Comparing the FAT-CAT webserver against other orthology prediction web servers

Description of benchmark dataset: We compared the FAT-CAT webserver against the top orthology databases providing broad taxonomic coverage (eggNOG, OrthoMCL, InParanoid, KEGG, PhylomeDB, OMA and OrthoDB). Comparisons involved significant and time-consuming manual analyses; the benchmark is therefore small and we do not make any claim that these results are statistically significant.

We selected seven proteins for this comparison: five vertebrate (four human and one chicken, all from the manually curated SwissProt database), one plant and one bacterial sequence from the human oral microbiome. Six proteins were from large multi-gene families with many distinct functional subtypes (i.e., paralogs): GPCRs, ion channels, transcription factors, Toll-like receptors, plant receptor-like proteins, inorganic pyrophosphatases. This dataset includes four sequences with promiscuous domains, to allow us to evaluate the robustness of orthology web servers to these data. Promiscuous domains are named for their propensity to combine with many different domains in distinct multi-domain architectures (MDAs, the ordered sequence of structural or functional domains). Promiscuous domains are quite common across all taxonomic lineages, but particularly so among eukaryotic species (see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2259109/). Promiscuous domains can result in deceptively significant E-values between proteins whose functions and evolutionary origins are quite distinct. When a promiscuous domain is present in a query sequence, ortholog identification or functional annotation based on a simple BLAST analysis can result in errors. Since most orthology prediction methods make use of local similarity scores derived by BLAST, we anticipated that these methods might not be robust to the presence of promiscuous domains. (The requirement of sharing a common multi-domain architecture is necessary but not sufficient for orthology. For instance, rhodopsin-like GPCRs all share the same multi-domain architecture (seven transmembrane domains, with the N-terminus outside and the C-terminus inside the cell); hundreds of paralogs exist in animal genomes.)

Web servers evaluated: We compared FAT-CAT (both High Recall and High Precision settings) against seven major orthology web servers: eggNOG, OrthoMCL, InParanoid, KEGG, PhylomeDB, OMA and OrthoDB. If a webserver provided two or more candidate orthology groups, we chose the group listed first (under the assumption that the order was intended as meaningful). KEGG allows two different modes of identifying possible orthologs: prior classification within KEGG to a KEGG Orthology (KO) group, and a consensus analysis of the top ten BLAST matches to an Orthology Cluster (OC); we evaluated both.

Evaluation protocol: We evaluated each of the webservers for their precision in classifying sequences to orthology groups evaluated based on agreement in multi-domain architecture (as indicated by Pfam-A hmmscan analysis and/or pairwise alignment of the test sequence and the predicted ortholog); and on agreement in functional subtype (when subtypes were clearly labelled and/or supported by SwissProt). Many predicted orthologs had minimal or no functional annotations; we attempted to analyze some of these (e.g., using BLAST and Pfam), but had to leave several as ambiguous. (Not all web servers make orthologs downloadable, and some only provide links to third-party databases; in some cases links were dead, so that we could not retrieve the predicted ortholog for analysis.) For each test sequence-webserver combination, there were four possible outcomes:

1. Perfect precision – no mixtures of different domain architectures, and no obvious paralogs (based on existing annotations, supplemented by BLAST analysis). (Indicated by a check mark in Supplementary Table 1.)
2. A small number of errors. (Indicated by a "e" in the table.)
3. A significant number of errors, whether due to the inclusion of different multi-domain architectures (indicated in the table by "D") and/or many paralogs (indicated by "P" in the table),
4. No matches found. (Indicated by a dash.)

Summary of findings: For these seven test sequences, FAT-CAT (using the High Precision settings) was the only web server to achieve perfect precision. Using high-recall parameter settings increased the number of predicted orthologs with a small number of paralogs, and included no sequences with different multi-domain architectures. OMA and PhylomeDB also had outstanding precision with very few errors.

In contrast, eggNOG, OrthoMCL, KEGG, InParanoid and OrthoDB had far lower precision on these test sequences. Each of these web servers attempted to predict a large number of orthologs from many species, but were not robust to the presence of promiscuous domains (mixing proteins with different multi-domain architectures) and a large fraction of their clusters mixed paralogs and orthologs.

Although this benchmark is small, its broad taxonomic scope and functional diversity are designed to allow us to anticipate these results will generalize to other inputs. However, FAT-CAT high precision comes at a considerable cost in computatioanl efficiency; all other web servers reviewed here return results quickly.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Protein accession/identifier	37700354	IPYR_HUMAN	KCNT1_CHICK	TLR1_HUMAN	FOXF1_HUMAN	5HT1A_HUMAN	429758323
Species of origin	rice	human	chicken	human	human	human	Actinomyces
FAT-CAT High Precision	✓	✓	✓	✓	✓	✓	✓
FAT-CAT High Recall	✓	✓	✓	e	e	✓	✓
PhylomeDB	D	✓	✓	✓	✓	✓	✓
OMA	✓	e	✓	D	✓	✓	✓
eggNOG	D	P	P	P,D	P	P	D
KEGG:KO	—	P	✓	e	P	P	—
KEGG:OC	D	P	P	P,D	P	P	D
InParanoid	D	e	P	D	e	✓ ?	D
OrthoMCL	D	e	P	P	✓	e	D
OrthoDB	—	P	P	P,D	✓	✓	D

Table 1. Results comparing FAT-CAT and major orthology web servers on seven test sequences.

Test sequences 2-6 are from the manually curated SwissProt, and show the SwissProt identifiers for those sequences. Sequences 1 and 7 show GenBank identifiers. FAT-CAT provides four preset parameter options; we show results for two here: High Precision and High Recall.

We confirmed that predicted orthologs agreed at the multi-domain architecture level using hmmscan vs Pfam-A for OMA, KEGG (both KO and OC), OrthoDB, OrthoMCL and eggNOG; these resources provide downloadable FASTA files (or make sequence retrieval relatively straightforward). For InParanoid and PhylomeDB, we analyzed a small number of sequences manually (submitting predicted orthologs to one of the Pfam web servers).

A check mark indicates perfect precision. The letter "D" indicates predicted orthologs disagreed at the level of multi-domain architecture; "P" indicates paralogs were included. A lower-case "e" indicates a small number of errors (either paralogs or disagreements in domain architecture). A "?" indicates ambiguous results. A dash indicates no results were returned from a web server for a query.

These results show the high precision of FAT-CAT (at both the High Precision and High Recall settings), PhylomeDB and OMA on this benchmark. In contrast, the other orthology web servers had greater difficulty in separating orthologs and paralogs, and were not robust to the presence of promiscuous domains.

Case 1 : Rice receptor-like protein (gi|37700354; Os03g48880.1)

GenBank entry: gi|37700354|gb|AAR00644.1| putative LRR receptor-like protein kinase
Taxonomy: *Oryza sativa subsp. japonica* (rice)
UniProt ID: Q851L1_ORYSJ
UniProt Description: Leucine Rich Repeat family protein
Sequence length: 508
Multi-domain architecture (Pfam):

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
LRRNT_2	Leucine rich repeat N-terminal domain	Family	n/a	39	67	45	66	21	42	43	20.4	0.00035	n/a
LRRNT_2	Leucine rich repeat N-terminal domain	Family	n/a	339	378	340	377	2	42	43	15.7	0.01	n/a

Summary: Pfam analysis of Os03g48880.1 shows it is composed entirely of leucine-rich repeats (LRRs). LRRs are one of the most common promiscuous domains, found in many different multi-domain architectures. In plant genomes, LRR domains are found in many proteins, including receptor-like proteins (RLPs) and receptor-like protein kinases (RLKs). RLKs have an N-terminal extracellular LRR region, a transmembrane helix and a cytoplasmic kinase domain at the C-terminus. RLPs are identical to RLKs along their extracellular domain and transmembrane helix, but end in a short cytoplasmic tail (i.e., they have no kinase domain). (RLPs are reviewed in Fritz-Laylin, Krishamurthy, Tor, Sjölander and Jones, "Phylogenomic analysis of the receptor-like proteins of rice and Arabidopsis", Plant Physiology, June 2005.) The GenBank annotation for 37700354 is “putative LRR receptor-like protein kinase” (i.e., an RLK), but this is clearly incorrect; no kinase domain is present. Although no transmembrane domain is predicted by transmembrane prediction tools, GenBank sequence 37700354 has close homologs to receptor-like proteins (RLPs) and is likely to be a member of the RLP family.

Given the limited functional data for most of the matches, we evaluated results of web server predictions almost entirely on the basis of the multi-domain architecture: sequences having a detectable kinase domain (based on Pfam analysis) were rejected as possible orthologs. Only two orthology web servers correctly restrict predicted orthologs to those sharing the same MDA: FAT-CAT (both High Precision and High Recall) and OMA. All other web servers merge proteins with diferent multi-domain architectures in their orthology groups.

Web servers with matches agreeing at the multi-domain architecture

FAT-CAT: Both **High Precision** and **High Recall** settings returned a single ortholog (A2XKY0_ORYSI, in *Oryza sativa subsp. Indica*) with the same multi-domain architecture (no kinase domain).

OMA returned one ortholog (UniProt: C5WNY6, from *Sorghum bicolor; Sb01g011385*) verified manually to be restricted to LRRs.

Web servers including matches with different multi-domain architectures

OrthoMCL classified the query to OG5_141123, containing 14 sequences from 5 taxa, including many containing kinase domains (e.g. , atha|NP_176789, Arabidopsis TMK1).

KEGG: OC: Two OCs (orthology clusters) were predicted: OC.324320 and OC.254553; both contained proteins with kinase domains.

PhylomeDB. The top-scoring sequence match found by the PhylomeDB BLAST search is an RLK (Phy00013M8_ARATH, P43298, TMK1_ARATH). Predicted orthologs in the tree for TMK1_ARATH are also RLKs.

InParanoid: many predicted orthologs contain kinase domains.

eggNOG returned two orthology groups. NOG273280 contained 6 sequences matching the LRR region, and having no kinase domains. KOG0619 had 10,006 proteins and contained many kinases.

Web servers with no matches

OrthoDB returned no matches.

KEGG:KO: The top BLAST hit (to a TMK-like protein from *Brachypodium distachyon*: 100844416) had not been classified to a KO (KEGG orthology) group.

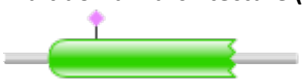
Links to results (URLs longer than one line have been shortened):

FAT-CAT (high precision)	http://makana.berkeley.edu/phylofacts/fatcat/2507/
FAT-CAT (high recall)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2889/
OrthoDB	no matches

KEGG (KO)	N/A
KEGG (OC)	http://www.genome.jp/tools-bin/ocv?entry=OC.254553; http://www.genome.jp/tools-bin/ocv?entry=OC.324320
OrthoMCL	http://tr.im/42876
OMA	http://tr.im/42877
PhylomeDB	http://phylomedb.org/?q=search_tree&seqid=Phy00013M8_ARATH
InParanoid	http://tr.im/42878
eggNOG	http://eggnog.embl.de/version_3.0/cgi/blast_results.py?type=blast&id=ljzlecn

Case Study 2: Human Inorganic pyrophosphatase (IPYR_HUMAN)

SwissProt ID: IPYR_HUMAN
Gene name: PPA1
Description: Inorganic pyrophosphatase
Sequence length: 289
Multidomain architecture (Pfam):



Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
Pyrophosphatase	Inorganic pyrophosphatase	Domain	n/a	45	229	45	228	1	**155**	156	186.1	2.7e-55	90

Summary: Inorganic phyrophosphatases are present in all genomes. Two paralogs are present in the human genome: PPA1 and PPA2. FAT-CAT and PhylomeDB had perfect precision on this test sequence; eggNOG, KEGG and OrthoDB merged many paralogous PPA2 genes with orthologous PPA1. InParanoid, OrthoMCL and OMA had a small number of errors.

Web servers containing no obvious paralogs (all presumed orthologs):

FAT-CAT high precision returned 14 PPA1 matches (including the query, for a total of 13 orthologs).
FAT-CAT high recall returned a cluster of 47 PPA1 matches (including the query, for a total of 46 orthologs).
PhylomeDB returned 12 orthologs below the duplication node, all PPA1.

Web servers with few/minor errors/questionable results

InParanoid matches appear to be mostly PPA1, with the single exception of LmjF11.0210 from *Leishmania major strain Friedlin* (XP_001681489), which appears more closely related to PPA2 than to PPA1. BLAST analysis of this protein against SwissProt (using the UniProt BLAST server) ranks IPYR2_HUMAN (2.0×10-62) and IPYR2_MOUSE (3.0×10-64) higher than IPYR_HUMAN (2.0×10-57) and IPYR_MOUSE (2.0×10-58).

OrthoMCL returned 118 sequences from 90 taxa. Spot checking these results shows two PPA2 genes: Horse sequence ecab|XP_001503241 annotated as "inorganic pyrophosphatase 2, mitochondrial-like" and macaque protein mmul|XP_001082969, annotated as "inorganic pyrophosphatase 2, mitochondrial". BLAST analysis versus SwissProt shows both have significantly stronger scores to human and mouse PPA2 than to human and mouse PPA1 genes.
OMA returned 69 one-to-one orthologs, 68 of which appeared to be PPA1. The single PPA2 is IPYR2_SCHPO (annotated by SwissProt as PPA2).

Web servers merging multiple paralogs (PPA1 and PPA2)

eggNOG returned 10 orthology groups, 6 of which merged PPA1 and PPA2.
OrthoDB – 106 sequences in 52 species -- merged PPA1 and PPA2.
KEGG:KO for query (K01507) merged PPA1 and PPA2.
KEGG: OC: KEGG OC OC.356015 has 276 sequences, merging PPA1 and PPA2.

Links to results (URLs longer than one line have been shortened):

FAT-CAT (high precision)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2849/
FAT-CAT (high recall)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2857/
OrthoDB	http://tr.im/4287c
KEGG (KO)	http://www.genome.jp/dbget-bin/www_bget?ko:K01507
KEGG (OC)	http://tr.im/4287d
OrthoMCL	http://tr.im/4287f
OMA	http://omabrowser.org/cgi-bin/gateway.pl?f=DisplayEntry&p1=HUMAN00769
PhylomeDB	http://phylomedb.org/?q=search_tree&seqid=Phy000808R_HUMAN
InParanoid	http://tr.im/4287g
eggNOG	http://eggnog.embl.de/version_3.0/cgi/blast_results.py?type=blast&id=sdzbpwa

Case 3: Human Potassium channel subfamily T member 1 (KCNT1_CHICK)

SwissProt ID: KCNT1_CHICK
UniProt accession: Q8QFV0
Taxonomy: *Gallus gallus* (chicken)
Sequence length: 1201
Multidomain architecture (Pfam):

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
Ion_trans_2	Ion channel	Domain	CL0030	254	333	257	325	2	70	79	46.8	1.7e-12	n/a
BK_channel_a	Calcium-activated BK potassium channel a ...	Family	n/a	473	578	474	578	2	101	101	121.7	9.9e-36	n/a

Summary: Potassium channel genes are present in all genomes. Multiple genes encode potassium channels in vertebrate genomes. KCNT1 in vertebrate genomes is closely related to the paralogous KCNT2; both forms have six transmembrane helices over the first ~300 amino acids. On this test sequence, FAT-CAT, OMA, PhylomeDB and KEGG (KO) had perfect precision. Web servers mixing KCNT1 and KCNT2 include eggNOG, InParanoid, KEGG (OC), OrthoDB and OrthoMCL.

Web servers containing no obvious paralogs and no sequences with different multi-domain architectures

FAT-CAT: High Precision. An orthology cluster of 8 sequences (including the query, for a total of 7 orthologs) was returned; all KCNT1.

FAT-CAT: High Recall: An orthology cluster of 15 sequences (including the query, for a total of 14 orthologs) was returned; all KCNT1.

OMA returned 41 one-to-one orthologs (including the query), all KCNT1.

PhylomeDB's top three matches found by BLAST are the query (KCNT1_CHICK); however, these lack associated trees. The fourth match is to Phy001R6CS_HUMAN (KCNT1_HUMAN), for which a tree is available. In that tree KCNT1_HUMAN is under a duplication event in the tree, so is the only ortholog predicted.

KEGG KO: The KO for the query, K04946, had 31 members (including the query); all appear to be orthologs. (Entry 100316919 from *Xenopus tropicalis* is annotated as KCNT2, but appears more closely related to KCNT1).

Web servers mixing paralogs with orthologs:

eggNOG returned 8 orthology groups, 6 of which mixed KCNT1 and KCNT2.

OrthoDB returned Group EOG64QV6J: 98 genes in 51 species, a mix of KCNT1 and KCNT2.

KEGG OC: KEGG OC results (OC.331359; 89 sequences) merge KCNT1 and KCNT2 sequences.

OrthoMCL returned 50 sequences in 28 species, a roughly equal mixture of KCNT1 and KCNT2.

InParanoid returned both vertebrate and invertebrate sequences. Spot-checking results showed several probable paralogs. Cluster 1491 shows a predicted ortholog in *Batrachochytrium dendrobatidis* JAM81, 34605; this sequence appears to be more closely related to KCMA1 genes than to either KCNT1 or KCNT2. Cluster 186 shows a predicted ortholog in *Caenorhabditis remanei*, RP47316; this has closer matches to human and mouse KCNT2 than to human and mouse KCNT1 (the score differences between RP47316 and KCNT1 and KCNT2 chicken are negligible). Cluster 1133 shows a match to *Leishmania major* protein LmjF01.0810; this protein has stronger scores to chicken, human and rat KCNT2 than to the KCNT1 genes in these species. Cluster 1403 shows a predicted ortholog to *Phytophthora sojae* sequence 136732; this protein has stronger scores to KCNT2 genes in chicken, human and rat than to KCNT1 genes in these species.

Links to results (URLs longer than one line have been shortened):

FAT-CAT (high precision)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2869/
FAT-CAT (high recall)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2838/
OrthoDB	http://tr.im/427xp
KEGG (KO)	http://www.genome.jp/dbget-bin/www_bget?ko:K04946
KEGG (OC)	http://tr.im/427xq
OrthoMCL	http://tr.im/427xt
OMA	http://tr.im/427xr
PhylomeDB	http://phylomedb.org/?q=search_tree&seqid=Phy001R6CS_HUMAN
InParanoid	http://tr.im/427xs
eggNOG	http://eggnog.embl.de/version_3.0/cgi/blast_results.py?type=blast&id=kjtaqgg

Case 4: Human Toll-like receptor 1 (TLR1_HUMAN)

SwissProt ID: TLR1_HUMAN
UniProt accession: Q15399
Sequence length: 786
Multidomain architecture (Pfam):

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
LRR_8	Leucine rich repeat	Repeat	CL0022	69	126	69	125	1	**60**	61	37.5	1.4e-09	n/a
LRR_8	Leucine rich repeat	Repeat	CL0022	445	501	445	500	1	**59**	61	30.9	1.6e-07	n/a
TIR	TIR domain	Family	CL0173	639	775	639	774	1	**140**	141	130.2	3.4e-38	n/a

Summary: Toll-like receptors (TLRs) in vertebrate genomes are related to insect Toll receptors; all are involved in the innate immune response. TLRs have an extracellular leucine rich repeat (LRR) region, a transmembrane domain and a cytoplasmic Toll-Interleukin Receptor (TIR) domain. Ten paralogous TLRs exist in the human genome (TLRs 1-10). The presence of a promiscuous domain (the extracellular LRR region) and recent lineage-specific expansions present distinct challenges to orthology prediction; most of the web servers tested merged paralogs with orthologs, and some included sequences with different multi-domain architectures (matching only the LRR region). On this test sequence, only FAT-CAT (with high precision settings) and PhylomeDB had perfect precision, with FAT-CAT returning 21 orthologs and PhylomeDB returning only 4. FAT-CAT (high-recall) and KEGG KO included on paralog each and no errors in multi-domain architecture. InParanoid and OMA had a small number of MDA errors (including sequences with partial homology, missing the TIR domain). Web servers merging many paralogs include KEGG (both OC), eggNOG, OrthoMCL and OrthoDB.

Web servers containing no obvious errors (all presumed orthologs, no differences in domain architectures):
FAT-CAT high precision returned a cluster of 22 orthologs (including the query), all TLR1.
PhylomeDB returned 4 sequences, all TLR1.

Web servers with a small number of paralogs and no MDA errors
FAT-CAT high recall returned a cluster of 31 orthologs (including the query) all have the same MDA; only one is a paralog (TLR6_DASNO, from nine-banded armadillo).

KEGG KO (K05398) returned 22 orthologs, of which one is a paralog (771173, a chicken TLR6).

Web servers merging multiple paralogs and/or sequences with different multi-domain architectures
eggNog returned 10 orthology groups, 7 of which contained paralogs. The top group – NOG272762 – had 83 proteins in 42 species, mixing numerous paralogs (TLR1, TLR6, and TLR10) and also including proteins lacking either LRR or TIR domains.

OrthoDB returned 96 genes in 48 species, mixing numerous paralogs (TLR1, TLR6, and TLR10), and also including four sequences lacking the TIR domain.

OrthoMCL returned 25 sequences in 13 species, mixing numerous paralogs (TLR1, TLR6, and TLR10).

KEGG OC (Orthology Cluster) results mixed paralogous genes (TLR1, TLR2, TLR6, and TLR10) and one sequence with a kinase domain (xtr:100487973).

InParanoid returned numerous matches, most of which appear credible. Two are missing the TIR domain (i.e., do not share the same multi-domain architecture): *Populus trichocarpa* protein 581697 (identical to gi|224124498) (Cluster 2587) and Tc00.1047053510329.100 (UniProt: Q4E3K3_TRYCC) from *Trypanosoma cruzi* (Cluster 1886).

OMA returned 29 sequences. Of these, 26 are probable orthologs. Three are missing the TIR domain so do not share the same MDA and cannot be orthologs: TUPBE04858, DIPOR14309 and DROVI00112.

Links to results (URLs longer than one line have been shortened):

FAT-CAT (high precision)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2848/
FAT-CAT (high recall)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2839/
OrthoDB	http://tr.im/4287j
KEGG (KO)	http://www.genome.jp/dbget-bin/www_bget?ko:K05398
KEGG (OC)	http://tr.im/4287k
OrthoMCL	http://tr.im/4287m
OMA	http://tr.im/428r6
PhylomeDB	http://tr.im/428r7
InParanoid	http://tr.im/4287n
eggNOG	http://tr.im/428qz

Case 5: Human Forkhead box protein F1 (FOXF1_HUMAN)

SwissProt ID: FOXF1_HUMAN
UniProt accession: Q12946
Sequence length: 379
Multidomain architecture (Pfam):

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
Fork_head	Fork head domain	Domain	n/a	48	143	48	140	1	93	96	131.2	1.1e-38	n/a

Summary: FOXF1 is one of dozens of forkhead box transcription factors in human, the closest paralog of which is FOXF2. Most web servers had problems differentiating orthologs and paralogs. Five web servers had perfect separation of orthologs and paralogs: FAT-CAT (high precision), OrthoDB, OrthoMCL, OMA, and PhylomeDB. Web servers including large numbers of paralogs include eggNOG and KEGG (particularly the KEGG OC).

Web servers containing no obvious paralogs

FAT-CAT high precision cluster had 9 members (including the query), all appear to be FOXF1**.**

OrthoDB returns 38 sequences in 38 species; all appear to be FOXF1.

OrthoMCL returned 11 sequences in 11 species; all appear to be FOXF1.

OMA returned 33 1-1 orthologs; all appear to be FOXF1.

PhylomeDB's top ranked tree contained an orthology group with the query sequence and no vertebrate orthologs (all homologs were from insect). The second-ranked tree included the query sequence and 11 vertebrate FOXF1 orthologs.

Web servers with a small number of paralogs

FAT-CAT high recall cluster included 3 FOXF2 paralogs – C1K2Z2_ORYLA (medaka), B1AB75_SCYCA (Small-spotted catshark) and D3ZU72_RAT.

InParanoid one-to-one orthologs included four paralogs: *Monosiga brevicollis* 24248 (Cluster 3214); BLAST against SwissProt at NCBI shows 4 human paralogs with stronger scores. *Schizosaccharomyces pombe* 972h-.31 SPBC32H8.11 (Cluster 1755); BLAST shows 9 human paralogs with stronger scores. *Caenorhabditis elegans* CE04965 (Cluster 3490); BLAST shows 13 human paralogs with stronger scores. *Caenorhabditis remanei* RP18324 (Cluster 3378); BLAST shows 12 human paralogs with stronger scores.

Web servers merging many paralogs

eggNOG returned 10 groups, of which six merged multiple paralogous groups (four with a mixture of FOXF1 and FOXF2, and two containing multiple subtypes).

KEGG:KO The KO containing the query included 58 sequences in 39 species; 10 were FOXF2.

KEGG OC (Orthology Cluster) group contained 718 sequences including numerous FOX paralogous subtypes (FOXA1, FOXA2, FOXI3, FOXD1, FOXB2, FOXC1, FOXL1 etc.).

Links to results (URLs longer than one line have been shortened):

FAT-CAT (high precision)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2845/
FAT-CAT (high recall)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2841/
OrthoDB	http://tr.im/428gr
KEGG KO	http://www.genome.jp/dbget-bin/www_bget?ko:K09399
KEG OC	http://tr.im/428qy
OrthoMCL	http://tr.im/428gv
OMA	http://omabrowser.org/cgi-bin/gateway.pl?f=DisplayEntry&p1=HUMAN06033
PhylomeDB	http://phylomedb.org/?q=search_tree&seqid=Phy0024HUJ_HUMAN
InParanoid	http://tr.im/428gx
eggNOG	http://eggnog.embl.de/version_3.0/cgi/blast_results.py?type=blast&id=xvqvuiz

Case 6: Human Serotonin receptor 1A (SwissProt: 5HT1A_HUMAN)

UniProt accession: P08908
Description: 5-hydroxytryptamine receptor 1A
Sequence length: 422
Multidomain architecture (Pfam):

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
7tm_1	7 transmembrane receptor (rhodopsin fami ...	Family	CL0192	53	400	53	400	1	257	257	327.1	5.8e-98	n/a

Summary: 5HT1A_HUMAN is a human serotonin receptor, one of ~1000 G protein-coupled receptors (GPCRs) in the human genome. GPCRs form one of the largest superfamilies found in animal genomes, recognizing hundreds of endogenous ligands (serotonin, dopamine, histamine, opioids, cannabinoids, odorants, pheromones, etc.). Numerous paralogous serotonin receptor subtypes exist; closely related receptors include dopamine, histamine, and adrenergic receptors. The Pfam 7TM_1 domain recognizes hundreds of GPCR subtypes. Web servers with perfect separation of orthologs and paralogs include FAT-CAT (both high recall and high precision), OrthoDB, OMA and PhylomeDB. Web servers merging many paralogous genes include eggNOG and KEGG (both KO and OC).

Web servers containing no obvious paralogs (all presumed orthologs):

FAT-CAT high precision returns a cluster of 16 orthologs (including the query).

FAT-CAT high recall returns a cluster of 20 orthologs (including the query).

OrthoDB returned 43 sequences in 40 species with no apparent errors.

OMA returns 49 1-1 orthologs with no obvious paralogs.

PhylomeDB's top-ranked tree includes a subtree with the query sequence and 10 orthologs – all 5HT1A.

Web servers with minor possible errors or some ambiguous results

InParanoid returned many predicted ortholog from both vertebrate and invertebrate genomes. Invertebrate orthologs were octopamine/tyramine receptors (accepted in this analysis as possible orthologs to vertebrate serotonin receptors). Most vertebrate matches appear correct; one was questionable: *Branchiostoma floridae* sequence 60377, Cluster 3785 has closer matches to other serotonin receptor subtypes than to 5HT1A.

OrthoMCL group OG5_133249 has 32 sequences in 24 species, divided between vertebrate and invertebrate (mostly insect) species. Spot checking revealed two possible paralogs. *Tetraodon nigroviridis* (green spotted puffer fish) sequence ENSTNIP00000016556 has closer matches to alpha adrenergic receptors and other GPCRs than to serotonin receptors of any subtype. *Apis mellifera* (honeybee) sequence NP_001011594 is a probable paralog, with closer BLAST matches to serotonin receptor subtypes 1B and 1F than to 1A.

Web servers merging multiple paralogs

KEGG KO. The KO for the query merges many paralogous serotonin receptor subtypes – 5HT1A, B, D, E, and F.

KEGG OC: The OC (Orthology Cluster) contains 1046 sequences with many paralogous receptors including adenosine, alpha adrengergic, dopamine, muscarinic acetylcholine, histamine, etc.

eggNOG returned 10 clusters, 6 of which merged many paralogous groups. The first cluster (NOG249628) includes multiple serotonin subtypes, alpha adrenergic, dopamine and octopamine receptors.

Links to results (URLs longer than one line have been shortened):

FAT-CAT (high precision)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2846/
FAT-CAT (high recall)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2842/
OrthoDB	http://tr.im/428hh
KEGG (KO)	http://www.genome.jp/dbget-bin/www_bget?K04153
KEGG (OC)	http://tr.im/428hg
OrthoMCL	http://tr.im/428hl
OMA	http://tr.im/428h7
PhylomeDB	http://phylomedb.org/?q=search_tree&seqid=Phy0008FRG_HUMAN
InParanoid	http://tr.im/428h6
eggNOG	http://eggnog.embl.de/version_3.0/cgi/blast_results.py?type=blast&id=vtfkmdr

Case 7: Metagenome sequence from human oral microbiome

GenBank accessions: gi|429758323|ref|ZP_19290840.1|

Taxonomy: *Actinomyces sp. oral taxon*
Sequence length: 429
Multi-domain architecture (Pfam):

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value
				Start	End	Start	End	From	To			
DUF1212	Protein of unknown function (DUF1212)	Family	CL0470	17	214	17	214	1	193	193	151.5	1.8e-44
DUF3815	Protein of unknown function (DUF3815)	Family	CL0470	277	407	278	405	**2**	**128**	130	65.2	5.1e-18

Summary: This protein of unknown function and unknown taxonomic origin has two Pfam domains: DUF1212 and DUF3815; both Pfam domains are promiscuous. Because the query had no known function, actual orthology calls were not possible. Instead, we analyzed predicted orthologs for agreement with the multi-domain architecture.

Web servers effectively separating sequences with different multi-domain architectures include FAT-CAT (both high recall and high precision settings), PhylomeDB, OMA and PhylomeDB. FAT-CAT retrieved the closest matches (including one with 75.8% identity), whereas matches predicted by other web servers all had <= 43% identity. OMA retrieved the largest number of homologs with the same multi-domain architecture. KEGG, eggNOG, OrthoMCL and OrthoDB all mixed proteins with different multi-domain architectures in orthology groups.

Web servers with no obvious errors (checking only for errors in multi-domain architecture).

FAT-CAT high recall: Six predicted orthologs, four from Actinomyces, all sharing the same multi-domain architecture; the closest ortholog (from Actinomyces) had 75.8% identity.

FAT-CAT high precision: Two predicted orthologs from Actinobacteria, both with the same multi-domain architecture. The closest match had 75.8% identity.

PhylomeDB: The top-scoring tree shows one candidate ortholog from *Rhodopseudomonas palustris*, an alphaproteobacterium (JPSR_RHOPA; tr|Q6NA56|Q6NA56_RHOPA). Pfam analysis of this protein shows the same MDA as query (DUF1212 followed by DUF3815). Analyzing the tree shows a duplication event just above the leaf for this sequence, so that no additional orthologs are supported by the tree.

OMA. The closest match in OMA (ARCHD00411; UniProt: D7BMN1; from *Arcanobacterium haemolyticum*) had 43% identity and a global alignment (same Pfam MDA). All 72 members of the cluster containing ARCHD00411 appear to have the same MDA.

Web servers with obvious errors (including sequences with different multi-domain architectures)

eggNOG. Three clusters are returned. Two clusters include members with only the DUF1212 domain -- COG2966 (516 proteins in 396 species) and bactNOG36857 (30 proteins in 30 species). For example, COG2966 includes 568813.SSUSC84_1020 (UniProt: C6GNW7_STRSX) a putative membrane protein from *Streptococcus suis* and bactNOG36857 includes 282458.SAR0591 (UniProt: Q6GJ83_STAAR), a putative membrane protein from *Staphylococcus aureus*, and 65393.PCC7424_2957 (UniProt: B7KA04_CYAP7), a putative integral membrane protein from *Cyanothece sp*.

OrthoDB: Group EOG68KPWS contains 2 genes in 2 species. Both contain only the DUF1212 domain (i.e., are missing the DUF3815 domain).

OrthoMCL: The closest match in the OrthoMCL database is from *Yersinia enterocolitica* (yent|YP_001004977.1). This protein has a DUF1212 domain only (as do several other members of the orthology cluster for this protein, OG5_138621).

KEGG OC: The Orthology Cluster produced by KEGG BLAST analysis (OC.777633) had 450 sequences spanning different multi-domain architectures. Many members contain only the DUF3815 domain (e.g., *Escherichia coli* K-12 MG1655: b4363 (UniProt: P0ADD2), annotated as a "conserved inner membrane protein", *Psychrobacter arcticus*:

Psyc_1692 (UniProt: Q4FR18), annotated as “ThrE family exporter small subunit”, and *Denitrovibrio acetiphilus*: Dacet_2128 (UniProt: D4H299)).

InParanoid: The closest match in InParanoid was to *E.coli* K12 protein NP_418784.4. Clicking on the link for that entry yielded no results. However, Pfam analysis of the protein shows only a DUF1212 domain.

Web servers with no results.

KEGG KO (Kegg Orthology). The closest homolog to the query is rmu:RMDY18_06380, with only 39% identity. This protein has no KO.

Links to results (URLs longer than one line have been shortened):

FAT-CAT (high precision)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2890/
FAT-CAT (high recall)	http://phylogenomics.berkeley.edu/phylofacts/fatcat/2859/
OrthoDB	http://tinyurl.com/bptt3vn
KEG OC	http://tinyurl.com/d9elfwf
OrthoMCL	http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=sequenceList&groupac=OG5_138621
OMA	http://tinyurl.com/bmfob6u
PhylomeDB	http://phylomedb.org/?q=search_tree&seqid=Phy001JPSR_RHOPA
InParanoid	http://tinyurl.com/csc7hlm
eggNOG	http://eggnog.embl.de/version_3.0/cgi/blast_results.py?type=blast&id=nycxvah