Database, Vol. 2010, Article ID baq002, doi:10.1093/database/baq002

Original article

Gene-oriented ortholog database: a functional comparison platform for orthologous loci

Meng-Ru Ho[1,2,3], Chun-houh Chen[4] and Wen-chang Lin[1,3,*]

[1]Institute of Biomedical Informatics, National Yang-Ming University, Taipei 112, [2]Bioinformatics Program, Taiwan International Graduate Program, [3]Institute of Biomedical Sciences and [4]Institute of Statistical Sciences, Academia Sinica, Taipei 115, Taiwan

***Corresponding author:** Tel: +886 2 2652 3967; Fax: +886 2 2785 7654; Email: wenlin@ibms.sinica.edu.tw

Submitted 27 August 2009; Revised 8 December 2009; Accepted 14 January 2010

The accumulation of complete genomic sequences enhances the need for functional annotation. Associating existing functional annotation of orthologs can speed up the annotation process and even examine the existing annotation. However, current protein sequence-based ortholog databases provide ambiguous and incomplete orthology in eukaryotes. It is because that isoforms, derived by alternative splicing (AS), often share higher sequence similarity to interfere the sequence-based identification. Gene-Oriented Ortholog Database (GOOD) employs genomic locations of transcripts to cluster AS-derived isoforms prior to ortholog delineation to eliminate the interference from AS. From the gene-oriented presentation, isoforms can be clearly associated to their genes to provide comprehensive ortholog information and further be discriminated from paralogs. Aside from, displaying clusters of isoforms between orthologous genes can present the evolution variation at the transcription level. Based on orthology, GOOD additionally comprises functional annotation from the Gene Ontology (GO) database. However, there exist redundant annotations, both parent and child terms assigned to the same gene, in the GO database. It is difficult to precisely draw the numerical comparison of term counts between orthologous genes annotated with redundant terms. Instead of the description only, GOOD further provides the GO graphs to reveal hierarchical-like relationships among divergent functionalities. Therefore, the redundancy of GO terms can be examined, and the context among compared terms is more comprehensive. In sum, GOOD can improve the interpretation in the molecular function from experiments in the model organism and provide clear comparative genomic annotation across organisms.

Database URL: http://goods.ibms.sinica.edu.tw/goods/

Introduction

Orthologs are defined as genes in different species that originated from a single genetic locus in the last common ancestor, i.e. homology following speciation (1–4). Taking orthology as the basis to infer functional annotation between orthologs can accelerate annotation process and is widely adopted. The Gene Ontology (GO) database (5) is a major collection possessing consistent descriptions of gene products from different databases. The GO database maintains three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Associated with the elaborate label of genes' functionality from the GO database, orthologs offer the ability to accurately convey annotation across organisms.

Several ortholog databases are now available online. Most of them, however, consider orthology from the aspect of protein sequences individually, including HomoloGene (http://www.ncbi.nlm.nih.gov/homologene), EnsemblCompara (6), Inparanoid (7,8), Roundup (9), OrthoMCL (10,11) and OrthoDB (12). There exist ambiguous and incomplete ortholog assignments because of the

Database, Vol. 2010, Article ID baq002, doi:10.1093/database/baq002

interference mediated by alternative splicing (AS) (13). Isoforms of one gene might be assigned to different orthologous clusters. For instance, one gene, *SORBS2*, belongs to two individual HomoloGene group ids (HIDs), HomoloGene:83295 and HomoloGene:33484, because its isoforms are separated into two different group ids. This implies that both orthologous clusters contain the ortholog information of *SORBS2*, but none of them is the complete orthologous cluster of *SORBS2* in HomoloGene. Moreover, the orthologous annotation might be inconsistent across databases without considering alternative splice variants collectively. For example, EnsemblCompara takes the longest protein of each gene as the representative to identify orthologs without considering alternative splice variants. The human ortholog of mouse gene, *Sorbs2*, is annotated as Ensembl:ENSP00000284776 in EnsemblCompara but Ensembl:ENSP00000284776 in Inparanoid. Ensembl Compara and Inparanoid contain inconsistent ortholog information of *SORBS2*. It is because that human *SORBS2* actually possesses two protein records, Ensembl: ENSP00000284776 and Ensembl:ENSP000000347852. Considering orthology from a representative protein only would derive this incomplete and inconsistent information. Furthermore, Roundup reports that the human ortholog of mouse Kcnq4 is only NCBI:NP_751895 while human *KCNQ4* actually owns two protein products, NCBI:NP_751895 and NCBI:NP_004691. Above evidence demonstrates that protein-sequence-based ortholog databases contain ambiguous and incomplete orthology.

Actually there exist well-known ortholog databases incorporating GO terms to illustrate functional evolution, such as Roundup (9) and YOGY (14). They, however, merely display text of terminal GO terms. Only showing text of end GO nodes is unable to reveal all relationships among related functionalities from the GO database which is structured as directed acyclic graphs. Take DNA binding (GO:0003677) and transcription factor activity (GO:0003700) as an illustration, these two GO terms, having parent–child relationship (Figure 1D), are annotated redundantly to the mouse gene, *Gtf2ird1*, in molecular function (Figure 1C). This kind of redundant annotations leads a propagation of annotated functions. Without the topology justification, the numerical comparison of functionalities among orthologous genes is doubted. Moreover, the textual comparison between GO terms merely tells their differences in the letter. Assisted with the topology of GO terms, the context among related GO terms is more comprehensive. For example, there is no common annotated GO ID of *GTF2IRD1* between human and mouse in molecular function (Figure 1C). However, their annotated terms do own direct linkage in topology (Figure 1D). Considering the parent–child relationship among these GO terms, these annotated functions of *GTF2IRD1* between human and mouse actually possess conservation. Therefore, the topology is helpful to reveal the potential connection among GO terms.

In this study, we present Gene-Oriented Ortholog Database (GOOD): a functional comparison platform for orthologous loci. Employing genomic locations of transcripts to cluster AS-derived isoforms prior to ortholog delineation eliminates the interference from AS. Displaying clusters of isoforms between orthologous genes can further show the evolution variety at the transcription level. Based on orthology, GOOD additionally comprises functional annotation from the GO database. This information can benefit species which lack of functional annotation such as chimpanzee. That is, functional annotation for a given species can be predicted using the GO annotation of orthologous transcription regions (genes) from other well-annotated species. Furthermore, GOOD not only lists the description of GO terms, but also presents graphical views of connections among them. Using graphs, GOOD simultaneously displays all relationships among nodes which are in the paths from the annotated GO term to the root node. That is, graphs of GO terms can further provide the comprehensive topology for users to reveal the divergence among related GO terms in different GO layers. Hence, we believe that GOOD can serve as a comprehensive comparison platform for orthologous genes.

Data construction and content

Data sources

In this study, we carefully chose species for ortholog analysis. Because GOOD emphasizes the interference caused by AS events which is more abundant in higher order eukaryotes, we first considered mammals. In addition, the algorithm we applied (13) is sensitive to the quality of genome assembly and transcript annotations. Therefore, species are sequentially included by their annotation levels. Organisms which are updated more frequently are regarded as better-annotated ones based on the release status of each genome announced in NCBI Genome database (http://www.ncbi.nlm.nih.gov/Genomes/) and UCSC website (http://genome.ucsc.edu/). Human, mouse and cow are then processed in order. When it comes to rat, the performance is significantly dropped down. Thus, we stopped including the remainder. All the chosen species are updated at least four times, and their latest update time is later than 2007. Examining other species with sequenced vertebrate genomes, zebrafish is also well studied. We eliminated it because the algorithm we used (13) has a limitation to apply directly to such a far distant species to all other chosen species. We further included chimpanzee to demonstrate that the GO annotation from a well-annotated gene can benefit to its un-annotated orthologous gene.

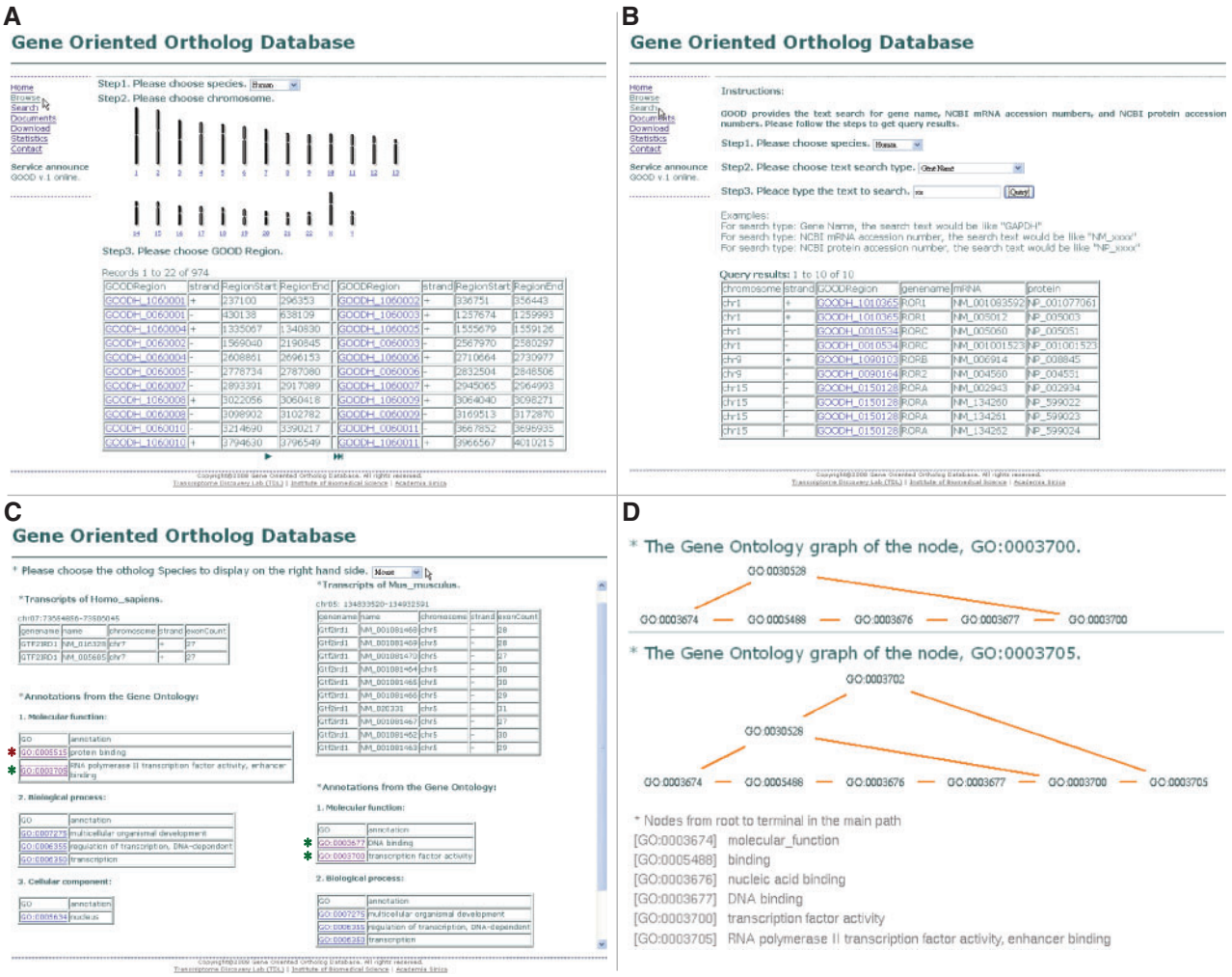

Figure 1. Snapshots of the GOOD web interface. Panels **A** and **B** are the two ways, browse and search functions, for users to select a genomic locus on the website. Users can browse according to genomic positions to look into a specific genomic locus. Or they can achieve the same purpose by searching text of a gene name or a NCBI accession number. Panel **C** demonstrates the simultaneous display of transcripts and GO annotation between orthologous genomic loci, *GTF2IRD1*. Transcripts are limited to NCBI reference sequence database, and GO terms are arranged with respect to three ontologies. Users can further click GO terms to see their topology. There are two graphs of GO terms shown in panel **D**.

We used the genome assemblies from human NCBI build 36.3 published on 26 March 2008, mouse NCBI build 37.1 published on 5 July 2007, and chimpanzee NCBI build 2.1 published on 5 October 2006. Cow genome assembly, published in October 2007 (bosTau4), is available on UCSC website. All the annotation of reference transcripts was downloaded from UCSC website. The latest versions of human, mouse, chimpanzee, and cow builds are termed hg18, mm9, panTro2 and bosTau4, respectively. We used the functional annotation from the GO database (http://www.geneontology.org/) contained in the file named go_200806-termdb-tables.tar.gz.

Orthology

The utility of GOOD is that it exposes functional difference between orthologous genomic loci. There are four eukaryotic organisms, *Homo sapiens*, *Mus musculus*, *Pan troglodytes* and *Bos taurus* in GOOD. To identify genomic transcription regions in those genomes, we used the annotation of reference transcripts from UCSC website. All AS-derived isoforms are clustered together by their genomic locations and associated with their transcription regions (genes) prior to the ortholog delineation. GOOD contains generated regions, 18 373 in human, 18 858 in mouse, 17 681 in chimpanzee and 9311 in cow. Each region has its own unique transcriptional representative, the processed transcription unit (PTU) (13). That is, AS information is analyzed to generate PTUs which are DNA sequences of genes without absolute introns. We then performed the alignment of PTUs between two chosen species to get reciprocal best hits (RBHs) pairs, putative orthologs.

Database, Vol. 2010, Article ID baq002, doi:10.1093/database/baq002

Those putative orthologs are mapped back to their respective genomes to depict an outline of the synteny map. Based on the presented synteny, we further enlarged orthologs from potential PTU pairs, possessing a best hit from one side only. The ortholog of duplicated genes caused by the homologous recombination and the split orthologous regions of embedded genes would be included via this step. Thus, the relationship between orthologous regions might not be one-to-one. We followed all steps to perform six combinations of four species. Hence, there are 17 588 human/chimpanzee orthologous pairs, 16 545 human/mouse orthologous pairs, 9314 human/cow orthologous pairs, 15 078 chimpanzee/mouse orthologous pairs, 9144 chimpanzee/cow orthologous pairs and 8937 mouse/cow orthologous pairs in GOOD.

Functional annotation and graphs

GOOD includes functional annotation from the GO database. The GO database possesses three structured controlled ontologies: biological process, cellular component and molecular function. The ontologies of the GO database are structured as directed acyclic graphs, which are similar to hierarchies but different in that a more specialized term (child) can be related to more than one less specialized term (parent). That is, one GO term might have multiple paths to its root node, one of the three ontologies. In addition, not all paths from the same term contain the same amount of internal nodes.

In this study, GOOD lists GO terms with respect to three ontologies for each genomic locus (Figure 1C). By this manner, transcripts and GO information of two orthologous genes are shown together. Moreover, the functional topology of GO terms, composed of parent GO terms to describe the annotated function, is presented graphically. The algorithm used to generate those graphs is illustrated as follows.

Graph Generation Algorithm (GGA)

Input: The queried GO term (Q),
Relationship table from GO (term2term: T2T),
Roots (cellular component: CC;
biological process: BP;
molecular function: MF)

Output: The relationship graph of the queried GO term

Step 1 Recursively query the parents of Q according to T2T, until the obtained parent belonging to {CC, BP, MF}

Step 2 Construct and list all the possible paths from Q to {CC, BP, MF}

Step 3 Find out the longest path (LP) from all the possible paths

Step 4 Horizontally depict LP on the center of a graph

Step 5 Find out the longest common path to LP from the remainder
Case 1: Upscore > Downscore
Attach the path onto the existing paths
Case 2: Upscore < Downscore
Attach the path under the existing paths
Case 3: Upscore = Downscore
Attach the path to the side that contains fewer paths
Until all paths are plotted

In GGA, the longest path (LP) of a queried GO term is first depicted on the center of a graph horizontally. Each edge of LP is equally assigned according to the width of a graph. Take the plotted LP as standard, the rest of paths are attached in order. The longest path of the remainder having the highest similarity with LP comes first. Furthermore, Upscore and Downscore are used to determine which side of the existing graph the coming path is attached to. Upscore is the number of overlapped terms between the existing upper GO terms of LP and the internal GO terms of the coming path. Downscore, similarly, is the overlapped number between the existing lower GO terms of LP and the internal GO terms of the coming path. Reiterate the process until all paths are plotted.

Actually, all edges (connections) in GO graphs are shrinkable. That is, GO graphs are mainly used to display the connection among terms. Since there is no clear definition to follow, those graphs are not tended to claim that the depicted lengths of edges are the absolute lengths of connections. For instance, it is difficult to number edges or even decide whether the edge between binding and nucleic acid binding is longer than the edge between nucleic acid binding and DNA binding. In this study, GO graphs are, therefore, adjusted to be most visible for user to reveal the linkage. First, let LP scattered on the center of a graph evenly to make sure the most complex path is expanded properly. Attach the rest of paths to LP in order. The ordering criterion is that the most common path of LP comes first from the remainder. This criterion minimizes the increasing length for adding a new path to an existing graph. In this way, the entropy of GO terms in a graph can be the lowest based on the same LP.

Web interface

GOOD is designed to reveal the function of genomic loci and further associated with orthology to infer functional evolution. Based on this main principle, users have to begin with a genomic locus selection to explore data. There are two ways for users to choose a specific genomic locus from the web interface, browse with genomic positions (Figure 1A) and search with a gene name or a NCBI accession number (Figure 1B). Both lead users to look into a specific genomic region (gene). Once getting into a

region, users first observe all reference transcripts derived from this genomic locus. Next come all GO annotated terms of this gene with respect to three main ontologies of the GO database (the left panel of Figure 1C). All GO terms are clickable to show their GO graphs (the up panel of Figure 1D). Later, orthology is introduced when users choose the other species. The transcripts and GO terms of the orthologous gene are shown in the right panel of Figure 1C. And, linking NCBI HomoloGene (http://www.ncbi.nlm.nih.gov/homologene) through the accession number of transcripts provides users to qualify the result and pass the annotation from characterized species to uncharacterized ones. This intuitive web interface leads users to compare transcripts and annotations from the GO database between selected othologous loci. Alternatively, user can download all relative data from the download page to do their own analysis.

Discussion

The limitation inherited from the algorithm

The orthology in GOOD is designated by the algorithm proposed by Ho *et al.* in 2008. According to the comparison among public databases, the performance of this algorithm is remarkable. But the authors also admit that this methodology relies on higher completeness of genome assembly and transcript annotations. That is, poorer data sources would derive poorer results. This makes the algorithm can merely applied to well-annotated species while all other methods also suffer the same issue more or less. Aside from that it also needs some modification when utilized between two far distant organisms, like human and zebrafish. The authors propose to apply this method several times for all relative species between two distant species instead of applying this method directly. Interpolating effective species between these two distant species makes PTUs sensitive enough to provide the comprehensive orthology. Inheriting these two limitations from the algorithm, GOOD now only contains four species. For expansion, we first consider including the species which is close to any of the four species in GOOD. That is, mammals with well-annotated genomes are all original candidates to be a new species in GOOD and further followed by birds, amphibians, echinoderms and fish. In this manner, we keep examining all possible species and include well-annotated species. As the sequencing technology is improved and more valid annotation is accumulated, there will be more species to enrich GOOD.

Gene-oriented ortholog presentation

AS in eukaryotic genomes plays an important role in augmenting biological complexity such that one gene can result in generating multiple proteins. Those proteins derived from AS share high similarity in sequences and then hinder protein-sequence-based ortholog identification. Existing ortholog databases ignore the AS-mediated interference. That might cause isoforms to be over-clustered into separate orthologous clusters and then orthology is ambiguous. In addition, orthology in some databases is presented by the representative protein which is incomplete orthologous information. There, then, is the inconsistency among various databases according different chosen rule of representatives. Surely, it is important to clarify the AS when presenting orthology. GOOD bases on gene loci to represent orthology. Each AS product can be associated with its own genomic region, and orthology can be inferred at the gene level. That is, based on the gene scale, the interpretation of othologs in GOOD is clear and complete. And, the transcript lists of orthologous genes can further recapitulate the transcription changes.

Graphs of the hierarchical-like structure

Combining functional annotation with orthologs can speed up the annotation process of genes' functionality. Three ontologies in the GO database are structured as directed acyclic graphs. Only considering the terminal terms therefore is not sufficient. For instance, redundant annotations, which mean that both child and parent nodes are assigned together, might be unaware. This can inflate the annotation number and lead unfair comparisons. Although the GO website lists the text of parent terms in a line-based structure to display the topology of a queried GO term, users still can not capture the entire topology at once. Once child terms have multiple parents, there exists a reticular relation among terms. That makes the line-based exploration more inconvenience to use. It is due to that lines are not adequate to present a net structure. Here, GOOD utilizes graphs to depict the topology of GO terms for users to catch the relationships among related terms comprehensively. A graph is a collection of points and lines connecting some subset of them. Consequently, graphs can make the topology of GO terms clear at a glance. With this sight, either the redundant annotations of the same gene or different levels of functional changes between orthologs can be pointed out. Even so, nodes in GO might be repeated and edges are still undetermined. Those make it meaningless to number terms or perform direct comparisons among graphs. It needs more solid explication among terms to accomplish more detail and specific comparisons.

Conclusion

Protein-sequence-based orthologs assignment is obstructed by alternative splicing events in Eukaryotes. From the gene-oriented presentation, isoforms can be clearly associated to their genes to provide comprehensive ortholog

Database, Vol. 2010, Article ID baq002, doi:10.1093/database/baq002

information and further be discriminated from paralogs. Furthermore, GOOD incorporates the GO database; therefore, not only the redundancy of annotation can be examined, but also functional annotation can be inferred to the target species based on orthology. Aside from, GO graphs provide topology views of GO terms. GO graphs make the functional comparison more precise and thorough. In this study, GOOD is presented as a gene-oriented comparison platform of functionalities based on orthology for researchers to derive interested molecular functions from experiments in model organisms to speed up the process of functional annotation.

Funding

This work was supported by research grants from Academia Sinica, Taiwan. Funding for open access charge: Academia Sinica, Taiwan.

Conflict of interest statement. None declared.

References

1. Sonnhammer,E.L.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
2. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
3. Fitch,W.M. (2000) Homology: a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
4. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
5. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
6. Vilella,A.J., Severin,J., Ureta-Vidal,A. *et al*. (2009) EnsemblCompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
7. Remm,M., Storm,C.E. and Sonnhammer,E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
8. O'Brien,K.P., Remm,M. and Sonnhammer,E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
9. Deluca,T.F., Wu,I.H., Pu,J. *et al*. (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*, **22**, 2044–2046.
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. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr *et al*. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
12. Kriventseva,E.V., Rahman,N., Espinosa,O. *et al*. (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.*, **36**, D271–D275.
13. Ho,M.R., Jang,W.J., Chen,C.H. *et al*. (2008) Designating eukaryotic orthology via processed transcription units. *Nucleic Acids Res.*, **36**, 3436–3442.
14. Penkett,C.J., Morris,J.A., Wood,V. *et al*. (2006) YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms. *Nucleic Acids Res.*, **34**, W330–W334.