MCB 5472
Lecture 3 Feb 10/14

(1) Types of homology
(2) BLAST

Homology references

"Homology a personal view on some of the problems"
Fitch WM (2000) Trends Genet. 16: 227-231

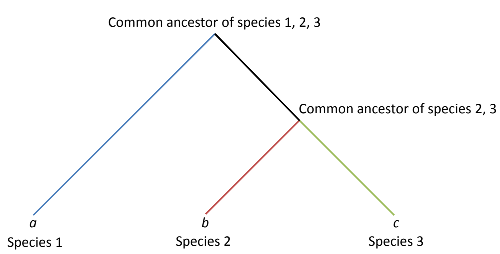
"Orthologs, paralogs, and evolutionary genomics"
Koonin EV (2005) Annu. Rev. Genet. 39: 309-338

What is homology?

- Owen 1843: "the same organ in different animals under every variety of form and function"
- Huxley (post Darwin): homology evidence of evolution
  - Similarity is due to descent from a common ancestor

What is homology?

- Homology is a statement about shared ancestry
  - Two things either share a common ancestor (are homologous) or do not



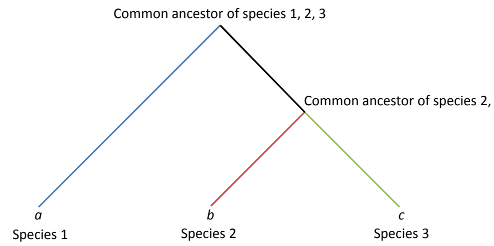
These are all homologs (common ancestor)

Ohno 1970: "Evolution by Gene Duplication"

- New genes arise by gene duplication
  - One copy retains ancestral function
  - Other copy diverges functionally
- "Homolog" as a single term therefore is a sloppy fit
  - What kind of ancestor to homologs share?

Fitch 1970: "Orthologs" and "Paralogs"

- "Orthologs": genes related by vertical descent



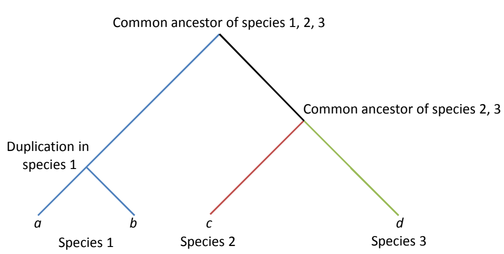
These are all homologs (common ancestor)
These are all orthologs (vertical descent)

Homology and Function

- Homology and function are two different concepts
- Strict orthology and functional conservation often correlate but this is not absolute
- Basis for annotating genomes based on similarity to previous work

Fitch 1970: "Orthologs" and "Paralogs"

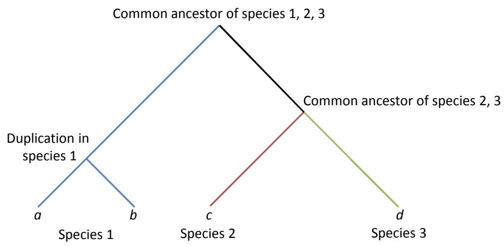
- "Orthologs": genes related by vertical descent
- "Paralogs": gene related by gene duplication



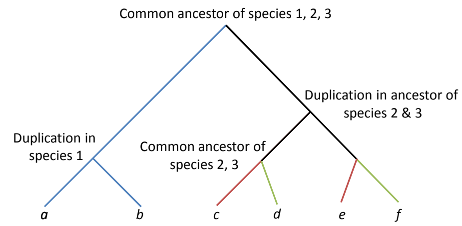
Genes *a*, *b*, *c* & *d* are homologs (common ancestor)
Genes *a* & *b* are paralogs (related by duplication)

Orthology/paralogy is somewhat relative

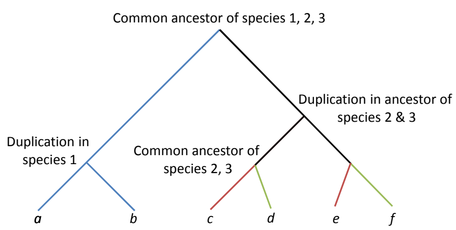
- Depends on the depth of duplication relative to common ancestry
- "Co-orthologs": paralogs formed in a lineage after speciation, relative to other lineages (Koonin 2005)



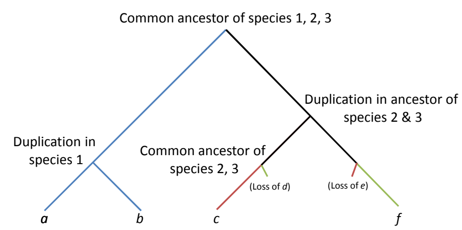
Genes *a* & *b* are paralogs (related by duplication)
Genes *a* & *b* are co-orthologs of genes *c* & *d*
(duplication followed speciation)



Genes *a* & *b* are paralogs (duplication)
Genes *a* & *b* are co-orthologs of *c*, *d*, *e* & *f*
(duplication followed speciation)



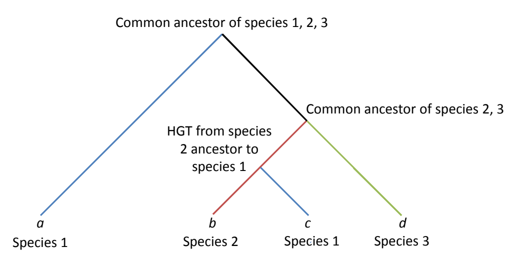
Genes *c* & *d* are orthologs (common ancestor)
Genes *e* & *f* are orthologs (common ancestor)
Genes *c* & *d* are paralogs of genes *e* & *f*
(duplication preceded speciation)



Genes *c* is a paralog of gene *f* even though it doesn't seem so
(duplication still preceded speciation followed by extinction)

Xenologs

- Bacteria exchange DNA between distant relatives by horizontal gene transfer (HGT)
  - Increasingly recognized in eukaroytes too
- Gene tree does not match species tree



Gene *c* is a xenolog relative to the others

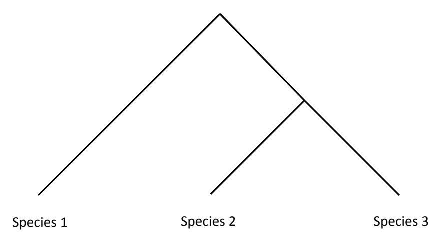
Other "-logs"

- Inparalogs: duplication follows speciation
- Outparalogs: duplication precedes speciation
- Synlogs: arising from organism fusion

- Orthology & paralogy can get quite complicated when multiple duplications happened at different moments in time
- Gene loss & HGT can always confound – one often has to rely on external evidence to recreate speciation
  - E.g., other genes not thought to be horizontally transferred, average signal of multiple genes

---

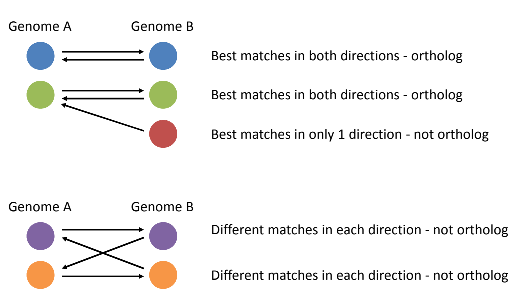
Discuss: how are these genes related to each other? Three possibilities



How to determine orthologs

- Most detailed: phylogenetic trees
  - Can be computationally expensive
- Reciprocal BLAST hit (RBH/BBH)
  - Simplest, computationally cheap, less accurate & more complicated with many genomes
- More complicated RBH clustering
  - OrthoMCL, Inparanoid

RBH orthologs



BLAST

- Standard method to identify homologous sequences
  - Not for comparing two sequences directly; use NEEDLE instead for this (global vs. local alignment methods)
- Requires database to query sequence against
- Probably the most common scientific experiment

Different BLAST types

- BLASTn: nucleotide vs nucleotide
- BLASTp: protein vs protein
- BLASTx: protein vs translated nucleotide
- tBLASTn: translated nucleotide vs protein
- tBLASTx: translated nucleotide vs translated nucleotide

- Nucleotides translated in all six open reading frames

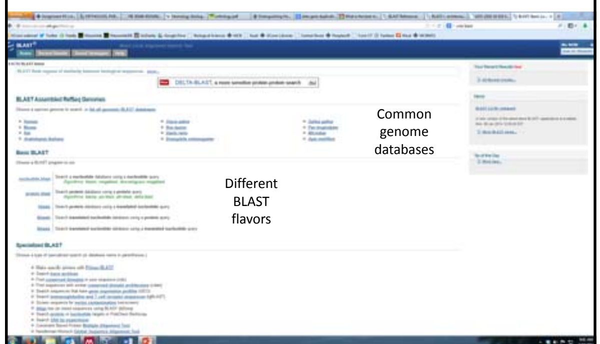
Implimentations

- `blastall`: older command line version
  - Atschul et al. 1990 J. Mol. Biol. 215:403-410
- `BLAST+:` newer command line version
  - Camacho et al. 2008 BMC Bioinformatics 10:421
  - Faster than `blastall`
- Web BLAST:
  - www.blast.ncbi.nlm.nih.gov/Blast.cgi
  - Web version of BLAST+

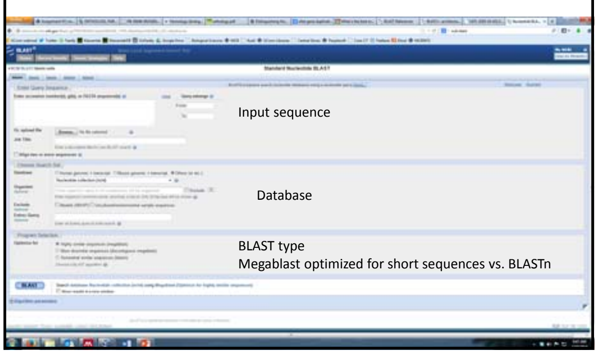
Databases

- All BLAST queries are done vs. a database
- Examples:
  - NCBI's "nr" queries against all of GenBank
  - WebBLAST has preformatted databases for different taxonomic groups, other NCBI divisions (e.g., Refseq, Genomes)
- Command line allows custom databases
  - e.g., lab genomes

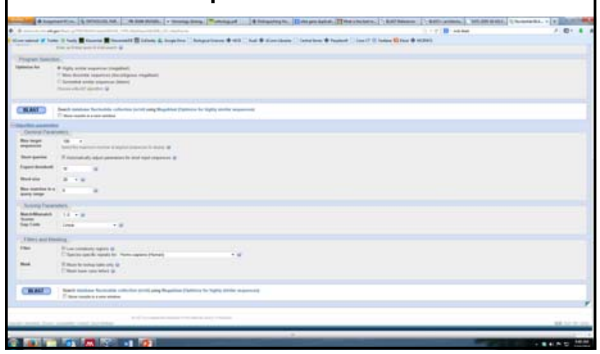
WebBLAST



WebBLAST (BLASTn)

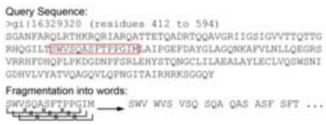


WebBLAST (BLASTn) parameters



BLAST: Step 1

- Break sequence into words
  - Protein: 2-3 amino acids
  - Nucleotide: 16-256 nucleotides



- Goal: exact word matches
  - Computational speedup

http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1001014

Substitution matrices

- Evolutionarily, some substitutions are more common than others
  - Some amino acids are common (e.g., Leu) and some are rare (e.g., Trp)
  - Some substitutions are more feasible than others (e.g., Leu -> Ile vs. Leu -> Arg)
- Substitution matrices therefore weight alignments by these probabilities

BLOSUM matrices

- Alignments of a set of divergent reference sequences
  - BLOSUM62: sequences 62% identical
  - BLOSUM80: sequences 80% identical
- Substitution frequency calculated for each reference set and used to derive substitution matrix
- Henikoff & Henikoff (1992) PNAS 89:10915-10919
- Also: M. Dayhoff's PAM matrices from 1978

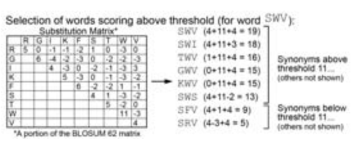
BLOSUM62 matrix

http://upload.wikimedia.org/wikipedia/commons/5/52/BLOSUM62.gif

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLAST: Step 2

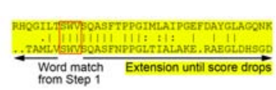
- Use substitution matrix to find synonymous words about some scoring threshold



http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1001014

BLAST: Step 3

- Find matching words in the database
- Extend word matches between query and matching sequence in both directions until extension score drops below threshold
  - First without gaps



http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1001014

BLAST: Step 4

- If initial alignment good enough, redo with gaps and calculate statistics

http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1001014

BLAST score

$$S = \left(\sum M_{ij}\right) - cO - dG$$

Score: S; Sum of scores from distance matrix: $\sum M_{ij}$; Penalty for opening gap: c; # of gaps: O; Per-residue for gap extension: d; Total length of gaps: G

Gap opening penalty typically significantly larger than gap extension penalty

Why?

Questions:

1. Why do gap opening and extension penalties differ?
2. Why is BLAST a local aligner vs. global

Local alignment

- Sequence extensions do not necessarily extend to sequence ends
  - Domains vs entire proteins
- Can be multiple query->reference matches
  - i.e., alignment can be broken, each with own statistics
- Can be multiple reference matches to the same query

Sequence masking

- Low-complexity regions can arise convergently
  - Small hydrophobic amino acids common in transmembrane helices
- Violates homology assumption, therefore often excluded from BLAST search

Comparing BLAST scores

- Different BLASTs can use different parameters, e.g., matrices & gap penalties
- "Bit scores" normalize for this

$$S' = (\lambda S - \ln K)/\ln 2$$

Bit score: S'; Matrix penalty: λ; Score: S; Gap penalty: K

E-values

- What is the likelihood that the sequence similarity is due to chance vs. actual homology?
- Larger databases are more likely to include chance matches

E-values

Length of the query sequence

Bit score

$$E = (n \times m)/(2^{S'})$$

E-value

Total # of residues in the database

E-values

- The E-value represents the likelihood of a random match >= the calculated score
- Smaller E-values therefore reflect greater probability of true homology
- Typically $1e^{-5}$ operationally used as a threshold for considering sequences as homologous

Summary

- Wednesday: applying BLAST
- Next week: expanding from one->many sequence comparisons to many->many