Quest for orthologs

Homology and common ancestry

Homology was first defined in biology by:

Sir Richard Owen in 1843

- homology : the same organ under every variety of form and function.

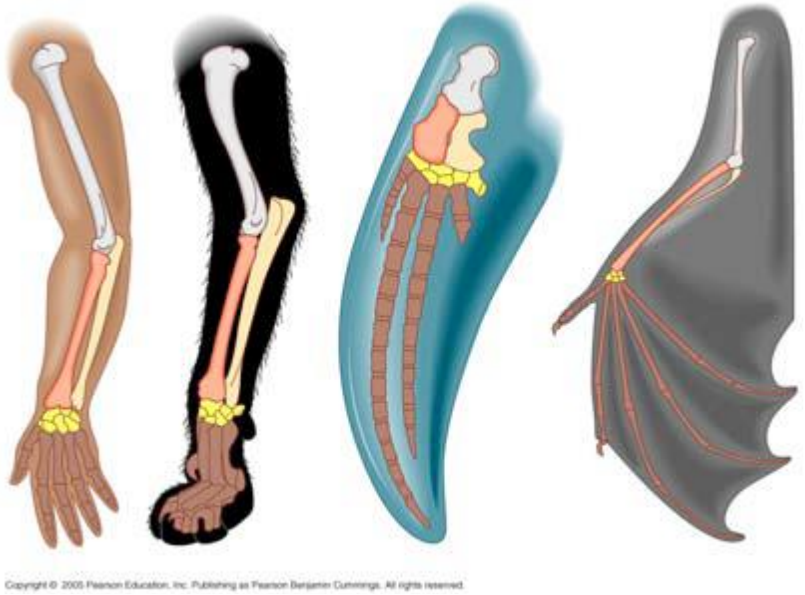
Charles* Darwin *: *Origin of Species* published on 24 November 1859

- homology: traits that are the same due to **common ancestry**,
- analogy: traits that are similar due to **evolutionary convergence**.

Homology: one of Owen/Darwin most impressive contributions to evolutionary thinking.

Homology

- Darwin used the example of **homologous structures**, or variations on a structure present in a **common ancestor**.



- For example, a human arm, a cat's leg, a whale's flipper, and a bat's wing all are adapted to different purposes, but share the **same bone structure**.
- This suggests one **common ancestor with that common structure**.

- Walter M. Fitch (1970): introduced the concepts and definitions of orthology and paralogy

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

Abstract

Fitch, W. M. (Dept. Physiological Chem., U. Wisconsin, Madison 53706) 1970. Distinguishing homologous from analogous proteins. Syst. Zool., 19:99–113.—This work provides a means by which it is possible to determine whether two groups of related proteins have a common ancestor or are of independent origin. A set of 16 random amino acid sequences were shown to be unrelated by this method. A set of 16 real but presumably unrelated proteins gave a similar result. A set of 24 model proteins which was composed of two independently evolving groups, converging toward the same chemical goal, was correctly shown to be convergently related, with the probability that the result was due to chance being $<10^{-21}$. A set of 24 cytochromes composed of 5 fungi and 19 metazoans was shown to be divergently related, with the probability that the result was due to chance being $< 10^{-9}$. A process was described which leads to the absolute minimum of nucleotide replacements required to account for the divergent descent of a set of genes given a particular topology for the tree depicting their ancestral relations. It was also shown that the convergent processes could realistically lead to amino acid sequences which would produce positive tests for relatedness, not only by a chemical criterion, but by a genetic (nucleotide sequence) criterion as well. Finally, a realistic case is indicated where truly homologous traits, behaving in a perfectly expectable way, may nevertheless lead to a ludicrous phylogeny.

Two subclasses of homology

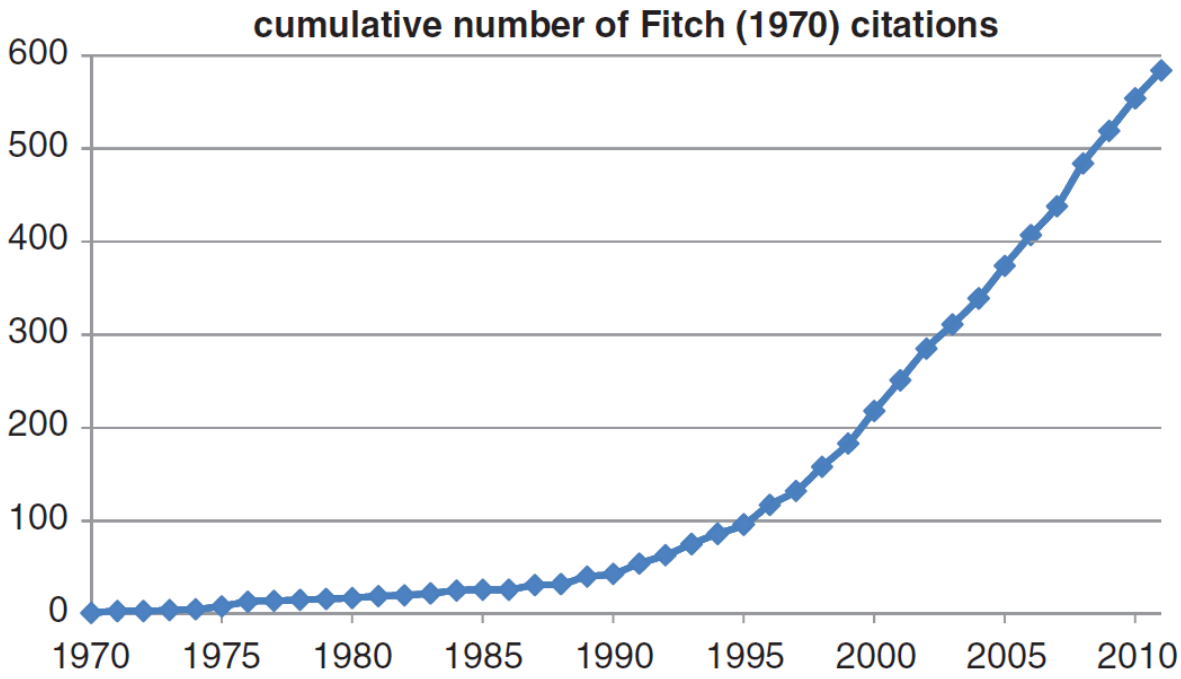
- "It has been pointed out before that **a phylogeny of birds and mammals** based upon a mixture of α and β hemoglobins would be biological nonsense since the initial dichotomy would be on the distinction between the α and β genes rather than between the birds and the mammals (Fitch and Margoliash, 1967). "

- "Therefore, there should be **two subclasses of homology**:
  - If the homology is the result of **gene duplication** so that both copies have descended side by side during the history of an organism, (for example, α and β hemoglobin) the genes should be called ***paralogous*** (para =in parallel).

  - If the homology is the result of **speciation** so that the history of the gene reflects the history of the species (for example a hemoglobin in man and mouse) the genes should be called ***orthologous*** (ortho = exact)."

- Fitch's paper: conceptual cornerstone of modern genomics

Fitch publication was poorly cited over the next 25 years



- **Cumulative dynamics of the citation of Fitch's 1970 article**.
- The citation data were from the ISI Web of Science (Koonin 2011).

  - by the end of 2010, it has been cited only 554 times!

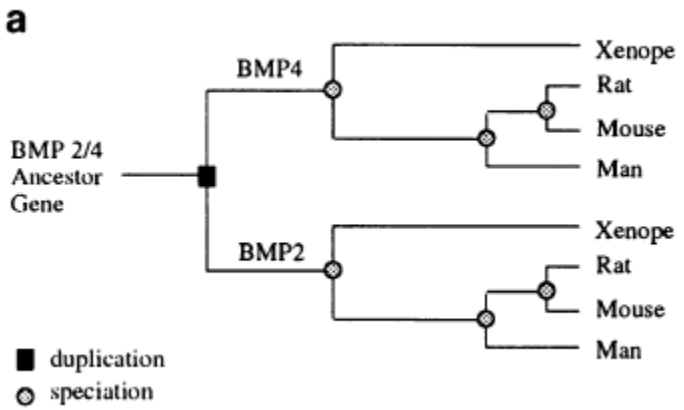
2360–2365 Nucleic Acids Research, 1994, Vol. 22, No. 12

HOVERGEN: a database of homologous vertebrate genes

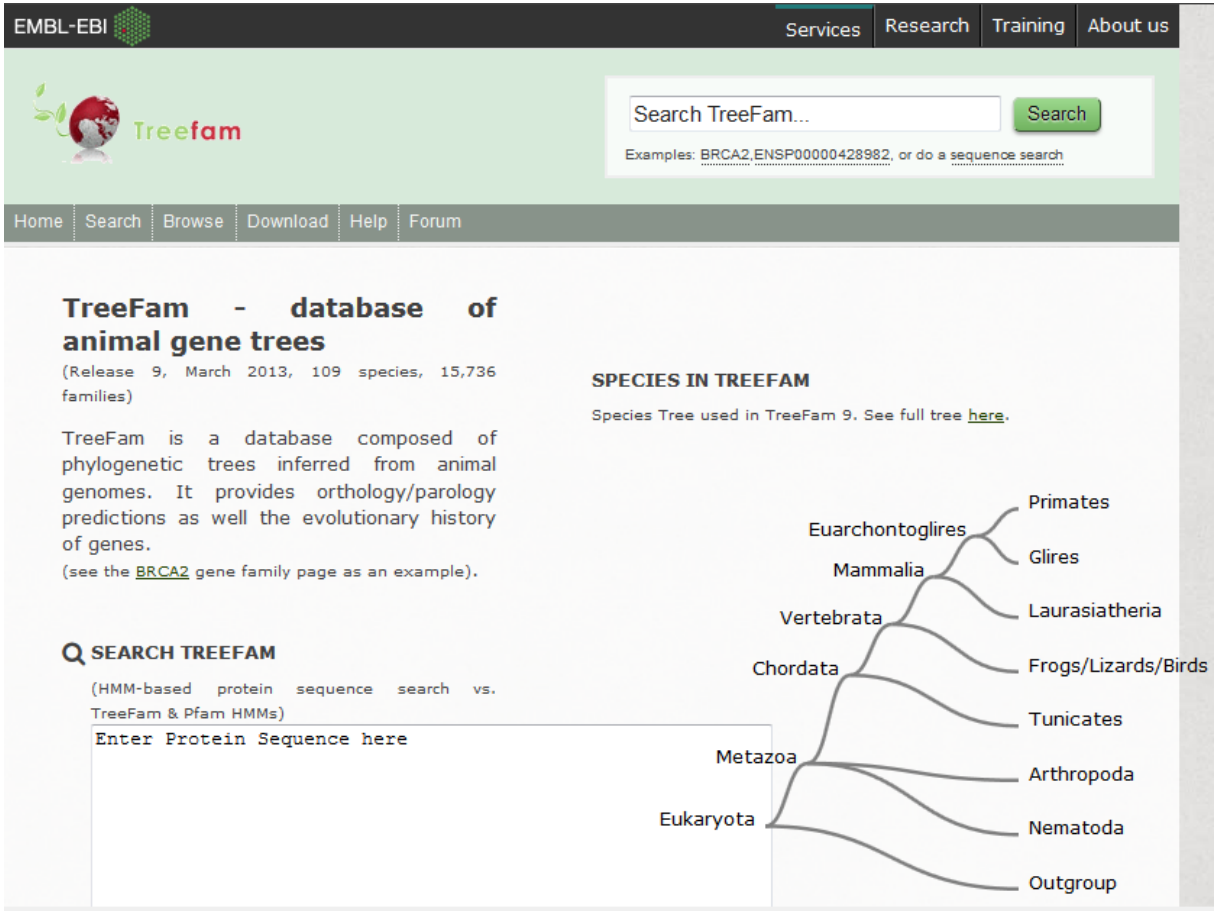
Laurent Duret*, Dominique Mouchiroud and Manolo Gouy
Laboratoire de Biométrie, Génétique et Biologie des Populations, Université Claude Bernard, Lyon I, URA-CNRS 243 Bat. 741, 43 Blvd du 11 Novembre 1918, 69622 Villeurbanne cedex, France

Received February 17, 1994; Revised and Accepted May 10, 1994

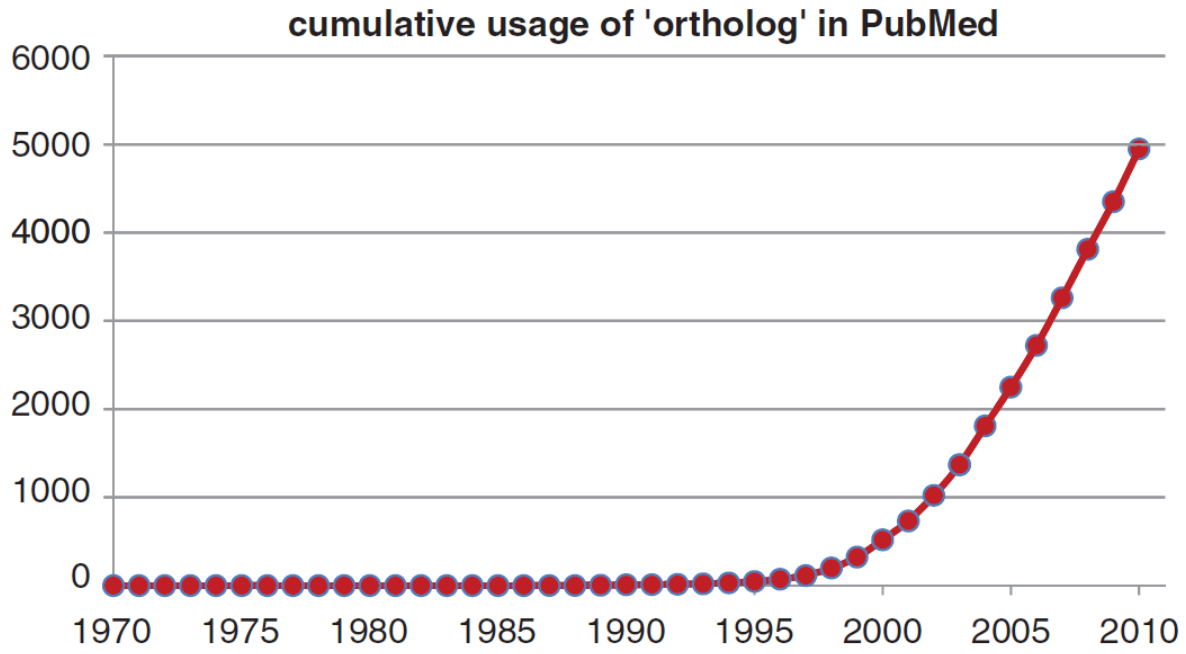
- Original HOGENOM has inferred orthologs from 5 phylogenetic trees.
  - focuses on gene families from **completely sequenced genomes**.
  - based completely upon **automatically** generated trees.



Molecular phylogeny takeoff



Cumulative usage of 'ortholog' in PubMed

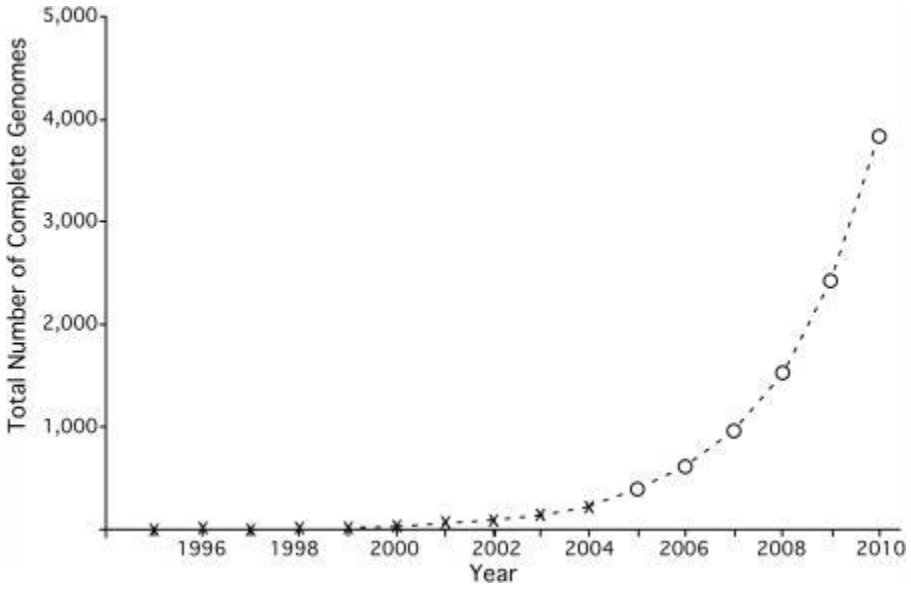


- **The usage of the term 'ortholog' in the title or abstract of scientific publications**. The usage data were from PubMed (Koonin 2011).

  - almost 90% (554/4947) of articles in PubMed that use the term do not cite Fitch.

Genome projects

- 1990: Human Genome Project is launched. The project aims to sequence all 3 billion letters of a human genome in 15 years.
- 1995: The first bacterium genome sequence is completed (*Haemophilus influenza*).



- 1996: An international team complete sequencing the genome of yeast, *Saccharomyces cerevisiae*.
- 1998: The genome of the nematode worm, *C. elegans*.
- 2000: The full genome sequence of the model organism *Drosophila melanogaster* (fruit fly) is completed.

Distinguishing orthologs from paralogs

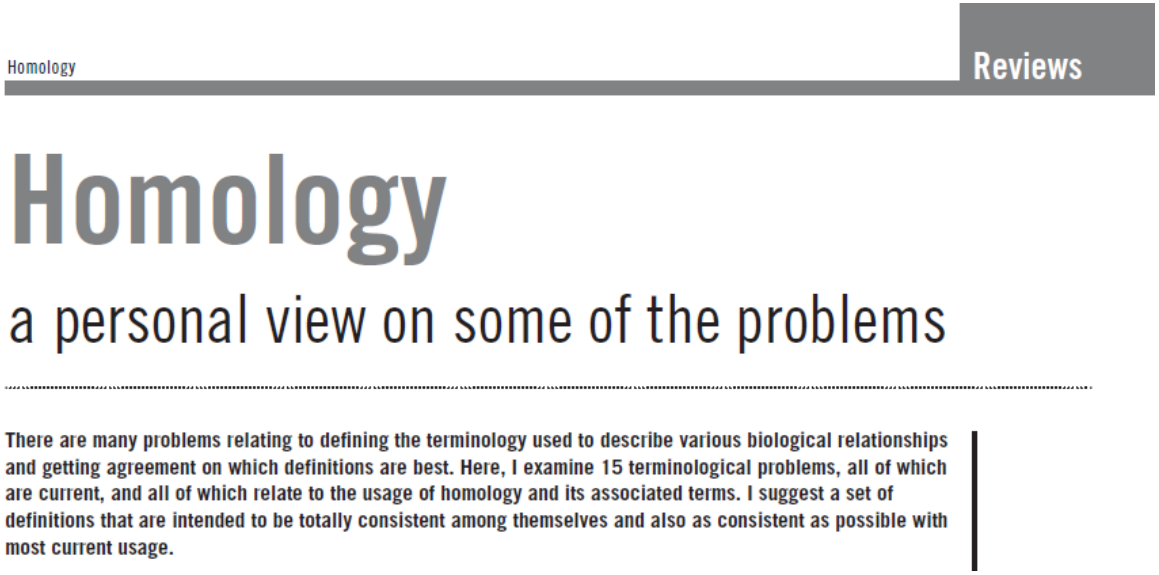
- EugeneV. Koonin

Distinguishing orthologs from paralogs is critical for at least three key tasks:

- reconstruction of **genome evolution** including genes losses, horizontal gene transfer and lineage-specific duplication;
- study of major aspects of the **evolutionary process** such as the distribution of selection pressure across genes;
- **transfer of functional information** from functionally characterized genes to uncharacterized homologs from other organisms which is the basis of genome annotation.

- at the center of almost every **comparative genomic** study.

2010: 30th anniversary



Homology

Reviews

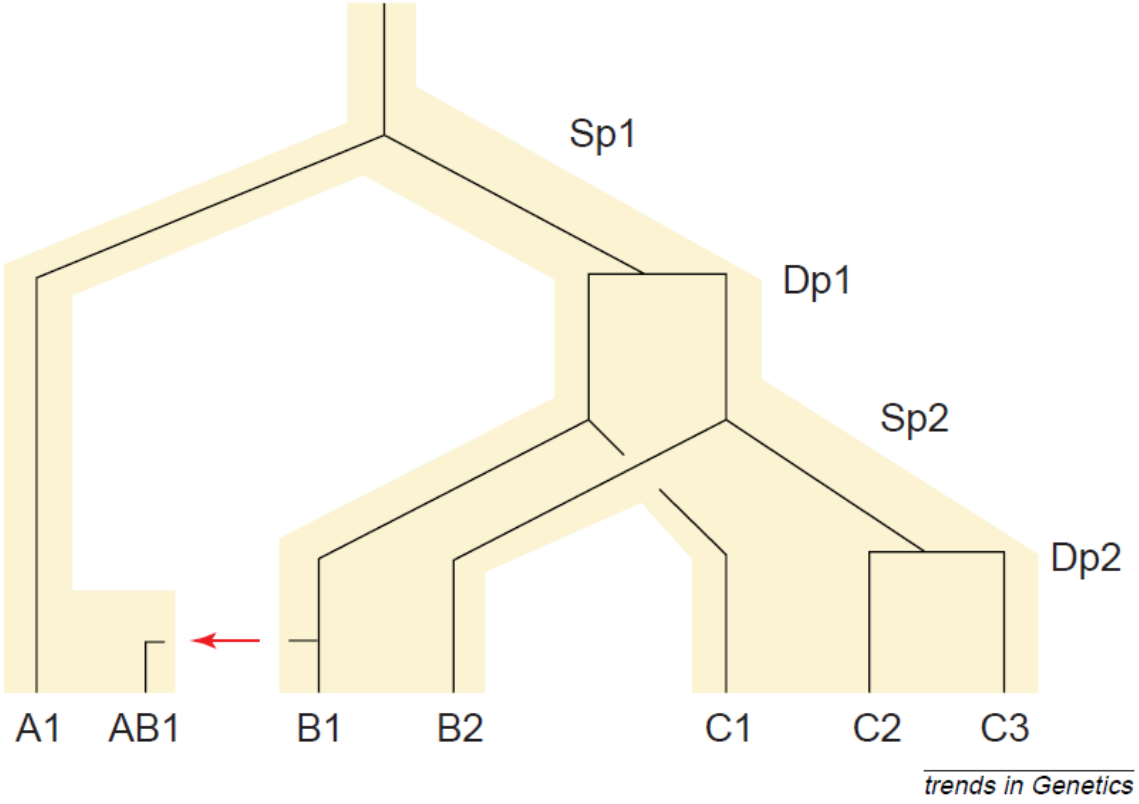
Homology

a personal view on some of the problems

There are many problems relating to defining the terminology used to describe various biological relationships and getting agreement on which definitions are best. Here, I examine 15 terminological problems, all of which are current, and all of which relate to the usage of homology and its associated terms. I suggest a set of definitions that are intended to be totally consistent among themselves and also as consistent as possible with most current usage.

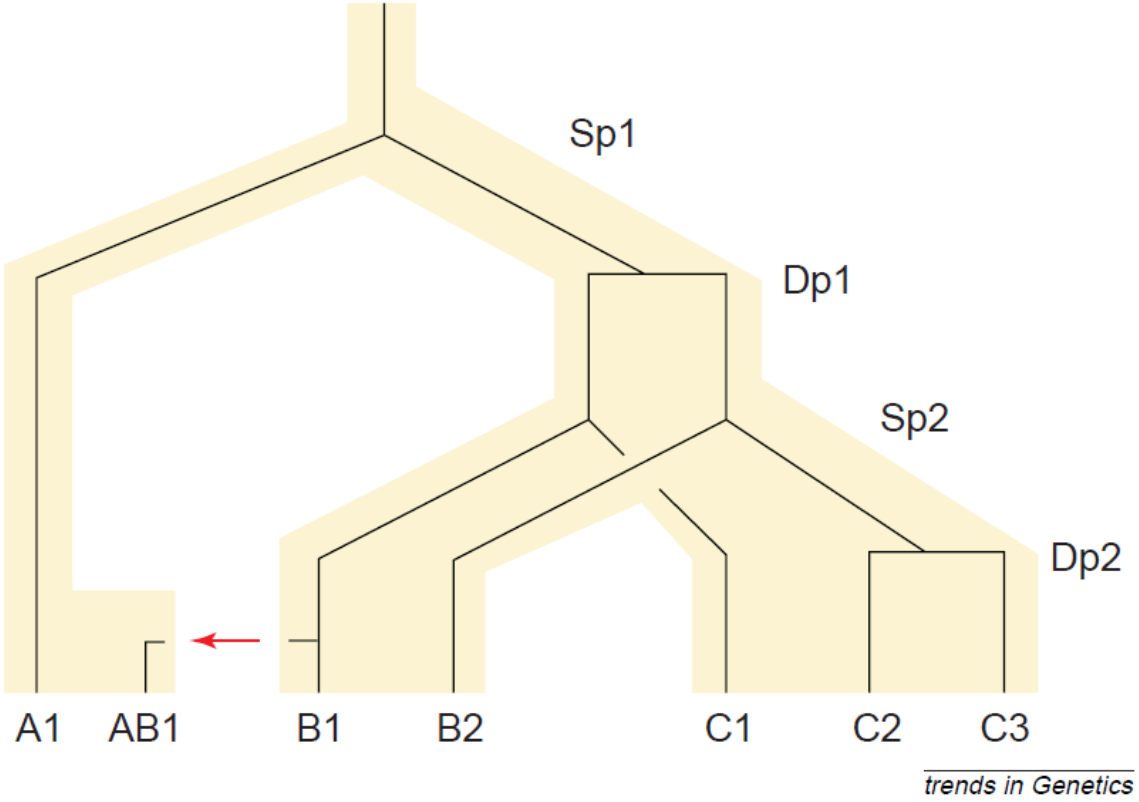
- For the 30th anniversary of his landmark paper, Fitch revisited the subject and published an equally lucid and succinct discussion of various aspects of homology.
  - the best reading on this subject!

Orthology, paralogy and Xenology



- **The evolution of a gene from a common ancestor descending to three populations A, B and C.**
  - Two speciation events : Sp1 and Sp2 (Y junction),
  - Two gene-duplication events: Dp1 and Dp2 (horizontal bar).

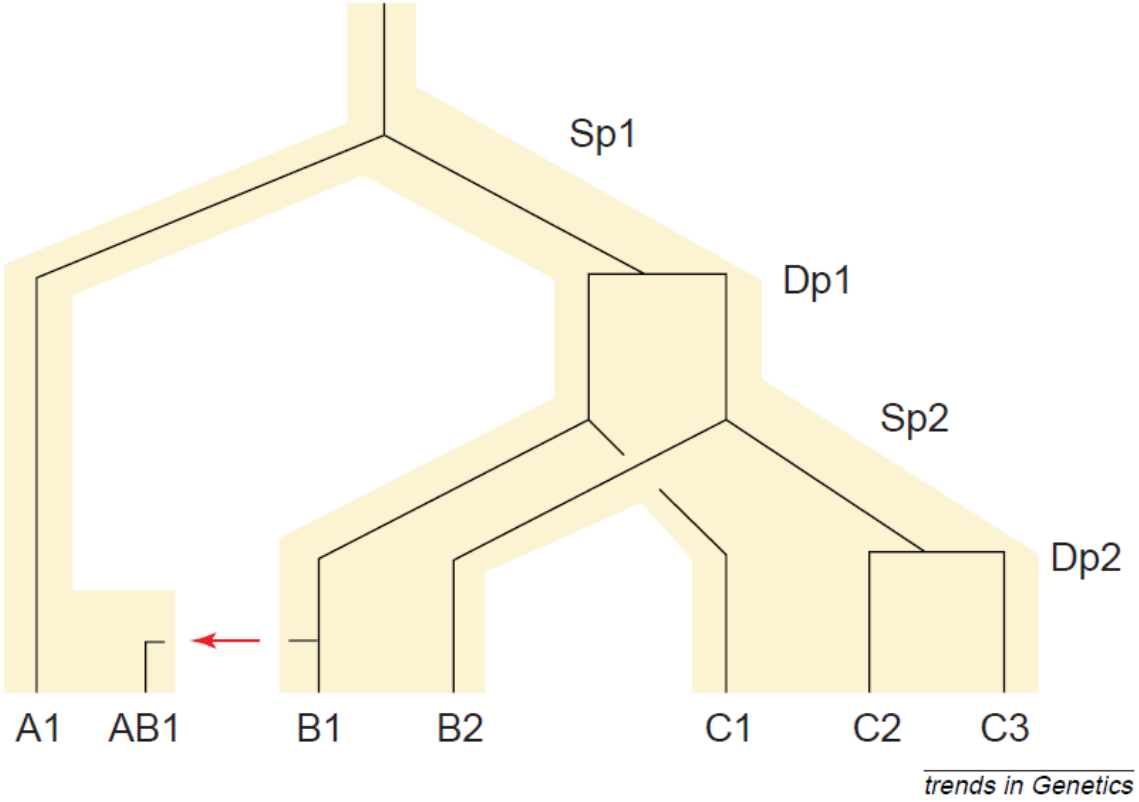
Orthology, paralogy and Xenology



- **orthologs**: two genes with last common ancestor at a Y junction (speciation)
- **Paralogs**: two genes last common ancestor at a horizontal bar junction (gene duplications)

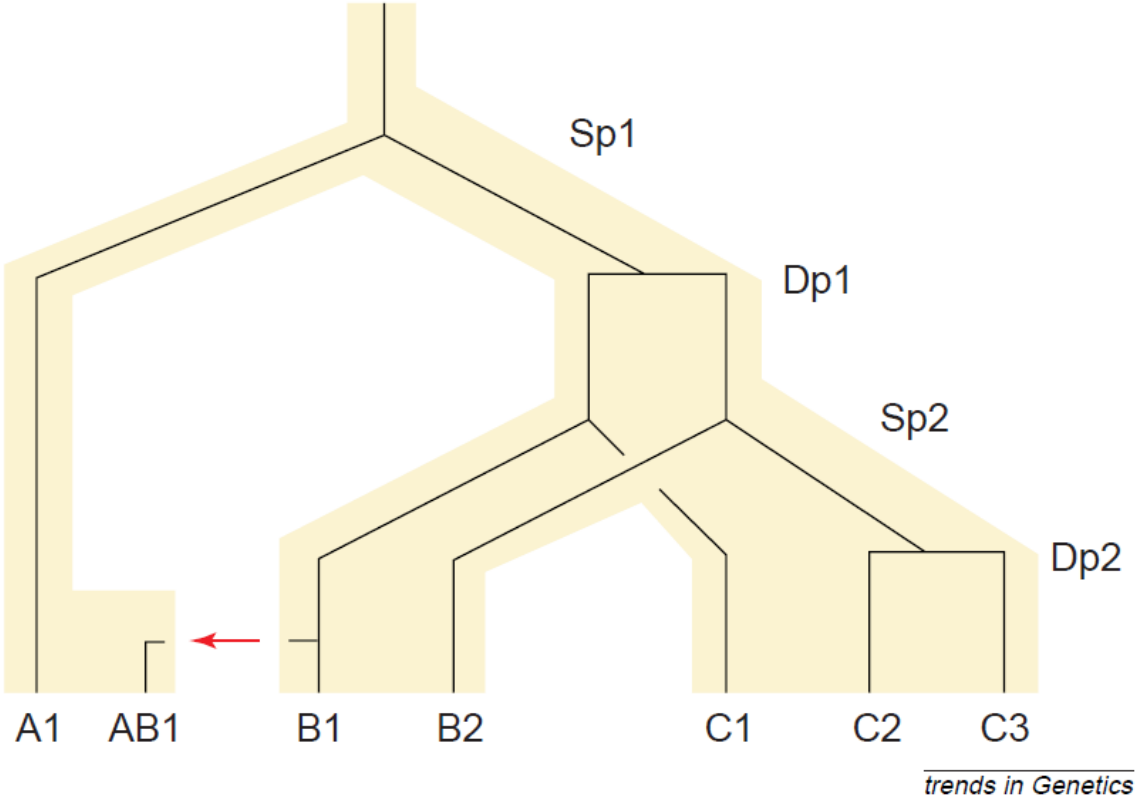
- C2 and C3 are paralogs, but are orthologous to B2.
- Both are paralogous to B1 but orthologous to A1.

Orthology, paralogy and Xenology



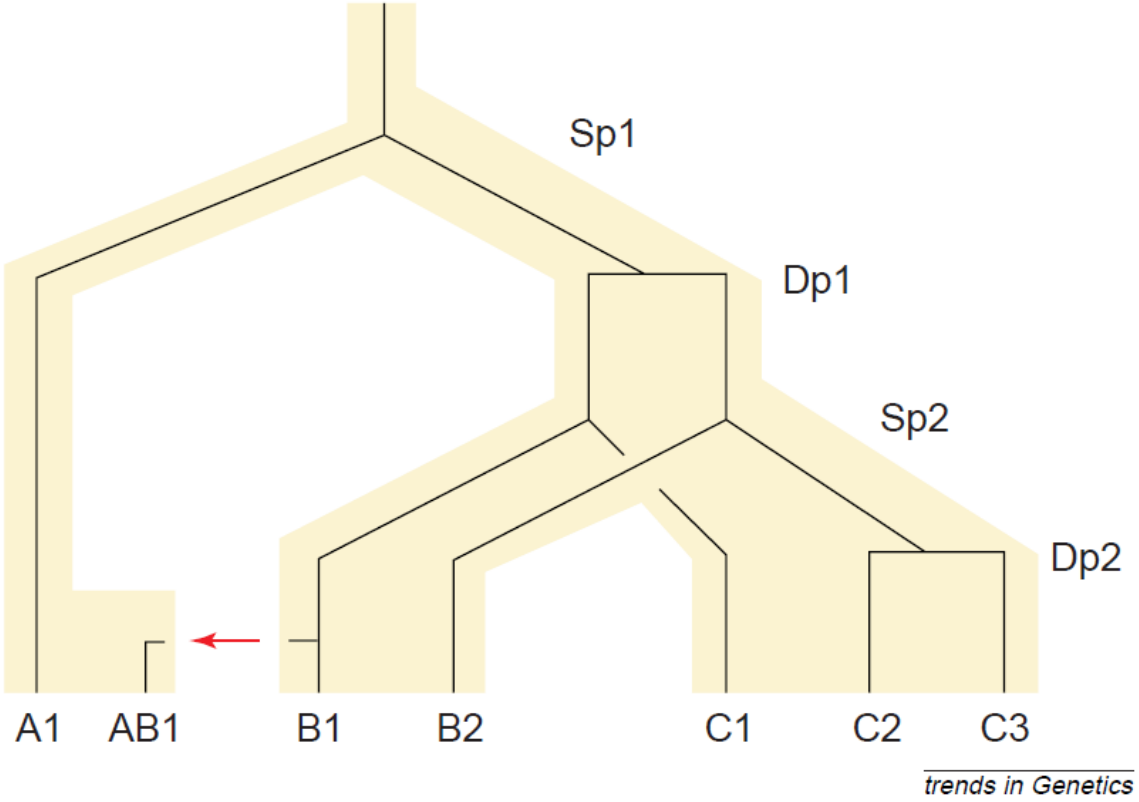
- Red arrow : transfer of the B1 gene from species B to species A. AB1 gene is xenologous to all six other genes.

Orthology, paralogy and Xenology



- Relationships are **reflexive** (A -> B1 implies B1 -> A1 where -> = 'is ortholog to')
- Relationships are not **transitive**. (C2 -> A1 -> C3 is true, but C2 -> C3 is false).

Orthology, paralogy and Xenology



Homology is an **abstraction**: it is a relationship, common ancestry, but which we can **only infer** with more or less certainty.

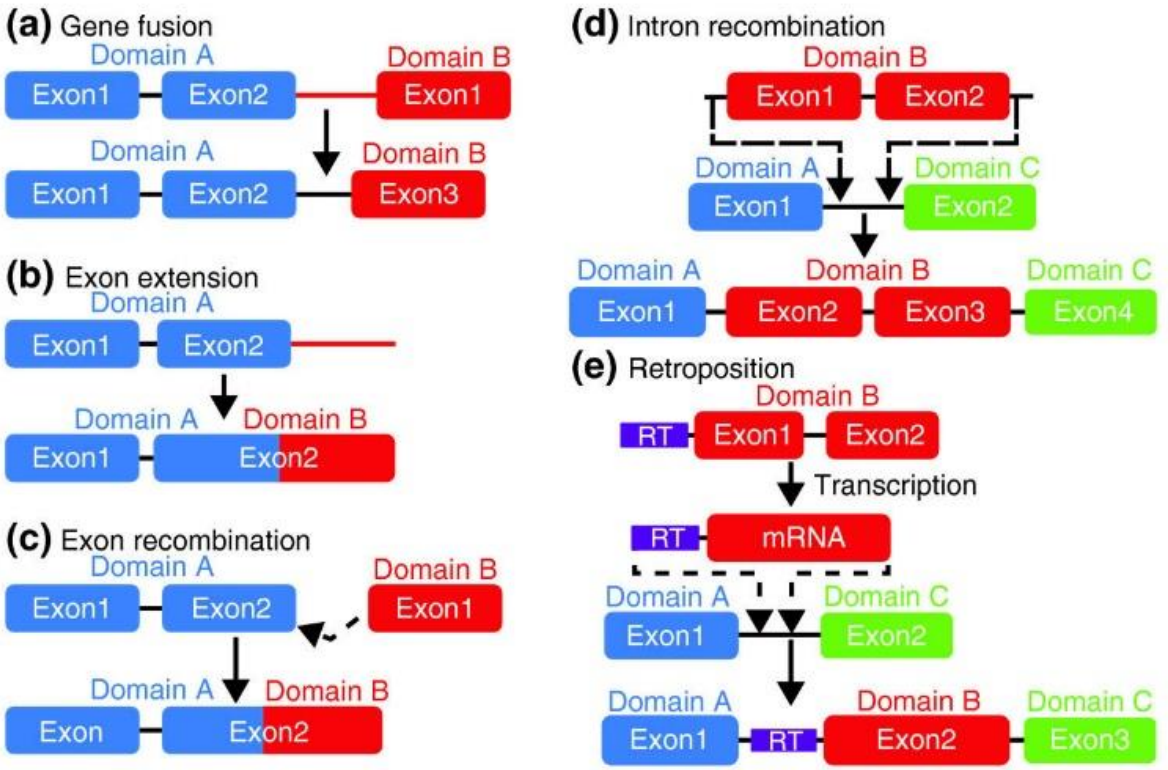
The bird/bat limbs problem

- **Are their forelimbs homologous or not**?

The bird/bat limbs problem

- The forelimbs of the bat and the bird are adapted to flight, but the evolution to flight occurred independently in each lineage.
- Their cenancestral limb is the forelimb of a flightless reptile that is itself the **reptilian cenancestor** of the birds and mammals.
  - Thus the limbs are (structurally) orthologous.
- On the other hand, the flight of birds and bats is (functionally) analogous.

The recombination problem



- not all parts of a gene have the same history => the gene is not the unit to which the terms orthology and paralogy apply.

- If the domain that is homologous to the domain 1 constitutes 20% of the protein then the protein is only 20% homologous to that domain (irrespective of its percent identity)
- This is the only situation where 'percent homology' has a legitimate meaning => **partial homology**.

Orthologs and Paralogs: Deriving Clusters of Orthologous Groups

Science. 1997 Oct 24;278(5338):631-7.

A genomic perspective on protein families.

Tatusov RL[1], Koonin EV, Lipman DJ.

COGs

Phylogenetic classification of proteins encoded in complete genomes

NCBI

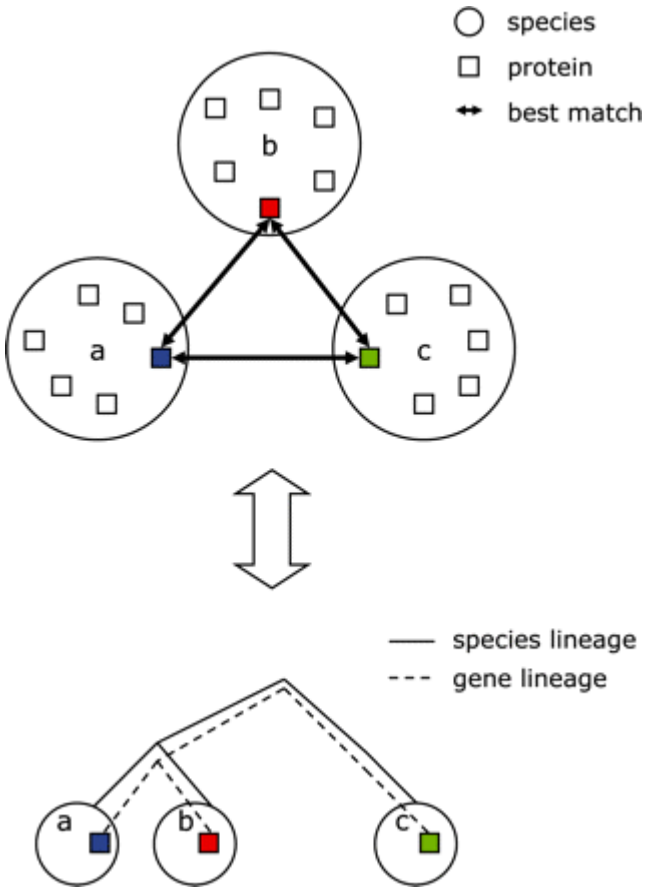
COGs on FTP

- 2003 COGs, 2014 update, HTML NEW
- 2003 COGs, 2014 update, data NEW
- 2003 COGs, original format
- 2003 KOGs, original format
- 2003 COGs
- arCOGs NEW
- NCVOGs
- mimiCOOGs
- 2011 POGs, annotated
- 2011 POGs, extended
- 2013 POGs
- COG software

Publications

- Original COG paper. *Science* 1997 Oct 24;278(5338):631-7
- 2003 database update. *BMC Bioinformatics* 2003 Sep 11;4(1):41
- 2003 eukaryotic KOGs. *Genome Biol.* 2004 Jan 15;5(2):R7.
- Cyanobacterial COGs. *Proc Natl Acad Sci U.S.A.* 2006 Aug 29;103(35):13126-13131.
- Lactic acid bacteria COGs. *Proc Natl Acad Sci U.S.A.* 2006 Oct 17;103(42):15611-15616.
- 2007 archaeal COGs. *Biol Direct* 2007 Nov 27;2:33.
- NCLDV COGs. *Virol J.* 2009 Dec 17;6:223.
- Improved COG algorithm. *Bioinformatics* 2010 Jun 15;26(12):1481-1487.
- 2011 phage COGs. *J Bacteriol.* 2011 Apr;193(8):1806-1814.
- Orthologs and BBH. *Genome Biol Evol.* 2012 Jan;4(12):1286-1294.
- 2012 archaeal COGs. *Biol Direct* 2012 Dec 14;7:46.
- 2013 phage COGs. *J Bacteriol.* 2013 Mar;195(5):941-950.
- mimiCOGs. *Virol J.* 2013 Apr 4;10:106.
- 2014 update of 2003 COGs. *Nucleic Acids Res.* 2015 Jan;43:D261-D269. NEW
- 2014 archaeal COGs. *Life* 2015 Mar 10;5(1):818-840. NEW

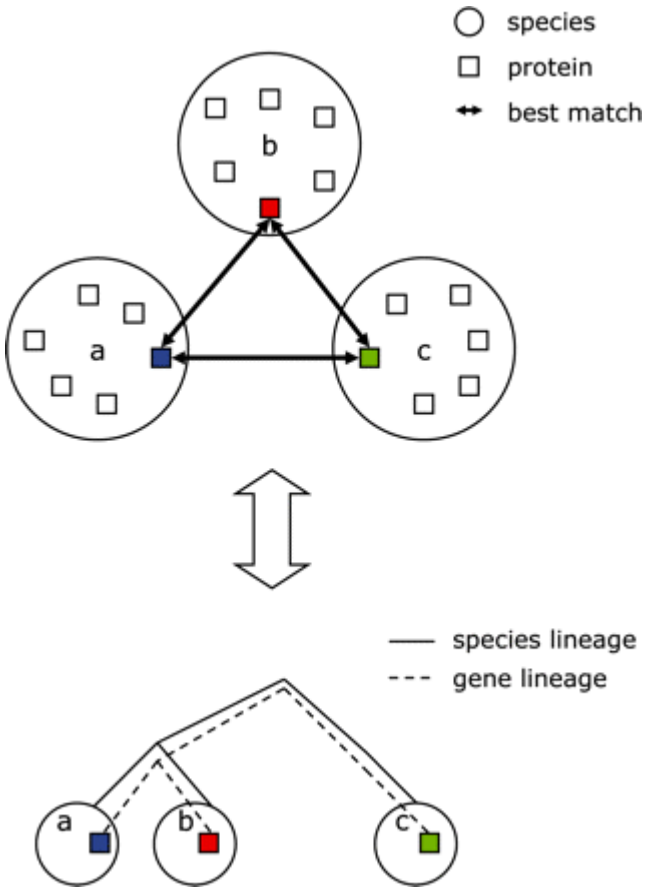
Orthologs and Paralogs: Deriving Clusters of Orthologous Groups



- Existence of one-to-many and many-to-many orthologous relationships,
  - identifying orthologs as the delineation of **clusters of orthologous groups** (COGs).

- Each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events.

Orthologs and Paralogs: Deriving Clusters of Orthologous Groups

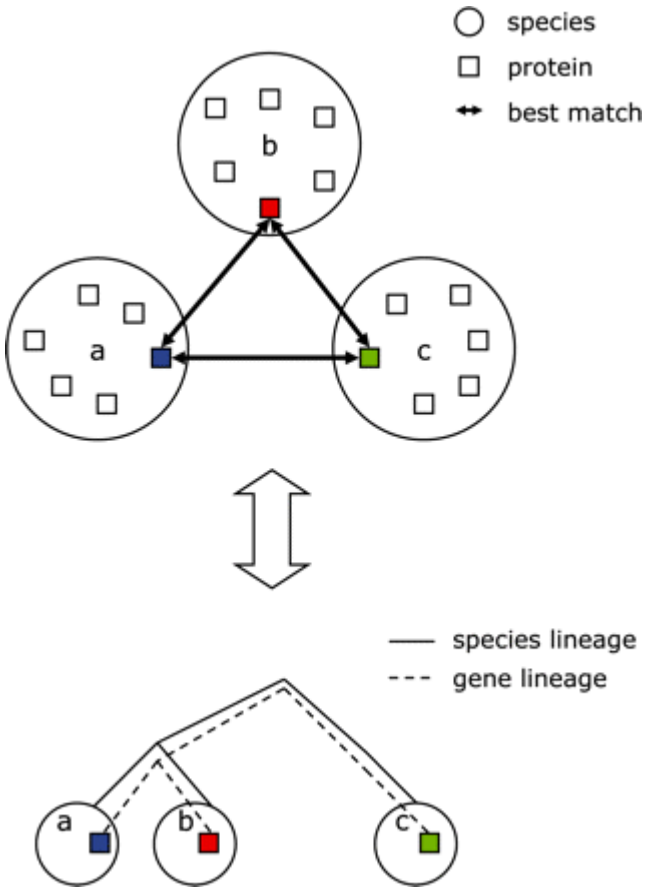


To delineate the COGs,

- All pairwise sequence comparisons were performed,
- For each protein, the best match in each of the other genomes was detected.
- COGs = merging adjacent triangles

- Triangle does not depend on the absolute level of similarity between the compared proteins and thus allows the detection of orthologs among both slowly and quickly evolving genes.
- Because of the existence of paralogs, the best marches that form the triangles are not necessarily symmetrical.

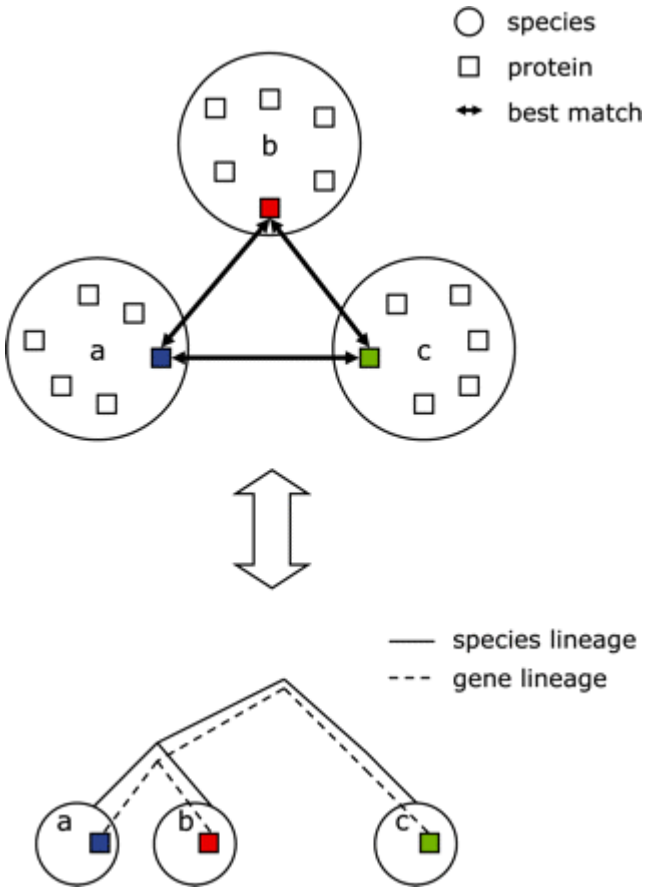
Orthologs and Paralogs: Deriving Clusters of Orthologous Groups



In certain cases, COGs may be lumped together.

- Multidomain proteins:
  - individual domains were isolated and a second iteration of the sequence comparison was performed
- Differential gene loss:
  - Some of the COGs may include proteins from different lineages that are paralogs rather than orthologs
  - the level of sequence similarity between the members of each cluster was analyzed, and clusters that seemed to contain two or more COGs were split.

Orthologs and Paralogs: Deriving Clusters of Orthologous Groups



Only five major, phylogenetically distant clades were used as independent contributors to COGs:

1. Gram-negative bacteria (*Escherichia coli* and *H. influenzae*),
2. Gram-positive bacteria (*Mycoplasma genitalium* and *M. pneumoniae*),
3. Cyanobacteria (*Synechocystis sp.*),
4. Archaea (Euryarchaeota) (*Methanococcus jannaschii*),
5. Eukarya (Fungi) (*Saccharomyces cerevisiae*)

Not regularly updated due to the manual labor required!

Measuring genome evolution

- (Huynen and Bork 1998)
- First automated method based on **best bi-directional hits** (BBH) between a pair of species

- **9 sequenced** Archaea and Bacteria that were publicly available!

- Genomes can be compared at a **variety of levels**:
  - the fraction of **orthologous** sequences between genomes,
  - the conservation of **gene order** between genomes,
  - the conservation of **spatial clustering** of genes (operon).

Identification of Orthologous Genes

- The **most straightforward approach** to identifying orthologous genes is to compare all genes in genomes with each other, and then to select pairs of genes with significant pairwise similarities.
  - A pair of sequences with the highest level of identity then is considered orthologous.

- **Auxiliary information for detection of orthology.**

- **Synteny**
  - the presence in both genomes of neighboring sequences that are also orthologs of each other.
  - But, the potential for using synteny for identifying orthologs is limited mainly to genomes that have speciated only relatively recently.

- **Third genome (triangle)**
  - If two genes from different genomes have the highest level of identity both to each other and to a single gene from a third genome, then this is a strong indication that they are orthologs.

Identification of Orthologous Genes: hampered by a variety of evolutionary processes

- Orthologs identification is **hampered** by a variety of evolutionary processes.
- **Sequence divergence**: homolog sequences can diverge ‘‘beyond recognition’’
- **Nonorthologous** gene displacement: two nonorthologous genes that are unrelated or only remotely related perform the same function in two organisms.
- **Gene loss**: If two genomes lose different paralogs of an ancestral gene that was duplicated before the speciation event, the remaining genes have highest sequence identity even though they are not orthologs
- **Horizontal gene transfer**: the genes still could be orthologs
  - But horizontal gene transfer and ancient gene duplications cannot be distinguished!
- **Orthology in multidomain proteins**: In multidomain proteins two levels of orthology can be distinguished: one is at the level of single domains, a second at the level of the whole protein.
  - In bacteria/Archaea: high occurrences of ‘‘gene fusion’’ or ‘‘gene splitting’’

From homologs to orthologs

- Orthologs are defined in the following manner:

  1. they have the highest level of **pairwise identity (BBH)** when compared with the identities of either gene to all other genes in the other's genome;

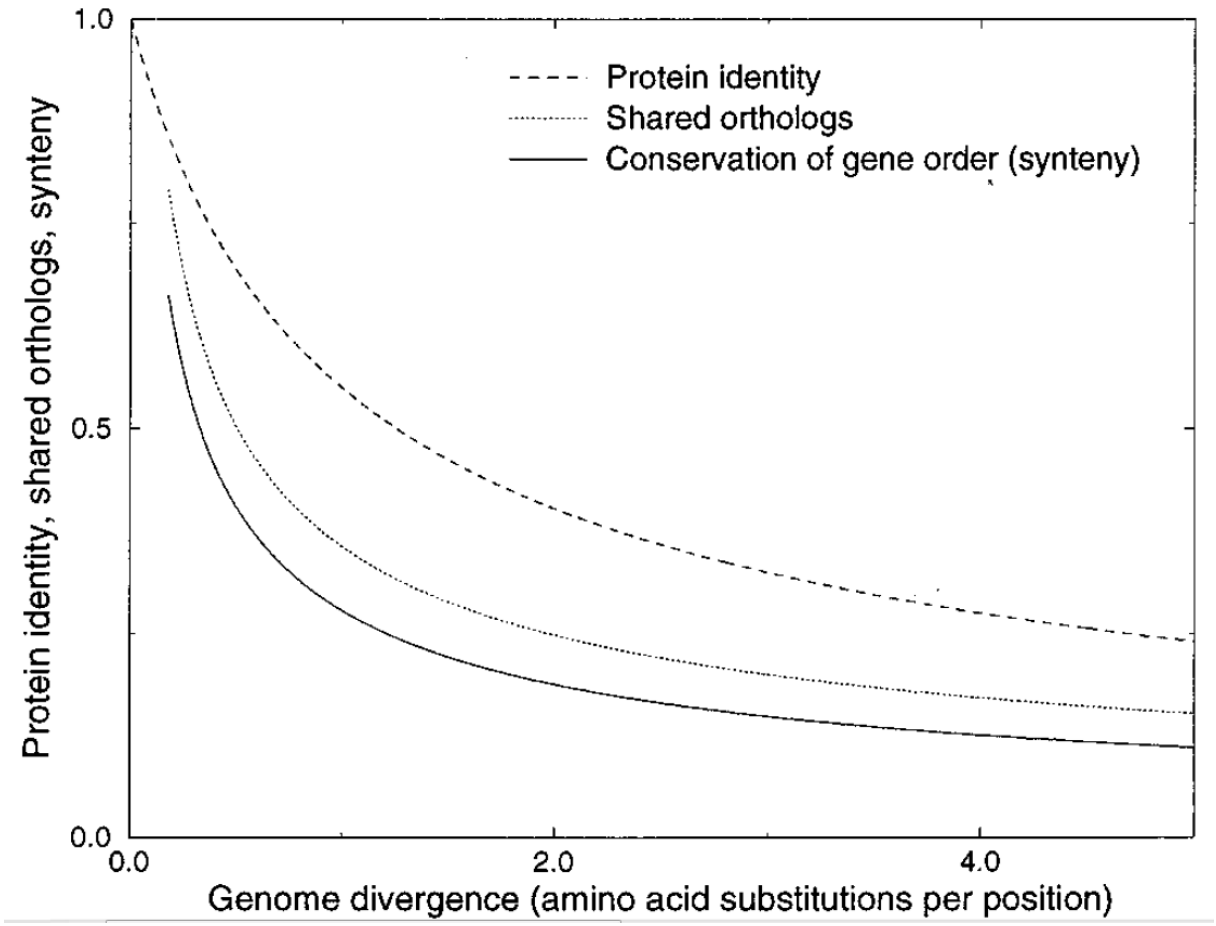
  2. the pairwise identity is **significant** ($E < 0.01$),

  3. the **similarity extends** to at least 60% of one of the genes.

- The region of similarity is not required to cover the majority of both genes to include the possibility of gene fusion and gene splitting.

- Method does not detect paralogs!

From homologs to orthologs



Relative rates of genome evolution (Huynen and Bork 1998)

INPARANOID

Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons (Remm et al. 2001)

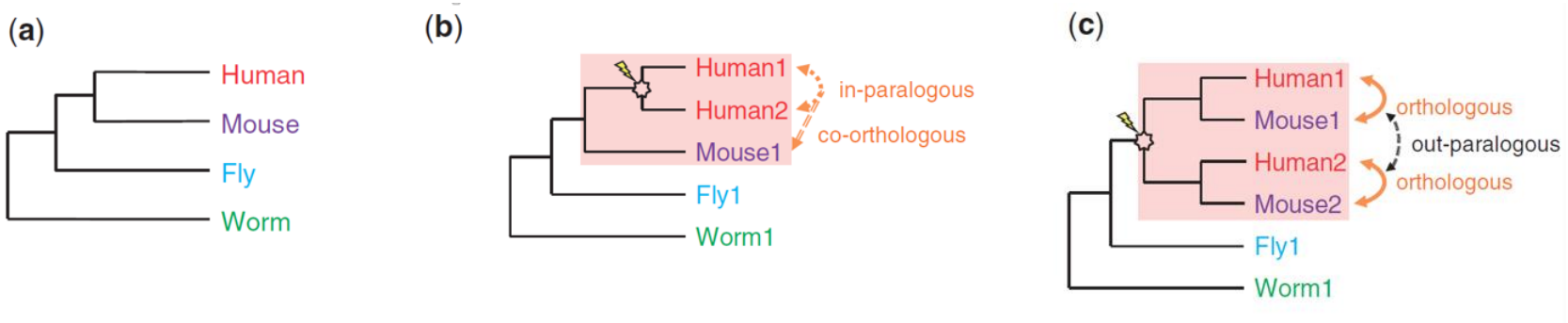
- Motivation = comparison of genome pairs:
  - Which genes in the human genome are sharing the exact same biological function with genes in simpler organisms?
  - Which are the human orthologs of a given Drosophila gene or which are the mouse orthologs of a given human gene?

Gene function!

Gene function

- **Orthologs:** Genes in two species that have are most likely to share the same function.
- **Recent duplication**
- If the sequences have duplicated after the speciation event. In this case there is **more than one ortholog** in one or both species (one-to-many or many-to-many relationship).
- In such cases, it is non-trivial to determine which of the orthologs is functionally equivalent to the ortholog in the other species. It may be only one, but several genes could also have redundant functions.
- There may also be paralogs that arose from a duplication event before the speciation.
    - These are therefore not orthologs.

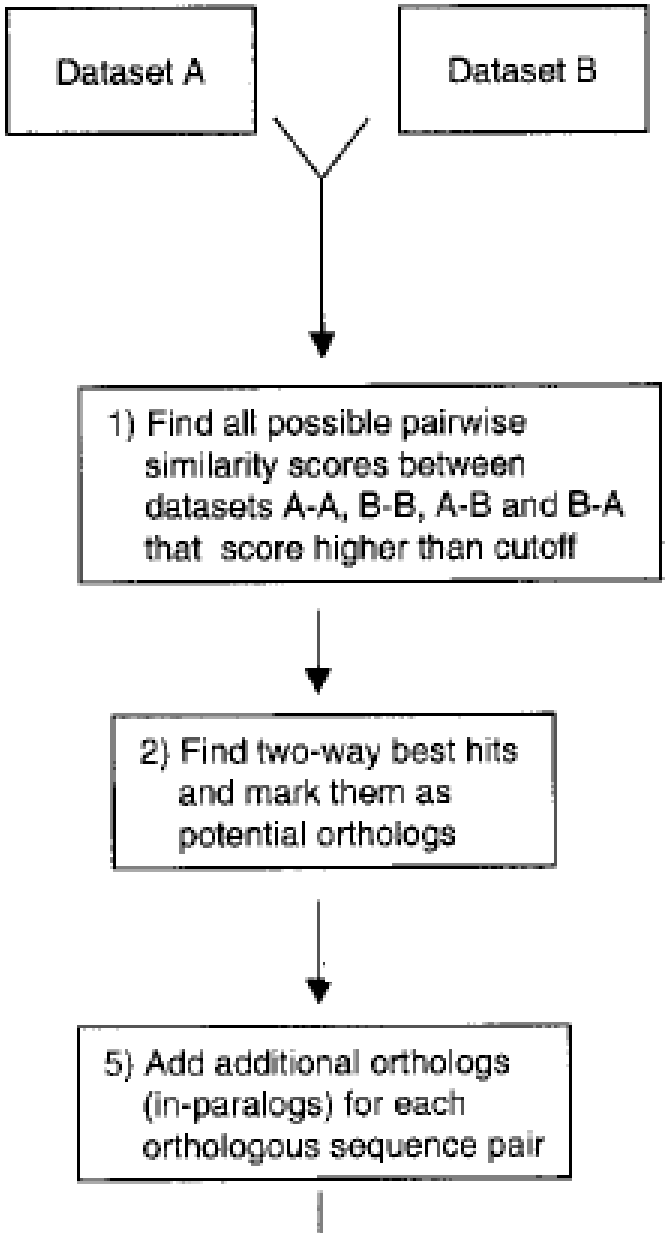
Distinction between paralogs



- **Distinction between paralogs**
  - Paralogs duplicated before a speciation event
  - Paralogs duplicated after a speciation event.
- In analogy with the phylogenetic concepts of **outgroup** and **in-group**.
  - **out-paralogs**: paralogs predating the speciation event
  - **in-paralogs**: paralogs that were duplicated after the speciation event (co-orthologs)
- **Automatic detection of orthologs and in-paralogs**
  - **Phylogenetic trees**, the natural way.
  - An alternative: **all-versus-all sequence comparison** between two genomes.

- **INPARANOID**, that identifies orthologs and in-paralogs between any given pair of genomes.

Identifies orthologs and in-paralogs

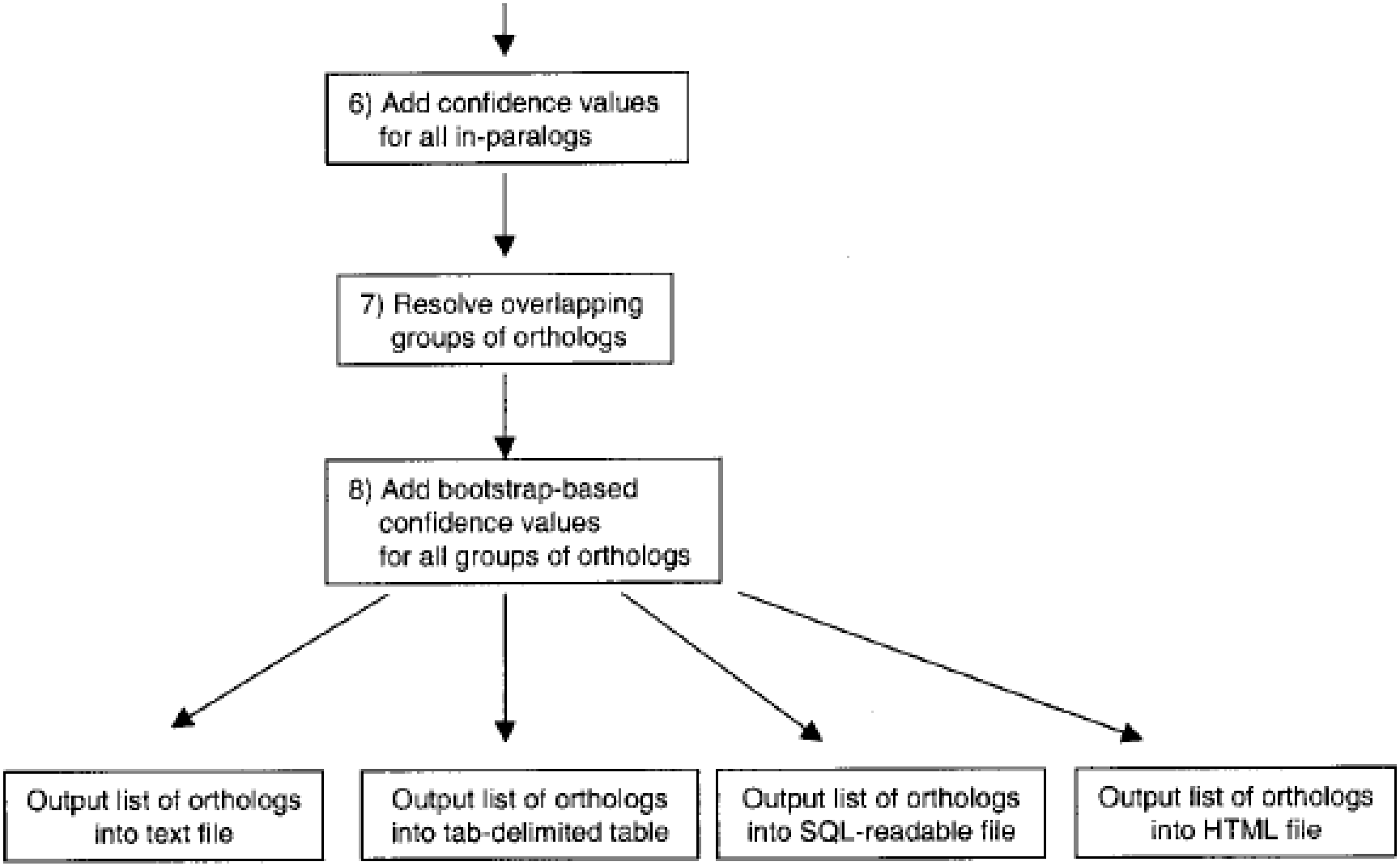


Two cut-off values

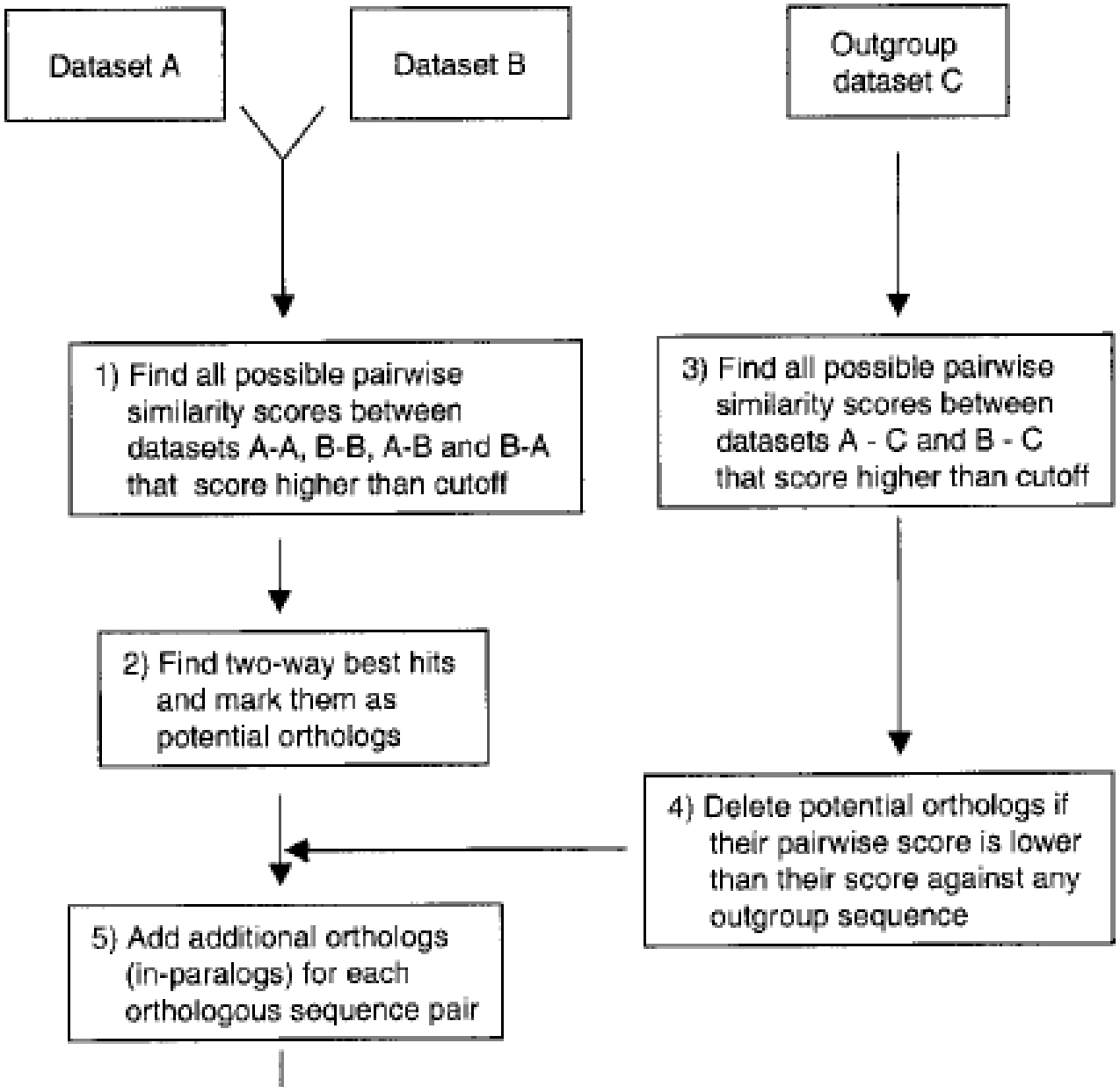
1. **a score** cut-off to separate significant scores from spurious matches
2. **an overlap** cut-off to avoid short, domain-level matches.

Orthologous sequences are expected to maintain the homology over the majority of their length (>50%)

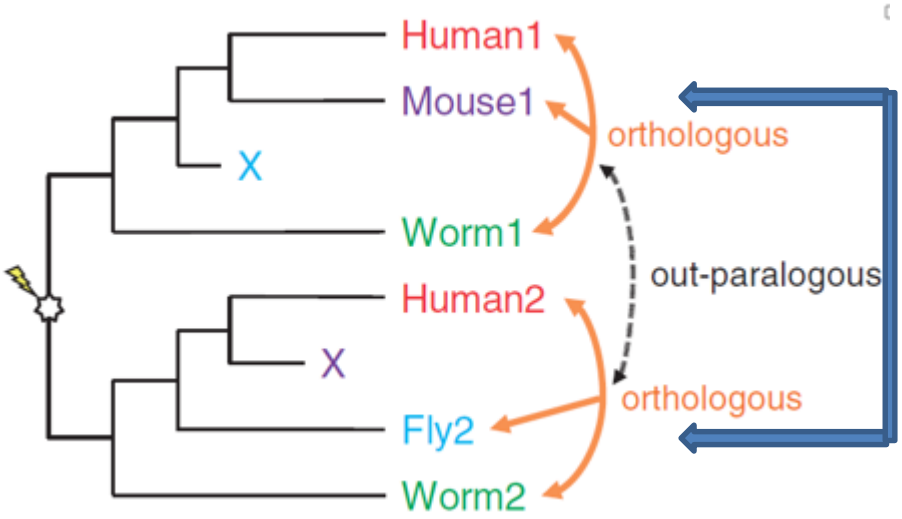
Identifies orthologs and in-paralogs



Identifies orthologs and in-paralogs



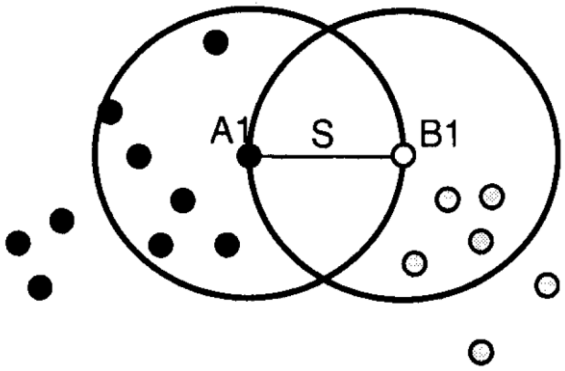
Identifies orthologs and in-paralogs



Clustering algorithm

- **Clustering algorithm:** Find non-overlapping groups of orthologous sequences using pairwise similarity scores.

- Mutually best hits (**BBH**) are marked as the **main ortholog pair** of a given ortholog group (A1, B1).



Each circle represents a sequence from species A (black) or species B (grey).

Main orthologs (BBH pairs) are denoted A1 and B1. With a similarity score = S.

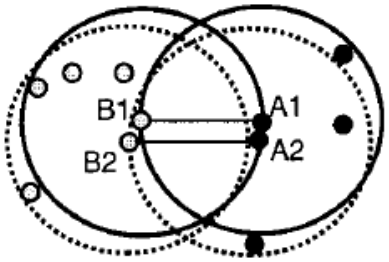
- The assumption for clustering of in-paralogs
  - the main ortholog is more similar to in-paralogs from the same species than to any sequence from other species.

- All **in-paralogs** with score S or better to the main ortholog are inside the circle with diameter S that is drawn around the main ortholog.

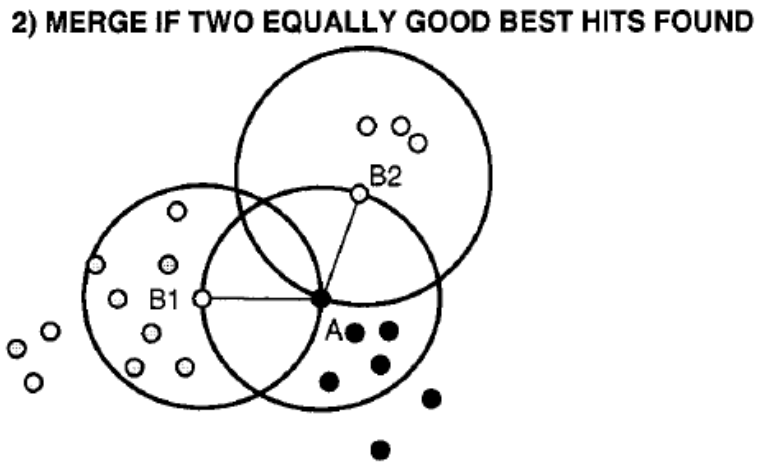
- Sequences outside the circle are **out-paralogs**.

- **Overlap between groups**
- **The rules for resolving overlapping groups of in-paralogs.**

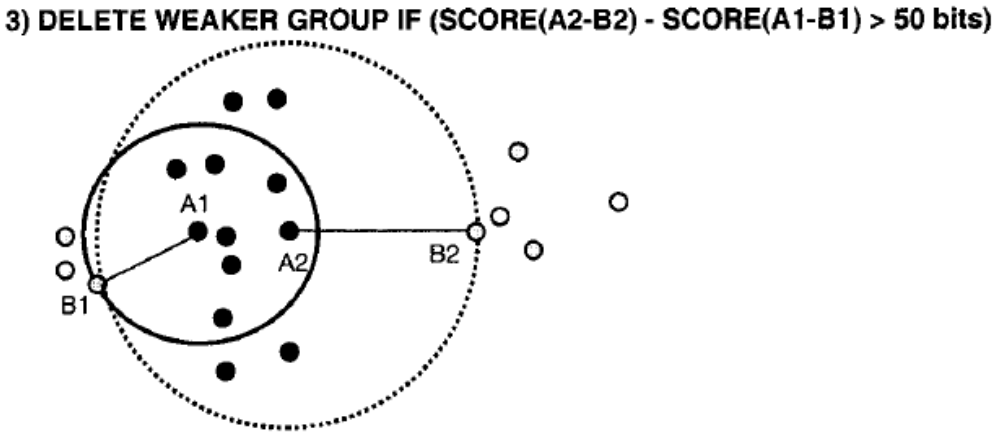
1) MERGE IF BOTH ORTHOLOGS ARE ALREADY CLUSTERED IN THE SAME GROUP



- **Overlap between groups**
- **The rules for resolving overlapping groups of in-paralogs.**

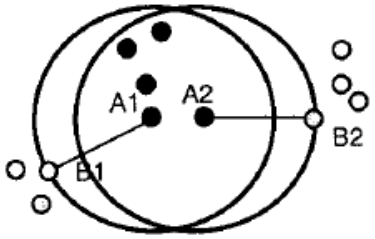


- **Overlap between groups**
- **The rules for resolving overlapping groups of in-paralogs.**

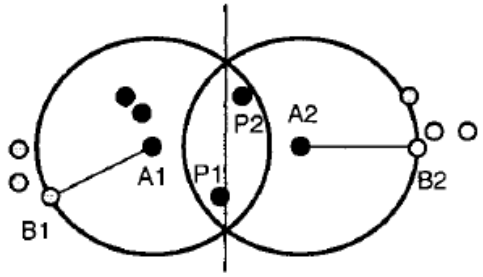


- **Overlap between groups**
- **The rules for resolving overlapping groups of in-paralogs.**

4) MERGE IF (SCORE(A1-A2) < 0.5 * SCORE(A1-B1))



5) DIVIDE IN-PARALOGS IN OVERLAPPING AREAS



- **Confidence values for in-paralogs**
- The confidence value simply shows how far a given sequence is from the main ortholog of the same species on a scale between 0% and 100%.

$$\text{Confidence for } A_p = 100\% \times (\text{score}AA_p - \text{score}AB)/(\text{score}AA - \text{score}AB)$$

$$\text{Confidence for } B_p = 100\% \times (\text{score}BB_p - \text{score}AB)/(\text{score}BB - \text{score}AB)$$

- where,
  - *Ap* is an in-paralog from dataset A,
  - *Bp* is an in-paralog from dataset B,
  - *A* is the main ortholog from dataset A,
  - *B* is the main ortholog from dataset B,
  - scoreXY is the similarity score between protein X and Y in bits.

- **Bootstrap values for groups of orthologs**

- Estimate the reliability of each orthologous group.
- The bootstrap values are calculated by comparing two pairwise sequence alignments.

- These two alignments are between main ortholog pair (*A1*, *B1*) and between an alternative, lower-scoring alignment (*A1*, *B2*).
- The columns in alignments between sequences (*A1*, *B1*) and (*A1*, *B2*) are sampled with replacement, considering an insertion as a single unit.
- The bootstrap value is expressed as the fraction of sampled alignments that support the hypothesis (*A1*, *B1*), and not (*A1*, *B2*.)
- Repeat for (*A1*, *B1*) versus (*A2*, *B1*)

InParanoid 8

InParanoid: ortholog groups with inparalogs

273 organisms: 3718323 sequences

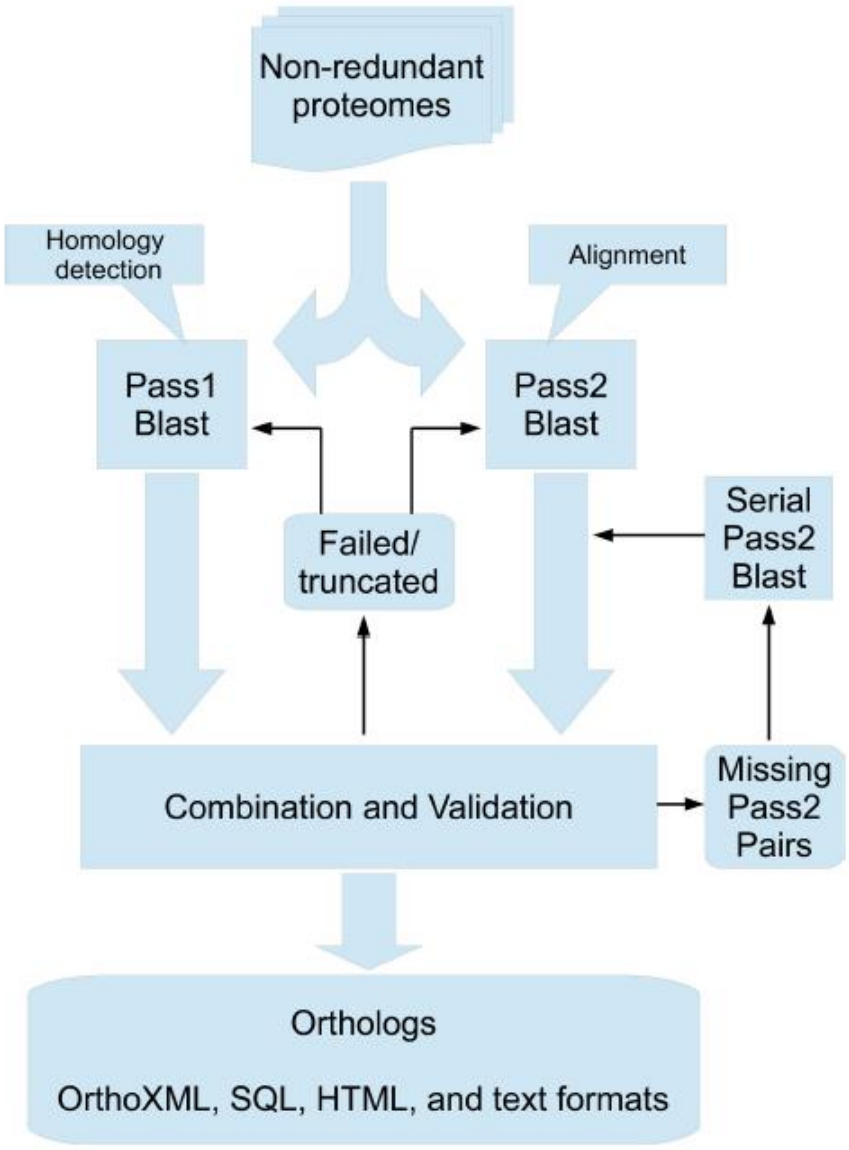
Version 8.0, Updated December 2013 (release notes)

BROWSE the database - Select two species and view all their orthologs
SEARCH BY SEQUENCE IDs - View orthologs of a specific gene or protein
TEXT SEARCH - Query InParanoid by keywords
BLAST SEARCH - Find orthologs in InParanoid similar to your protein sequence
DOWNLOAD DATA - Obtain tables, html, orthoXML, sequences and core data
SUMMARY OF INPARANOID - Statistics of the database and genomes used
ORTHOPHYLOGRAM - Phylogenetic tree based on the average fraction of InParanoid orthologs between species.

Stand-alone InParanoid Program

InParanoid Version 4.1 is available here

InParanoid 8 workflow



In version 4.1, the two BLAST passes are run after each other

- the first run to find all homologs between two species (avoids false low-complexity matches),
- and then a second run is launched per query sequence to make **accurate alignments** with only the homologs found in pass 1.

Workflow used for generating InParanoid 8.

BLAST runs are launched for all pairs of proteomes, running both passes in parallel.

OrthoMCL

- OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes (Li et al. 2003; Chen et al. 2006)

- Approach similar to INPARANOID, but differs primarily in the requirement that
- recent paralogs must be more similar to each other than to any sequence from other species.

- To resolve the many-to-many orthologous relationships inherent in comparisons across multiple
- genomes, OrthoMCL applies MCL.

- **Challenges for Comparative Eukaryotic Genomics**

1. Compared to prokaryotes, eukaryotic genomes tend to exhibit a much **higher rate** of duplicative gene family expansion.
   - Difficult to distinguish functional redundancy from functional divergence.
     - Genes that have evolved from relatively "ancient" duplication events may have diverged to acquire new functions
       - these homologs should not be clustered with true ortholos.
2. Complicated domain architecture of many proteins.
   - Multidomain proteins with different functions may be mistakenly clustered into a single group because they share domains.
3. Incompleteness of genome sequence data.

OrthoMCL

- OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes (Li et al. 2003; Chen et al. 2006)

- Approach similar to INPARANOID, but differs primarily in the requirement that
- recent paralogs must be more similar to each other than to any sequence from other species.

- To resolve the many-to-many orthologous relationships inherent in comparisons across multiple
- genomes, OrthoMCL applies MCL.

- **Challenges for Comparative Eukaryotic Genomics**

1. Compared to prokaryotes, eukaryotic genomes tend to exhibit a much **higher rate** of duplicative gene family expansion.
   - Difficult to distinguish functional redundancy from functional divergence.
     - Genes that have evolved from relatively "ancient" duplication events may have diverged to acquire new functions
       - these homologs should not be clustered with true ortholos.
2. Complicated domain architecture of many proteins.
   - Multidomain proteins with different functions may be mistakenly clustered into a single group because they share domains.
3. Incompleteness of genome sequence data.

Identification of Orthologous Groups by OrthoMCL

- **Identification of Orthologous Groups by OrthoMCL**

1. The OrthoMCL procedure starts with all-against-all BLASTP comparisons of a set of protein sequences from genomes of interest.
2. Putative orthologous relationships are identified between pairs of genomes by **reciprocal best similarity pairs**.
3. For each putative ortholog, **probable "recent" paralogs** are identified as sequences within the same genome that are (reciprocally) more similar to each other than either is to any sequence from another genome.

- A P-value cut-off of 1e-5 was chosen for putative orthologs or paralogs (based on empirical studies).

- **Weighted Graph**

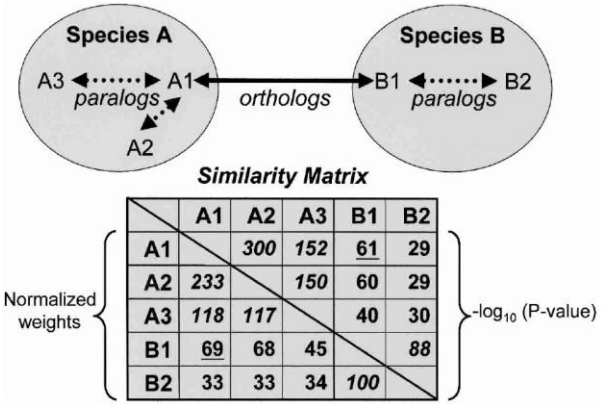
4. Putative orthologous and paralogous relationships are converted into a **graph** in which the nodes represent protein sequences, and the weighted edges represent their relationships.

- Weights are initially computed as the average -log10 (P-value) of BLAST results for each pair of sequences.

- **Weight normalization**

5. Because the high similarity of "recent" paralogs relative to orthologs can bias the clustering process, edge weights are then **normalized** to reflect the average weight for all ortholog pairs in these two species (or "recent" paralogs when comparing within species).

Identification of Orthologous Groups by OrthoMCL



	A1	A2	A3	B1	B2
A1		*300*	*152*	61	29
A2	*233*		*150*	60	29
A3	*118*	*117*		40	30
B1	69	68	45		*88*
B2	33	33	34	*100*	

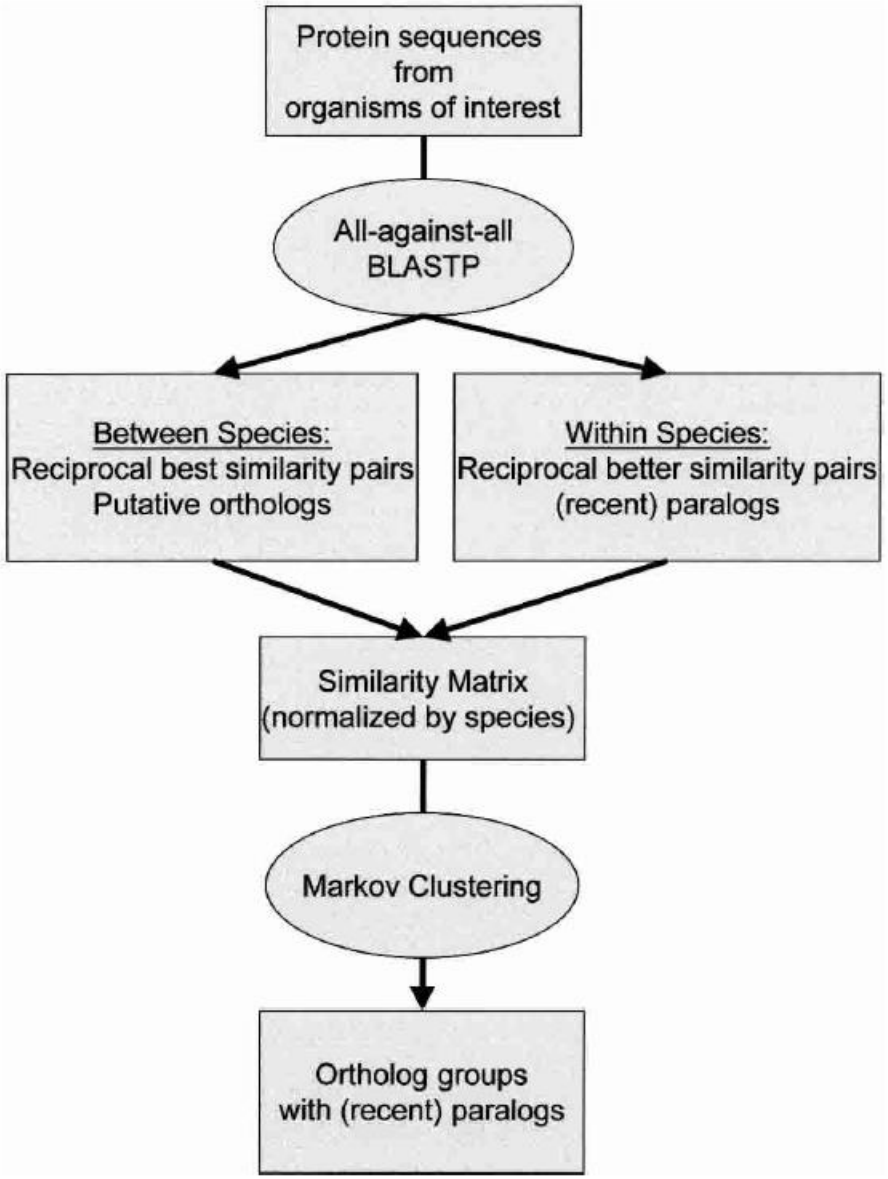
- **Illustration of sequence relationships and similarity matrix construction.**
  Dotted arrows represent "recent" paralogy;
  Solid arrows represent orthology.

  The upper right half of the matrix contains initial weights calculated as average -log10 (P-value) from pairwise WU-BLASTP similarities.

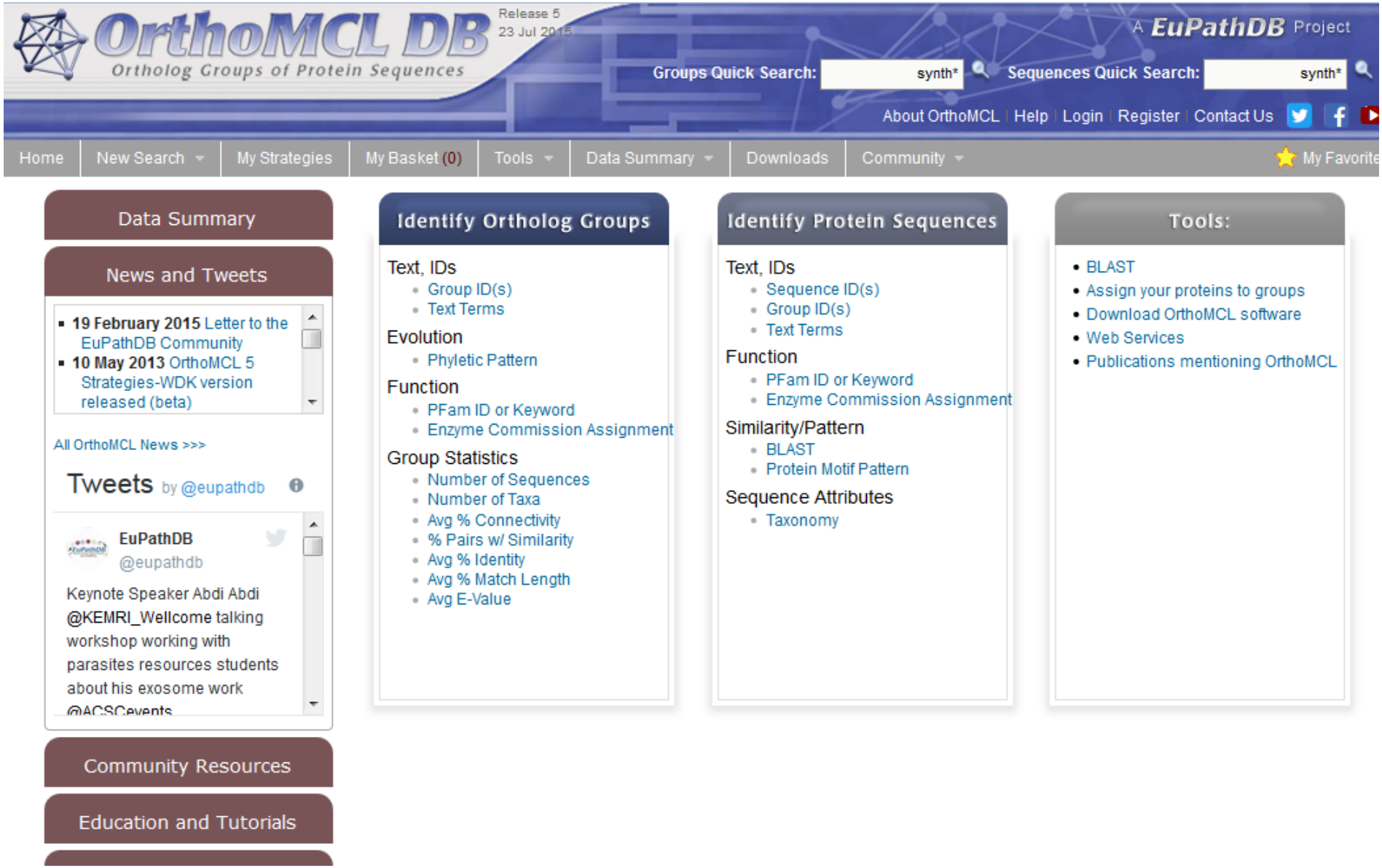
  The lower left half contains corrected weights supplied to the MCL algorithm;

  - The edge weight connecting each pair of sequences w_{ij} is divided by W_{ij}/W, where
  - W represents the average weight among all ortholog (underlined) and "recent" paralog (italicized) pairs,
  - W_{ij} represents the average edge weight among all ortholog pairs from species i and j.

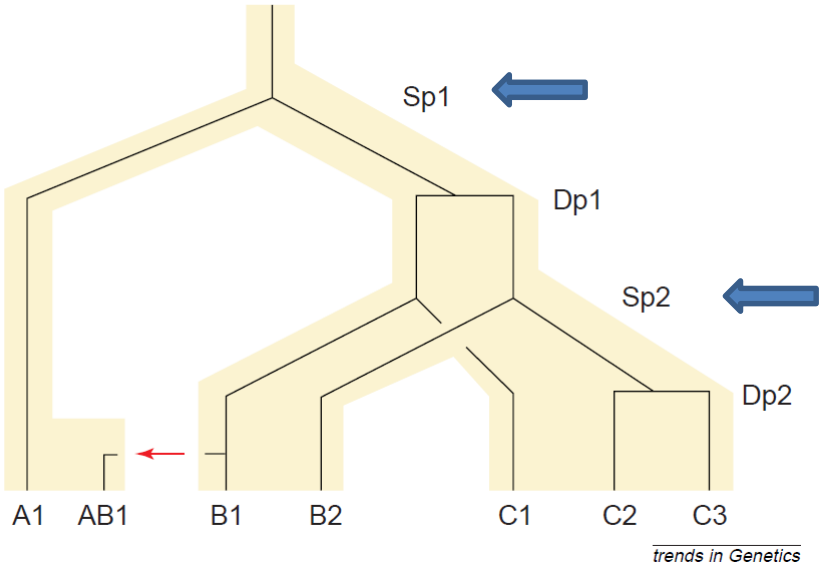
Identification of Orthologous Groups by OrthoMCL



Overlap between groups



Hierarchy of orthologous groups



- An important feature of orthology and paralogy classification, it is relative to a particular ancestor, as orthology of genes is defined by their descent from a common ancestor gene by speciation.
- The more distantly related species are considered the more general (inclusive) orthologous groups become, because all lineage-specific duplications since this last common ancestor should be considered as co-orthologs.
- When closely related species are considered, orthologous groups become more fine-grained (more 1 : 1 relations), as there was less time for gene duplications to occur.

eggNOG: automated construction and annotation of orthologous groups of genes (Powell et al. 2014)

- **eggnog**
  - can be updated without the requirement for manual curation,
  - covers more genes and genomes than existing databases,
  - contains a **hierarchy of orthologous groups** to balance phylogenetic coverage and resolution
  - provides automatic function annotation of similar quality to that obtained through manual inspection

Assemble proteins into orthologous groups using an automated procedure similar to the original **COG/KOG approach**.

- Briefly,

1. compute all-against-all Smith–Waterman similarities among all proteins in eggNOG (low complexity filtering).
2. group recently duplicated sequences into **in-paralogous groups**, which are then treated as single units to ensure that they will be assigned to the same orthologous groups.

   To form the **in-paralogous groups**,

   1. assemble highly related genomes into **clades** (strains of a particular species or close pairs such as human and chimpanzee).
   2. In these clades, join into **in-paralogous groups** all proteins that are more similar to each other (within the clade), than to any other protein outside the clade.
3. start assigning orthology between proteins, by joining triangles of reciprocal best hits (3 different species). At this step In-paralogous groups are represented by their bestmatching member.

Refinements

- This procedure occasionally causes an orthologous group to be split in two;
  - identified by an abundance of reciprocal best hits between groups, which are then joined.
  - Next, relax the triangle criterion and allow remaining unassigned proteins to join a group by simple bidirectional best hits.
- **Identification of gene fusion events**: proteins that bridge unrelated orthologous groups.
  - The different parts of the fusion protein are assigned to their respective orthologous groups.

  This step is crucial for the analysis of eukaryotic multi-domain proteins.

To construct a **hierarchy of orthologous groups**, the procedure described above was applied to several subsets of organisms.

- To make a set of course-grained orthologous groups across all three domains of life, we constructed **non-supervised orthologous groups** (NOGs) from the genes that could not be mapped to a COG or KOG.
- Focusing on eukaryotic genes, we constructed more fine-grained **eukaryotic NOGs** (euNOGs) from the genes that could not be mapped to a KOG.
- Finally, we build sets of NOGs of increasing resolution for five eukaryotic clades: fungi (fuNOGs), metazoans (meNOGs), insects (inNOGs), vertebrates (veNOGs) and mammals (maNOGs).

eggNOG


EggNOG 4.5.1
Search protein or OG
Navigation
Home
Sequence search
eggNOG-mapper
New
(genome-wide functional annotation)
Downloads
API
Methods
Viral OGs
Discover EggNOG 4.5.1
A database of orthologous groups and functional annotation
Organisms
2,031
Viruses
352
Orthologous Groups
190k
Trees & Algs.
1.9M
Search
What's new in version 4.5.1 (Nov 2016)
Added new tool eggNOG-mapper for fast functional annotations of sequence collections
Minor improvements in phylogenetic tree visualization (e.g. Show original sequence names)

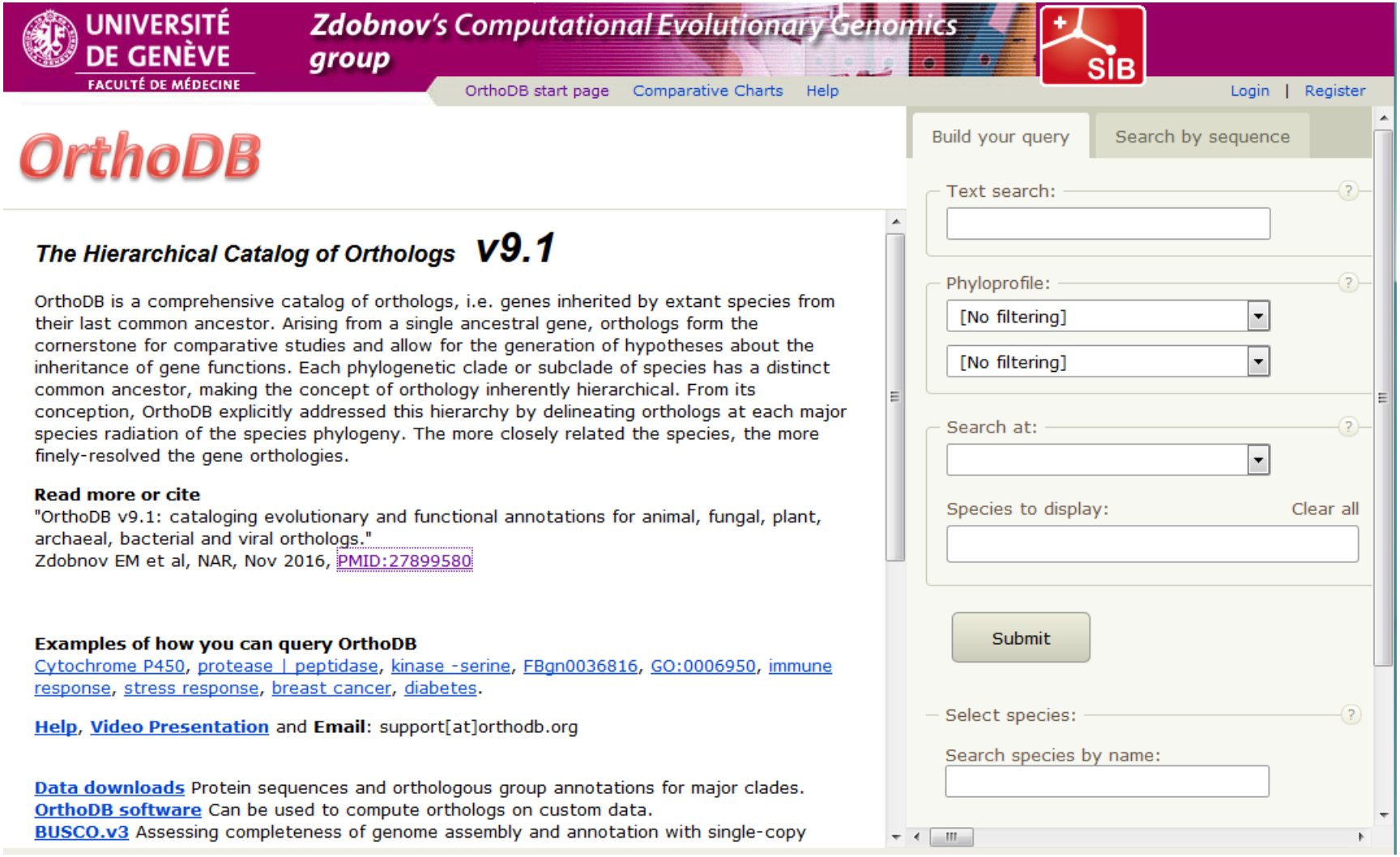
OrthoDB: the hierarchical catalog of eukaryotic orthologs
(Kriventseva et al. 2007; Waterhouse et al. 2011; Zdobnov et al. 2017)

Own implementation of COG-like and Inparanoid-like ortholog identification procedures from all-against-all sequence comparisons across multiple species, and explicitly delineate the hierarchy of the orthologous groups, consistently applying the procedure to the sets of species with varying levels of relatedness according to the species tree.

- **Orthology delineation**
- based on all-against-all protein sequence comparisons using the Smith-Waterman algorithm
- clustering of best reciprocal hits from highest scoring ones to 10^{-6} e-value cutoff for triangulating Best-Reciprocal-Hits, (BRH) or 10^{-10} cutoff for unsupported BRH, and requiring a sequence alignment overlap of at least 30 amino acids across all members of a group.
- The orthologous groups were expanded by genes that are more similar to each other within a proteome than to any gene in any of the other species, and by very similar copies that share over 97% sequence identity.

- The outlined procedure was first applied to all species considered, and then to each subset of species according to the radiation of the phylogenetic tree.

OrthoDB


UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Zdobnov's Computational Evolutionary Genomics group
SIB
OrthoDB start page
Comparative Charts
Help
Login | Register
OrthoDB
The Hierarchical Catalog of Orthologs v9.1
OrthoDB is a comprehensive catalog of orthologs, i.e. genes inherited by extant species from their last common ancestor. Arising from a single ancestral gene, orthologs form the cornerstone for comparative studies and allow for the generation of hypotheses about the inheritance of gene functions. Each phylogenetic clade or subclade of species has a distinct common ancestor, making the concept of orthology inherently hierarchical. From its conception, OrthoDB explicitly addressed this hierarchy by delineating orthologs at each major species radiation of the species phylogeny. The more closely related the species, the more finely-resolved the gene orthologies.
Read more or cite
"OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs."
Zdobnov EM et al, NAR, Nov 2016, PMID:27899580
Examples of how you can query OrthoDB
Cytochrome P450, protease | peptidase, kinase -serine, FBgn0036816, GO:0006950, immune response, stress response, breast cancer, diabetes.
Help, Video Presentation and Email: support[at]orthodb.org
Data downloads Protein sequences and orthologous group annotations for major clades.
OrthoDB software Can be used to compute orthologs on custom data.
BUSCO.v3 Assessing completeness of genome assembly and annotation with single-copy
Build your query
Search by sequence
Text search:
Phyloprofile:
[No filtering]
[No filtering]
Search at:
Species to display:
Clear all
Submit
Select species:
Search species by name:

Hieranoid

Hieranoid: infers orthologs between multiple species by progressively applying the pairwise InParanoid method.

(Schreiber and Sonnhammer 2013; Kaduk and Sonnhammer 2017)

- The progressive idea takes its cue from the "**progressive alignment**" approach.
  - Orthology relationships are inferred at the nodes of a bifurcating guide tree, the species tree.

- Using a hierarchical progressive approach, Hieranoid combines the advantages of
  - **graph-based** methods in that it is computationally less expensive
  - **tree-based methods** in that it produces tree-structured hierarchical groups.

- This progressive approach results in a linear computational complexity.
-
- The reduced computational complexity makes Hieranoid attractive for the analysis of **very large datasets**, which is timely given that thousands of genomes are currently being sequenced.

Hieranoid

Hieranoid: infers orthologs between multiple species by progressively applying the pairwise InParanoid method.

(Schreiber and Sonnhammer 2013; Kaduk and Sonnhammer 2017)

- The progressive idea takes its cue from the "**progressive alignment**" approach.
  - Orthology relationships are inferred at the nodes of a bifurcating guide tree, the species tree.

- Using a hierarchical progressive approach, Hieranoid combines the advantages of
  - **graph-based** methods in that it is computationally less expensive
  - **tree-based methods** in that it produces tree-structured hierarchical groups.

- This progressive approach results in a linear computational complexity.
-
- The reduced computational complexity makes Hieranoid attractive for the analysis of **very large datasets**, which is timely given that thousands of genomes are currently being sequenced.

Input

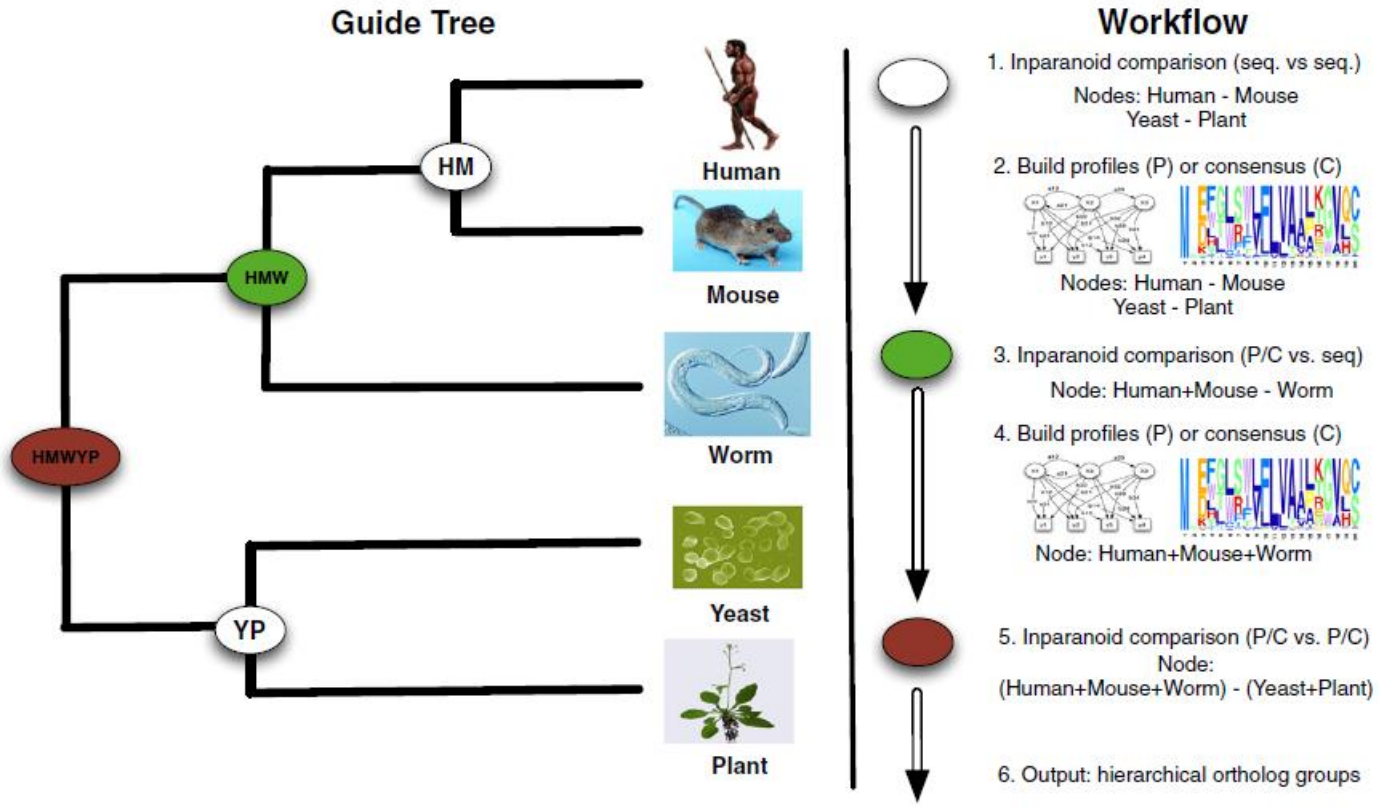
1. a set of proteome sequences from the species under study (FASTA or SeqXML format)
2. a guide tree connecting the species (NCBI taxonomy or user defined) in Newick format.

Progressive orthology inference strategy

- Guide tree: determine order of pairwise comparisons
- Leaves: the species under study
- Inner nodes: hypothetical ancestors or pseudospecies (result of the pairwise orthology inference of the two daughter nodes)

  - Possible pairwise comparisons:
    - a pair of species,
    - a species and a pseudospecies,
    - two pseudospecies.

Hieranoid workflow



Hieranoid workflow

- **Building an initial set of homologs**
-
- Adapted from InParanoid
- Species versus species
  - replace BLAST by USEARCH

- Species versus pseudospecies and pseudospecies versus pseudospecies
  - Consensus sequences
    - A consensus sequence is calculated for each ortholog group (residues with the highest occurrence frequency).
    - USEARCH can be used
  - Profile HMMs
    - HMMs are calculated using hmmbuild with default parameters
    - HHSearch to perform profile–profile searches

- Hieranoid reduces the number of required profile–profile searches by performing
  1. Initial sequence–sequence search using consensus sequences to get a list of potential hits.
  2. Profile–profile search between the query and the top hits.

Hieranoid workflow

Building an initial set of homologs

Adapted from InParanoid

- Species versus species
  - replace BLAST by USEARCH

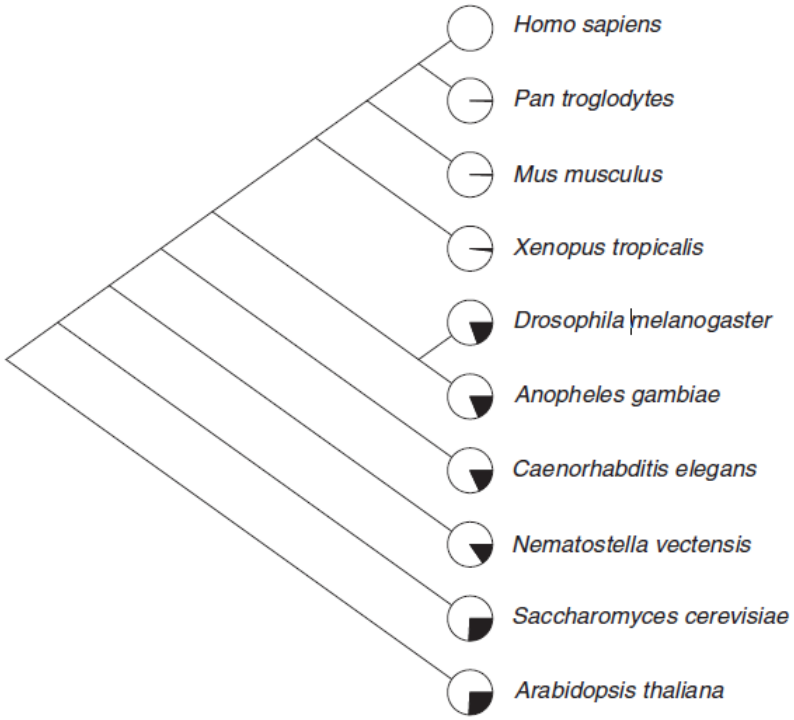
- Species versus pseudospecies and pseudospecies versus pseudospecies
  - Consensus sequences
    - A consensus sequence is calculated for each ortholog group (residues with the highest occurrence frequency).
    - USEARCH can be used
  - Profile HMMs
    - HMMs are calculated using hmmbuild with default parameters
    - HHSearch to perform profile–profile searches

- Hieranoid reduces the number of required profile–profile searches by performing
  1. Initial sequence–sequence search using consensus sequences to get a list of potential hits.
  2. Profile–profile search between the query and the top hits.

Orthology inference

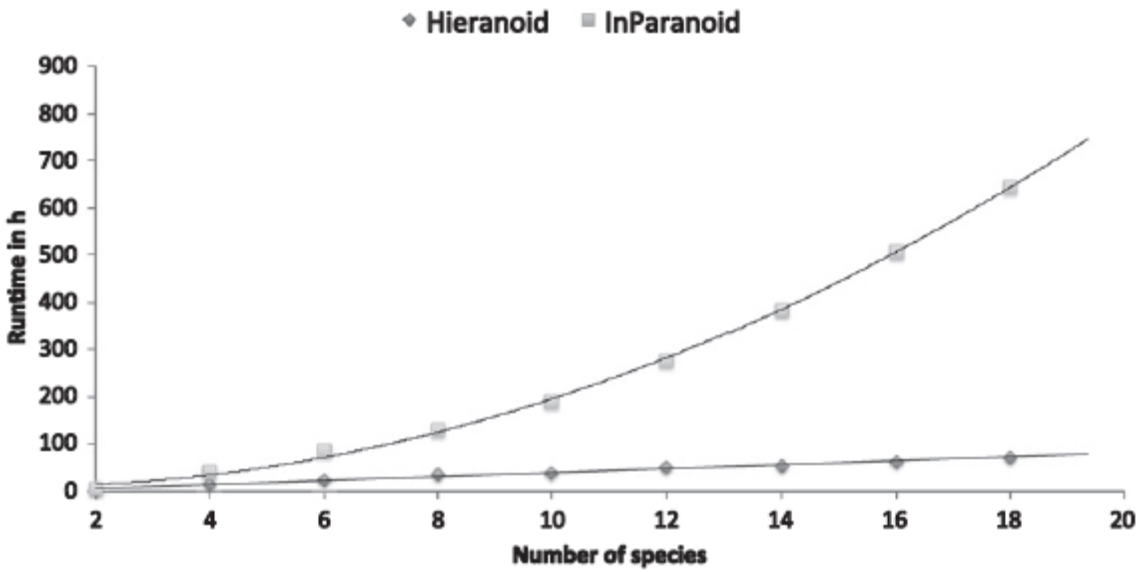
- Once an initial set of putative homologs is built, Hieranoid infers orthologs and inparalogs with InParanoid algorithm.

Comparison to InParanoid



- **Comparison of ortholog inferences from InParanoid and Hieranoid consensus**
-
- The pie charts along the guide tree represent the agreement of inferred orthologs for human versus other species comparisons (fraction of matching pairwise orthology assignments between Hieranoid and InParanoid relative to the union of all their orthology assignments.

Hieranoid versus InParanoid runtime comparison

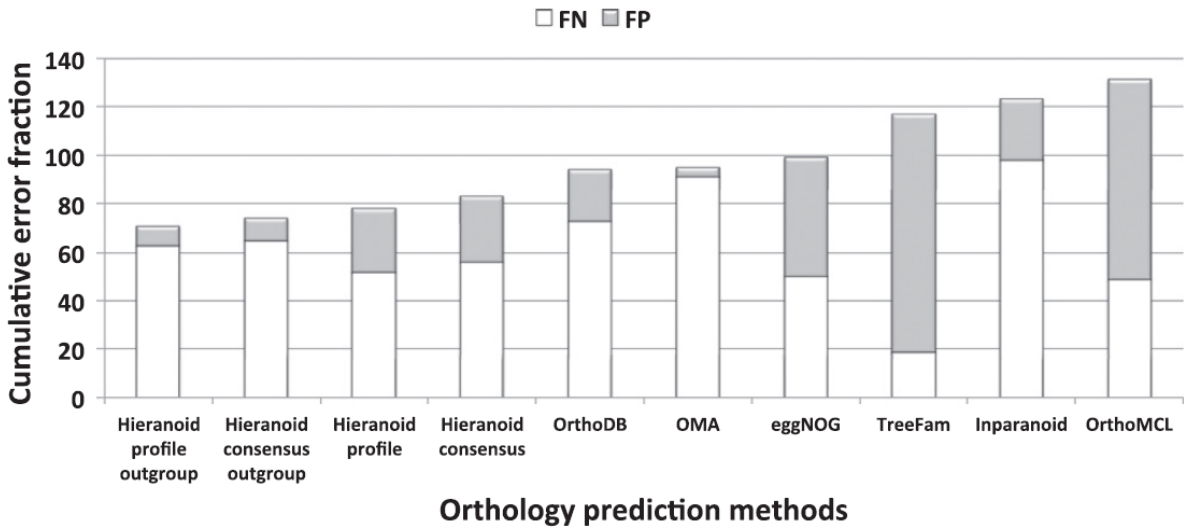


Hieranoid versus InParanoid runtime comparison.

- Hieranoid performs n−1 pairwise comparisons
- InParanoid performs n(n−1)/2 comparisons

(n number of specie)s

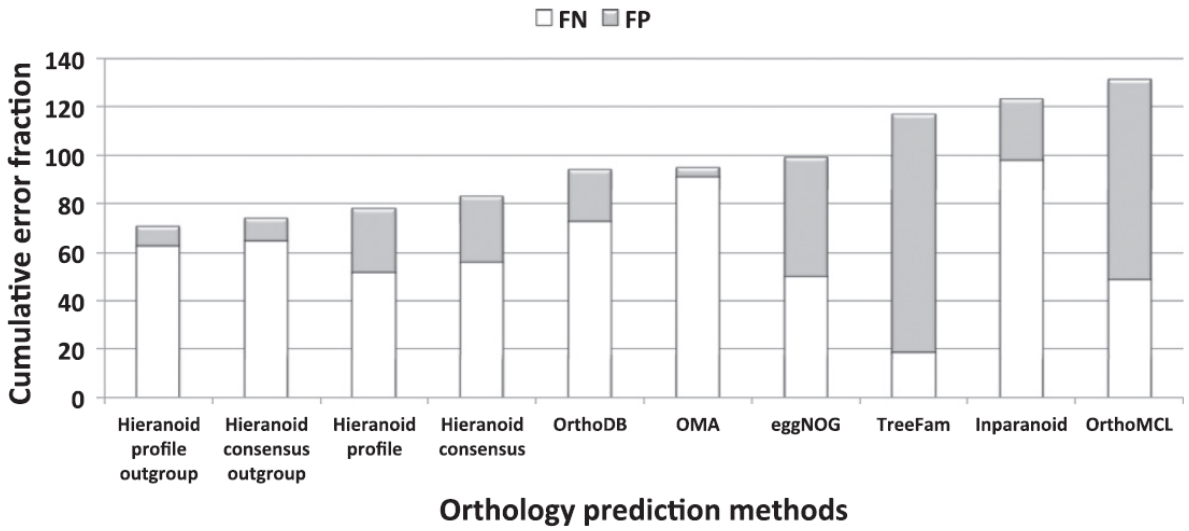
Performance comparison with other orthology inference methods



- We used the following counting scheme:
- For each true pair in an orthobench group, we count how often this pair has not been inferred by one of the methods (**false negative**), given that both sequences were included in the input data of the database.
- For each ortholog group in a database, we count how often a protein pair is inferred as being orthologous, but is not orthologous in the benchmark dataset (**false positive**), given that both sequences are in included in the orthobench input data.

1. one group of methods with a low level of false negatives but a high level of false positives (OrthoMCL, TreeFam)
2. another group with the reverse trend (Hieranoid, InParanoid, OMA, OrthoDB).
3. eggNOG had about equal levels of both types of errors.

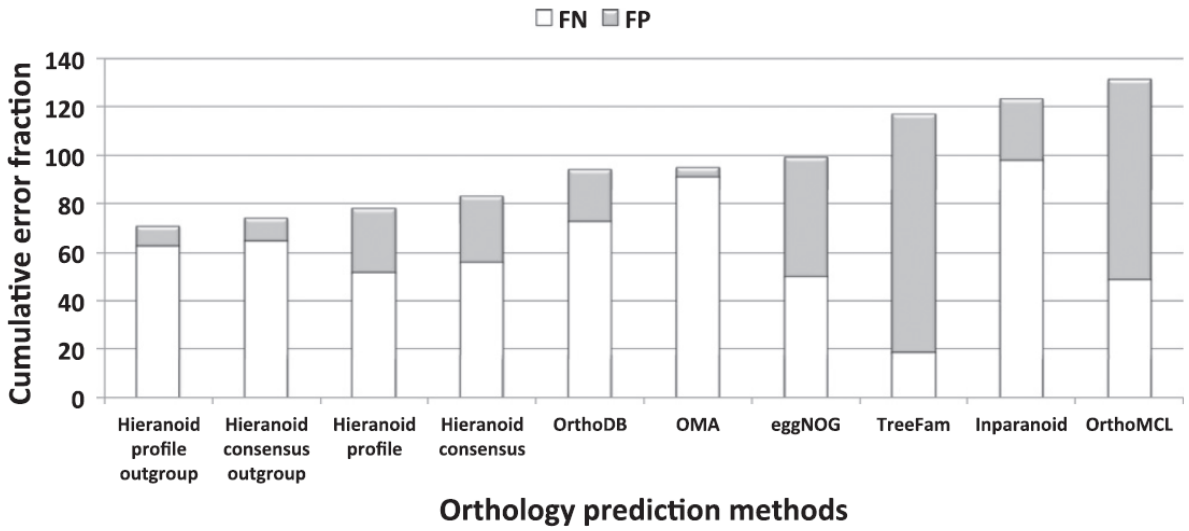
Performance comparison with other orthology inference methods



- Hieranoid, it **misses more** orthology relationships than eggNOG, OrthoMCL, and TreeFam and makes **more false positives** than OMA.
- However, as can be seen from the stacked error bars, Hieranoid shows the **overall lowest error rate.**

- This “hybrid” tree/graph method outperforms other methods that are classical graph-based or tree-based methods.
- a better compromise between these two types of errors.

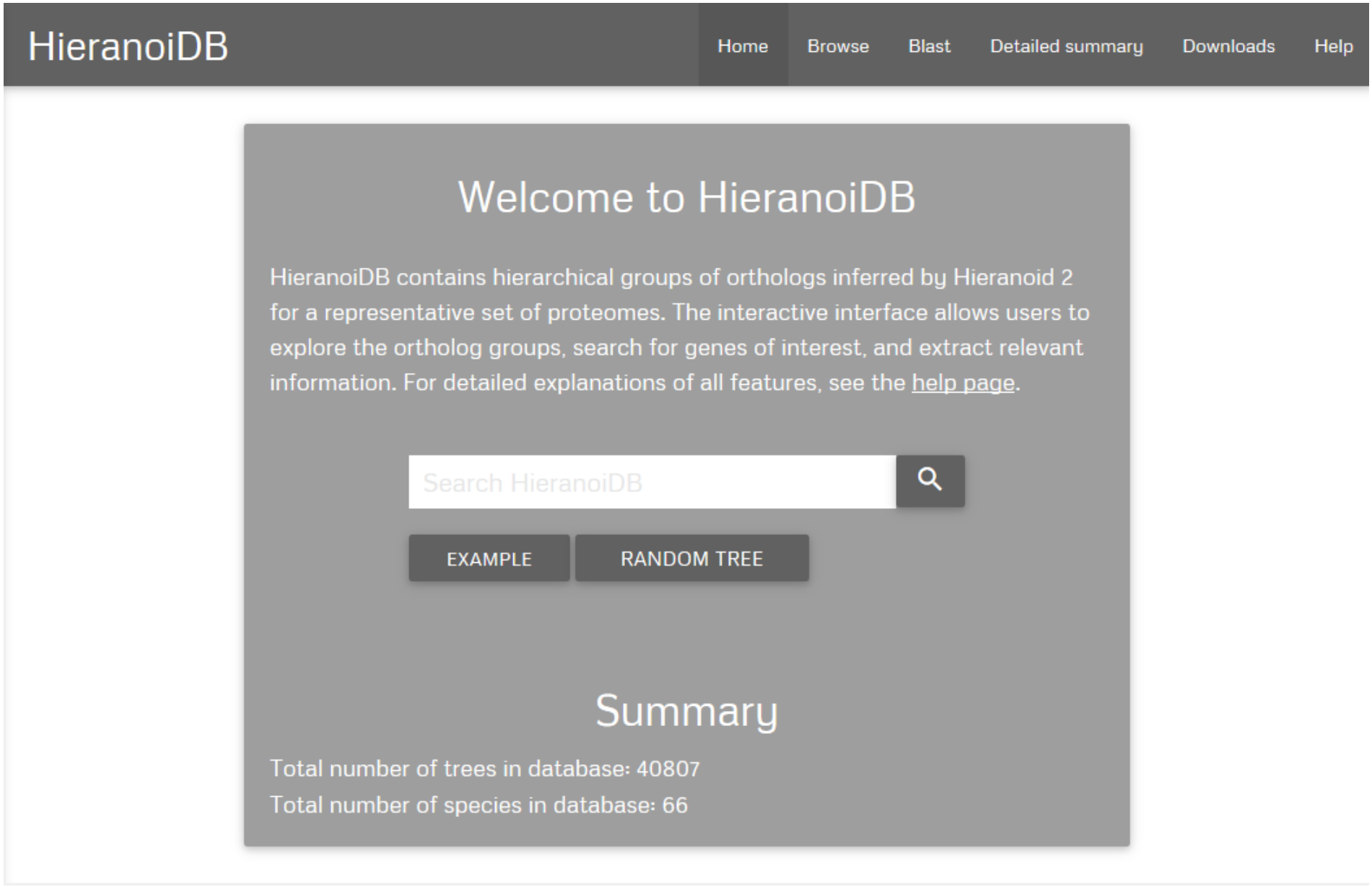
Performance comparison with other orthology inference methods



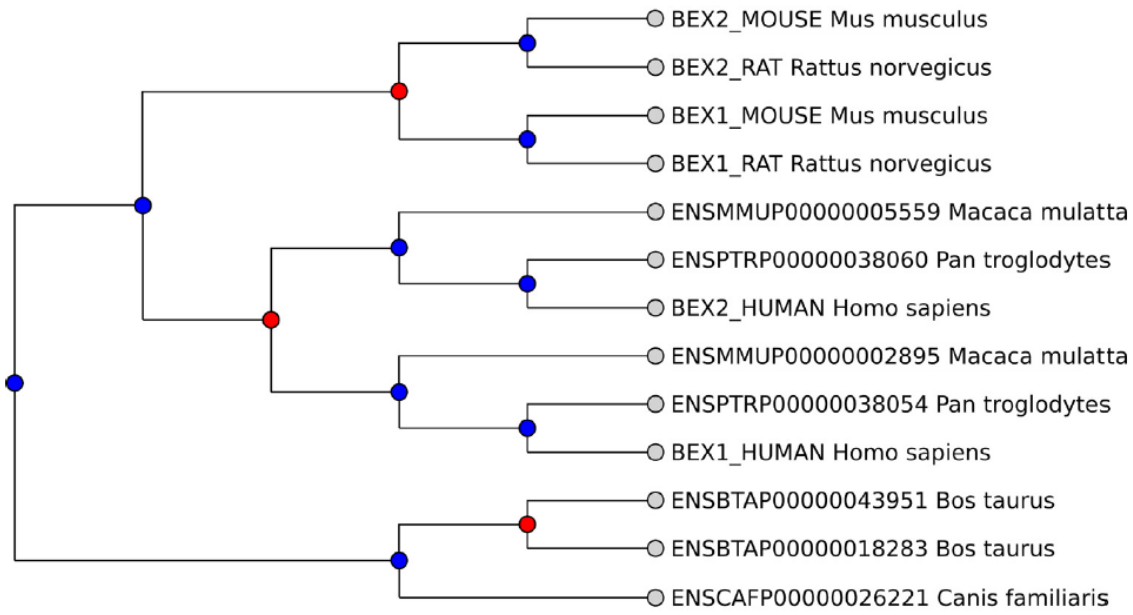
- Hieranoid, it **misses more** orthology relationships than eggNOG, OrthoMCL, and TreeFam and makes **more false positives** than OMA.
- However, as can be seen from the stacked error bars, Hieranoid shows the **overall lowest error rate.**

- This “hybrid” tree/graph method outperforms other methods that are classical graph-based or tree-based methods.
- a better compromise between these two types of errors.

Performance comparison with other orthology inference methods



Performance comparison with other orthology inference methods



Example of HieranoiDB ortholog tree with BEX1 and BEX2 proteins. Blue nodes are speciations and red nodes are duplications.

Methods for Orthology Inference

- Orthology prediction methods can be classified based on the methodology they use to infer orthology into:
    1. **graph-based methods**, which cluster orthologs based on sequence similarity of proteins,
        1. Pairwise species methods
        2. Multi-species graph-based methods
    2. **tree-based methods**, which not only cluster, but also reconcile the protein family tree with a species tree.
        1. Multi-species tree-based methods
    3. Hybrid and other approaches

Methods for Orthology Inference

- *Pairwise species methods* : BHR, InParanoid, RoundUp…:
- Based on these methods, orthologs are **best bi-directional hits** (BBH) between a pair of species.
- **BRH** (Huynen and Bork 1998) is the first automated method and does not detect paralogs.

Pairwise species methods

- Pairwise species methods
- Orthologs are **best bi-directional hits** (BBH) between a pair of species.
  - **BRH** (Huynen and Bork 1998) is the first automated method and does not detect paralogs.
  - **InParanoid** (Remm et al. 2001; Sonnhammer and Östlund 2015) implements an additional step for the detection of paralogs (in-paralogs).
  - **RoundUp** (van der Heijden et al. 2007) uses evolutionary distances instead of BBH.
- These methods are disadvantageous for long evolutionary distances.

Multi-species graph-based methods

- ***Multi-species graph-based methods***
- Due to the fast implementation and high scalability, there are many graph-based methods for multi-species comparisons.
  - All of them use similar sequence-similarity search algorithms.
  - But are quite diverse regarding the clustering algorithms.

- COG (), eggNOG (Powell et al. 2014), and OrthoDB (Waterhouse et al. 2011), share the same methodology: they identify three-way BBHs in three different species and then **merge triangles that share a common side**.

- OrthoMCL (Li et al. 2003), uses a Markov clustering procedure to cluster BBH into OGs.

- OMA (Altenhoff et al. 2011), removes from the initial graph BBHs characterized by **high evolutionary distance**; a concept similar to RoundUp.
  - clustering based on **maximum weight cliques**.
  - **hierarchical groups** (OGs in different taxonomic levels)
  - ‘‘pure orthologs’’ (generate groups of **one-to-one orthologs** without paralogs).

Multi-species tree-based methods

- Multi-species tree-based methods
- Tree-based prediction methods can be separated into approaches that
  - do use **tree reconciliation** EnsemblCompara (Vilella et al. 2009), TreeFam (animal genomes (Ruan et al. 2008)), and PhylomeDB (Huerta-Cepas et al. 2014))
  - do not use **tree reconciliation** (LOFT (van der Heijden et al. 2007)).
- Tree-based methods also initially use homology searches; however, their criteria are more relaxed, as the orthology is resolved through tree topology.

- Although a **reconciled phylogenetic tree** is the most appropriate illustration of orthology/paralogy assignment, there are a few caveats to such an approach, namely their scalability and sensitivity to data quality.

Hybrid and other approaches

- **Hybrid and other approaches**
- Phylogenetic and heuristic approaches can be combined with each other or with synteny information, to yield hybrid approaches that **attempt to overcome the shortcomings** of using either method alone.
- **Ortholuge** uses a phylogenetic approach to refine clusters made by a heuristic algorithm, noting cases where relative gene divergence is atypical between two compared species and an outgroup species and therefore suggests paralogy rather than orthology.
- **EnsemblCompara** further integrates the tree reconciliation and BBH pair-linking approaches by starting with gene trees made from the initial clusters produced by heuristic algorithms, and reconciling these with the species tree.
- **HomoloGene** is another hybrid approach that uses pairwise gene comparisons but follows a guide tree to compare more closely related organisms first, and also adds gene neighborhood conservation.

- Other approaches do not fall into any of the above categories, including a method that uses **topological distance** in a species tree as a factor in a linkage equation to find dense clusters in a multipartite graph (edges are not restricted to BBHs) and a machine-learning predictor of orthology using a set of graph features that, in addition to sequence similarity and synteny, also includes gene co-expression and protein interaction networks.

Assessing the performance of these orthology inference methods

Standardized benchmarking in the quest for orthologs (Altenhoff et al. 2016)

- Because the true **evolutionary history of genes is unknown**, assessing the performance of the orthology inference methods is not straightforward.

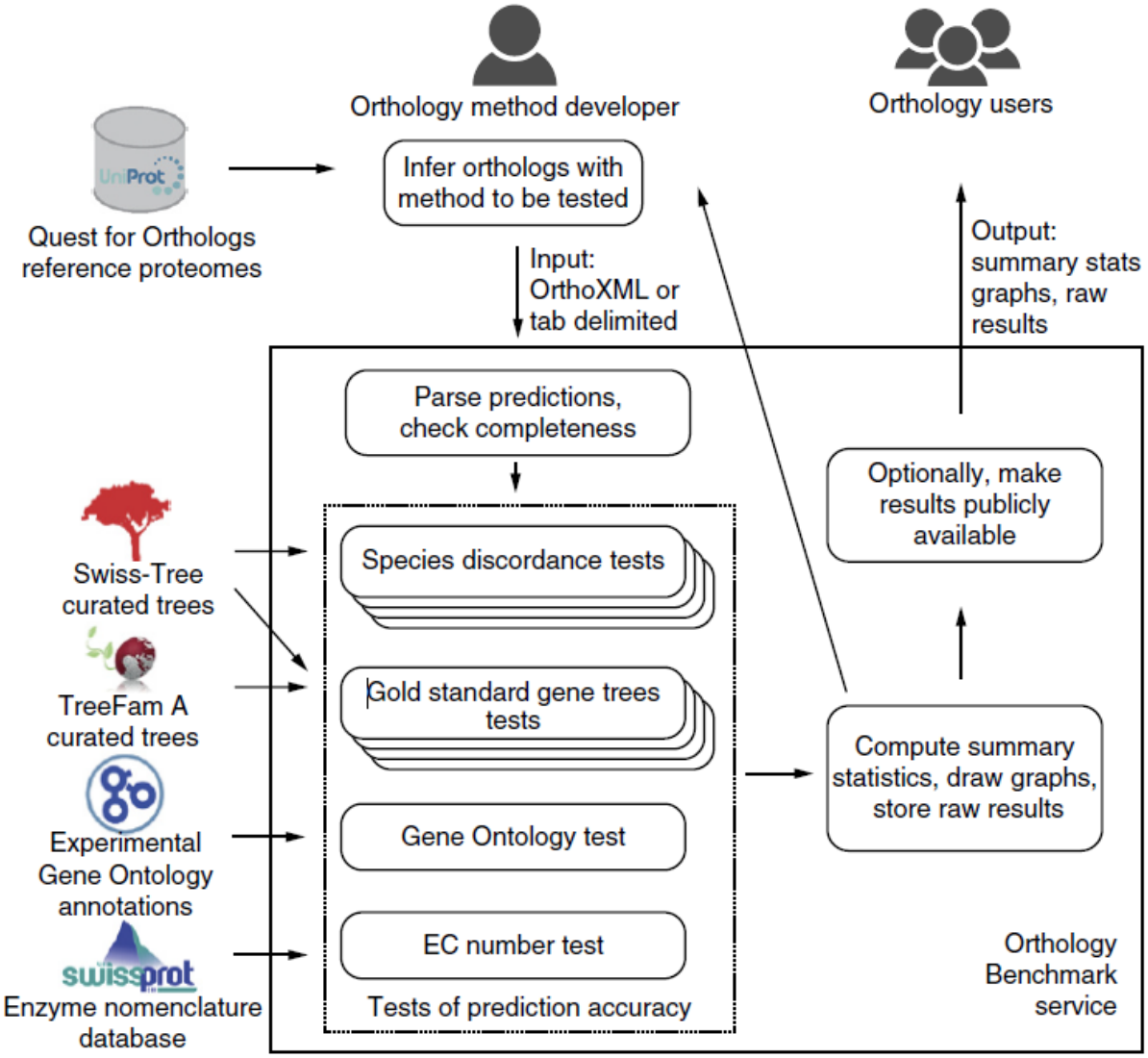
- Several indirect approaches have been proposed.
- **Functional conservation**: used several measures of **functional conservation** (coexpression levels, protein–protein interactions and protein domain conservation) to benchmark orthology inference methods.
- **Consensus among** different orthology methods.
- **Phylogenetic benchmark:** measuring the concordance between gene trees reconstructed from putative orthologs and undisputed species trees.
- **Gold standard**: reference sets, either manually curated or derived from trusted resources
- **Simulation:** simulated genomes to assess orthology inference in the presence of varying amounts of duplication, lateral gene transfer and sequencing artifacts.

Orthology Benchmark service

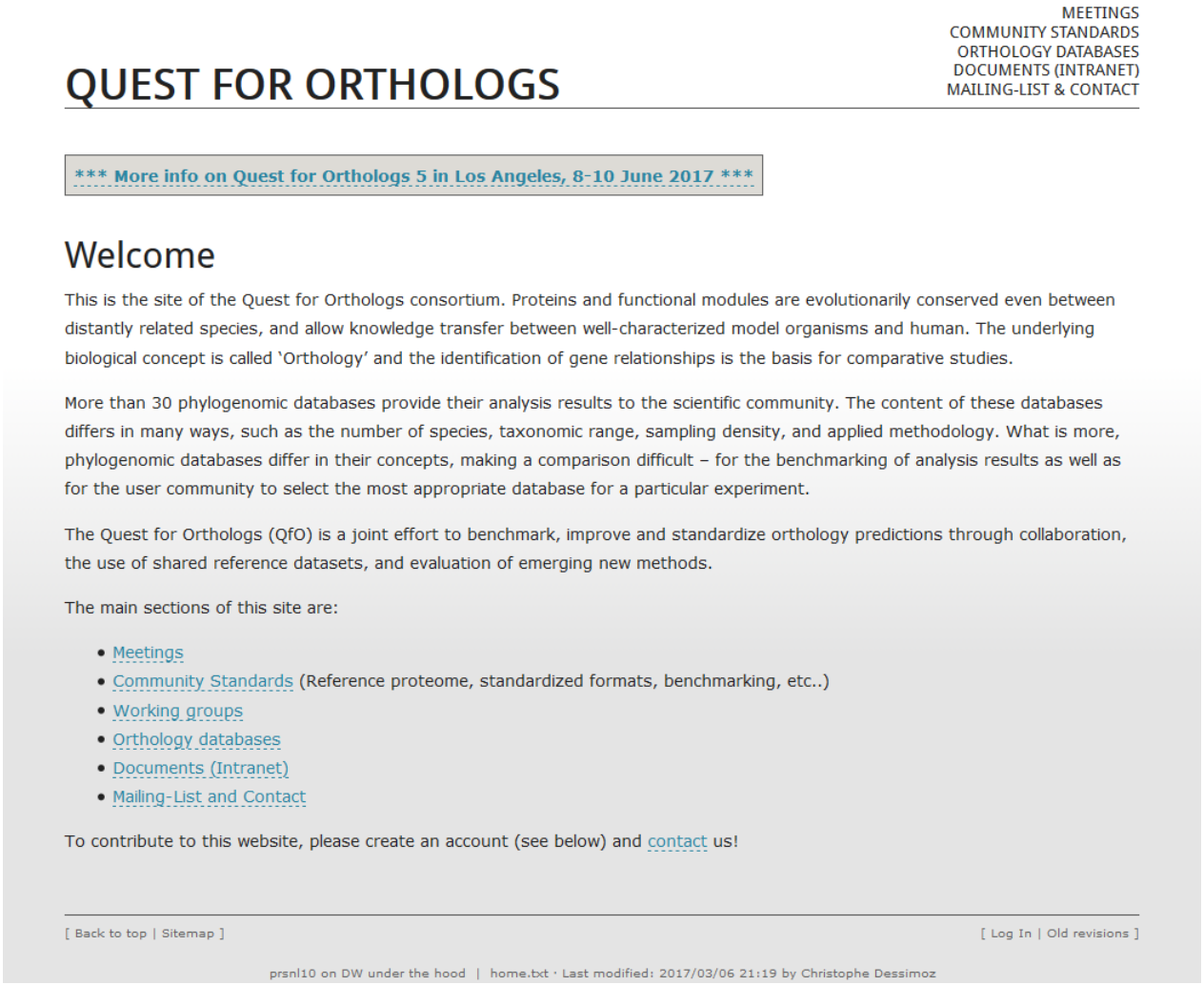
- **Orthology Benchmark service**

- The **Orthology Benchmark service** enables systematic comparison of a new method with state-of-the-art approaches on to a wide range of benchmarks.
- It replaces current practice, which typically includes fewer methods, fewer tests and less empirical data.
- By relying on a common set of data for all methods, the benchmark service ensures that the results obtained by different methods are **directly comparable**.
- The only caveat is that, since proteomes vary in quality and analytical difficulty, the results on the benchmark data set may not entirely reflect the quality of the orthology assignments otherwise provided by each resource.

Orthology Benchmark service



Quest for Orthologs



MEETINGS
COMMUNITY STANDARDS
ORTHOLOGY DATABASES
DOCUMENTS (INTRANET)
MAILING-LIST & CONTACT

QUEST FOR ORTHOLOGS

*** More info on Quest for Orthologs 5 in Los Angeles, 8-10 June 2017 ***

Welcome

This is the site of the Quest for Orthologs consortium. Proteins and functional modules are evolutionarily conserved even between distantly related species, and allow knowledge transfer between well-characterized model organisms and human. The underlying biological concept is called 'Orthology' and the identification of gene relationships is the basis for comparative studies.

More than 30 phylogenomic databases provide their analysis results to the scientific community. The content of these databases differs in many ways, such as the number of species, taxonomic range, sampling density, and applied methodology. What is more, phylogenomic databases differ in their concepts, making a comparison difficult – for the benchmarking of analysis results as well as for the user community to select the most appropriate database for a particular experiment.

The Quest for Orthologs (QfO) is a joint effort to benchmark, improve and standardize orthology predictions through collaboration, the use of shared reference datasets, and evaluation of emerging new methods.

The main sections of this site are:

- Meetings
- Community Standards (Reference proteome, standardized formats, benchmarking, etc..)
- Working groups
- Orthology databases
- Documents (Intranet)
- Mailing-List and Contact

To contribute to this website, please create an account (see below) and contact us!

[Back to top | Sitemap] [Log In | Old revisions]

prsnl10 on DW under the hood | home.txt · Last modified: 2017/03/06 21:19 by Christophe Dessimoz

- The **Quest for Orthologs** (QfO) is a joint effort to benchmark, improve and standardize orthology predictions through collaboration, the use of shared reference datasets, and evaluation of emerging new methods.