D190–D195 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 9 November 2009*
doi:10.1093/nar/gkp951

eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations

J. Muller[1], D. Szklarczyk[1,2], P. Julien[3], I. Letunic[1], A. Roth[4], M. Kuhn[1], S. Powell[1], C. von Mering[4], T. Doerks[1], L. J. Jensen[2] and P. Bork[1,5,*]

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark, [3]The Center for Integrative Genomics, University of Lausanne, Lausanne, [4]University of Zurich and Swiss Institute of Bioinformatics, Winterthurerstrasse 190, 8057 Zurich, Switzerland and [5]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, 13092 Berlin, Germany

Received September 15, 2009; Revised October 9, 2009; Accepted October 12, 2009

ABSTRACT

The identification of orthologous relationships forms the basis for most comparative genomics studies. Here, we present the second version of the eggNOG database, which contains orthologous groups (OGs) constructed through identification of reciprocal best BLAST matches and triangular linkage clustering. We applied this procedure to 630 complete genomes (529 bacteria, 46 archaea and 55 eukaryotes), which is a 2-fold increase relative to the previous version. The pipeline yielded 224 847 OGs, including 9724 extended versions of the original COG and KOG. We computed OGs for different levels of the tree of life; in addition to the species groups included in our first release (i.e. fungi, metazoa, insects, vertebrates and mammals), we have now constructed OGs for archaea, fishes, rodents and primates. We automatically annotate the non-supervised orthologous groups (NOGs) with functional descriptions, protein domains, and functional categories as defined initially for the COG/KOG database. In-depth analysis is facilitated by precomputed high-quality multiple sequence alignments and maximum-likelihood trees for each of the available OGs. Altogether, eggNOG covers 2 242 035 proteins (built from 2 590 259 proteins) and provides a broad functional description for at least 1 966 709 (88%) of them. Users can access the complete set of orthologous groups via a web interface at: http://eggnog.embl.de.

INTRODUCTION

Next-generation sequencing technologies are now generating a vast amount of sequence data. This leads to a dramatic increase in the number of predicted protein sequences, which serve as a starting point for structural, functional and phylogenomic studies. In such studies, high-throughput comparative analyses are often required to transfer information between organisms, for which the concept of orthology is crucial. The original definition by Fitch (1) describes orthologs as genes that diverged through a speciation event, as opposed to paralogs, which diverged after a duplication event. This has been extended and refined by introducing the concepts of orthologous groups (OGs) (2), in-paralogs and out-paralogs (3,4). In practice, however, the identification and classification of homologous genes remain very difficult and rely on operational definitions. An enormous effort is being put into the development of different approaches to establish orthologous relationships between genes from different genomes. This includes several algorithms using the simple graph-based methods, including reciprocal-best-hit approach (5), identification of best-hit triangles (2,6–8) and clustering-based approaches (9–11) as well as tree-based methods (12–16).

In addition to the quality of the grouping of genes, the practical usability of OGs is determined by the ability to provide a robust functional annotation. Thus, newer projects not only aggregate orthology information from various sources to allow comparison between methods but also aim to provide annotation tools (17,18). Nevertheless, evolutionary genealogy of genes: non-supervised OGs (eggNOG) (19) and the COG/KOG/arCOG resources (2,6,7) are still the only databases

*To whom correspondence should be addressed. Tel: +49 6221 3878526; Fax: +49 6221 387519; Email: bork@embl.de
Present address:
M. Kuhn, Biotechnology Center, TU Dresden, 01062 Dresden, Germany.

providing explicit functional annotations for the OGs at different hierarchical levels, whereby the COG/KOG resource is based on a robust manual expert annotation, which eggNOG is using and automatically extending (19).

Here, we describe the new features of the second version of eggNOG, a resource that provides OGs from the three domains of life at several levels of resolution. eggNOG v2 contains twice as many species and proteins as the previous version, additional hierarchical levels allowing higher resolution for a number of taxonomic groups, new annotation sources and an extended interface for an in-depth analysis of orthologous relationships.

CONSTRUCTION OF HIERARCHICAL OGs

The automated procedure described previously (19) has been used to assemble proteins into OGs from 630 complete genomes (529 bacteria, 46 archaea and 55 eukaryotes). Complete proteomes were downloaded from the RefSeq (20), Ensembl (21), GiardiaDB (22) or TAIR (23) databases. This particular data set also forms the basis for STRING v8 (24) and STITCH v2 (25), allowing for easy integration across these databases.

Altogether, the protein data set covers 2 590 259 proteins of which 2 242 035 (87%) were included in at least one of 224 847 OGs generated by eggNOG. The growing number of species and proteins included in this release drastically increased the computational time. All-against-all similarity searches have therefore been performed using Basic Local Alignment Search Tool (BLAST) (26) instead of the Smith–Waterman algorithm (27).

Compared to the 4873 COGs and the 4850 KOGs that are constructed across all three domains of life and for all eukaryotes, respectively, this procedure assembles additional proteins into NOGs (440 359 proteins into 59 497 NOGs and 181 427 into 17 845 euNOGs). These complement the published COGs and KOGs built respectively for 66 and seven species (6), which are extended in eggNOG to cover 630 species encompassing, respectively, 1 547 381 and 483 043 proteins.

To provide a higher resolution of OGs in frequently used taxonomic groupings, we applied our procedure to several subsets of organisms separately. We updated the previously computed more fine-grained NOGs at the level of fungi (fuNOGs), metazoans (meNOGs), insects (inNOGs), vertebrates (veNOGs) and mammals (maNOGs) and added groups for archaea (arNOGs), fishes (fiNOGs), rodents (roNOGs) and primates (prNOGs).

Extending the automated annotation of protein function

An important feature of eggNOG is the functional annotations of the OGs. Our original pipeline, providing functional descriptions for the NOGs, is now complemented by an automatic inference of functional categories (FCs) which were taken from the COG database (2). The 25 FCs available from the COG resource have been widely used to assess comparative genomics studies and will enable higher-order analyses of OGs identified in any data set.

We use two complementary methods to infer FCs of OGs based on the 4617 COGs (used for NOGs and arNOGs) and 4381 KOGs (used for all other OGs). The first method uses Support Vector Machines (SVM) trained on the COGs and KOGs to classify NOGs into the 25 FCs based on feature vectors. Two feature vectors were created for each OG. One was built from functional information mapped onto the eggNOG protein data set, including KEGG pathways and modules (28), GO terms (29), SMART domains (30), PFAM domains (31), UniProt keywords (32) and words from UniProt/RefSeq (20) description lines. The second feature vector includes also words from MEDLINE abstracts referring to a particular protein (24). Each attribute in the feature vector encodes the fraction of proteins in the group having the feature in question.

The second method for assigning FCs makes use of the hierarchical structure of eggNOG, namely that the same proteins can be assigned to OGs at several levels in the tree of life (e.g. a KOG and a meNOG). In case an FC could not be assigned to a NOG by the SVM method, we check if most of the proteins in the NOG belong to a common functionally annotated COG or KOG, in which case we transfer the FCs from the coarse-grained level (COGs or KOGs) to the more fine-grained one (e.g. arNOGs or meNOGs). The assignment of an FC to a single NOG is achieved on the basis of a coverage value determined by the occurrence of that FC (via the proteins shared with the reference level) in respect to the total number of proteins in that NOG.

ANNOTATION RESULTS

In addition to providing functional annotations via description lines for many NOGs (19), we are now able to predict functional categories as well. At the universal level, our function annotation pipeline provides

Table 1. Annotation statistics at different taxonomic levels

Level	OG count	Description line		Functional categories	
		Annotated	(%)	Annotated	(%)
COG + NOG	64 370	4474 + 14 956	30.2	2824 + 6262	14.1
arNOG	9809	4144	42.2	4540	46.3
KOG + euNOG	22 695	4288 + 7566	52.2	3514 + 4120	33.6
fuNOG	9976	5661	56.7	5775	57.9
meNOG	22 691	16 636	73.3	13 490	59.5
inNOG	8049	5034	62.5	5810	72.2
veNOG	21 357	16 722	78.3	13 291	62.2
fiNOG	13 674	8903	65.1	9580	70.1
maNOG	20 222	16 959	83.9	13 075	64.7
roNOG	14 038	11 918	84.9	10 547	75.1
prNOG	17 966	14 773	82.2	13 124	73.0

At the levels for COGs (universal) and KOGs (eukaryotes) the additional automatically generated non-supervised orthologous groups NOGs and euNOGs, respectively, are separated.

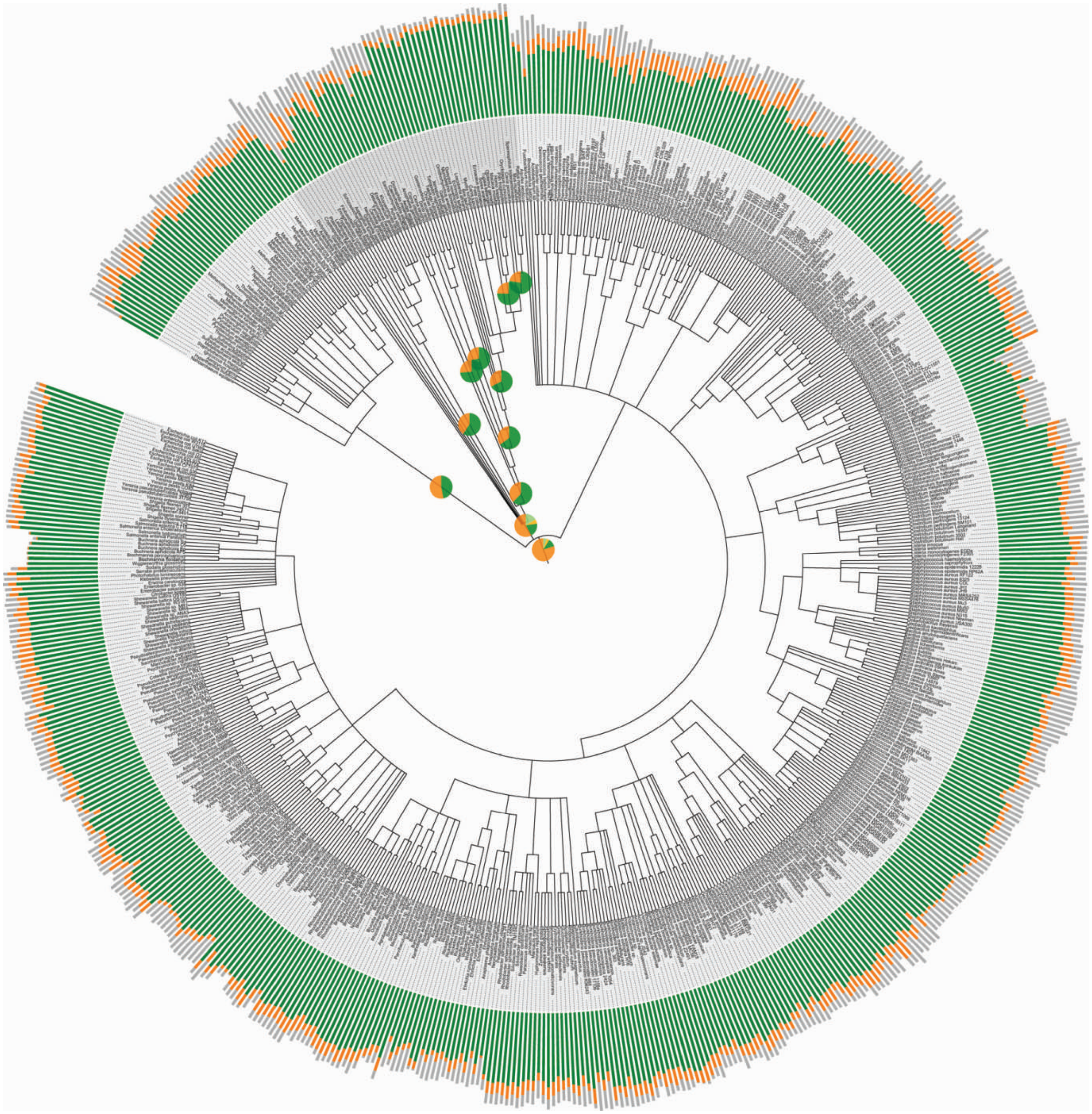
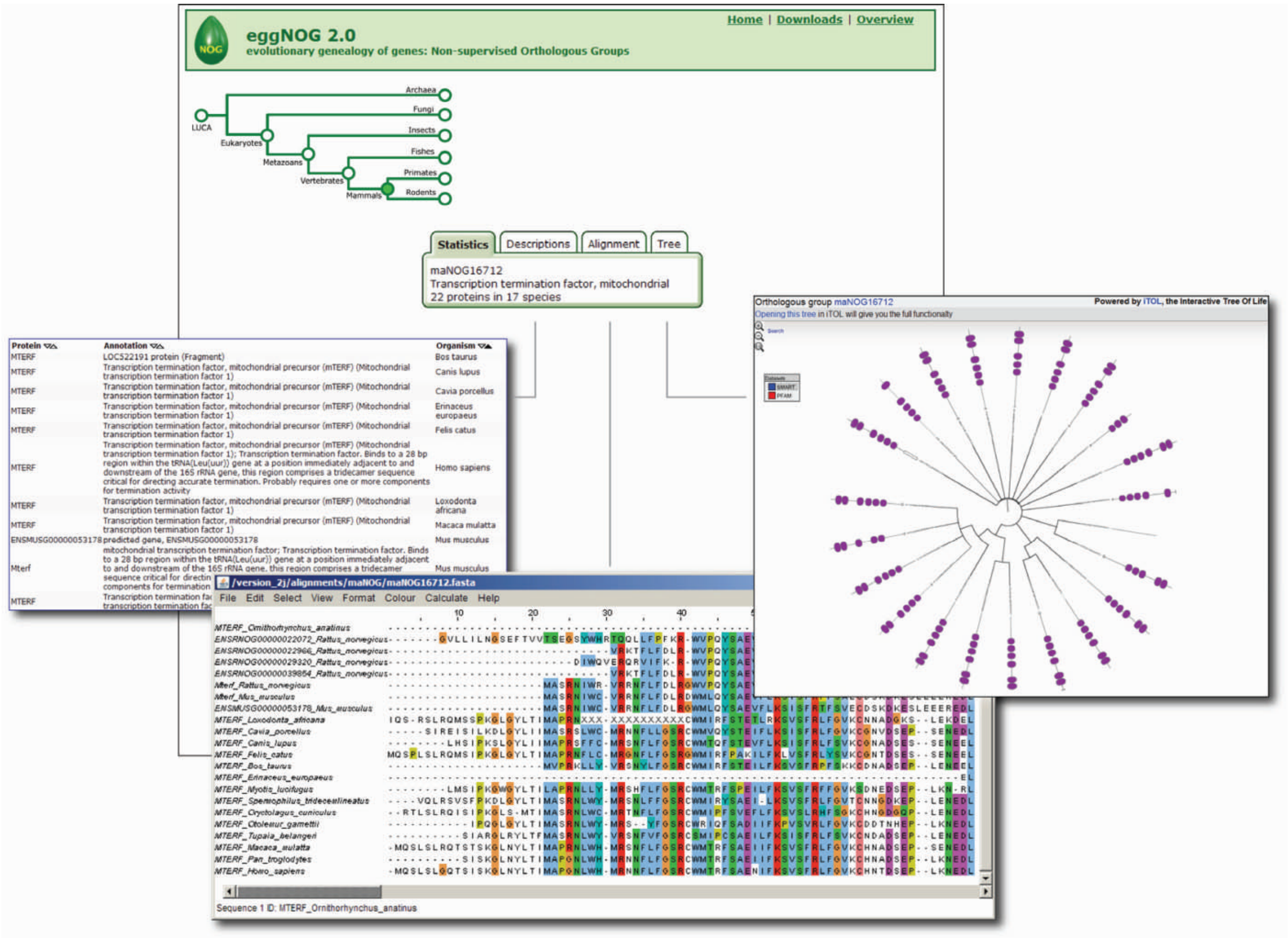


Figure 1. Statistics on the content of the eggNOG database. The eggNOG assignments for 630 complete genomes were mapped onto the tree of life. The stacked bar charts outside the tree show the proportion of genes from each genome that can be assigned to a functionally annotated orthologous group (green), an unannotated orthologous group (orange) or no orthologous group (gray). The length of each bar is proportional to the logarithm of the number of genes in the respective genome. The pie charts inside the tree show the fractions of orthologous groups at each level in the hierarchy that could be annotated with a functional category (green for NOGs, light green for extended COGs and KOGs) or not (orange for NOGs, light orange for extended COGs and KOGs). An interactive version is available in the 'Overview' section at: http://eggnog.embl.de. This figure was made using iTOL.

description lines for 14 956 (25%) and an FC for 6262 (11%) of the 59 497 coarse-grained NOGs. At the eukaryotic level, 7566 euNOGs (52%) have a description line and 4120 (34%) have an FC. In addition, eggNOG contains 137 782 more fine-grained OGs of which 100 750 (73%) and 89 232 (65%) have been annotated with a description line and an FC, respectively (Table 1).

This enables us to assign 2 242 035 of the 2 590 259 genes (87% of the genes in the analyzed genomes) to an OG and to provide at least a broad functional description or FC for 1 966 709 of them (78% of the genes that could be assigned to an OG). The corresponding numbers for each set of OGs as well as for each individual genome are summarized in Figure 1.



maNOG16712
Transcription termination factor, mitochondrial
22 proteins in 17 species

Opening this tree in iTOL will give you the full functionality

Protein	Annotation	Organism
MTERF	LOC522191 protein (Fragment)	Bos taurus
MTERF	Transcription termination factor, mitochondrial precursor (mTERF) (Mitochondrial transcription termination factor 1)	Canis lupus
MTERF	Transcription termination factor, mitochondrial precursor (mTERF) (Mitochondrial transcription termination factor 1)	Cavia porcellus
MTERF	Transcription termination factor, mitochondrial precursor (mTERF) (Mitochondrial transcription termination factor 1)	Erinaceus europaeus
MTERF	Transcription termination factor, mitochondrial precursor (mTERF) (Mitochondrial transcription termination factor 1)	Felis catus
MTERF	Transcription termination factor, mitochondrial precursor (mTERF) (Mitochondrial transcription termination factor 1); Transcription termination factor. Binds to a 28 bp region within the tRNA(Leu(uur)) gene at a position immediately adjacent to and downstream of the 16S rRNA gene, this region comprises a tridecamer sequence critical for directing accurate termination. Probably requires one or more components for termination activity	Homo sapiens
MTERF	Transcription termination factor, mitochondrial precursor (mTERF) (Mitochondrial transcription termination factor 1)	Loxodonta africana
MTERF	Transcription termination factor, mitochondrial precursor (mTERF) (Mitochondrial transcription termination factor 1)	Macaca mulatta
ENSMUSG00000053178	predicted gene, ENSMUSG00000053178	Mus musculus
Mterf	mitochondrial transcription termination factor; Transcription termination factor. Binds to a 28 bp region within the tRNA(Leu(uur)) gene at a position immediately adjacent to and downstream of the 16S rRNA gene, this region comprises a tridecamer sequence critical for directin components for termination	Mus musculus
MTERF	Transcription termination fac transcription termination fac	

/version_2j/alignments/maNOG/maNOG16712.fasta

```
MTERF_Ornithorhynchus_anatinus              -------------------------------------------------
ENSRNOG00000022072_Rattus_norvegicus        -------GVLLILNGSEFTVVTSEGSYWHRTQQLLFPFKR-WVPQYSAE
ENSRNOG00000022966_Rattus_norvegicus        ------------------------------VRKTFLFDLR-WVPQYSAE
ENSRNOG00000029320_Rattus_norvegicus        ------------------------DIWQVERQRVIFK-R-WVPQYSAE
ENSRNOG00000039854_Rattus_norvegicus        ------------------------------VRKTFLFDLR-WVPQYSAE
Mterf_Rattus_norvegicus                     ---------------------MASRNIWR-VRRNFLFDLRGWVPQYSAE
Mterf_Mus_musculus                          ---------------------MASRNIWC-VRRNFLFDLRDWMLQYSAE
ENSMUSG00000053178_Mus_musculus             ---------------------MASRNIWC-VRRNFLFDLRGWMLQYSAEVFLKSISFRTFSVECDSKDKESLEEEREDL
MTERF_Loxodonta_africana                    IQS-RSLRQMSSPKGLGYLTIMAPRNXXX-XXXXXXXXXXCWMIRFSTETLRKSVSFRLFGVKCNNADGKS--LEKDEL
MTERF_Cavia_porcellus                       -----SIREISILKDLGYLIIMASRSLWC-MRNNFLLGSRCWMVQYSTEIFLKSISFRLFGVKCGNVDSEP--SENEDL
MTERF_Canis_lupus                           --------LHSIPKSLGYLIIMAPRSFFC-MRSNFLFGSRCWMTQFSTEVFLKSISFRLFSVKCGNADSES--SENEEL
MTERF_Felis_catus                           MQSPLSLRQMSIPKGLGYLTIMAPRNFLC-MRGNFLFGSRGWMIRFPAKILFKLVSFRLYSVKCGNTDSES--SENEEL
MTERF_Bos_taurus                            ---------------------MVPRKLLY-VRSNYLFGSRCWMIRFSTEILFKSVSFRPFSKKCDNADSEP--LENEEL
MTERF_Erinaceus_europaeus                   -----------------------------------------------------------------------------EL
MTERF_Myotis_lucifugus                      --------LMSIPKGWGYLTILAPRNLLY-MRSHFLFGSRCWMTRFSPEILFKSVSFRFFGVKSDNEDSEP--LKN-RL
MTERF_Spermophilus_tridecemlineatus         ----VQLRSVSFPKDLGYLTIMASRNLWY-MRSNLFFGSRCWMIRYSAEI-LKSVSFRLFGVTCNNGDKEP--LENEDL
MTERF_Oryctolagus_cuniculus                 --RTLSLRQISIPKGLS-MTIMASRNLWC-MRTNFLFGSRCWMIPFSVEFLFKSVSLRHFSGKCHNGDGQP--LENEDL
MTERF_Otolemur_garnettii                    -----------IPQGLGYLTIMASRNLWY-MRS--YFGSRCWRIQFSADIIFKPVSVRLFGVKCDDTNHEP--LKNEDL
MTERF_Tupaia_belangeri                      ----------SIARGLRYLTFMASRNLWY-VRSNFVFGSRCSMIPCSAEILFKSISFRLFSVKCNDADSEP--LENEDL
MTERF_Macaca_mulatta                        -MQSLSLRQTSTSKGLNYLTIMAPRNLWH-MRSNFLFGSRCWMTRFSAEIIFKSVSFRLFGVKCHNADSEP--SENEDL
MTERF_Pan_troglodytes                       ----------SISKGLNYLTIMAPGNLWH-MRNNFLFGSRCWMTRFSAEIIFKSVSFRLFGVKCHNADSEP--LKNEDL
MTERF_Homo_sapiens                          -MQSLSLGQTSISKGLNYLTIMAPGNLWH-MRNNFLFGSRCWMTRFSAENIFKSVSFRLFGVKCHNTDSEP--LKNEDL
```

Sequence 1 ID: MTERF_Ornithorhynchus_anatinus

Figure 2. Screenshot of the detailed results page. The eggNOG database was queried for the term 'mTERF', the mitochondrial precursor of the transcription termination factor 1. The navigation tree at the top of the page allows the user to change the view to more coarse-grained orthologous groups, for example, the mammalian orthologous groups. The tab menu, shown here, enables several in-depth interactions with the new data (i.e. MSA or phylogenetic trees, here displayed with SMART domains).

Extended features in eggNOG v2.0

To facilitate the in-depth analysis of the orthologous relationships within the groups of proteins, we now provide precomputed high-quality Multiple Sequence Alignments (MSAs) and maximum-likelihood trees via the web interface (Figure 2).

Numerous methods are available to build MSAs [e.g. ClustalW (33), Muscle (34), MAFFT (35) and PRANK (36)] but some programs appear to be more suitable for particular protein families than others (37). Thus, we applied a new approach, named Automated QUality improvement for multiple sequence Alignments (AQUA) (Muller *et al.*, submitted for publication), which combines existing tools to deliver high-quality MSAs.

The construction of the different phylogenetic trees was carried out using the following steps. One hundred bootstrap replicates were created from the MSA using the SEQBOOT program from the Phylip package (38). Following this, PhyML (39) was used to find the maximum-likelihood tree for each of the 100 bootstrap replicates and for the original alignment using default parameters. Finally, a consensus tree was constructed, using the CONSENSE program from the Phylip package. We used ReadSeq (40) to convert between the different sequence file formats used by those programs.

ACCESS OPTIONS

The eggNOG resource can be queried via a web interface; data can be downloaded under the Creative Commons Attribution 3.0 License at: http://eggnog.embl.de or via FTP at: ftp://eggnog.embl.de/eggNOG/2.0/. Gene and protein names, database identifiers, amino acid sequences, or OG names can be used to query the database. As a default, the most fine-grained OGs available are displayed for maximal resolution. The user can navigate among the different levels of orthology using an available guide-tree of organisms to find the desired balance between phylogenetic coverage and functional specificity within our hierarchy of OGs. Through the new interface, users can access different information panels encompassing the detailed list of proteins belonging to a particular OG as well as the corresponding MSA and phylogenetic tree. The MSA can be interactively displayed using the Jalview applet (41) or downloaded in FASTA format. The phylogenetic trees are accessed through a dedicated iTOL (42) viewer together with mapped PFAM and SMART domains, via the ATV program applet (43), or can be downloaded in Newick format.

CONCLUSIONS/PERSPECTIVES

With 630 genomes covered, an increased OG hierarchy, and a high coverage of newly categorized functional annotation, the new version of eggNOG is one of the most comprehensive and complete resources for deciphering the orthologous relationships between proteins from various species. The changes and improvements in the interface and the availability of the OGs for download will not only facilitate the daily use of the database, but also the integration of eggNOG in high-throughput comparative genomics studies. Our future plans include the addition of more complete genomes and development of a more scalable and flexible pipeline for generating the groups.

FUNDING

EMBL, the European Commission Programme, Eurasnet EU [Grant LSHG-CT-2005-518238 (FP6), IMPACT 213037 (FP7)]; the Novo Nordisk Foundation Center for Protein Research, the Swiss Institute of Bioinformatics; and the University of Zurich (partial, through its Research Priority Program 'Systems Biology and Functional Genomics'). This work was supported in part by the Bundesministerium fuer Bildung und Forschung (Nationales Genomforschungsnetz Foerderkennzeichen 01GS08169). Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
3. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
4. Sonnhammer,E.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
5. Berglund,A.C., Sjolund,E., Ostlund,G. and Sonnhammer,E.L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.*, **36**, D263–D266.
6. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
7. Makarova,K.S., Sorokin,A.V., Novichkov,P.S., Wolf,Y.I. and Koonin,E.V. (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct.*, **2**, 33.
8. Kriventseva,E.V., Rahman,N., Espinosa,O. and Zdobnov,E.M. (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.*, **36**, D271–D275.
9. Roth,A.C., Gonnet,G.H. and Dessimoz,C. (2008) Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, **9**, 518.
10. Li,L., Stoeckert,C.J. Jr. and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
11. Uchiyama,I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res.*, D343–D346.
12. van der Heijden,R.T., Snel,B., van Noort,V. and Huynen,M.A. (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*, **8**, 83.
13. Wapinski,I., Pfeffer,A., Friedman,N. and Regev,A. (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, **23**, i549–i558.
14. Huerta-Cepas,J., Bueno,A., Dopazo,J. and Gabaldon,T. (2008) PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.*, **36**, D491–D496.
15. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
16. Datta,R.S., Meacham,C., Samad,B., Neyer,C. and Sjolander,K. (2009) Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res.*, W84–W89.
17. Eyre,T.A., Wright,M.W., Lush,M.J. and Bruford,E.A. (2007) HCOP: a searchable database of human orthology predictions. *Brief Bioinform.*, **8**, 2–5.
18. Kuzniar,A., Lin,K., He,Y., Nijveen,H., Pongor,S. and Leunissen,J.A. (2009) ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Res.*, **37**, W428–W434.
19. Jensen,L.J., Julien,P., Kuhn,M., von Mering,C., Muller,J., Doerks,T. and Bork,P. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, **36**, D250–D254.
20. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
21. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
22. Aurrecoechea,C., Brestelli,J., Brunk,B.P., Carlton,J.M., Dommer,J., Fischer,S., Gajria,B., Gao,X., Gingle,A., Grant,G. *et al.* (2009) GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis. *Nucleic Acids Res.*, **37**, D526–D530.
23. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
24. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
25. Kuhn,M., von Mering,C., Campillos,M., Jensen,L.J. and Bork,P. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.
26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
27. Saebo,P.E., Andersen,S.M., Myrseth,J., Laerdahl,J.K. and Rognes,T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.*, **33**, W535–W539.
28. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

29. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
30. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
31. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
32. The Universal Protein Resource (UniProt). (2009) *Nucleic Acids Res.*, **37**, D169–D174.
33. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
34. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
35. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.
36. Loytynoja,A. and Goldman,N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635.
37. Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.
38. Felsenstein,J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
39. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
40. Gilbert,D. (2003) Sequence file format conversion with command-line readseq. *Curr. Protoc. Bioinformatics*, Appendix 1, Appendix 1E.
41. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
42. Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.
43. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.