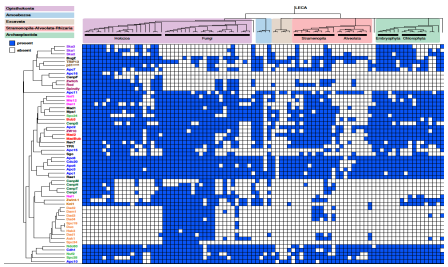
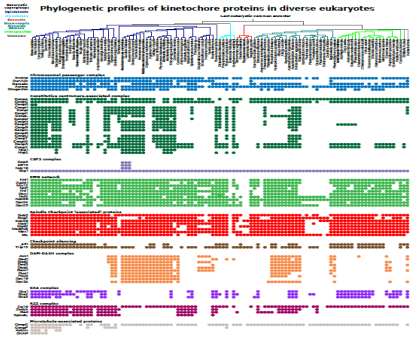
"Large Scale"/Automatic Orthology (& gene family) Inference

- Introduction
- Orthology between two species
  - Bidirectional Best Hits (BBH)
  - Inparanoid
- Networks/graphs of (BBH) blast hits
  - Families & MCL,
  - OrthoMCL,
  - COG (EggNOG)
- Full phylogenomics pipelines
  - First collect families (blast networks)
  - Non-strict reconciliation
    - notung
    - Compara / treebest (also graph of blast hits)
- Final thoughts

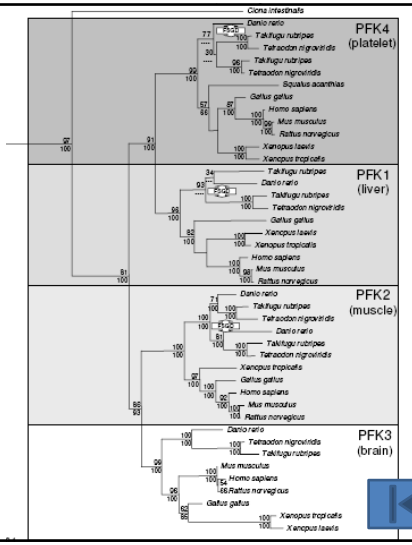
- I've talked about the importance of e.g. gene duplication and gene loss for genome evolution and there is a lot of evidence for this from studying individual gene families (NB a lot of individual gene families have been studied!!!)
- However we /also want to qauntify these patterns look for trends etc. Hence also do it on a large scale
- **Need** for automatic orthology,
- … but remains an *unsolved problem*

Presence/absence of kinetochore subunits across species = orthologs! Revealed complex ancestor and indepdent loss: Can we do this for all complexes/pathways?





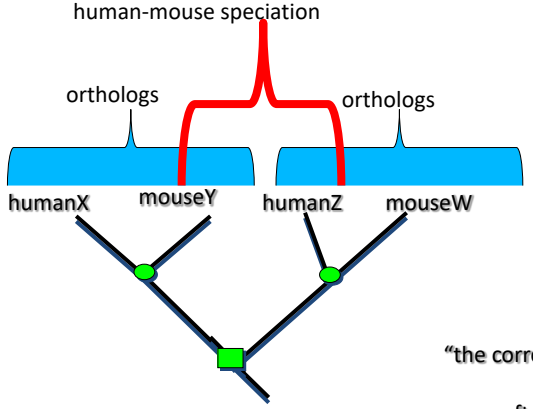
Gene duplications at the base of vertebrates which genes have maintained in duplo, triplet or quadruplet relative to invertebrates? = orthology



Automatic methods for Orthology between two species, bidirectional best hits & inparanoid

- Oldest automatic methods
- Still used
- Illustrate how a method from a set of blast hits is used to infer evolutionary history (i.e. a phylogenetic tree)

NOT orthologous relative to the human-mouse speciation



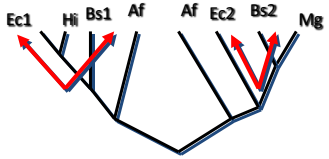
Fitch 1970
Two genes in two species are orthologous if they derive from one gene in their *last* common ancestor; **This means that the "node"' in the tree where they diverge is a speciation node.**

"the corresponding gene"

often have the same function

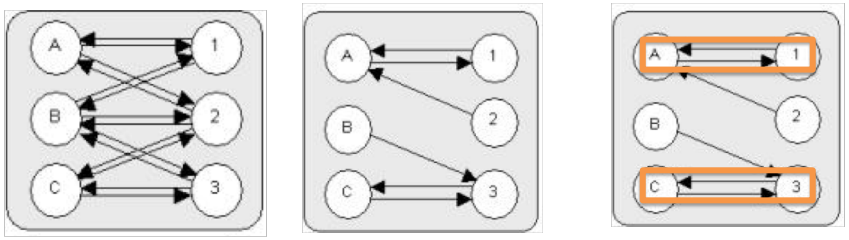
Genes can diverge by
- Speciation, or
- Duplication

Bidirectional Best Hits (BBH) / Reciprocal Best Hits (RBH)



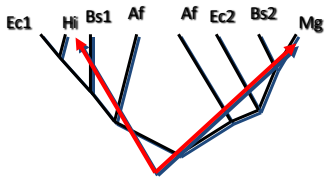
Extracting tree-like information from pairwise similarities

```
Ec1 Bs1  50%
Ec1 Bs2  35%
Ec2 Bs1  33%
Ec2 Bs2  48%
```



http://web.mit.edu/manoli/www/thesis/Chapter1.html

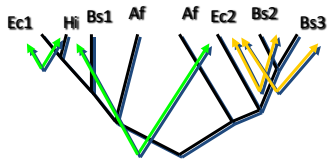
BBH issues A: differential gene loss (False Positives)



The more genomes, the more frequent this becomes …

Mg Hi 35%

BBH issues A: ignores inparalogs (False negative's)

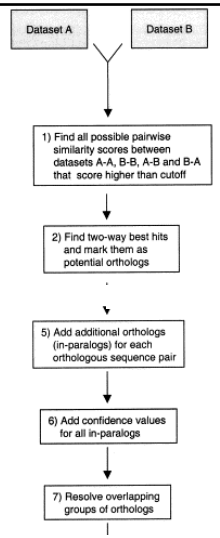


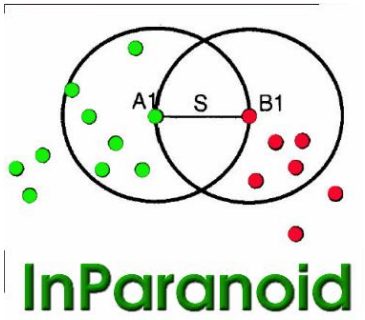
Ec1 Hi 70% Ec2 Bs2 48%
Ec2 Hi 38% Ec2 Bs3 51%

(Bs2 Bs3 70%)

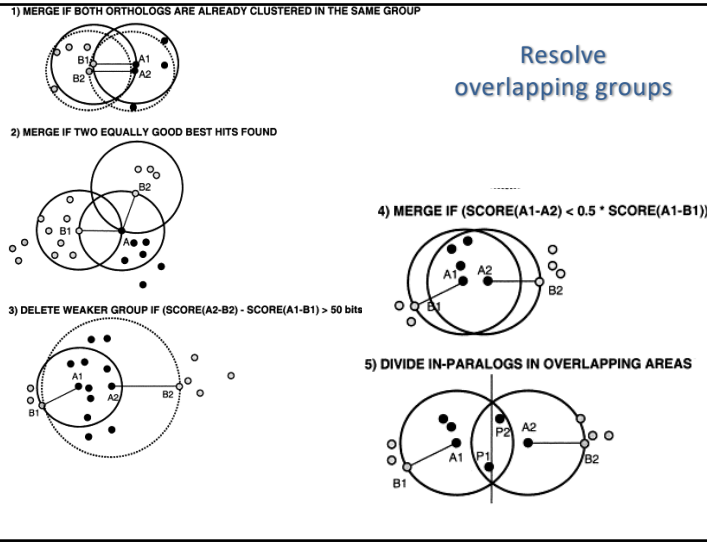
Prevalence? Depends on e.g. evo distance, group vs pairwise orthology
At least 16% prokaryotes, Much higher in eukaryotes

(reason for development of INPARANOID)





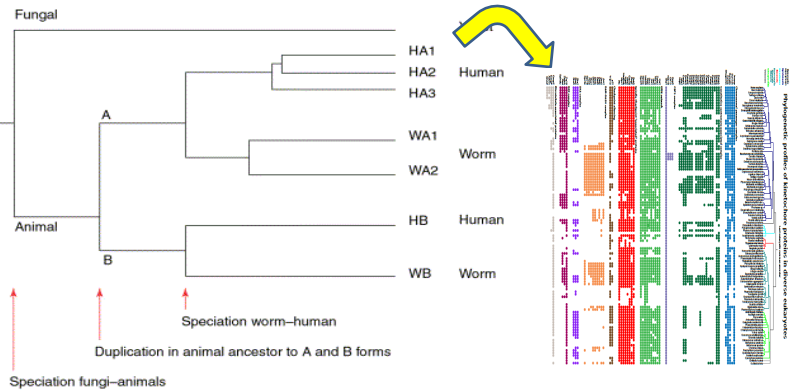
Large scale blast, relatively strict score cut-off &
Overlap criteria > 50%



Resolve overlapping groups

Orthologous groups from homology/blast networks/graphs

Orthology is defined between pairs of species, but for many questions you think about a set of species, i.e. what to put in the excel-sheet .

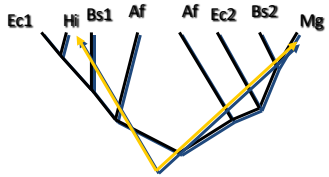


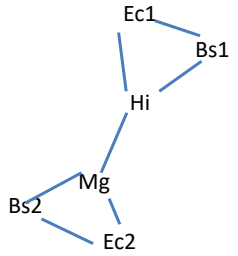
Orthologous groups

- Work around to the non-transitivity of the concept of orthology is: “Group orthology”
- Conceptually: all proteins that are directly descended from one protein in the last common ancestor of all species in the set are considered orthologous to each other (i.e. includes inparalogs relative to this potentially quite ancient speciation)

b

Another reason to go >2 species in analyzing blast graphs is to solve the BBH issue differential loss

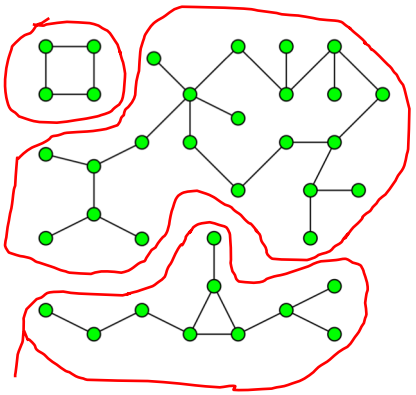




i.e. Mg and Hi do not share a BBH hit!
Making their orthology less likely …

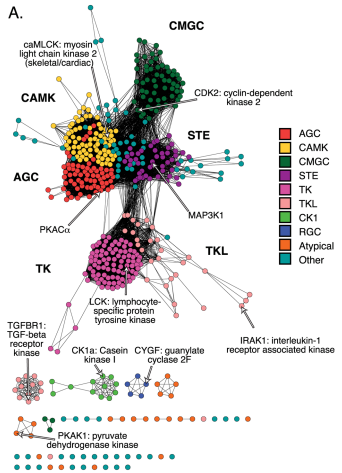
From a network to families *or* orthologous groups

- Orthology is a specification of "the kind of homology", so as a first step generate homologs and then subdivide them into orthologs (via e.g. trees)?
- Automatically generating orthologs: first automatically generate gene families to make trees
- Homology is transitive, so when creating families for generating automatically trees or for phylogenetic profiles, you can just link them up by defining **connected components?**



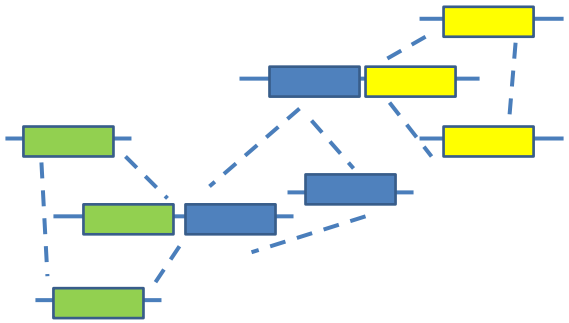
https://en.wikipedia.org/wiki/Connected_component_(graph_theory)

Representing blast hits or BBH's as graphs/networks



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004345

"problem type 1": fusion/fission

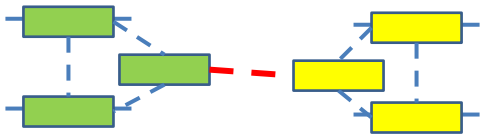


How to solve fusion/fission?

- Disallow "fusion proteins" to bring in new stuff (somehow)(but how do you detect fusion proteins?)
- Filter hits on spanning e.g. >70% of length query (and/or target).

- Work on restricted taxon sets (e.g. ENSEMBL COMPARA, oomycetes)
- Look at fusion cases by hand (COGs)

Problem type 2: “false positive links”

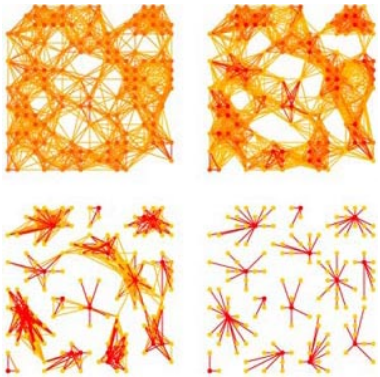


- In single linkage a few (random) FP links snowball and connect
- Sources of FP links:
  - false positives FP’s statistics/e-value true but ~“multiple testing” (blast E-values are not exact but heuristics) a.k.a. bad luck
  - “convergent signal” Disorder, coiled coil, TM
  - Low complexity

Solution “false positive links”

- Very conservative e-values
- Filter low complexity
- Take low complexity into e-value into account (modern blast)
- Filter coiled / coil (infrequent)
- Filter disorder (never seen done).
- Work at restricted taxon sets (e.g. ensembl COMPARA, oomycetes)

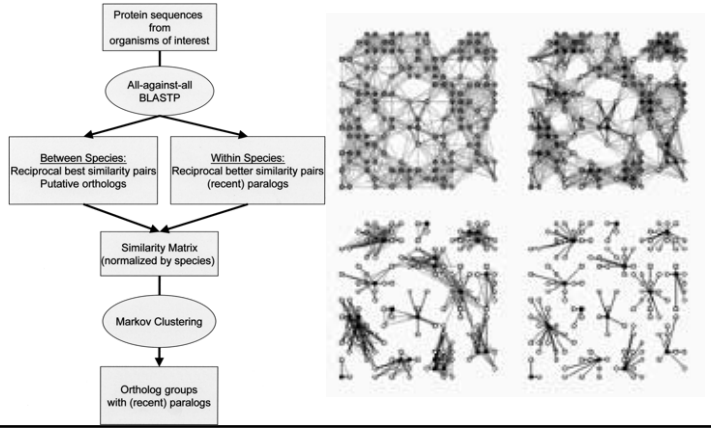
An often used solution (also for orthologs) to create families from blast-graphs: MCL



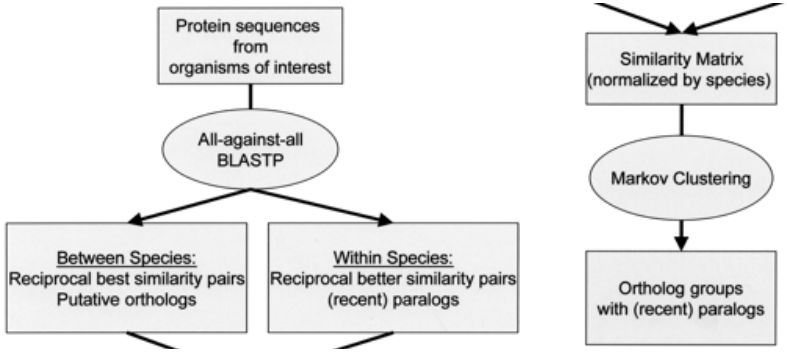
MCL **Markov Cluster algorithm**

- Simulate many random walks (or flow) within the whole graph,
- strengthen flow where it is already strong, and weaken it where it is weak.
- By repeating the process an underlying cluster structure will gradually become visible.
- Yields a number of regions with strong internal flow (clusters), separated by ‘dry’ boundaries with hardly any flow.
- Inflation parameter. higher inflation parameter leads to higher granularity
- **So the idea is that this removes e.g. “false edges” and ~forces a fusion protein to go one or the other side.**

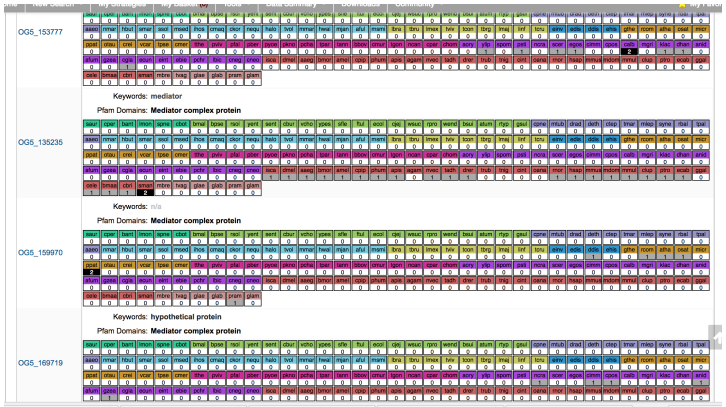
Ortho MCL: overview



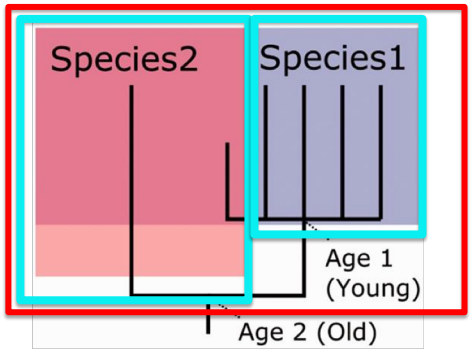
Ortho MCL: how the graph is seeded



OrthoMCL Starts from blast so diverged / short orthologs are difficult (preview to COO)



Another problem for orthoMCL (and other network based methods) oversplitting"



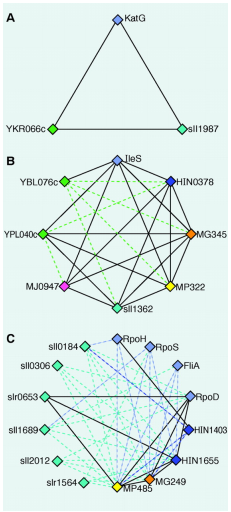
two incomplete orthologous groups

correct evolutionary orthologous group

https://academic.oup.com/gbe/article/8/6/1812/2574026

Graph based orthology: COG

- 1. Perform the all-against-all protein sequence comparison.
- 2. Detect and collapse obvious paralogs, that is, proteins from the same genome that are more similar to each other than to any proteins from other species.
- 3. Detect triangles of mutually consistent, genome-specific best hit (BeTs), taking into account the paralogous groups detected at step 2. This approach is most likely to be informative when the BeTs forming a triangle come from widely different lineages, i.e. demands on a triangle.
- 4. Merge triangles with a common side to form COGs.



COG, the final two steps: manual curation for fusion

5. A case-by-case analysis of each COG. This analysis serves to eliminate false-positives and to identify groups that contain ***multidomain proteins*** by examining the pictorial representation of the BLAST search outputs. The sequences of detected multidomain proteins are split into single-domain segments and steps 1–4 are repeated with these sequences *(iterative!),* which results in the assignment of individual domains to COGs in accordance with their distinct evolutionary affinities.

COG, the final two steps: manual curation for "missed" differential loss or other complications

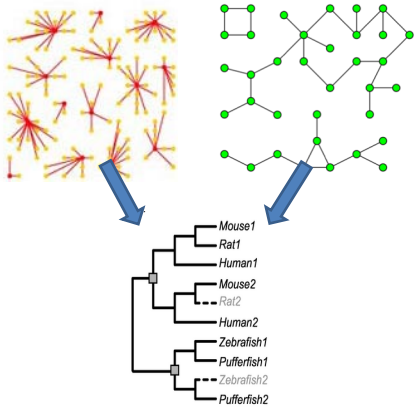
6. Examination of large COGs that include multiple members from all or several of the genomes using **phylogenetic trees, cluster analysis and visual inspection of alignments**; as a result, some of these groups are split into two or more smaller ones that are included in the final set of COGs.

The manual curation of COGs also allowed each COG to be annotated with a function

```
COG0001   H    Glutamate-1-semialdehyde aminotransferase
COG0002   E    N-acetyl-gamma-glutamylphosphate reductase
COG0003   P    Anion-transporting ATPase, ArsA/GET3 family
COG0004   P    Ammonia channel protein AmtB
COG0005   F    Purine nucleoside phosphorylase
COG0006   E    Xaa-Pro aminopeptidase
COG0007   H    Uroporphyrinogen-III methylase
```

Full phylogenomics pipelines: define families & automatic tree reconciliation

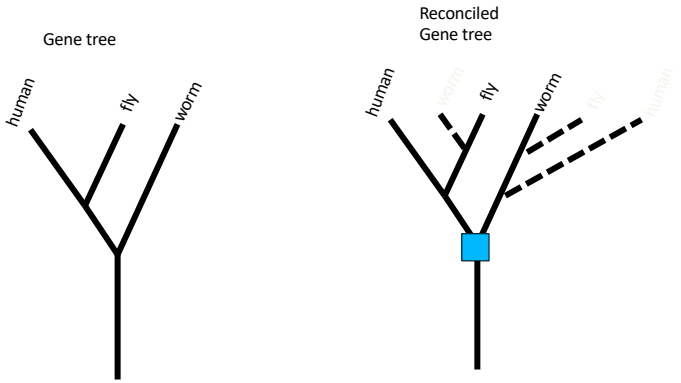
- First a graph of homologs and then either via **MCL** or via "**connected component"** have "families"
- Then make trees of these families and do automatic **tree reconciliation**



Methods to go from trees to orthologs (automatic tree reconciliation and species tree aware gene tree reconstruction)

- First: methods that were strict (see next slides): http://pbil.univ-lyon1.fr/software/RAP/RAP.htm (Phylogenetic Tree Reconciler (Réconciliateur d'Arbres Phylogénétiques))
- Currently: programs that take uncertainty into account and also weigh the amount of "genome evolution" that a topology implies NOTUNG, TREEBEST, SYNERGY, TREEFIX

The problem of too strict tree reconciliation (as implemented in naïve first generation software)



NOTUNG

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 3/4, 2000
Mary Ann Liebert, Inc.
Pp. 429–447

NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees

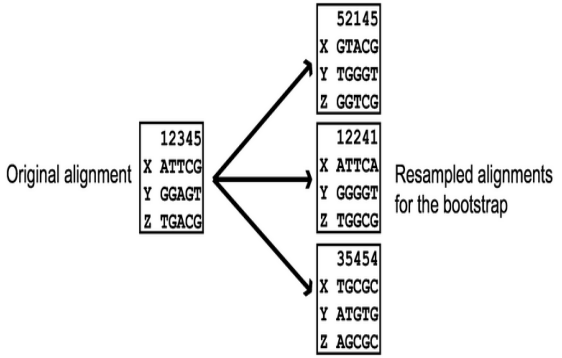
KEVIN CHEN,[1] DANNIE DURAND,[2] and MARTIN FARACH-COLTON[3]

- Use bootstrap ensemble to find clustering that is more consistent with species tree and a minimum bootstrap value above which a clustering cannot be overridden:

How to assess confidence in tree

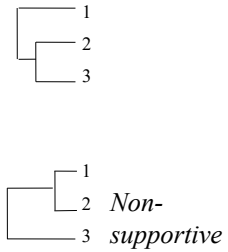
- bootstrap:
  - Select multiple alignment columns with replacement
  - Recalculate tree
  - Compare branches with original tree
  - Repeat 100-1000 times, so calculate 100-1000 different trees
  - How often is branching preserved for each internal node?
  - Uses samples of the data

The Bootstrap

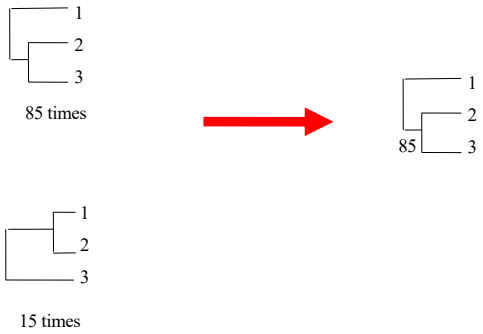


The Bootstrap

	1	2	3	4	5	6	7	8
Original	C	C	V	K	V	I	Y	S
	M	A	V	R	L	I	F	S
	M	C	L	R	L	L	F	T
			2x	1x		3x		2x
Resampled	3	4	3	8	6	6	8	6
	V	K	V	S	I	I	S	I
	V	R	V	S	I	I	S	I
	L	R	L	T	L	L	T	L



The Bootstrap



NOTUNG

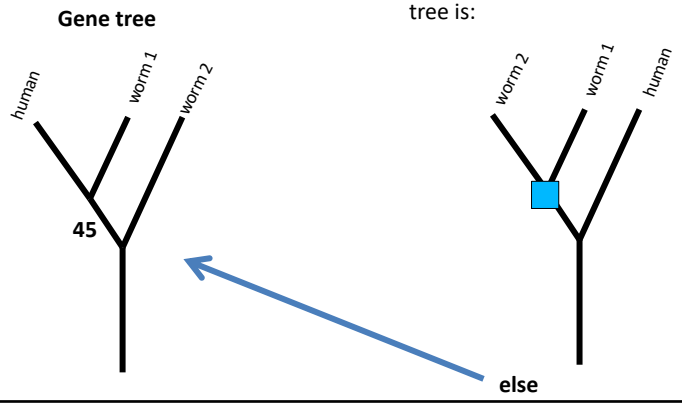
JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 3/4, 2000
Mary Ann Liebert, Inc.
Pp. 429–447

NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees

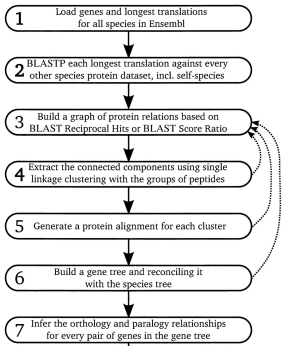
KEVIN CHEN,[1] DANNIE DURAND,[2] and MARTIN FARACH-COLTON[3]

- Use bootstrap ensemble to find clustering that is more consistent with species tree and a minimum bootstrap value above which a clustering cannot be overridden:

If clustering worm1, human has bootstrap < parameter (e.g. 75), **then** the reconciled tree is:

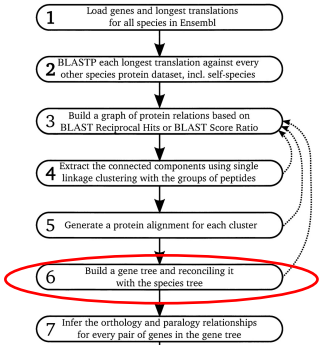


ENSEMBL COMPERA (including treebest)



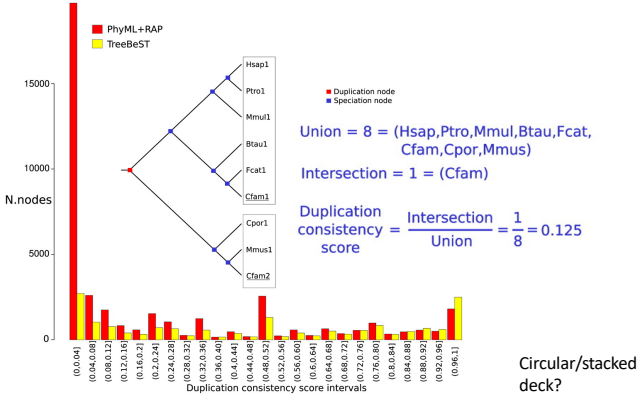
Genome Res. 2009 Feb;19(2):327-35. Epub 2008 Nov 24.
EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E.

Tree reconciliation: treebest



1. Merge several input trees into one tree by minimizing number of duplications and losses (neighbour-joining synonymous distance (dS) tree, NJ non-synonymous distance (dN), NJ p-distance, max-likelihood tree under the WAG model and ML under the HKY model.)
2. calculate the probability of a gene tree in the context of species evolution and multiplies this with the probability of sequence evolution. PhyML typed search is then applied to search for the max-likelihood tree.

Species overlap benchmark



Union = 8 = (Hsap,Ptro,Mmul,Btau,Fcat, Cfam,Cpor,Mmus)

Intersection = 1 = (Cfam)

$$\text{Duplication consistency score} = \frac{\text{Intersection}}{\text{Union}} = \frac{1}{8} = 0.125$$

Circular/stacked deck?

Some final points

- Automatic tree reconciliation is nice, but of which sequences are you making trees? →back to graph based methods?
- Choice of orthology should depend on question
- (parallel) HGT? (serial) Endosymbiosis?
- Insufficient use of e.g. profile searches or knowledge e.g. PFAM (too many methods start at blast -> too many false negatives (?) e.g. orthoMCL med11)
- Some of kind of manual curation is perhaps inevatible
- Some kind of levels of orthology is needed. (should you start at the top or at the bottom?)