more reliable reconstruction of phylogenetic structure.

The EggNOG database (Powell et al. 2014) is an excellent resource providing orthologous groups shared at different taxonomic ranks including several prokaryotes. Here we present Phylen, a simple and automated software package written in R that reconstructs phylogenies by interacting with the EggNOG database. First, a set of orthologous groups available at the EggNOG database is selected and automatically downloaded or, alternatively, an external set of orthologous groups can be provided formatted as a Hidden Markov Model (HMM) file. Second, genome annotations in GFF3 format (such as those from Prokka annotation software (Seemann 2014)) are parsed to extract translated coding sequences. Third, genomes are screened against these orthologous groups using HMMER3 (Eddy 2011). Forth, "core" coding sequences are extracted and multiple sequence alignment is performed over each recovered gene set using MAFFT (Katoh and Standley 2013). Fifth, alignments are concatenated into a single supergene and phylogenetic reconstruction is performed using Maximum-Likelihood or distance methods (Fig. 1A). Phylen outputs one multi-fasta alignment per gene, one supergene multi-fasta alignment file, one tree file in Newick format and an object of class "phylo" which can be further analysed using the R packeges ape (Paradis, Claude, and Strimmer 2004) and phangorn (Schliep 2011).

Phylen has been already used by our group for building the *Helicobacter* genus phylogeny (Fresia et al. 2017) from a set of 40 universal marker genes (Mende et al. 2013), and to reconstruct core genome phylogenies of *Leptospira* genus (Puche et al. 2017; Thibeaux et al. 2018) from orthologous groups defined in the EggNOG database (spiNOG) (Powell et al. 2014). Additionaly, here we screened 93 *Epsilonproteobacteria* genomes against 4513 orthologous groups from the EggNOG database (eproNOG) to obtain the phylogenetic tree shown in Fig. 1B. In the near future we plan to add more functionalities such as different multiple sequence alignment algorithms and tools for alignment quality check and trimming.

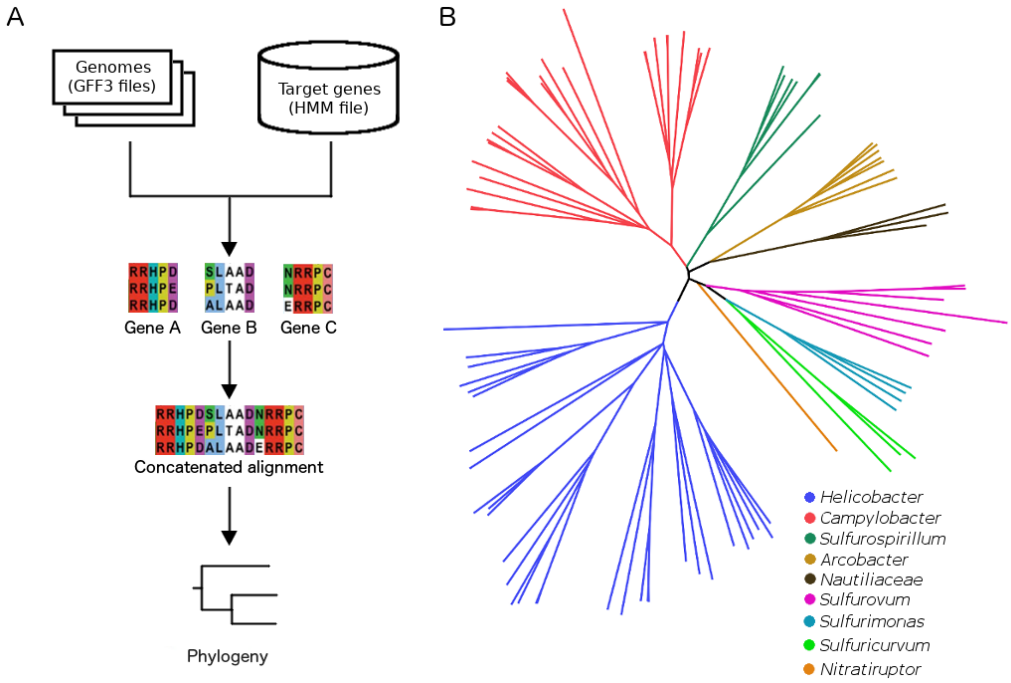

Figure 1: A) Schematic workflow of Phylen. B) Phylogeny of *Epsilonproteobacteria* obtained with the eggNOG database (eproNOG orthologs).

Phylen has been designed to facilitate the reconstruction of high-resolution phylogenies at any desired taxonomic rank, and from any set of genes like taxon-specific markers or the whole core genome. Phylogenetic reconstruction is a standard kick-off analysis in almost every comparative genomics project and despite many methods have been developed, Phylen is unique as it integrates the highly accessed EggNOG database (for phylogenetic marker genes) with the R environment as a widely used programming interface for microbial genomics and data analysis. Phylen depends on the R package phangorn (Schliep 2011) for phylogenetic reconstruction and external tools including HMMER3 (Eddy 2011) as gene search engine and MAFFT (Katoh and Standley 2013) for multiple sequence alignment.

Acknowledgements

I.F. was supported by ANII (Uruguay) postgraduation grant POS_NAC_2016_1_131079. We thank Pablo Fresia and Daniela Costa for testing Phylen.

References

Eddy, Sean R. 2011. "Accelerated Profile Hmm Searches." *PLoS Computational Biology* 7 (10). Public Library of Science:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Fresia, Pablo, Ronald Jara, Rafael Sierra, Ignacio Ferrés, Gonzalo Greif, Gregorio Iraola, and Luis Collado. 2017. "Genomic and Clinical Evidence Uncovers the Enterohepatic Species Helicobacter Valdiviensis as a Potential Human Intestinal Pathogen." *Helicobacter* 22 (5). Wiley Online Library. https://doi.org/10.1111/hel.12425.

Katoh, Kazutaka, and Daron M Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4). Society for Molecular Biology; Evolution:772–80. https://doi.org/10.1093/molbev/mst010.

Mende, Daniel R, Shinichi Sunagawa, Georg Zeller, and Peer Bork. 2013. "Accurate and Universal Delineation of Prokaryotic Species." *Nature Methods* 10 (9). Nature Publishing Group:881. https://doi.org/10.1038/nmeth.2575.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. "APE: Analyses of Phylogenetics and Evolution in R Language." *Bioinformatics* 20 (2). Oxford University Press:289–90. https://doi.org/10.1093/bioinformatics/btg412.

Powell, Sean, Kristoffer Forslund, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Jaime Huerta-Cepas, Toni Gabaldon, et al. 2014. "EggNOG V4. 0: Nested Orthology Inference Across 3686 Organisms." *Nucleic Acids Research* 42 (D1). Oxford University Press:D231–D239. https://doi.org/10.1093/nar/gkt1253.

Puche, Rafael, Ignacio Ferrés, Lizeth Caraballo, Yaritza Rangel, Mathieu Picardeau, Howard Takiff, and Gregorio Iraola. 2017. "Leptospira Venezuelensis Sp. Nov., a New Member of the Intermediates Group Isolated from Rodents, Cattle and Humans." *International Journal of Systematic and Evolutionary Microbiology*. https://doi.org/10.1099/ijsem.0.002528.

Schliep, Klaus Peter. 2011. "Phangorn: Phylogenetic Analysis in R." *Bioinformatics* 27 (4). Oxford University Press:592. https://doi.org/10.1093/bioinformatics/btq706.

Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14). Oxford University Press:2068–9. https://doi.org/10.1093/bioinformatics/btu153.

Thibeaux, Roman, Gregorio Iraola, Ignacio Ferrés, Emilie Bierque, Dominique Girault, Marie-Estelle Soupé-Gilbert, Mathieu Picardeau, and Cyrille Goarant. 2018. "Deciphering the Unexplored Leptospira Diversity from Soils Uncovers Genomic Evolution to Virulence." *Microbial Genomics* 4 (1). Microbiology Society. https://doi.org/10.1099/mgen.0.000144.