# FPC-DB: The Fungal Protein Cluster Database



Jaehee Jung<sup>1</sup>, Gangman Yi<sup>1</sup>, Serenella A. Sukno<sup>2</sup>, and Michael R. Thon <sup>2</sup> <sup>1</sup>Texas A&M University, Department of Computer Science, College Station, TX. 77845, USA. <sup>2</sup> Centro Hispano-Luso de Investigaciones Agrarias (CIALE), Department of Microbiology and Genetics, University of Salamanca, Salamanca 37007, Spain.



**Protein Cluster View** 

#### 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 volutionary 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 othologous pretens tend to have conserved function in different species, however numerous examples that demonstrate functional diversification among orthologous genese 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 (PFC-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 comparitive views. The proteins from fungal whole-genome sequencing projects are annotated with InterPro terms, SwissProt Expoveds, and Gene Ontology terms using an automated functional classification server previously developed in our group. The proteins are then clustere into putative gene families using the MCL application. Each protein cluster includes multiple sequence alignments, phylogenetic trees and a summary of functional actegories found in the cluster. Users can perform queries using proteins IDs, protein functional actegories, 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 actegories 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, senseling 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 neighborjoining 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.

|     |     | 11  | _   | 39  | Neurospora crassa       |
|-----|-----|-----|-----|-----|-------------------------|
|     | 3   |     |     |     | Magnaporthe grisea      |
|     |     |     | 57  |     | Fusarium graminearur    |
|     |     |     | 41  |     | Stagonospora podoru     |
|     | 1   |     |     |     | 87                      |
|     |     |     | 20  |     | Scierotinia scierotioru |
| 500 | 400 | 300 | 200 | 100 | Botrytis cinerea        |

 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  | or          | Family | InterPro Categories   |  |  |  |
|----------|-------------|--------|---|--|--|--|
|          | Contraction | ID     |   |  |  |  |
| 0.00E+00 | С           | 63     | Di-copper centre-containing; Tyrosinase   |  |  |  |
| 0.00E+00 | С           | 68     | Chitin-binding, type 1; Polysaccharide deacetylase; Cellulose-binding region  |  |  |  |
| 0.00E+00 | С           | 70     | Carbohydrate-binding WSC  |  |  |  |
| 0.00E+00 | с           | 20     | Heterokaryon incompatibility; Tetratrico peptide-like helical; Kinesin light chain; Disease<br>resistance protein; Protein prenyltransferase; NACHT nucleoside triphosphatase |  |  |  |
| 0.00E+00 | С           | 330    | UNKNOWN   |  |  |  |
| 0.00E+00 | С           | 328    | UNKNOWN   |  |  |  |
| 0.00E+00 | С           | 107    | Proteinase inhibitor, propeptide; Peptidase S8 and S53, subtilisin, kexin, sedolisin  |  |  |  |
| 1.98E-06 | С           | 2      | Major facilitator superfamily MFS 1   |  |  |  |
| 2.18E-05 | С           | 30     | Glycoside hydrolase, family 61; Cellalose-binding region, fungal  |  |  |  |
| 2.92E-04 | с           | 45     | General substrate transporter, Sagar transporter superfamily; Major facilitator<br>superfamily  |  |  |  |
| 3.36E-08 | С           | 3      | CFEM  |  |  |  |
| 5.67E-06 | С           | 24     | Heterokaryon incompatibility  |  |  |  |
| 8.94E-04 | С           | 142    | UNKNOWN   |  |  |  |
| 0.00E+00 | Е           | 115    | Integrase, catalytic region; Polynacleotidyl transferase, Ribonaclease H fold; Peptidase<br>sopartic, catalytic; Reverse transcriptase, RNA-dependent DNA polymerase          |  |  |  |
| 0.00E+00 | E           | 1646   | UNKNOWN   |  |  |  |
| 0.00E+00 | E           | 1360   | Ankyrin; Heterokaryon incompatibility   |  |  |  |
| 0.00E+00 | Е           | 428    | Gly coside hydrolase; Pecto lytic enzyme, Pectin lyase fold; Virulence factor, pectin lyase fold  |  |  |  |
| 1.78E-05 | E           | 927    | Phosphotransferase system, HPr serine phosphory lation site   |  |  |  |
|          | E           | 382    | UNKNOWN   |  |  |  |
| 2.98E-0/ |             | 1.77   |   |  |  |  |

|   | Zuci ji uge   |
|---|---|
| FPC-DB beta   | and the second se |
|   |   |
| 60 tem :<br>245 tem :   | ex : activity or 60.0000123 (188409)<br>ex : PGDV, as or PH000577   |
| Renky 1<br>Species 1  | er i Harrygrupp an einer sente, recur jaccia. 1<br>er i H23, 12 H30<br>er i Sacharmyon omvilae, 0402  |
| abancol seach<br>10 : Fender 10<br>14 : Fonzein 10<br>14 : Fonzein 10<br>14 : Fonzein 10  |   |
| Constitute and example<br>AMD 1 (Mic2) and Mic2P (Stretts, 1<br>OK 5 Cold 2 and Mic2P (Stretts, 1<br>MICT 1 OKID 2 and Mic2P, ADMIN: 1<br>MICT 1 OKID 2 and Mic2P, ADMIN: 1   |   |
| [num] = Species name ( DNV pan spoly to 22 and PR )<br>1 = Advisor georgia<br>2 = Advisor georgia<br>3 = Advisor georgia<br>4 = Aspecifika Christian<br>4 = Aspecifika Christian<br>5 = Aspecifika Christian  |   |
| Appropriate region     Analyzing the means     Analyzing the terminal     Betwarms Acclusion     Conditional     Conditional     Conditional     Conditional     Conditional     Conditional     Conditional  |   |
| <ul> <li>11 - Outpationers inmits an</li></ul>  |   |
| 19 = Lobienmynes ekonperan<br>20 = Magnaporto gries<br>21 = Rescutorya halteri<br>22 = Rescutorya casa<br>23 = Poto spatementi<br>24 = Poto spatementi<br>24 = Stata  |   |
| 25 = Social estimativa constant<br>26 = Social estimativa admitter<br>27 = Social estimativa<br>28 = Vandenza Burgera<br>29 = Vandenza Burgera<br>20 = Vantenza Repútica  |   |
| Bearch for proteins by Ir<br>functional domain), (<br>lescription.<br>Jse Boolean search opera  | nterPro term (protein family or<br>Gene Ontology Term, or<br>ators (AND, OR, NOT)   |
| imit searches to species.   | ·   |
|   |   |
|   |   |
|   | ₩   |
| Search Res  | ults Summary  |
|   |   |
| Funge Protein Coulour Datasance   |   |
| User (garry ) Postalini Protessee   |   |
| Source () (and a second ()) (and ()) (a                | anne (14 Canada ann (14) Canada (14) Canada (14) Canada (14)  |
| Tatal number of Clutters<br>Tatal number of Species   | Sant rout<br>15<br>21   |
| Tatal number of selected induits<br>Tatal number of GO terms accounted to proteins in protein. 3D order<br>Tatal number of GO terms by query<br>Tatal number of DRL terms executed to proteins in protein. 3D order<br>Tatal number of DRL terms of accounted to proteins in protein. 3D order<br>Tatal number of DRL terms of accounted to protein | 209<br>14<br>14<br>99   |
| e search results summ<br>erview of your search re<br>proteins, clusters, GO te<br>ults.   | ary page gives you a quick<br>sults. Quickly find the number<br>rms, etc. that are in your search   |
|   | Ļ   |
| ch Results Are  | e Organized In Views  |
| ew presents the same sea  | rch results, but organized in differe   |
|   |   |
|   |   |
|   |   |
|   |   |
| Futu  | re Plans  |
| SwissProt keyword anno  | otations to proteins  |
| each cluster for evidence   | e of positive selection   |
| tifu orthologg/porologg i   | n each cluster  |

### References

v

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.

|    |   | C-DE  | B beta   |   |  |   |
|----|---|---|--|---|--|---|
|    | in Der geny Protein pr<br>Reuts ty : Sootmark :   | ninaan<br>Jarey Yammery) (1) (year c<br>Andarant)   | mmelj(s) ( <del>Ann a</del> sserj(s  | ( <del>1988-00 1000)</del> (1/1) (                              | (sanatani) () (satani) () (setan   | (Bra                                    |
|    | Cluster Inform<br>Protain List in Outler<br>(year/Instancial in Outle   | ation   |  |   |  |   |
|    | View and Exempland for<br>(Your Poten Multiple Sou<br>Phylogenetic Tree   | Hultiple Sequence Algories  | 1715 GD-0000795 GD-0000796<br>© 0  | 60 eccentas<br>0  |  |   |
|    |   | 10211<br>1000 1<br>1000 1<br>1000 0<br>1100 0<br>1100 0<br>1100 0<br>1100 0<br>1000 0 | 000000000000000000000000000000000000000  | 000000000000000000000000000000000000000                         |  |   |
| 1  | EXCLUSION APPLIED<br>(2014) Application operating<br>(2014) Address generating<br>(2014) Applied classific<br>(2014) Applied class  |   | 0 0  | 0   |  | - 2                                     |
|    | 124. Candidate document<br>124. Candidate document<br>125. Candidate document<br>126. Candidate and<br>127. Candidate and<br>127. Candidate and<br>127. Candidate and<br>127. Candidate and<br>127. Statistica and<br>12  | er<br>Innans var. nenformans (5-35)<br>speci<br>recall<br>N   | ESA  |   |  |   |
|    | 10.11.1 Codemonitors from<br>(MSL) Respective plane<br>(MSL) Researchers Intel<br>(MSL) Researchers Intel<br>(MSL) Restructions Intel<br>(MSL) Restructions (Intel<br>(MSL) Sectionshorm(Intel<br>(MSL) Sectionshorm(Intel<br>(MSL) Sectionshorm(Intel<br>(MSL) Sections advects<br>(INT) Sections advects<br>(INT) Sections advects  | gaporan<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I<br>I  |  |   |  |   |
|    | (VC), Vandemadouyna (     (VL) - Yarowa ipojeca     (VL) - Yarowa ipojeca     (O) term Descriptions     (O) 900113 hadroid rodd   | vingens<br>viese complex<br>soon report, DNA damage rec   | ugutun   | Annotated with<br>Percentage of p<br>92,31%<br>92,31%<br>96,32% | 60 term in this chaster<br>Hamber of proteins<br>12<br>13<br>13  |   |
|    | GD 200755 middae diven<br>GD 200755 barrier aptur<br>Dowrkael GD tem : (Dow   | inat)   |  | 92.31%<br>92.31%  | 12   | -                                       |
|    |   | Clust   | ter Ta   | ble V   | iew  |   |
|    |   | C-DE  | B beta   |   |  |   |
|    | BOACH ABOY<br>Der gary Pestelan an<br>Routs by :<br>Boolewsk :  | HETHCOS HELP<br>Hease<br>Jerry Summary] (1) (Sear C<br>Sector 24)   | 105.08 0000  | (7<br>) ( <del>764-00 1000)</del> (7) (                         | (same tang) (r) (ga tang) (r) (ga tan  | E)ru                                    |
|    | Number of pro   |   | r ID         (fotal Clusters           BYU CAL CEL CEL CEL CEL CEL CEL         CEL CEL CEL           3         2         0         1         0         0   | 25)<br>0 DHA EDJ GEE HLA LEL<br>0 0 1 0 6<br>0 0 0 1 1 6        | NGR NYT NCR PGLI PST SCT SPO SSC LMA VP<br>0 2 4 0 4 0 0 0 1 1<br>0 1 1 0 0 5 1 2 0 0  | 0<br>0                                  |
|    | 34<br>139<br>124 0<br>209 0<br>201 0<br>2000 0<br>201 0<br>2010 |   |  |   | 0         0         0         1         0  |   |
| // | 412 0<br>415 1<br>447 0<br>702 0<br>702 0<br>700 0<br>813 0<br>814 0<br>815 0     |   |  |   |  | 000000000000000000000000000000000000000 |
|    | 1113 0<br>1273 1<br>1394 0<br>1403 0<br>1403 0<br>1500 0  |   |  |   | 0         1         0  | 000000000000000000000000000000000000000 |
|    | 2122<br>2222<br>2346<br>2355<br>2355<br>2355  |   |  |   |  |   |
|    | _   | Prot  | tein L   | ist Vi  | iew  |   |
|    |   | C-DE  | beta   | a j   |  | 1                                       |
| ١  | Der cory / Polain pr<br>Rouits ty : 5<br>Bostmark : 5   | denes<br>Gery Summery (1) (year (<br>Bookerank)   | honey](7) ( <del>yes hono</del> y)(7   | ( <del>yias (0 1011)</del> )(7) (                               | (kane Tanic) [7] (30 Tahic) [7] (30 Tan  | £)ru                                    |
|    | Protein ID Ord  | er<br>mittes<br>10 00   | 1 Series<br>0.000079   |   | Margen Sons<br>Philippes - AAA, ATPasa, John<br>Philippes - AAA, ATPasa, John  |   |
|    | 99,0015403411 12  | 000000000000000000000000000000000000000   | instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instantin<br>Instan |   | Peor 201 AMA Alfree, see<br>Peor 201 St. proteams, P6<br>Peor 201 AMA Alfree, see<br>Peor 201 AMA Alfree, see<br>Peor 201 AMA Alfree, see<br>Peor 201 AMA Alfree, see  |   |
|    | MP_001540370.1 1271<br>MP_001540300.1 5000  |   |  |   | PROTOCO : Populane, MAR<br>PROTOCO : Populaci, MAR<br>PROTOCO : Deputy (C.2), BR<br>PROTOCO : Braziliana Statement<br>PROTOCO : Braziliana Statement, Sampage, Sal   |   |
|    | 10-00154040.1 MIN   | 33  | 1:0000075<br>1:cpte checkpare<br>1:0000724<br>Dire-chand brook repair via he   | ndopa roontinging   | PROTOCY : Page S14, CpP<br>PROTOCY : ANA, ATPase, core<br>PROTOCY : Page, S16, C<br>PROTOCY : Page, S16, C<br>PROTOCY : Page, S16<br>PROTOCY : PROTOCY : PROT |   |
|    | 10,0034070.1 D  | 888   | ciacons<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrain<br>constrai |   | PROVIDE AMAY ATTAIN JAN<br>PROVIDE PROJECT AMA ATTAIN OF<br>PROJECT AMA ATTAIN OF<br>PROJECT AMA ATTAIN OF<br>PROJECT AMAY ATTAIN JAN  |   |
|    | n'anteant u   | 00<br>05  | orazin<br>Hotobel<br>A-dested DNA polymerase ad  | tuny .  | PROTECT : 145, proteasane, 745<br>PROTECT : Pap, Aquati, cit<br>PROTECT : Pap, Aquati, cit<br>PROTECT : Pap, Aquati, AS  |   |
|    | 99_00350000.1 12  | 00<br>00<br>00<br>00<br>00<br>00<br>00<br>00<br>00<br>00<br>00<br>00<br>00  | n code dhedquard<br>Indoctives<br>simular<br>Indoctives<br>Indoctives<br>Indoctives<br>Indoctives<br>Indoctives<br>Indoctives<br>Indoctives<br>Indoctives<br>Indoctives  |   | 100, JP0000000, JP00<br>P001250 1: AAA, ATPanc, John<br>P001250 1: AAA, ATPanc, John   |   |
|    |   | Prote   | ein De   | tail V  | /iew   |   |
|    | Protein information   |   |  |   |  |   |
|    | Species<br>Protein ID<br>LOCUS ID   | Ajellentyces ca<br>XP 001537373<br>HCAG 07645   | pozlatas<br>1  |   |  | _                                       |
|    | Cluster ID<br>GO term   | 12<br>GO:000075<br>GO:000075  | cell cycle checkpoint<br>chromatia   |   |  |   |
|    | Interpro term   | 1PR003959 : AJ<br>1PR003960 : AJ<br>1PR003593 : AJ<br>1PR005937 : 26  | AA_ATPase_core<br>AA_ATPase_sub<br>AA+_ATPase_core<br>S_protessome_P45   |   |  |   |