ELSEVIER

DRUG DISCOVERY TODAY TECHNOLOGIES

Drug Discovery Today: Technologies

Vol. 3, No. 2 2006

Editors-in-Chief
Kelvin Lam – Pfizer, Inc., USA
Henk Timmerman – Vrije Universiteit, The Netherlands

Genomic technologies

Overview and comparison of ortholog databases

Andrey Alexeyenko, Julia Lindberg, Åsa Pérez-Bercoff, Erik L.L. Sonnhammer*

Stockholm Bioinformatics Center, Albanova, Stockholm University, SE-106 91, Stockholm, Sweden

Orthologs are an indispensable bridge to transfer biological knowledge between species, from protein annotations to sophisticated disease models. However, orthology assignment is not trivial. A large number of resources now exist, each with its own idiosyncrasies. The goal of this review is to compare their contents and clarify which database is most suited for a certain task.

Section Editor:
Andreas Russ – University of Oxford, Oxford, UK

Introduction

Genome sequencing projects have produced the complete proteomes for hundreds of prokaryotic and dozens of eukaryotic species. When comparing proteomes it is important to correctly identify ORTHOLOGS and PARALOGS. Orthology is a strong indication of functional conservation and, therefore, provides the best functional annotation of experimentally undetermined proteins. This review of available ortholog databases and orthology analysis methods reveals great differences between them. Starting with databases that extract orthologs from sequence similarity, we proceed to discuss approaches that use additional information such as synteny and other patterns of concomitant evolution. Finally, we examine databases that use phylogenetic trees to infer orthologs. Although the more sophisticated methods are more reliable, they are limited because of their computational complexity and never reach the same coverage as the simpler methods.

Orthologs are genes in different species that derive from a single gene in their last common ancestor. They are created by speciation events, whereas paralogs are created by gene duplication. If the duplication was more recent than the speciation, they are called INPARALOGS, whereas if it was more ancient they are called OUTPARALOGS [1] (See Glossary).

Strictly speaking, orthology is a pairwise relation – a speciation happens between a pair of species. It is rare that several species in a group derive from more or less simultaneous speciation events, hence doing simultaneous orthology analysis across multiple species generally leads to conflicts.

To illustrate this, the tree in Fig. 1 shows Human1 and Human2 as inparalogs in relation to Worm1. This means that they are both orthologs to Worm1. Adding mouse, which is much closer related to human than to worm, leads to a situation where not all genes in the group are orthologs to each other. For instance, Human2 is not an ortholog to Mouse1, but an outparalog. It is a major challenge in orthology detection to find all inparalogs without including outparalogs, and this is the reason for the diversity among the different databases. Some have focused on small, pure groups, whereas others aim at large groups, accepting the inclusion of outparalogs. A point-by-point comparison of the most common ortholog databases is found in Table 1.

Orthologs from pairwise genome comparisons

The first large-scale effort to build a multi-species ortholog database based on pairwise similarity is clusters of orthologous groups (COGs; [2]). It uses a special clustering algorithm, in which seed clusters are formed when consistent reciprocally best hits are found between three species. Other genes and/or species might be added to the cluster afterward using

*Corresponding author: Erik L.L. Sonnhammer (Erik.Sonnhammer@sbc.su.se)
URL: http://www.sbc.su.se

1740-6749/$ DOI: 10.1016/j.ddtec.2006.06.002

Glossary

Homologs: genes with shared ancestry.
Inparalogs: genes that derive from a duplication event after a speciation of interest. Inparalogs are together orthologs to the corresponding orthologous gene/genes in the other species.
Orthologs: genes in two species that have directly evolved from a single gene in the last common ancestor and are likely to be functionally related.
Outparalogs: genes that derive from a duplication event before a speciation event of interest, thus not orthologs according to definition.
Outgroup: one or more species that are phylogenetically distant to the taxonomic group of interest (the ingroup).
Paralogs: homologous genes related by a duplication event. Might be in the same or in different genome.

fairly relaxed criteria. The initial version included unicellular (mainly prokaryotic) organisms only, and it now contains 66 species. Using the same approach, they later released the eukaryotic KOGs based on seven eukaryotic species, of which three were unicellular [3].

The COGs/KOGs clusters are built in such a way that they are often contaminated with outparalogs. For example, cluster KOG1383 (http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi?KOG1383) appears to contain two ortholog clusters stemming from two genes that existed before the divergence of eukaryotes. One cluster only contains genes from yeast and *Arabidopsis*, whereas the other cluster contains genes from three more eukaryotes. This makes it unlikely that all genes in the KOG would have the same function. Indeed, the *Arabidopsis* outparalogs have different functional

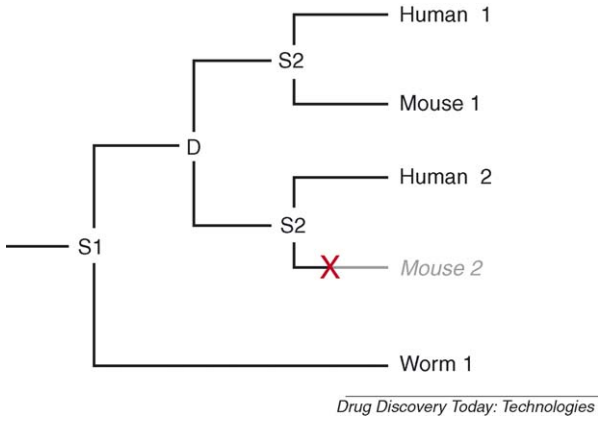

Figure 1. A schematic gene tree with two speciations (S1, S2), one duplication (D), and one gene loss (X). Human1 and Mouse1 are orthologs to each other, as the last event in their common evolution was a speciation. Human1, Mouse1, and Human2 are all orthologs relative to Worm1, and can thus be called inparalogs. However, Mouse1 and Human2 are not orthologs, but outparalogs, as the last event in their common evolution was duplication D. This illustrates why including multiple species with differing evolutionary distances typically leads to outparalogs in the same cluster.

definitions in GenBank ('calmodulin binding', 'glutamate decarboxylase 2', 'carboxy-lyase').

The EGO database (former TOGA [4]) was also built with the COG technology. It is based on TIGR gene sequences (partly assembled from EST data) and thus uses similarity of DNA rather than of amino acid sequences. Incomplete genomes are allowed, which means that wrong assignments might be done if the true best hit is missing. The applicability of EGO is also limited because of the long absence of updates.

The InParanoid [5] algorithm was specifically designed to find all inparalogs (hence the name) in ortholog groups between two species. A 'seed ortholog' is here the reciprocally best matching protein, whereas inparalogs are paralogs closer to the seed ortholog than the seed orthologs are to each other. A unique feature is that confidence values are provided for both the seed orthologs and the inparalogs; hence a user can select the strongest orthologs only. The latest version of the InParanoid database contains 21 eukaryotic organisms plus *Escherichia coli*. In addition to the main resource, the Ortho-Disease Web site [6] provides orthologs to the human disease genes listed in OMIM. The InParanoid program and technology is widely used by model organism databases (e.g. FlyBase, WormBase, SGD, and so forth) to cross-reference each other. For this and many other purposes, orthology analysis of two proteomes at the time is suitable. It simplifies the procedure and also avoids problems of conflicting evolutionary history that can occur when involving other species. However, for some purposes it is desirable to build ortholog groups of several species. To this end, the MultiParanoid algorithm [7] can assemble InParanoid clusters into multi-species groups.

OrthoMCL was built in a fashion similar to InParanoid in terms of gathering inparalogs. A major difference however is that OrthoMCL provides the possibility of building ortholog groups of multiple species. Clustering of the orthologs is done using the Markov Clustering algorithm (MCL), which is based on probability and graph flow theory [8]. It is a rather robust method for avoiding outparalogs that occur when distantly related species are mixed with sets of closely related species; nevertheless it does not eliminate this kind of erroneous predictions completely [7]. Another aspect is that alternative splice variants are included, which can lead to high redundancy or that different splice forms of the same gene end up in different ortholog groups.

The ortholog assignments in the KEGG database have a focus on similarity in molecular function. They are based on protein sequence comparison, information from the COGs database, and expert classifications of protein families [9,10]. There are possibilities to get HOMOLOG data of different qualities. In the KEGG GENES catalog, orthologs and paralogs have been identified through manual curation of sequence similarity data. Orthologous groups for incomplete genomes (DGENES) and ESTs (EGENES) are also available but have been

Table 1. Comparison of ortholog databases

Ortholog resource	EGO	COG/KOG	HomoloGene	InParanoid/MultiParanoid	HOPS
Number of species	82	66/7	18	22/4	n/a
Pros	Includes also partially sequenced species.	Has become a standard for 'uniform-function' protein groups. Easy addition of new genomes without recalculating the whole set. Manual curation. Provides species-specific expansions.	Classifies ortholog groups by the last common ancestor. Synteny is used to assist similarity. Provides specifies-specific expansions and supplementary sequence features.	Includes genomes for all major eukaryotic clades. Precise and exhaustive ortholog delineation for pairwise proteome analysis.	Domain oriented, integrated in the Pfam server. Graphical user interface. Includes also partially sequenced species.
Cons	DNA based and incompatible with other systems. Inclusion of incomplete genomes can lead to false assignments. Not updated for a long time.	Contains many outparalogs.	Systematically wrong in spreading inparalogs over different clusters.	No tree view provided. Only one Prokaryote (*E. coli*).	Only pairwise orthology between 2x3 eukaryotic clades. No prokaryotes. Not downloadable, only runs in web browser. Not queryable.
References	http://www.tigr.org/tdb/tgi/ego/	http://www.ncbi.nlm.nih.gov/COG/	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene	http://www.inparanoid.cgb.ki.se/, http://www.multiparanoid.cgb.ki.se/	http://www.pfam.cgb.ki.se/HOPS/

Ortholog resource	KEGG	OrthoMCL	PhIGs	Ensembl Compara	MGD
Number of species	355	55	34	25	21 (focus on 3)
Pros	Manually curated, taking into account function information. Ortholog information for ESTs and incomplete genomes. Pathway linked clusters.	Multiple species comparisons.	Tree view provided.	Manually curated. Synteny is used to assist similarity.	Ortholog classification supported by scientific publications. Expert knowledge considered.
Cons	Generates unexpectedly large clusters.	Some clusters contain out-paralogs. Includes multiple splice variants of genes.	Poor website. Clusters not downloadable.	Does not support multiple inparalogs.	Limited in species. Mainly mouse, rat and human. Varying quality.
References	http://www.genome.jp/kegg/	http://www.orthomcl.cbil.upenn.edu/cgi-bin/OrthoMclWeb.cgi	http://www.phigs.jgi-psf.org/	http://www.ensembl.org	http://www.informatics.jax.org/searches/homology_form.shtml

Table 1 (*Continued*)

Ortholog resource	HOGENOM	HOVERGEN	INVHOGEN	TreeFam	OrthologID
Number of species	263	n/a	n/a	20 complete	3
Pros	Includes many complete genomes. Tree based. Using the graphical tool FamFetch, tree patterns can be searched.	Contains all vertebrate sequences from UniProt. Tree based. FamFetch	Contains all invertebrate sequences from UniProt. Tree based. FamFetch	Manually curated based on trees. Ortholog pairs can be downloaded. Uses both known and novel data from the Ensembl databases.	Improved with 'character' analysis.
Cons	No manual curation. Does not use Ensembl gene data.	No manual curation. Does not use Ensembl gene data.	No manual curation. Does not use Ensembl gene data.	Only contains animal taxa, with the exception of some plant and fungal species used as outgroups.	Only plant proteomes. The trees convey evolutionary relations, but not orthology. Not down-loadable.
References	http://www.pbil.univ-lyon1.fr/databases/hogenom.html	http://www.pbil.univ-lyon1.fr/databases/hovergen.html	http://www.bi.uni-duesseldorf.de/~invhogen/invhogen.html	http://www.treefam.org	http://www.nypg.bio.nyu.edu/orthologid/

n/a, not applicable.

automatically generated through BLAST [11] comparisons to KEGG GENES. Another way of retrieving orthologous and paralogous relationships from the KEGG database is by searching the Sequence Similarity Database (SSDB), where the user can determine search criteria such as type of organism, one way or reciprocal best hits and what similarity score threshold to use. However, deciding on ortholog assignment is left to the user's discretion. Thus, the expert assignments found in the KEGG orthology tables (KO) seem to be the main virtue of the resource. However, even these classifications should not be universally trusted as we found the method to group proteins that could never be considered orthologs by sequence similarity.

PhiGs uses a graph-based method guided by known phylogenetic relationships to cluster orthologs [12]. Comparisons between multiple species are performed in this method. Ortholog clusters are created at each of the nodes on the chordate species tree to overcome the problem with erroneous ortholog predictions because of the inclusion of incomplete genomes (Paramvir Dehal, per. commun.). This means that for each gene, several clusters are available. Which cluster to choose depends on your purpose with the analysis and is preferably done by looking at the trees. Unfortunately it is not yet possible to download neither the software nor the ortholog groups, which makes it difficult to evaluate this resource properly.

The ortholog clusters available in MGD at the research community Mouse Genome Informatics (MGI) were identified using a combination of computational analysis and manual curation [13]. Most of the orthology assignments were extracted from scientific publications and all are supported by at least one reference. Depending on the reference of the ortholog assignment, the method for identification and the quality of the cluster varies. The database is focused on the orthologous relationships between mouse, human and rat but also provides some information about orthology in 15 other mammalian species.

Orthologs from synteny

A different approach for predicting orthologous relationships is to look for conserved physical location of genes on the chromosomes, that is, synteny. The Ensembl Compara database and the HomoloGene database are resources where this approach has been adopted.

The Ensembl Compara database is primarily based on best reciprocal hits for pairs of species, but a region of 1 Mb around each such hit is analyzed for synteny. Other ortholog pairs might be assigned in this region if the gene order is conserved and the identity is above 40% (Xose Fernandez, pers. commun.). A drawback with this algorithm is that it only keeps the best hit and does not consider multiple inparalogs.

The Homologene project exists since the pre-genomic era when it was based on incomplete proteomes. The

building procedure now accounts for chromosomal synteny, DNA sequence features and other information (http://www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html). The clusters are labeled by the last common ancestor of the species they contain. Additionally, species-specific expansions (paralog groups that exist only in one organism) are provided. The latest version [14] is announced to include inparalogs; however, these are generally assigned to different clusters. Typically, the ortholog groups are split into narrow slices with one inparalog in each. This gives a fragmented picture of the orthologous genes, and the lack of cross-links makes it impossible to find complete sets of inparalogs.

Orthologs from trees

HOVERGEN and HOGENOM contain protein families of vertebrate and complete genomes [15,16] with tree-based orthology assignments. The families were created by BLAST-based clustering [17] of sequences from UniProt, which were aligned using ClustalW. Trees were built from the alignments using RAP that produces a reconciled tree from gene and species trees. INVHOGEN [18] was created through the same procedure applied to invertebrate sequences, except that the trees were constructed with the IQPNNI method [19]. An interesting feature for all three databases is that one can use the graphical user interface FamFetch to search the databases for tree patterns specified by the user. For example, one can query the databases for trees with no duplication events and hence exclude paralogs whereas only detecting orthologs. However, there is no function for bulk download of orthologs. Instead, the idea is to visually inspect reconciled trees and make manual judgments of orthology and paralogy.

TreeFam is a manually curated database of trees with genes from animal taxa [20]. Baker's yeast (*Saccharomyces cerevisiae)*, fission yeast (*Schizosaccharomyces pombe*) and thale cress *(Arabidopsis thaliana)* were also included to serve as OUTGROUPS. The families were based on seed clusters from the PhIGs database that were expanded by both BLAST and HMM [21] searching. A neighbor-joining tree was built for all families to generate the TreeFam-B database, and a set of in-house tools was used to infer speciation and duplication nodes in the trees to produce orthology assignments. Some of the trees were manually curated to construct the TreeFam-A database, which is more reliable. In release 1.1, TreeFam-A contained 690 families and TreeFam-B 11646.

A plant-specific ortholog database called OrthologID was recently built [22] from the three finished plant genomes (*A. thaliana*, *P. trichocarpa*, *O. sativa*), using the alga *Chlamydomonas* as outgroup. The gene family clusters were built from BLAST hits and subjected to parsimony tree analysis. Cluster-specific diagnostic characters, that is, patterns of several nucleotides/amino acids, were also considered as evidence of relatedness. The relation to the *Chlamydomonas* outgroup seems to be the main criterion to define the clustering level of the groups. Probably as a consequence of this, the gene trees often contain multiple subtrees with 2–3 species, each looking like a distinct ortholog group. Such clusters would thus contain outparalogs. For instance, the genes of the cluster 10084 http://www.nypg.bio.nyu.edu/orthologid/cgi-bin/query.cgi?genename=At4g15440+&qtype=ortholog emerged because of a series of duplication events in a common ancestor of the angiosperms, that is, after the split from *Chlamydomonas* but before the divergence of the three plant species. Descendants of those 3–4 copies are retained in all of the contemporary organisms. OrthologID might, thus, serve a resource of plant gene evolution but does not provide strict ortholog groups. The data is not downloadable; the intended usage is by submitting a query protein that is rapidly placed in a cluster tree.

The HOPS database [23] uses gene trees to extract orthologs from two species with the Orthostrapper [24] algorithm. This method looks for ortholog groups between two species that cluster below an outgroup. By application to bootstrap trees, confidence values are calculated for the orthology assignments. It was applied to the entire Pfam database, in which all eukaryotic sequences were divided into six clades on two levels. At the *Eukaryota* level, these were *Metazoa*, *Viridiplantae*, and *Fungi*. At the metazoan level they were *Chordata*, *Nematoda* and *Arthropoda*. A unique feature of HOPS is that because it is based on Pfam, it is possible to analyze a protein's orthologous relationships for each domain separately. For this, a graphical user interface NIFAS [26] was built into the Stockholm Pfam web server.

Conclusions

More than a dozen resources providing orthology analysis are currently available. We here provide glimpses of their different values and qualities. Which resource is the most suited depends on the purpose that the user has in mind. The most common purpose is probably detailed functional annotation transfer between genes. Here, it is crucial to obtain the set of genes that are most likely to have the same function. This means that the gene set should contain all orthologs but exclude outparalogs. Inparalogs that emerged after the speciation event are all true orthologs and should all be included for completeness, as it is normally unknown whether functional divergence has occurred among them. However, outparalogs can be detrimental to annotation transfer – if closer relatives to the gene exist, then why consider proteins that might have acquired a special function earlier?

To exemplify this, Fig. 2 shows a case with neuroendocrine disease genes for which different databases provide different sets of orthologs. KOGs generally has by far the largest clusters, often comprising several distinct ortholog groups, whereas Ensembl Compara and Homologene have the smallest, normally with only one ortholog/species. The opti-

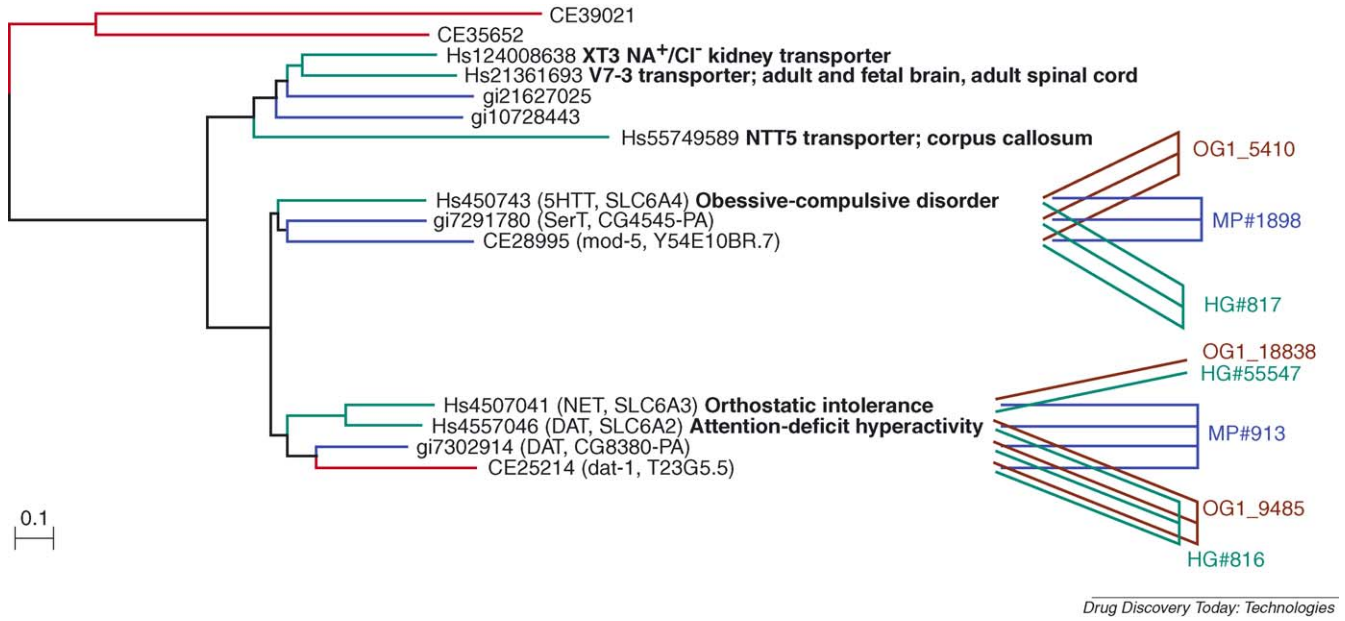

Figure 2. Orthology analysis of neuroendocrine disease genes.

This tree corresponds to KOGs cluster #3659, with human (green branches), fly (blue branches) and worm (red branches) proteins. Ortholog clusters from other databases have been marked to the right: OG*: OrthoMCL; MP*: MultiParanoid; HG*: Homologene.

It is apparent that the KOGs cluster contains three distinct ortholog clusters. For the three human genes with an OMIM-associated disease (Hs4507041, Hs4557046, and Hs4507043) only the nearest fly and worm genes are true orthologs. The upper 7 genes of the KOG tree derive from earlier duplication events and are thus outparalogs to the human disease genes. The annotation to the KOG cluster ('Sodium-neurotransmitter symporter') is very general and does not reflect the functions that can be discerned from an invertebrate point of view.

The correct fly and worm orthologs for Hs4507043 were found in MultiParanoid (MP#1898), OrthoMCL (OG1_5410) and HomoloGene (HG#817). The fly and worm genes SerT and mod-5 are the optimal counterparts of the human obsessive-compulsive disorder gene. This provides the more detailed function 'Na[+],Cl[−]-dependent serotonin transporter'.

For Hs4507041 and Hs4557046, only MultiParanoid provided the correct orthology group. OrthoMCL and HomoloGene only clustered Hs4557046 with the fly and worm orthologs, whereas the other human inparalog was found in another vertebrate-specific cluster (not shown). Researchers leaning on these two resources would be unaware of the close relation between the two disease genes, as well as the existence of fly and worm orthologs to the latter.

The MultiParanoid orthologs and inparalogs can also be found in the InParanoid database, as every MultiParanoid assignment was based on pair-wise InParanoid clusters ([human-worm], [human-fly] and [fly-worm]). We note that the genes CE39021 and CE35652 do not have human orthologs in MultiParanoid. InParanoid only clustered them with *C. briggsae* genes – they seem to represent an ancient lineage currently present only in nematodes.

The tree was built by neighbor-joining in Belvu using *Scoredist* distances [25]. Fragmentary genes in the KOG were excluded, and updated gene identifiers are shown (The KOG database uses old identifiers for some of the genes in this cluster). Human gene annotations were taken from OMIM.

mal level of clustering usually lies in between these extremes, for example, as in InParanoid or OrthoMCL.

Another purpose for using orthologs is in the study of gene family evolution. Here the tree-based databases such as HOGENOM, TreeFam, and HOPS offer a significant advantage. Unfortunately, the orthology information is rarely retrievable for whole proteomes in these resources, so they are not applicable on a large scale. But the user interfaces are powerful and a skilled user can learn much about single families by manually exploring these resources.

The final purpose is an expert user that wants to create his/her own ortholog database using proprietary data. Here the choice of resources is poorer as many algorithms and/or programs are not available. The InParanoid software is popular because it is robust and has always been available as open source code. The OrthoMCL programs are also available since the last publication but might need fine-tuning of the parameters for proper sensitivity and specificity. A TreeFam tree builder and tree viewer programs are also available for download. Many of the present technologies use, to some extent, manual curation of automatically prepared ortholog groups, which cannot be exactly reproduced outside of the respective research group. This represents one of the greatest challenges of the field: to educate scientists about what orthologs and paralogs really are. With this knowledge, fewer orthology mis-assignments and fewer flawed experiments would take place.

Related articles

Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338

Gogarten, J.P. and Olendzenski, L. (1999) Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* 9, 630–636

Outstanding issues

- Multi-species orthology relationships are based on currently complete genomes and might be altered when more genomes are sequenced.
- Many proteins have multiple domains – but the domain structure is not considered in most ortholog databases.
- How to simultaneously maximize coverage and minimize the amount of outparalogs?
- In cases with multiple inparalogs, what is the functional redundancy/diversity among them? Have some inparalogs diverged in function?

References

1 Sonnhammer, E.L.L. and Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18, 619–620

2 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637

3 Tatusov, R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinform.* 4, 1–14

4 Lee, Y. *et al.* (2002) Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* 12, 493–502

5 Remm, M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052

6 O'Brien, K.P. *et al.* (2004) OrthoDisease: a database of human disease orthologs. *Hum. Mutat.* 24, 112–119

7 Alexeyenko A. *et al.* Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* (in press)

8 Li, L. *et al.* (2003) OrthoMCL: identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13, 2178–2189

9 Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30

10 Kanehisa, M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354–D357

11 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

12 Dehal, P. and Boore, J.L. (2005) Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLOS Biol.* 3, 1700–1708

13 Eppig, J.T. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice – a community resource for mouse biology. *Nucleic Acids Res.* 33, D471–D475

14 Wheeler, D.L. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 34, D173–D180 (Database issue)

15 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

16 Duret, L. *et al.* (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22, 2360–2365

17 Perriere, G. *et al.* (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.* 10, 379–385

18 Paulsen, I. and von Haeseler, A. (2006) INVHOGEN: a database of homologous invertebrate genes. *Nucleic Acids Res.* 34, 349–353 (Database issue)

19 Vinh, le S. and von Haeseler, A. (2004) IQPNNI: moving fast through tree space and stopping in time. *Mol. Biol. Evol.* 21, 1565–1571

20 Li, H. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 34, 572–580 (Database issue)

21 Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* 14, 755–763

22 Chiu, J.C. (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22, 699–707

23 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

24 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

25 Sonnhammer, E.L.L. and Hollich, V. (2005) Scoredist: a simple and robust protein sequence distance estimator. *BMC Bioinform.* 6, 108

26 Storm, C.E.V. and Sonnhammer, E.L.L. (2001) NIFAS: Visual analysis of domain structure evolution. *Bioinformatics* 17, 343–348