FPC-DB: The Fungal Protein Cluster Database

Jaehee Jung[1], Gangman Yi[1], Serenella A. Sukno [2], and Michael R. Thon [2]

[1]Texas A&M University, Department of Computer Science,
College Station, TX. 77845, USA.

[2] Centro Hispano-Luso de Investigaciones Agrarias (CIALE), Department of Microbiology and Genetics,
University of Salamanca, Salamanca 37007, Spain.

Summary

Our ability to predict gene function for newly sequenced fungal genomes is based to a large extent on comparisons to other organisms. It is important when making these comparisons to take into account the large evolutionary distances, sometimes hundreds of millions of years, that separate many of the fungal taxa for which whole genome sequences exist. It is commonly accepted that orthologous proteins tend to have conserved function in different species, however numerous examples that demonstrate functional diversification among orthologous genes exist. The phylogenetic context of a gene family must be taken into account when transferring knowledge of gene function from other species. The Fungal Protein Cluster Database (FPC-DB) is designed to enable users to quickly locate information about the evolutionary history and the functions of fungal proteins and gene families and present the information in a number of comparative views. The proteins from fungal whole-genome sequencing projects are annotated with InterPro terms, SwissProt keywords, and Gene Ontology terms using an automated functional classification server previously developed in our group. The proteins are then clustered into putative gene families using the MCL application. Each protein cluster includes multiple sequence alignments, phylogenetic trees and a summary of functional categories found in the cluster. Users can perform queries using proteins IDs, protein functional categories, and species names as search terms. The results are displayed in a variety of forms that allow users to compare the occurrence of gene and their functional categories between species. Users can also view a detailed page for each protein cluster and simultaneously view a phylogenetic tree and a matrix of functional categories, enabling them to identify cases of lineage specific gains and losses of functional categories.

More Information

http://bioinformatics.usal.es

http://fcg.tamu.edu

Data Sources and Analyses

- Automated functional annotation of fungal proteins using AAPFC (Jung et al 2007). The proteins are annotated with protein domains (InterPro) and Gene Ontology Terms (GO terms).
- Proteins are clustered to form gene families using the MCL application.
- A phylogenetic tree is constructed for each cluster using the neighbor-joining method.

Export and Analyze

- Export Data in Table Format: import table to excel for comparative analyses of gene content across species. Display and analyze changes in gene family size using other programs.
- Export protein and DNA sequences of each cluster: perform your own alignments and phylogenetic analyses.
- Data from the cluster table view can be used for downstream analyses in other programs. In this example, data from the cluster table view were exported to the CAFE program, which was used to reconstruct and test gene family sizes in internal branches of the species tree.
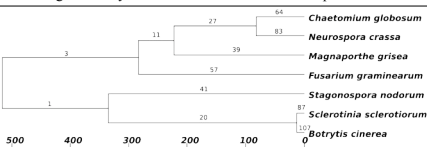
- Data were exported from the cluster table view and imported to a spreadsheet with copy/paste. The data were saved in a format suitable for analysis with CAFE.
- CAFE reconstructs ancestral gene family sizes and performs a test to help identify significant expansions/contractions in size. Numbers over branches represent the number of families with significant expansions or contractions.

P-value	Expansion or Contraction	Family ID	InterPro Categories
0.00E+00	C	63	Di-copper centre-containing; Tyrosinase
0.00E+00	C	68	Chitin-binding, type 1; Polysaccharide deacetylase; Cellulose-binding region
0.00E+00	C	70	Carbohydrate-binding WSC
0.00E+00	C	20	Heterokaryon incompatibility; Tetratricopeptide-like helical; Kinesin light chain; Disease resistance protein; Protein prenyltransferase; NACHT nucleoside triphosphatase
0.00E+00	C	330	UNKNOWN
0.00E+00	C	328	UNKNOWN
0.00E+00	C	107	Proteinase inhibitor, propeptide; Peptidase S8 and S53, subtilisin, kexin, sedolisin
1.98E-06	C	2	Major facilitator superfamily MFS_1
2.18E-05	C	30	Glycoside hydrolase, family 61; Cellulose-binding region, fungal
2.92E-04	C	45	General substrate transporter; Sugar transporter superfamily; Major facilitator superfamily
3.36E-08	C	3	CFEM
5.67E-06	C	24	Heterokaryon incompatibility
8.94E-04	C	142	UNKNOWN
0.00E+00	E	115	Integrase, catalytic region; Polynucleotidyl transferase, Ribonuclease H fold; Peptidase aspartic, catalytic; Reverse transcriptase; RNA-dependent DNA polymerase
0.00E+00	E	1646	UNKNOWN
0.00E+00	E	1360	Ankyrin; Heterokaryon incompatibility
0.00E+00	E	428	Glycoside hydrolase; Pectolytic enzyme, Pectin lyase fold; Virulence factor, pectin lyase fold
1.78E-05	E	927	Phosphotransferase system, HPr serine phosphorylation site
3.87E-07	E	382	UNKNOWN
4.09E-05	E	175	Glycoside hydrolase, family 71; Peptidase S8 and S53, subtilisin, kexin, sedolisin

This table shows each significant cluster from the Leotiomycetes lineage.

FPC-DB Query Page

- Search for proteins by InterPro term (protein family or functional domain), Gene Ontology Term, or description.
- Use Boolean search operators (AND, OR, NOT)
- Limit searches to species.

Search Results Summary

The search results summary page gives you a quick overview of your search results. Quickly find the number of proteins, clusters, GO terms, etc. that are in your search results.

Search Results Are Organized In Views

Each view presents the same search results, but organized in different ways.

Future Plans

- Add SwissProt keyword annotations to proteins
- Test each cluster for evidence of positive selection
- Test each cluster for significant changes in gene copy number
- Identify orthologs/paralogs in each cluster

References

Jung, J. and Thon, M.R. (2006) Automatic annotation of protein functional class from sparse and imbalanced data sets, *Lecture Notes in Computer Science*, 4316, 65-77.

De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution, *Bioinformatics*, 22, 1269-1271.
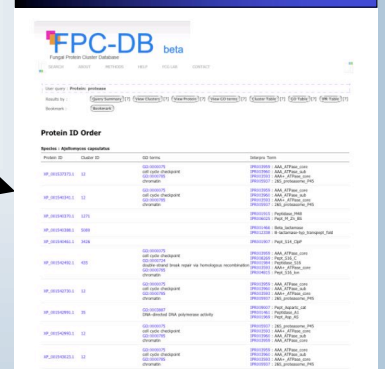
Protein Cluster View
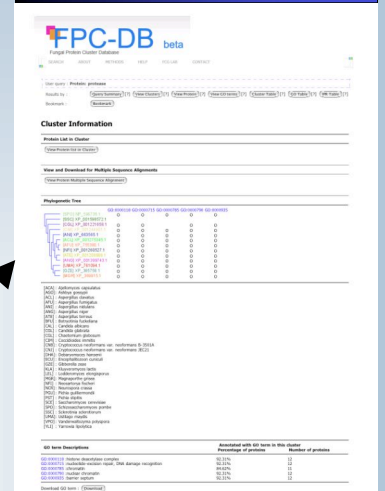
Cluster Table View
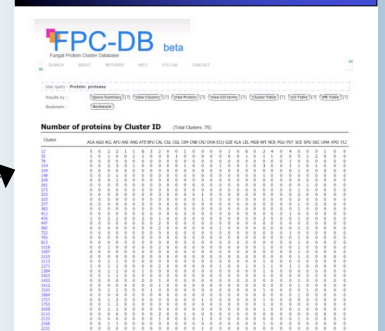
Protein List View

Protein Detail View