W330–W334 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl311

YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms

Christopher J. Penkett, James A. Morris, Valerie Wood and Jürg Bähler*

Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK

Received February 13, 2006; Revised March 6, 2006; Accepted April 11, 2006

ABSTRACT

We present YOGY a web-based resource for orthologous proteins from nine eukaryotic organisms: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Plasmodium falciparum*, *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*. Using a gene name from any of these organisms as a query, this database provides comprehensive, combined information on orthologs in other species using data from five independent resources: KOGs, Inparanoid, HomoloGene, OrthoMCL and a table of curated fission and budding yeast orthologs. Associated Gene Ontology (GO) terms of orthologs can also be retrieved for functional inference. Integrating these different and complementary datasets provides a straightforward tool to identify known and predicted orthologs of proteins from a variety of species. This resource should be useful for bench scientists looking for functional clues for their genes of interest as well as for curators looking for information that can be transferred based on orthology and for rapidly identifying the relevant GO terms as an aid to literature curation. YOGY is accessible online at http://www.sanger.ac.uk/PostGenomics/S_pombe/YOGY/.

INTRODUCTION

It is common practice to obtain useful clues about the function and evolution of a protein of interest by identifying homologous proteins in other organisms (1–3). There are three types of homology with biological relevance (4). Orthology is most useful for insight into related gene functions as it arises from a common protein in an ancestral organism rather than from gene duplication (paralogy) or horizontal transfer of genes (xenology).

Several methods are available to identify orthologous proteins in different organisms. KOGs [euKaryotic Orthologous Groups; Ref. (5)] is a homology database derived from seven eukaryotic genomes, which uses the principle of BLAST best hits between three proteins from different organisms (6,7); since many eukaryotic proteins contain multiple domains, some common modules are masked (5). Inparanoid contains 26 datasets from 23 eukaryotic organisms; it can distinguish true homologs (orthologs and in-paralogs) from out-paralogs that arose from gene duplications prior to the divergence of two species (8–10). HomoloGene is a system for automated detection of homologs among the annotated proteins of 18 eukaryotic genomes; it is integrated with other databases at the NCBI including PubMed, Entrez and GEO (11). A recent addition to orthology resources is OrthoMCL, which can group orthologs from multiple genomes into a single cluster [currently 55 organisms; Refs (12,13)]. Finally, a curated list of orthologs between *Schizosaccharomyces pombe* (fission yeast) and *Saccharomyces cerevisiae* (budding yeast) is also available. This dataset has been compiled by inspecting multiple alignments and clusters of protein families on a protein-by-protein basis, taking into account experimental evidence, domain organization, protein length and species distribution (14).

These various homology resources have different advantages and complement each other. For example, no method is optimal for both specificity and coverage; assessing the results from multiple resources can thus increase confidence in orthology calls. Ortholog identification and subsequent extraction of relevant functional data on a gene by gene basis can be time consuming and confusing, owing to a lack of integration of the various resources. We have designed a web server called YOGY (eukarYotic OrtholoGY) that integrates results from the homology databases described above. Information from all these data sources is stored in a combined database to ease the search for and interpretation of orthologs. Gene Ontology [GO; Ref. (15)] annotations

*To whom correspondence should be addressed. Tel: +44 0 1223 496948; Fax: +44 0 1223 496802; Email: jurg@sanger.ac.uk

supported by manual evidence codes are included to provide functional insight into uncharacterized proteins. All of this information can be searched with a web interface in a single step.

IMPLEMENTATION

YOGY is implemented in a MySQL relational database running on a UNIX server. Data for the external resources have been downloaded from the associated FTP and websites for import into our database (Supplementary Data). The data model has been validated to identify and remove potential problems such as many-to-many relationships. It uses Perl scripts together with the Perl DBI module for file import. For queries, we have designed a web interface using the CGI module of Perl, hosted on an Apache server. The Perl GD graphics module is used for bar charts.

Genes and proteins from the following nine organisms can be searched using gene names or systematic identifiers from the corresponding Model Organism Database (MOD), Ensembl (16), NCBI (11) or UniProt (17): *Homo sapiens, Mus musculus* (18), *Rattus norvegicus* (19), *Arabidopsis thaliana* (20), *Drosophila melanogaster* (21), *Caenorhabditis elegans* (22), *Plasmodium falciparum* (23), *S.pombe* (23), and *S.cerevisiae* (24). Where possible, identifiers from the appropriate MOD are shown throughout the output so that proteins from the five data sources can be evaluated for consistency; this is useful as the different homology resources use identifiers from a variety of databases. Because of the ambiguity of many identifiers, legacy naming systems and revisions to gene structures and gene complements, it is not always possible to be certain whether some apparent differences in orthology calls are, in fact, equivalent proteins. Whilst we have made every effort to map these identifiers automatically using resources from the MOD, the International Protein Index (25), UniProt (17) and the NCBI Entrez Gene database (11), any discrepancies should be checked manually by the user. It is possible to use incomplete names with a wildcard option, providing a list of genes and one-line descriptions for further search.

GO terms annotated to the identified orthologs can also be retrieved. Only associations using experimental and curator validated evidence codes are included. The option to show GO terms is switched off by default due to the increased time required to download GO data. Options are provided to display GO terms in separate tables at the end of each resource, or in a single table at the end of the output.

The output is provided in a tabulated HTML format (Figure 1). The first table contains general information for the protein of interest including description and links to the corresponding MOD and the UniProt database, if this accession number is available. For *S.pombe*, links to gene expression profiles during the cell cycle [C; Ref. (26)], eiotic differentiation [M; Ref. (27)] and stress conditions [S; Ref. (28)] are also provided. The data sources which provide positive orthology results for the gene of interest are then specified with links to the corresponding outputs.

The orthology results are presented in a standard output format for each dataset. At the top is information about the query protein cluster(s), followed by a list of available orthologs ordered by organism together with links to the ortholog resource (Figure 1B–E). Links to UniProt are also provided if the accession number is available. Below, each data source is mentioned in the order given in the output page.

For KOGs, the summary table starts with the unique KOG name together with a link to the website. The next column displays a bar chart of the ortholog numbers for each organism, revealing the phylogenetic pattern for the KOGs (Figure 1B). This chart also provides a link to a list of other KOGs that share the same phylogenetic pattern, which provides insight into gene preservation and loss in different lineages. The summary table also indicates the functional classification, with a link to other KOGs in this classification, and a one-line description for the KOG. The orthologs are displayed in a list below the summary table, together with links for each protein or domain to the corresponding KOG cluster alignments and to the relevant protein page at NCBI (Figure 1B).

For Inparanoid, we have excluded orthologs from largely unannotated organisms, which are not in the other homology resources; this reduces the output page to 18 organisms (20 databases, as both mouse and rat include two datasets). The bar chart on top shows the phylogenetic pattern for the orthologs (Figure 1C). The list underneath shows the orthologs for the query protein, links to the Inparanoid protein clusters for each organism, the Inparanoid score and a link to the protein page in the corresponding MOD (Figure 1C). Inparanoid uses a sophisticated methodology to distinguish between in- and out-paralogs (8); we have downloaded the tables from the Inparanoid website and present these precalculated datasets on the YOGY website.

For HomoloGene, the summary at the top provides a link to the query protein cluster at NCBI and a phylogenetic bar chart. Each ortholog is then presented by organism with links to the relevant NCBI pages.

For OrthoMCL, we have again excluded orthologs from largely unannotated organisms and prokaryotes (except *Escherichia coli*, which is also included in Inparanoid) reducing the output to 24 organisms. The summary table includes a link to the OrthoMCL cluster and a phylogenetic bar chart (Figure 1D). This table is followed by a list of orthologs in the cluster with a link to the original protein sequence used for clustering and a link to the relevant MOD (Figure 1D). For some of the less well-characterized yeasts, which have no MOD, a link is provided to either the 'Yeast Gene Order Browser' or Génolevures that both provide graphical representations of conserved genome location (29,30).

For the curated yeast ortholog dataset, only fission and budding yeast proteins are included. The output provides the lists of orthologs together with links to the *S.pombe* GeneDB (23)and SGD (24) databases (Figure 1E).

If selected, either multiple tables or one table at the end provide a summary for all GO terms found for the query protein and its orthologs. This includes the term name, the aspect (P: Biological Process; C: Cellular Component and F: Molecular Function), the evidence codes, and the corresponding organisms together with the accession numbers of orthologs containing the GO term for each organism (Figure 1F). GO terms with the evidence code 'Inferred from Electronic Annotation'(IEA) are not included as these have not

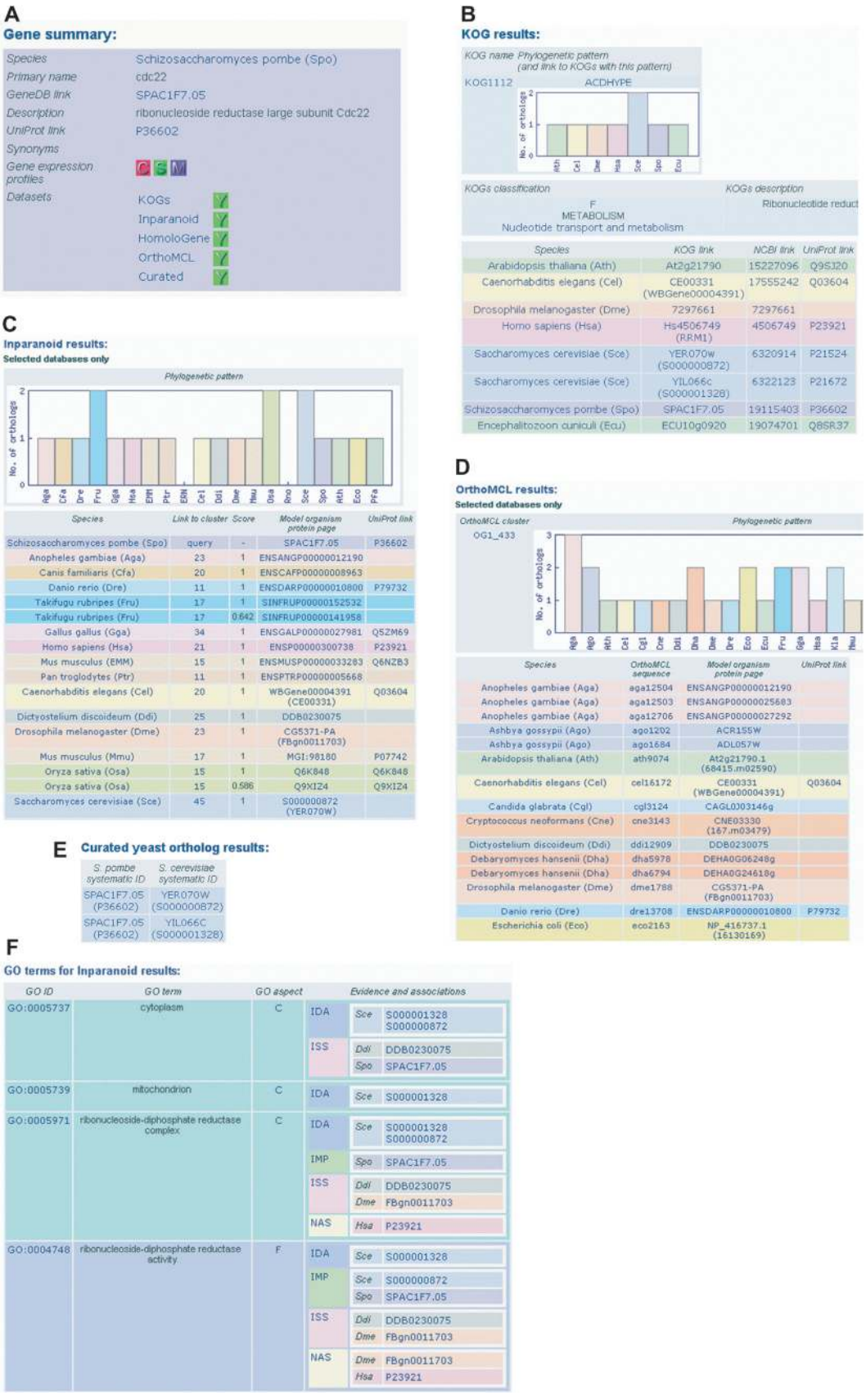

Figure 1. An example of the output of YOGY for the *S.pombe* protein Cdc22, which shows: (**A**) the summary table, (**B**) part of the KOGs table, (**C**) part of the Inparanoid table, (**D**) part of the OrthoMCL table, (**E**) the curated yeast ortholog table and (**F**) part of the table of associated GO terms for Inparanoid orthologs.

been assessed by an annotator, and tend to be to higher level terms.

In the future, we plan to make further changes to improve the display and integration of the different orthology resources. In the longer term, GO annotations will be represented on the GO tree structure, which will allow for the rapid identification of redundant and non-overlapping annotations from the various model organisms.

CONCLUSION

The described integrated database together with the accompanying search site provides a straightforward resource to identify orthologs from all specialized databases that are currently most useful; these ortholog databases have been built using different methods that complement each other, and the integrated results give a rich picture of orthology based on combined evidence from the independent resources. The GO annotations of orthologs can provide additional evidence on orthology and help to infer functional information for genes with limited annotation. This resource will be regularly updated to include the latest information from the independent data sources.

SUPPLEMENTARY DATA

Supplementary Data are available at *NAR* Online.

ACKNOWLEDGEMENTS

The authors thank Matloob Qureshi and members of the Bähler laboratory and the Pathogen Sequencing Unit for discussions and help with programming. The work in the group is funded by a Cancer Research UK [CUK] Grant No. C9546/A6517 and by DIAMONDS, an EC FP6 Lifescihealth STREP (LSHB-CT-2004-512143). Funding to pay the Open Access publication charges for this article was provided by Cancer Research UK.

Conflict of interest statement. None declared.

REFERENCES

1. Henikoff,S., Greene,E.A., Pietrokovski,S., Bork,P., Attwood,T.K. and Hood,L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.
2. Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T. *et al.* (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
3. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
4. Fitch,W.M. (2000) Homology: a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
5. 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.
6. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
7. 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.
8. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
9. Sonnhammer,E.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
10. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
11. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
12. Chen,F., Mackey,A.J., Stoeckert,C.J.,Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
13. Li,L., Stoeckert,C.J.,Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
14. Wood,V. (2006) *Schizosaccharomyces pombe* comparative genomics; from sequence to systems. In Sunnerhagen,P. and Piskur,J. (eds), *Comparative Genomics Using Fungi as Models (Series: Topics in Current Genetics)*. Springer, Berlin, Vol 15, pp. 233–285.
15. 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.
16. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox.T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
17. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
18. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S., Bello,S.M. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res.*, **33**, D471–D475.
19. de la Cruz,N., Bromberg,S., Pasko,D., Shimoyama,M., Twigger,S., Chen,J., Chen,C.F., Fan,C., Foote,C., Gopinath,G.R. *et al.* (2005) The Rat Genome Database (RGD): developments towards a phenome database. *Nucleic Acids Res.*, **33**, D485–D491.
20. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
21. Grumbling,G. and Strelets,V. (2006) FlyBase: anatomical data, images and queries. *Nucleic Acids Res.*, **34**, D484–D488.
22. Schwarz,E.M., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Canaran,P., Chan,J., Chen,N., Chen,W.J., Davis,P. *et al.* (2006) WormBase: better software, richer content. *Nucleic Acids Res.*, **34**, D475–D478.
23. Hertz-Fowler,C., Peacock,C.S., Wood,V., Aslett,M., Kerhornou,A., Mooney,P., Tivey,A., Berriman,M., Hall,N., Rutherford,K. *et al.* (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, **32**, D339–D343.
24. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al.* (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **32**, D311–D314.
25. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
26. Rustici,G., Mata,J., Kivinen,K., Lio,P., Penkett,C.J., Burns,G., Hayles,J., Brazma,A., Nurse,P. and Bähler,J. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nature Genet.*, **36**, 809–817.

27. Mata,J., Lyne,R., Burns,G. and Bähler,J. (2002) The transcriptional program of meiosis and sporulation in fission yeast. *Nature Genet*., **32**, 143–714.
28. Chen,D., Toone,W.M., Mata,J., Lyne,R., Burns,G., Kivinen,K., Brazma,A., Jones,N. and Bähler,J. (2003) Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell*, **14**, 214–229.
29. Byrne,K.P. and Wolfe,K.H. (2006) Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic Acids Res*., **34**, D452–455.
30. Sherman,D., Durrens,P., Iragne,F., Beyne,E., Nikolski,M. and Souciet,J.L. (2006) Génolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts. *Nucleic Acids Res*., **34**, D432–D435.