Review

The quest for orthologs: finding the corresponding gene across genomes

Arnold Kuzniar[1], Roeland C.H.J. van Ham[1], Sándor Pongor[2] and Jack A.M. Leunissen[1]

[1] Laboratory of Bioinformatics, Wageningen University, Dreijenlaan 3, 6703 HA Wageningen, the Netherlands
[2] Protein Structure and Bioinformatics, International Centre for Genetic Engineering and Biotechnology, AREA Science Park, Padriciano 99, 34012 Trieste, Italy

Orthology is a key evolutionary concept in many areas of genomic research. It provides a framework for subjects as diverse as the evolution of genomes, gene functions, cellular networks and functional genome annotation. Although orthologous proteins usually perform equivalent functions in different species, establishing true orthologous relationships requires a phylogenetic approach, which combines both trees and graphs (networks) using reliable species phylogeny and available genomic data from more than two species, and an insight into the processes of molecular evolution. Here, we evaluate the available bioinformatics tools and provide a set of guidelines to aid researchers in choosing the most appropriate tool for any situation.

The concept of orthology

In the early days of comparative biology, relationships between different species were studied using morphological characters. With the emergence of sequencing techniques and, in particular, the high-throughput techniques of the past decade, the amount of molecular characters in the form of fully sequenced genomes from a diverse range of organisms has increased enormously. A wide array of bioinformatics tools has been developed to interpret the sequence data from evolutionary and functional perspectives [1]. The knowledge of molecular phylogenies in general and orthology in particular has become an integral component of many genome-scale studies of gene content, conserved gene order and gene expression, regulatory networks, metabolic pathways and in functional genome annotation [2–12].

The concept of homology (see Glossary) is fundamental to make inferences about evolutionary processes such as

Glossary

Conserved gene neighborhood (CGN): refers to conserved genomic segments containing orthologous genes in a similar collinear order between species. Sometimes, the term *conserved synteny* is used instead, which originally denoted gene loci on the same chromosome regardless of whether or not they are genetically linked. Respecting the original definition of 'synteny' and its etymology, we therefore use the term 'conserved gene neighborhood' [79].
Co-orthologs: two or more sequences in one lineage that are collectively orthologous to one or more sequences in another lineage owing to a lineage-specific duplication(s).
Homology: refers to a testable hypothesis that characters in different species sharing significant sequence similarity (at least 30–35% as a rule of thumb for protein sequences) descend from a single common ancestral character. Sequences that are evolutionarily related to each other in this way are known as homologs. Note that homology is independent of the size and molecular nature of a biological sequence.
Horizontal gene transfer (HGT): an evolutionary process that involves transfer of genetic material between species but does not follow the vertical descent from a parental lineage to its offspring. HGT is an important phenomenon in the evolution of prokaryotes and eukaryotes [66–68].
In-paralogs: paralogs that result from a lineage-specific duplication(s) subsequent to a given speciation event (sometimes termed 'recent' paralogs). They are likely to have retained similar functions within a species.
Non-transitivity of phylogenetic relationships: orthology, paralogy and xenology are strictly pairwise and non-transitive relationships between (groups of) genes. This can best be understood using the following example: if two genes, *a* and *b,* are equally (co-) orthologous to gene *c,* it does not imply that *a* and *b* must also be orthologous to each other [14]. Therefore, an OG must always be hierarchical and defined with respect to the last common ancestor of the investigated genes (taxonomic position).
Orthologous group of genes (OG): a collection of homologous genes from at least two species. After a duplication event, an OG might group paralogs and orthologs together. Therefore, an OG must be defined within a phylogenetic tree in the context of speciation and duplication events to guarantee the non-transitivity of phylogenetic relationships. If an OG consists of single-copy orthologous genes, then all of the genes can be grouped together because the phylogenetic relationships between all of them are equivalent.
Orthologs: homologous sequences derived by a speciation event from a single ancestral sequence in the last common ancestor of the species being compared. Orthologs typically perform equivalent functions in closely related species.
Out-paralogs: paralogs resulting from a duplication(s) preceding a given speciation event (sometimes termed 'ancient' paralogs). They are likely to have different functions.
Paralogs: homologous sequences derived by a duplication event from a single sequence. Paralogous relationships occur both within and between genomes. Paralogs can evolve novel functions and are likely to have mechanistically distinct but biologically related functions.
Subtree-neighbors: homologs in a rooted gene tree that are found at a particular level (parent node) of the tree [38].
Super-orthologs: a subset of orthologs selected on a rooted gene tree such that only speciation events are assigned to each internal node on their connecting path [38].
Ultra-paralogs: a subset of paralogs selected on a rooted tree such that its internal nodes connecting them represent only duplication events (in-paralogs) [38].
Xenologs: homologous sequences, the history of which involves transfer of genetic information between species (see horizontal gene transfer or HGT). They often appear as true orthologs in genome comparisons and might exhibit variable functions [80].

Corresponding author: Leunissen, J.A.M. (jack.leunissen@wur.nl).

0168-9525/$ – see front matter doi:10.1016/j.tig.2008.08.009 Available online 24 September 2008

speciation, gene duplication or horizontal gene transfer (HGT). At the beginning of the 1970 s, Walter Fitch divided homology into orthology and paralogy according to the distinct evolutionary processes, namely speciation and gene duplication, respectively [13,14]. Thus, orthologs are homologous genes that relate through speciation from a single ancestral gene present in their latest common ancestor, whereas paralogs are homologs that arose through gene duplication. Nonetheless, an understanding of homology, orthology and paralogy has been challenged by other important evolutionary processes such as HGT and gene fusion or fission events, which are thought to have enabled the formation of complex phylogenetic networks [15,16]. Several terms (e.g. in-paralogs, out-paralogs, super-orthologs or ultra-paralogs) have been coined to further refine the various evolutionary origins of sequence similarities. The term 'orthology' is often misunderstood to refer to functionally equivalent genes in different species; but, it is strictly an evolutionary concept, rather than a functional one [14]. Orthologs have primarily been used as evolutionary markers for inferring species phylogenies because they follow species divergence [17,18], but they can be used to link functionally equivalent genes across genomes and, as such, enable the function of an unknown protein to be inferred using known (i.e. functionally characterized) orthologs in other species [5,19]. However, the main caveats of using orthologs in function annotation are domain shuffling, presence or absence of a domain, lineage-specific gene duplication and gene loss [20]. Controlled vocabularies (ontologies) have emerged to describe biological functions (e.g. gene functions, mode and site of action within a cell) in a standardized form and have intensively been used to link heterogeneous datasets of various molecular databases [21–23]. For example, databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/), BioCyc (http://biocyc.org/) or IMG (http://img.jgi.doe.gov/) integrate molecular data on pathways, enzymes and substrates associated with orthologous genes (proteins) from diverse genomes [24–26].

Here, we review the computational tools (i.e. programs and databases) commonly used to infer orthologous relationships between genes and proteins (Boxes 1–3). Here, we compare the orthology detection tools and demonstrate the advantages and/or limitations of these methods using real examples of gene families and evolutionary scenarios. Also proposed is a set of guidelines to aid researchers in selecting the correct tool in a given situation.

Classification and critical review of orthology detection methodologies

For the purpose of this review, a classification scheme that recognizes both conceptual and practical differences among orthology detection tools available to date has been introduced (Figure 1). The different tools are grouped along methodological lines: those based on trees (tree-based methods), graphs (network or graph-based methods) or both (hybrid methods). From a practical point of view, this classification distinguishes between *ab initio* and post-processing tools. The former example infers orthologs in entire sets of genes (proteins) of two or more species and

Box 1. Tree-based methods

Correlation Coefficient-based Clustering (COCO-CL)
The COCO-CL program takes the non-transitivity of phylogenetic relations within a set of homologous proteins into account using a hierarchical numbering scheme. [31]. It uses a heuristics based upon Pearson's correlation matrix of sequence distances to decides upon speciation and duplication events without a species tree. Sets containing out-paralogs are recursively split into two smaller subsets until no additional out-paralogs are found, thus forming a hierarchy of sets. Each split is flagged as either speciation or duplication according to its reliability (bootstrap) score. Pros: COCO-CL infers orthologs and paralogs from pre-computed homologs in a hierarchical framework without a species tree. The COCO-CL program and refined COG dataset are freely available. Cons: COCO-CL does not implement a tree-reconciliation algorithm.

Orthostrapper and Hierarchical grouping of Orthologous and Paralogous Sequences (HOPS)
The Orthostrapper program uses a heuristic sequence similarity search to infer orthologs with confidence values from a set of bootstrapped gene trees [32]. Orthostrapper does not use a species tree in a strict sense. Instead, sequences are assigned to a taxonomic group. The HOPS database provides orthology assignments for eukaryotic Pfam domains [33]. Pros: HOPS provides domain-based orthologs. The Orthostrapper program is freely available. Cons: HOPS dataset is not available for download and the web server does not work.

Levels of Orthology From Trees (LOFT)
The LOFT program addresses the non-transitivity of phylogenetic relations within phylogenetic trees [34]. It implements two algorithms to infer speciation or duplication events in a given gene tree. Besides the SDI tree-reconciliation algorithm, LOFT offers an alternative approach, the so-called 'species-overlap' rule, especially when the species tree is not known. This simple heuristics implies that a speciation event is only assigned to an internal node if its branches contain mutually exclusive sets of species. LOFT makes a use of a hierarchical numbering scheme for orthologous groups (similar to that found in COCO-CL). Pros: LOFT infers orthologs and paralogs from pre-computed homologs in a hierarchical framework without a species tree. The LOFT program comes with a GUI. Both the program and the refined COG dataset are freely available. Cons: LOFT cannot be executed without the GUI as a command line tool. The 'species-overlap' is not adjustable.

Réconciliateur d'Arbres Phylogénétiques (RAP)
Originally, the RAP tree-reconciliation program (http://pbil.univ-lyon1.fr.) [35] was used to infer orthologs in HOVERGEN and HOBACGEN [36,37] databases. Pros: The algorithm can handle unresolved trees and take both bootstrap values and branch lengths into account for the reliability of trees. The RAP program is freely available. Cons: RAP cannot be used as a command line tool.

Speciation Duplication Inference (SDI) and Resample Inference of Orthologs (RIO)
The SDI tree-reconciliation algorithm requires properly rooted and completely binary input trees to infer speciation and duplication events reliably. The orthology assignments in the RIO database [38] were made by using the *Pfam* protein domains and SDI algorithm on bootstrap re-sampled gene trees [39]. A confidence (orthology bootstrap) score is given for each database hit. High scores indicate 'true' orthology, whereas low values indicate absence of orthologs. Three novel homology concepts were introduced to enhance function prediction of genes (Box 1; super-orthologs, ultra-paralogs and subtree-neighbors). Pros: RIO provides phylogenetic resolution for domain-based orthologs with confidence scores. The SDI algorithm is freely available. Cons: RIO data are not available and the web server is not operational. SDI cannot root the input trees and requires fully resolved trees

the latter two use pre-computed homologs to infer orthologs and paralogs. Furthermore, a distinction is made between the methods that use exclusively primary sequence data and those that also use auxiliary information, such as conserved gene neighborhood (CGN). Although CGN might assist in finding additional orthologs when inference of homology is hampered by low sequence similarity [27], or in distinguishing true orthologs from single-copy paralogs (out-paralogs) in the presence of reciprocal gene losses [28,29], it is applicable only to closely

Box 2. Graph-based method

Nearest neighbor
We use the term 'nearest neighbor' to collectively designate all approaches that apply an operational definition of orthology. Even though the approaches do not necessarily imply phylogenetic proximity [40], they are commonly used as first-pass approximations to find putative orthologs using some 'flavor' of the 'best' genome-wide matches between two species. These methods include best hit (BeT), reciprocal best hit (RBH), bi-directional best hit (BBH), symmetrical best hit (SymBeT) and reciprocal smallest distance (RSD) [6,41–45]. The nearest-neighbor methods might also address one-to-many and many-to-many orthologous relations depending on which definition is used and how it is implemented in the computation. The key concepts are best understood using graph theory (Figure I). Clearly, the RBH approach using different similarity measures might result in distinct, but largely overlapping, sets of orthologs.

Clusters of Orthologous Groups (COGs) of proteins
The COG approach extends best BLAST hits (BeTs) to multiple proteomes by using congruent 'triangles' of BeTs from at least three different species [5,6]. These minimal COGs are then merged by a single linkage into larger groups (protein families). The database consists of two sections for unicellular (mainly prokaryotes) and eukaryotic proteomes (euKaryotic Orthologous Groups or KOGs) from 66 fully sequenced genomes. Pros: The COG database is a widely used resource for functional annotation of genomes, mainly owing to availability and manual curation. COGs are functionally annotated. The COG database stores orthologous groups from prokaryotic and eukaryotic genomes. Cons: The 'triangles' of the COG are disadvantageous in the presence of gene losses. The COG approach does not differentiate between in- and out-paralogs automatically; therefore, one needs to investigate the pre-computed phylogenetic trees for duplication and speciation events. The automatic clustering procedure creates exclusive clusters, thus, multi-domain proteins must be handled manually. The database has not been updated since 2003

Eukaryotic Gene Orthologs (EGO)
The EGO (previously known as TIGR Orthologous Gene Alignments or TOGA) database (http://compbio.dfci.harvard.edu/tgi/tgi/ego/) is constructed by an orthology detection procedure similar to that of the COG system [44], but instead of proteins, it uses virtual assemblies of transcripts, which provide evidence of a gene at the transcription level. Pros: The EGO database is freely available and contains more genomes (89) than COG. Cons: It has similar disadvantages as the COG approach and it does not have functional annotations.

InParanoid
The InParanoid program distinguishes between in-paralogs and out-paralogs for two proteomes without using phylogenetic trees [45]. Instead, the method implements a set of heuristic rules to merge, delete and separate predicted orthologous groups. First, the main orthologs are identified as protein pairs having the highest symmetric BLAST score and are used as 'seeds' for finding all in-paralogs for each species. InParanoid and OrthoDisease databases store orthology assignments mainly of eukaryotic species (35) [46,47]. Pros: InParanoid addresses one-to-many and many-to-many orthologous relationships between two proteomes. It also enables an out-group species. Confidence values are assigned to individual in-paralogs and orthologous groups as a whole. The program and the database are freely available. Cons: InParanoid is limited to pair-wise proteome comparisons and does not permit overlapping clusters in the presence of a hybrid protein.

MultiParanoid
The MultiParanoid program (http://www.sbc.su.se/~andale/multiparanoid/html/index.html) constructs multi-species orthologous groups of proteins from all possible pairwise species InParanoid comparisons. The clustering is less stringent (a single-linkage approach) than that of the approach of COG [48]. Pros: MultiParanoid constructs multi-species orthologous groups. The program and the dataset of four eukaryotic species is freely available. Cons: MultiParanoid can be used for only a few species, which diverged at roughly the same time point from a common ancestor, otherwise the approach becomes inclusive for out-paralogs. It does not address the non-transitivity of phylogenetic relations. The web server is broken; a major update is planned (JL, personal communication).

Ortholuge
The Ortholuge program (URL: http://www.pathogenomics.ca/ortholuge) is designed to improve the specificity of RBH-based orthology predictions by handling gene-loss events for both bacterial and eukaryotic species [49]. The method is similar to InParanoid but it uses phylogenetic distance ratios instead of BLAST similarities. Pros: Ortholuge can use pre-computed (tentative) orthologs or construct a dataset using an RBH-based BLAST approach. It is freely available. Cons: Ortholuge predictions of orthologs are incomplete in the presence of single gene loss. Ortholuge is limited to pair-wise proteome comparisons.

OrthoMCL and OrthoMCL–DB
The OrthoMCL pipeline integrates a Markov Cluster algorithm (MCL) for grouping proteins into multi-species orthologous groups (S. van Dongen, PhD thesis, University of Utrecht, 2000) [50]. First, 'seed' orthologs and in-paralogs are found using a similar approach to that of InParanoid and clustered using the MCL algorithm. Similarities between proteins are calculated as normalized BLAST P-values. The OrthoMCL-DB database stores orthologs of mainly eukaryotic genomes (87 species) [51]. Pros: The OrthoMCL program constructs multi-species orthologous groups, which can be queried by phylogenetic patterns (presence and absence of species). The program and the database are freely available. Cons: OrthoMCL does address the non-transitivity of phylogenetic relations within orthologous groups. It might group out-paralogs and orthologs together in the presence of gene losses and does not handle hybrid proteins. The groups do not have function annotations.

Reciprocal Smallest Distance (RSD) and RoundUp
The RSD approach combines local and global sequence alignments and maximum likelihood estimation of evolutionary distances together to predict orthologous proteins [43]. The RoundUp repository encompasses pairwise species orthologs from >250 genomes at various threshold levels of BLAST *E*-values and sequence divergence [52]. Pros: RSD uses explicit evolutionary model to calculate distances between proteins. The RoundUp database covers wide range of species. Cons: RSD cannot compare more than two genomes simultaneously and does not permit the use of an out-group species.

Best Unambiguous Subset (BUS)
The BUS algorithm detects groups of orthologs between two genomes using a single linkage graph clustering (M. Kellis, PhD thesis, Massachusetts Institute of Technology, 2003). Graph edges are weighted by the amino acid sequence identity and the overall length of BLAST matches. An orthologous group consists only of genes that have 'best' matches within the group and no 'best' matches of any gene are outside that group. Pros: BUS makes a use of CGN to find additional putative orthologs, and can handle incomplete (draft) genomes. Cons: BUS is limited to pair-wise genome comparisons and is not available online.

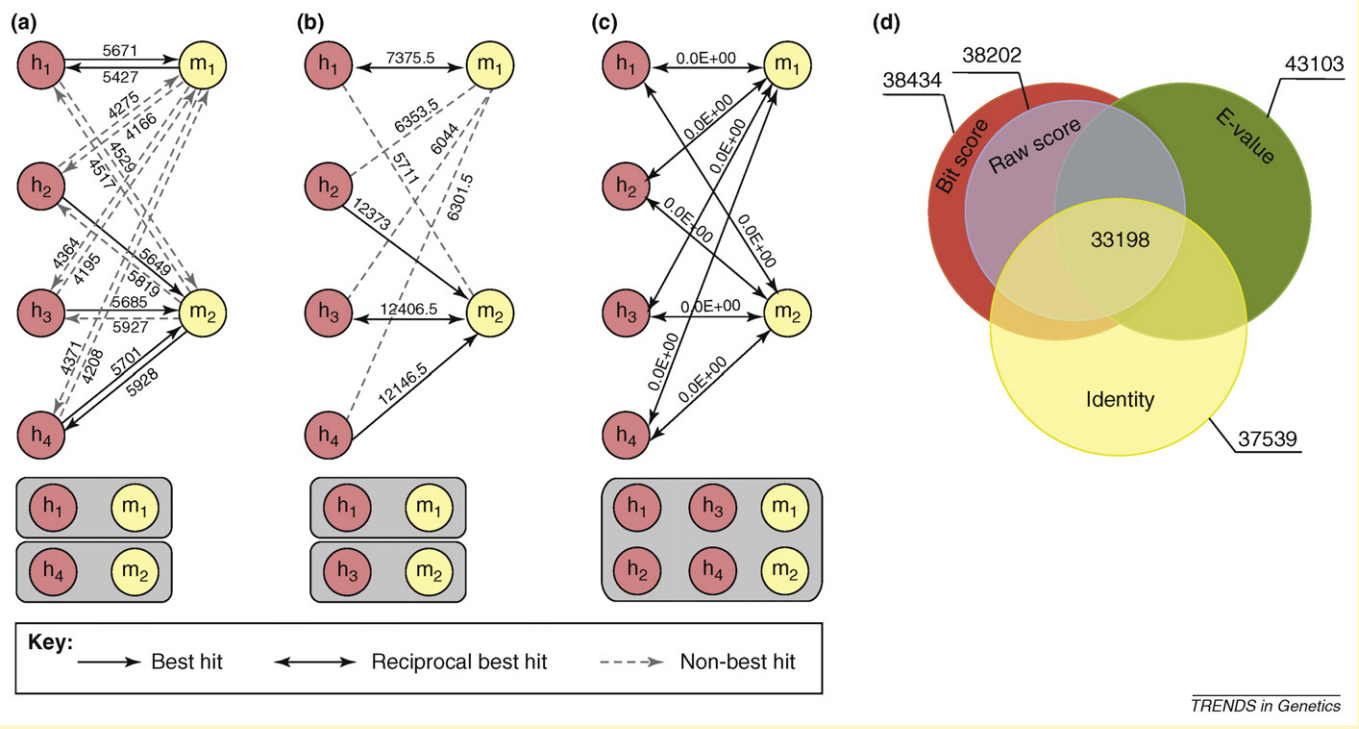


Figure I. Different sets of putative orthologs defined as reciprocal best hits. Three graphs of human (h_1–h_4) and mouse (m_1- and m_2) mucin-5 proteins are constructed using three different protein similarity measures: **(a)** asymmetric BLAST raw score; **(b)** symmetric Smith-Waterman score; and **(c)** symmetric BLAST E-value. The corresponding set of predicted orthologs is shown below each graph. Clearly, the reciprocal best hit approach using different similarity measures might result in different but largely overlapping sets of orthologs. **(d)** Venn diagram of four different sets of orthologs, using BLAST identity, E-value, raw and bit score, are inferred from complete human and mouse proteomes (Refseq version 29). The total number of orthologs is indicated for the sets and four-way intersection. Graph nodes correspond to RefSeq protein accessions: h_1, XP_001717932; h_2, NP_059981 (*Muc5ac*); h_3, NP_002449 (*Muc5b*); h_4, XP_001719401; m_1, NP_034974 (*Muc5ac*); m_2, NP_083077 (*Muc5b*).

related species [30]. The merits and pitfalls of various orthology detection tools are summarized in Boxes 1–3 [5–7,31–58].

Tree-based methods

Tree-based methods infer orthologous and paralogous relationships from phylogenetic trees. First, one must collect homologous sequences, construct a multiple-sequence alignment and phylogenetic tree(s) and then, the relationships can be analyzed either in the presence or absence of 'known' phylogenetic relations between species (e.g. mouse, rat and human). Because a gene tree does not necessarily have the same topology as the species tree, owing to evolutionary processes such as gene loss and HGT, tree-reconciliation techniques, which infer speciation (orthologs) and duplication (paralogs) events from reconciled trees, have been commonly used to account for these differences [35,39,59,60]. However, this approach can only be used when the species tree is reliable. This poses the question of how one deals with those cases in which the phylogenetic relationships between species are not known. Recently, two methods, namely the Correlation Coefficient-based Clustering (COCO-CL; http://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/COCOCL/) and the Levels of Orthology From Trees (LOFT; http://www.cmbi.ru.nl/LOFT/), have been proposed to distinguish between orthologs and paralogs in a gene tree without using a corresponding species tree [31,34].

The current tree-based methods have several shortcomings. First, phylogenetic-tree reconstruction algorithms rarely produce completely reliable trees. Ambiguities in either a gene tree or a species tree result in a spurious inference of duplication and speciation events. However, one can use sampling methods, such as bootstrap [61] or Markov Chain Monte Carlo (MCMC) [62] methods to assess the reliability of the tree. Second, the tree-based algorithms require properly rooted trees, which are commonly rooted by the midpoint in the tree or by the careful manual selection of an out-group species. Midpoint rooting approaches are often problematic for protein families in which members evolve at different rates, whereas the manual selection of out-groups might be impractical and difficult to automate, especially for large-scale genome analyses [39]. Alternatively, the trees can be rooted by an approach that minimizes dissimilarity between the gene and species trees [60]. Third, a plausible phylogenetic gene tree depends on a biologically correct multiple-sequence alignment. Therefore, incorrect alignments draw false conclusions about evolution. Finally, algorithms for phylogenetic-tree construction and multiple-sequence alignment scale poorly with the increasing amount of sequence data available and are not suitable for complete genomes. Although the computational cost can be reduced with heuristic algorithms, or deploying parallel algorithms on distributed systems, it is challenging to construct reliable sequence alignments and trees for large gene

Box 3. Hybrid methods

Ensembl Compara
The database provides comparative genome and proteome data for >30 eukaryotic species, mainly mammals [53]. The orthology prediction pipeline combines both BLAST-based RBHs and a phylogenetic tree reconciliation. Pros: The orthology uses a phylogenetic approach for handling gene losses. Orthologous relationships are labeled as one-to-one, one-to-many and many-to-many. Moreover, additional orthologs can be inferred in the genome context using whole-genome alignments. The Ensembl Compara database is regularly updated, freely available and accessible through several interfaces. Cons: The approach does not consider alternative transcripts for a gene, but assumes that a gene is best represented by the longest transcript or translation product.

HomoloGene
The HomoloGene database provides automatically predicted homologs of 19 completely sequenced eukaryotes (animals, plants and fungi) and includes cross-references to other resources on experimentally verified protein functions, conserved domains and phenotypic data [54]. The clustering procedure uses pre-computed BLAST protein similarities and CGN and is guided by a species phylogeny (starting from closely-related species). Aligned protein sequences are linked to their corresponding DNA sequences, from which non-synonymous-to-synonymous nucleotide substitution ratios are calculated to prevent inclusion of out-paralogs into groups. Paralogs are identified as sequences that are more similar within species than between species. Pros: HomoloGene groups are constructed using explicit species phylogeny and CGN and do not group unrelated proteins together in the presence of a hybrid protein. The database is regularly updated and freely available. Cons: HomoloGene groups are exclusive and lack plausible function annotations (only labeled by the last common ancestor of group members). The clustering procedure is not available.

OrthoParaMap (OPM)
The OPM package (http://www.tc.umn.edu/~cann0010/Software.html) integrates comparative genomic positional databased on BLAST comparisons and gene phylogenies to infer evolutionary processes in gene families from two species [55]. Unlike tree-reconciliation methods, OPM does not use a species tree but a conserved gene neighborhood (CGN) to decide upon speciation and duplication events. Pros: OPM incorporates CGN and distinguishes between segmental and tandem duplicates. The program is freely available. Cons: OPM cannot be used for more than two genomes simultaneously.

Phylogenetically inferred groups (PhIGs)
The PhIGs database (http://phigs.org) provides protein clusters, protein family trees and synteny maps for 23 completely sequenced genomes of fungi and metazoans [56]. Protein clusters are constructed using all-versus-all BLAST comparisons, calculations of protein distances from refined alignments and a hierarchical clustering guided by a species tree. A maximum likelihood protein family tree is inferred for each protein cluster. Pros: The clustering procedure takes species phylogeny into account. The web server provides visualization of synteny maps. Cons: Trees must be examined manually to infer speciations and duplications. The database has not been updated since its first release and is not available for download.

Phylogenetic orthologous groups (PHOGs)
The PHOG database stores clusters of orthologous groups (PHOGs) at various levels of the species tree from mainly prokaryotic genomes [57]. PHOGs are constructed by traversing the species tree from the leaves towards the root and finding BBH-based BLAST hits for each pair of species (proteomes). Only the highest-scoring protein pairs (seeds) within newly created PHOGs are aligned by Smith-Waterman algorithm and used in the next iteration. Pros: The PHOG approach constructs orthologous groups at various levels using species phylogeny. It incorporates automatic detection and handling of fusion events in multi-domain proteins. Cons: The database server is not available online.

Phylogenetic orthology and paralogy (PhyOP)
The PhyOP orthology prediction pipeline explicitly handles multiple transcripts per gene to reliably infer orthology and paralogy relationships between genes for recently diverged species [7]. First, clusters of transcripts are constructed using single linkage clustering based on BLAST protein similarities, protein-to-transcript mappings and synonymous nucleotide substitutions. In the next step, clusters are used to infer phylogenies of transcripts using a modified least-square distance-based method. A set of heuristic rules is applied to the phylogenies to detect orphan genes and to distinguish between functional genes and pseudogenes. Pros: The PhyOP pipeline takes multiple-transcripts per gene into account to predict orthologs. It can distinguish between functionally active and inactive genes (pseudogenes). Moreover, PhyOP is particularly useful in predicting orthologous genes for incomplete (draft quality) genomes. The program is available upon request. Cons: The PhyOP can only be used for two closely related genomes.

TreeFam
TreFam is a database of curated (TreeFam-A) and automatically constructed (TreeFam-B) animal gene families, phylogenetic trees, inferred orthologs and paralogs for fully sequenced animal genomes [58]. First, TreeFam clusters are created by hierarchical clustering of all-versus-all BLAST similarities and then gene family trees are constructed using several different approaches including maximum likelihood and neighbor-joining. Orthologs and paralogs are inferred using the Duplication/Loss Inference (DLI) tree-reconciliation algorithm, which uses the taxonomy tree of NCBI as a species tree. Pros: The orthology prediction uses a phylogenetic approach for handling gene losses. Speciation, duplication and gene-loss events are displayed in the phylogenetic trees. Experts manually correct errors in the automatically constructed trees. All data and software can be freely downloaded. Besides a web interface, users can access the TreeFam database directly. Cons: A gene is represented by one transcript.

families that have complex histories. In summary, tree-based methods provide phylogenetic resolution at multiple levels of a gene tree and are suitable to infer orthologs and paralogs from any protein (domain) family database available. However, these approaches are computationally intensive for large datasets, not easily automated owing to the need to choose appropriate outgroup species and depend on the pre-defined protein families.

Graph-based methods

Graph-based methods are suitable for orthology inferences from two or more complete genomes (proteomes). Unlike tree-based methods, they do not construct multiple-sequence alignments and phylogenetic trees, but rely on pairwise sequence similarities calculated between all sequences involved and an operational definition of orthology, for example, reciprocal best hits (RBHs) (Box 2). The choice of a sequence-similarity search algorithm [e.g. basic local alignment search tool (BLAST) or Smith–Waterman] and a scoring scheme for pairwise alignments has a bearing on the sensitivity and specificity of orthology predictions [63]. Some graph-based methods use clustering techniques (e.g. single-linkage, complete-linkage or Markov Cluster algorithm [64]) to extend nearest neighbors to more than two species and construct multi-species orthologous groups (OGs) of particular granularity [65]. These approaches use the definition of orthology liberally because orthologs and paralogs are often grouped together in an

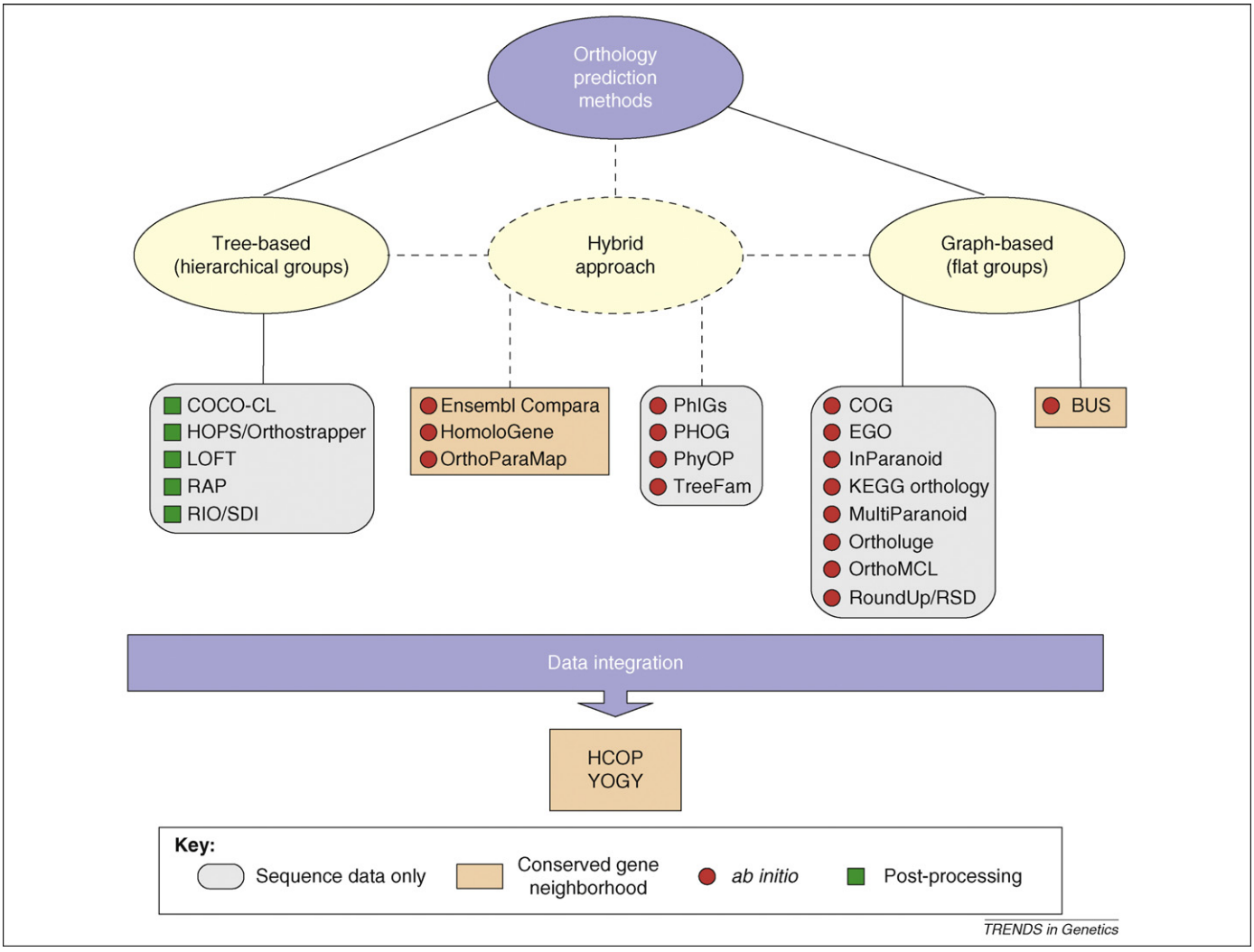


Figure 1. Classification of orthology detection methods. Three main categories are recognized according to the data representations they operate on, including tree-based, graph-based and hybrid methods (see main text for a full description). Further distinctions are based on conserved gene order (CGN) and *ab initio* or post-processing approaches. Data integration does not offer a new algorithmic approach *per se*, but is used to merge multiple datasets, which include both experimentally verified and automatically predicted orthologs, into a unified, consolidated collection. The examples of integrated databases include HUGO gene nomenclature committee (HGNC) Comparison of Orthology Predictions (HCOP; http://www.genenemas.org/) and Eukaryotic Orthology (YOGY; http://www.sanger.ac.uk/PostGenomics/S_pombe/YOGY) [75,81]. A comparison of tree-based, graph-based and hybrid methods is given in Boxes 1–3, respectively.

OG, in which all members are collapsed down to the last common ancestor of all species in that OG. However, this is not a concern for graph-based methods that analyze two species at once (either in the presence or absence of an outgroup) [43,45,49].

Hybrid methods

Hybrid methods make use of both tree and graph representations at various stages of processing; for example, to refine OGs within a hierarchical framework of phylogenetic trees or to guide the clustering procedure using a species tree [7,53–58]. Although all hybrid methods must incorporate phylogenies of some form, they are not required to use CGN (Figure 1). Because the hybrid approaches combine tree and graph-based methods by using the phylogenetic resolution of the former and the scalability of the latter, they are suitable for genome-wide analyses. Besides the advantages, one must be aware of which of these methods do not provide a phylogenetic resolution at multiple levels in *de novo* generated OGs [54,56].

Caveats of orthology detection

Mosaics of proteins

The fusion, fission, shuffling, gain and loss of protein domains are common processes in protein evolution, which give rise to protein chimeras or hybrids (i.e. a protein that consists of at least two distinct, non-homologous sequence regions, either in the form of a single domain or as a full-length protein). Hybrid proteins can complicate orthology assignments in a way illustrated by the bifunctional dihydrofolate reductase-thymidylate synthase gene (*DHFR-TS1)* from *Arabidopsis thaliana* (Figure 2). Importantly, OGs delineated without considering the possibility of hybrids run the risk of containing proteins that do not have a common evolutionary ancestry. Clearly, a hybrid protein can be legitimately similar to more than one OG. Therefore, grouping proteins into overlapping (non-exclusive) OGs is likely to provide more reliable and informative gene trees and a more complete representation of phylogenetic and functional relationships among the proteins than exclusive grouping schemes (wherein a protein sequence is assigned to its most similar neighbors based

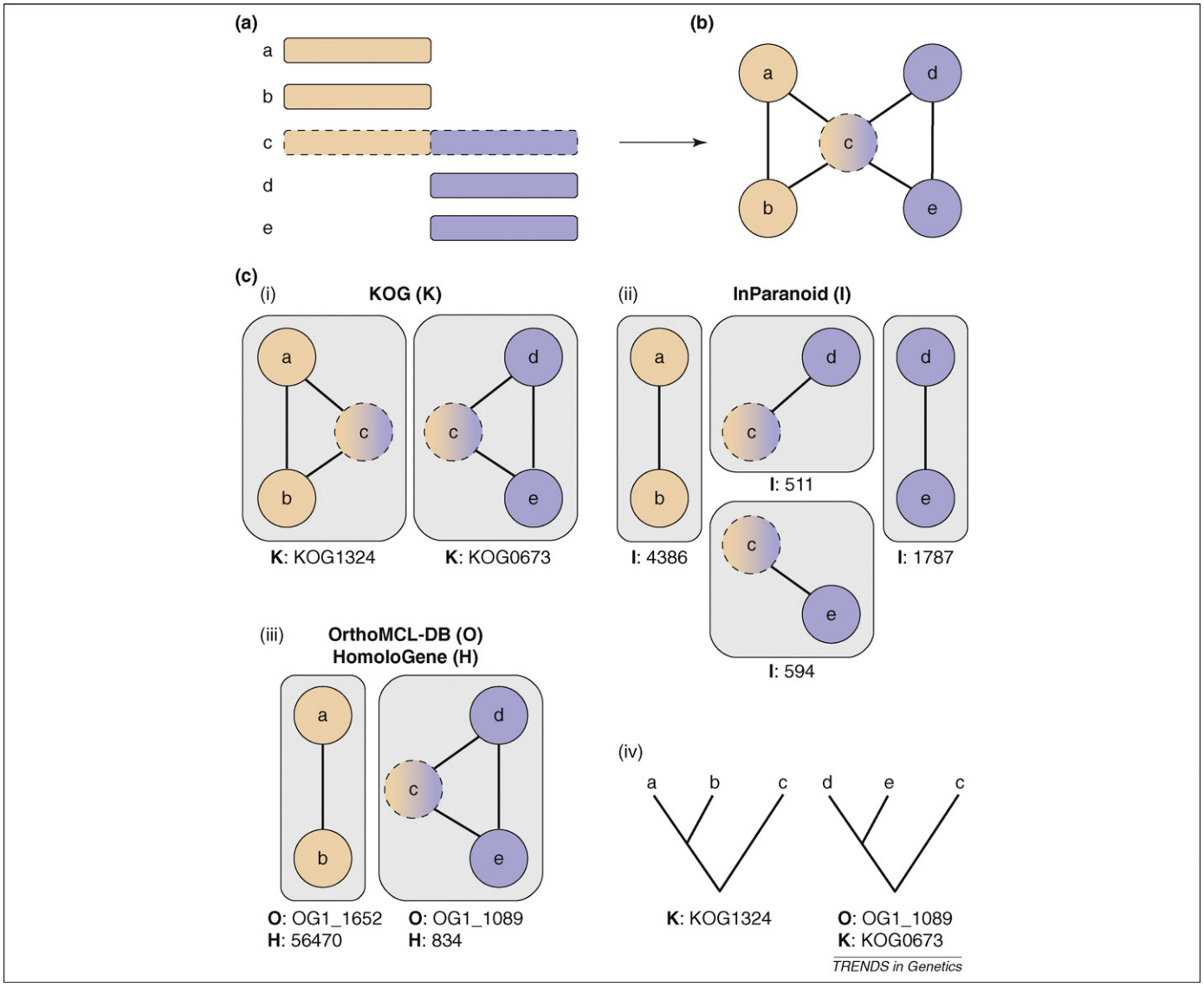
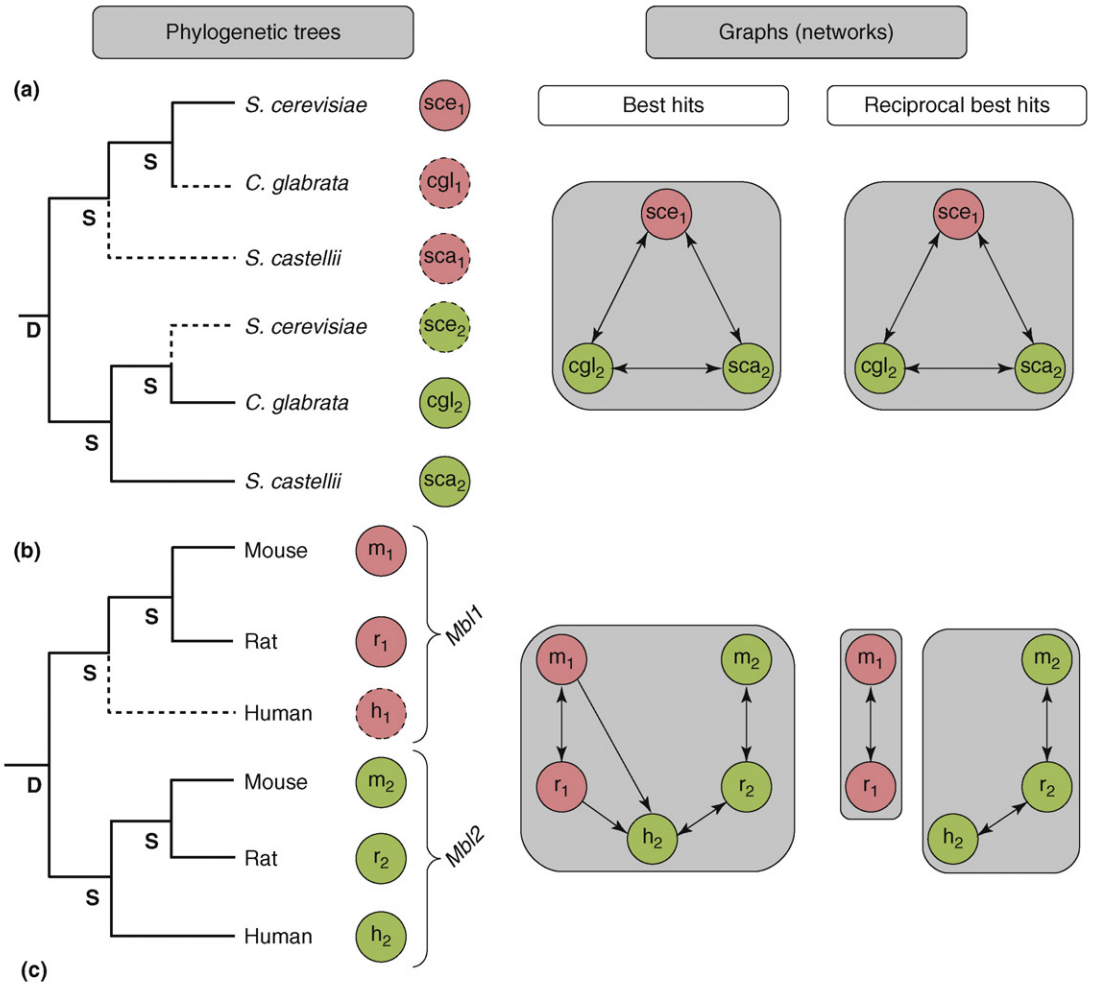


Figure 2. Partial homology to a hybrid (fusion) protein causes distinct orthologous groups to overlap. **(a)**The five proteins involved in overlapping [labeled (a-e)] are depicted as rectangles and grouped together into two overlapping groups (a,b,c and c,d,e), where protein c is the hybrid having partial homology to both groups. **(b)** The protein similarity graph of significant similarities between the proteins. Two phylogenetically unrelated protein groups are joined together. **(c)** Diagram illustrating how different databases handle the grouping of these proteins: (i) KOG (K); (ii) InParanoid (I); and (iii) HomoloGene (H), OrthoMCL-DB (O). In the current example, only the KOG database reflects the orthologous relationships between the proteins correctly, leading to a reliable inference of the protein phylogenies (iv). It should be emphasized that a phylogenetic gene tree cannot be constructed from the protein similarity graph in panel (b), because this group includes proteins that have no mutual sequence similarity at all [(a,b) versus (d,e)]. Graph nodes correspond to UniProt accessions: a, dihydrofolate reductase of *Drosophila melanogaster* (fruit fly), P17719; b, dihydrofolate reductase of Homo sapiens (human), P00374; c, bifunctional dihydrofolate reductase-thymidylate synthase 2 of *Arabidopsis thaliana* (thale cress), Q05762; d, thymidylate synthase of fruit fly, O76511; e, thymidylate synthase of human, P04818.

on partial homology), which are used by most orthology detection tools. For example, the Resample Inference of Orthologs (RIO; http://rio.janelia.org) and the Hierarchical Grouping of Orthologous and Paralogous Sequences (HOPS; http://pfam.cgb.ki.se/HOPS/) databases consider protein domains as the basic units for orthology (domain-centric view) [33,38], whereas the Phylogenetic Ortholog Groups (PHOG) database organizes proteins into overlapping OGs (protein-centric view) in which hybrid proteins are automatically flagged. [57]. Moreover, alternative splicing, errors in gene structures and low-complexity regions create problems analogous to those of hybrid proteins. Interestingly, the Phylogenetic Orthology and Paralogy (PhyOP) program is the only approach that explicitly handles genes with multiple transcripts during orthology detection [7]. Although attempts have been made to solve the problems described above, most tools currently in use were designed for single-domain proteins; therefore, all orthology data might need additional manual refinements on a case-by-case basis.

Horizontal gene transfer

HGT is an important phenomenon in the evolution of prokaryotes and eukaryotes [66–68]. Genes inherited through HGT are known as xenologs [69]. A phylogenetic inference without awareness of xenologs often leads to confounding outcomes and might indicate, for example, very close phylogenetic relationships between two distantly related organisms that have recently exchanged a gene. Moreover, HGT introduces an additional problem in



(c)

Tool	Method	Orthology prediction: Correct	Orthology prediction: Incomplete	Orthology prediction: Incorrect	Comment
Orthostrapper	Tree-based	✓			a
COCO-CL	Tree-based	✓			b
RAP	Tree-based	✓			b
Ensembl Compara	Hybrid	✓			
HomoloGene	Hybrid	✓			
Treefam	Hybrid	✓			
PhIG	Hybrid		✓		c
InParanoid	Graph-based	✓			
RoundUp	Graph-based	✓			
Reciprocal best hits	Graph-based	✓			
Ortholuge	Graph-based		✓		d
OrthoMCL-DB v2.0	Graph-based		✓	✓	e
EGO	Graph-based		✓	✓	f
Best hits	Graph-based			✓	g
OrtoMCL-DB v1.0	Graph-based			✓	h

Figure 3. Comparison of orthology detection methods in the presence of gene losses. The relationships between genes are shown from a tree (left) and a graph (right) perspective. **(a)** A reconciled gene tree (midpoint rooted) of single-copy genes (general transcriptional co-repressors) from three yeast species (*Saccharomyces cerevisiae*, *Saccharomyces castellii* and *Candida glabrata*) is inferred using known species phylogeny (for details, see Ref. [28]). Genes of *S. cerevisiae* and *S. castellii* are not orthologs but paralogs owing to the reciprocal gene loss in these species. The graph-based (nearest neighbor) approaches cannot distinguish between out-paralogs and orthologs (sce_1 is in one group with cgl_2 and sca_2). **(b)** A reconciled gene tree (midpoint rooted) of mannose-binding lectin genes (experimentally verified) from mouse, rat and human. Both rodents have two paralogous genes (*Mbl1* and *Mbl2*), whereas human has only one gene (*Mbl2*) owing to a single gene loss [82]. **(c)** The table summarizes the results of 15 different orthology prediction methods using the example of *Mbl1* and *Mbl2* genes. Orthology predictions are classified into three quality categories: (i) correct, the inference must be correct for all genes; (ii) incomplete, some orthologous relationships might be absent; and (iii) incorrect, out-paralogs and orthologs are grouped

classification (i.e. xenologs must be distinguished from other types of homologs). None of the methods that are compared in Boxes 1–3 explicitly detects xenologs, which usually requires a careful phylogenetic analysis taking phylogenetic incongruence, mobile elements, insertion and deletion patterns and atypical sequence composition into account [70,71]. Most methods that can infer HGT are only capable of detecting examples of recently acquired genes. To detect early HGT events, using the phylogenetic distribution of protein families across all domains of life might prove effective [72,73].

Gene loss and 'incomplete' genomes

Gene losses in genomes are an important source of false-positive orthology predictions. An analysis of fungal genomes has indicated that, by incorporating the information of CGNs into orthology detection, approximately half of the predicted one-to-one orthologs are, in fact, out-paralogs owing to reciprocal gene losses [29]. Therefore, out-paralogs might erroneously be inferred as orthologs when true orthologs are physically absent. Given the two real examples of gene losses in Figure 3, it is demonstrated that, unlike tree-reconciliation, a graph-based approach cannot distinguish between orthologs and out-paralogs in the presence of multiple gene loss evens (Figure 3a). In another case of a single gene loss, however, some graph-based methods (e.g. InParanoid (http://inparanoid.sbc.su.se/), RoundUp (https://rodeo.med.harvard.edu/tools/roundup/) and RBH can provide reliable orthology assignments, which are equivocal to those of all tree-based and most hybrid methods compared (Figure 3b,c). An out-group species is commonly used to identify false-positive orthologs. However, this has both advantages and disadvantages because the added sequence might provide extra resolution and specificity, but it might also decrease the sensitivity by removing authentic orthologs [45] (Figure 3). Similarly, using 'triangles' of best hits among three species is particularly disadvantageous for the Clusters of Orthologous Groups (COG; http://www.ncbi.nlm.nih.gov/COG/) of proteins in which a gene of one species is lost because such COGs will, consequently, be discounted. [19]. In principle, the tree-based methods are more robust in the presence of gene losses and varying rates of evolution than graph-based methods. This is as a result of the fact that the former group defines an orthologous relationship in the global context of all homologs and a well-established species phylogeny, whereas the latter considers pairwise nearest neighbor relations between genes from only two species. In other words, an orthology relationship must be defined in a given context, especially in terms of taxonomic sampling. However, even then, one cannot be completely certain that genes inferred as orthologs are in fact out-paralogs [38]. In two databases, namely Ensembl Compara (http://www.ensembl.org.) and TreeFam (http://www.treefam.org), gene losses are addressed explicitly using reconciled trees [53,58].

Semantics and limitations of phylogenetic concepts

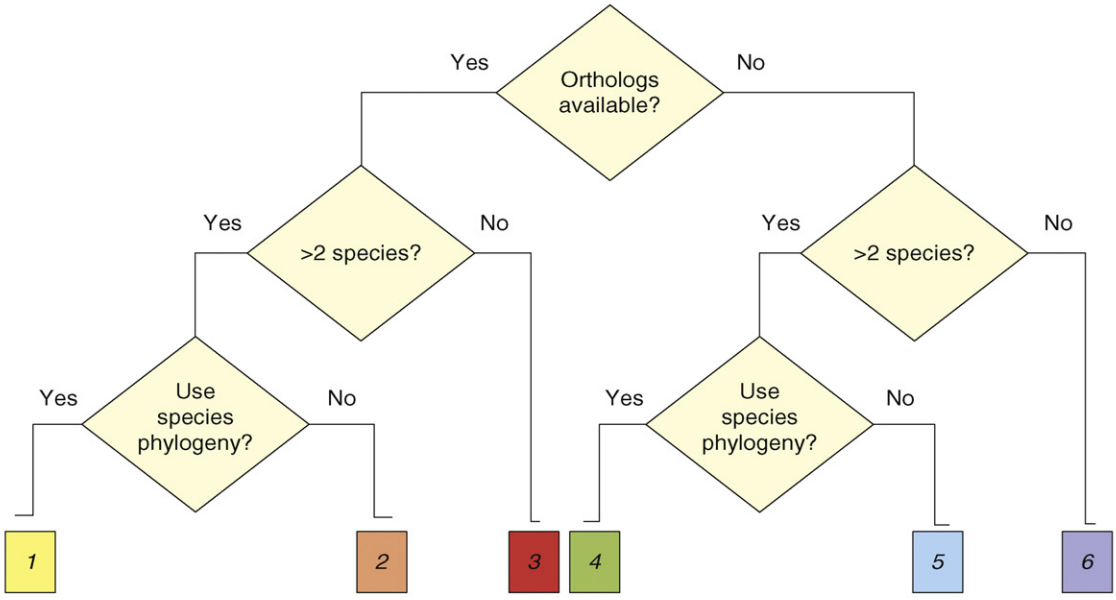
How does the language used to describe the relationships between genes complicate matters? Orthologs and paralogs are defined with respect to one event of speciation and duplication, respectively, whereas terms such as co-orthologs, in-paralogs, out-paralogs, super-orthologs and ultra-paralogs reflect a particular sequence (pattern) of speciation and/or duplication events. In principle, new terms could be associated with some other patterns in a phylogenetic tree as well, but this would be impractical for large trees. Moreover, from a visual perspective, large trees are not suitable for retrieving a subset of genes with desired properties (e.g. a taxonomic coverage or a pattern). One way to approach this problem is to convert a gene tree into one that can facilitate these 'gene-centric' queries for large-scale genome studies; for example, by means of the hierarchical numbering of OGs (similar to the way enzymes are classified [21]) used by the COCO-CL and LOFT programs [31,34]. Because the phylogenetic relationships are strictly non-transitive, an OG must always be hierarchical and defined with respect to the last common ancestor of the investigated genes (taxonomic position). In general, trees are sufficient for most evolutionary scenarios; however, the complex background of some sequences (e.g. mosaics of proteins or xenologs) requires another kind of representation, such as a graph (network), which, unlike a tree, accounts for many-to-many relations. Therefore, it seems reasonable to use both a tree and a graph (network) interchangeably in phylogenetic inferences, instead of using either one exclusively [15,16].

'Gold' standards in benchmarks

Orthology methods can be judged using several criteria including phylogenetic congruence, functional conservation and computational complexity (e.g. scalability, run times or memory usage). These benchmarks are often hampered by several factors including lack of 'gold' standards, availability of results, heterogeneous datasets, taxonomic biases, differences in the underlying methodologies and sparse documentation of the methods [74]. Amidst the flood of raw data, reliable functional annotations have only been found for a few model organisms, making the extrapolation of the results to distant species difficult owing to the high level of sequence divergence. Some orthology detection tools perform better than others in predicting a particular kind of functional conservation (e.g. co-expression, pathways or protein–protein interactions) using functional genomic data [12]. A common observation is that the tree-based orthology prediction methods generally exhibit low sensitivity and high specificity, whereas the graph-based methods show high sensitivity and low specificity [33,48,65]. Of the graph-based tools, InParanoid and OrthoMCL (http://orthomcl.cbil.upenn.edu/cgi-bin/OrthoMclWeb.cgi) perform best with respect to consistency of protein function and domain

together (e.g. *Mbl1* gene in the *Mbl2* group). Meaning of the letters (a–g) present in the 'Comment' column: a, zebrafish is used as an out-group; b, default parameters are used; c, human *Mbl2* gene (protein) is apart from mouse and rat *Mbl2* orthologs; d, mouse, rat and human *Mbl2* orthologs (transcripts) are absent; e, human *Mbl2* and mouse and rat *Mbl1* genes (proteins) are in one cluster (OG2_81338); f, human *Mbl2* and mouse *Mbl1* genes (transcripts) are in one cluster (#1119333); g, mouse and rat *Mbl1* genes link to paralogous human *Mbl2* gene; h,*Mbl1* and *Mbl2* genes (proteins) are in one cluster (OG1_4283). Graph nodes correspond to accessions: sce_1, YBR112C (UniProt: P14922); cgl_2, CAGL0D01364g (UniProt: Q6FWC0); sca_2, 705.55; m_1, UniProt: P39039, RefSeq: NM_010775; r_1, UniProt: P19999, RefSeq: NM_012599; m_2, UniProt: P41317, RefSeq: NM_010776; r_2, UniProt: P08661, RefSeq: NM_022704; h_2, UniProt: P11226, RefSeq: NM_000242.

architecture [12,65]. In contrast to functional benchmarks, phylogenetic benchmark sets of true orthologous relationships between sequences are not available yet. Although several attempts have been made to provide manually curated and consolidated sets of orthologs, mainly of vertebrate species [58,75], the following issues, in our opinion, should be addressed systematically. First, orthology is a testable hypothesis about the evolutionary



Database/Data files	1	2	3	4	5	6
COG & KOG		✓				
COGs refined by COCO-CL		✓				
COGs refined by LOFT	✓	✓				
EGO		✓				
Ensembl Compara	✓		✓			
HCOP	✓		✓			
HomoloGene	✓					
Inparanoid			✓			
KEGG orthology			✓			
Multiparanoid		✓				
OrthoMCL-DB		✓				
RoundUp			✓			
TreeFam	✓					
Programs	**1**	**2**	**3**	**4**	**5**	**6**
COCO-CL					✓	
Inparanoid						✓
LOFT				✓	✓	
Multiparanoid					✓	
OrthoParaMap						✓
Ortholuge						✓
OrthoMCL					✓	
Orthostrapper				✓		
PhyOP						✓
RAP				✓		
RSD						✓
SDI				✓		

TRENDS in Genetics

Figure 4. A decision tree for choosing the appropriate orthology detection tool. Databases and programs are listed in the table below the tree. Each tool is assigned (by a check mark) to a leaf in the tree, corresponding to a particular decision. Note: some tools are not listed here because of the limited availability or access.

descent through speciation; therefore, the orthology detection tools should be evaluated using reliable species phylogenies in the context of known evolutionary processes. For instance, simulation studies of sequence (genome) evolution involving events of gene loss might be helpful in establishing reliable orthologous relationships [49]. Furthermore, CGN might be considered for another benchmark because most orthologs tend to be found in CGN, especially if the rate of genomic rearrangements is low [34]. Second, it is not clear how to construct alignments of distant homologs consisting of multiple domains in shuffled order and how to model sequence rearrangements such as domain fusions, fissions or losses in phylogenetic inferences [20]. As a result, orthology is usually addressed using either a domain-centric or a protein-centric view. Third, orthology data cannot be exploited efficiently without thorough integration of sequence data from genomes to proteomes, distinguishing between *in silico* predicted from experimentally verified gene products and using standard and stable identifiers for database entries. Finally, standardized protocols, rules and definitions should be established and documented when manual curation is used to decide upon whether two sequences are orthologs or not.

Computation of orthologs

The large number of fully sequenced genomes raises several questions for further research, including the scalability of the orthology detection algorithms and the availability of reliable and up-to-date orthology databases (see pros and cons of the databases in Boxes 1–3). The scalability is only an issue if the number of genomes (proteomes) being compared at once is large, owing to high demands on computer resources. In fact, most graph-based methods are suitable only for pair-wise proteome comparisons (sometimes including an out-group). Clearly, these approaches do not consider all sequence data and phylogenetic information available, therefore, they are more error-prone than the tree-based methods. On the contrary, hybrid methods attempt to address the scalability and reliability by incorporating phylogenenies at various steps of the clustering process, and by using more species (genomes) to increase the reliability of orthology predictions. Therefore, fast and scalable sequence similarity search and clustering algorithms are essential for further inferences of orthologies in the hundreds of genomes available [64].

Recommendations and conclusions

The basis for most current bioinformatics tools used to detect orthology relies on three major computational principles. The proposed classification aids researchers in recognizing the essential design principles and main attributes of newly developed orthology detection tools and in designing benchmarks by means of a careful analysis of the results.

Although the different tools and approaches provide superior solutions for a variety of scenarios, the choice of methods depends on the purpose, availability and phylogenetic background (e.g. number and diversity of species or known relationships between species) of OGs (Figure 4). When biologists are interested in identifying orthologs, they might want, for example, to find functionally equivalent genes (proteins) involved in a particular biological process (e.g. cell cycle) or metabolic pathway (e.g. lipid metabolism), to study fundamental processes and mechanisms of genome evolution (e.g. speciation, duplication or HGT), fate of genes and biological functions (e.g. gain and loss), or the genetic background of complex traits and inheritable diseases. Although this list is probably far from being complete, we propose the following guidelines to choose the appropriate tool. First, one should use publicly available databases of orthologs, query them with sequences (species) of interest and, upon the availability of orthologous sequences, decide whether to use the precomputed orthologs or to make the inferences partially (i.e. using post-processing programs) or entirely *de novo* (i.e. using *ab initio* programs). Several databases are available and updated regularly, including InParanoid, OrthoMCL-DB, Ensembl Compara, HomoloGene (http://www.ncbi.nlm.nih.gov), TreeFam and HCOP [46,51,53,54,58,75]. In the next step, one should address whether the context of many species is important for the research or not, which also closely relates to the trade-off between sensitivity and specificity. If this is not a concern, then a graph-based (nearest neighbor) method is usually reliable for inferring orthologs between two closely related genomes, even in the presence of a single gene loss; otherwise, a tree-based method should be used for robust handling of multiple gene losses (Figure 3). Alternatively, multi-species OGs constructed by a graph-based approach can be used when the phylogenetic resolution is not required. Finally, if the phylogenetic relationships between species of interest are known, a choice should be made between a tree-based and a hybrid method, depending on the desired phylogenetic resolution of OGs.

Orthology detection methods seek to extend the limits of sequence comparisons by extracting information from sequence similarity networks and phylogenetic trees or by using auxiliary information of structural (conserved gene neighborhoods) and functional (ontologies) origins.

Hybrid orthology detection methods, which have addressed several shortcomings of the tree-based and graph-based methods, are likely to provide enriched context of phylogenetic and functional relationships by using both a tree and a graph representation in the computation. The application of network propagation algorithms seems especially promising for detecting relevant functional relationships among proteins by incorporating various external sources of knowledge [76–78].

At present, the number of published complete genomes approaches nearly 1000 (http://www.genomesonline.org) and hundreds more are being sequenced. The orthology detection tools reviewed here represent a valuable foundation and guide for further manual analyses. However, a scalable, fully automated procedure for inferring orthologs across genomes of all kingdoms of life still remains an elusive goal for current comparative genomics.

Acknowledgements

The authors are grateful to Jack Franklin and Simon Fisher for their help in shaping the manuscript and to the anonymous reviewers for their valuable comments.

References

1 Ouzounis, C.A. *et al.* (2003) Classification schemes for protein structure and function. *Nat. Rev. Genet.* 4, 508–519
2 Eisen, J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8, 163–167
3 Jeffroy, O. *et al.* (2006) Phylogenomics: the beginning of incongruence? *Trends Genet.* 22, 225–231
4 Delsuc, F. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375
5 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637
6 Tatusov, R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41
7 Goodstadt, L. and Ponting, C.P. (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLOS Comput. Biol.* 2, e133
8 Bandyopadhyay, S. *et al.* (2006) Systematic identification of functional orthologs based on protein network comparison. *Genome Res.* 16, 428–435
9 Mazurie, A. *et al.* (2005) An evolutionary and functional assessment of regulatory network motifs. *Genome Biol.* 6, R35
10 Grigoryev, D.N. *et al.* (2004) Orthologous gene-expression profiling in multi-species models: search for candidate genes. *Genome Biol.* 5, R34
11 Mao, F. *et al.* (2006) Mapping of orthologous genes in the context of biological pathways: An application of integer programming. *Proc. Natl. Acad. Sci. U. S. A.* 103, 129–134
12 Hulsen, T. *et al.* (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.* 7, R31
13 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113
14 Fitch, W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.* 16, 227–231
15 Rivera, M.C. and Lake, J.A. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431, 152–155
16 Doolittle, W.F. and Bapteste, E. (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 104, 2043–2049
17 Blair, J.E. and Hedges, S.B. (2005) Molecular phylogeny and divergence times of deuterostome animals. *Mol. Biol. Evol.* 22, 2275–2284
18 Ciccarelli, F.D. *et al.* (2006) Toward automatic reconstruction of a highly resolved Tree of Life. *Science* 311, 1283–1287
19 Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338
20 Sjölander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 20, 170–179
21 International Union of Biochemistry and Molecular Biology, eds (1992) *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*, Academic Press Inc
22 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
23 Ruepp, A. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* 32, 5539–5545
24 Kanehisa, M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354–D357
25 Krummenacker, M. *et al.* (2005) Querying and computing with BioCyc databases. *Bioinformatics* 21, 3454–3455
26 Markowitz, V.M. *et al.* (2007) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.* 36, D528–D533
27 Simillion, C. *et al.* (2004) Recent developments in computational approaches for uncovering genomic homology. *Bioessays* 26, 1225–1235
28 Scannell, D.R. *et al.* (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440, 341–345
29 Scannell, D.R. *et al.* (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8397–8402
30 Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5849–5856
31 Jothi, R. *et al.* (2006) COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics* 22, 779–788
32 Storm, C.E.V. and Sonnhammer, E.L.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 18, 92–99
33 Storm, C.E.V. and Sonnhammer, E.L.L. (2003) Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res.* 13, 2353–2362
34 van der Heijden, R.T. *et al.* (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 8, 83
35 Dufayard, J.F. *et al.* (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21, 2596–2603
36 Duret, L. *et al.* (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22, 2360–2365
37 Perriere, G. *et al.* (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.* 10, 379–385
38 Zmasek, C.M. and Eddy, S.R. (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3, 14
39 Zmasek, C.M. and Eddy, S.R. (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17, 821–828
40 Koski, L.B. and Golding, G.B. (2001) The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* 52, 540–542
41 Hirsh, A.E. and Fraser, H.B. (2001) Protein dispensability and rate of evolution. *Nature* 411, 1046–1049
42 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
43 Wall, D.P. *et al.* (2003) Detecting putative orthologs. *Bioinformatics* 19, 1710–1711
44 Lee, Y. *et al.* (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* 12, 493–502
45 Remm, M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052
46 O'Brien, K.P. *et al.* (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33, D476–D480
47 O'Brien, K.P. *et al.* (2004) OrthoDisease: a database of human disease orthologs. *Hum. Mutat.* 24, 112–119
48 Alexeyenko, A. *et al.* (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*,22, e9–e15. DOI: 10.1093/bioinformatics/btl213
49 Fulton, D.L. *et al.* (2006) Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics* 7, 270
50 Li, L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189
51 Chen, F. *et al.* (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34, 363–368
52 Deluca, T.F. *et al.* (2006) Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics* 22, 2044–2046
53 Hubbard, T.J.P. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.* 35, D610–D661
54 Wheeler, D.L. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 36, D13–D21
55 Cannon, S.B. and Young, N.D. (2003) OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics* 4, 35
56 Dehal, P.S. and Boore, J.L. (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* 7, 201
57 Merkeev, I.V. *et al.* (2006) PHOG: a database of supergenomes built from proteomecomplements. *BMC Evol. Biol.* 6, 52
58 Li, H. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 34, D572–D580
59 Goodman, M. *et al.* (1979) A parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* 28, 132–163
60 Page, R.D. and Charleston, M.A. (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* 7, 231–240
61 Felsenstein, J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22, 521–565

62 Larget, B. and Simon, D.L. (1999) Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Mol. Biol. Evol.* 16, 750–759
63 Hulsen, T. *et al.* (2006) Testing statistical significance scores of sequence comparison methods with structure similarity. *BMC Bioinformatics* 7, 444
64 Enright, A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584
65 Chen, F. *et al.* (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2, e383
66 Koonin, E.V. *et al.* (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* 55, 709–742
67 Lerat, E. *et al.* (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* 3, e130
68 Loftus, B. *et al.* (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature* 433, 865–868
69 Hillis, D.M. (1994) Homology in molecular biology. In *Homology, the Hierarchical Basis of Comparative Biology* (Hall, B.K., ed.), pp. 339–368, Academic Press
70 Sundin, G.W. (2007) Genomic insights into the contribution of phytopathogenic bacterial plasmids to the evolutionary history of their hosts. *Annu. Rev. Phytopathol.* 45, 129–151
71 Gupta, R.S. (2001) The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int. Microbiol.* 4, 187–202
72 Kunin, V. *et al.* (2005) The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* 15, 954–959
73 Kunin, V. and Ouzounis, C.A. (2003) GeneTRACE-reconstruction of gene content of ancestral species. *Bioinformatics* 19, 1412–1416
74 Sonego, P. *et al.* (2007) A protein classification benchmark collection for machine learning. *Nucleic Acids Res.* 35, D232–D236
75 Eyre, T.A. *et al.* (2007) HCOP: a searchable database of human orthology predictions. *Brief. Bioinform.* 8, 2–5
76 Kuang, R. *et al.* (2005) Motif-based protein ranking by network propagation. *Bioinformatics* 21, 3711–3718
77 Noble, W.S. (2005) Identifying remote protein homologs by network propagation. *FEBS J.* 272, 5119–5128
78 Carroll, S. and Pavlovic, V. (2006) Protein classification using probabilistic chain graphs and the gene ontology structure. *Bioinformatics* 22, 1871–1878
79 Passarge, E. *et al.* (1999) Incorrect use of the term synteny. *Nat. Genet.* 23, 387
80 Okuda, Y. *et al.* (2003) Occurrence, horizontal transfer and degeneration of *VDE* intein family in Saccharomycete yeasts. *Yeast* 20, 563–573
81 Penket, C.J. *et al.* (2006) YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms. *Nucleic Acids Res.*,34, W330–W334. DOI: 10.1093/nar/gkl311
82 Sastry, R. *et al.* (1995) Characterization of murine mannose-binding protein genes Mbl1 and Mbl2 reveals features common to other collectin genes. *Mamm. Genome* 6, 103–110