Published online 30 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D263–D266***
doi:10.1093/nar/gkm1020

InParanoid 6: eukaryotic ortholog clusters with inparalogs

Ann-Charlotte Berglund[1], Erik Sjölund[2], Gabriel Östlund[2] and Erik L. L. Sonnhammer[2,*]

[1]Linnaeus Centre for Bioinformatics, Uppsala University, BMC Box 598, 75124, Uppsala and
[2]Stockholm Bioinformatics Center, Albanova, Stockholm University, SE-10691 Stockholm, Sweden

Received September 15, 2007; Revised October 23, 2007; Accepted October 27, 2007

ABSTRACT

The InParanoid eukaryotic ortholog database (http://InParanoid.sbc.su.se/) has been updated to version 6 and is now based on 35 species. We collected all available 'complete' eukaryotic proteomes and *Escherichia coli*, and calculated ortholog groups for all 595 species pairs using the InParanoid program. This resulted in 2 642 187 pairwise ortholog groups in total. The orthology-based species relations are presented in an ortho-phylogram. InParanoid clusters contain one or more orthologs from each of the two species. Multiple orthologs in the same species, i.e. inparalogs, result from gene duplications after the species divergence. A new InParanoid website has been developed which is optimized for speed both for users and for updating the system. The XML output format has been improved for efficient processing of the InParanoid ortholog clusters.

INTRODUCTION

Many analyses in comparative genomics depend on correct mapping of orthologs between species. Orthologs are defined as genes in different species deriving from a single gene in the last common ancestor (1), and are therefore likely to have the same function. If an ortholog undergoes duplication in one species, the copies are referred to as inparalogs (2). Inparalogs are by definition co-orthologs to one or more orthologs in another species. In contrast, two genes deriving from a duplication that predated the speciation event between the species are referred to as outparalogs. The InParanoid program was developed to identify clusters of inparalogs while avoiding inclusion of outparalogs.

InParanoid is one of the first comprehensive ortholog databases (3,4), but nowadays more than 15 different ortholog databases exist (5). A reason for the multitude of ortholog databases is that different research questions have different needs. For instance, the COGs database (6) contains very large clusters of orthologs that often contain outparalogs (7). At the other extreme, the Homologene database (8) often places inparalogs in different clusters. For some applications, one extreme or the other may be appropriate. However, the average user is normally interested in simply finding all orthologs in species Y to a gene in species X, including all inparalogs but excluding outparalogs. InParanoid was developed to optimally serve this type of user.

Two benchmarks have recently been published that try to objectively assess the quality of different ortholog databases (9,10). In both these tests, which look either at accuracy of functional annotation or at inferred accuracy, InParanoid was top ranked. This suggests that InParanoid is successful at balancing the false-negative and false-positive rate, and is appropriate as a general-purpose orthology tool.

The InParanoid program has been upgraded to version 2.0. This release contains a number of fixes to minor bugs that could lead to incorrect cluster merging or bootstrap values. These problems were however rare.

We here present InParanoid 6, comprising 34 eukaryotic species and one prokaryotic outgroup. The website has been completely reconstructed and has new front- and back-ends, yet looks very similar to the old site. The new design makes it much faster for the user, and allows easier updating of the system. With the new back-end, we can handle much larger datasets in the future without performance problems.

DATA AND IMPLEMENTATION

The data was gathered from three different sources: Ensembl, NCBI and model organism databases (MODs). We only considered eukaryotic genomes sequenced to a coverage greater than 6X, with <1% unknown amino acids (X in the protein sequences). Most MOD data was packaged and uploaded by the staffs at TAIR, WormBase, FlyBase, ZFIN, dictyBase, SGD and MGI to us directly, but three MODs were downloaded from their repositories. Before running InParanoid, each proteome was made non-redundant by keeping only the longest transcript from each gene. If this is not done first, different transcripts from the

*To whom correspondence should be addressed. Tel: +46 8 55378567; Fax: +46 8 55378214; Email: Erik.Sonnhammer@sbc.su.se

same gene can end up in different clusters if they exist in more than one species. Below we only list the non-redundant number of proteins for each species.

Nine of the proteomes were uploaded to us by MOD staff. Together, we have formed an informal consortium of MODs that want to cross-reference each other using orthologs from InParanoid. We particularly welcome this system as it allows us to use the most complete and recent set of proteins for each organism, and ensures that we use identifiers that work in the MODs so that web links to proteins are valid. We hope that more MODs will join in and provide their proteomes in a new and robust XML format that will be introduced for the next release.

From Ensembl, data was obtained for *Aedes aegypti* (transcripts for 15 419 genes), *Anopheles gambiae* (13 277), *Apis mellifera* (13 448), *Bos taurus* (22 280), *Canis familiaris* (19 314), *Ciona intestinalis* (14 278), *Gallus gallus* (16 715), *Gasterosteus aculeatus* (20 879), *Homo sapiens* (22 983), *Macaca mulatta* (22 045), *Monodelphis domestica* (19 597), *Pan troglodytes* (20 982), *Rattus norvegicus* (23 299), *Takifugu rubripes* (22 008), *Tetraodon nigroviridis* (28 005) and *Xenopus tropicalis* (18 473). *Apis mellifera* was taken from Ensembl release 38 and all other proteomes from release 43.

From NCBI, we obtained Candida glabrata (5192), Cryptococcus neoformans (6487), Debaromyces hansenii (6318), Entamoeba histolytica (9772), Escherichia coli K12 (4243), Entamoeba histolytica (9772), Kluyveromyces lactis (5336), Yarrowia lipolytica (6544).

The MODs uploaded proteomes for Arabidopsis thaliana (26 819), Caenorhabditis briggsae (19 334), Caenorhabditis elegans (20 084), Caenorhabditis remanei (25 595), Danio rerio, (12 303), Dictyostelium discoideum (13 523), Drosophila melanogaster (13 854), Mus musculus (23 132), Saccharomyces cerevisiae (5792). We obtained from other MODs Oryza sativa (77 853) (from http://www.gramene.org), Drosophila pseudoobscura (9871) (from http://www.flybase.org), and Schizosaccharomyces pombe (5003) (http://www.sanger.ac.uk).

InParanoid clustering

NCBI–Blast comparisons using these datasets were performed between each pair of species, involving four whole proteome runs per species pair (normal runs both ways plus two self-self runs). For the 35 proteomes this amounts to 595 species pairs, requiring 1225 whole-proteome Blast searches. These were executed on the SBC compute cluster comprising about 300 Linux nodes. The pairwise Blast results were used as the input for the InParanoid ortholog clustering procedure (3).

The output from InParanoid 6 is available as XML, SQL, HTML and native format for downloading at the InParanoid homepage, and is searchable via the web interface. The XML format was defined in the RELAX NG schema language.

INPARANOID CONTENT

The 35 species present in the InParanoid database result in 595 pairwise ortholog lists. The information in these lists was used to generate a phylogenetic tree that reflects the level of orthology between the different species. We calculated the orthology distance from species A to B, dAB, by

$$\frac{(\text{proteins\_A} - \text{proteins\_A\_orthologous\_to\_B})}{\text{proteins\_A}}$$

and used the average orthology distances (dAB + dBA)/2 to construct a UPGMA tree, shown in Figure 1. This so-called 'orthophylogram' shows quantitatively the level of orthology between different clades. In general, it agrees with the standard taxonomic species tree, but we noted a few exceptions. Opossum (*M. domesticus*), a marsupial mammal, is clustered together with placental mammals, and the zebrafish *D. rerio* clustered as an outgroup to the land animals rather than together with other fish. The latter anomaly is very minor as all fish are still neighbors in the tree, but the placement of opossum is surprising. If this placement is correct, then marsupials could have evolved from a particular lineage of placental mammals. Another difference is found in the yeast clade. In the taxonomy, *K. lactis*, *S. cerevisiae* and *D. hansenii* are clustered together, while *C. glabrata* is placed outside this group. These are arranged differently in the orthophylogram in that *C. glabrata* has traded place with *D. hansenii*, which now is placed as an outgroup to *K. lactis*, *S. cerevisiae* and *C. glabrata*. InParanoid's grouping is supported by 25S rDNA sequences (11). Surprisingly, the green plants are placed as a subgroup among single-cell organisms, next to the fungal group.

It is worth noting that on average only 91.2% of the proteins in *H. sapiens* and chimpanzee *P. troglodytes* are orthologous. The individual figures are 88.4% for human and 94% for chimpanzee. This is surprisingly low since the genome-wide nucleotide divergence between human and chimpanzee is estimated to only 1.23% (12). The much higher difference observed for orthologs is not due to unique proteins in either proteome, as the fraction of homologs reported by Inparanoid is 96.7% for human and 99.3% for chimpanzee. Rather, it reflects that the sequences were too divergent to be considered orthologs. This is, however, often caused by incomplete sequencing or errors in gene annotation.

The average number of inparalogs per cluster ranges from 1.001 (in *Drosophila pseudoobscura* when compared to *D. melanogaster*) to 7.160 (in *O. sativa* when compared to *D. rerio*). The overall mean number of inparalogs per species was 1.54, and the median was 1.25. The distribution of cluster sizes is shown in Figure 2. The highly duplicated genome of *O. sativa* is responsible for all average cluster sizes of four, and generates a separate peak in the distribution around five. In fact, *O. sativa* had on average more than four inparalogs per ortholog group when compared to every other non-plant species. It is surprising that the average number of inparalogs in *O. sativa* was so high when compared with *D. rerio*; when compared with *D. rerio*'s phylogenetic neighbors the number was only around five. Although the rice proteome clearly contains the largest number of genes, our figures are probably somewhat overestimated. Evidence for this is that we were not able to find shared

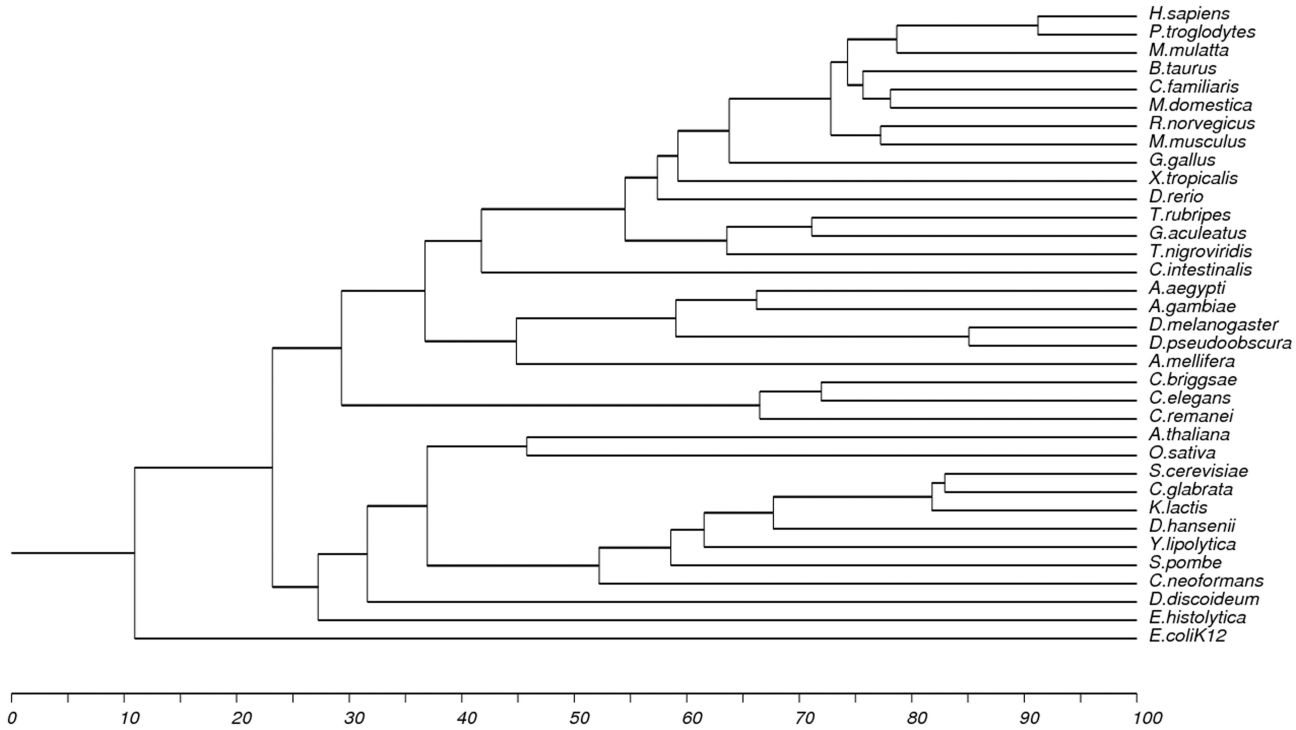

Figure 1. Orthophylogram of all 35 species in InParanoid 6. This UPGMA tree is based on the average fraction of orthologs between species. For instance, on average 91.2% of the proteins in *H. sapiens* and *P. troglodytes* are orthologous. The tree topology generally corresponds to the standard taxonomy, but a few exceptions were noted (see text).

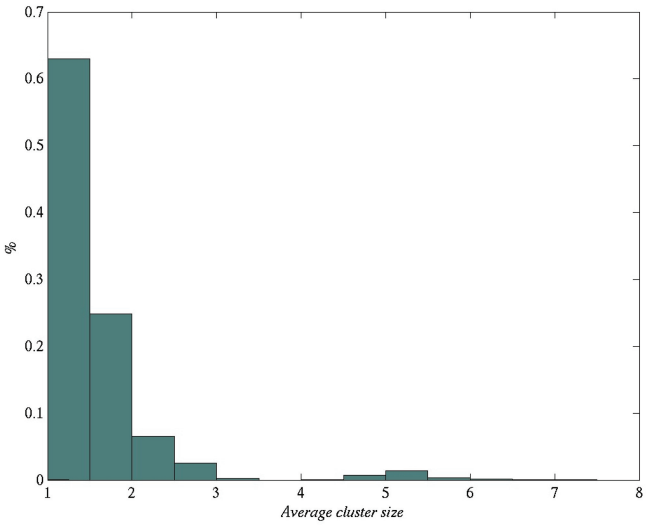

Figure 2. Histogram of average number of inparalogs/cluster per species for all species–species comparisons in InParanoid 6. The peak around five inparalogs/cluster is entirely caused by *O. sativa*, rice.

gene identifiers between any rice proteins in the MOD. This problem will be resolved in the future by collaborating directly with the rice MOD staff to get a better-annotated rice proteome.

DATA AVAILABILITY

The InParanoid database is freely available at http://inparanoid.sbc.su.se. In addition to the data which is available to search/browse using the web interface, fasta files containing all proteins, protein description files, ortholog tables in raw, SQL and XML format are available for each pairwise InParanoid analysis. The InParanoid program is freely available upon request to inparanoid@sbc.su.se.

ACKNOWLEDGEMENTS

We thank Tomas Ohlson for database assistance and all the MOD staff that have provided their data. This study was funded by grants from Stockholm University, Royal Institute of Technology, Pharmacia, and the Knut and Alice Wallenberg Foundation.
Funding to pay the Open Access publication charges for this article was provided by Stockholm University.

Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Sonnhammer,E.L.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
3. Remm,M., Storm,C.E.V. and Sonnhammer,E.L.L. (2001) Automatic clustering of orthologs and In-paralogs from pairwise species comparisons. *J. Mot. Behav.*, **314**, 1041–1052.
4. O'Brien, Remm,M and Sonnhammer,E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.

5. Alexeyenko,A., Lindberg,J., Perez-Bercoff,A. and Sonnhammer,E.L.L. (2006) Overview and comparison of ortholog databases. *Drug Discovery Today: Technol.*, **3**, 137–143.
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. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
7. Dessimoz,C., Boeckmann,B., Roth,A.C. and Gonnet,G.H. (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res.*, **34**, 3309–3316.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
9. Hulsen,T., Huynen,M.A., de Vlieg,J. and Groenen,P.M. (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.*, **7**, R31.
10. Chen,F., Mackey,A.J., Vermunt,J.K. and Roos,D.S. (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, **2**, e383.
11. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., De Montigny,J., Marck,C., Neuvéglise,C. *et al.* (2004) Genome evolution in yeasts. *Nature*, **439**, 35–44.
12. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.