Automated tools for the generation and interpretation of single gene trees at a broad taxonomic scale

Denis Baurain[1], **Mick Van Vlierberghe**[1], **Arnaud Di Franco**[2], **Herve Philippe**[2]

[1] InBioS, University of Liege, Belgium; [2] SETE, UMR CNRS 5321, Moulis, Fance

mvanvlierberghe@doct.ulg.ac.be

Abstract

IDENTIFYING orthology relationships among sequences is fundamental in phylogenomics; indeed, those are essential to understand evolution, diversity of life and ancestry among organisms. To build alignments of orthologous sequences, phylogenomic pipelines often start with a step of all-vs-all similarity search followed by a clustering with an algorithm such as *OrthoFinder* [Emms and Kelly (2015) *Genome Biol* 16:157]. For it to be as accurate as possible, proteomes of good quality are needed but their availability is limited to a small subset of the living beings. Therefore, large-scale taxonomic phylogenomic analyses imply the enrichment of preexisting orthologous groups with transcriptomic or genomic data and the need for robust tools for identifying orthologues from heterogeneous sequence data. To this end, we have developed a novel tool, "Forty-Two", along the lines of HaMStR [Ebersberger et al. (2009) *BMC Evol Biol* 9:157], whose aim is to add (and optionally align) sequences to thousands of preexisting multiple sequence alignments (MSA) while controlling for orthology relationships and potentially contaminating sequences. "Forty-Two" uses advanced heuristics based on a multiple Best Reciprocal Hit (multi-BRH) strategy against reference proteomes to distinguish orthologous and paralogous sequences among homologues. It is fully functional and has already been used in two high-profile phylogenomic manuscripts (under review) dealing with the animal tree of life. Here, we present the principles and algorithms underlying "Forty-Two" as well as the results of an extensive test suite of its features, in order to support its release to the public.

Workplan

Benchmark test for recovery of depleted orthologuous groups (MSA)

1. Dataset
   (a) 57 organisms (quality proteomes) in 8 taxonomic groups (Figure 1)
   (b) Define orthologous groups (MSAs)
   (c) Classify orthogroups (Table 1)
2. 'Forty-Two'
   (a) For each depleted orthogroup:
   (b) Try to add back the removed sequences
   (c) Compute statistics (TPs, FPs, TNs, FNs)
   (d) Two runs of removal: groups and & supergroups

		Max copy mean		
		1.1	1.2	2.0
Distribution	Universal	17	33	36
	Widespread	76	130	59
	Sparse	32	60	37

Table 1: *Number of orthogroups (MSAs) according to classification criteria. A taxonomic criterion (few intra-group representatives ("Sparse"), almost all intra-group representatives ("Widespread") and all the representatives ("Universal")) and a maximum copy mean number criterion (1.1, 1.2 and 2.0)*

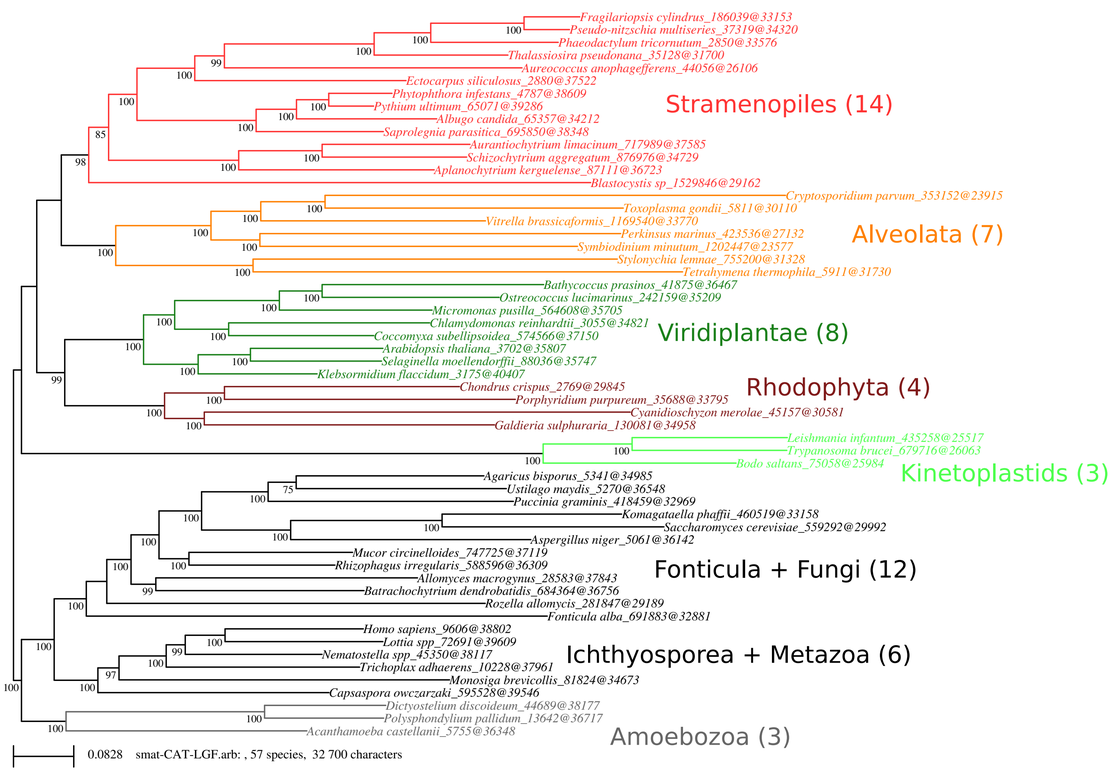

Figure 1: *RAxML tree built using our 57 selected species (model PROTCATLGF) showing the phylogenetic groups used for the test.*

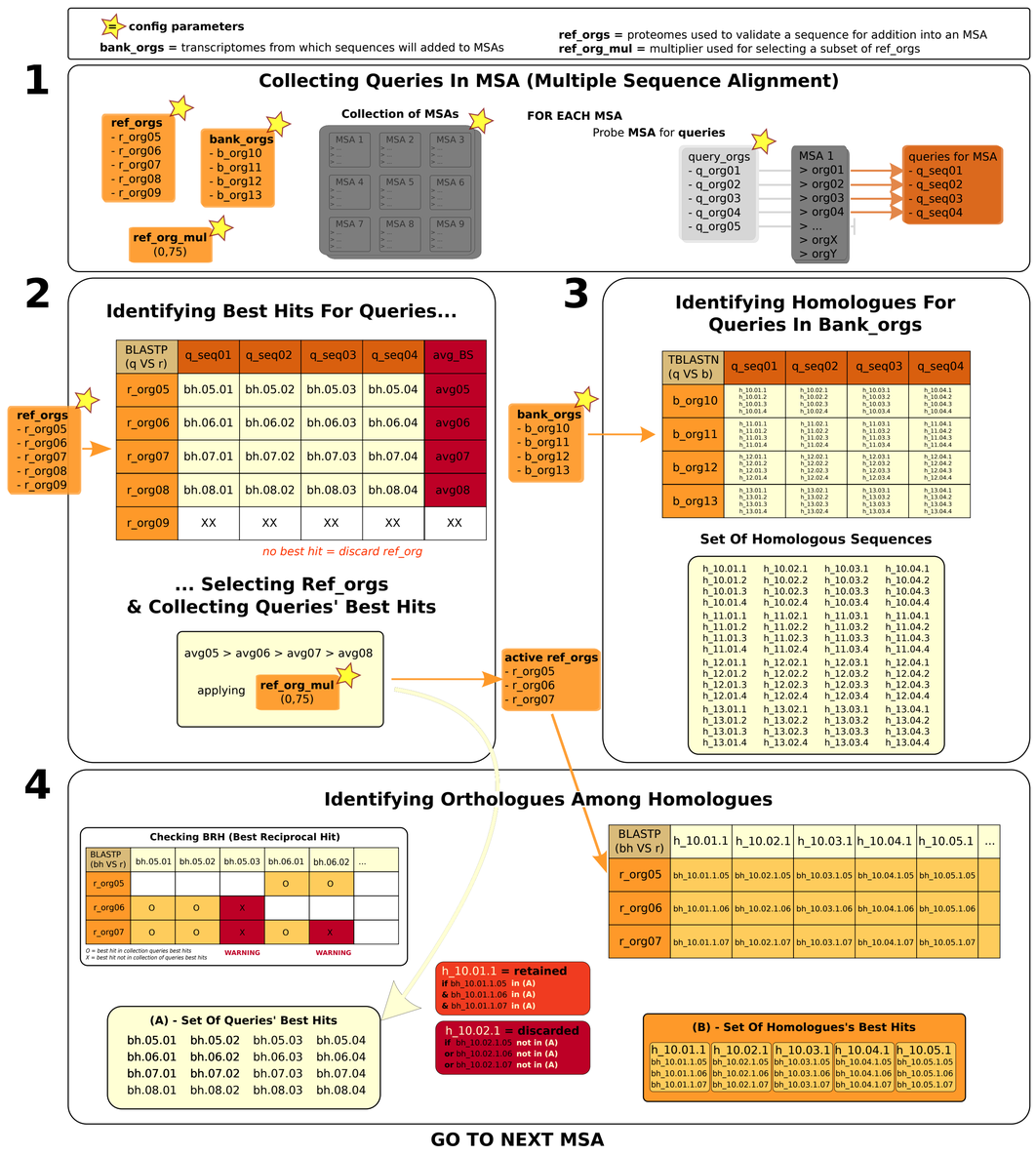

Figure 2: *Diagram showing heuristics behind 'forty-two'.*

Figure 3: *PR-curves for MSAs with few intra-group representatives ("Sparse"), almost all intra-group representatives ("Widespread") and all the representatives ("Universal", 57 species).*

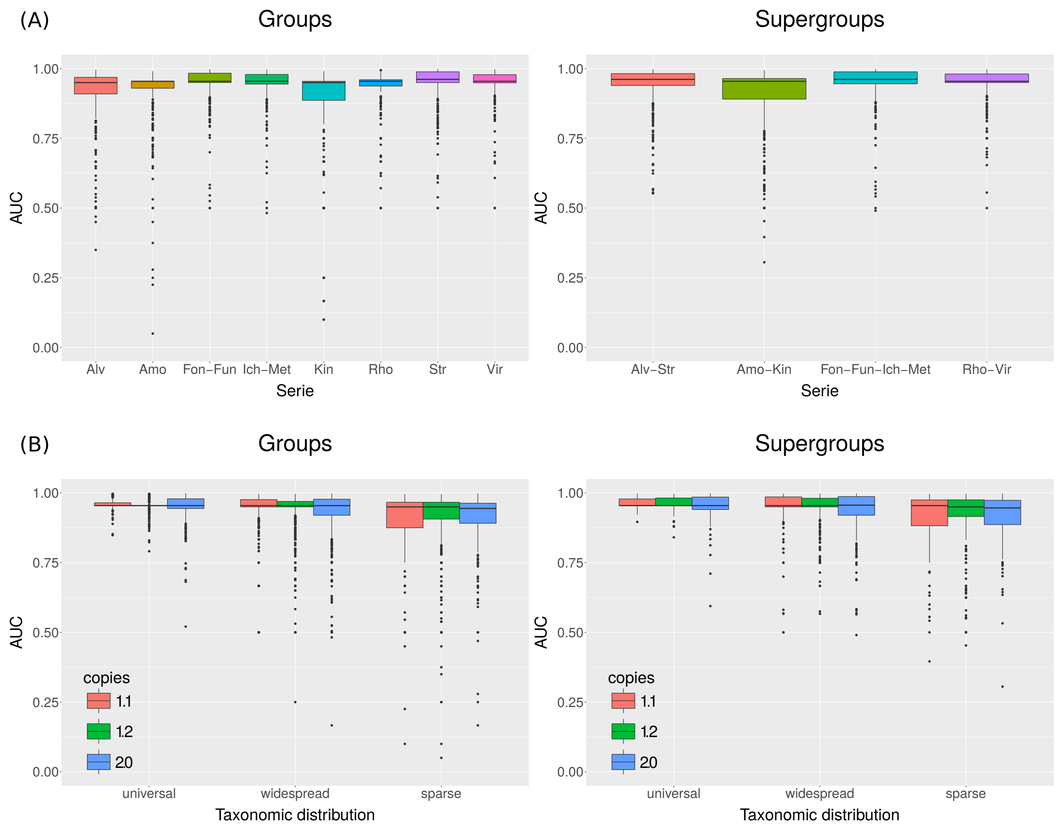

Figure 4: *Distribution of computed areas under curve for each orthogroup for each each of the two runs ("groups" and "supergroups") relative to (A) "groups" and "supergroups" and to (B) taxonomic distribution within orthogroups and number copies per MSA.*

Perspectives

Further testing is planned; indeed it will be interesting to try to enrich public orthogroups with biased taxonomic sampling and also to compare performances versus *HaMStR* [*BMC Evolutionary Biology* 2009 9:157] which is also able to search for othologs in ESTs.

M. Van Vlierberghe is a Ph.D. student funded by the F.R.S.-FNRS.