Clustering analyses of eukaryotic proteomes

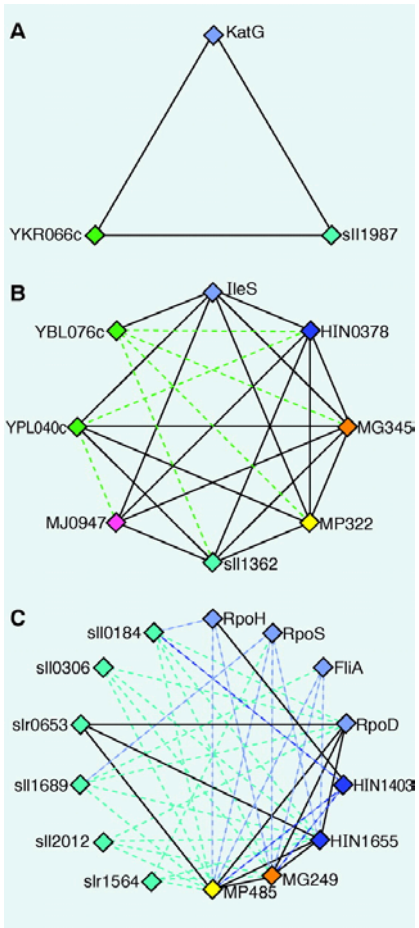
Sergei Mekhedov
NCBI

What you will have to listen about

- Introduction: orthologs, paralogs, COGs, and KOGs
- Public KOG db: curated clusters for 7 species
- Orthology domains
- Relative vs. absolute KOGs
- Impossibility to continue the way KOGs were made before
- A semi-automatic approach to create curated KOGs
- Current status of KOG project for 50 eukaryotic species
- Efficient editing interface is a necessary prerequisite for accurate KOGs
- Problem of proteome updates in KOGs
- Features of new KOGnitor
- A way to improve gene predictions in available proteomes
- HMMs for KOGs – the way to save time on a long run
- Superfamilies and KOGs
- KOG applications
- Phylogenetic profile search
- What you need to be a real parasite

Tatusov RL, Koonin EV, Lipman DJ.
Science. 1997, 278(5338):631-637.
A genomic perspective on protein families.

Examples of simple COGs based on BeTs



BeTs are best hits of a query protein in a particular proteome

solid lines – symmetrical BeTs

broken lines – asymmetrical BeTs

from Tatusov et al., 1997

COG

Clusters of Orthologous Groups

KOG

Eukaryotic Clusters of Orthologous Groups

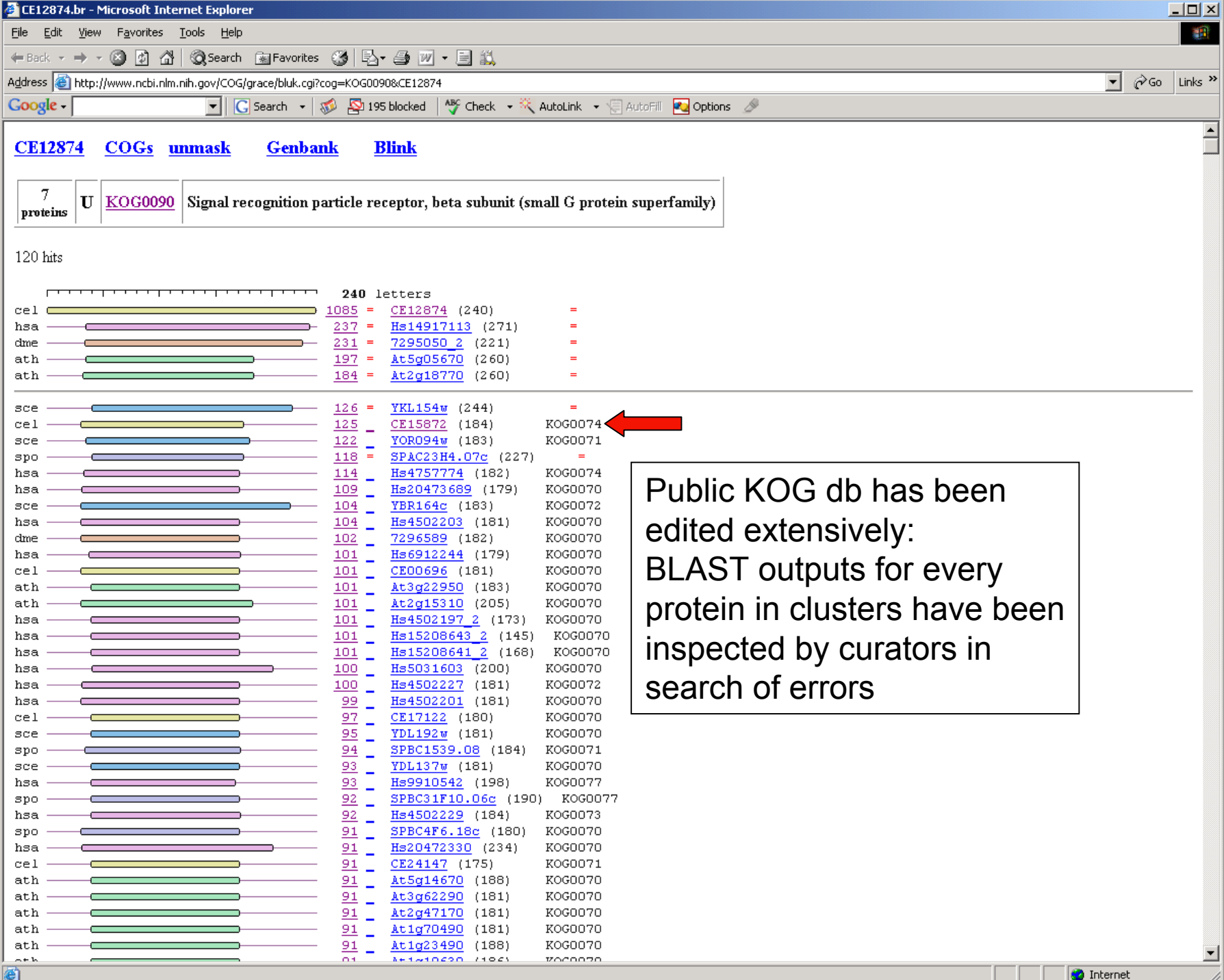
Public version of KOG database was created as a result of 2+ years of hard work by 20 programmers and computational biologists at NCBI. It included protein sequences from 7 eukaryotic species

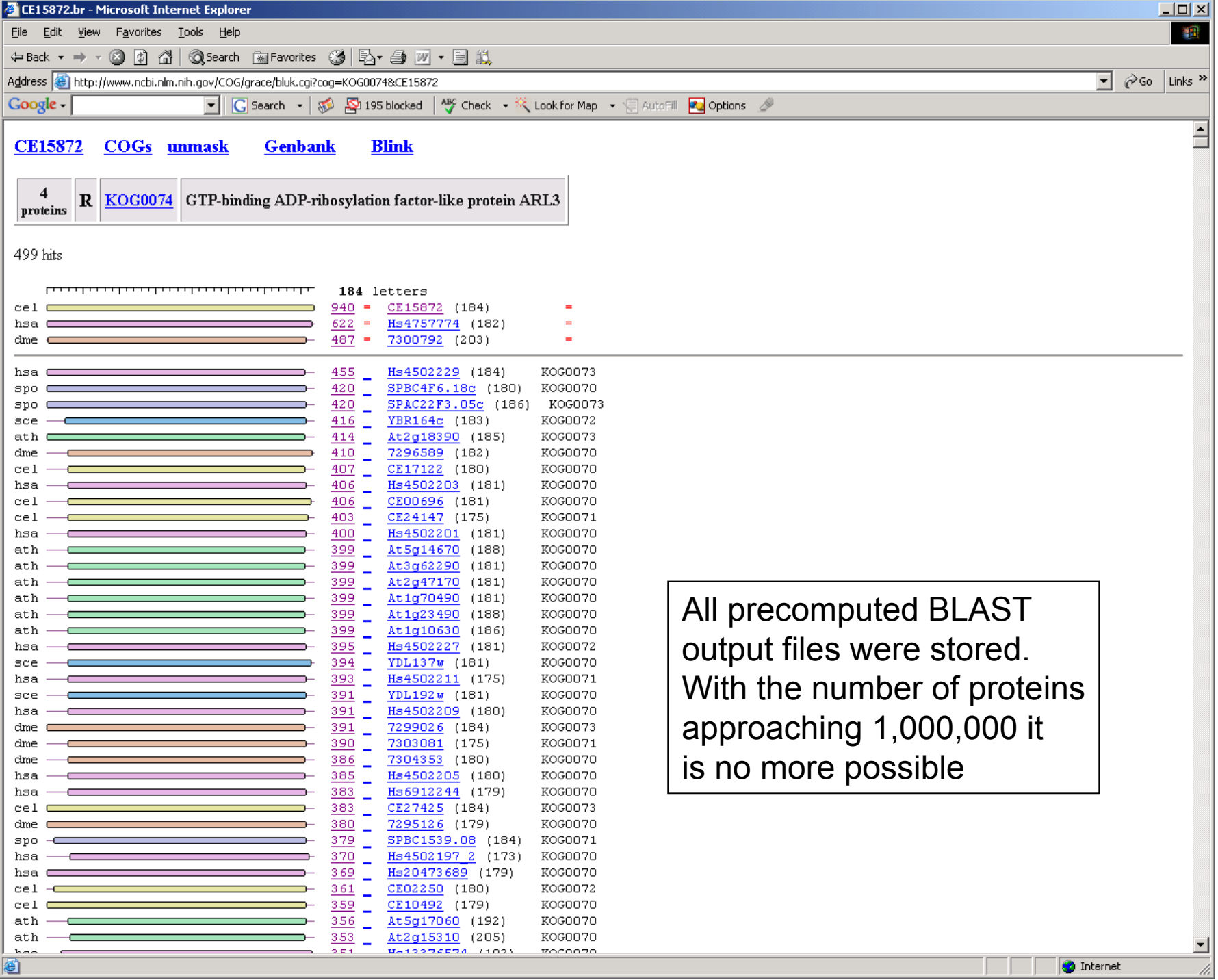
Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, SverdlovAV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003 The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003 4(1):41

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2004

A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.Genome Biol. 2004;5(2):R7.

http://www.ncbi.nlm.nih.gov/COG/new/


CE12874.br - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.ncbi.nlm.nih.gov/COG/grace/bluk.cgi?cog=KOG0090&CE12874
CE12874 COGs unmask Genbank Blink
7 proteins
U
KOG0090
Signal recognition particle receptor, beta subunit (small G protein superfamily)
120 hits
240 letters
cel 1085 = CE12874 (240) =
hsa 237 = Hs14917113 (271) =
dme 231 = 7295050_2 (221) =
ath 197 = At5g05670 (260) =
ath 184 = At2g18770 (260) =
sce 126 = YKL154w (244) =
cel 125 _ CE15872 (184) KOG0074
sce 122 _ YOR094w (183) KOG0071
spo 118 = SPAC23H4.07c (227) =
hsa 114 _ Hs4757774 (182) KOG0074
hsa 109 _ Hs20473689 (179) KOG0070
sce 104 _ YBR164c (183) KOG0072
hsa 104 _ Hs4502203 (181) KOG0070
dme 102 _ 7296589 (182) KOG0070
hsa 101 _ Hs6912244 (179) KOG0070
cel 101 _ CE00696 (181) KOG0070
ath 101 _ At3g22950 (183) KOG0070
ath 101 _ At2g15310 (205) KOG0070
hsa 101 _ Hs4502197_2 (173) KOG0070
hsa 101 _ Hs15208643_2 (145) KOG0070
hsa 101 _ Hs15208641_2 (168) KOG0070
hsa 100 _ Hs5031603 (200) KOG0070
hsa 100 _ Hs4502227 (181) KOG0072
hsa 99 _ Hs4502201 (181) KOG0070
cel 97 _ CE17122 (180) KOG0070
sce 95 _ YDL192w (181) KOG0070
spo 94 _ SPBC1539.08 (184) KOG0071
sce 93 _ YDL137w (181) KOG0070
hsa 93 _ Hs9910542 (198) KOG0077
spo 92 _ SPBC31F10.06c (190) KOG0077
hsa 92 _ Hs4502229 (184) KOG0073
spo 91 _ SPBC4F6.18c (180) KOG0070
hsa 91 _ Hs20472330 (234) KOG0070
cel 91 _ CE24147 (175) KOG0071
ath 91 _ At5g14670 (188) KOG0070
ath 91 _ At3g62290 (181) KOG0070
ath 91 _ At2g47170 (181) KOG0070
ath 91 _ At1g70490 (181) KOG0070
ath 91 _ At1g23490 (188) KOG0070
Public KOG db has been edited extensively:
BLAST outputs for every protein in clusters have been inspected by curators in search of errors
Internet



All precomputed BLAST output files were stored. With the number of proteins approaching 1,000,000 it is no more possible

PUBLIC KOG database contains three types of clusters

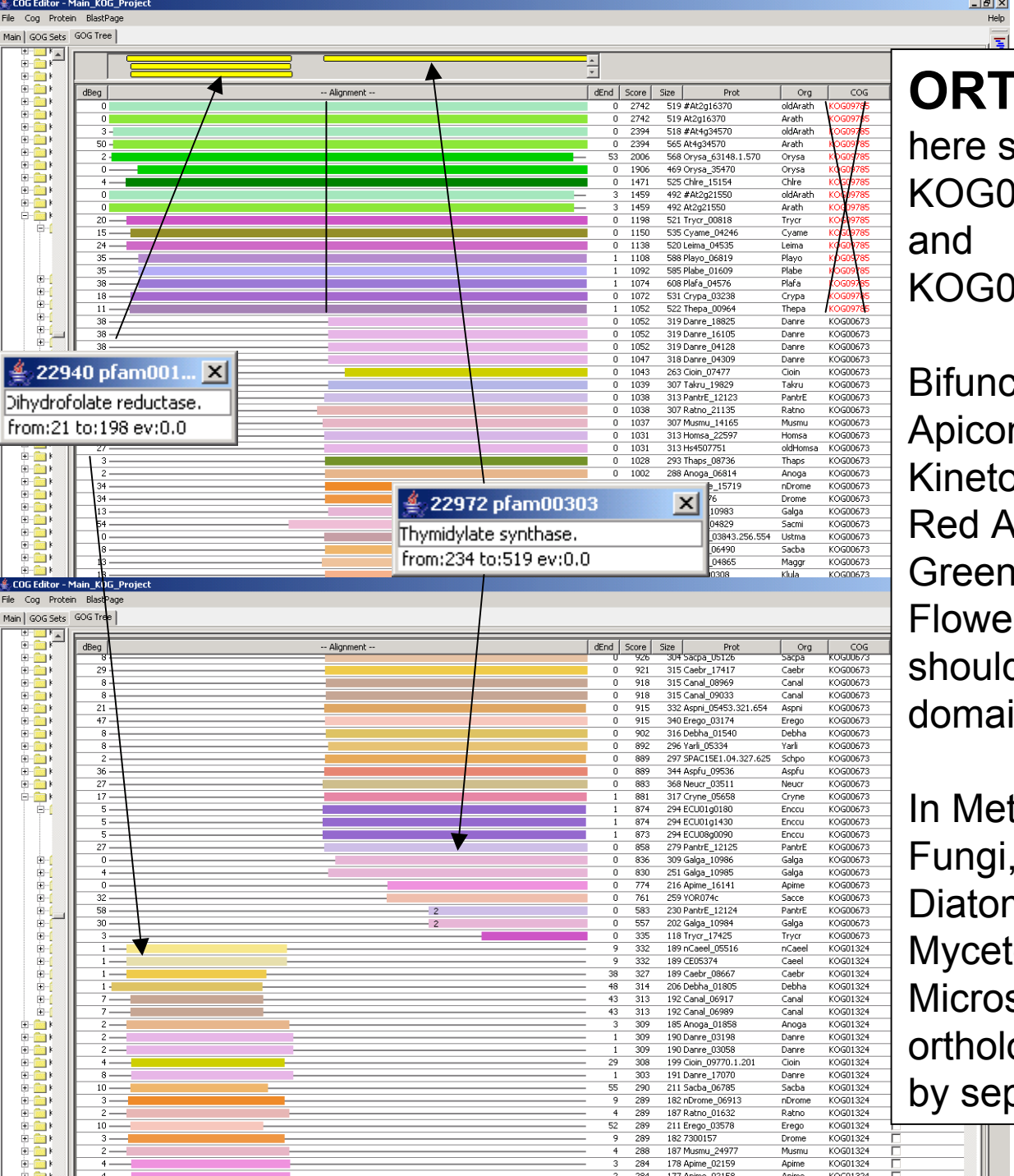
	sequences	in KOGs/LSEs		KOG	TWOGs	LSEs	total clusters
H.sapiens	38,638	26,324	68%	K 4,597	T 602	L 1,373	=6,572
D.melanogaster	13,703	10,517	77%	K 4,351	T 491	L 303	=5,145
C.elegans	20,751	17,101	82%	K 4,235	T 225	L 845	=5,305
S.pombe	5,035	4,233	84%	K 2,668	T 296	L 40	=3,004
S.cerevisiae	6,387	4,841	76%	K 2,762	T 293	L 132	=3,187
E.cuniculi	2,000	1,443	72%	K 1,073	T 6	L 29	=1,108
A.thaliana	26,406	24,154	91%	K 3,285	T 129	L 1,458	=3,872
Total	112,920	88,613	78%	4,952 KOGs	1,021 TWOGs	4,180 LSEs	

Interface for KOG editing

- Changes are implemented on the fly
- No precomputed BLASTP results for all proteins, but only for a small selection AND only temporarily
- Real time BLASTP
- Precomputed RPS BLAST results
- Precomputed HMM search results
- Easy ways to switch to multiple alignments

KOG db deals not with orthologs but with

ORTHOLOGY DOMAINS



ORTHOLOGY DOMAINS

here should form two KOGs:
KOG01324 Dihydrofolate reductase
and
KOG00673 Thymidylate synthase

Bifunctional enzymes occuring in
Apicomplexans,
Kinetoplastids,
Red Algae,
Green Algae, and
Flowering Plants
should be split into orthology domains along the vertical line

In Metazoa,
Fungi,
Diatoms,
Mycetozoa, and
Microsporidia these orthology domains are represented by separate polypeptides

Current status of KOG db: 50 Eukaryotic species; 46 species KOGnitorized, 1 updated;
712,427 protein sequences; two versions are present for 5 proteomes (+ 123,981 sequences)
21,816 clusters: 18,628 KOGs, 3,188 LSEs+mixed clusters

species name		Taxa	sequences	In KOGs/LSEs	%	KOGs	sequence data published	source
Chlamydomonas reinhardtii		Chlorophyta (green algae)	19,922	10,194	51%	4,653	No	JGI
Cyanidioschyzon merolae		Rhodophyta (red algae)	5,040	3,769	74%	2,735		JGI
Dictyostelium discoideum		Mycetozoa	13,677	9,162	67%	4,573		GB
Encephalitozoon cuniculi	OLD	Microsporidia	2,003	1,459	73%	1,132		GB
Entamoeba histolytica		Entamoebidae	10,088	7,728	77%	2,930		TIGR
Giardia lamblia		Diplomonadida group	6,573	2,105	32%	1,242	No	GB
Leishmania major		Euglenozoa	8,071	7,558	93%	4,502		TIGR
Trypanosoma brucei		Euglenozoa	8,291					GB
Trypanosoma cruzi		Euglenozoa	19,666	16,844	86%	5,023		GB
Thalassiosira pseudonana		Bacillariophyta (diatoms)	11,406	9,311	81%	4,033		JGI
Cryptosporidium parvum		Apicomplexa	3,410	2,212	66%	1,780		GB
Plasmodium berghei		Apicomplexa	11,782	7,004	59%	3,461	No	GB
Plasmodium falciparum		Apicomplexa	5,279	3,365	64%	3,403		GB
Plasmodium yoelii		Apicomplexa	7,868					GB
Theileria parva		Apicomplexa	4,080					GB
Arabidopsis thaliana	NEW	Viridiplantae	26,759	25,781	96%	6,953		GB
Arabidopsis thaliana	OLD	Viridiplantae	27,138	25,410	94%	6,925		GB
Oryza sativa		Viridiplantae	57,272	38,260	67%	6,701		TIGR
Aspergillus fumigatus		Fungi; Ascomycota; Pezizomycotina	10,029	8,959	89%	5,232	No	GB
Aspergillus nidulans		Fungi; Ascomycota; Pezizomycotina	10,317	9,629	93%	5,283	No	GB
Gibberella zeae		Fungi; Ascomycota; Pezizomycotina	12,221	11,153	91%	5,910	No	GB
Magnaporthe grisea		Fungi; Ascomycota; Pezizomycotina	11,381	9,043	79%	5,397		GB
Neurospora crassa		Fungi; Ascomycota; Pezizomycotina	10,478	8,206	78%	5,559		GB
Candida albicans		Fungi; Ascomycota; Saccharomycotina	14,274	10,952	77%	4,094		GB
Candida glabrata		Fungi; Ascomycota; Saccharomycotina	5,216	4,992	96%	3,790	No	GB
Debaryomyces hansenii		Fungi; Ascomycota; Saccharomycotina	6,350	5,700	90%	4,048		GB
Eremothecium gossypii		Fungi; Ascomycota; Saccharomycotina	4,757	4,619	97%	3,810	No	GB
Kluyveromyces lactis		Fungi; Ascomycota; Saccharomycotina	5,366	5,058	94%	3,980		GB
Kluyveromyces waltii		Fungi; Ascomycota; Saccharomycotina	5,240	5,055	96%	3,919	No	MIT
Yarrowia lipolytica		Fungi; Ascomycota; Saccharomycotina	6,596	5,728	87%	4,018		GB
Saccharomyces bayanus		Fungi; Ascomycota; Saccharomycotina	12,040	8,488	70%	5,879		MIT
Saccharomyces cerevisiae	OLD	Fungi; Ascomycota; Saccharomycotina	6,387	6,146	96%	4,427		GB
Saccharomyces mikatae		Fungi; Ascomycota; Saccharomycotina	10,350	8,481	82%	5,828		MIT
Saccharomyces paradoxus		Fungi; Ascomycota; Saccharomycotina	10,601	8,375	79%	6,005		MIT
Schizosaccharomyces pombe	OLD	Fungi; Ascomycota; Schizosaccharomycetes	5,064	4,461	88%	3,360		GB
Cryptococcus neoformans		Fungi; Basidiomycota; Hymenomycetes	6,640	5,471	82%	3,702	No	GB
Ustilago maydis		Fungi; Basidiomycota; Ustilaginomycetes	6,685	5,221	78%	3,779	No	GB
Anopheles gambiae		Metazoa; Insecta; Diptera; Nematocera	13,895	12,405	89%	5,946		GB
Apis mellifera		Metazoa; Insecta; Hymenoptera	16,997	14,785	87%	4,906	No	EMBL
Drosophila melanogaster	NEW	Metazoa; Insecta; Diptera; Brachycera	18,873					GB
Drosophila melanogaster	OLD	Metazoa; Insecta; Diptera; Brachycera	13,936	11,552	83%	6,241		GB
Caenorhabditis briggsae		Metazoa; Nematoda	19,523	17,963	92%	6,938		GB
Caenorhabditis elegans	NEW	Metazoa; Nematoda	21,193					GB
Caenorhabditis elegans	OLD	Metazoa; Nematoda	21,246	20,461	96%	7,175		GB
Ciona intestinalis		Metazoa; Chordata; Ascidiacea	16,139	12,101	75%	5,684		JGI
Danio rerio		Vertebrata; Teleostomi	19,893	19,676	98%	4,426		GB
Gallus gallus		Vertebrata; Aves	28,751	26,613	93%	6,894	No	EMBL
Homo sapiens	NEW	Vertebrata; Mammalia	27,524					GB
Homo sapiens	OLD	Vertebrata; Mammalia	39,436	29,835	77%	9,304		GB
Mus musculus		Vertebrata; Mammalia	26,464	24,784	94%	9,284		GB
Pan troglodytes	GB	Vertebrata; Mammalia	22,225					GB
Pan troglodytes	EMBL	Vertebrata; Mammalia	38,845	36,984	95%	8,923		EMBL
Rattus norvegicus		Vertebrata; Mammalia	21,531					GB
Takifugu rubripes		Vertebrata; Teleostomi	33,357	31,879	96%	6,909		GB
Tetraodon nigroviridis		Vertebrata; Teleostomi	28,262	25,520	90%	7,033	No	GB

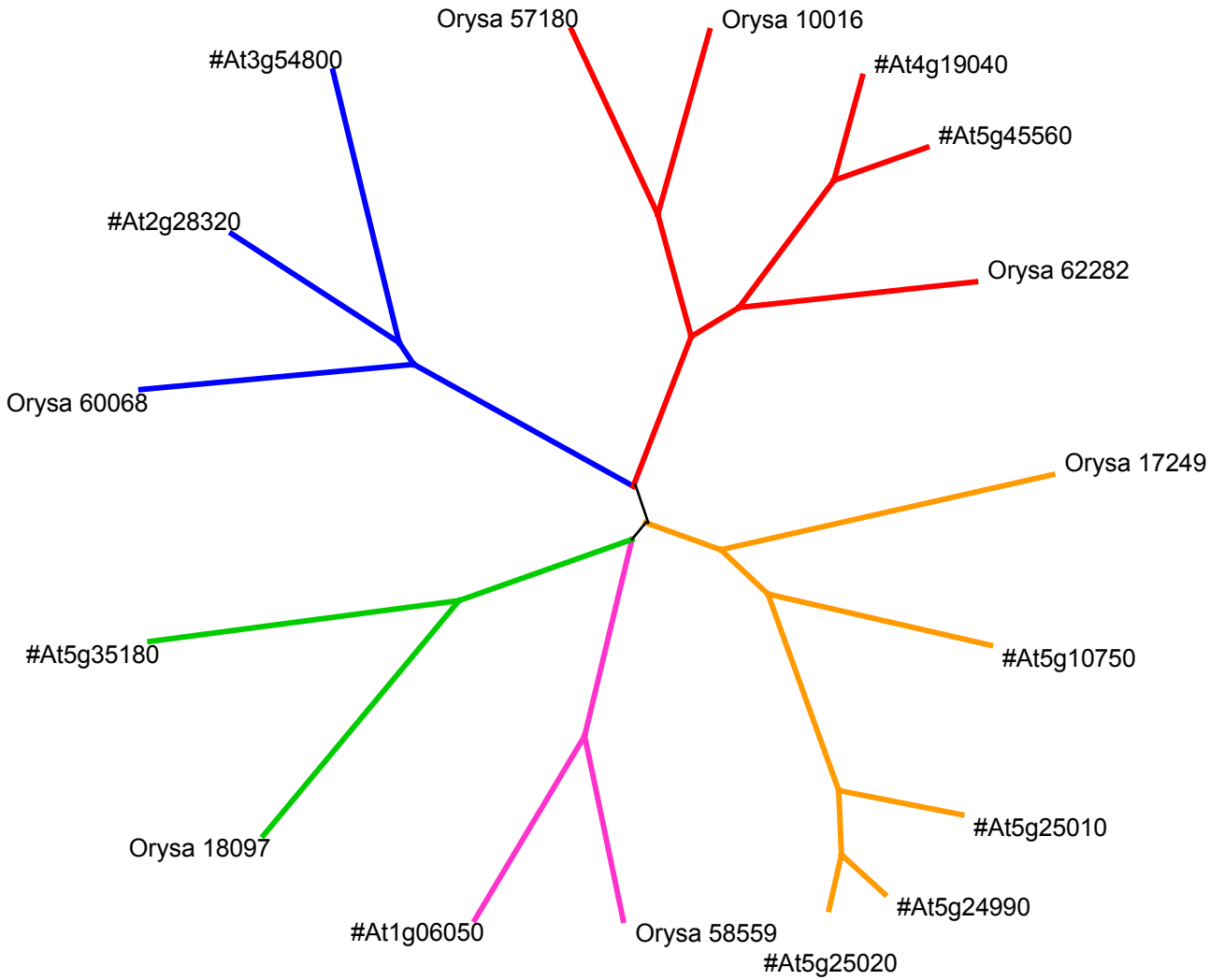
These are the results of ongoing research KOG project. These results should be considered preliminar

Absolute KOGs are not possible

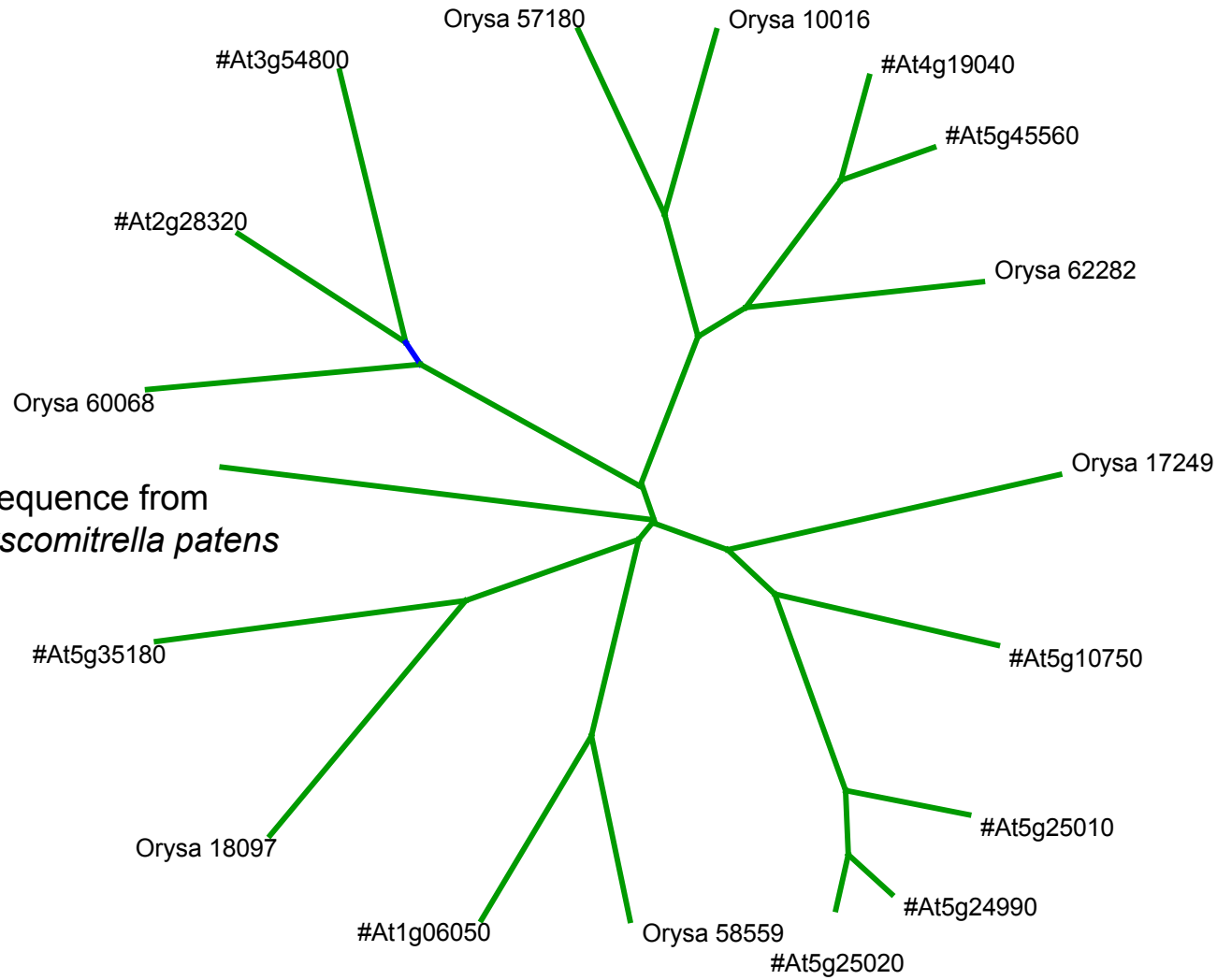
KOGs are relative to the set of proteomes used in their construction

This sad truth constitutes the major difficulty in creating algorithms that would allow to add proteomes to KOGs automatically

5 related KOGs

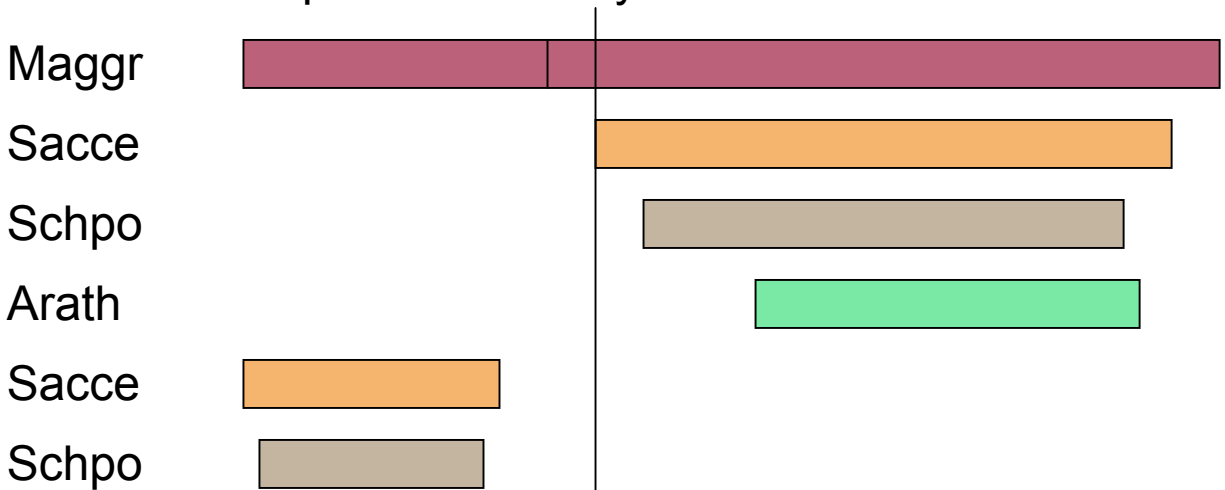


5 KOGs → 1 KOG rooted by one sequence from a new genome

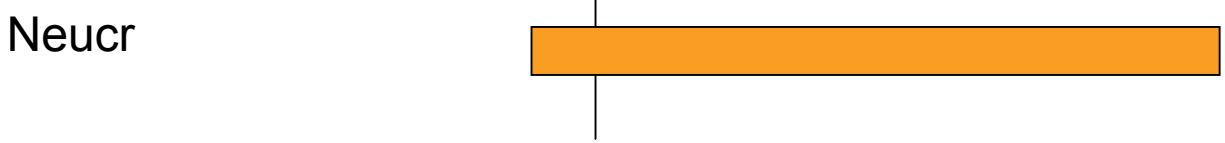


Even dividing sequences into orthology domains depends on the proteome set

A chimeric sequence from *M.grisea* was divided between two KOGs when its proteome was added to KOGs with just two fungal species. It was not possible to find division point accurately



Subsequently, *N.crassa* proteome is added. KOGnitor suggests to divide non-chimeric protein according to BLAST hits to divided Maggr sequence

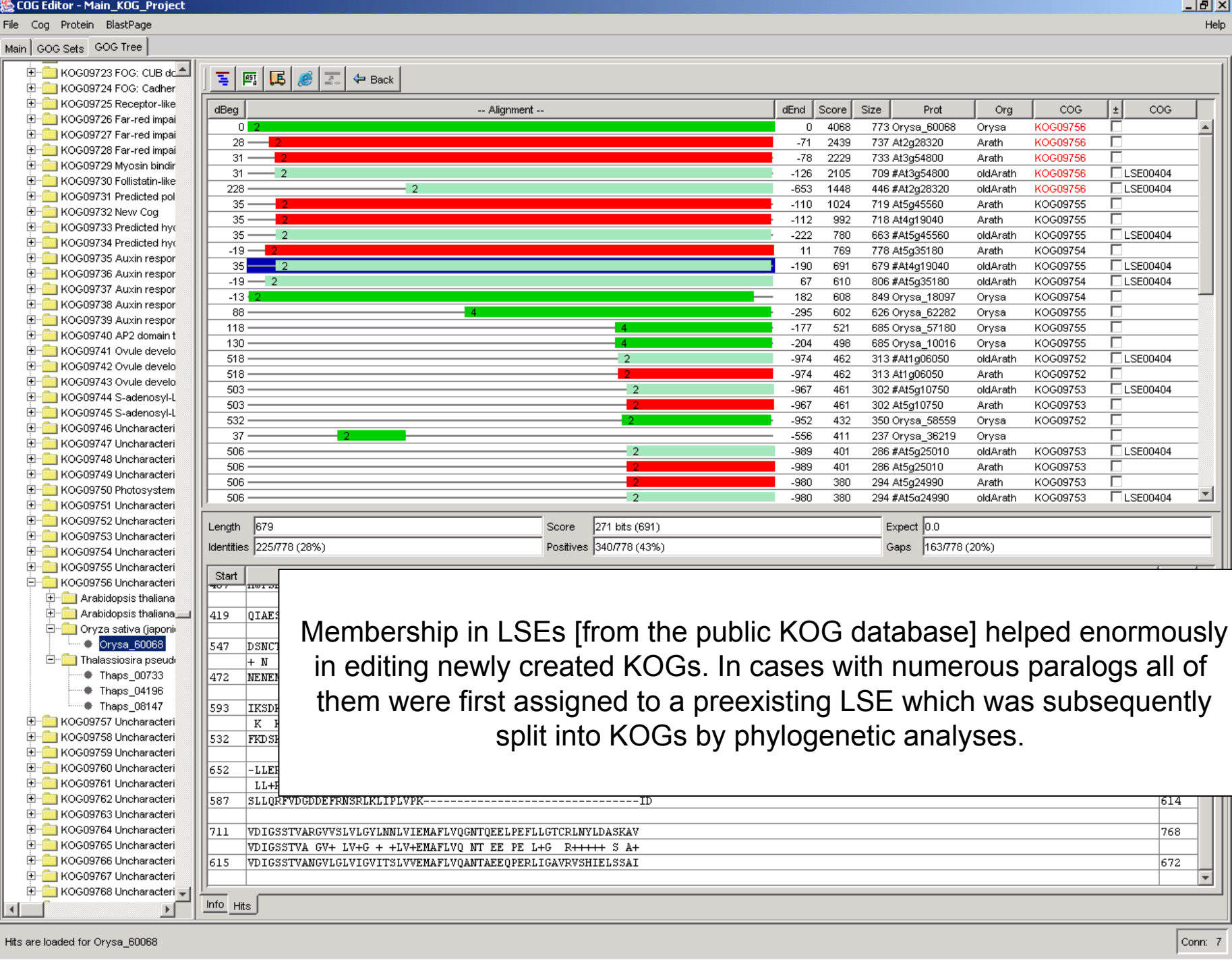


Due to extensive manual curation the public KOG db project contains substantial added value if compared with results of automatic clustering using BeTs.

It would be unwise and even silly to discard these results.

One of important features of public KOGs were LSEs.

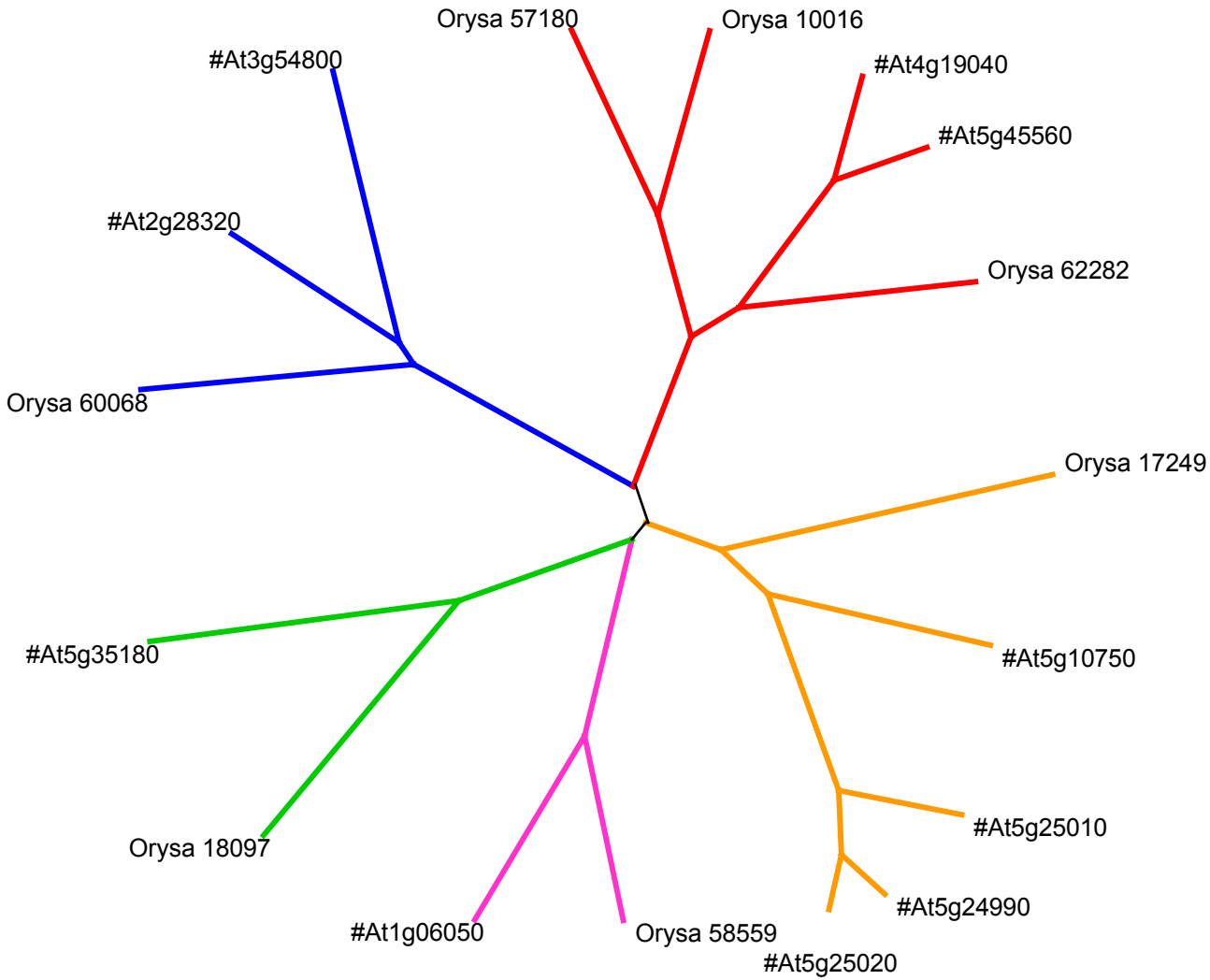
Lineage Specific Expansions (LSE)



Membership in LSEs [from the public KOG database] helped enormously in editing newly created KOGs. In cases with numerous paralogs all of them were first assigned to a preexisting LSE which was subsequently split into KOGs by phylogenetic analyses.

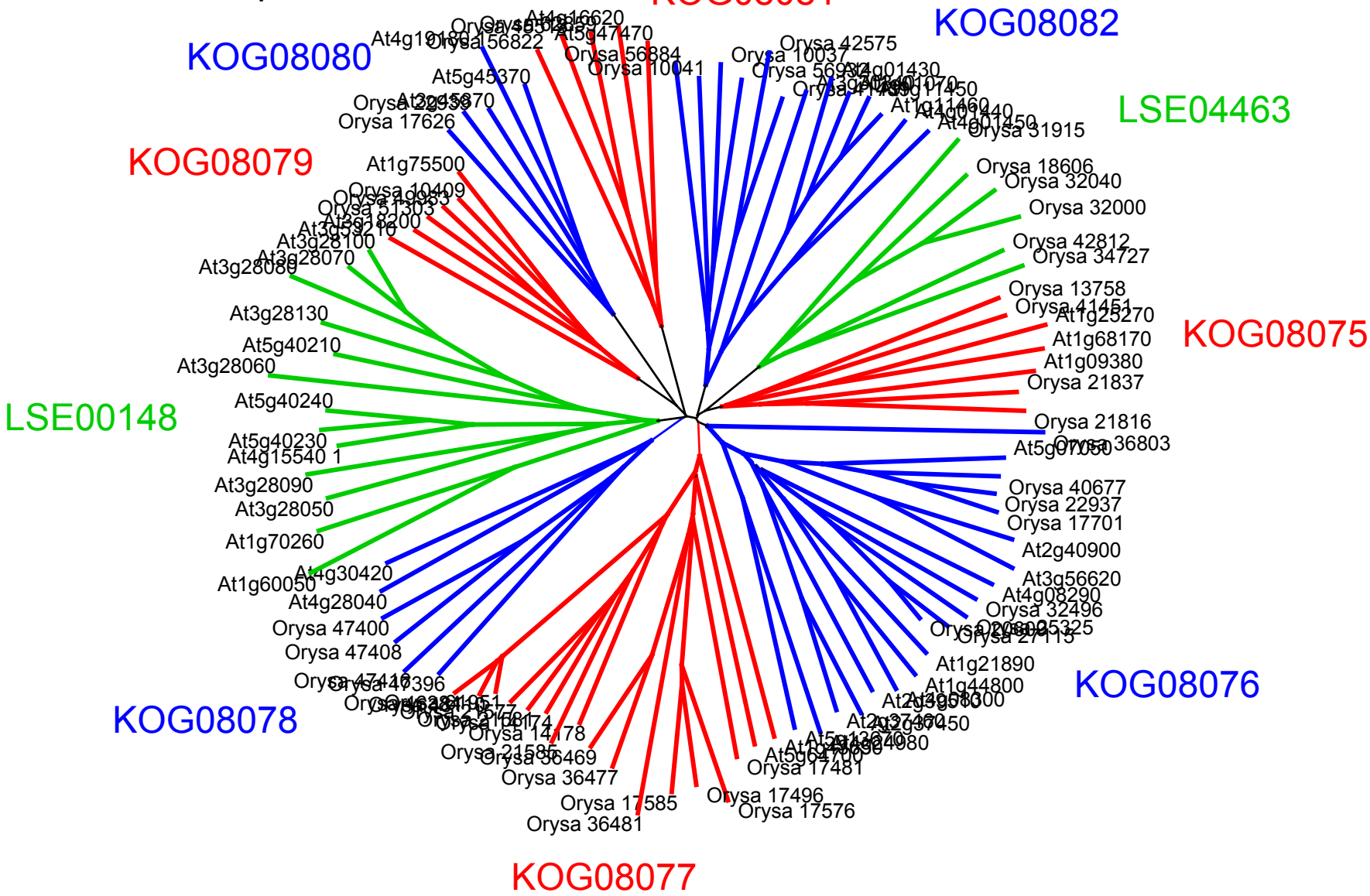
LSE00404 → 5 KOGs

Uncharacterized proteins



LSE00148 → 8 KOGs, 2 LSEs

Nodulin-like proteins



With the number of protein sequences approaching 1,000,000 BLAST all-against-all does not seem to be possible

New KOGnitor

Proteomes, one at a time are added to KOGs with KOGnitor program (Y.Wolf, unpublished) which uses BLASTP computed for every protein in the proteome against a selected set of KOGnitorized proteomes.

KOGnitor creates a list of sequences that potentially should form new KOGs

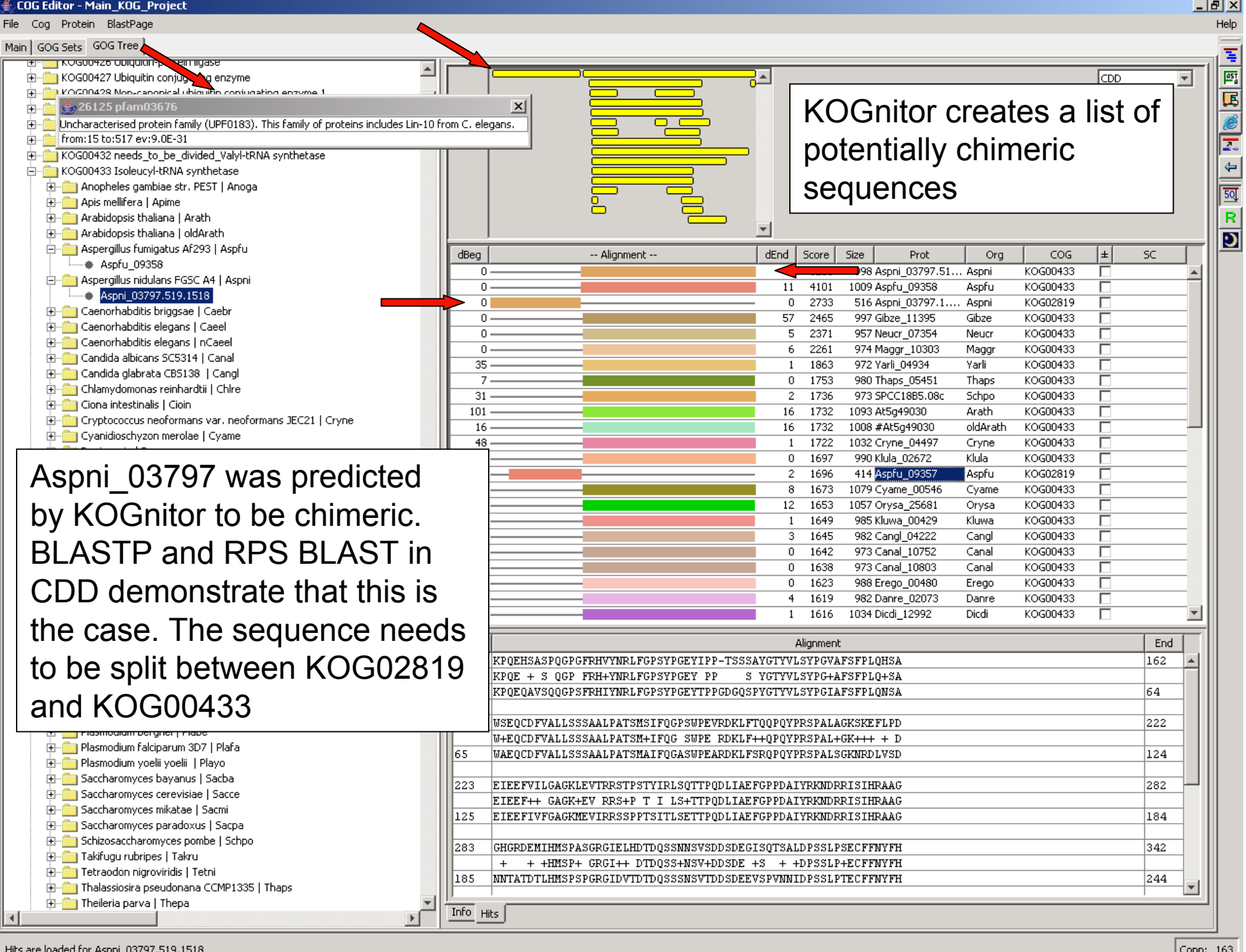
and

a list of sequences that are potentially chimeric and have to be split into separate orthology domains

Unfortunately, due to unidirectional hits and necessarily relaxed BLAST score cutoffs KOGnitor produces high rate of false positives. We checked completely one proteome, *Cryptosporidium parvum* and found 14% of false positives. This rate is unacceptable and we developed a plan to fix it by simultaneously applying two independent methods of sequence analysis – KOGnitor and HMM search

A plan to create curated set of KOGs semi-automatically by constructing expanding KOG_HMM profiles:

1. Build HMM profiles for previously curated KOGs
2. Add proteomes to KOGs, one at a time with KOGnitor
3. Confirm suggested new KOGs and divisions of potentially chimeric sequences into separate orthology domains by manual curation
4. Run HMM search using KOG_HMM profiles in KOG db
5. Compare results of KOGnitor and HMM search and manually check ONLY conflicting sequences. The great majority of sequences would not require checking.
6. Whenever necessary, add the new sequences to HMM profiles such that all legitimate members of KOGs are recognized by updated HMM profiles
7. Such an approach would result not only in curated set of KOGs, but also in creating a library of KOG_HMM which presumably will be used more and more with accumulation of new genomes. We anticipate that BLAST searches will be largely replaced in the future by profile seaches when new proteomes are KOGnitorized.


COG Editor - Main_KOG_Project
File Cog Protein BlastPage Help
Main GOG Sets GOG Tree
26125 pfam03676
Uncharacterised protein family (UPF0183). This family of proteins includes Lin-10 from C. elegans.
from:15 to:517 ev:9.0E-31
KOG00427 Ubiquitin conjug ng enzyme
KOG00432 needs_to_be_divided_Valyl-tRNA synthetase
KOG00433 Isoleucyl-tRNA synthetase
Anopheles gambiae str. PEST | Anoga
Apis mellifera | Apime
Arabidopsis thaliana | Arath
Arabidopsis thaliana | oldArath
Aspergillus fumigatus Af293 | Aspfu
Aspfu_09358
Aspergillus nidulans FGSC A4 | Aspni
Aspni_03797.519.1518
Caenorhabditis briggsae | Caebr
Caenorhabditis elegans | Caeel
Caenorhabditis elegans | nCaeel
Candida albicans SC5314 | Canal
Candida glabrata CBS138 | Cangl
Chlamydomonas reinhardtii | Chlre
Ciona intestinalis | Cioin
Cryptococcus neoformans var. neoformans JEC21 | Cryne
Cyanidioschyzon merolae | Cyame
Plasmodium falciparum 3D7 | Plafa
Plasmodium yoelii yoelii | Playo
Saccharomyces bayanus | Sacba
Saccharomyces cerevisiae | Sacce
Saccharomyces mikatae | Sacmi
Saccharomyces paradoxus | Sacpa
Schizosaccharomyces pombe | Schpo
Takifugu rubripes | Takru
Tetraodon nigroviridis | Tetni
Thalassiosira pseudonana CCMP1335 | Thaps
Theileria parva | Thepa
CDD
KOGnitor creates a list of potentially chimeric sequences
dBeg -- Alignment -- dEnd Score Size Prot Org COG ± SC
0 98 Aspni_03797.51... Aspni KOG00433
0 11 4101 1009 Aspfu_09358 Aspfu KOG00433
0 0 2733 516 Aspni_03797.1.... Aspni KOG02819
0 57 2465 997 Gibze_11395 Gibze KOG00433
0 5 2371 957 Neucr_07354 Neucr KOG00433
0 6 2261 974 Maggr_10303 Maggr KOG00433
35 1 1863 972 Yarli_04934 Yarli KOG00433
7 0 1753 980 Thaps_05451 Thaps KOG00433
31 2 1736 973 SPCC18B5.08c Schpo KOG00433
101 16 1732 1093 At5g49030 Arath KOG00433
16 16 1732 1008 #At5g49030 oldArath KOG00433
48 1 1722 1032 Cryne_04497 Cryne KOG00433
0 1697 990 Klula_02672 Klula KOG00433
2 1696 414 Aspfu_09357 Aspfu KOG02819
8 1673 1079 Cyame_00546 Cyame KOG00433
12 1653 1057 Orysa_25681 Orysa KOG00433
1 1649 985 Kluwa_00429 Kluwa KOG00433
3 1645 982 Cangl_04222 Cangl KOG00433
0 1642 973 Canal_10752 Canal KOG00433
0 1638 973 Canal_10803 Canal KOG00433
0 1623 988 Erego_00480 Erego KOG00433
4 1619 982 Danre_02073 Danre KOG00433
1 1616 1034 Dicdi_12992 Dicdi KOG00433
Aspni_03797 was predicted by KOGnitor to be chimeric. BLASTP and RPS BLAST in CDD demonstrate that this is the case. The sequence needs to be split between KOG02819 and KOG00433
Alignment End
KPQEHSASPQGPGFRHVYNRLFGPSYPGEYIPP-TSSSAYGTYVLSYPGVAFSFPLQHSA 162
KPQE + S QGP FRH+YNRLFGPSYPGEY PP S YGTYVLSYPG+AFSFPLQ+SA
KPQEQAVSQQGPSFRHIYNRLFGPSYPGEYTPPGDGQSPYGTYVLSYPGIAFSFPLQNSA 64
WSEQCDFVALLSSSAALPATSMSIFQGPSWPEVRDKLFTQQPQYPRSPALAGKSKEFLPD 222
W+EQCDFVALLSSSAALPATSM+IFQG SWPE RDKLF++QPQYPRSPAL+GK+++ + D
65 WAEQCDFVALLSSSAALPATSMAIFQGASWPEARDKLFSRQPQYPRSPALSGKNRDLVSD 124
223 EIEEFVILGAGKLEVTRRSTPSTYIRLSQTTPQDLIAEFGPPDAIYRKNDRRISIHRAAG 282
EIEEF++ GAGK+EV RRS+P T I LS+TTPQDLIAEFGPPDAIYRKNDRRISIHRAAG
125 EIEEFIVFGAGKMEVIRRSSPPTSITLSETTPQDLIAEFGPPDAIYRKNDRRISIHRAAG 184
283 GHGRDEMIHMSPASGRGIELHDTDQSSNNSVSDDSDEGISQTSALDPSSLPSECFFNYFH 342
+ + +HMSP+ GRGI++ DTDQSS+NSV+DDSDE +S + +DPSSLP+ECFFNYFH
185 NNTATDTLHMSPSPGRGIDVTDTDQSSSNSVTDDSDEEVSPVNNIDPSSLPTECFFNYFH 244
Info Hits
Hits are loaded for Aspni_03797.519.1518

KOG db contains comprehensive records about potentially chimeric (misassembled) protein sequences which can be used to improve gene predictions in numerous proteomes ☺.
There are thousands of such sequences ☹

Is there a way to automatically confirm predicted break points?

An independent way to confirm predicted chimeric sequences: TBLASTN in dbEST

RID=1131319008-32328-119676246285.BLASTQ2, - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi

Chimeric Aspni_03797 sequence was used as query in TBLASTN in dbEST

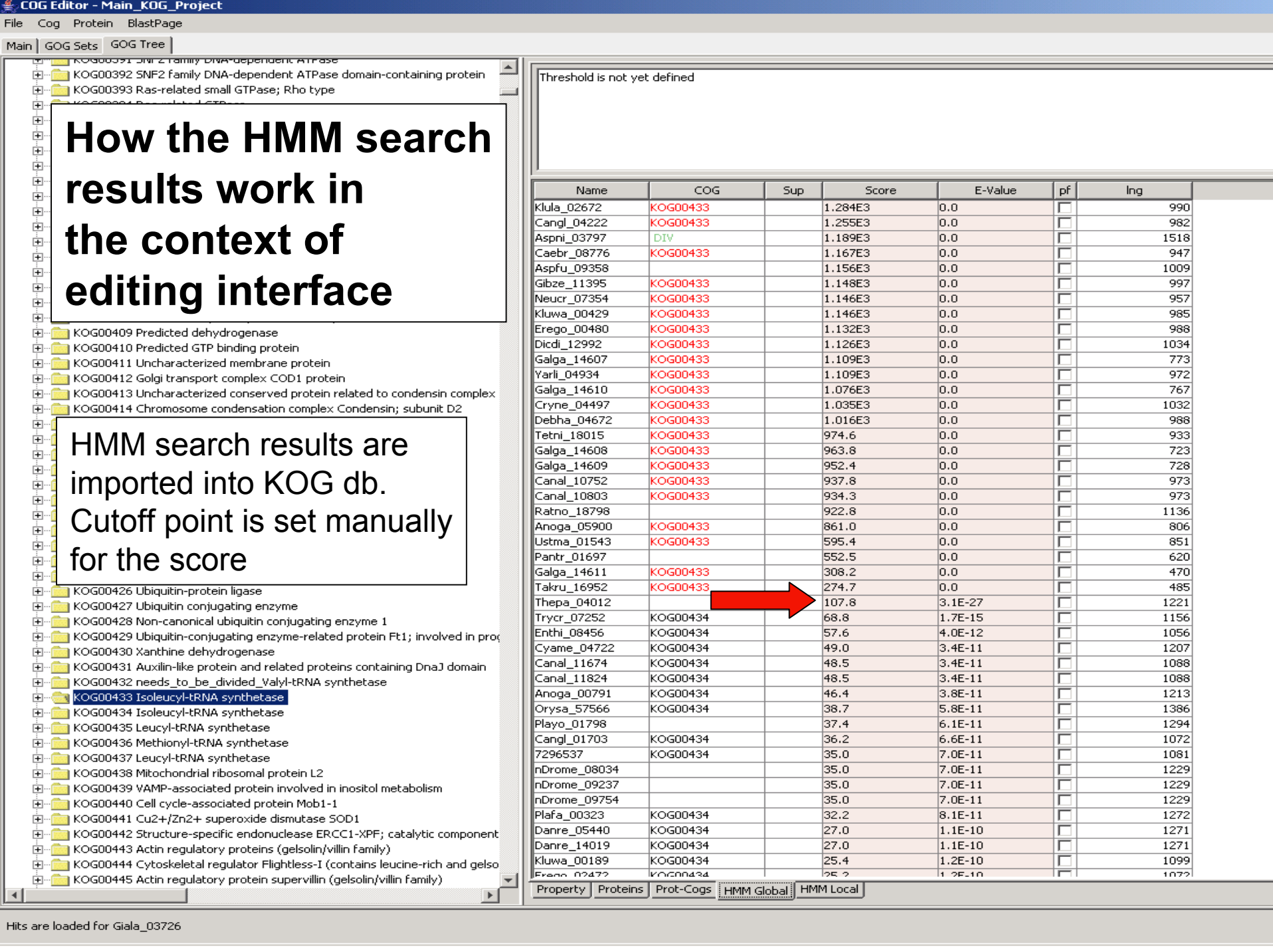
Distribution of 1064 Blast Hits on the Query Sequence

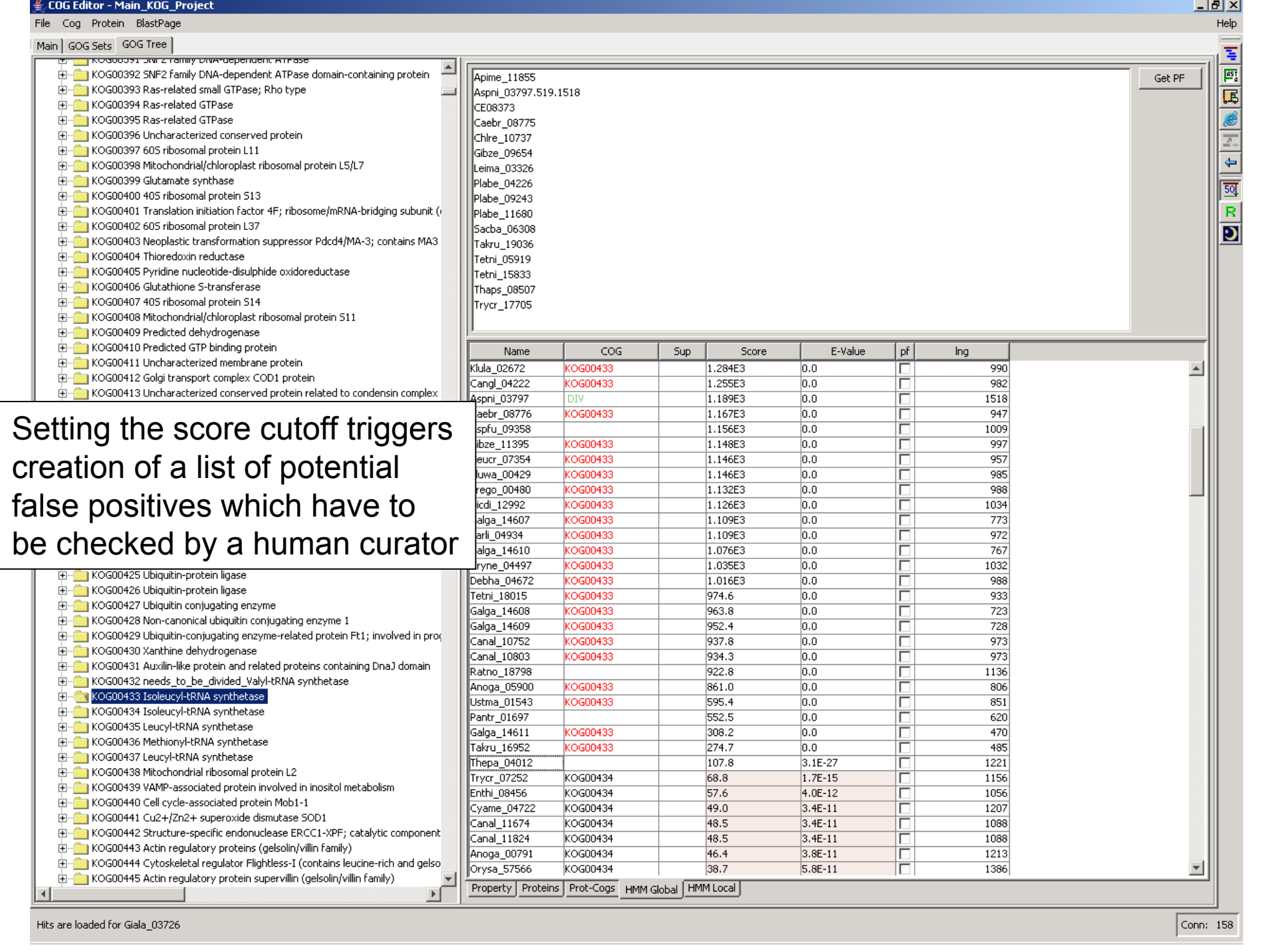
PREDICTED BREAK POINT

Sequences producing significant alignments:		Score (Bits)	E Value
gi|70826686|gb|DR710376.1|	Asn_12133 Aspergillus niger pBlues...	364	4e-98
gi|45923932|gb|CF816054.1|	EST695436 Coccidioides posadasii s...	353	7e-95
gi|48559074|gb|CO028230.1|	EST806614 Coccidioides posadasii s...	335	2e-89
gi|25129738|gb|CA581347.1|	EST001022 Mycelium and yeast cells...	303	9e-80
gi|48516926|gb|CO010037.1|	EST798372 Coccidioides posadasii s...	300	7e-79

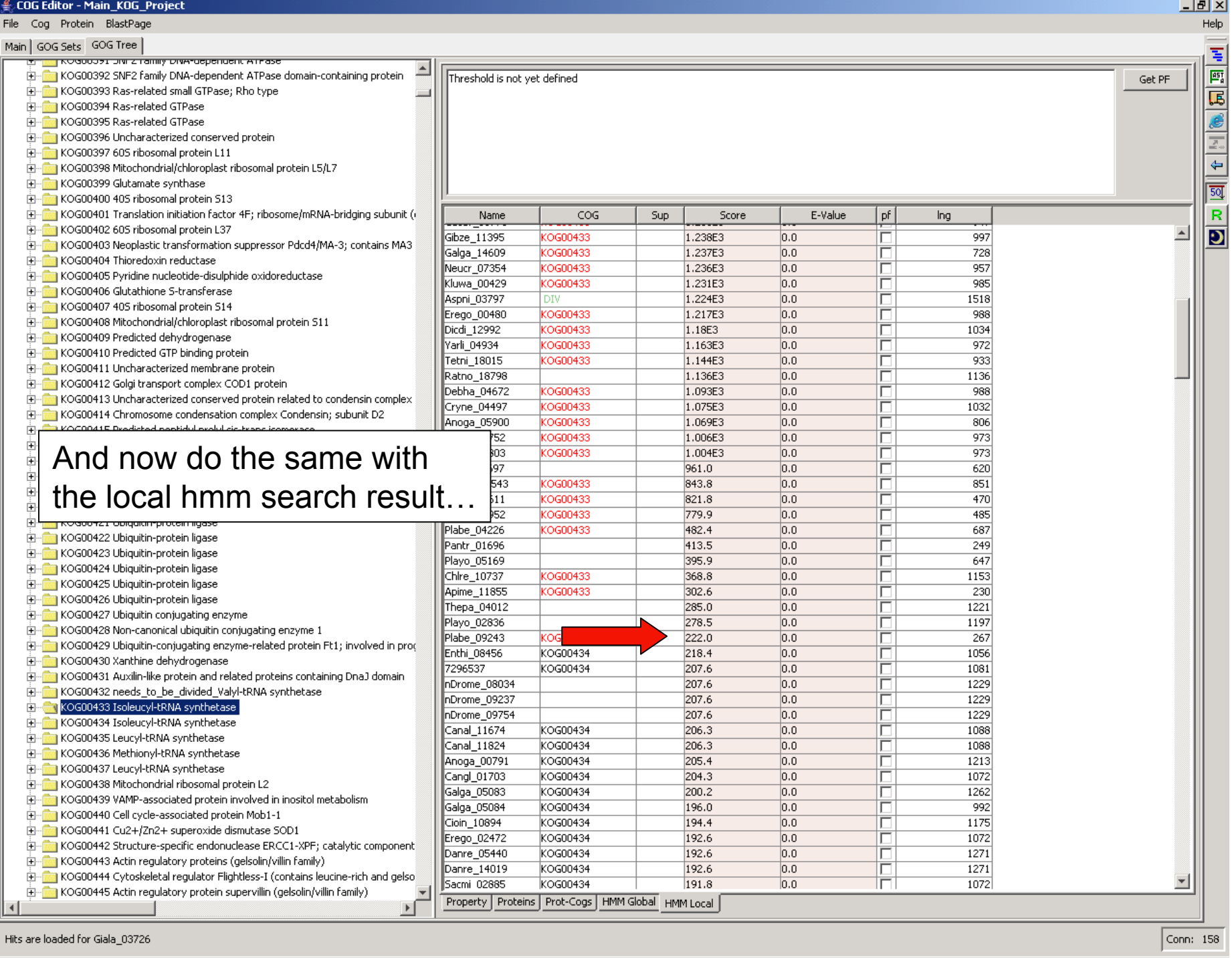
HMM profiles were built for curated versions (containing less than 50, typically a set of 12+ proteomes) of ca. 15,000 KOGs.

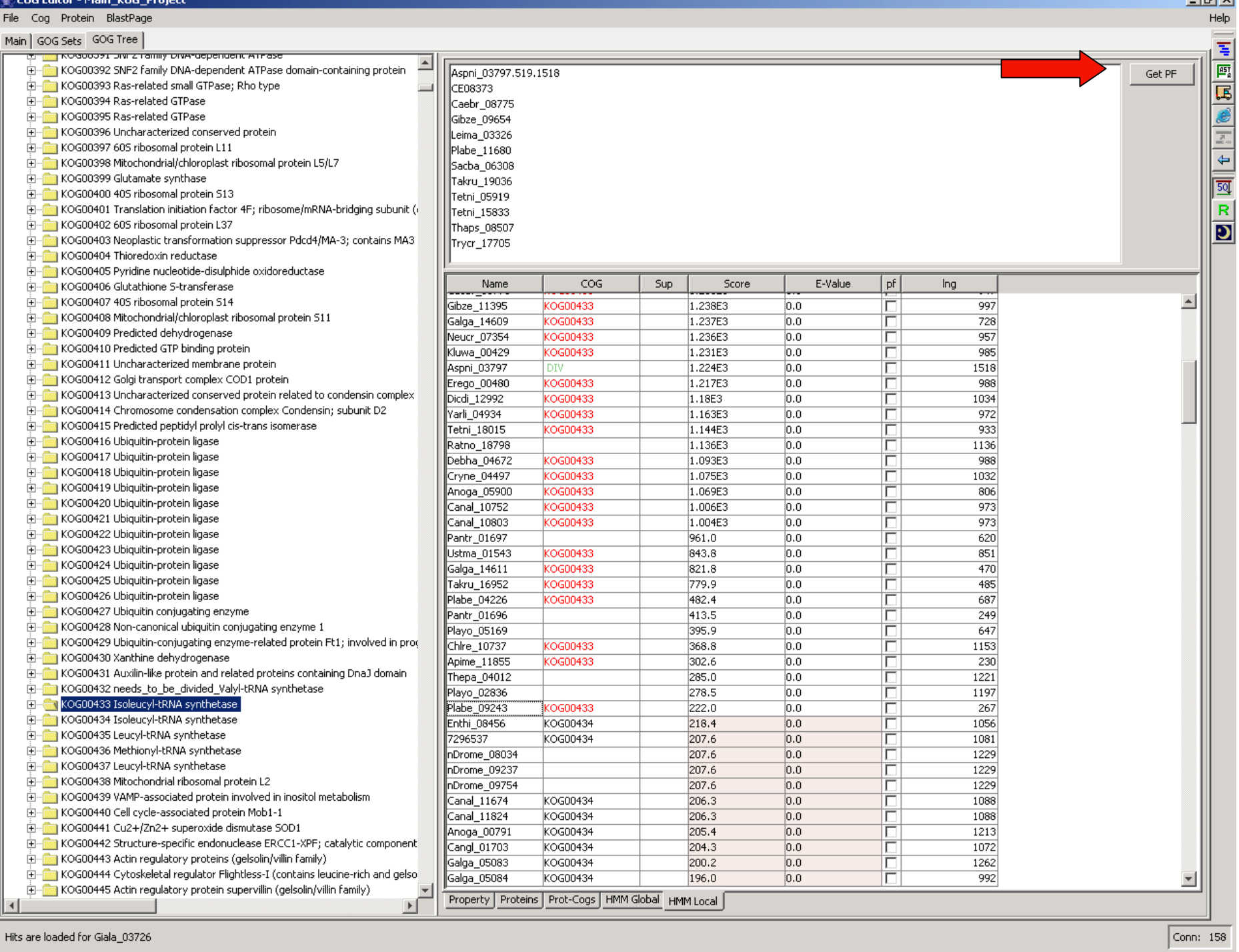
HMM search results in KOG db with HMM profiles as queries provide a new tool to detect conflicts between KOGnitor and HMM search and thus sort out false positives

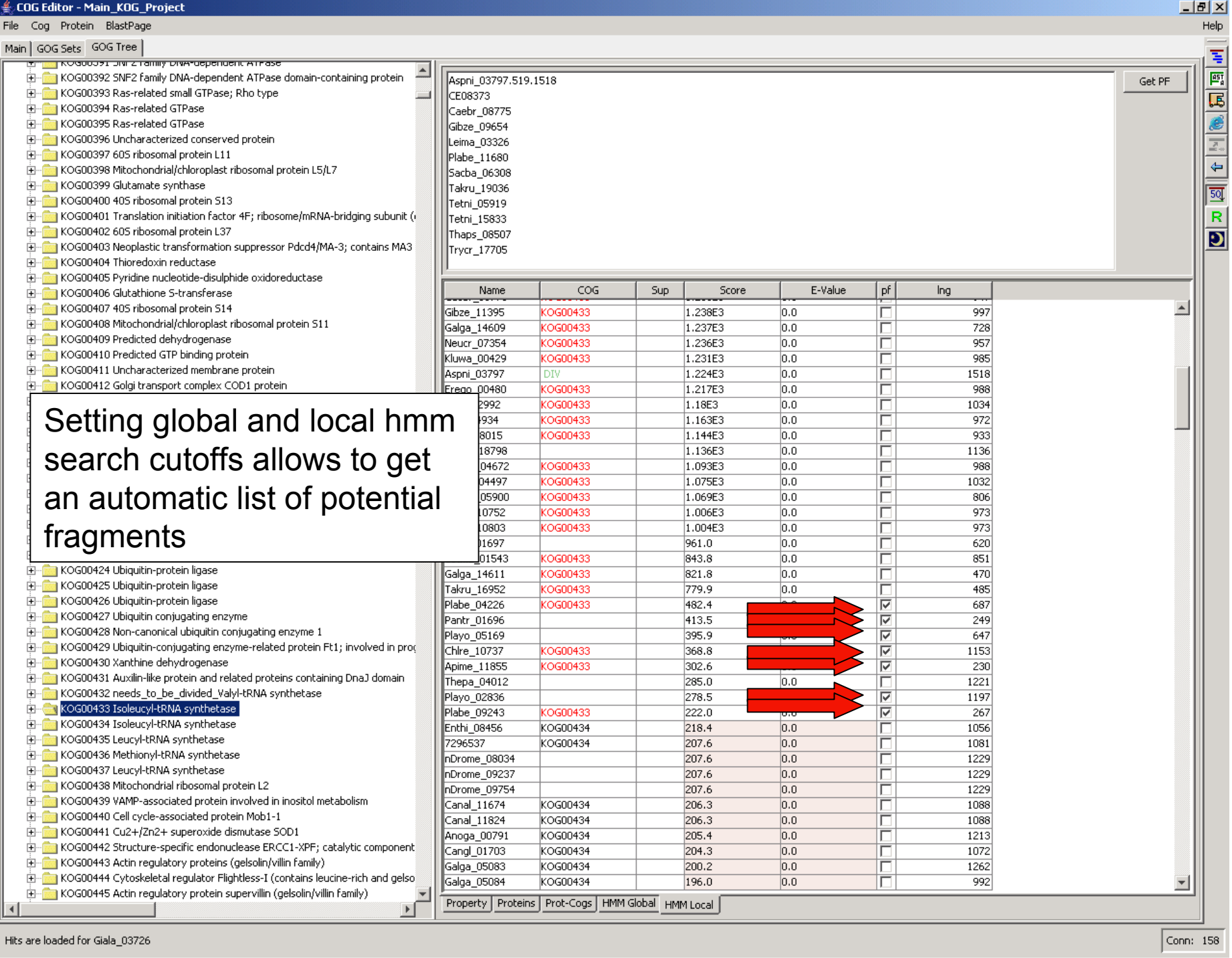

COG Editor - Main_KOG_Project
File Cog Protein BlastPage
Main GOG Sets GOG Tree
KOG00392 SNF2 family DNA-dependent ATPase domain-containing protein
KOG00393 Ras-related small GTPase; Rho type
How the HMM search results work in the context of editing interface
KOG00409 Predicted dehydrogenase
KOG00410 Predicted GTP binding protein
KOG00411 Uncharacterized membrane protein
KOG00412 Golgi transport complex COD1 protein
KOG00413 Uncharacterized conserved protein related to condensin complex
KOG00414 Chromosome condensation complex Condensin; subunit D2
HMM search results are imported into KOG db. Cutoff point is set manually for the score
KOG00426 Ubiquitin-protein ligase
KOG00427 Ubiquitin conjugating enzyme
KOG00428 Non-canonical ubiquitin conjugating enzyme 1
KOG00429 Ubiquitin-conjugating enzyme-related protein Ft1; involved in pro
KOG00430 Xanthine dehydrogenase
KOG00431 Auxilin-like protein and related proteins containing DnaJ domain
KOG00432 needs_to_be_divided_Valyl-tRNA synthetase
KOG00433 Isoleucyl-tRNA synthetase
KOG00434 Isoleucyl-tRNA synthetase
KOG00435 Leucyl-tRNA synthetase
KOG00436 Methionyl-tRNA synthetase
KOG00437 Leucyl-tRNA synthetase
KOG00438 Mitochondrial ribosomal protein L2
KOG00439 VAMP-associated protein involved in inositol metabolism
KOG00440 Cell cycle-associated protein Mob1-1
KOG00441 Cu2+/Zn2+ superoxide dismutase SOD1
KOG00442 Structure-specific endonuclease ERCC1-XPF; catalytic component
KOG00443 Actin regulatory proteins (gelsolin/villin family)
KOG00444 Cytoskeletal regulator Flightless-I (contains leucine-rich and gelso
KOG00445 Actin regulatory protein supervillin (gelsolin/villin family)
Threshold is not yet defined
Name COG Sup Score E-Value pf lng
Klula_02672 KOG00433 1.284E3 0.0 990
Cangl_04222 KOG00433 1.255E3 0.0 982
Aspni_03797 DIV 1.189E3 0.0 1518
Caebr_08776 KOG00433 1.167E3 0.0 947
Aspfu_09358 1.156E3 0.0 1009
Gibze_11395 KOG00433 1.148E3 0.0 997
Neucr_07354 KOG00433 1.146E3 0.0 957
Kluwa_00429 KOG00433 1.146E3 0.0 985
Erego_00480 KOG00433 1.132E3 0.0 988
Dicdi_12992 KOG00433 1.126E3 0.0 1034
Galga_14607 KOG00433 1.109E3 0.0 773
Yarli_04934 KOG00433 1.109E3 0.0 972
Galga_14610 KOG00433 1.076E3 0.0 767
Cryne_04497 KOG00433 1.035E3 0.0 1032
Debha_04672 KOG00433 1.016E3 0.0 988
Tetni_18015 KOG00433 974.6 0.0 933
Galga_14608 KOG00433 963.8 0.0 723
Galga_14609 KOG00433 952.4 0.0 728
Canal_10752 KOG00433 937.8 0.0 973
Canal_10803 KOG00433 934.3 0.0 973
Ratno_18798 922.8 0.0 1136
Anoga_05900 KOG00433 861.0 0.0 806
Ustma_01543 KOG00433 595.4 0.0 851
Pantr_01697 552.5 0.0 620
Galga_14611 KOG00433 308.2 0.0 470
Takru_16952 KOG00433 274.7 0.0 485
Thepa_04012 107.8 3.1E-27 1221
Trycr_07252 KOG00434 68.8 1.7E-15 1156
Enthi_08456 KOG00434 57.6 4.0E-12 1056
Cyame_04722 KOG00434 49.0 3.4E-11 1207
Canal_11674 KOG00434 48.5 3.4E-11 1088
Canal_11824 KOG00434 48.5 3.4E-11 1088
Anoga_00791 KOG00434 46.4 3.8E-11 1213
Orysa_57566 KOG00434 38.7 5.8E-11 1386
Playo_01798 37.4 6.1E-11 1294
Cangl_01703 KOG00434 36.2 6.6E-11 1072
7296537 KOG00434 35.0 7.0E-11 1081
nDrome_08034 35.0 7.0E-11 1229
nDrome_09237 35.0 7.0E-11 1229
nDrome_09754 35.0 7.0E-11 1229
Plafa_00323 KOG00434 32.2 8.1E-11 1272
Danre_05440 KOG00434 27.0 1.1E-10 1271
Danre_14019 KOG00434 27.0 1.1E-10 1271
Kluwa_00189 KOG00434 25.4 1.2E-10 1099
Property Proteins Prot-Cogs HMM Global HMM Local
Hits are loaded for Giala_03726



Setting the score cutoff triggers creation of a list of potential false positives which have to be checked by a human curator

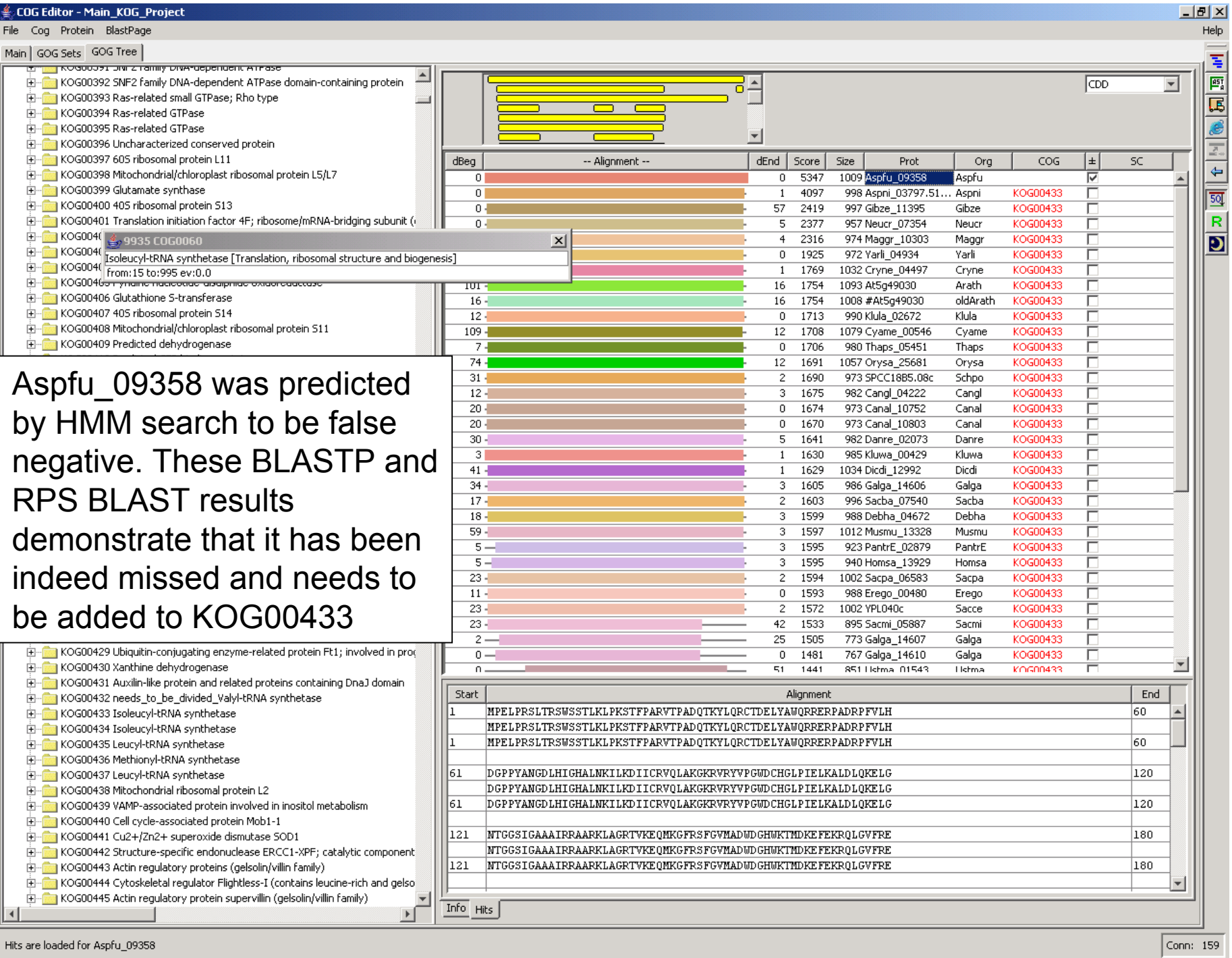

COG Editor - Main_KOG_Project
File Cog Protein BlastPage Help
Main GOG Sets GOG Tree
KOG00392 SNF2 family DNA-dependent ATPase domain-containing protein
KOG00393 Ras-related small GTPase; Rho type
KOG00394 Ras-related GTPase
KOG00395 Ras-related GTPase
KOG00396 Uncharacterized conserved protein
KOG00397 60S ribosomal protein L11
KOG00398 Mitochondrial/chloroplast ribosomal protein L5/L7
KOG00399 Glutamate synthase
KOG00400 40S ribosomal protein S13
KOG00401 Translation initiation factor 4F; ribosome/mRNA-bridging subunit
KOG00402 60S ribosomal protein L37
KOG00403 Neoplastic transformation suppressor Pdcd4/MA-3; contains MA3
KOG00404 Thioredoxin reductase
KOG00405 Pyridine nucleotide-disulphide oxidoreductase
KOG00406 Glutathione S-transferase
KOG00407 40S ribosomal protein S14
KOG00408 Mitochondrial/chloroplast ribosomal protein S11
KOG00409 Predicted dehydrogenase
KOG00410 Predicted GTP binding protein
KOG00411 Uncharacterized membrane protein
KOG00412 Golgi transport complex COD1 protein
KOG00413 Uncharacterized conserved protein related to condensin complex
KOG00414 Chromosome condensation complex Condensin; subunit D2
KOG00422 Ubiquitin-protein ligase
KOG00423 Ubiquitin-protein ligase
KOG00424 Ubiquitin-protein ligase
KOG00425 Ubiquitin-protein ligase
KOG00426 Ubiquitin-protein ligase
KOG00427 Ubiquitin conjugating enzyme
KOG00428 Non-canonical ubiquitin conjugating enzyme 1
KOG00429 Ubiquitin-conjugating enzyme-related protein Ft1; involved in pro
KOG00430 Xanthine dehydrogenase
KOG00431 Auxilin-like protein and related proteins containing DnaJ domain
KOG00432 needs_to_be_divided_Valyl-tRNA synthetase
KOG00433 Isoleucyl-tRNA synthetase
KOG00434 Isoleucyl-tRNA synthetase
KOG00435 Leucyl-tRNA synthetase
KOG00436 Methionyl-tRNA synthetase
KOG00437 Leucyl-tRNA synthetase
KOG00438 Mitochondrial ribosomal protein L2
KOG00439 VAMP-associated protein involved in inositol metabolism
KOG00440 Cell cycle-associated protein Mob1-1
KOG00441 Cu2+/Zn2+ superoxide dismutase SOD1
KOG00442 Structure-specific endonuclease ERCC1-XPF; catalytic component
KOG00443 Actin regulatory proteins (gelsolin/villin family)
KOG00444 Cytoskeletal regulator Flightless-I (contains leucine-rich and gelso
KOG00445 Actin regulatory protein supervillin (gelsolin/villin family)
And now do the same with the local hmm search result…
Threshold is not yet defined
Get PF
Name COG Sup Score E-Value pf lng
Gibze_11395 KOG00433 1.238E3 0.0 997
Galga_14609 KOG00433 1.237E3 0.0 728
Neucr_07354 KOG00433 1.236E3 0.0 957
Kluwa_00429 KOG00433 1.231E3 0.0 985
Aspni_03797 DIV 1.224E3 0.0 1518
Erego_00480 KOG00433 1.217E3 0.0 988
Dicdi_12992 KOG00433 1.18E3 0.0 1034
Yarli_04934 KOG00433 1.163E3 0.0 972
Tetni_18015 KOG00433 1.144E3 0.0 933
Ratno_18798 1.136E3 0.0 1136
Debha_04672 KOG00433 1.093E3 0.0 988
Cryne_04497 KOG00433 1.075E3 0.0 1032
Anoga_05900 KOG00433 1.069E3 0.0 806
KOG00433 1.006E3 0.0 973
KOG00433 1.004E3 0.0 973
961.0 0.0 620
KOG00433 843.8 0.0 851
KOG00433 821.8 0.0 470
KOG00433 779.9 0.0 485
Plabe_04226 KOG00433 482.4 0.0 687
Pantr_01696 413.5 0.0 249
Playo_05169 395.9 0.0 647
Chlre_10737 KOG00433 368.8 0.0 1153
Apime_11855 KOG00433 302.6 0.0 230
Thepa_04012 285.0 0.0 1221
Playo_02836 278.5 0.0 1197
Plabe_09243 KOG 222.0 0.0 267
Enthi_08456 KOG00434 218.4 0.0 1056
7296537 KOG00434 207.6 0.0 1081
nDrome_08034 207.6 0.0 1229
nDrome_09237 207.6 0.0 1229
nDrome_09754 207.6 0.0 1229
Canal_11674 KOG00434 206.3 0.0 1088
Canal_11824 KOG00434 206.3 0.0 1088
Anoga_00791 KOG00434 205.4 0.0 1213
Cangl_01703 KOG00434 204.3 0.0 1072
Galga_05083 KOG00434 200.2 0.0 1262
Galga_05084 KOG00434 196.0 0.0 992
Cioin_10894 KOG00434 194.4 0.0 1175
Erego_02472 KOG00434 192.6 0.0 1072
Danre_05440 KOG00434 192.6 0.0 1271
Danre_14019 KOG00434 192.6 0.0 1271
Sacmi_02885 KOG00434 191.8 0.0 1072
Property Proteins Prot-Cogs HMM Global HMM Local
Hits are loaded for Giala_03726
Conn: 158


File
Cog
Protein
BlastPage
Help
Main
GOG Sets
GOG Tree
KOG00392 SNF2 family DNA-dependent ATPase domain-containing protein
KOG00393 Ras-related small GTPase; Rho type
KOG00394 Ras-related GTPase
KOG00395 Ras-related GTPase
KOG00396 Uncharacterized conserved protein
KOG00397 60S ribosomal protein L11
KOG00398 Mitochondrial/chloroplast ribosomal protein L5/L7
KOG00399 Glutamate synthase
KOG00400 40S ribosomal protein S13
KOG00401 Translation initiation factor 4F; ribosome/mRNA-bridging subunit
KOG00402 60S ribosomal protein L37
KOG00403 Neoplastic transformation suppressor Pdcd4/MA-3; contains MA3
KOG00404 Thioredoxin reductase
KOG00405 Pyridine nucleotide-disulphide oxidoreductase
KOG00406 Glutathione S-transferase
KOG00407 40S ribosomal protein S14
KOG00408 Mitochondrial/chloroplast ribosomal protein S11
KOG00409 Predicted dehydrogenase
KOG00410 Predicted GTP binding protein
KOG00411 Uncharacterized membrane protein
KOG00412 Golgi transport complex COD1 protein
KOG00413 Uncharacterized conserved protein related to condensin complex
KOG00414 Chromosome condensation complex Condensin; subunit D2
KOG00415 Predicted peptidyl prolyl cis-trans isomerase
KOG00416 Ubiquitin-protein ligase
KOG00417 Ubiquitin-protein ligase
KOG00418 Ubiquitin-protein ligase
KOG00419 Ubiquitin-protein ligase
KOG00420 Ubiquitin-protein ligase
KOG00421 Ubiquitin-protein ligase
KOG00422 Ubiquitin-protein ligase
KOG00423 Ubiquitin-protein ligase
KOG00424 Ubiquitin-protein ligase
KOG00425 Ubiquitin-protein ligase
KOG00426 Ubiquitin-protein ligase
KOG00427 Ubiquitin conjugating enzyme
KOG00428 Non-canonical ubiquitin conjugating enzyme 1
KOG00429 Ubiquitin-conjugating enzyme-related protein Ft1; involved in pro
KOG00430 Xanthine dehydrogenase
KOG00431 Auxilin-like protein and related proteins containing DnaJ domain
KOG00432 needs_to_be_divided_Valyl-tRNA synthetase
KOG00433 Isoleucyl-tRNA synthetase
KOG00434 Isoleucyl-tRNA synthetase
KOG00435 Leucyl-tRNA synthetase
KOG00436 Methionyl-tRNA synthetase
KOG00437 Leucyl-tRNA synthetase
KOG00438 Mitochondrial ribosomal protein L2
KOG00439 VAMP-associated protein involved in inositol metabolism
KOG00440 Cell cycle-associated protein Mob1-1
KOG00441 Cu2+/Zn2+ superoxide dismutase SOD1
KOG00442 Structure-specific endonuclease ERCC1-XPF; catalytic component
KOG00443 Actin regulatory proteins (gelsolin/villin family)
KOG00444 Cytoskeletal regulator Flightless-I (contains leucine-rich and gelso
KOG00445 Actin regulatory protein supervillin (gelsolin/villin family)
Aspni_03797.519.1518
CE08373
Caebr_08775
Gibze_09654
Leima_03326
Plabe_11680
Sacba_06308
Takru_19036
Tetni_05919
Tetni_15833
Thaps_08507
Trycr_17705
Get PF
Name
COG
Sup
Score
E-Value
pf
lng
Gibze_11395 KOG00433 1.238E3 0.0 997
Galga_14609 KOG00433 1.237E3 0.0 728
Neucr_07354 KOG00433 1.236E3 0.0 957
Kluwa_00429 KOG00433 1.231E3 0.0 985
Aspni_03797 DIV 1.224E3 0.0 1518
Erego_00480 KOG00433 1.217E3 0.0 988
Dicdi_12992 KOG00433 1.18E3 0.0 1034
Yarli_04934 KOG00433 1.163E3 0.0 972
Tetni_18015 KOG00433 1.144E3 0.0 933
Ratno_18798 1.136E3 0.0 1136
Debha_04672 KOG00433 1.093E3 0.0 988
Cryne_04497 KOG00433 1.075E3 0.0 1032
Anoga_05900 KOG00433 1.069E3 0.0 806
Canal_10752 KOG00433 1.006E3 0.0 973
Canal_10803 KOG00433 1.004E3 0.0 973
Pantr_01697 961.0 0.0 620
Ustma_01543 KOG00433 843.8 0.0 851
Galga_14611 KOG00433 821.8 0.0 470
Takru_16952 KOG00433 779.9 0.0 485
Plabe_04226 KOG00433 482.4 0.0 687
Pantr_01696 413.5 0.0 249
Playo_05169 395.9 0.0 647
Chlre_10737 KOG00433 368.8 0.0 1153
Apime_11855 KOG00433 302.6 0.0 230
Thepa_04012 285.0 0.0 1221
Playo_02836 278.5 0.0 1197
Plabe_09243 KOG00433 222.0 0.0 267
Enthi_08456 KOG00434 218.4 0.0 1056
7296537 KOG00434 207.6 0.0 1081
nDrome_08034 207.6 0.0 1229
nDrome_09237 207.6 0.0 1229
nDrome_09754 207.6 0.0 1229
Canal_11674 KOG00434 206.3 0.0 1088
Canal_11824 KOG00434 206.3 0.0 1088
Anoga_00791 KOG00434 205.4 0.0 1213
Cangl_01703 KOG00434 204.3 0.0 1072
Galga_05083 KOG00434 200.2 0.0 1262
Galga_05084 KOG00434 196.0 0.0 992
Property
Proteins
Prot-Cogs
HMM Global
HMM Local
Hits are loaded for Giala_03726
Conn: 158

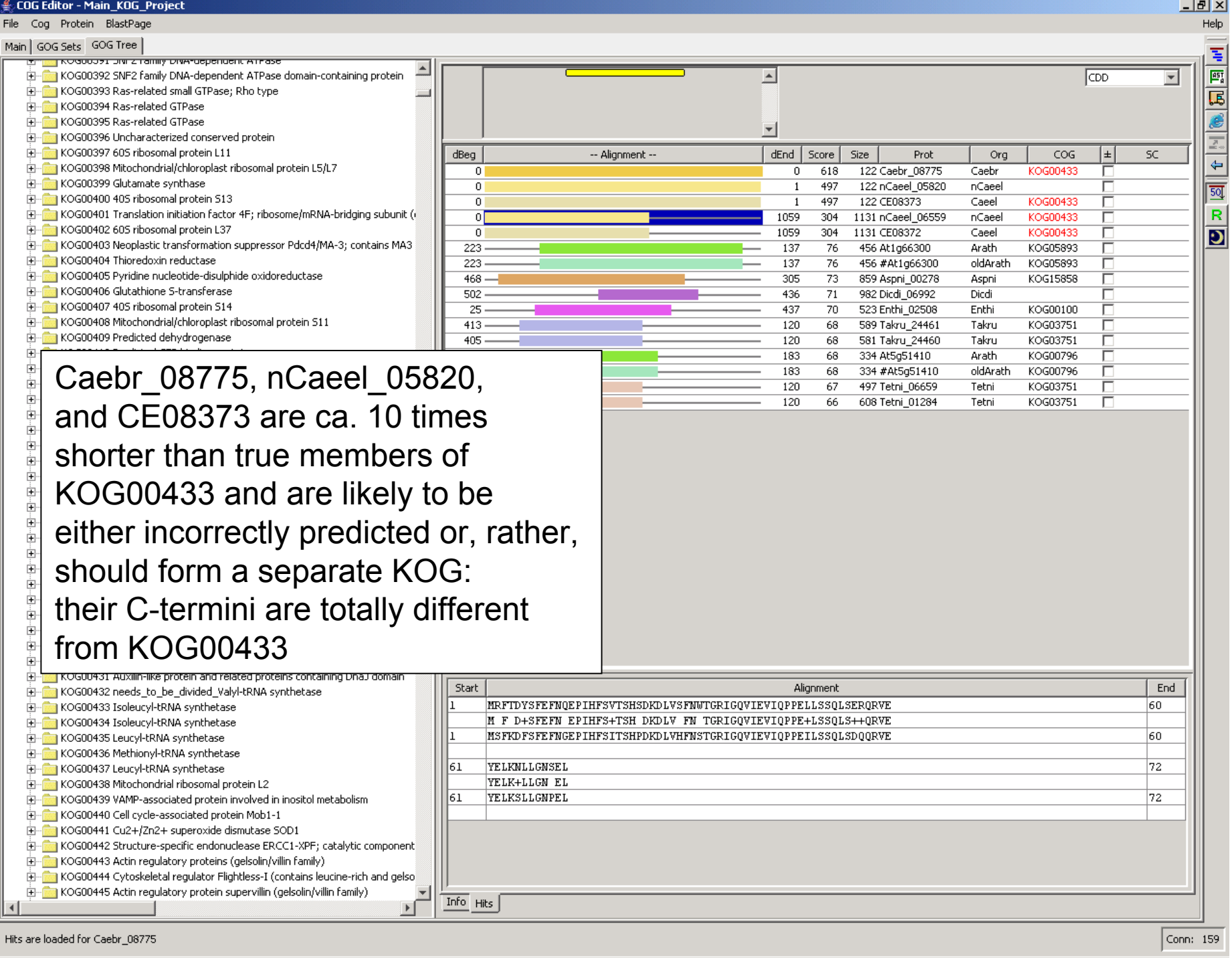

COG Editor - Main_KOG_Project
File Cog Protein BlastPage Help
Main GOG Sets GOG Tree
KOG00392 SNF2 family DNA-dependent ATPase domain-containing protein
KOG00393 Ras-related small GTPase; Rho type
KOG00394 Ras-related GTPase
KOG00395 Ras-related GTPase
KOG00396 Uncharacterized conserved protein
KOG00397 60S ribosomal protein L11
KOG00398 Mitochondrial/chloroplast ribosomal protein L5/L7
KOG00399 Glutamate synthase
KOG00400 40S ribosomal protein S13
KOG00401 Translation initiation factor 4F; ribosome/mRNA-bridging subunit
KOG00402 60S ribosomal protein L37
KOG00403 Neoplastic transformation suppressor Pdcd4/MA-3; contains MA3
KOG00404 Thioredoxin reductase
KOG00405 Pyridine nucleotide-disulphide oxidoreductase
KOG00406 Glutathione S-transferase
KOG00407 40S ribosomal protein S14
KOG00408 Mitochondrial/chloroplast ribosomal protein S11
KOG00409 Predicted dehydrogenase
KOG00410 Predicted GTP binding protein
KOG00411 Uncharacterized membrane protein
KOG00412 Golgi transport complex COD1 protein
KOG00424 Ubiquitin-protein ligase
KOG00425 Ubiquitin-protein ligase
KOG00426 Ubiquitin-protein ligase
KOG00427 Ubiquitin conjugating enzyme
KOG00428 Non-canonical ubiquitin conjugating enzyme 1
KOG00429 Ubiquitin-conjugating enzyme-related protein Ft1; involved in prog
KOG00430 Xanthine dehydrogenase
KOG00431 Auxilin-like protein and related proteins containing DnaJ domain
KOG00432 needs_to_be_divided_Valyl-tRNA synthetase
KOG00433 Isoleucyl-tRNA synthetase
KOG00434 Isoleucyl-tRNA synthetase
KOG00435 Leucyl-tRNA synthetase
KOG00436 Methionyl-tRNA synthetase
KOG00437 Leucyl-tRNA synthetase
KOG00438 Mitochondrial ribosomal protein L2
KOG00439 VAMP-associated protein involved in inositol metabolism
KOG00440 Cell cycle-associated protein Mob1-1
KOG00441 Cu2+/Zn2+ superoxide dismutase SOD1
KOG00442 Structure-specific endonuclease ERCC1-XPF; catalytic component
KOG00443 Actin regulatory proteins (gelsolin/villin family)
KOG00444 Cytoskeletal regulator Flightless-I (contains leucine-rich and gelso
KOG00445 Actin regulatory protein supervillin (gelsolin/villin family)
Setting global and local hmm search cutoffs allows to get an automatic list of potential fragments
Aspni_03797.519.1518
CE08373
Caebr_08775
Gibze_09654
Leima_03326
Plabe_11680
Sacba_06308
Takru_19036
Tetni_05919
Tetni_15833
Thaps_08507
Trycr_17705
Get PF
Name COG Sup Score E-Value pf lng
Gibze_11395 KOG00433 1.238E3 0.0 997
Galga_14609 KOG00433 1.237E3 0.0 728
Neucr_07354 KOG00433 1.236E3 0.0 957
Kluwa_00429 KOG00433 1.231E3 0.0 985
Aspni_03797 DIV 1.224E3 0.0 1518
KOG00433 1.217E3 0.0 988
KOG00433 1.18E3 0.0 1034
KOG00433 1.163E3 0.0 972
KOG00433 1.144E3 0.0 933
1.136E3 0.0 1136
KOG00433 1.093E3 0.0 988
KOG00433 1.075E3 0.0 1032
KOG00433 1.069E3 0.0 806
KOG00433 1.006E3 0.0 973
KOG00433 1.004E3 0.0 973
961.0 0.0 620
KOG00433 843.8 0.0 851
Galga_14611 KOG00433 821.8 0.0 470
Takru_16952 KOG00433 779.9 0.0 485
Plabe_04226 KOG00433 482.4 687
Pantr_01696 413.5 249
Playo_05169 395.9 647
Chlre_10737 KOG00433 368.8 1153
Apime_11855 KOG00433 302.6 230
Thepa_04012 285.0 0.0 1221
Playo_02836 278.5 1197
Plabe_09243 KOG00433 222.0 267
Enthi_08456 KOG00434 218.4 0.0 1056
7296537 KOG00434 207.6 0.0 1081
nDrome_08034 207.6 0.0 1229
nDrome_09237 207.6 0.0 1229
nDrome_09754 207.6 0.0 1229
Canal_11674 KOG00434 206.3 0.0 1088
Canal_11824 KOG00434 206.3 0.0 1088
Anoga_00791 KOG00434 205.4 0.0 1213
Cangl_01703 KOG00434 204.3 0.0 1072
Galga_05083 KOG00434 200.2 0.0 1262
Galga_05084 KOG00434 196.0 0.0 992
Property Proteins Prot-Cogs HMM Global HMM Local
Hits are loaded for Giala_03726
Conn: 158

What trouble did KOGnitor make to KOG00433?

We take the list of potential false positives and potential false negatives and compute BLASTP with output in xml format. The resulting files are loaded into KOG db temporarily to check all the sequences in question.



Aspfu_09358 was predicted by HMM search to be false negative. These BLASTP and RPS BLAST results demonstrate that it has been indeed missed and needs to be added to KOG00433


COG Editor - Main_KOG_Project
File Cog Protein BlastPage Help
Main GOG Sets GOG Tree
KOG00392 SNF2 family DNA-dependent ATPase domain-containing protein
KOG00393 Ras-related small GTPase; Rho type
KOG00394 Ras-related GTPase
KOG00395 Ras-related GTPase
KOG00396 Uncharacterized conserved protein
KOG00397 60S ribosomal protein L11
KOG00398 Mitochondrial/chloroplast ribosomal protein L5/L7
KOG00399 Glutamate synthase
KOG00400 40S ribosomal protein S13
KOG00401 Translation initiation factor 4F; ribosome/mRNA-bridging subunit
KOG00402 60S ribosomal protein L37
KOG00403 Neoplastic transformation suppressor Pdcd4/MA-3; contains MA3
KOG00404 Thioredoxin reductase
KOG00405 Pyridine nucleotide-disulphide oxidoreductase
KOG00406 Glutathione S-transferase
KOG00407 40S ribosomal protein S14
KOG00408 Mitochondrial/chloroplast ribosomal protein S11
KOG00409 Predicted dehydrogenase
KOG00431 Auxilin-like protein and related proteins containing DnaJ domain
KOG00432 needs_to_be_divided_Valyl-tRNA synthetase
KOG00433 Isoleucyl-tRNA synthetase
KOG00434 Isoleucyl-tRNA synthetase
KOG00435 Leucyl-tRNA synthetase
KOG00436 Methionyl-tRNA synthetase
KOG00437 Leucyl-tRNA synthetase
KOG00438 Mitochondrial ribosomal protein L2
KOG00439 VAMP-associated protein involved in inositol metabolism
KOG00440 Cell cycle-associated protein Mob1-1
KOG00441 Cu2+/Zn2+ superoxide dismutase SOD1
KOG00442 Structure-specific endonuclease ERCC1-XPF; catalytic component
KOG00443 Actin regulatory proteins (gelsolin/villin family)
KOG00444 Cytoskeletal regulator Flightless-I (contains leucine-rich and gelso
KOG00445 Actin regulatory protein supervillin (gelsolin/villin family)
CDD
dBeg -- Alignment -- dEnd Score Size Prot Org COG ± SC
0 0 618 122 Caebr_08775 Caebr KOG00433
0 1 497 122 nCaeel_05820 nCaeel
0 1 497 122 CE08373 Caeel KOG00433
0 1059 304 1131 nCaeel_06559 nCaeel KOG00433
0 1059 304 1131 CE08372 Caeel KOG00433
223 137 76 456 At1g66300 Arath KOG05893
223 137 76 456 #At1g66300 oldArath KOG05893
468 305 73 859 Aspni_00278 Aspni KOG15858
502 436 71 982 Dicdi_06992 Dicdi
25 437 70 523 Enthi_02508 Enthi KOG00100
413 120 68 589 Takru_24461 Takru KOG03751
405 120 68 581 Takru_24460 Takru KOG03751
183 68 334 At5g51410 Arath KOG00796
183 68 334 #At5g51410 oldArath KOG00796
120 67 497 Tetni_06659 Tetni KOG03751
120 66 608 Tetni_01284 Tetni KOG03751
Caebr_08775, nCaeel_05820, and CE08373 are ca. 10 times shorter than true members of KOG00433 and are likely to be either incorrectly predicted or, rather, should form a separate KOG: their C-termini are totally different from KOG00433
Start Alignment End
1 MRFTDYSFEFNQEPIHFSVTSHSDKDLVSFNWTGRIGQVIEVIQPPELLSSQLSERQRVE 60
M F D+SFEFN EPIHFS+TSH DKDLV FN TGRIGQVIEVIQPPE+LSSQLS++QRVE
1 MSFKDFSFEFNGEPIHFSITSHPDKDLVHFNSTGRIGQVIEVIQPPEILSSQLSDQQRVE 60
61 YELKNLLGNSEL 72
YELK+LLGN EL
61 YELKSLLGNPEL 72
Info Hits
Hits are loaded for Caebr_08775
Conn: 159

Action after checking potential false positives list in KOG00433:

- 10 sequences are removed, 7 of which formed three new KOGs;
- 4 sequences must be added to the HMM profile

NEW KOG

Leima_03326 and Trycr_17705 were false positives, were removed and formed a new KOG
which also includes "XP_823314 71748518 hypothetical protein Tb10.26.1000 [Trypanosoma brucei]"
which is now present in the GeneBank but was absent from the downloaded T.brucei proteome

NEW KOG

Thaps_08507 was added to KOG00433 by a spurious unidirectional hit, it dragged Sacba_06308 into
KOG00433; both were removed and formed a new KOG

NEW KOG

Caebr_08775, nCaeel_05820, and CE08373 are ca. 10 times shorter than true members of
KOG00433 and are likely to be either incorrectly predicted or, rather, should form a separate KOG:
their C-termini are totally different from KOG00433

REMOVE

Tetni_05919, Tetni_15833, Gibze_09654 were false positive dragged into KOG00433 by spurious BLAST hits;
they were removed from KOG00433 and were made free

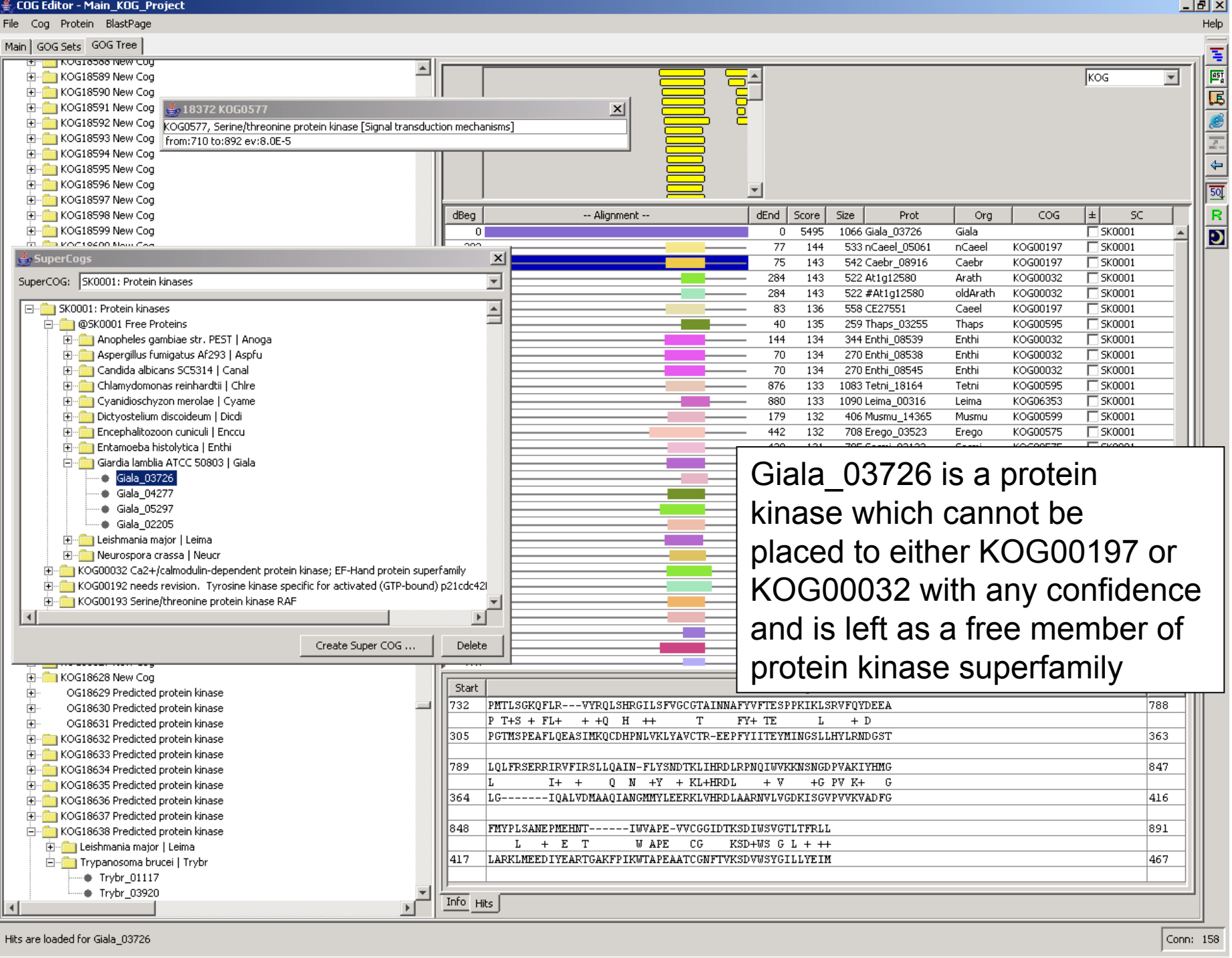
ADD TO HMM

Plabe_04226, Playo_05169, Plabe_09243, and Plafa_01680 are legitimate members of KOG00433 and MUST be
added to the HMM profile

Plabe_11680, Apime_11855, Chlre_10737, and Takru_19036 are fragments and are legitimate members of KOG00433

With precomputed XML BLAST files it takes a few minutes

Superfamilies and KOGs


COG Editor - Main_KOG_Project
File Cog Protein BlastPage Help
Main GOG Sets GOG Tree
KOG18589 New Cog
KOG18590 New Cog
KOG18591 New Cog
KOG18592 New Cog
KOG18593 New Cog
KOG18594 New Cog
KOG18595 New Cog
KOG18596 New Cog
KOG18597 New Cog
KOG18598 New Cog
KOG18599 New Cog
18372 KOG0577
KOG0577, Serine/threonine protein kinase [Signal transduction mechanisms]
from:710 to:892 ev:8.0E-5
KOG
SuperCogs
SuperCOG: SK0001: Protein kinases
SK0001: Protein kinases
@SK0001 Free Proteins
Anopheles gambiae str. PEST | Anoga
Aspergillus fumigatus Af293 | Aspfu
Candida albicans SC5314 | Canal
Chlamydomonas reinhardtii | Chlre
Cyanidioschyzon merolae | Cyame
Dictyostelium discoideum | Dicdi
Encephalitozoon cuniculi | Enccu
Entamoeba histolytica | Enthi
Giardia lamblia ATCC 50803 | Giala
Giala_03726
Giala_04277
Giala_05297
Giala_02205
Leishmania major | Leima
Neurospora crassa | Neucr
KOG00032 Ca2+/calmodulin-dependent protein kinase; EF-Hand protein superfamily
KOG00192 needs revision. Tyrosine kinase specific for activated (GTP-bound) p21cdc42
KOG00193 Serine/threonine protein kinase RAF
Create Super COG ...
Delete
dBeg -- Alignment -- dEnd Score Size Prot Org COG ± SC
0 0 5495 1066 Giala_03726 Giala SK0001
77 144 533 nCaeel_05061 nCaeel KOG00197 SK0001
75 143 542 Caebr_08916 Caebr KOG00197 SK0001
284 143 522 At1g12580 Arath KOG00032 SK0001
284 143 522 #At1g12580 oldArath KOG00032 SK0001
83 136 558 CE27551 Caeel KOG00197 SK0001
40 135 259 Thaps_03255 Thaps KOG00595 SK0001
144 134 344 Enthi_08539 Enthi KOG00032 SK0001
70 134 270 Enthi_08538 Enthi KOG00032 SK0001
70 134 270 Enthi_08545 Enthi KOG00032 SK0001
876 133 1083 Tetni_18164 Tetni KOG00595 SK0001
880 133 1090 Leima_00316 Leima KOG06353 SK0001
179 132 406 Musmu_14365 Musmu KOG00599 SK0001
442 132 708 Erego_03523 Erego KOG00575 SK0001
KOG18628 New Cog
OG18629 Predicted protein kinase
OG18630 Predicted protein kinase
OG18631 Predicted protein kinase
KOG18632 Predicted protein kinase
KOG18633 Predicted protein kinase
KOG18634 Predicted protein kinase
KOG18635 Predicted protein kinase
KOG18636 Predicted protein kinase
KOG18637 Predicted protein kinase
KOG18638 Predicted protein kinase
Leishmania major | Leima
Trypanosoma brucei | Trybr
Trybr_01117
Trybr_03920
Start
732 PMTLSGKQFLR---VYRQLSHRGILSFVGCGTAINNAFYVFTESPPKIKLSRVFQYDEEA 788
P T+S + FL+ + +Q H ++ T FY+ TE L + D
305 PGTMSPEAFLQEASIMKQCDHPNLVKLYAVCTR-EEPFYIITEYMINGSLLHYLRNDGST 363
789 LQLFRSERRIRVFIRSLLQAIN-FLYSNDTKLIHRDLRPNQIWVKKNSNGDPVAKIYHMG 847
L I+ + Q N +Y + KL+HRDL + V +G PV K+ G
364 LG-------IQALVDMAAQIANGMMYLEERKLVHRDLAARNVLVGDKISGVPVVKVADFG 416
848 FMYPLSANEPMEHNT------IWVAPE-VVCGGIDTKSDIWSVGTLTFRLL 891
L + E T W APE CG KSD+WS G L + ++
417 LARKLMEEDIYEARTGAKFPIKWTAPEAATCGNFTVKSDVWSYGILLYEIM 467
Info Hits
Giala_03726 is a protein kinase which cannot be placed to either KOG00197 or KOG00032 with any confidence and is left as a free member of protein kinase superfamily
Hits are loaded for Giala_03726
Conn: 158

Most important practical application of KOG db would be transfer of gene annotation across species

Query: gi|58385127 ENSANGP00000016957 [Anopheles gambiae str. PEST]
Matching gi: 55240814

Show identical | Best hits | C

Animal proteins from KOG04771 are not annotated;
There are no hits to fungal proteins

25 BLAST hits to 17 unique species

0 Archaea | 0 Bacteria | 24 Metaz

Keep only | Cut-Off 100 | Select | Reset | New search by GI: 58385127 | Go

199 aa

SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
338	20	NP_651884	24651718	CG11563-PA [Drosophila melanogaster]
338	20	AAM29402	21064345	RE07994p [Drosophila melanogaster]
338	20	EAL28137	54638735	GA11068-PA [Drosophila pseudoobscura]
259	8	CAF97276	47221358	unnamed protein product [Tetraodon nigroviridis]
241	8	NP_001...	62460412	hypothetical protein LOC506421 [Bos taurus]
230	8	CAG33450	48146455	HSPC111 [Homo sapiens]
229	8	XP_518115	55625438	PREDICTED: hypothetical protein XP_518115 [Pan troglodytes]
229	8	XP_536424	57085179	PREDICTED: similar to Protein CGI-117 [Canis familiaris]
227	8	NP_057475	7705451	hypothetical protein LOC51491 [Homo sapiens]
225	8	BAC40235	26353210	unnamed protein product [Mus musculus]
225	8	NP_848720	30519925	hypothetical protein LOC28126 [Mus musculus]
225	8	XP_225176	27682843	PREDICTED: similar to DNA segment, Chr 13, Wayne State University 177, expressed [Rattus norvegicus]
221	8	AAH12213	15126561	DNA segment, Chr 13, Wayne State University 177, expressed [Mus musculus]
221	8	NP_001...	50540130	hypothetical protein LOC436807 [Danio rerio]
190	8	AAH68216	45872610	Unknown (protein for MGC:76225) [Xenopus tropicalis]
185	8	AAH73261	49255993	MGC80623 protein [Xenopus laevis]
175	8	AAH32424	34783430	Unknown (protein for MGC:40298) [Homo sapiens]
172	8	XP_782922	72007783	PREDICTED: similar to Protein CGI-117 [Strongylocentrotus purpuratus]
169	7	CAE57069	39578894	Hypothetical protein CBG24963 [Caenorhabditis briggsae]
168	8	XP_414551	50754941	PREDICTED: similar to Hypothetical protein HSPC111 [Gallus gallus]
158	7	NP_492248	17511059	ZK265.6 [Caenorhabditis elegans]
129	8	XP_343880	62655980	PREDICTED: similar to DNA segment, Chr 13, Wayne State University 177, expressed [Rattus norvegicus]
109	8	XP_688990	68356182	PREDICTED: similar to Protein CGI-117 [Danio rerio]
101	8	XP_357941	63586355	PREDICTED: similar to DNA segment, Chr 13, Wayne State University 177, expressed [Mus musculus]
100	3	XP_810459	71417059	hypothetical protein, conserved [Trypanosoma cruzi]

gi|7705451 view - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?pid=7705451

Google Search 195 blocked Check

Go Links

Blink search with the human protein does not help much either

Show identical | Best hits | Common Tree | Taxonomy Report | 3D structures | CDD-Search | GI list

34 BLAST hits to 19 unique species Sort by taxonomy proximity

0 Archaea | 0 Bacteria | 32 Metazoa | 0 Fungi | 0 Plants | 0 Viruses | 2 Other Eukaryotae

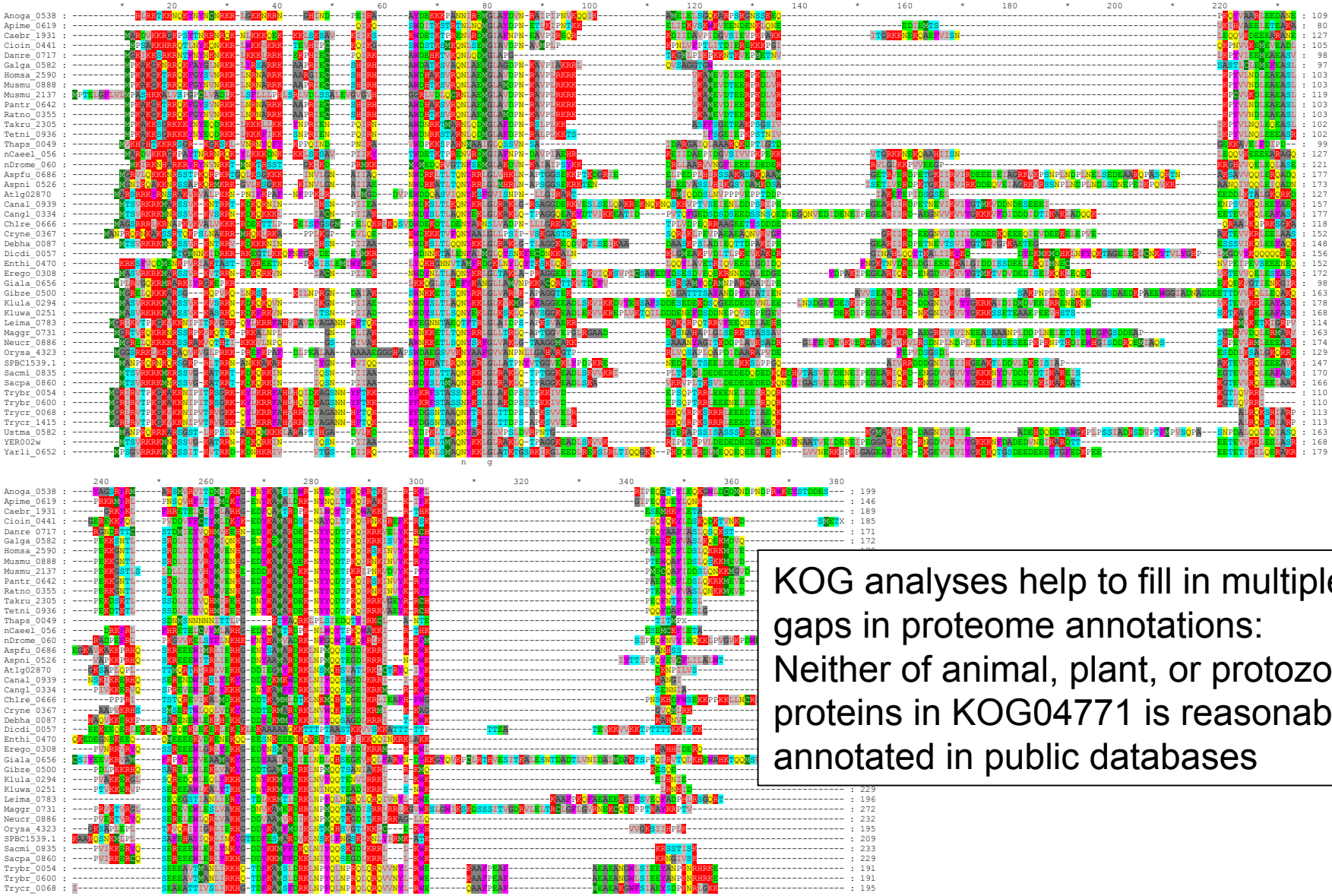
Keep only | Cut-Off 100 | Select | Reset | New search by GI: 7705451 | Go

178 aa

SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
915	28	XP_518115	55625438	PREDICTED: hypothetical protein XP_518115 [Pan troglodytes]
912	30	CAG33450	48146455	HSPC111 [Homo sapiens]
882	21	XP_536424	57085179	PREDICTED: similar to Protein CGI-117 [Canis familiaris]
863	21	NP_001...	62460412	hypothetical protein LOC506421 [Bos taurus]
857	22	NP_848720	30519925	hypothetical protein LOC28126 [Mus musculus]
857	22	BAC40235	26353210	unnamed protein product [Mus musculus]
848	22	AAH12213	15126561	DNA segment, Chr 13, Wayne State University 177, expressed [Mus musculus]
848	22	XP_225176	27682843	PREDICTED: similar to DNA segment, Chr 13, Wayne State University 177, expressed [Rattus norvegicus]
665	30	AAH32424	34783430	Unknown (protein for MGC:40298) [Homo sapiens]
597	18	XP_414551	50754941	PREDICTED: similar to Hypothetical protein HSPC111 [Gallus gallus]
477	17	AAH73261	49255993	MGC80623 protein [Xenopus laevis]
473	22	XP_357941	63586355	PREDICTED: similar to DNA segment, Chr 13, Wayne State University 177, expressed [Mus musculus]
473	17	AAH68216	45872610	Unknown (protein for MGC:76225) [Xenopus tropicalis]
463	15	CAF97276	47221358	unnamed protein product [Tetraodon nigroviridis]
413	15	NP_001...	50540130	hypothetical protein LOC436807 [Danio rerio]
404	22	XP_343880	62655980	PREDICTED: similar to DNA segment, Chr 13, Wayne State University 177, expressed [Rattus norvegicus]
340	30	AAF36105	7106760	HSPC185 [Homo sapiens]
280	15	XP_688990	68356182	PREDICTED: similar to Protein CGI-117 [Danio rerio]
251	9	XP_782922	72007783	PREDICTED: similar to Protein CGI-117 [Strongylocentrotus purpuratus]
227	8	XP_313729	58385127	ENSANGP00000016957 [Anopheles gambiae str. PEST]
197	7	CAE57069	39578894	Hypothetical protein CBG24963 [Caenorhabditis briggsae]
186	7	NP_492248	17511059	ZK265.6 [Caenorhabditis elegans]
163	8	EAL28137	54638735	GA11068-PA [Drosophila pseudoobscura]
160	8	NP_651884	24651718	CG11563-PA [Drosophila melanogaster]
160	8	AAM29402	21064345	RE07994p [Drosophila melanogaster]
111	3	XP_810459	71417059	hypothetical protein, conserved [Trypanosoma cruzi]
108	7	NP_509607	25152153	F49E2.5b [Caenorhabditis elegans]
108	7	NP_001...	71991080	F49E2.5h [Caenorhabditis elegans]
108	7	NP_001...	71991088	F49E2.5i [Caenorhabditis elegans]
108	7	NP_001...	71991094	F49E2.5j [Caenorhabditis elegans]
108	7	NP_509608	25152156	F49E2.5c [Caenorhabditis elegans]
108	7	NP_509605	25152150	F49E2.5a [Caenorhabditis elegans]
108	7	NP_509606	25152159	F49E2.5d [Caenorhabditis elegans]
105	3	CAJ02832	68124716	hypothetical protein, conserved [Leishmania major]

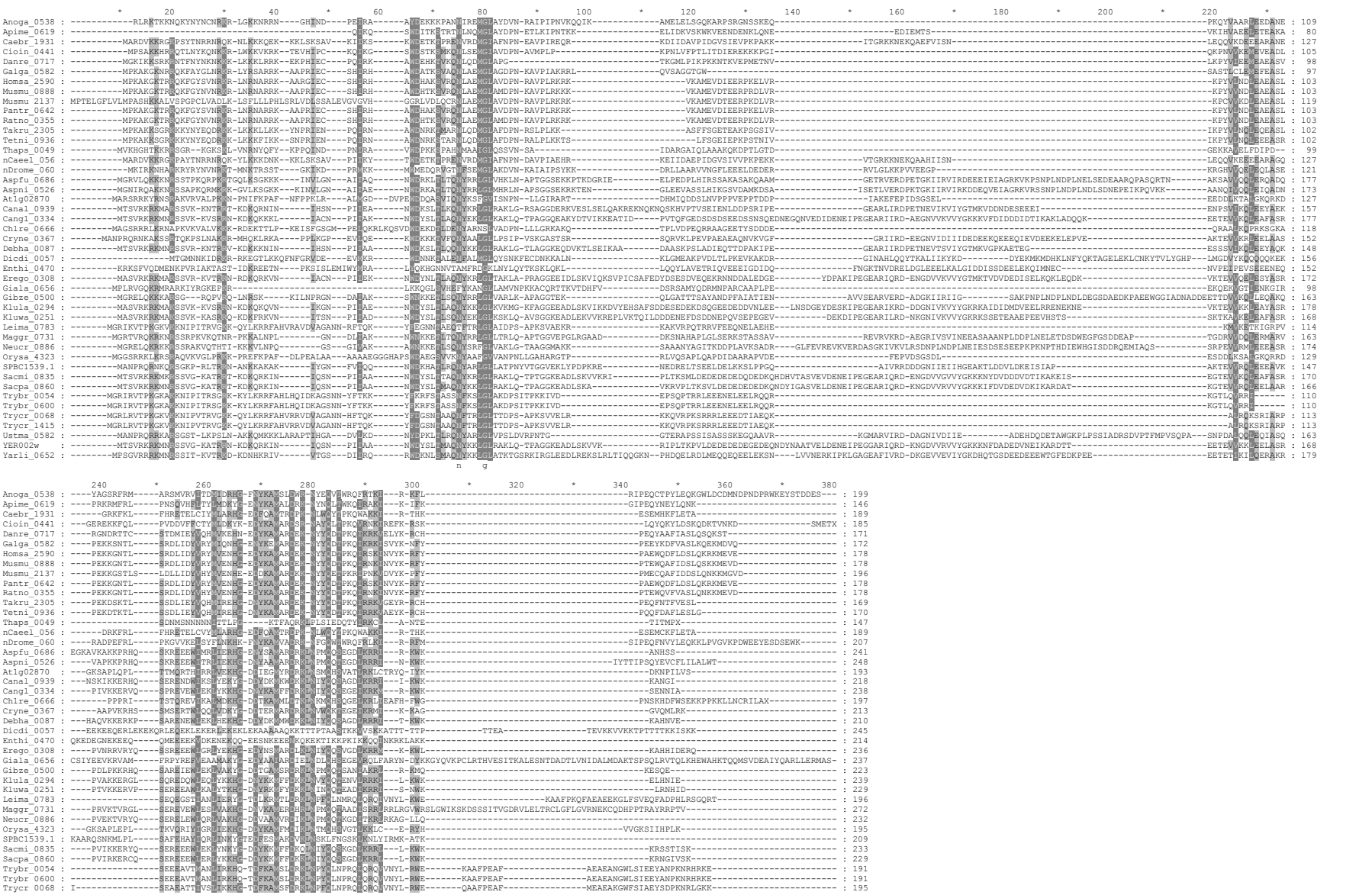
Done Internet

KOG04771 Nucleolar protein (NOP16) involved in 60S ribosomal subunit biogenesis

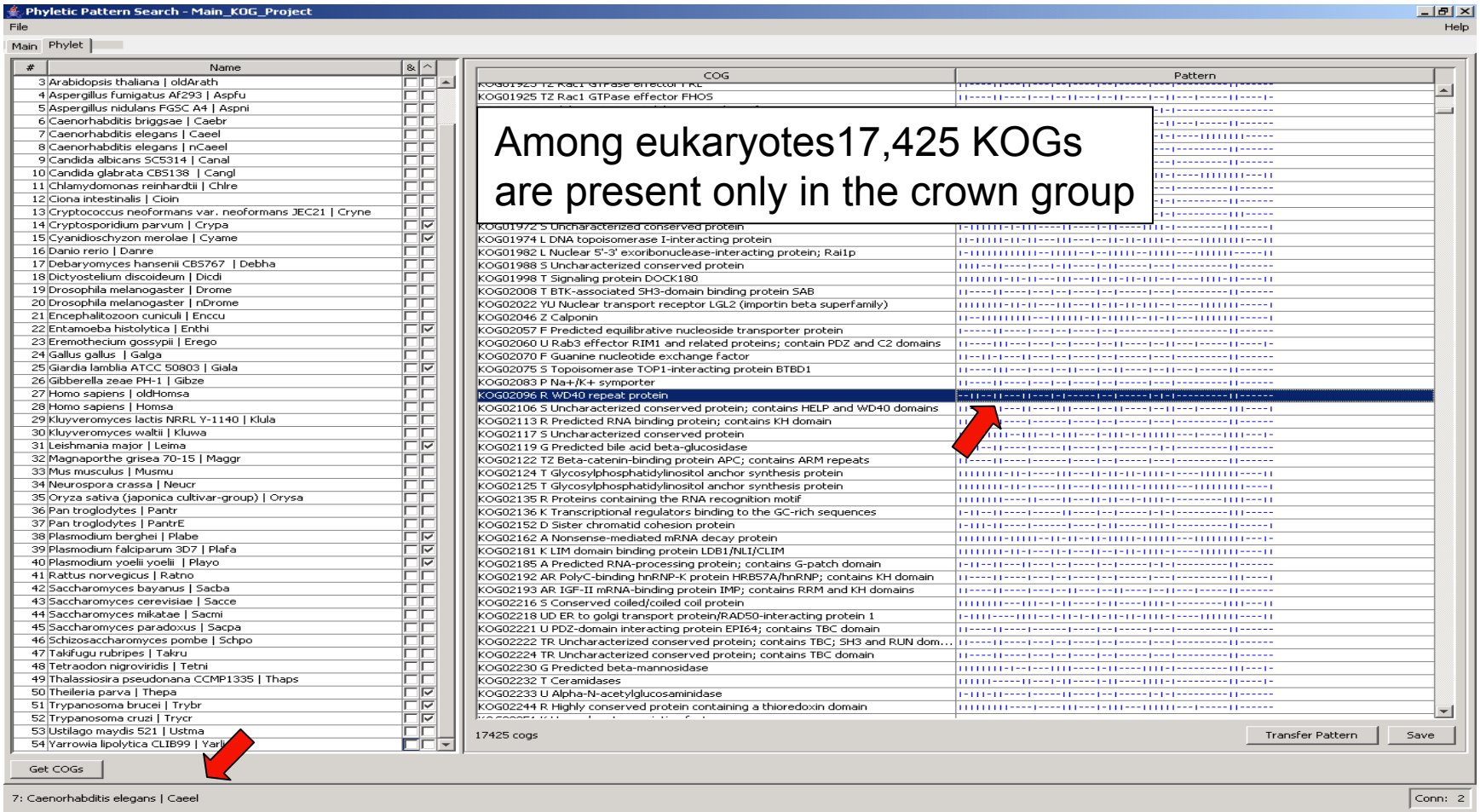


KOG analyses help to fill in multiple gaps in proteome annotations: Neither of animal, plant, or protozoan proteins in KOG04771 is reasonably annotated in public databases

KOG04771 Nucleolar protein (NOP16) involved in 60S ribosomal subunit biogenesis
Conserved residues



Phylogenetic pattern searches



Only 157 KOGs are represented in all KOGnitorized proteomes

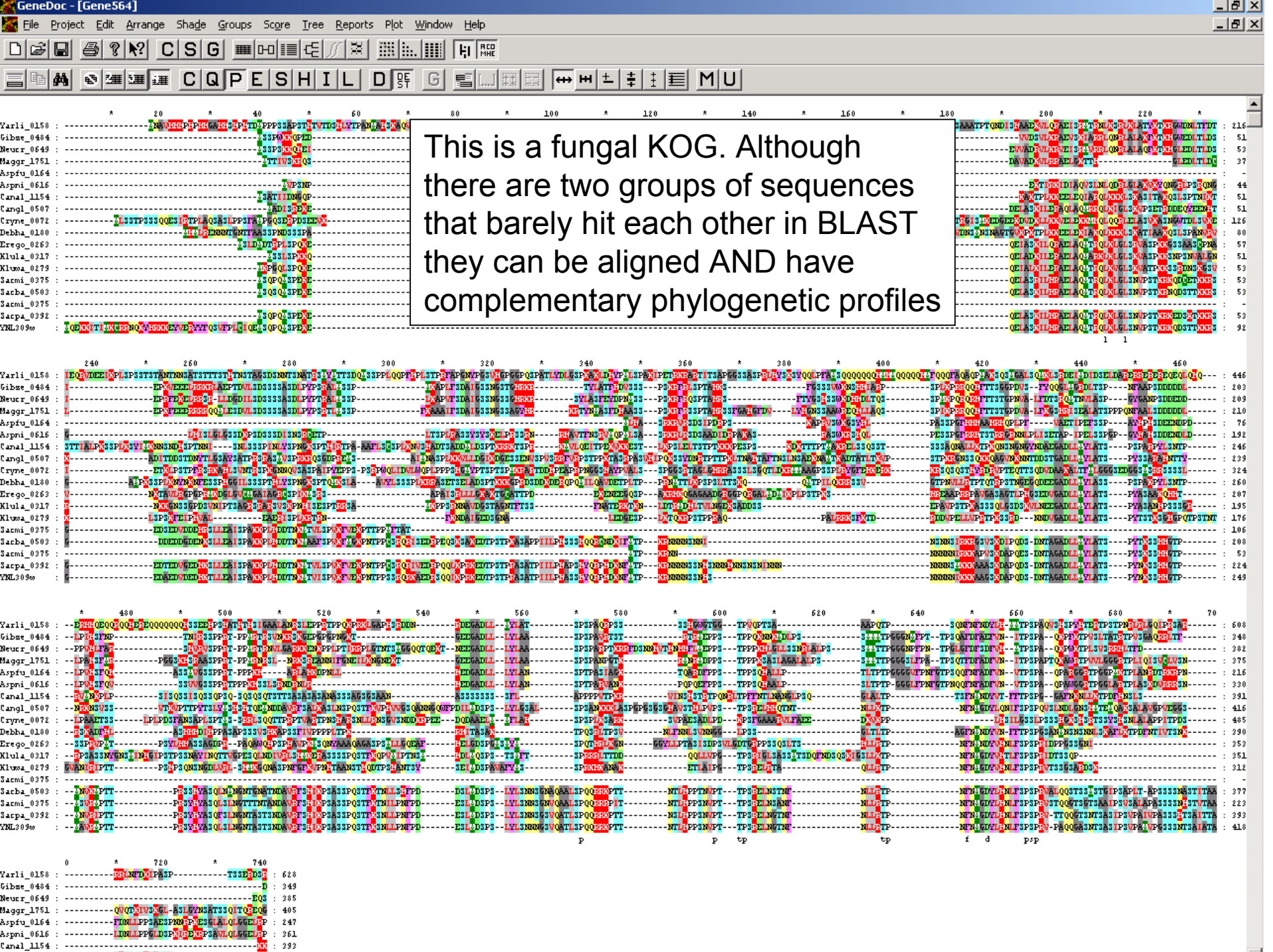
KOG00010 OR Ubiquitin-like protein
KOG00032 T Ca2+/calmodulin-dependent protein kinase; EF-Hand protein superfamily
KOG00034 T Ca2+/calmodulin-dependent protein phosphatase (calcineurin subunit B); EF-Hand superfamily protein
KOG00050 AD mRNA splicing protein CDC5 (Myb superfamily)
KOG00070 U GTP-binding ADP-ribosylation factor Arf1
KOG00077 U Vesicle coat complex COPII; GTPase subunit SAR1
KOG00084 TU GTPase Rab1/YPT1; small G protein superfamily; and related GTP-binding proteins
KOG00101 O Molecular chaperones HSP70/HSC70; HSP70 superfamily
KOG00102 O Molecular chaperones mortalin/PBP74/GRP75; HSP70 superfamily
KOG00131 A Splicing factor 3b; subunit 4
KOG00166 U Karyopherin (importin) alpha
KOG00173 O 20S proteasome; regulatory subunit beta type PSMB7/PSMB10/PUP1
KOG00175 O 20S proteasome; regulatory subunit beta type PSMB5/PSMB8/PRE2
KOG00188 J Alanyl-tRNA synthetase
KOG00190 O Protein disulfide isomerase (prolyl 4-hydroxylase beta subunit)
KOG00191 O Thioredoxin/protein disulfide isomerase
KOG00206 R P-type ATPase
KOG00211 T Protein phosphatase 2A regulatory subunit A and related proteins
KOG00213 A Splicing factor 3b; subunit 1
KOG00219 L Mismatch repair ATPase MSH2 (MutS family)
KOG00232 C Vacuolar H+-ATPase V0 sector; subunits c/c'
KOG00242 Z Kinesin-like protein
KOG00264 B Nucleosome remodeling factor; subunit CAF1/NURF55/MSI1
KOG00270 S WD40 repeat-containing protein
KOG00284 A Polyadenylation factor I complex; subunit PFS2
KOG00304 A mRNA deadenylase subunit
KOG00307 U Vesicle coat complex COPII; subunit SEC31
KOG00328 J Predicted ATP-dependent RNA helicase FAL1; involved in rRNA maturation; DE
KOG00331 A ATP-dependent RNA helicase
KOG00335 A ATP-dependent RNA helicase
KOG00355 B DNA topoisomerase type II
KOG00358 O Chaperonin complex component; TCP-1 delta subunit (CCT4)
KOG00360 O Chaperonin complex component; TCP-1 alpha subunit (CCT1)
KOG00361 O Chaperonin complex component; TCP-1 eta subunit (CCT7)
KOG00362 O Chaperonin complex component; TCP-1 theta subunit (CCT8)
KOG00366 O Protein geranylgeranyltransferase type II; beta subunit
KOG00371 T Serine/threonine protein phosphatase 2A; catalytic subunit
KOG00374 TR Serine/threonine specific protein phosphatase PP1; catalytic subunit
KOG00397 J 60S ribosomal protein L11
KOG00417 O Ubiquitin-protein ligase
KOG00425 O Ubiquitin-protein ligase
KOG00446 UR Vacuolar sorting protein VPS1; dynamin; and related proteins
KOG00469 J Elongation factor 2
KOG00481 L DNA replication licensing factor; MCM5 component
KOG00531 T Protein phosphatase 1; regulatory subunit; and related proteins
KOG00580 D Serine/threonine protein kinase
KOG00583 T Serine/threonine protein kinase
KOG00594 R Protein kinase PCTAIRE and related kinases
KOG00657 G Glyceraldehyde 3-phosphate dehydrogenase
KOG00676 Z Actin and related proteins
KOG00706 T Predicted GTPase-activating protein
KOG00712 O Molecular chaperone (DnaJ superfamily)

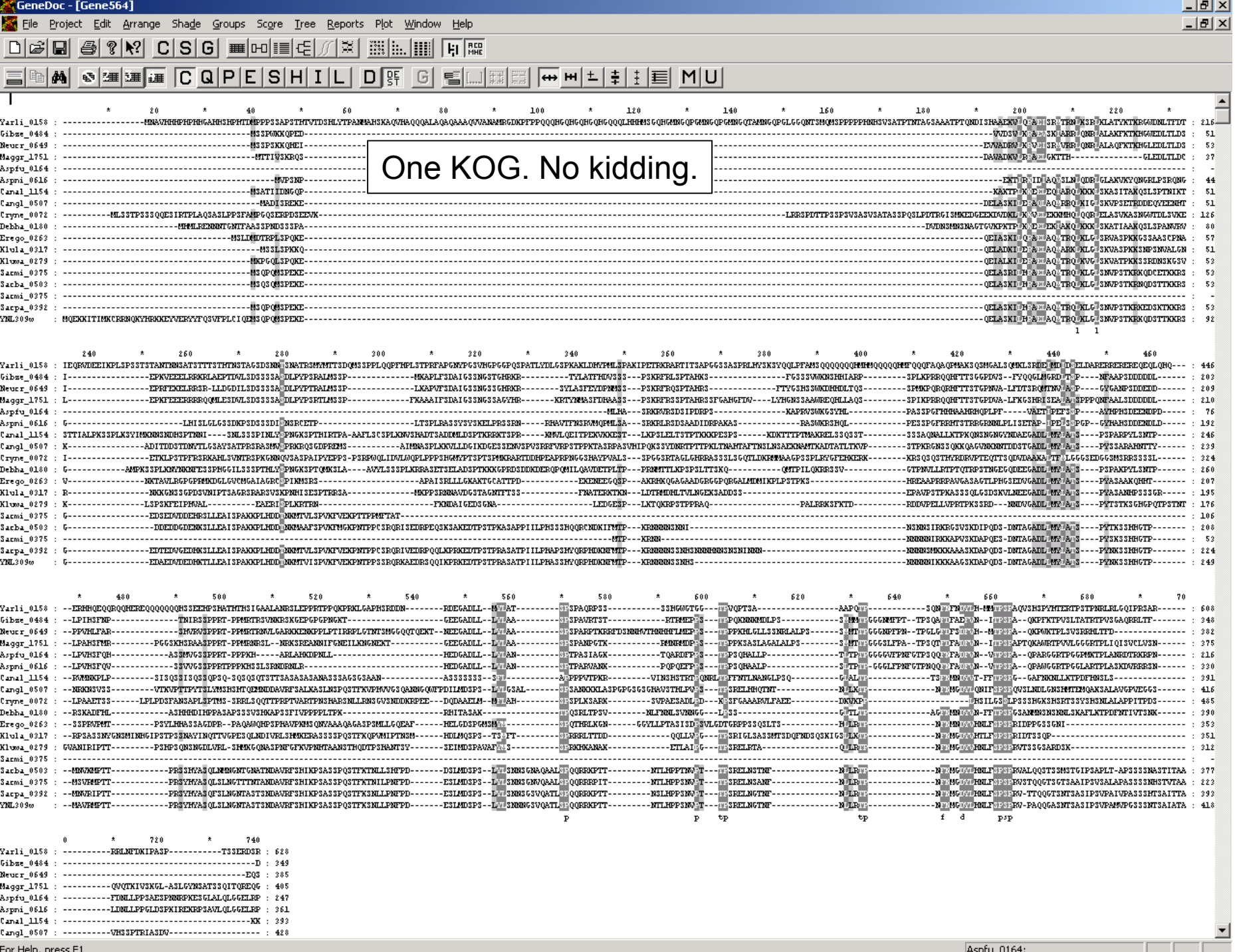
Because it is really hard to believe that some organisms do not have conventional ribosomal proteins we have to assume that analysis of phylogenetic profiles is the way to detect potentially missed genes

Giardia lamblia, Plasmodium falciparum, Plasmodium bergeri, Cryptosporidium parvum, Encephalitozoon cuniculi, Leishmanija major, and *Trypanosoma cruzi* share only 413 KOGs:

```
     Functional categories in KOG db
  102 Translation
   64 Posttranslational modification, protein turnover, chaperones
   43 RNA processing and modification
   31 Replication
   30 General function prediction only
   26 Transcription
   24 Intracellular trafficking, secretion
   15 Signal transduction mechanisms
   12 Energy production and conversion
   11 Unknown function
   11 Chromatin structure and dynamics
    9 Cytoskeleton
    7 Cell cycle control, cell division, chromosome partitioning
    6 Lipid transport and metabolism
    5 Amino acid transport and metabolism
    5 Nucleotide transport and metabolism
    3 Cell wall/membrane/envelope biogenesis
    2 Coenzyme transport and metabolism
    1 Inorganic ion transport and metabolism
    1 Secondary metabolites biosynthesis, transport and catabolism
```

Detection of complementary phylogenetic profiles in related KOGs will allow to make more aggressive hypotheses about orthology relationships and thus predict functions for greater number of proteins


GeneDoc - [Gene564]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
This is a fungal KOG. Although there are two groups of sequences that barely hit each other in BLAST they can be aligned AND have complementary phylogenetic profiles


GeneDoc - [Gene564]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
One KOG. No kidding.
Yarli_0158
Gibze_0484
Meurr_0649
Maggr_1751
Aspfu_0164
Aspni_0616
Canal_1154
Cangl_0507
Cryne_0072
Debha_0180
Erego_0263
Klula_0317
Kluwa_0279
Sacmi_0375
Sacba_0503
Sarpa_0392
YNL309w

Current status of KOG db: 50 Eukaryotic species; 46 species KOGnitorized, 1 updated;

712,427 protein sequences; two versions are present for 5 proteomes (+ 123,981 sequences)
21,816 clusters: 18,628 KOGs, 3,188 LSEs+mixed clusters

species name		Taxa	sequences	In KOGs/LSEs	%	KOGs	sequence data published	source
Chlamydomonas reinhardtii		Chlorophyta (green algae)	19,922	10,194	51%	4,653	No	JGI
Cyanidioschyzon merolae		Rhodophyta (red algae)	5,040	3,769	74%	2,735		JGI
Dictyostelium discoideum		Mycetozoa	13,677	9,162	67%	4,573		GB
Encephalitozoon cuniculi	OLD	Microsporidia	2,003	1,459	73%	1,132		GB
Entamoeba histolytica		Entamoebidae	10,088	7,728	77%	2,930		TIGR
Giardia lamblia		Diplomonadida group	6,573	2,105	32%	1,242	No	GB
Leishmania major		Euglenozoa	8,071	7,558	93%	4,502		TIGR
Trypanosoma brucei		Euglenozoa	8,291					GB
Trypanosoma cruzi		Euglenozoa	19,666	16,844	86%	5,023		GB
Thalassiosira pseudonana		Bacillariophyta (diatoms)	11,406	9,311	81%	4,033		JGI
Cryptosporidium parvum		Apicomplexa	3,410	2,212	66%	1,780		GB
Plasmodium berghei		Apicomplexa	11,782	7,004	59%	3,461	No	GB
Plasmodium falciparum		Apicomplexa	5,279	3,365	64%	3,403		GB
Plasmodium yoelii		Apicomplexa	7,868					GB
Theileria parva		Apicomplexa	4,080					GB
Arabidopsis thaliana	NEW	Viridiplantae	26,759	25,781	96%	6,953		GB
Arabidopsis thaliana	OLD	Viridiplantae	27,138	25,410	94%	6,925		GB
Oryza sativa		Viridiplantae	57,272	38,260	67%	6,701		TIGR
Aspergillus fumigatus		Fungi; Ascomycota; Pezizomycotina	10,029	8,959	89%	5,232	No	GB
Aspergillus nidulans		Fungi; Ascomycota; Pezizomycotina	10,317	9,629	93%	5,283	No	GB
Gibberella zeae		Fungi; Ascomycota; Pezizomycotina	12,221	11,153	91%	5,910	No	GB
Magnaporthe grisea		Fungi; Ascomycota; Pezizomycotina	11,381	9,043	79%	5,397		GB
Neurospora crassa		Fungi; Ascomycota; Pezizomycotina	10,478	8,206	78%	5,559		GB
Candida albicans		Fungi; Ascomycota; Saccharomycotina	14,274	10,952	77%	4,094		GB
Candida glabrata		Fungi; Ascomycota; Saccharomycotina	5,216	4,992	96%	3,790	No	GB
Debaryomyces hansenii		Fungi; Ascomycota; Saccharomycotina	6,350	5,700	90%	4,048		GB
Eremothecium gossypii		Fungi; Ascomycota; Saccharomycotina	4,757	4,619	97%	3,810	No	GB
Kluyveromyces lactis		Fungi; Ascomycota; Saccharomycotina	5,366	5,058	94%	3,980		GB
Kluyveromyces waltii		Fungi; Ascomycota; Saccharomycotina	5,240	5,055	96%	3,919	No	MIT
Yarrowia lipolytica		Fungi; Ascomycota; Saccharomycotina	6,596	5,728	87%	4,018		GB
Saccharomyces bayanus		Fungi; Ascomycota; Saccharomycotina	12,040	8,488	70%	5,879		MIT
Saccharomyces cerevisiae	OLD	Fungi; Ascomycota; Saccharomycotina	6,387	6,146	96%	4,427		GB
Saccharomyces mikatae		Fungi; Ascomycota; Saccharomycotina	10,350	8,481	82%	5,828		MIT
Saccharomyces paradoxus		Fungi; Ascomycota; Saccharomycotina	10,601	8,375	79%	6,005		MIT
Schizosaccharomyces pombe	OLD	Fungi; Ascomycota; Schizosaccharomycetes	5,064	4,461	88%	3,360		GB
Cryptococcus neoformans		Fungi; Basidiomycota; Hymenomycetes	6,640	5,471	82%	3,702	No	GB
Ustilago maydis		Fungi; Basidiomycota; Ustilaginomycetes	6,685	5,221	78%	3,779	No	GB
Anopheles gambiae		Metazoa; Insecta; Diptera; Nematocera	13,895	12,405	89%	5,946		GB
Apis mellifera		Metazoa; Insecta; Hymenoptera	16,997	14,785	87%	4,906	No	EMBL
Drosophila melanogaster	NEW	Metazoa; Insecta; Diptera; Brachycera	18,873					GB
Drosophila melanogaster	OLD	Metazoa; Insecta; Diptera; Brachycera	13,936	11,552	83%	6,241		GB
Caenorhabditis briggsae		Metazoa; Nematoda	19,523	17,963	92%	6,938		GB
Caenorhabditis elegans	NEW	Metazoa; Nematoda	21,193					GB
Caenorhabditis elegans	OLD	Metazoa; Nematoda	21,246	20,461	96%	7,175		GB
Ciona intestinalis		Metazoa; Chordata; Ascidiacea	16,139	12,101	75%	5,684		JGI
Danio rerio		Vertebrata; Teleostomi	19,893	19,676	98%	4,426		GB
Gallus gallus		Vertebrata; Aves	28,751	26,613	93%	6,894	No	EMBL
Homo sapiens	NEW	Vertebrata; Mammalia	27,524					GB
Homo sapiens	OLD	Vertebrata; Mammalia	39,436	29,835	77%	9,304		GB
Mus musculus		Vertebrata; Mammalia	26,464	24,784	94%	9,284		GB
Pan troglodytes	GB	Vertebrata; Mammalia	22,225					GB
Pan troglodytes	EMBL	Vertebrata; Mammalia	38,845	36,984	95%	8,923		EMBL
Rattus norvegicus		Vertebrata; Mammalia	21,531					GB
Takifugu rubripes		Vertebrata; Teleostomi	33,357	31,879	96%	6,909		GB
Tetraodon nigroviridis		Vertebrata; Teleostomi	28,262	25,520	90%	7,033	No	GB

Some conclusions

- Clusters of orthology domains (KOGs) will form a platform for comparative genomics and evolutionary studies
- At present KOG db cannot be updated completely automatically
- A new version of eukaryotic clusters of orthology domains is being developed; it will contain ca. 25,000 clusters for 50 proteomes and HMM profiles for all clusters
- An approach to add proteomes semi-automatically is developed which is based on picking conflicts between BLASTP (KOGnitor) and HMM searches
- KOG db will serve as a tool for annotation of newly sequenced genomes, for fixing errors and filling in gaps in existing annotations, for detecting incorrect gene models and genes missed by gene predictions

Acknowledgements

Sergei Smirnov, NCBI

Yuri Wolf , NCBI

Alex Sorokin , NCBI

Eugene Koonin , NCBI

ERROR - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?pid=29246175 Go Links

Google Search 195 blocked Check AutoLink AutoFill Options

Database connection error. Please try again later

Done Internet