THE FAT-CAT WEB SERVER

QUICK-START GUIDE

Last revised 2/10/2013

Helpful Tip

Look for this icon to find helpful hints on using FAT-CAT and interpreting results.

1.7 The FAT-CAT Pipeline

The FAT-CAT pipeline starts with the submission of a protein sequence and parameter selection and proceeds through family and subtree HMM scoring to ortholog identification and functional annotation. The FAST-CAT variant differs from the default FAT-CAT pipeline in stage 3 (indicated by red arrows).

In stage 1, the query is scored against family HMMs in the PhyloFacts database for proteins sharing the same multi-domain architecture (MDA, shown at top) and HMMs constructed for Pfam domains (shown at bottom). Families meeting stage 1 criteria are passed to stage 2. In this toy example, PhyloFacts trees for two Pfam domains and a tree for the multi-domain architecture meet stage 1 criteria and are passed to stage 2.

In stage 2, we score all the HMMs in the tree. The subtree node corresponding to the top-scoring HMM is examined to determine its suitability as a source of orthologs to the query. For each top-scoring node that meets these criteria we identify a (typically larger) enclosing clade supported by one or more orthology methods. Enclosing clades are passed to Stage 3 for ortholog identification.

In stage 3, sequences in the enclosing clades are retrieved and analyzed for agreement with pre-defined minimum sequence identity and overlap to the query. FAT-CAT (blue arrows) evaluates the pairwise alignment between the query and each sequence and identifies all supporting evidence supporting the orthology. FAST-CAT (red arrows) uses a fast k-tuple comparison to select the most similar sequences from the enclosing clade, constructs an MSA including the query, estimates a phylogenetic tree, and extracts a subtree of the phylogenetically closest sequences. Alignment analysis is then restricted to this smaller subset. Sequences meeting these criteria are passed to stage 4.

In stage 4, we derive a weighted consensus functional annotation for the query based on orthologs selected in stage 3. Annotations from close orthologs are given higher weight than those from more distant orthologs, and manually curated annotations are given higher weight than those that are derived computationally.

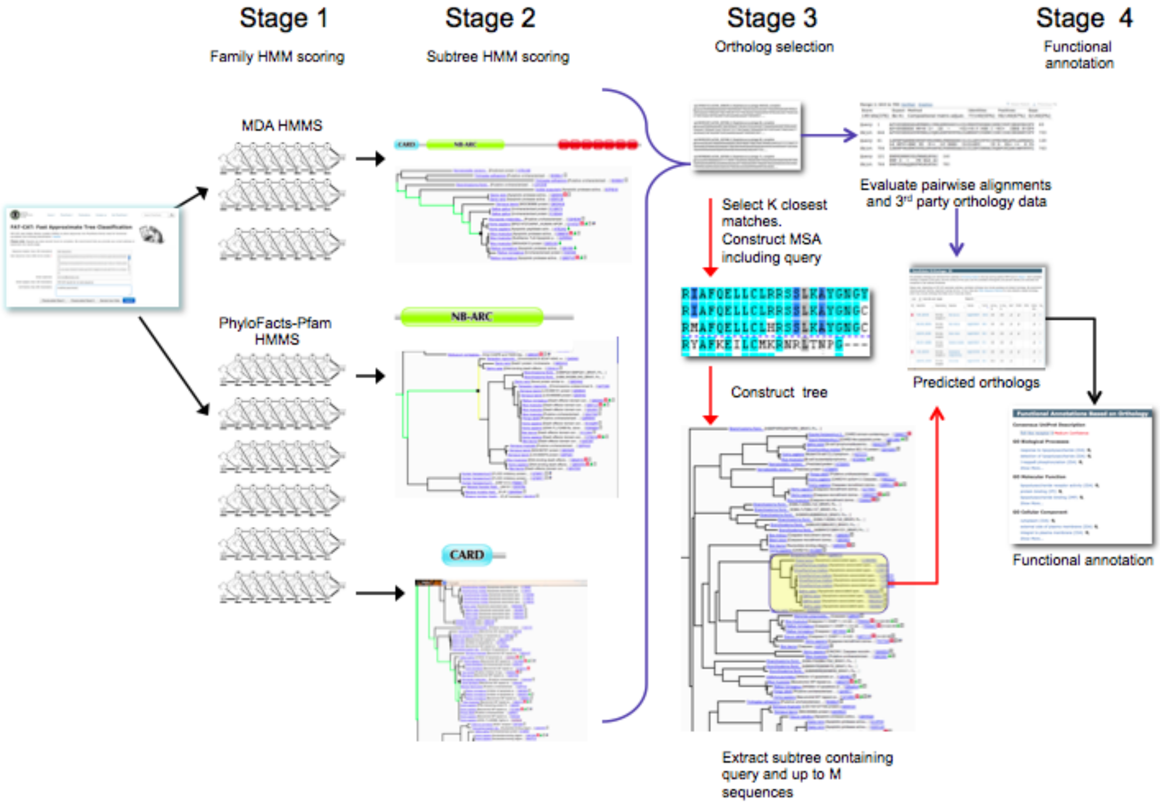
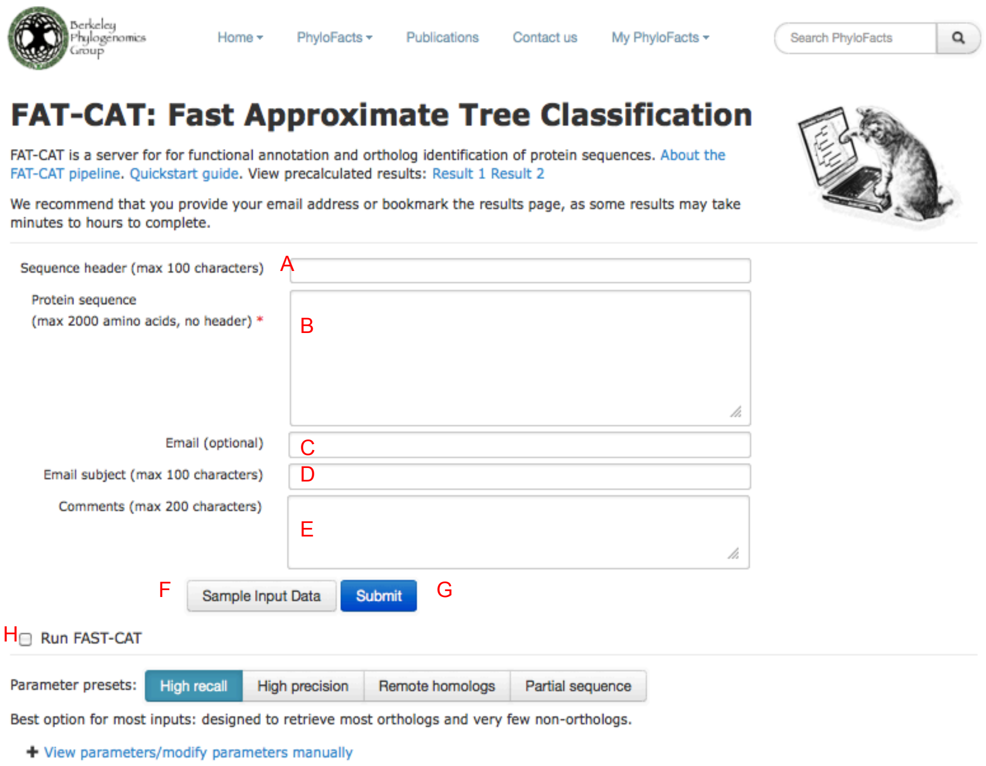


Figure 1. FAT-CAT pipeline. The FAST-CAT variant is shown in red.

2. SUBMITTING A SEQUENCE TO THE FAT-CAT WEBSERVER



FAT-CAT: Fast Approximate Tree Classification

FAT-CAT is a server for for functional annotation and ortholog identification of protein sequences. About the FAT-CAT pipeline. Quickstart guide. View precalculated results: Result 1 Result 2

We recommend that you provide your email address or bookmark the results page, as some results may take minutes to hours to complete.

Sequence header (max 100 characters)

Protein sequence (max 2000 amino acids, no header) *

Email (optional)

Email subject (max 100 characters)

Comments (max 200 characters)

□ Run FAST-CAT

Parameter presets: High recall | High precision | Remote homologs | Partial sequence

Best option for most inputs: designed to retrieve most orthologs and very few non-orthologs.

✚ View parameters/modify parameters manually

Figure 2. FAT-CAT input form at http://phylogenomics.berkeley.edu/phylofacts/fatcat/.

2.1 The input form (figure 1).

A: Header line (optional). Recommended so that you can remember what sequence you submitted.

B: Protein sequence (required). Input your sequence in raw format – no header line, just the amino acids .

C: Email address (optional). If you'd like results to be sent to you by email, input your email address. Otherwise, bookmark the results page displayed after you click Submit.

D: Email subject (optional): If you enter your email address, you may optionally fill in the "Email subject" section, e.g., to include the sequence identifier; the default subject line will read "PhyloFacts FAT-CAT Job <number> has completed".

E: Comments (optional). Comments will be stored in the Job Summary section of the Results.

F: Sample Input Data. Clicking on this button populates the form so that you can see what kind of input is expected in each section.

G: Submit. Click on this when you're ready to launch the job. It will bring you to a progress page which you can bookmark and where you can track the progress of your job.

H: FAST-CAT. FAST-CAT is a beta version of FAT-CAT (in other words, we haven't debugged it completely) that is designed for speed.

I: Parameter presets. We provide four different combinations of parameters designed for different types of input sequences. If you click on each button you'll see a brief description of the types of inputs these settings are designed to handle. See the Decision Tree section of this guide for additional information.

J: View/Modify parameters. Click on this link to view and edit any individual parameters.

2.2 What types of sequences does FAT-CAT accept?

FAT-CAT expects raw amino acid sequence – no header line – in the input box. You can put the header line into the corresponding section of the input form (optional). Nucleotide data are NOT accepted. The maximum sequence length is 2000 amino acids.

Most bioinformatics web servers allow either FASTA input (in which the first line is a “header line” starting with a “>” symbol) or raw sequence, which omits the header line. These are shown below.

```
Example FASTA input:

>sp|Q9XT58|ADRB3_SHEEP Beta-3 adrenergic receptor OS=Ovis aries GN=ADRB3 PE=3 SV=2
MAPWPPGNSSLTPWPDIPTLAPNTANASGLPGVPWAVALAGALLALAVLATVGGNLLVIV
AIARTPRLQTMTNVFVTSLATADLVVGLLVVPPGATLALTGHWPLGVTGCELWTSVDVLC
VTASIETLCALAVDRYLAVTNPLRYGALVTKRRARAAVVLVWVVSAAVSFAPIMSKWWRV
GADAEAQRCHSNPRCCTFASNMPYALLSSSVSFYLPLLVMLFVYARVFVVATRQLRLLRR
ELGRFPPEESPPAPSRSGSPGPAGPYASPAGVPSYGRRPARLLPLREHRALRTLGLIMGT
FTLCWLPFFVVNVVRALGGPSLVSGPTFLALNWLGYANSAFNPLIYCRSPDFRSAFRRLL
CRCPPEEHLAAASPPRAPSGAPTVLTSPAGPRQPSPLDGASCGLS

The raw sequence for this entry is:

MAPWPPGNSSLTPWPDIPTLAPNTANASGLPGVPWAVALAGALLALAVLATVGGNLLVIV
AIARTPRLQTMTNVFVTSLATADLVVGLLVVPPGATLALTGHWPLGVTGCELWTSVDVLC
VTASIETLCALAVDRYLAVTNPLRYGALVTKRRARAAVVLVWVVSAAVSFAPIMSKWWRV
GADAEAQRCHSNPRCCTFASNMPYALLSSSVSFYLPLLVMLFVYARVFVVATRQLRLLRR
ELGRFPPEESPPAPSRSGSPGPAGPYASPAGVPSYGRRPARLLPLREHRALRTLGLIMGT
FTLCWLPFFVVNVVRALGGPSLVSGPTFLALNWLGYANSAFNPLIYCRSPDFRSAFRRLL
CRCPPEEHLAAASPPRAPSGAPTVLTSPAGPRQPSPLDGASCGLS

The header line is: >sp|Q9XT58|ADRB3_SHEEP Beta-3 adrenergic receptor OS=Ovis aries
GN=ADRB3 PE=3
```

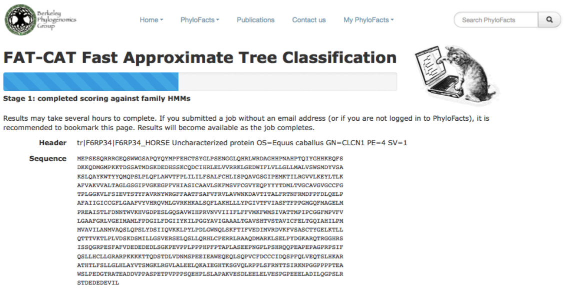
Figure 3. FASTA and raw sequence formats. FAST-CAT requires raw sequence input (amino acid only).

2.3 Submitting your email address or bookmarking the results page.

We encourage you to submit your email address, as many FAT-CAT jobs can take an hour or longer to return. If you don’t want to give your email address, then bookmark the page that displays after you click Submit. Or copy the URL and save it for your records.

2.4 Tracking the progress of your job.

After you click Submit, you will be brought to a page where you can track the progress of your submitted job. You will be reminded to bookmark that page if you haven’t submitted your email address.



FAT-CAT Fast Approximate Tree Classification

Stage 1: completed scoring against family HMMs

Results may take several hours to complete. If you submitted a job without an email address (or if you are not logged in to PhyloFacts), it is recommended to bookmark this page. Results will become available as the job completes.

Header tr|F6RP34|F6RP34_HORSE Uncharacterized protein OS=Equus caballus GN=CLCN1 PE=4 SV=1

Sequence

```
MEPSESQRRRGEQSWWGSAPQYQYMPFEHCTSYGLPSENGGLQHRLWRDAGHHPNAHPTQIYGHHKEQFS
DKKQDMGMPKKTDSSATMDSKDEDHSSKCQDCIHRLELVVRRKLGEDWIFLVLLGLLMALVSWSMDYVSA
KSLQAYKWTYYQMQPSLPLQFLAWVTFPLILILFSALFCHLISPQAVGSGIPEMKTILRGVVLKEYLTLK
AFVAKVVALTAGLGSGIPVGKEGPFVHIASICAAVLSKFMSVFCGVYEQPYYYTDMLTVGCAVGVGCCFG
TPLGGKVLFSIEVTSTYFAVRNYWRGFFAATFSAFVFRVLAVWNKDAVTITALFRTNFRMDFPFDLQELP
AFAIIGICCGFLGAAFVYVHRQVMLGVRKHKALSQFLAKHLLLYPGIVTFVIASFTFPPGMGQFMAGELM
PREAISTLFDNNTWVKHVGDPESLGQSAVWIHPRVNVVIIIFLFFVMKFWMSIVATTMPIPCGGFMPVFV
LGAAFGRLVGEIMAMLFPDGILFDGIIYKILPGGYAVIGAAALTGAVSHTVSTAVICFELTGQIAHILPM
MVAVILANMVAQSLQPSLYDSIIQVKKLPYLPDLGWNQLSKFTIFVEDIMVRDVKFVSASCTYGELKTLL
QTTTVKTLPLVDSKDSMILLGSVERSELQSLLQRHLCPERRLRAAQDMARKLSELPYDGKARQTRGGHRS
ISSQGRPESFAFVDEDEDEDLSGKPEVPPLPPPHPFPTAPLASEEPNGPLPSHRQQPEAPEPAGPRPSIF
QSLLHCLLGRARPKKKKTTQDSTDLVDNMSPEEIEAWEQEQLSQPVCFDCCCIDQSPFQLVEQTSLHKAR
ATHTLFSLLGLHLAYVTSMGKLRGVLALEELQKAIEGHTKSGVQLRPPLSFRNTTSIRKNPGGPPPPTEA
WSLPEDGTRATEADDVPPASPETPVPPPSQEHPLSLAPAKVESDLEELELVESPGPEEELADILQGPSLR
STDEDEDEVIL
```

3. Selecting and modifying FAT-CAT pipeline parameters

Pipeline parameters for stages 1 through 3 are designed to accommodate different types of input sequences.

3.1 A decision tree to guide parameter selection.

The decision tree shown in Figure 4 is included to help you choose among pre-set parameter settings.

High Recall settings – the FAT-CAT default – are designed to handle sequences that are full-length, contain no promiscuous domains and have no close paralogs (e.g., with high sequence identities).

High Precision settings are designed for cases where a sequence is known to contain a promiscuous domain or that is in a protein superfamily with close paralogs (e.g., with >70% identity).

Partial Sequence settings are designed to handle sequences missing a region that its close homologs (and presumed orthologs) contain. Some of these may represent splice variants or gene model errors.

By contrast, the Remote Homolog Detection settings handle cases where few or no homologs, or only matches that align over local regions, can be found by BLAST. In this last case, however, we remind users that results will represent distant homologs that may not have the same function and may not be actual orthologs.

3.2 What if you don't know anything about your query?

The decision tree shown in Figure 4 assumes you know something about your query. But what if you have no idea how to answer those questions? You have two options.

Option 1: Go ahead and submit your sequence to FAT-CAT and see what happens. If too many results are returned or if you see paralogs in the results, resubmit with the High Precision settings. If too few results are returned, resubmit with Remote Homolog detection settings, or modify individual program parameters to relax constraints.

Option 2: Use BLAST and Pfam to do a little bioinformatics analysis to get some information about your query.

Is your sequence complete? If your sequence is annotated as a partial sequence or fragment, it's not complete. But many sequences that are annotated as complete are actually incomplete. To figure out which is which, try running BLAST and see what kinds of results display. For instance, if BLAST matches include many sequences that are roughly the same length as your query and align well along their entire length, your sequence is probably complete.

However, if BLAST detects close homologs with high sequence identity (e.g., >70% identity) but your sequence is shorter than these close homologs, your sequence may be partial due either to a gene model error, alternative splicing or some other cause.

Does your sequence contain a promiscuous domain? Try submitting your query to Pfam (e.g., at http://pfam.sanger.ac.uk/) to see what Pfam domains are present. For each Pfam domain found, check the Domain Arrangements tab – if any Pfam domains in your query are found in many different domain arrangements, your query has a promiscuous domain.

Does your sequence have close paralogs? If BLAST analysis lists many proteins with high sequence identity (e.g., >70%) and obviously different functions, your query may be in a superfamily with recent duplication events. Examples of this type include different ion channels, G-protein coupled receptors, toll-like receptors, globins, and defensins.

Tip: You can submit your sequence to FAT-CAT using the default settings, set for high recall. If too many results are returned or paralogs are included in the candidate orthologs, resubmit using the High Precision settings. If too few results are returned or you see orthologs are included in the "Other Sequence Matches" tab, modify parameters manually to relax sequence identity or alignment overlap requirements. You can also try the Remote Homolog Detection settings.

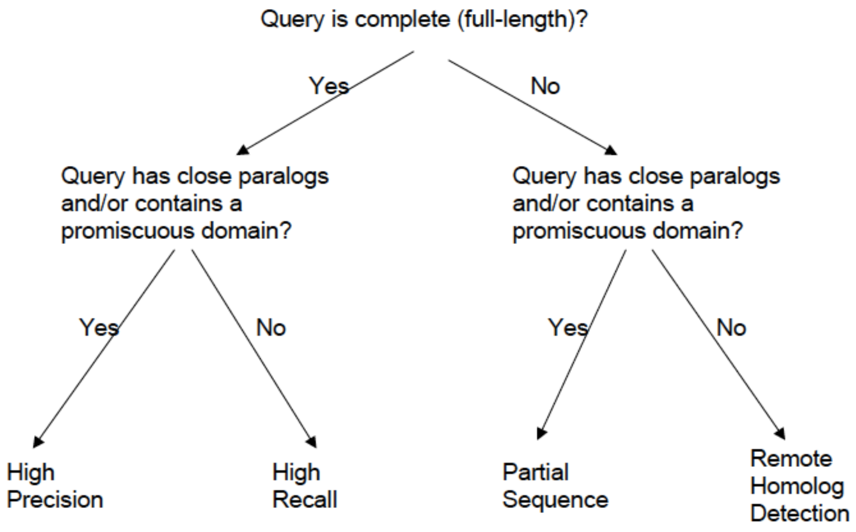

Query is complete (full-length)?
Yes
No
Query has close paralogs and/or contains a promiscuous domain?
Query has close paralogs and/or contains a promiscuous domain?
Yes
No
Yes
No
High Precision
High Recall
Partial Sequence
Remote Homolog Detection

Figure 4. Decision Tree for selection from the four parameter setting presets.

High Recall settings are designed to handle sequences that are full-length, contain no promiscuous domains and have no close paralogs.

High Precision settings are designed to handle sequences containing a promiscuous domain (e.g., a kinase domain, leucine-rich repeat, or other commonly found structural domain) or having closely related paralogs.

The Partial Sequence settings are designed to handle sequences that have a gene model error or that are splice variants but where closely related homologs can be detected by BLAST.

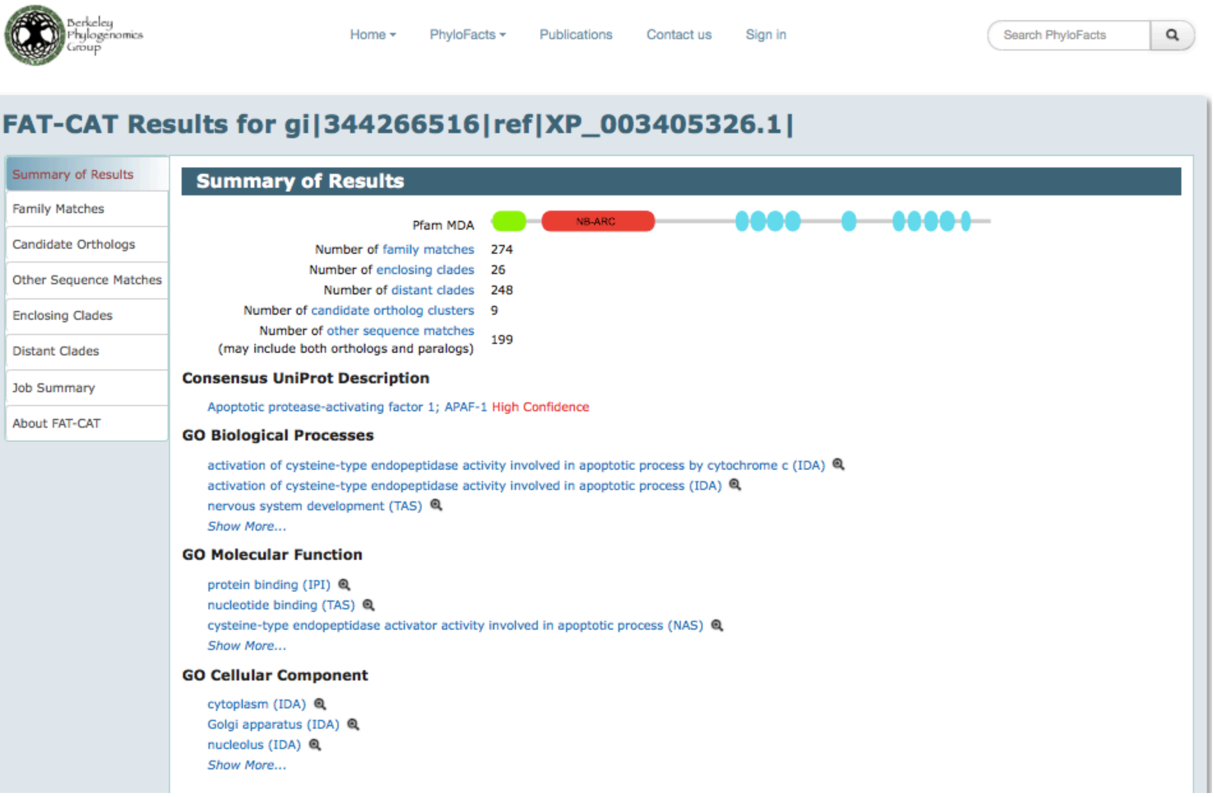
The Remote Homolog Detection settings handle cases where few or no homologs can be found by BLAST. In this last case, however, we remind users that results will represent distant homologs that may not have the same function and may not be actual orthologs.

4. Interpreting FAT-CAT results: A case study.

The figure below presents FAT-CAT results for gi|344266516|ref|XP_003405326.1, a predicted apoptotic protease-activating factor 1 isoform 1 from *Loxodonta africana* (African elphant). This result corresponds to PhyloFacts job 2570, and is available at http://makana.berkeley.edu/phylofacts/fatcat/2570/.

In this example, we used the FAT-CAT program defaults (as of 2/8/13), designed for high recall.

4.1 Summary of results



FAT-CAT Results for gi|344266516|ref|XP_003405326.1|

Summary of Results

Number of family matches	274
Number of enclosing clades	26
Number of distant clades	248
Number of candidate ortholog clusters	9
Number of other sequence matches (may include both orthologs and paralogs)	199

Consensus UniProt Description

Apoptotic protease-activating factor 1; APAF-1 High Confidence

GO Biological Processes

activation of cysteine-type endopeptidase activity involved in apoptotic process by cytochrome c (IDA)
activation of cysteine-type endopeptidase activity involved in apoptotic process (IDA)
nervous system development (TAS)
Show More...

GO Molecular Function

protein binding (IPI)
nucleotide binding (TAS)
cysteine-type endopeptidase activator activity involved in apoptotic process (NAS)
Show More...

GO Cellular Component

cytoplasm (IDA)
Golgi apparatus (IDA)
nucleolus (IDA)
Show More...

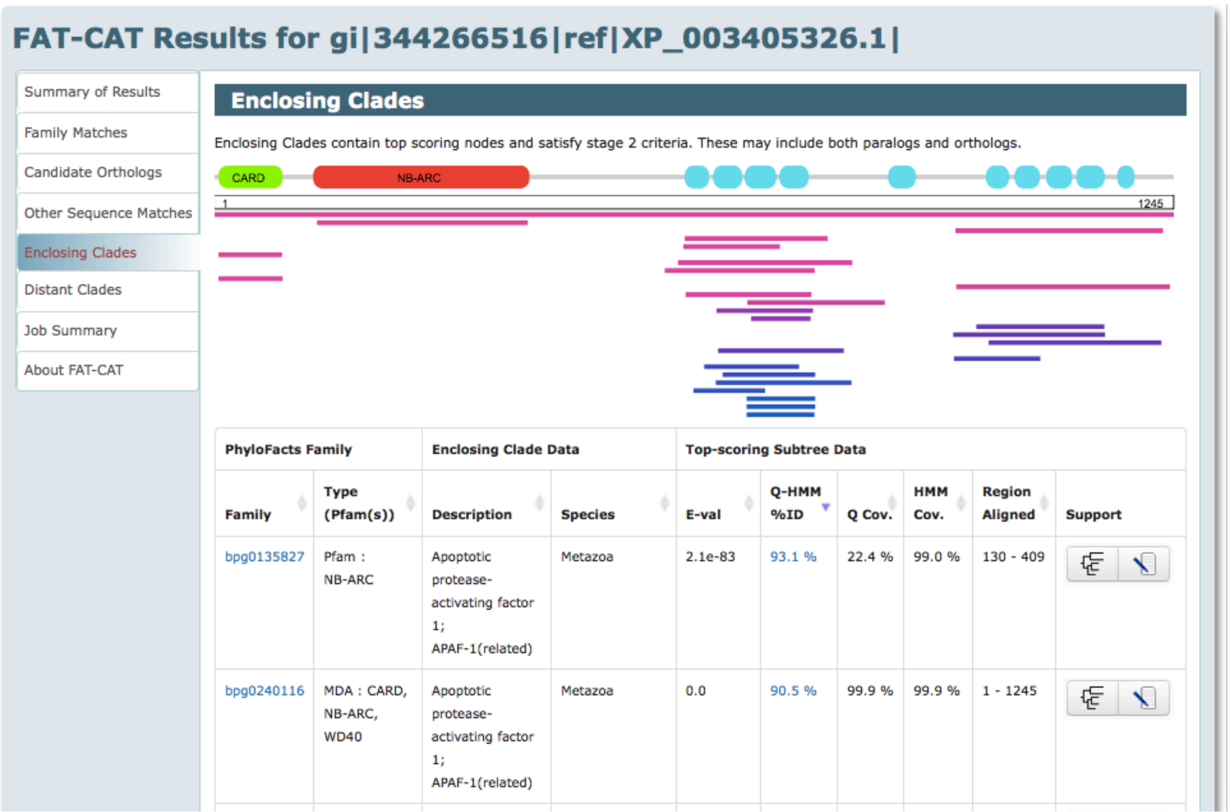
Figure 5. Summary of Results: overview of results, including the Pfam multi-domain architecture for the query produced by scanning Pfam-A HMMs.

The FAT-CAT pipeline identified 274 families matching Stage 1 criteria, orthologs from nine different genomes (candidate ortholog clusters), and 199 additional homologs that failed one or more criteria for orthology.

Predicted functional annotations for the query derived from orthologs satisfying stage 3 criteria are displayed.

The Job Summary tab displays the input sequence, all pipeline parameters and any comments in the Input Form.

4.2 Enclosing clades



FAT-CAT Results for gi|344266516|ref|XP_003405326.1|

Enclosing Clades

Enclosing Clades contain top scoring nodes and satisfy stage 2 criteria. These may include both paralogs and orthologs.

PhyloFacts Family		Enclosing Clade Data		Top-scoring Subtree Data					
Family	Type (Pfam(s))	Description	Species	E-val	Q-HMM %ID	Q Cov.	HMM Cov.	Region Aligned	Support
bpg0135827	Pfam : NB-ARC	Apoptotic protease-activating factor 1; APAF-1(related)	Metazoa	2.1e-83	93.1 %	22.4 %	99.0 %	130 - 409	
bpg0240116	MDA : CARD, NB-ARC, WD40	Apoptotic protease-activating factor 1; APAF-1(related)	Metazoa	0.0	90.5 %	99.9 %	99.9 %	1 - 1245	

Figure 6. Enclosing clades: Enclosing clade data passing stage 2 criteria, displaying the top two matches sorted by the sequence identity between the query and the top-scoring subtree HMM.

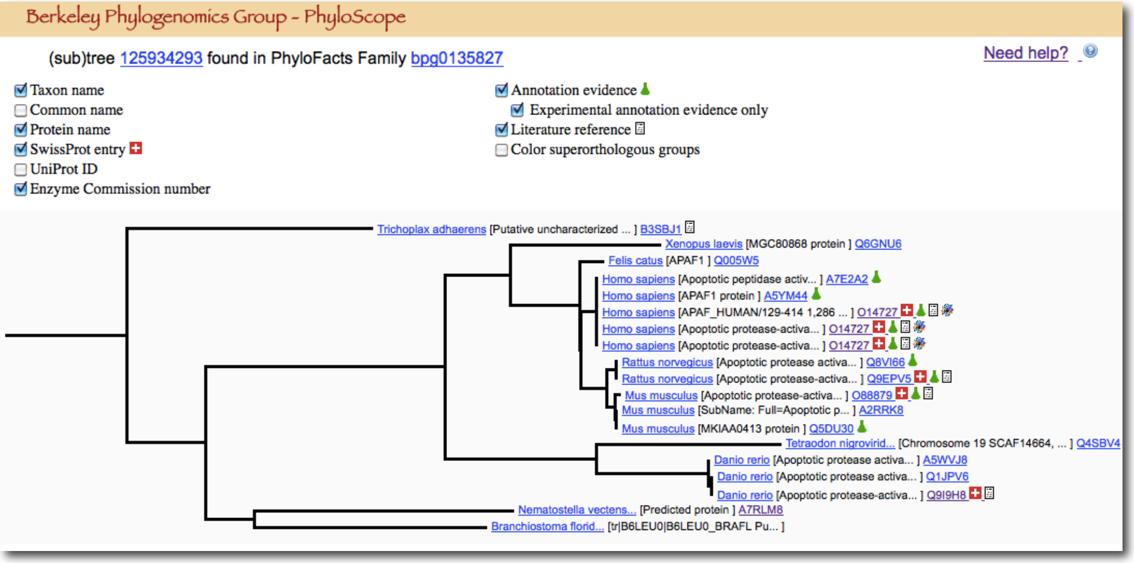
The top-scoring HMM matches the PhyloFacts-Pfam NB-ARC domain HMM.

The second top-ranked HMM matches along the entire multi-domain architecture (MDA).

Clicking on the Q-HMM %ID will display the alignment between the query and HMM.

Viewing the tree for an enclosing clade. Clicking on the tree icon at the far right of the enclosing clade data table launches the PhyloScope viewer, displaying the enclosing clade and highlighting a path to the top-scoring node (based on stage 2 HMM scoring). In many cases this phylogenetic placement provides an approximate taxonomic classification as well as a functional classification, as shown in this case study.

Tree icons on FAT-CAT pages are linked to the PhyloScope tree viewer. Click on these icons to view trees for enclosing clades.



Berkeley Phylogenomics Group - PhyloScope

(sub)tree 125934293 found in PhyloFacts Family bpg0135827

Need help?

☑ Taxon name
☐ Common name
☑ Protein name
☑ SwissProt entry
☐ UniProt ID
☑ Enzyme Commission number

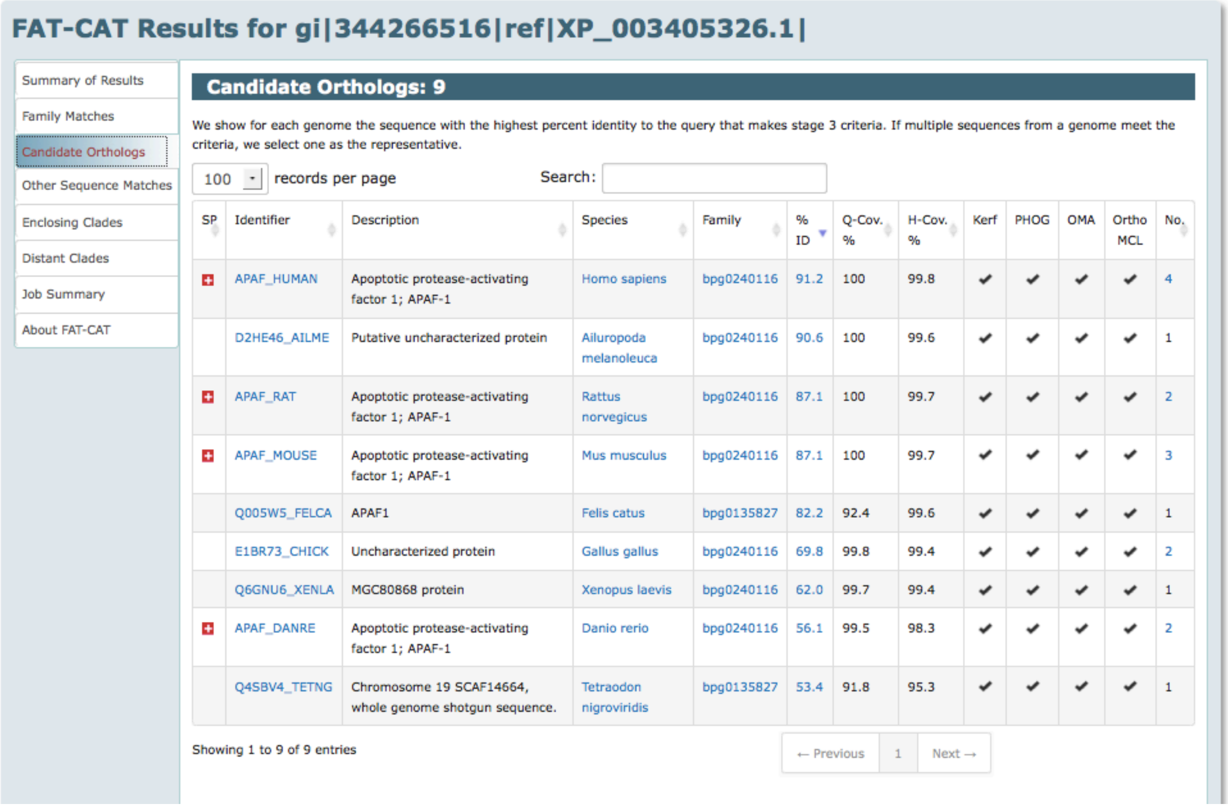
☑ Annotation evidence
☑ Experimental annotation evidence only
☑ Literature reference
☐ Color superorthologous groups

Trichoplax adhaerens [Putative uncharacterized ...] B3SBJ1
Xenopus laevis [MGC80868 protein] Q6GNU6
Felis catus [APAF1] Q005W5
Homo sapiens [Apoptotic peptidase activ...] A7E2A2
Homo sapiens [APAF1 protein] A5YM44
Homo sapiens [APAF_HUMAN/129-414 1,286 ...] O14727
Homo sapiens [Apoptotic protease-activa...] O14727
Homo sapiens [Apoptotic protease-activa...] O14727
Rattus norvegicus [Apoptotic protease activa...] Q8VI66
Rattus norvegicus [Apoptotic protease-activa...] Q9EPV5
Mus musculus [Apoptotic protease-activa...] O88879
Mus musculus [SubName: Full=Apoptotic p...] A2RRK8
Mus musculus [MKIAA0413 protein] Q5DU30
Tetraodon nigrovirid... [Chromosome 19 SCAF14664, ...] Q4SBV4
Danio rerio [Apoptotic protease activa...] A5WVJ8
Danio rerio [Apoptotic protease activa...] Q1JPV6
Danio rerio [Apoptotic protease-activa...] Q9I9H8
Nematostella vectens... [Predicted protein] A7RLM8
Branchiostoma florid... [tr|B6LEU0|B6LEU0_BRAFL Pu...]

Figure 7. PhyloScope viewer

Trees displayed in PhyloScope are decorated with icons indicating experimental support for GO annotations, PDB structures and biological literature; icons are linked to the databases providing these data. A link to the PhyloScope online help is provided in the upper right corner of the PhyloScope viewer.

4.4 Candidate orthologs



Candidate Orthologs: 9

We show for each genome the sequence with the highest percent identity to the query that makes stage 3 criteria. If multiple sequences from a genome meet the criteria, we select one as the representative.

100 records per page   Search:

SP	Identifier	Description	Species	Family	% ID	Q-Cov. %	H-Cov. %	Kerf	PHOG	OMA	Ortho MCL	No.
✚	APAF_HUMAN	Apoptotic protease-activating factor 1; APAF-1	Homo sapiens	bpg0240116	91.2	100	99.8	✓	✓	✓	✓	4
	D2HE46_AILME	Putative uncharacterized protein	Ailuropoda melanoleuca	bpg0240116	90.6	100	99.6	✓	✓	✓	✓	1
✚	APAF_RAT	Apoptotic protease-activating factor 1; APAF-1	Rattus norvegicus	bpg0240116	87.1	100	99.7	✓	✓	✓	✓	2
✚	APAF_MOUSE	Apoptotic protease-activating factor 1; APAF-1	Mus musculus	bpg0240116	87.1	100	99.7	✓	✓	✓	✓	3
	Q005W5_FELCA	APAF1	Felis catus	bpg0135827	82.2	92.4	99.6	✓	✓	✓	✓	1
	E1BR73_CHICK	Uncharacterized protein	Gallus gallus	bpg0240116	69.8	99.8	99.4	✓	✓	✓	✓	2
	Q6GNU6_XENLA	MGC80868 protein	Xenopus laevis	bpg0240116	62.0	99.7	99.4	✓	✓	✓	✓	1
✚	APAF_DANRE	Apoptotic protease-activating factor 1; APAF-1	Danio rerio	bpg0240116	56.1	99.5	98.3	✓	✓	✓	✓	2
	Q4SBV4_TETNG	Chromosome 19 SCAF14664, whole genome shotgun sequence.	Tetraodon nigroviridis	bpg0135827	53.4	91.8	95.3	✓	✓	✓	✓	1

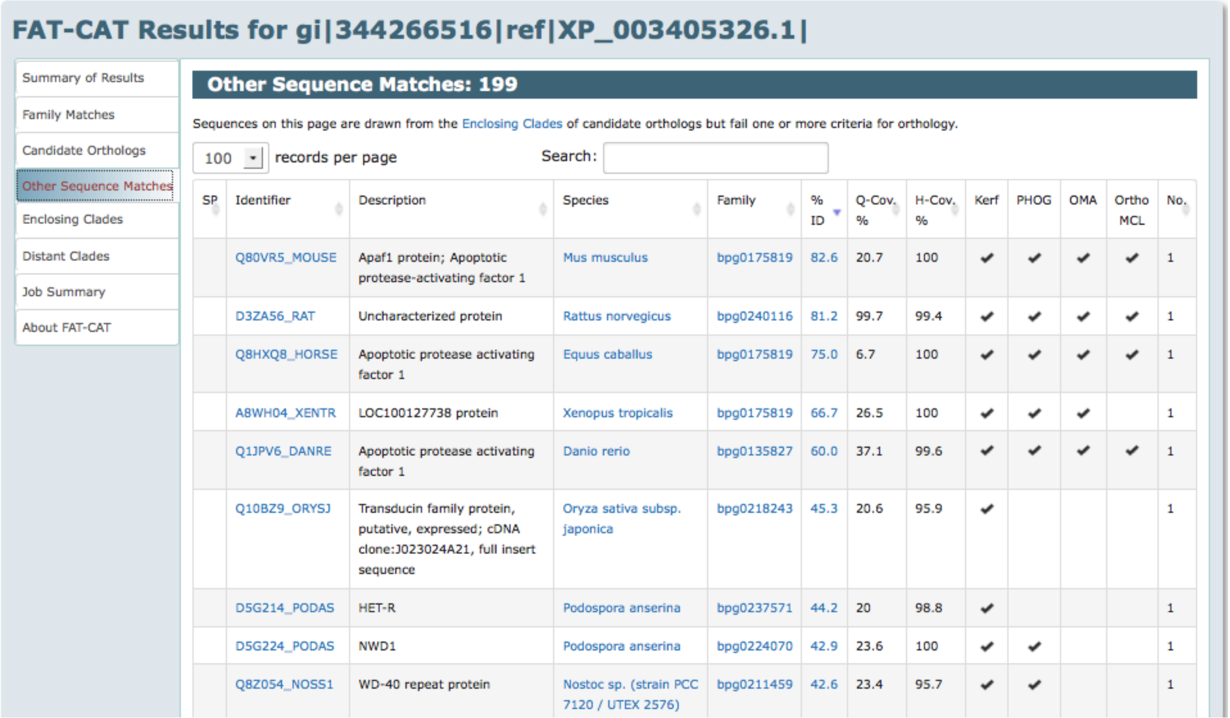
Showing 1 to 9 of 9 entries

Figure 8. Candidate Orthologs. In this example, nine candidate orthologs are identified; all are supported by all four orthology methods used. If multiple sequences are found from the same genome with the same sequence identity to the query, we pick one as the representative; if one of the cluster is in the SwissProt database, we use that as the representative, else we pick the one with the highest sequence identity.

The Candidate Orthologs data table includes data to help you evaluate the support for each proposed ortholog, and to explore data associated with that ortholog:

- Alignment statistics are provided, including the percent identity, query coverage and hit coverage. Clicking on the value in the %ID column will display the alignment between the query and candidate ortholog.
- Tool-tipping the species will display the common name for that species; clicking will bring you to the NCBI taxonomy.
- Clicking on the sequence identifier will bring you to the UniProt page for that sequence. SwissProt sequences are indicated by the red flag at far left.
- The bpg accession is linked to the family containing the enclosing clade in which the candidate ortholog was found.

4.5 Other Sequence Matches



FAT-CAT Results for gi|344266516|ref|XP_003405326.1|

Other Sequence Matches: 199

Sequences on this page are drawn from the Enclosing Clades of candidate orthologs but fail one or more criteria for orthology.

SP	Identifier	Description	Species	Family	% ID	Q-Cov. %	H-Cov. %	Kerf	PHOG	OMA	Ortho MCL	No.
	Q80VR5_MOUSE	Apaf1 protein; Apoptotic protease-activating factor 1	Mus musculus	bpg0175819	82.6	20.7	100	✔	✔	✔	✔	1
	D3ZA56_RAT	Uncharacterized protein	Rattus norvegicus	bpg0240116	81.2	99.7	99.4	✔	✔	✔	✔	1
	Q8HXQ8_HORSE	Apoptotic protease activating factor 1	Equus caballus	bpg0175819	75.0	6.7	100	✔	✔	✔	✔	1
	A8WH04_XENTR	LOC100127738 protein	Xenopus tropicalis	bpg0175819	66.7	26.5	100	✔	✔	✔		1
	Q1JPV6_DANRE	Apoptotic protease activating factor 1	Danio rerio	bpg0135827	60.0	37.1	99.6	✔	✔	✔	✔	1
	Q10BZ9_ORYSJ	Transducin family protein, putative, expressed; cDNA clone:J023024A21, full insert sequence	Oryza sativa subsp. japonica	bpg0218243	45.3	20.6	95.9	✔				1
	D5G214_PODAS	HET-R	Podospora anserina	bpg0237571	44.2	20	98.8	✔				1
	D5G224_PODAS	NWD1	Podospora anserina	bpg0224070	42.9	23.6	100	✔	✔			1
	Q8Z054_NOSS1	WD-40 repeat protein	Nostoc sp. (strain PCC 7120 / UTEX 2576)	bpg0211459	42.6	23.4	95.7	✔	✔			1

Figure 9. Other sequence matches. Sequences displayed on this page have been rejected as candidate orthologs due to failing one or more orthology criteria.

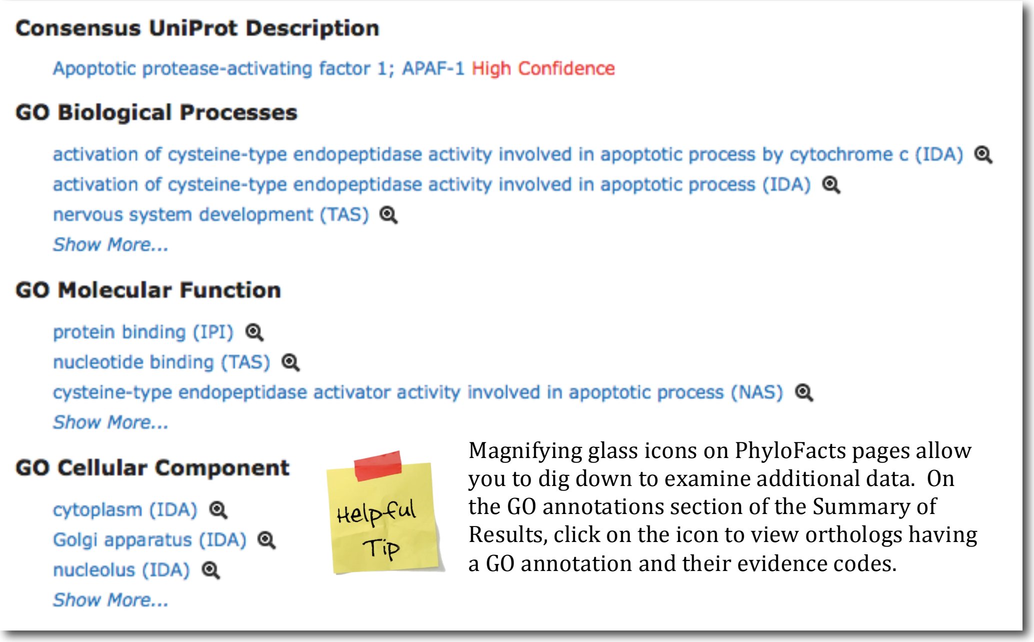
Why are some sequences rejected as orthologs despite having high sequence identity?

Sequences in an enclosing clade that are listed in Other Sequence Matches are typically rejected as orthologs due to not meeting alignment criteria. In this example, four of the top five fail the query coverage requirements, i.e., they align along only a subregion. The second sequence listed, from *Rattus norvegicus*, has 81% sequence identity and near-perfect overlap (for both the query and hit). However, another sequence from the rat genome was selected as the correct ortholog due to having higher sequence identity (87%).

4.6 How are functional annotations derived?

The **Functional Annotations** section of the Summary of Results displays predicted functions for the query derived from putative orthologs.

We derive a weighted consensus over the UniProt description of the protein's function, weighting close orthologs and manually curated annotations more than distant orthologs and annotations derived automatically. We also display the union over all Gene Ontology annotations (biological process, molecular function and cellular location) along with the best Evidence Code available for that annotation; clicking on the annotation will bring up a list of orthologs having that annotation.



Consensus UniProt Description

Apoptotic protease-activating factor 1; APAF-1 High Confidence

GO Biological Processes

activation of cysteine-type endopeptidase activity involved in apoptotic process by cytochrome c (IDA) 🔍
activation of cysteine-type endopeptidase activity involved in apoptotic process (IDA) 🔍
nervous system development (TAS) 🔍
Show More...

GO Molecular Function

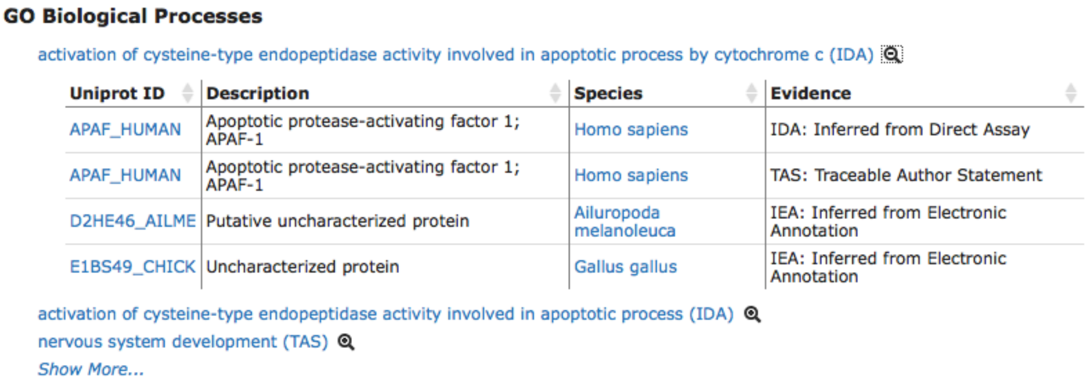
protein binding (IPI) 🔍
nucleotide binding (TAS) 🔍
cysteine-type endopeptidase activator activity involved in apoptotic process (NAS) 🔍
Show More...

GO Cellular Component

cytoplasm (IDA) 🔍
Golgi apparatus (IDA) 🔍
nucleolus (IDA) 🔍
Show More...

Magnifying glass icons on PhyloFacts pages allow you to dig down to examine additional data. On the GO annotations section of the Summary of Results, click on the icon to view orthologs having a GO annotation and their evidence codes.

Figure 10. Functional annotations and GO annotations derived for the query from orthologs identified by FAT-CAT.



GO Biological Processes

activation of cysteine-type endopeptidase activity involved in apoptotic process by cytochrome c (IDA) 🔍

Uniprot ID	Description	Species	Evidence
APAF_HUMAN	Apoptotic protease-activating factor 1; APAF-1	Homo sapiens	IDA: Inferred from Direct Assay
APAF_HUMAN	Apoptotic protease-activating factor 1; APAF-1	Homo sapiens	TAS: Traceable Author Statement
D2HE46_AILME	Putative uncharacterized protein	Ailuropoda melanoleuca	IEA: Inferred from Electronic Annotation
E1BS49_CHICK	Uncharacterized protein	Gallus gallus	IEA: Inferred from Electronic Annotation

activation of cysteine-type endopeptidase activity involved in apoptotic process (IDA) 🔍
nervous system development (TAS) 🔍
Show More...

Clicking on the magnifying glass next to a GO annotation shows the orthologs having that annotation along with the evidence codes.