BeNeLux Bioinformatics Conference – Brussels, December 2013
Abstract ID: **000** Category: Abstract template

IDENTIFYING LOSSES AND EXPANSIONS OF SELECTED GENE FAMILIES IN INCOMPLETE GENOMIC DATASETS

Arnaud Di Franco[1,2 *], *Marc Hanikenne*[2,3], *Denis Baurain*[1,2].
[1]*Eukaryotic Phylogenomics, Department of Life Sciences, University of Liège, B-4000 Liège, Belgium ;* [2]PhytoSYSTEMS, University of Liège, B-4000 Liège, Belgium ; [3]*Functional Genomics and Plant Molecular Imaging, Center for Protein Engineering (CIP), Department of Life Sciences, University of Liège, B-4000 Liège, Belgium ; * arnaud.difranco@gmail.com*

Plantae (Archaeplastida) are a natural group of organisms with plastids of primary endosymbiotic origin. Within this group, members of the red algae show evidence of a reduction of their genomic content. In this work, we designed a bioinformatics approach to investigate the few, sometimes incomplete, genomic datasets available for red algae, with the purpose of pointing out possible gene family losses and expansions. Our pipeline first populates a relational database with precomputed orthology relationships between green plant genomes and red algal datasets and then efficiently queries the database for computing statistics of losses and expansions for a series of gene families of interest.

INTRODUCTION

Primary plastids have been acquired by eukaryotes through a unique event of endosymbiosis with a cyanobacteria. This ancestral photosynthetic eukaryote gave rise to three monophyletic groups, glaucophytes, green plants and red algae. The latter have generally smaller genomes with less coding genes compared to the two others (especially green plants). We suppose that those losses are due to specific environmental conditions having affected the ancestor of red algae. To prove our hypothesis, we developed a pipeline to examine presence/absence of gene families of interest in this group for which only few genomes are available.

METHODS

We built an inventory containing the genes of interest by searching UniprotKB with specific keywords. Each gene was assigned to one or more categories depending on its functional annotations. Meanwhile, we collected the genomes of several green plants and defined orthologous groups (OGs). To consider only the gene families that were present in the common ancestor of Plantae (POGs), we used a taxonomic filter ensuring a minimal representation of each child lineage. Then, we used our inventory to partition the POGs into annotated and anonymous gene families. To identify the genes of interest in red algae, we built Hidden Markov Model profiles (pHMMs) from the aligned POGs and searched for each pHMM in the red algal genomic datasets. Finally, we used a BLAST Best Reciprocal Hit criterion (BRH) to discard matches to paralogous genes, in which, to be retained as orthologous, each red algal hit had to match back to a green plant sequence belonging to the pHMM.

Orthology relationship were loaded into the database, along with the tales about the inventory, functional categories and pHMMs properties. At first, we queried the database to retrieve the groups of pHMMs corresponding to the various categories of our inventory. These real groups (RGs) were used to look for orthologous genes into red algae datasets. To study losses, counts of red algal orthologs were converted to boolean values whereas all orthologs were counted individually to study family expansions. Then, to verify whether the observed losses/expansions were significant, we compared the counts obtained for each category to background distribution generated from 1000 control groups of pHMMs (CGs). CGs pHMMs were selected so as to match the properties of the RGs pHMMs. Hence, for each pHMMs properties, we performed a Kolmogorov-Smirnov test (KS) of its distribution in each of our RGs versus all of our pHMMs. To select the properties that we have to take into account when assembling the CGs, we sorted them by the geometric mean of their p-values over the different RGs. This allowed us to retained the seven most critical properties. Each CG was assemble as follows: (1) a number of pHMMs equal to the size of the RG are picked at random, (2) A KS is carried out for each of the seven properties between the RG and the candidate CG, (3) if the geometric mean of the seven KS p-values is greater than 0.05, the CG is accepted; otherwise the algorithm goes back to step 1.

RESULTS & DISCUSSION

Losses/expansions were investigate for each of our functional categories in five genomes of red algae. While gene losses were difficult to confidently identify, some interesting family expansions were detected (Figure 1). Further, we investigated if the approach also worked on transcriptomic datas.

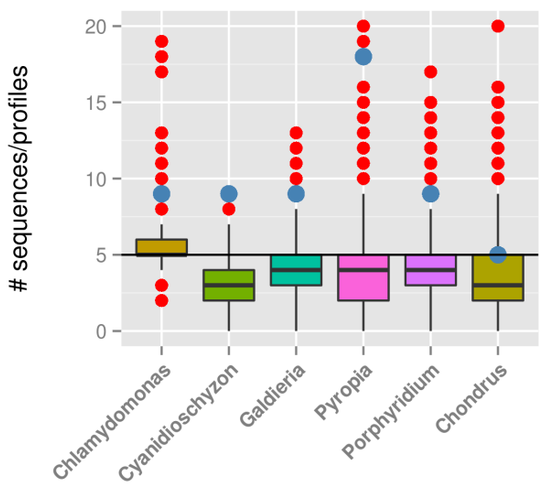

Figure 1: Expansions in iron transport related gene families in five species of red algae.