D556–D560 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 12 November 2010*
doi:10.1093/nar/gkq1109

PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions

Jaime Huerta-Cepas, Salvador Capella-Gutierrez, Leszek P. Pryszcz, Ivan Denisov, Diego Kormes, Marina Marcet-Houben and Toni Gabaldón*

Bioinformatics and Genomics Programme. Centre de Regulació Genòmica. Doctor Aiguader, 88. 08003 Barcelona, Spain

Received September 15, 2010; Revised and Accepted October 18, 2010

ABSTRACT

The growing availability of complete genomic sequences from diverse species has brought about the need to scale up phylogenomic analyses, including the reconstruction of large collections of phylogenetic trees. Here, we present the third version of PhylomeDB (http://phylomeDB.org), a public database for genome-wide collections of gene phylogenies (phylomes). Currently, PhylomeDB is the largest phylogenetic repository and hosts 17 phylomes, comprising 416 093 trees and 165 840 alignments. It is also a major source for phylogeny-based orthology and paralogy predictions, covering about 5 million proteins in 717 fully-sequenced genomes. For each protein-coding gene in a seed genome, the database provides original and processed alignments, phylogenetic trees derived from various methods and phylogeny-based predictions of orthology and paralogy relationships. The new version of phylomeDB has been extended with novel data access and visualization features, including the possibility of programmatic access. Available seed species include model organisms such as human, yeast, *Escherichia coli* or *Arabidopsis thaliana*, but also alternative model species such as the human pathogen *Candida albicans*, or the pea aphid *Acyrtosiphon pisum*. Finally, PhylomeDB is currently being used by several genome sequencing projects that couple the genome annotation process with the reconstruction of the corresponding phylome, a strategy that provides relevant evolutionary insights.

Introduction

Recent developments in sequencing technologies have radically changed the way in which many biologists perform their research. Although, genome sequencing used to be a costly and challenging analysis only reserved for a handful of model species, it is nowadays a technique that can be applied at a reasonable price and effort to significantly larger sets of organisms. As a result, the rate at which new genome sequences are deposited in public databases is growing. The comparison of genomes from diverse species within a common evolutionary framework, i.e. phylogenomics (1), constitutes a powerful tool to address many relevant questions including, among many others, the reconstruction of evolutionary relationships across different species (2), the prediction of function of uncharacterized proteins (3) and the establishment of orthology and paralogy relationships among homologous genes (4). Such comparative studies are often accompanied by large-scale phylogenetic analyses that comprise the reconstruction of evolutionary relationships of a large number of gene families, an approach that has been enabled by recent developments in hardware and phylogenetic algorithms (5). Currently, a number of public repositories for automatically-generated collections of phylogenetic trees do exist (6–10). Such collections vary in size and phylogenetic scope depending on their specific focuses. Most existing approaches rely on a clustering phase to define families of homologous sequences to which methods for phylogenetic reconstruction are applied. Alternatively, a phylogenetic reconstruction pipeline can be recursively applied to every gene in a genome so that each sequence is used as a seed in the process of phylogenetic reconstruction. The resulting collection of trees, representing the full complement of evolutionary histories of all genes encoded in a given genome, has been dubbed with the term 'phylome' (11).

*To whom correspondence should be addressed. Tel: +34 93 316 02 81; Fax: +34 93 316 00 99; Email: tgabaldon@crg.es; tonigaes@gmail.com

This strategy, which resembles more closely the gene-centered approach in classical phylogenetics, is computationally more costly than a family-based approach but it is mostly free of the difficulties of properly establishing family boundaries in clustering approaches. Moreover, a gene-centric approach ensures maximum coverage of the seed genome and, therefore, is more appropriate when the focus is set on a particular organism. The analysis of phylomes has proven to be very powerful in addressing a variety of questions, including the reconstruction of ancestral metabolisms (12), the evaluation of alternative species phylogenies (13,14), the identification of horizontal gene transfers (15), the inference of functional implications of massive waves of gene duplications (6,16) and the detection of orthology and paralogy relationships (17). To facilitate the exploitation of such genome-wide collections of phylogenetic trees, alignments and phylogeny-based orthology and paralogy predictions, we developed phylomeDB (7). Here we describe the new features of the current version of PhylomeDB, including recent improvements in the phylogenetic pipeline used to reconstruct new phylomes. In addition, we highlight the potential use of PhylomeDB resources for the annotation of newly sequenced genomes.

An improved phylome reconstruction pipeline

Phylomes stored in PhylomeDB are reconstructed using slight variations of an automated phylogenetic pipeline. This pipeline, first described in (6), is under constant improvement. In addition, each phylome is reconstructed under a specific set of parameters, which mainly depend on the phylogenetic scope of the sequences included. The specific details of how each phylome was generated, as well as the source of the raw data, are provided in the corresponding phylome description page. In brief, the pipeline proceeds as follows: for each protein-coding gene in the seed genome, the pipeline starts with a homology search in a database containing proteins encoded in the seed genome and in a set of selected fully-sequenced genomes. This set defines the taxonomic scope of the phylome and is adjusted to optimally address specific evolutionary questions. Homology searches are performed using the Smith–Waterman algorithm (18) and filtered according to specific e-value and overlap cut-offs. For genes encoding multiple splice variants the longest isoform is used. Subsequently, sets of homologous sequences are aligned. Several changes have been introduced in this step of the pipeline. First, instead of a single multiple sequence alignment method (e.g. MUSCLE), homologous sequences are aligned using three different programs: MUSCLE v3.7 (19), MAFFT v6.712b (20) and DIALIGN-TX (21). Moreover, alignments are performed in forward and reverse direction (i.e. using the Heads or Tails approach (22). The six resulting alignments are then combined with M-Coffee (23). This allows alignments to be trimmed not only based on their gap content but also on the pairing consistency across different alignments, using the program trimAl v1.2 (24). The resulting processed alignment is used to reconstruct phylogenetic trees using Neighbor Joining (NJ) and Maximum Likelihood (ML) methods. In the case of ML reconstruction, an additional improvement has been introduced in the model selection step. Instead of evaluating the likelihood of a tree under each evaluated model by means of a full ML reconstruction, the improved pipeline evaluates the likelihood on the topology obtained by NJ, allowing branch-length optimization. This allows exploring more models, from which only a selection [the best ranking ones according to the AIC criterion (25)] are used for a full ML approach. Confirming earlier observations (6,26), this procedure was found to provide the same result as a full ML approach in >98% of the cases (results not shown).

Phylome reconstruction and annotation of newly sequenced genomes

Although phylomes were initially reconstructed only for publicly available genomes, we soon realized its potential for ongoing genome annotation projects. In particular, the availability of a complete collection of phylogenies for the predicted gene set could be used to reliably assign orthology and paralogy relationships in related species, predict putative functions and detect large family expansions. Moreover, a phylome readily provides a detailed overview of the evolution of the targeted genome, and can be used to address particular questions of interest, such as the phylogenetic position of the seed species, the subset of genes under recent positive selection, the incidence of horizontal gene transfer, etc. Such strategy was applied to the annotation of the pea aphid *Acyrtosiphon pisum* genome (27), which, to our knowledge, constitutes the first example of a genome whose annotation is based on extensive phylogenetic analyses. The analysis of the *A. pisum* phylome (16) allowed the computation of the full catalogue of phylogeny-based orthology and paralogy relationships with other arthropod genomes and the assignment of putative functions based on annotated *Drosophila* one-to-one orthologs. Moreover, it enabled the discovery of a set of interesting gene expansions related to the particular diet and developmental biology of this insect. In another project, a phylome was reconstructed for two strains of the recently-sequenced halophilic bacterium *Salinibacter ruber* (15). The aim here was to understand the process of sympatric speciation by detecting differences in patterns of horizontal gene transfer, duplication and positive selection between two strains that were isolated simultaneously from the same environmental sample. Other ongoing sequencing projects are currently taking advantage of the PhylomeDB resources to address a variety of questions in different organisms. In this context, a password-protected pre-release section has been created to allow restricted access prior to publication (see below). We encourage interested groups and consortia to contact us for further details.

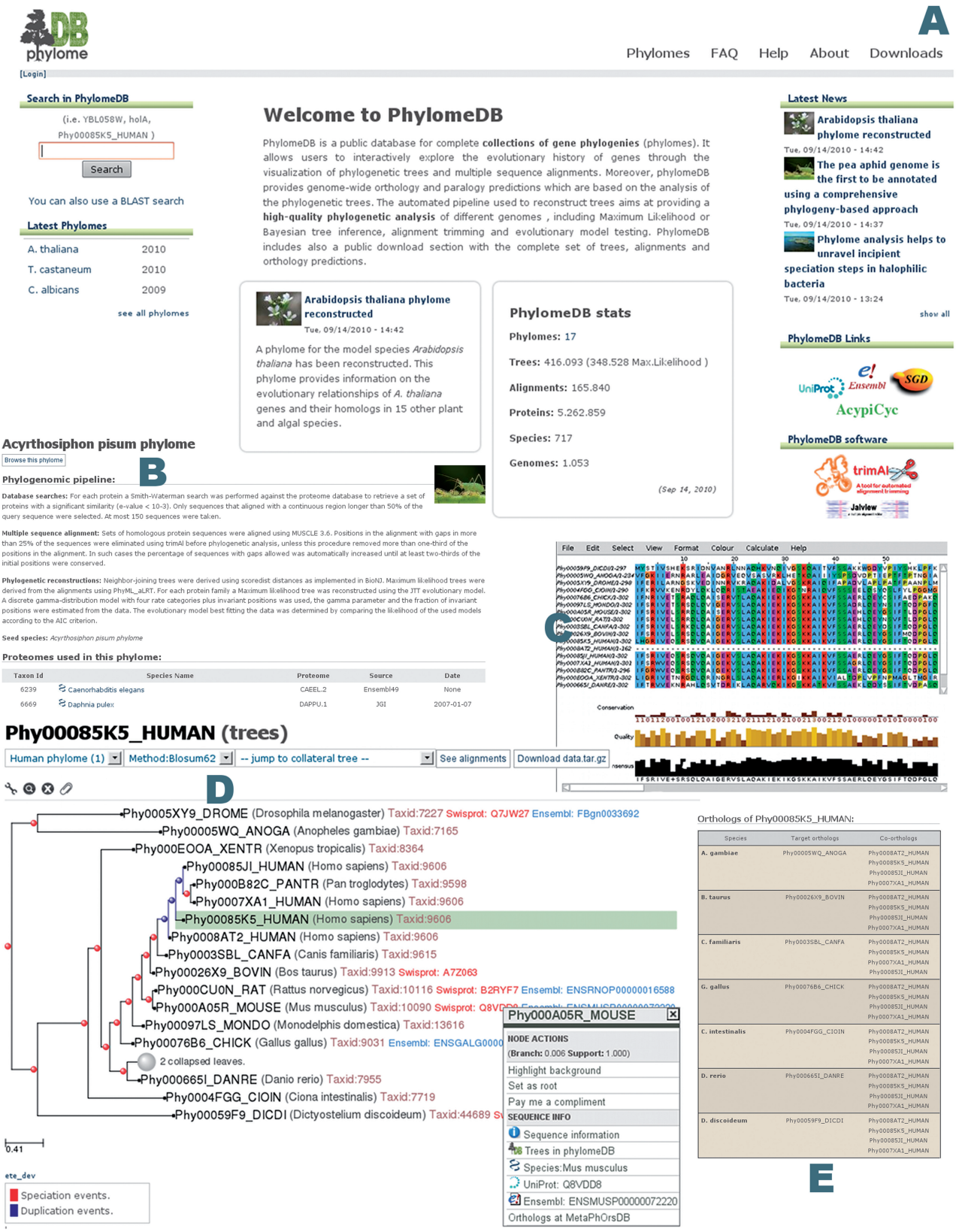

Figure 1. PhylomeDB web interface. (**A**) Entry page of phylomeDB, containing latest news, PhylomeDB statistics, login options and the main search panel. (**B**) Phylome description page. In the example, the phylogenetic pipeline and proteomes used to reconstruct the *A. pisum* phylome are shown. A Link is provided to browse all phylome resources (**C**) A multiple sequence alignment visualized through the integrated Jalview plugin. (**D**) Interactive tree visualization panel showing a particular tree from the human phylome. This panel allows online manipulation of tree topology and provides links to other related phylogenies in phylomeDB. Each sequence in the tree is also linked to external databases such as Ensembl, UniProtUniProt or MetaPhors. (**E**) Orthology prediction panel associated to each active tree. Phylogeny-based orthologs and paralogs of the seed sequence are shown sorted by species.

New features in PhylomeDB v3.0

Visualization of alignments and phylogenetic trees

Current version of PhylomeDB uses ETE v2.0 (28) to visualize phylogenetic trees, and Jalview Lite v2.4 (29) and trimAl v1.2 (24) for the visualization of raw and processed alignments. The visualization of trees and alignments is interactive and users can choose among various display options, such as collapsing parts of a tree or enabling/disabling the display of alternative IDs. Moreover, text strings (e.g. a species or a protein name) can be searched within phylogenetic trees. In addition, further information for each sequence, including hyperlinks to other databases, can be accessed by clicking on the corresponding node.

PhylomeDB unique sequence ID system

A new unique ID system has been developed for this version of PhylomeDB. This solves previous issues related to the inclusion of newer versions of existing proteomes and makes the information on the sequence species source more intuitive. In addition, it ensures that the same gene will receive the same ID in subsequent genome versions, unless the sequence has been updated. All PhylomeDB IDs (e.g. Phy0008C1X_HUMAN) start with the code 'Phy', followed by an alphanumerical string of length 7, an underscore symbol '_', and an alphanumeric species code. This species code corresponds to that assigned by UniProtKB in the 'controlled vocabulary of species' (30) or, when no code is present in UniProt, to the NCBI taxonomic ID (http://www.ncbi.nlm.nih.gov/Taxonomy/). For consistency, older PhylomeDB IDs are still associated to their corresponding sequences and are still searchable. Finally, PhylomeDB IDs are regularly mapped to IDs from other databases such as Ensembl (31) and UniProt (30) and corresponding conversion tables are provided in the download section.

Cross-links to other databases

The possibility of external linkage to phylomeDB has been improved and now phylomeDB is linked from many sequence, process and organism reference databases, including UniProt (30), EnsemblCompara (8), Saccharomyces Genome Database (SGD) (32), AphidBase (33), DeathBase (34) and PeroxisomeDB (35). Links to specific entries in phylomeDB are easily customizable and detailed instructions are provided in the help section of PhylomeDB. Thus, phylomeDB can also be regarded as a complementary resource providing evolutionary information for sequences maintained in other databases, and we encourage administrators of other databases to consider this possibility.

Seed and collateral trees

A sequence entry is now associated not only with the trees in which that sequence is used as a seed but also to all other trees that contain that sequence. These so-called 'collateral' trees may include phylogenies from the same phylome (e.g. trees in which a paralogous protein was used as a seed), but also trees from other phylomes that contain that sequence. This provides users with additional information on the evolution of the sequence of interest and may serve to evaluate whether a given scenario (e.g. an orthology relationship) is supported also by alternative trees. Indeed, partially overlapping phylogenetic trees from PhylomeDB and other phylogenetic databases are explored by the MetaPhOrs database (http://orthology.phylomedb.org) to provide consistency-based confidence scores to phylogeny-based orthology and paralogy predictions (36).

Data download and programmatic access

A FTP-based download section has been developed to provide easy access to files containing all alignments, trees and orthology and paralogy predictions associated to every public phylome. In addition, ID conversion tables to UniProt, Ensembl and other major sequence repositories are also provided. Alternatively, for each phylomeDB entry, a compressed folder containing all information associated with the corresponding sequence can be downloaded. Finally, an Application User Interface (API) for accessing phylomeDB is available through the ETE software (28), a python programming toolkit that assists in the automated manipulation, analysis and visualization of hierarchical trees. PhylomeDB interface uses ETE to handle tree manipulation, for the interactive visualization, and to operate with the main MySQL database. By using ETE libraries implemented in the API, users can connect to phylomeDB and search for pre-computed gene phylogenies, download complete phylomes or obtain the orthology and paralogy predictions provided by the database. This allows programmatic access to PhylomeDB, as well as the automation of any downstream analysis. An added advantage of using phylomeDB API is that phylogenetic trees can be directly downloaded as ETE tree objects, thus enabling visualization or complex tree exploration within the same scripts. Interactive web tree visualization is also scheduled to be released soon as part of the ETE package.

Pre-release section

A log-in protected private section has been created to store pre-release versions of phylomes, so that they can be used on-line before publication. This option is mainly used now by genome sequencing consortia that generate phylomes within their annotation pipeline. PhylomeDB has currently 13 private phylomes that will be released during the following months.

CONCLUSIONS/PERSPECTIVES

With 17 public phylomes comprising 416 093 phylogenetic trees, the new version of PhylomeDB constitutes one of the major and most comprehensive public repositories of phylogenetic trees (compared to 122 002 trees in ensemble, including ensemble compara v59 and ensemble genomes release 5). Regular updates of model-species based phylomes are planned when newest genome releases include significant improvements in terms of sequence quality and coverage. Contrary to most other databases,

PhylomeDB follows a gene-centric approach and stores phylomes focused on selected seed genomes. This ensures maximum coverage of the targeted genomes and allows specifically designing the taxonomic scope in order to address different questions. A particular application of PhylomeDB is to provide support for large-scale phylogenetic analyses used in genome annotation projects. This has provided added value in terms of evolutionary insights and allows going beyond standard blast-based automatic annotation of genes. Finally, one of the main aims of PhylomeDB is to provide phylogeny-based orthology and paralogy predictions, covering about 5 000 000 proteins in 717 fully-sequenced genomes.

FUNDING

Spanish Ministry of Science (GEN2006-27784-E/PAT, BFU2009-09168). Funding for open access charge: CRG, Spanish Ministry of Science and Innovation.

Conflict of interest statement. None declared.

REFERENCES

1. Eisen,J.A. and Fraser,C.M. (2003) Phylogenomics: intersection of evolution and genomics. *Science*, **300**, 1706–1707.
2. Delsuc,F., Brinkmann,H. and Philippe,H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.*, **6**, 361–375.
3. Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.
4. Kuzniar,A., van Ham,R.C., Pongor,S. and Leunissen,J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.*, **24**, 539–551.
5. Gabaldón,T., Marcet-Houben,M., Huerta-Cepas,J. and Russe,A. (2008) Reconstruction and analysis of large-scale phylogenetic data: challenges and opportunities. *Computational Biology: New Research*. New York, Nova Science Publishers, pp. 129–146.
6. Huerta-Cepas,J., Dopazo,H., Dopazo,J. and Gabaldón,T. (2007) The human phylome. *Genome Biol.*, **8**, R109.
7. Huerta-Cepas,J., Bueno,A., Dopazo,J. and Gabaldón,T. (2008) PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.*, **36**, D491–496.
8. 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.
9. Muller,J., Szklarczyk,D., Julien,P., Letunic,I., Roth,A., Kuhn,M., Powell,S., von Mering,C., Doerks,T., Jensen,L.J. *et al.* (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–195.
10. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Heriche,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–740.
11. Sicheritz-Ponten,T. and Andersson,S.G. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res.*, **29**, 545–552.
12. Gabaldón,T. and Huynen,M.A. (2003) Reconstruction of the proto-mitochondrial metabolism. *Science*, **301**, 609.
13. Comas,I., Moya,A. and Gonzalez-Candelas,F. (2007) From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-Proteobacteria as a test case. *Syst. Biol.*, **56**, 1–16.
14. Marcet-Houben,M. and Gabaldón,T. (2009) The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PLoS One*, **4**, e4357.
15. Peña,A., Teeling,H., Huerta-Cepas,J., Santos,F., Yarza,P., Brito-Echeverria,J., Lucio,M., Schmitt-Kopplin,P., Meseguer,I., Schenowitz,C. *et al.* (2010) Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting Salinibacter ruber strains. *ISME J.*, **4**, 882–895.
16. Huerta-Cepas,J., Marcet-Houben,M., Pignatelli,M., Moya,A. and Gabaldón,T. (2010) The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for Acyrthosiphon pisum genes. *Insect Mol. Biol.*, **19(Suppl. 2)**, 13–21.
17. Gabaldón,T. (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.*, **9**, 235.
18. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
19. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
20. Katoh,K., Asimenos,G. and Toh,H. (2009) Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.*, **537**, 39–64.
21. Subramanian,A.R., Hiran,S., Steinkamp,R., Meinicke,P., Corel,E. and Morgenstern,B. DIALIGN-TX and multiple protein alignment using secondary structure information at GOBICS. *Nucleic Acids Res.*, **38(Suppl.)**, W19–W22.
22. Landan,G. and Graur,D. (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.*, **24**, 1380–1383.
23. Wallace,I.M., O'Sullivan,O., Higgins,D.G. and Notredame,C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.
24. Capella-Gutierrez,S., Silla-Martinez,J.M. and Gabaldón,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
25. Akaike,H. (1973) Information theory and extension of the maximum likelihood principle. In Petrov,B.N. and Csaki,F. (eds), *Proceedings of the 2nd International Symposium on Information Theory*. Academiai Kiado, Budapest, Hungary, pp. 267–281.
26. Keane,T.M., Creevey,C.J., Pentony,M.M., Naughton,T.J. and McLnerney,J.O. (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.*, **6**, 29.
27. Consortium,I.A.G. (2010) Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS Biol.*, **8**, e1000313.
28. Huerta-Cepas,J., Dopazo,J. and Gabaldón,T. (2010) ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, **11**, 24.
29. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
30. Consortium,T.U. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
31. Kersey,P.J., Lawson,D., Birney,E., Derwent,P.S., Haimel,M., Herrero,J., Keenan,S., Kerhornou,A., Koscielny,G., Kahari,A. *et al.* Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.
32. Engel,S.R., Balakrishnan,R., Binkley,G., Christie,K.R., Costanzo,M.C., Dwight,S.S., Fisk,D.G., Hirschman,J.E., Hitz,B.C., Hong,E.L. *et al.* Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res.*, **38**, D433–D436.
33. Legeai,F., Shigenobu,S., Gauthier,J.P., Colbourne,J., Rispe,C., Collin,O., Richards,S., Wilson,A.C., Murphy,T. and Tagu,D. AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Mol. Biol.*, **19(Suppl. 2)**, 5–12.
34. Diez,J., Walter,D., Munoz-Pinedo,C. and Gabaldón,T. (2010) DeathBase: a database on structure, evolution and function of proteins involved in apoptosis and other forms of cell death. *Cell Death Differ.*, **17**, 735–736.
35. Schluter,A., Real-Chicharro,A., Gabaldón,T., Sanchez-Jimenez,F. and Pujol,A. (2010) PeroxisomeDB 2.0: an integrative view of the global peroxisomal metabolome. *Nucleic Acids Res.*, **38**, D800–D805.
36. Pryszcz,L., Huerta-Cepas,J. and Gabaldón,T. (2010) MetaPhOrs: orthology and paralogy predictions from multiple phylogennetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, in press; doi:10.1093/nar/gkq953.