functional Genomics

Conte M.G[1], Laporte M.A[2], Perin C[3], Rouard M[1]

[1]Bioversity International - Commodities for Livelihoods programme Parc Scientifique Agropolis II, 34397 Montpellier - Cedex 5, France
[2] CEFE - UMR 5175 - 1919 Route de Mende - F34293 Montpellier cedex 5, France
[3] CIRAD, Department BIOS, UMR DAP - TA40/03, 34398 Montpellier, France

Nowadays, most of the manual annotation in biology is done on gene sequences or protein patterns but relatively little is done for gene families at large. However, a proper catalogue of homeomorphic gene families, genes that evolved from a common ancestor and sharing full-length sequence similarity and common domain architecture, would be a valuable resource for evolution studies and orthologs inference. GreenPhylDB v2.0a contains groups of protein-coding gene sequences automatically clustered from 12 complete genomes of plants (fig. 1) that cover most of the taxonomy of green plants. Each cluster is first manually checked and then analyzed by a phylogeny approach to predict orthologs. We add value with several annotations, including family names defined via a consensus from existing gene and protein pattern annotations (e.g. UniProt, InterPro, Pirsf, Kegg, GO) for the sequences composing the clusters.
Here, we present our methodology and annotation tool for the curation of Protein-coding gene sequence families, a critical step before any phylogenetic analysis.

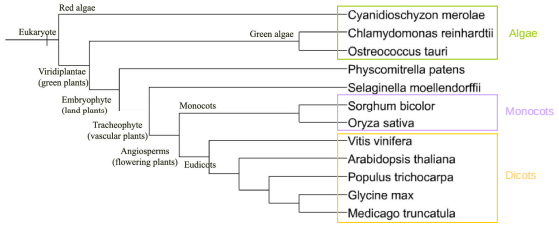


Figure 1. List of plant genomes analysed in GreenPhylDB. By integrating genomes representing on a broad taxonomy (from algae to angiosperms), we expect to define consistent and comprehensive set of homeomorphic plant families.

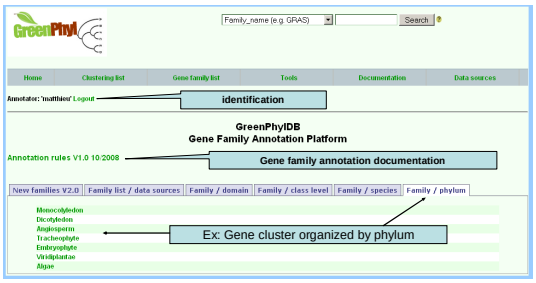


Figure 2. The database contains approximately 25,000 clusters spread over 4 levels of stringency. Cluster lists can be displayed using various filters such as phylum, species, protein domains specificity etc (fig. 7).

Clusters annotation in 3 steps

1

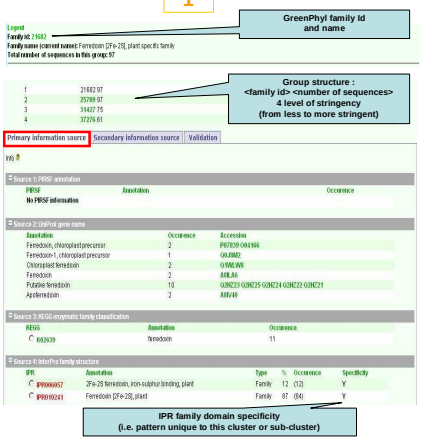


Figure 3. Primary source of information. This section sums up high quality annotations available in external databases for protein sequences of a cluster. Some calculations have been made to spot clusters with specific InterPro family motifs. Family names rely mostly on this information.

2

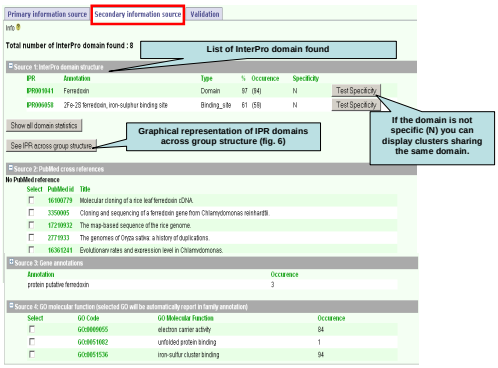


Figure 4. Secondary source of information. If the primary source is not sufficient, complementary information may come from non-family InterPro domain and their distribution in the sub-clusters (fig. 6). For a non-specific domain, a test button appears allowing searches within clusters with sequences bearing the same IPR motif. According to the number of occurrence, a domain shuffling may be identified. In such case, an IPR domain could be finally considered as specific. Clusters can be flagged by proposed Pubmed and GO references (linked to IPR and UniProt entries).

3

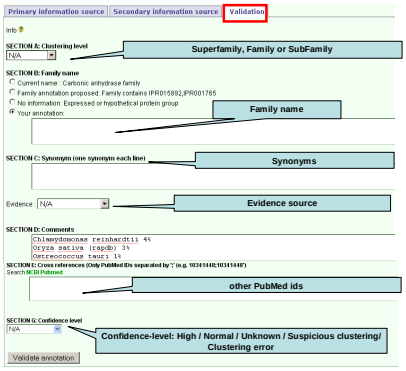


Figure 5. Validation step. Curators will finally propose a name and synonyms for this cluster and decide if the cluster may be considered as a superfamily, family or subfamily. The evidence used to define the name and a confidence level have to be specified. A free text area is provided to encourage additional comments on the cluster. The curator validates the annotation which will be versioned and monitored by an admin before insertion into the database.

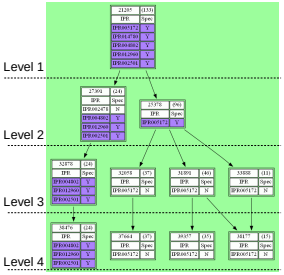
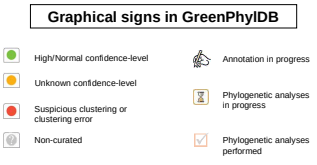


Figure 6. Example of distribution of InterPro domains across subclassification for a specific cluster GreenPhyl DB . These schemas are automatically generated by the interface



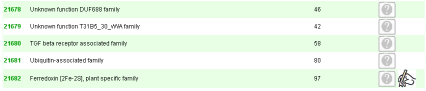


Figure 7. Example of a cluster list composed of the family id, family name, number of sequences, confidence-level and status of annotation.

References: (http://greenphyl.cirad.fr)

Conte MG, Gaillard S, Droc G, Périn C (2008). Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. BMC Genomics, 2008, 183

Conte MG, Gaillard S, Lanau N, Rouard M, Périn C (2008). GreenPhylDB: a database for plant comparative genomics. Nucleic Acids Research. 2008 January; 36 D991–D998

Contacts:
Mathieu Rouard (m.rouard@cgiar.org)
Matthieu Conte (m.conte@cgiar.org)