FAT-CAT Webserver
Quick-Start Guide

The FAT-CAT webserver uses hidden Markov models (HMMs) to place sequences into PhyloFacts trees for functional annotation and ortholog identification. This Quick Start Guide gives an overview of how to submit sequences to FAT-CAT and how to interpret the results. Additional details on tuning pipeline parameters are provided in the Online Help.

Topics

The FAT-CAT input form (Figure 1) uses a standard interface for bioinformatics web servers such as BLAST and Pfam: just paste in your protein sequence (in either FASTA format or raw sequence) and click *Submit*.

Submitting a sequence to the FAT-CAT webserver.

Required: Paste the raw amino acid sequence into the Raw sequence box
Optional:
- If you want to receive results by email enter your email address. Otherwise, bookmark the results page that displays after you click Submit.
- You may enter an email subject and/or comments or leave these fields blank.
- If you want to use the FAT-CAT default settings then press submit; alternatively select one of the other preset parameters or manually modify the parameters and then press Submit.

Don't get enough results? Too many results?

Try program default parameters first, and resubmit with revised settings as needed**.** Program default parameters (High Recall) are intended to provide some results for most user inputs. If you submit a sequence and FAT-CAT returns too many matches, or mixes paralogs with orthologs, resubmit the query using High Precision settings. If insufficient results are returned, you can look at the Other Sequence Matches tab of the results, or resubmit with the Remote Homolog settings . See Figure 2 for additional details.

Figure 1 The FAT-CAT Webserver input form

A **Sequence header:** *Optional.* Input the query sequence name, identifier or accession.

B **Raw sequence:** *Required. Protein sequences only.*
Lengths accepted: between 30-2000 amino acids.

C **Email:** *Optional.* Users providing email addresses will receive a link to results once the job completes. Alternatively, you can click *Submit* and then bookmark the next page. Job results will display on that page as they become available, allowing you to track the progress of your submission.

D **Email subject:** *Optional.* If you leave this field empty, the email subject line will include a FAT-CAT job ID but no identifying details about the query. Users who plan to submit multiple sequences to FAT-CAT may wish to enter brief details for each submission (e.g., the sequence identifier or accession) to make it easier to scan your email for specific sequence inputs.

E **Comments:** *Optional.* If you modify any program parameters, you may want to record those modifications in this section. Comments entered here are stored in the Job Summary tab of the FAT-CAT results.

F **Submit:** Click on the Submit button to submit your sequence.

G **Pre-calculated results and sample input:** The pre-calculated results allow you to view the results of different query sequences. The sample input button allows you to populate the form to see the expected formats.

H **Preset parameters:** We provide four preset parameters: high recall, high precision, remote homologs and partial sequence. Figure 2 provides a decision tree to help you choose among these presets.

I **View/modify parameter settings.** Clicking on this link will display the parameters that can be modified manually. See Figure 4.
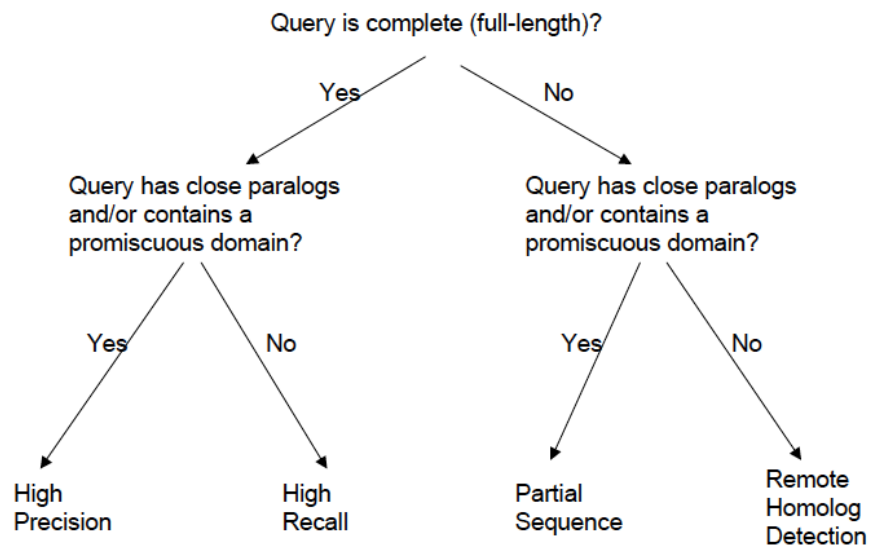
Selecting and modifying program parameters

Figure 2. Decision tree for selecting from FAT-CAT parameter preset options.
The four preset options for FAT-CAT – High Precision, High Recall, Partial Sequence and Remote Homolog Detection -- are designed to handle different types of input sequences.

Interpreting FAT-CAT results

Figure 6 shows a sample FAT-CAT results page, displaying the Family Matches for F7G2Z_MONDO, a predicted THO complex subunit 1 from *Monodelphis domestica* (opossum). The left column provides tabs to different results, including predicted functions, orthologs, and other FAT-CAT outputs.

FATCAT Results for tr|F7G2Z0|F7G2Z0_MONDO

Family Matches

Results shown here meet the following criteria (program defaults shown): significant E-Value; minimum length of aligned region >25aa; PhyloFacts-Pfam HMM overlap >70%; MDA HMM overlap >70% (bidirectional). Families are further investigated in Stages 2 and 3 to identify orthologous clades containing orthologs to the query. All other matches will be displayed in Other Sequences Matches.

Accession	Type	Consensus MDA	Description	Species	E-value	Region Aligned	HMM Coverage	Query Coverage
bpg0232621	MDA	efThoc1, Death	THO complex subunit 1; Tho1(related)	Bilateria	1.9e-289	4 - 650	95.1	98.3
bpg0206475	MDA	efThoc1	AT5g09860/MYH9_7; THO complex subunit 1(related)	rosids	3.3e-47	221 - 503	43.2	42.9
bpg0218621	MDA	efThoc1	Os03g0110400 protein; THO complex subunit 1, putative, expressed(related)	Poaceae	3.3e-45	226 - 506	41.3	42.6

Figure 3 The FAT-CAT Family Matches page. This tab display s all families that passed the criteria for Stage 1. The first column provides a link to the PhyloFacts family page for that match, where you can view data for the family and examine the phylogenetic tree.

FATCAT Results for tr|F7G2Z0|F7G2Z0_MONDO

Functional Annotations Based on Orthology

Consensus UniProt Description

THO complex subunit 1; Tho1 High Confidence

GO Biological Processes

apoptotic process (IDA) 🔍
regulation of DNA-dependent transcription, elongation (IDA) 🔍
intronless viral mRNA export from host nucleus (IDA) 🔍
Show More...

GO Molecular Function

protein binding (IPI) 🔍
RNA binding (IEA) 🔍
DNA binding (IEA) 🔍

GO Cellular Component

cytoplasm (IDA) 🔍
THO complex part of transcription export complex (IDA) 🔍
nucleolus (IDA) 🔍
Show More

Figure 4. Functional annotations tab of FAT-CAT results. The "Consensus UniProt Description" is a weighted consensus across all the orthologs identified by the pipeline, giving higher weight to close orthologs and to sequence annotations from the manually curated SwissProt database. The confidence score assigned to a UniProt annotation is based on defined cutoffs. GO annotations listed are retrieved from all orthologs and display the best GO Evidence Code for each annotation. If you click on the GO annotation you will see a list of the orthologs that have that annotation.

FATCAT Results for tr|F7G2Z0|F7G2Z0_MONDO

Other Sequence Matches: 30

Sequences in this group are primarily paralogs or sequences that have different multi-domain architectures. Occassionally, actual orthologs may be included among "Other Sequence Matches" due to having failed one of our Quality Assurance criteria. Gene model errors and sequence fragments often show up here. See additional details.

| 100 | records per page | | | Search: | | | | | | | |

SP	Identifier	Description	Species	Family	% ID	Q-Cov. %	H-Cov. %	Kerf	PHOG	OMA	Ortho MCL
	Q6Y1I8_PANTR	THO complex 1	Pan troglodytes	bpg0111022	95.0	30.4	100	✔	✔	✔	
	Q6TUH4_RAT	LRRGT00070; THO complex subunit 1	Rattus norvegicus	bpg0111022	87.7	75.2	99.0	✔	✔	✔	
✚	THOC1_RAT	THO complex subunit 1; Tho1	Rattus norvegicus	bpg0111022	72.8	50.8	97.4	✔	✔	✔	

Figure 5. Other Sequence Matches tab of FAT-CAT results. This tab displays sequences that did not pass defined ortholog criteria. Depending on the specifics of the query sequence and parameter settings, sequences shown on this page may be a mixture of orthologs and paralogs. Many candidate orthologs are rejected due to low overlap of the query (Q-Cov) or low overlap of the hit (H-Cov), or the sequence identity may fall below the minimum required. (These parameters are defined in Stage 3.)
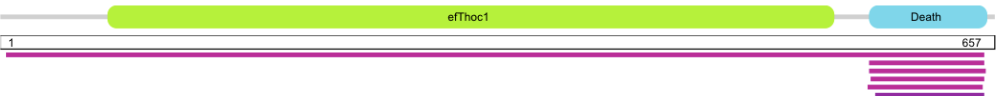
FATCAT Results for tr|F7G2Z0|F7G2Z0_MONDO

Orthologous Clades

Subtrees displayed here meet the following default criteria (unless modified by the user): the clade contains the top-scoring node (TSN); the alignment of the query to the TSN HMM meet specified coverage and percent identity criteria; sequence in the clade appear to be mutually orthologous (based on a consensus across multiple methods); >e;90% of the sequences in the clade satisfy query-ortholog criteria (alignment overlap and percent identity).

PhyloFacts Family		Enclosing Clade Data		Top-scoring Subtree Data						
Family	Type (Pfam(s))	Description	Species	E-value	Q-HMM %ID	Query Coverage	HMM Coverage	Region Aligned	Support	
bpg0232621	MDA : efThoc1, Death	THO complex subunit 1; Tho1(related)	Bilateria	0.0	95.3 %	99.8 %	99.8 %	1 - 657		
bpg0110975	Pfam : Death	THO complex subunit 1; Tho1	Chordata	1e-53	91.4 %	12.2 %	98.8 %	571 - 651		

Figure 6. Orthologous Clades tab. This tab displays PhyloFacts family clades containing the top-scoring node found in Stage 2 and for which a minimum fraction of the sequences pass Stage 3 criteria. The Pfam multi-
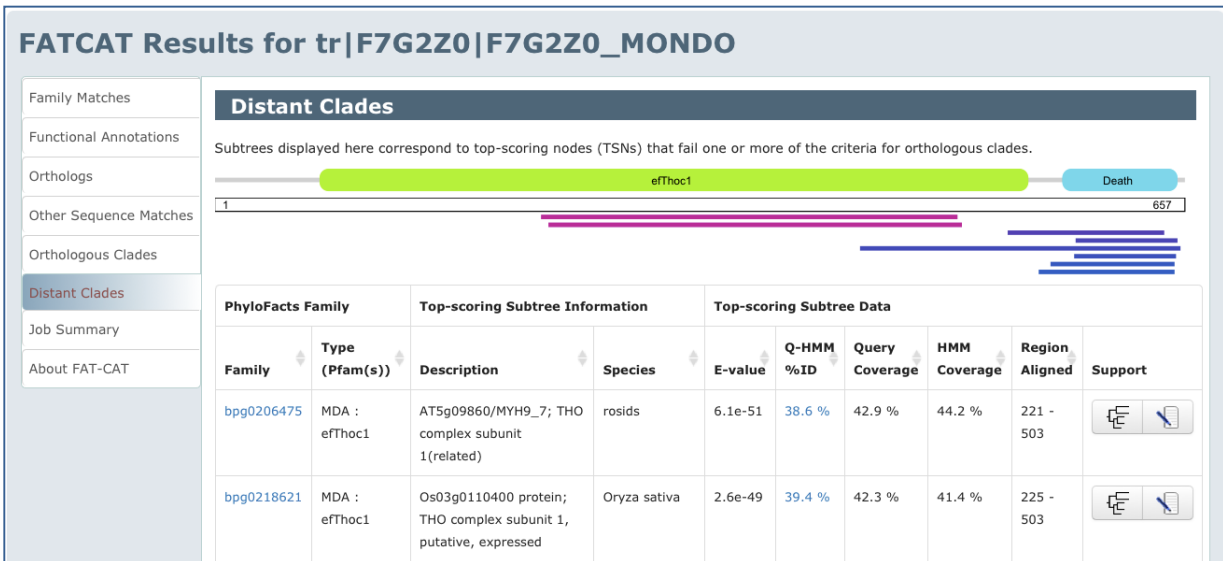
Figure 7. Distant clades tab. Subtrees for top-scoring nodes found in stage 2 that fail any criteria, either because the top-scoring node is outside any orthologous groups or because the sequence identity between the query and HMM falls below the defined minimum, or some other alignment criterion. Clicking on the Q-HMM %ID value will display the pairwise alignment between the query and the HMM. Clicking on the tree icon at far right will launch the PhyloScope tree viewer to display the tree. The page icon adjacent is linked to a page with additional data about that subtree.
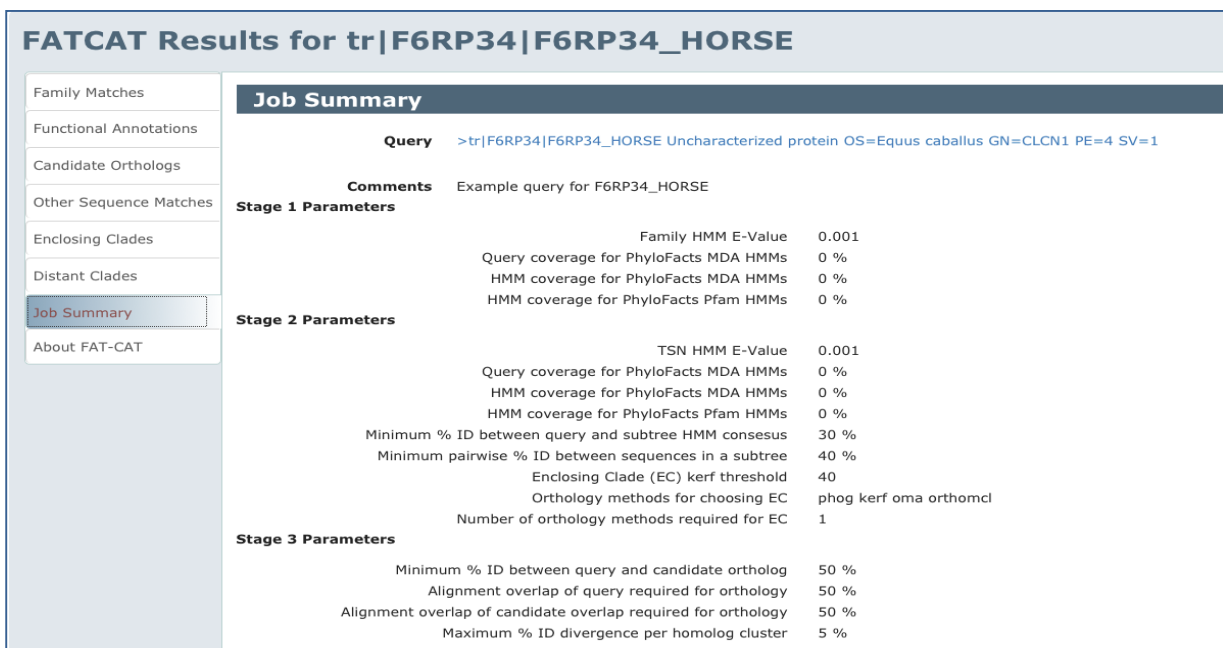

Figure 8 The FAT-CAT Job Summary page. The query sequence, pipeline parameters and comments on the Input Form are stored here.