

# Cyanorak - Help file main page

Software written by Alexis Dufresne; Help file written by Frédéric Partensky

Cyanorak is an in-house database system, accessible through a web front-end, which was set up to manually refine the annotation of orthologous protein families (or ‘clusters’) of the first 11 sequenced marine *Synechococcus* (BL107, CC9311, CC9605, CC9902, RCC307, RS9916, RS9917, WH5701, WH7803, WH7805, WH8102) as well as the 3 first published *Prochlorococcus* genomes (MED4, SS120 and MIT9313). It makes it possible to export gene/protein Fasta files or Artemis annotation files for any of these 14 genomes. When present, the cluster manual annotation (including changes in gene names or products, gene starts, addition of unmodelled genes, etc.) will prevail over previous annotations as are available in Genbank or other databanks. This is particularly useful for a number of genomes (in particular from the Venter Institute) which have been released quickly after sequencing, and often only have had an automatic annotation in Genbank. Comparative analyses of these 14 genomes are reported in Dufresne et al. 2008, *Genome Biology*, in press. See also Six et al. 2007, [Genome Biology 8: R259](#) for a specific description and comparison of phycobilisome genes.

The current version of Cyanorak is read-only and allows to retrieve any cluster of orthologs via different menus. If you see any obvious mistake in the annotation of a cluster of orthologs, please contact Frédéric Partensky ([partensky@sb-roscoff.fr](mailto:partensky@sb-roscoff.fr)) and we will manually correct the annotation within a few days. Indeed Cyanorak is in constant evolution. Each time we modify a cluster annotation, it is immediately included in the Artemis annotation files of all genomes which possess a member of this cluster.

## Main Page

### 1) Menu “Find a gene”

Find all genes entries containing a specified term (ORF ID, gene name, part or totality of a product). Search in the annotation data as are currently deposited in Genbank.

The scrolldown menu allows you to enter either:

- an ORF ID (or locus tag, e.g. SYNW2000; BL107\_10222; PMM0107, etc.)
- a gene name (e.g. *psbA*, *uvrA*, etc.)
- or a gene product (e.g. shikimate kinase, ribulose-phosphate 3-epimerase)

Then click “Go”

Note : you can enter complete or partial names:

*Example 1* : If you enter “uvr” as a gene name, you will have access to all *uvr* genes: *uvrA-D*

*Example 2* : If you enter ‘ribulose’ or even ‘ribul’ as a product, you will have access to all hits which have ‘ribulose’ or ‘ribul’ in the product definition, e.g. ‘ribulose bisphosphate carboxylase, small chain’ or ‘ribulose-phosphate 3-epimerase’, etc.).

## 2) Menu “Find a cluster”

Find all clusters entries containing a specified term (ORF ID, gene name, part or totality of a product). Search in the annotation annotations made **manually** by annotators often **after** the deposition of genomes into Genbank. The scrolldown menu allows you to enter either:

- a cluster number (i.e. a number allocated to one of the clusters of orthologs defined in cyanorak, e.g. 1, 60, 8002)
- an ORF ID (or locus tag, e.g. SYNW2000, BL107\_10222; PMM0107) of one gene of a particular cluster
- a gene name (e.g. psbA, uvrA, etc.). given to a cluster
- or a gene product (e.g. shikimate kinase, ribulose-phosphate 3-epimerase) given to a cluster

Then click “Go”

Note : you can enter complete or partial names (see examples above)

## 3) Menu “Find a COG”

Find all Clusters of Orthologous Genes ([COG](#)) entries containing a specified term (COG number). Search in the annotation data as are currently deposited in Genbank..

Enter a COG number (e.g. COG0703), then click “Go”

## 4) Menu “Find a CyOG”

Find all CyOG entries containing a specified term (CyOG number).A CyOG is a Cyanobacterial Cluster of Orthologous Genes, as defined by [Mulkidjanian AY, et al. \(2006\) PNAS 103:13126-13131](#)).

Enter a CyOG number (e.g. CyOG00650), then click “Go”

## 5) Menu “Find an EC number”

Find all Enzyme Commission (EC) number entries containing a specified term (EC number). Search in the annotation data as are currently deposited in Genbank.

Enter a EC number (e.g. 2.7.1.71), then click “Go”

## 5) Menu Blast site

Allows accessing a Blast site with all 23 marine published picocyanobacteria genomes

## 6) Menu Export : Opens the Export page (see below)

## Cyanorak - Helpfile Export page

**List of genes and clusters:** provides a complete list of genes available in Cyanorak with their ORF ID (or locus tag), cluster number and product description

⇒ copy/paste the page into e.g. MS Excel for formatting it into separate columns.

**List of phyletic patterns:** provides a complete list of clusters available in Cyanorak indicating their description (if any) and distribution in the different genomes

Note: import this file in e.g. MS Excel for formatting

**Protein sequences:** provides the complete list of proteins (under fasta format) for any of the genomes of the scrolldown menu

**Artemis annotation file:** provides the annotated genome file (under Artemis format) for any of the genomes of the scrolldown menu

⇒ Note that these Artemis files contain only CDS (no tRNA or rRNA, but those can easily be retrieved for Genbank annotation files).

## Cyanorak – Helpfile Search cluster Page

1) When using the “find a cluster” menu, you are directed to the “search cluster” page:

**Top of page:** cluster manual annotation (if any), which includes the following fields:

- Cluster: indicates the cluster number (always present)
- Gene name: indicates the manually assigned gene name for this cluster (if any)
- Product: indicates the manually assigned product description for this cluster (if any)
- EC number (if any): this information is not included in the exported Artemis file
- GO terms (if any): idem
- Comment: includes comments on this cluster (if any): idem
- Annotator: mentions the author of the last manual cluster annotation (if any)
- Date : indicates the date the last manual cluster annotation (if any)

**Middle of page:**

- Phyletic pattern, showing the distribution and number of members of this protein family in the different genomes. Note that paralogs or multiple copies of a gene which are not sufficiently distinct to be differentiated are included in the same cluster. However, if they can be differentiated (i.e. are more different in a given genome as they are in different genomes), they are found in different clusters.
- Link to retrieve sequences of amino acids (AA) or nucleotides (NT) of all members of the clusters in fasta format
- Genome context: showing the immediate genome environment of all members of the cluster. A random color code is generated with members of a same cluster having the same color.

- History: showing changes in the assignment of a particular gene to clusters

**Bottom of page:** listing of individual members of the clusters. This includes the following fields:

- ORF ID (= locus tag): these are the same as in Genbank, unless genes were missed in the genome file deposited Genbank. In that case, ORF ID is a number attributed by our software having the same strain-specific prefix as other ORFs from the same strain followed by a number always larger than previously attributed ones.
- Strain: composed of a 3-letter code (Pro for *Prochlorococcus* or Syn for *Synechococcus*) followed by the strain number.
- Gene name: if there is a gene annotation in Genbank then it is shown in this field. Otherwise the field contains the ORF ID.
- Product: Contains the product annotation as in Genbank. Note that it can be different from the cluster annotation (top of screen) which (if present) will be included in the Artemis file.
- COG number: contains a link to a page describing the COG for which this gene has a hit (if any). The page also lists all other clusters matching the same COG
- CyOG number: same with CyOGs
- Blast link: points toward pre-computed results of BLAST searches that have been completed for every protein sequence in the Entrez Proteins data domain.