Poster H-86

GREENPHYL: A Generic Phylogenomic Pipeline for Ortholog Prediction Between two Model Plant Species Arabidopsis thaliana and Oryza sativa

Authors:
Matthieu Conte (*CIRAD*)
Sylvain Gaillard (*CIRAD*)
Brigitte Courtois (*CIRAD*)
Manuel Ruiz (*CIRAD*)
Emmanuel Guiderdoni (*CIRAD*)
Christophe Perin (*CIRAD*)

Short Abstract: Gene ortholog identification is now a major objective for mining the increasing amount of sequence data provided by full or partial genome sequencing projects. Currently, most of the methods available for functional prediction are based on sequence similarity even if this kind of annotation transfer is often misleading. We developed GREENPHYL, an optimized phylogenomic pipeline which can generate weighted-supported orthologs and paralogs relationships to provide a robust and accurate way for gene function assignment.

Long Abstract:
Gene ortholog identification is now a major objective for mining the increasing amount of sequence data provided by full or partial genome sequencing projects. Comparative and functional genomics in non model species urgently need a way to identify orthologs in model species to shorten gene function inference. Currently, most of the methods available for functional prediction are based on sequence similarity since sequence similarity often implies function similarity. Unfortunately, sequence similarity does not always imply ortholog correspondence and thus direct annotation transfer is often misleading. Gene functions change as the result of evolution, and reconstructing the evolutionary history of genes should be a more accurate way to differentiate orthologs from paralogs. Phylogenomic takes into account phylogenetic information from high-throughput genome annotation but pipelines for automatic detection of orthologs are still scarce and suffer from several limitations. We developed GREENPHYL, an optimized phylogenomic pipeline which can generate ortholog and paralog relationships, adapting several steps to provide robust and accurate gene function assignments. GREENPHYL follows the well established phylogenomic analysis structure: Family clustering (MCLpipeline, Interproscan, LEON) - Alignment (MAAFT, RASCAL) – Tree construction (PHYML) – Tree analysis (RIO). Each of these open source softwares and parameters have been optimized to infer an optimal evolutionary tree. Moreover, contrasting with other phylogenomic analysis pipeline, GREENPHYL includes an automatic analysis of the generated tree. This detailed analysis allows detection of orthologs, paralogs and also other interesting phylogenomic relationships like subtree-neighboring. These predictions are supported by bootstrap values and can be easily filtered. Full genome sequences of the two plant model species, Oryza sativa and Arabidopsis thaliana are now available. We evaluated GREENPHYL performances against a set of published genes

already functionally characterized in these two species. GREENPHYL achieved high accuracy level in predicting ortholog/paralog relationships for experimentally characterized proteins. Our pipeline detects interesting new relationships that can be used as a starting point for functional characterisation analysis and comparative genomics. GREENPHYL is also able to detect specific family expansion or specific species clades. Our results illustrate the power of GREENPHYL for gene function assignment and comparative genomics. Whole phylogenomic analysis of Oryza sativa and Arabidopsis thaliana proteomes is now in progress.