

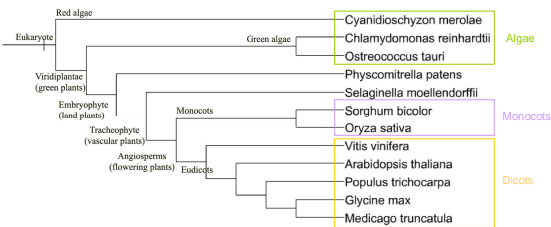
# functional Genomics

Conte M.G<sup>1</sup>, Laporte M.A<sup>2</sup>, Perin C<sup>3</sup>, Rouard M<sup>1</sup>

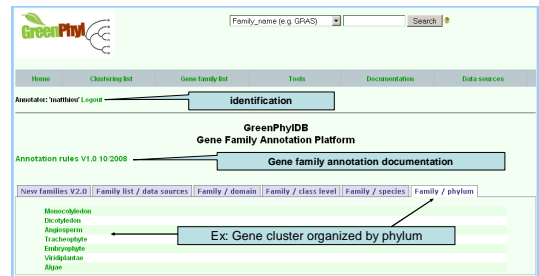
<sup>1</sup>Bioversity International - Commodities for Livelihoods programme Parc Scientifique Agropolis II, 34397 Montpellier - Cedex 5, France  
<sup>2</sup>CEFE - UMR 5175 - 1919 Route de Mende - F34293 Montpellier cedex 5, France  
<sup>3</sup>CIRAD, Department BIOS, UMR DAP - TA40/03, 34398 Montpellier,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

GreenPhyl family id and name

Group structure : <family id>-<number of sequences>-4 level of stringency (from less to more stringent)

Primary information source | Secondary information source | Validation

Legend: Family id: 1001, Family name: Content names: Ferredoxin [2p-2], plant specific family, Total number of sequences in this group: 97

1	2182.97
2	2039.97
3	1442.75
4	3729.61

Source 1: IPR annotation

IPR ID	Annotation	Occurrence
IPR00001	Ferredoxin	11

Source 2: UniProt gene names

Annotation	Occurrence	Accession
Ferredoxin, chloroplast/precursor	2	PF01291-040100
Ferredoxin, chloroplast/precursor	1	Q9M862
Chloroplast ferredoxin	2	Q9M868
Ferredoxin	2	Q9M868
Plaque ferredoxin	11	Q9M821-Q9M825, Q9M824, Q9M822, Q9M821
Arabidopsis	2	At04010

Source 3: NCBI taxonomic family classification

REFS	Annotation	Occurrence
100020	Ferredoxin	11

Source 4: InterPro family structure

IPR	Annotation	Type	% Occurrence	Specificity
IPR00002	2p-2b ferredoxin, non-subunit binding plant	Family	12 (12)	Y
IPR00001	Ferredoxin [2p-2], plant	Family	87 (86)	Y

IPR family domain specificity (i.e. pattern unique to this cluster or sub-cluster)

**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**

Primary information source | Secondary information source | Validation

Total number of InterPro domain found: 8

List of InterPro domain found

IPR	Annotation	Type	% Occurrence	Specificity
IPR00001	Ferredoxin	Domain	87 (84)	N
IPR00002	2p-2b ferredoxin, non-subunit binding site	Binding site	61 (58)	N

Graphical representation of IPR domains across group structure (fig. 5)

If the domain is not specific (N) you can display clusters sharing the same domain.

Source 2: PubMed cross references

Source 3: Gene annotations

Annotation	Occurrence
Gene: cytochrome ferredoxin	3

Source 4: GO molecular function (selected GO will be automatically filtered by stringency)

GO Code	GO Molecular Function	Occurrence
GO:0060605	electron carrier activity	84
GO:0061062	undefined protein binding	1
GO:0061026	non-catalytic cluster binding	84

**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**

Primary information source | Secondary information source | Validation

SECTION B: Clustering level

SECTION D: Family name

SECTION C: Synonyms (one synonym each item)

Evidence source

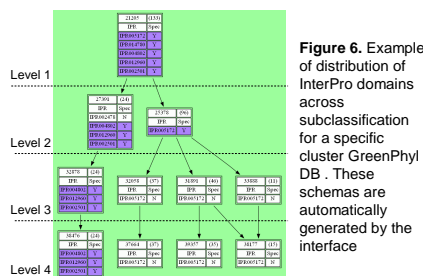
SECTION E: Comments

SECTION F: Cross references (Only PubMed IDs separated by ";" (e.g. 10341448;10341447))

SECTION G: Confidence level

Confidence-level: High / Normal / Unknown / Suspicious clustering / Clustering error

**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.



### Graphical signs in GreenPhylDB

- High/Normal confidence-level
- Unknown confidence-level
- Suspicious clustering or clustering error
- Non-curated
- Annotation in progress
- Phylogenetic analyses in progress
- Phylogenetic analyses performed

21629	Unknown function DUF688 family	45	
21629	Unknown function T3185_30_WA family	42	
21600	TGF beta receptor associated family	58	
21601	Ubiquitin-associated family	80	
21602	Ferredoxin [2p-2], plant specific family	97	

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