Bioinformatics 1
lecture 13

- Database searches. Profiles
- Orthologs/paralogs
- Tree of Life term projects

Various ways to do database searches

- Purpose of database search (what you want)
  - phylogenetic analysis (accurate distances)
  - functional inference (motif match)
  - 3D modeling (accurate alignment)
- Protein versus DNA search
  - Highly similar seqs --> DNA
  - Non-coding --> DNA
  - Otherwise --> protein
- Degree of similarity

forms of BLAST

BLAST	query	database
blastn	nucleotide	nucleotide
blastp	protein	protein
tblastn	protein	translated DNA
blastx	translated DNA	protein
tblastx	translated DNA	translated DNA
psi-blast	protein, profile	protein
phi-blast	pattern	protein
transitive blast*	any	any

*not really a blast. A way of using any blast.

Variants of PAM, BLOSUM

When you run an alignment, you get to choose the substitution matrix. How do you know which to use?

larger evolutionary timestep ↑

BLOSUM45	distant homologs, %ID < 50	Distance relationships, gap penalty = lower
BLOSUM62	homologs, <%ID>=62	Moderate similarity, optimal for mixed MSA
BLOSUM90	close homologs, few gaps	Close homologs, gap penalty = high

**Gap penalties optimized against gold-standard alignments (with the same %ID range) such as the alignments in BAliBase

Psi-BLAST: Blast with profiles

Psi-BLAST algorithm.

(Cycle 1) Normal BLASTp search. Align all sequences below e-value cutoff.

(Cycle 2) (a) Construct a **profile** from the results of **Cycle 1.**

(b) Search the database using the profile. Align.

(Cycle n=3...) (a) Construct a **profile** from the results of **Cycle n-1.**

(b) Search the database using the profile. Align.

etc.

Psi-BLAST is **much more *sensitive*** than BLAST.

Psi-BLAST is vulnerable to ***low-complexity***.

Psi-BLAST is usually the first step in **homology modeling**.

Aligning sequence to profile

```
S(i,j) = 0
do aa=1,20
      S(i,j) = S(i,j) + P(aa,i)*B(aa,s(j))
enddo
```

profile1@i P(aali)

sequence2@j

BLOSUM score

Aligning profile to profile

```
S(i,j) = 0
do aai=1,20
   do aaj=1,20
      S(i,j) = S(i,j) + P(aai,i)*P(aaj,j)*B(aai,aaj)
   enddo
enddo
```

No need to normalize, since $\sum_{aa_i} \sum_{aa_j} P(aa_i|i)*P(aa_j|j) = 1$

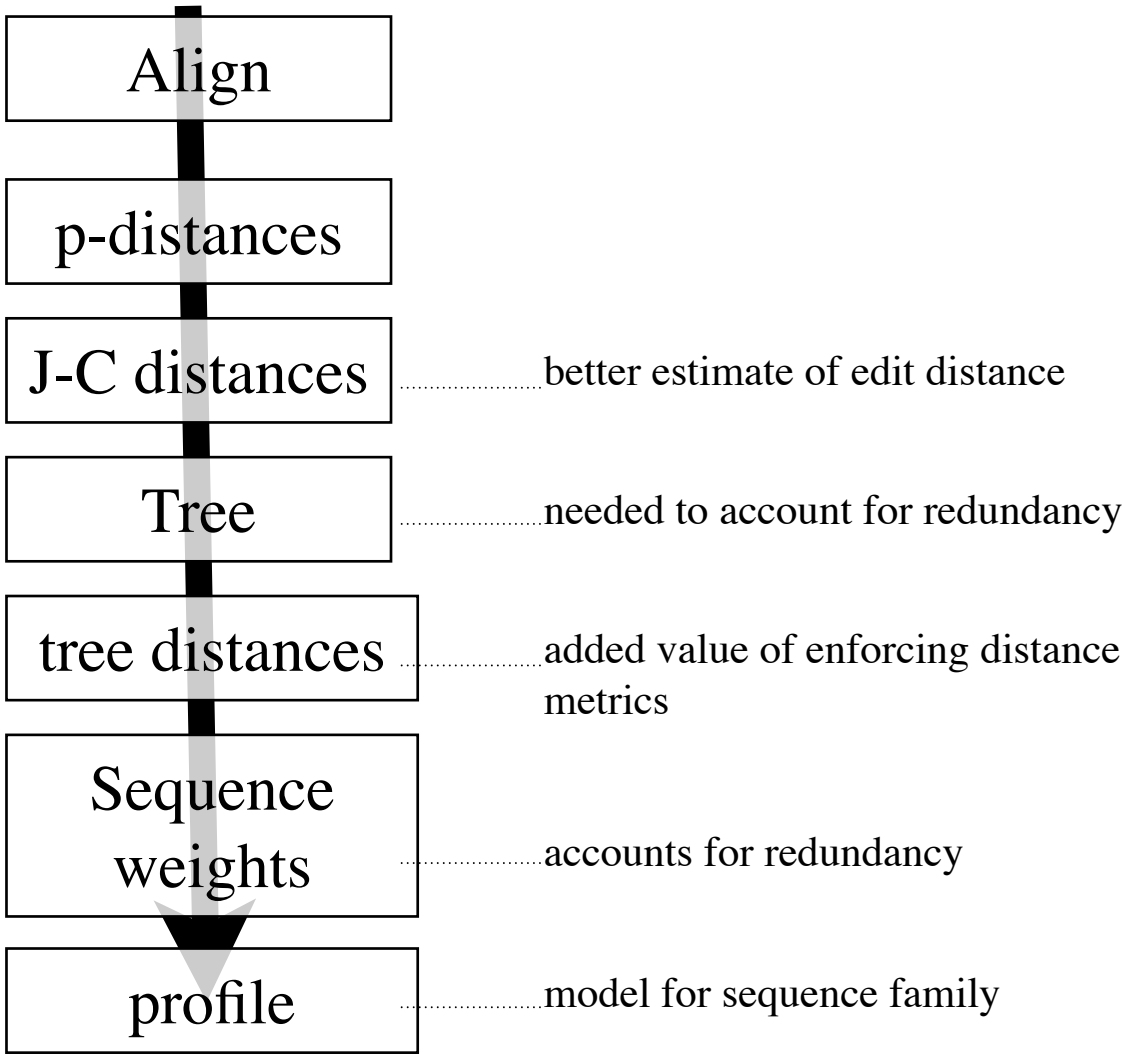
PHI-BLAST -- Patterned Hit Initiated BLAST

Table 1. Detection of subtle protein sequence relationships using PHI-BLAST

Conserved domain or motif under investigation	Pattern[a]	GenBank (30) accession no. of query	Top non-trivial relevant hit found by PHI-BLAST		Top non-trivial relevant hit found by BLAST	
			Accession no.	*E*-value	Accession no.	*E*-value
A. P-loop ATPase domain in apoptosis regulators and plant stress response proteins	[GA]xxxxGK[ST]	231729	2213598	0.038	2961373	4.7
B. ATPase domain in mismatch repair protein MutL, type II topoisomerases, histidine kinases, and HS90 molecular chaperones	hxhxDxGxG	127552	488200	0.017	2495364	1.8
C. Nucleotidyltransferase domain in archaeal tRNA nucleotidyltransferases	DhDhhh	2826366	2650333	0.061	2650333	8.6
D. Motif VI of superfamily II helicases in archaeal homologs of bacterial DNA primases	QxxGRx[GA]R	2128723	2499099	0.54		

Steps in building a profile

from a set of homolog sequences



In class exercise: **build a profile**

Us “Alignment of Animal Sequences” in Geneious Sample Documents

(1) Run Tree

(2) Look at tree distances. (“patristic”)

(3) Calculate **sequence weights** based on the sum of distances,. $w_A = \sum_i D_{iA}$

Then normalize (divide by sum of weights $\sum_i w_i$).

(4) Sum the probabilities of each AA in the **nth column**. (pick one)
For example: P(Q) = sum of w_i over sequences that contain an Q.

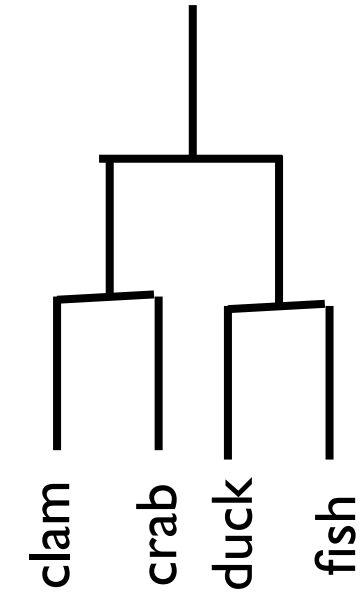
(5) Convert each P() to a LLR using equal probability AAs (0.05) as the expected value. Use a **pseudocount** of 0.02

$$LLR = \log((P(n)+0.02)/(0.05)) / \log(2)$$

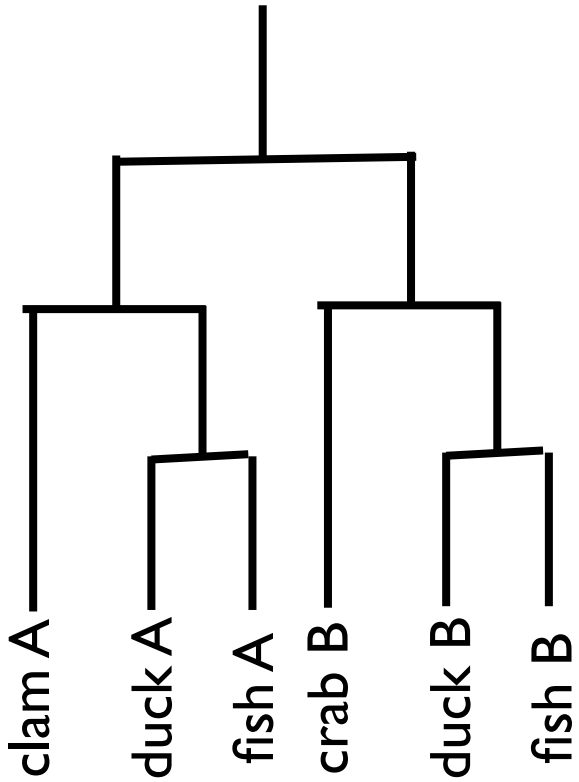
(6) Stack letters, **Logo style**. Height of letter = bits.

Orthologs/paralogs

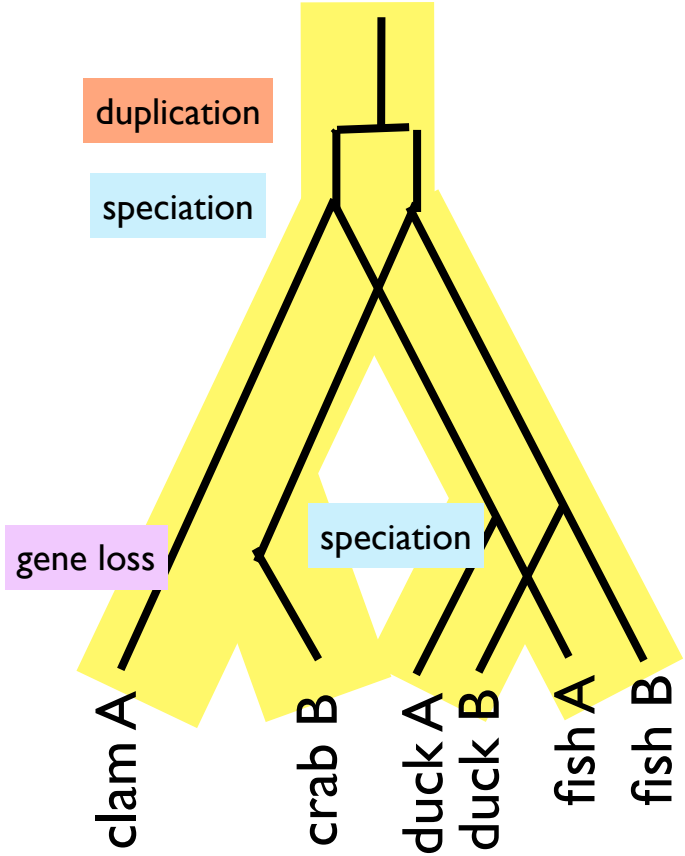
Orthologs: homologs originating from a speciation event
Paralogs: homologs originating from a gene duplication event.



Species tree



Sequence tree

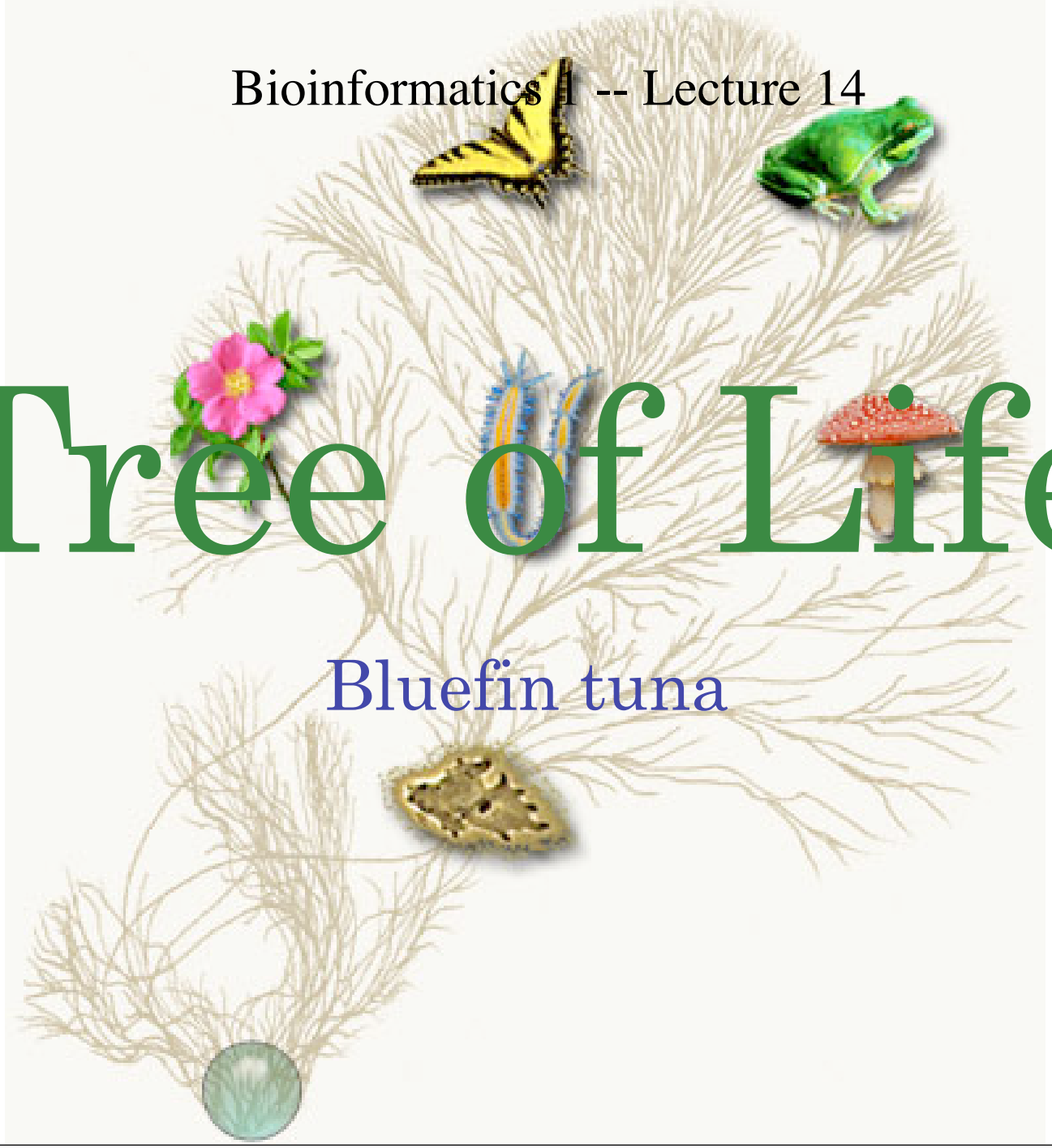


reconciled trees

How do I know it's a paralog?

- If it's a paralog, then at some point in evolutionary history, a species existed with two identical genes in it.
  - One may have been lost since then. (Descendants are still paralogs!)
  - Paralogs can be from different species.
- Paralogous genes have more than the expected sequence divergence.
  - Because they are more likely to have different functions
  - Because they diverged earlier than the speciation event.
- Without species information or functional information, it's impossible to tell

Bioinformatics 1 -- Lecture 14

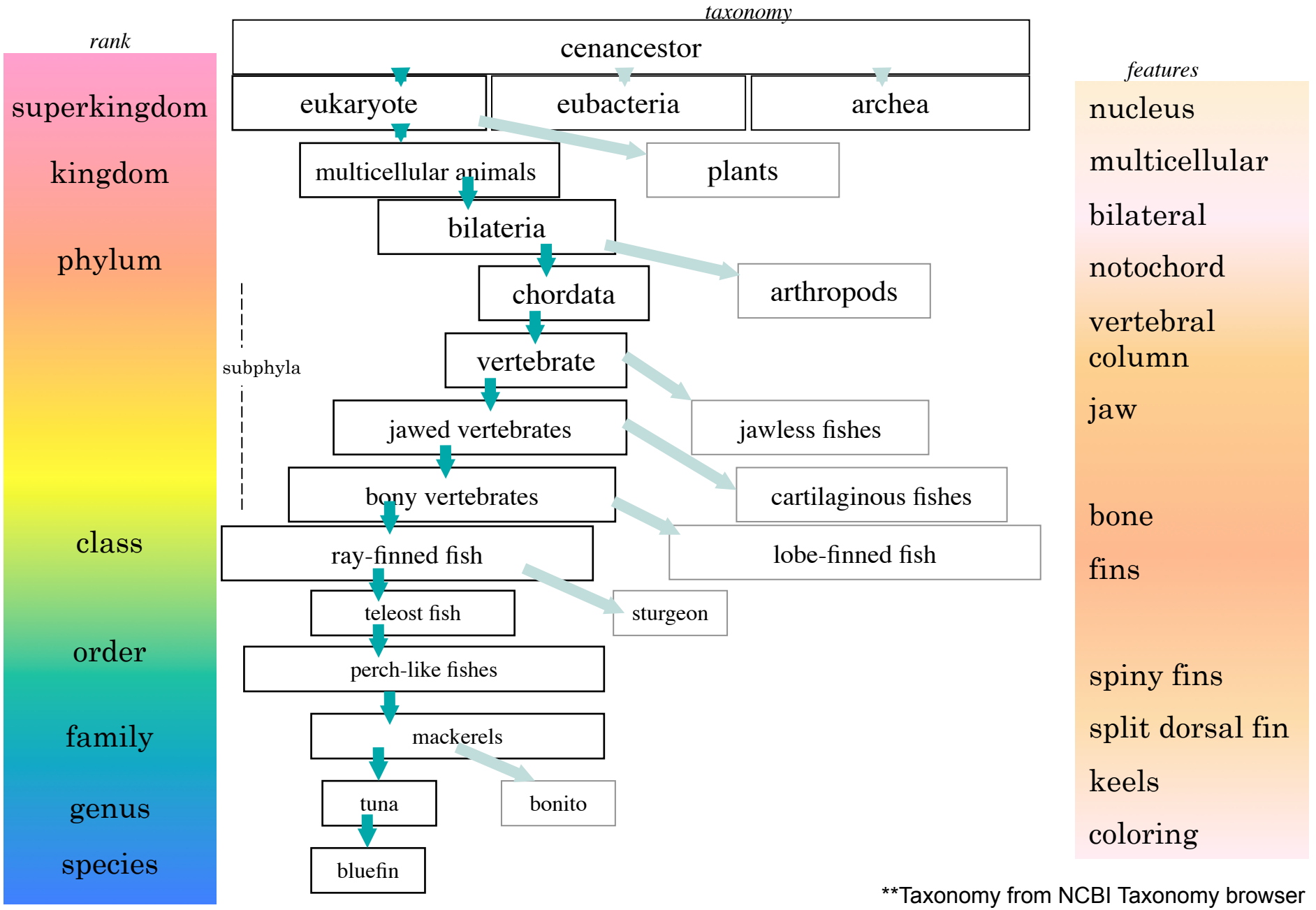


Tree of Life

Bluefin tuna

tolweb.org

Location of **bluefin tuna** in tree of life

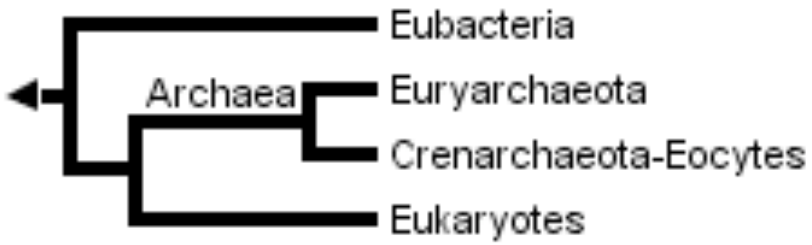


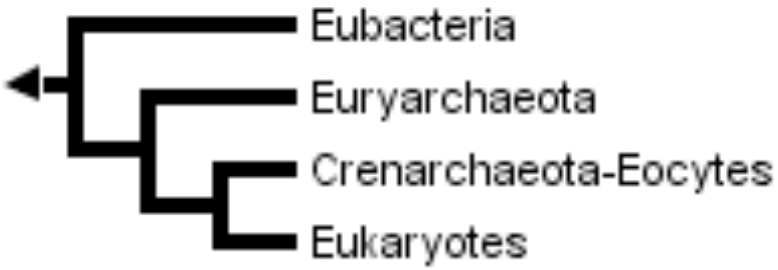
**Taxonomy from NCBI Taxonomy browser

In class exercise: explore TOL

- tolweb.org
- Start from the root of the tree
  - Find homo sapiens starting
  - Where do the following first appear
    - two eyes
    - jaw
    - four limbs
    - five fingers
    - no tail

tolweb.org/tree

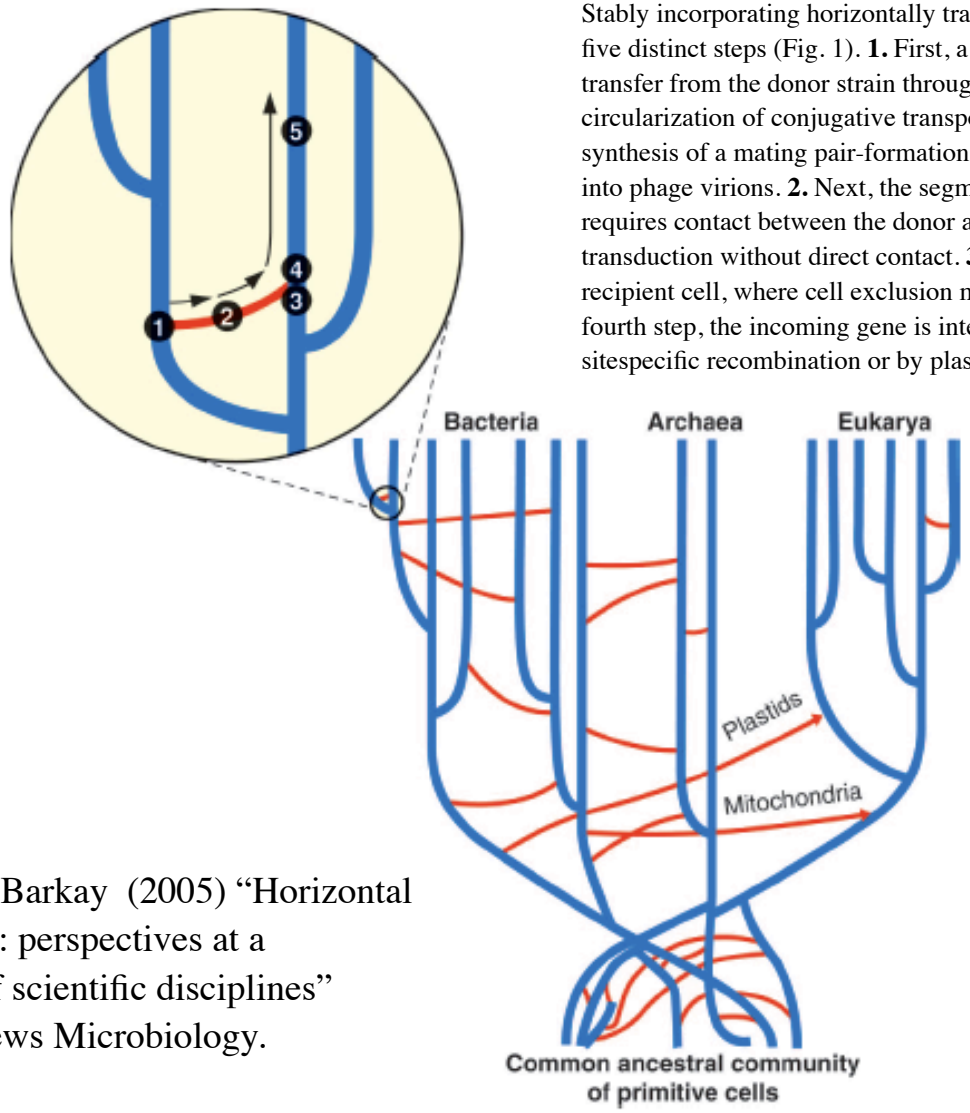




Earliest branching... still controversial.

Life is not strictly a tree --

horizontal gene transfer



BF Smets, T Barkay (2005) "Horizontal gene transfer: perspectives at a crossroads of scientific disciplines" Nature Reviews Microbiology.

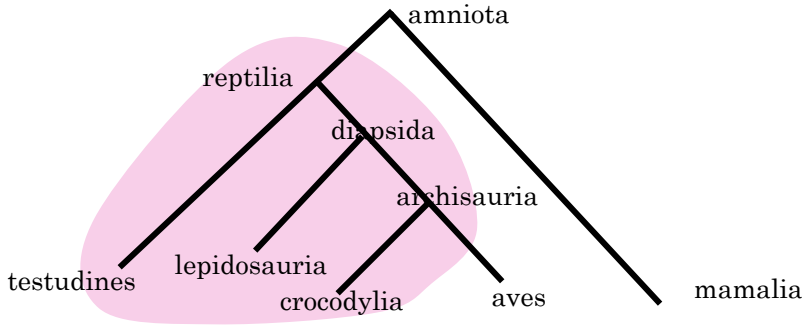
Discrete Steps Needed for Stability of Gene Transfer

Stably incorporating horizontally transferred genes into a recipient genome involves five distinct steps (Fig. 1). **1.** First, a particular segment of DNA or RNA is prepared for transfer from the donor strain through one of several processes, including excision and circularization of conjugative transposons, initiation of conjugal plasmid transfer by synthesis of a mating pair-formation protein complex, or packaging of nucleic acids into phage virions. **2.** Next, the segment is transferred either by conjugation, which requires contact between the donor and recipient cells, or by transformation and transduction without direct contact. **3.** During the third step, genetic material enters the recipient cell, where cell exclusion may abort the transfer. **4.** Otherwise, during the fourth step, the incoming gene is integrated into the recipient genome by legitimate or sitespecific recombination or by plasmid circularization and complementary strand synthesis. Barriers to transfer during this step come from restriction modification systems, failure to integrate and replicate within the new host genome, and incompatibility with resident plasmids. **5.** In the final step, transferred genes are replicated as part of the recipient genome and transmitted to daughter cells in stable fashion over successive generations. Researchers from different disciplines tend to focus on specific stages within this five-step sequence. Thus, evolutionary biologists who examine microbial genomes for evidence of past transfers tend to look at HGTs from the perspective of step five. Molecular biologists are more likely to examine the details of the transfer events, while microbial ecologists look more broadly when they describe the magnitude and diversity of the mobile gene pool, sometimes called the mobilome.

Classification is not phylogeny

...not always

- **Paraphyly** (paraphyletic) -- a taxonomic classification that does not contain all of the descendents of the common ancestor.
  "Reptile" is paraphyletic because does not include birds which diverged after the common ancestor.
- **Monophyletic** -- a classification that includes all descendents of a common ancestor.



Phylogenetics using Mitochondrial sequences

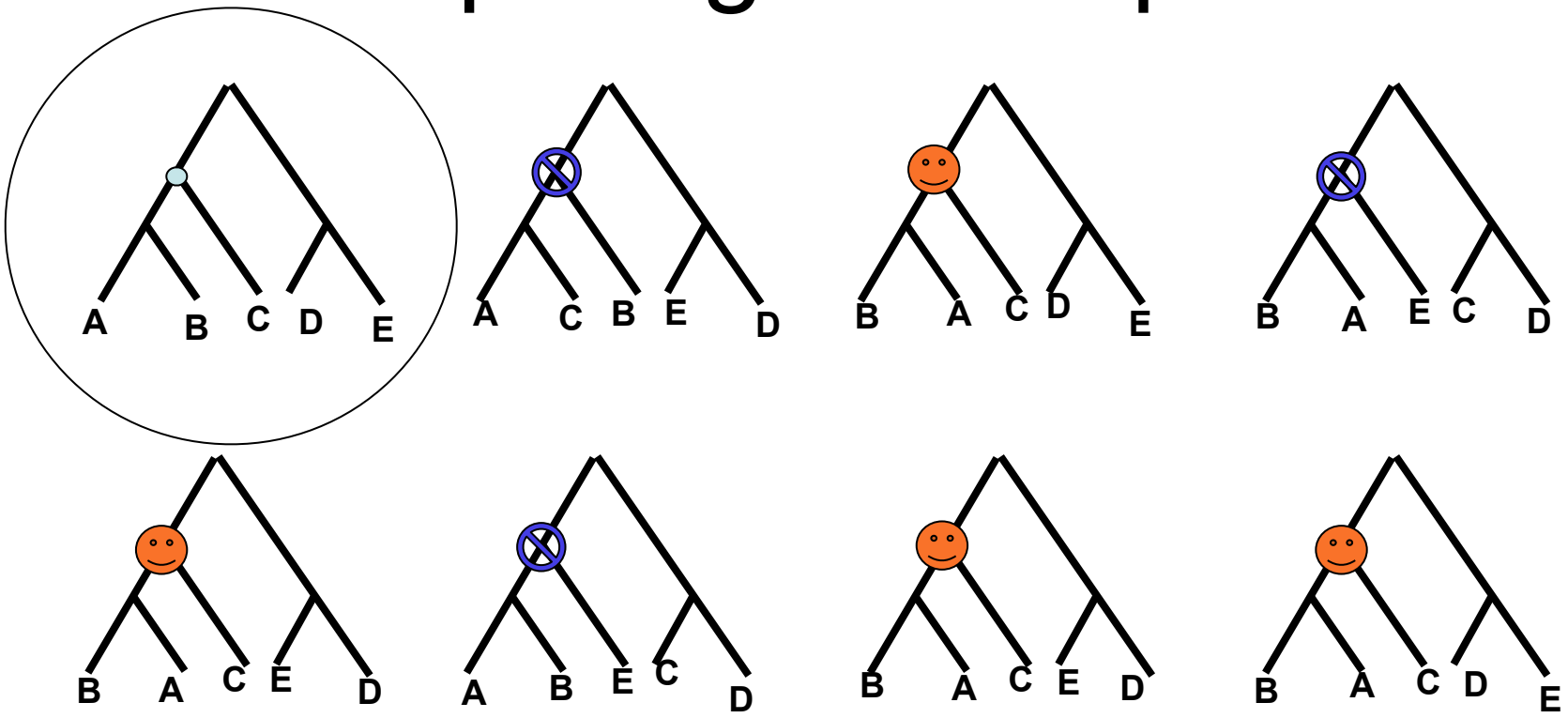
- **Blastn** search against '**nr**' using *thunnus thynnus* mitochondrial DNA (whole genome = 16526 bp)
- Used **Blast tree view** to find a clade of about 10 species.
- Downloaded those species in GenBank format
- Imported these into **Geneious**. Ran **ClustalW**
- Selected **10 non-overlapping regions** and calculated a tree for each.
- Assigned confidences to branchpoints by hand using **Bootstrap** analysis.

“Boot strap analysis”

- A method to validate a phylogenetic tree, branchpoint by branchpoint.
- Requires a means to generate independent trees. (For example trees generated from different regions of the mitochondrial genome.)
- Choose the representative tree as the ‘parent’. Calculate the following:

For each branchpoint in the parent tree,
  For each tree, ask
    Is there a branchpoint having the same subclade contents (i.e. same taxa, any order)
Bootstrap value = number of trees having the branchpoint / total trees.

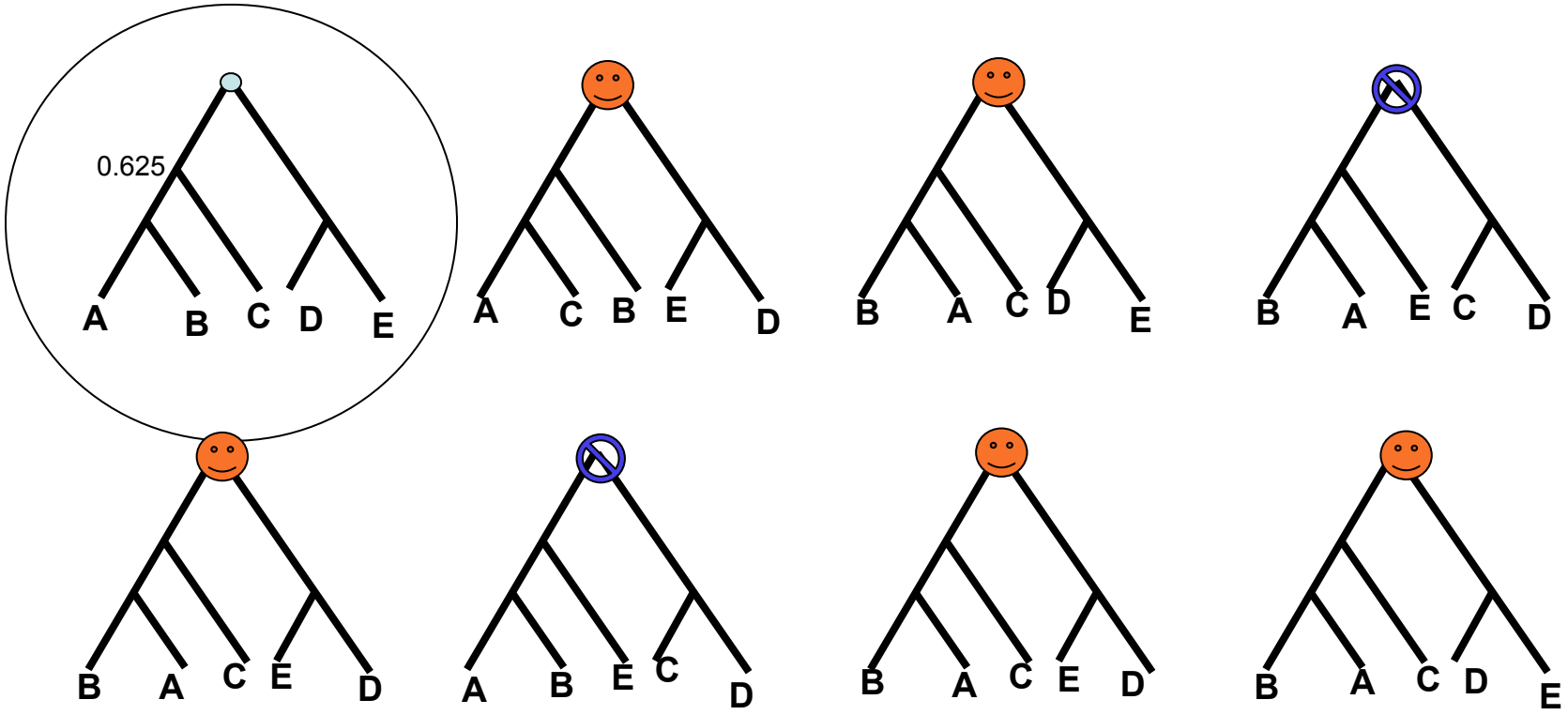
Comparing branchpoints



- = P((A,B),C) = 5/8

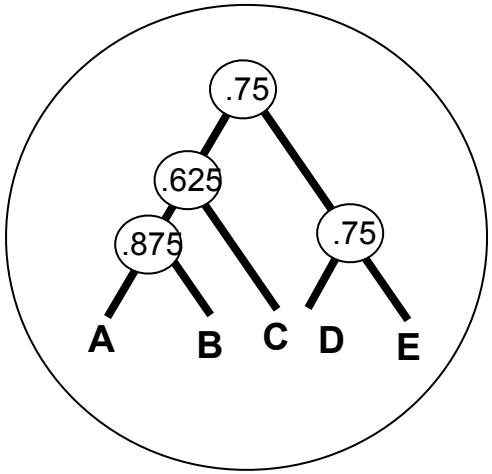
For each branchpoint in the parent tree,
For each tree, ask
Is there a branchpoint having the same subclade contents (i.e. same taxa, any order)
Bootstrap value = number of trees having the branchpoint / total trees.

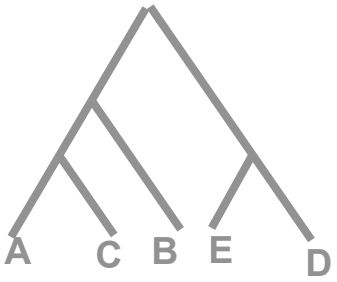
Comparing branchpoints

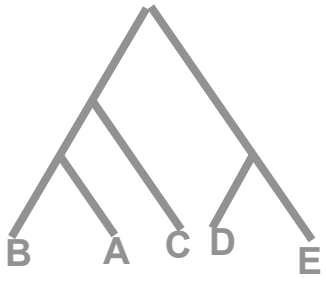


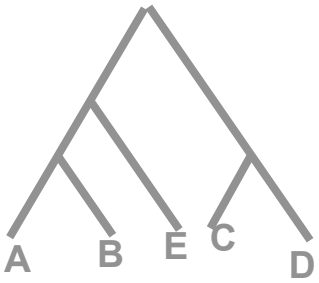
- = P((A,B,C),(D,E))=6/8

Bootstrap values for this data

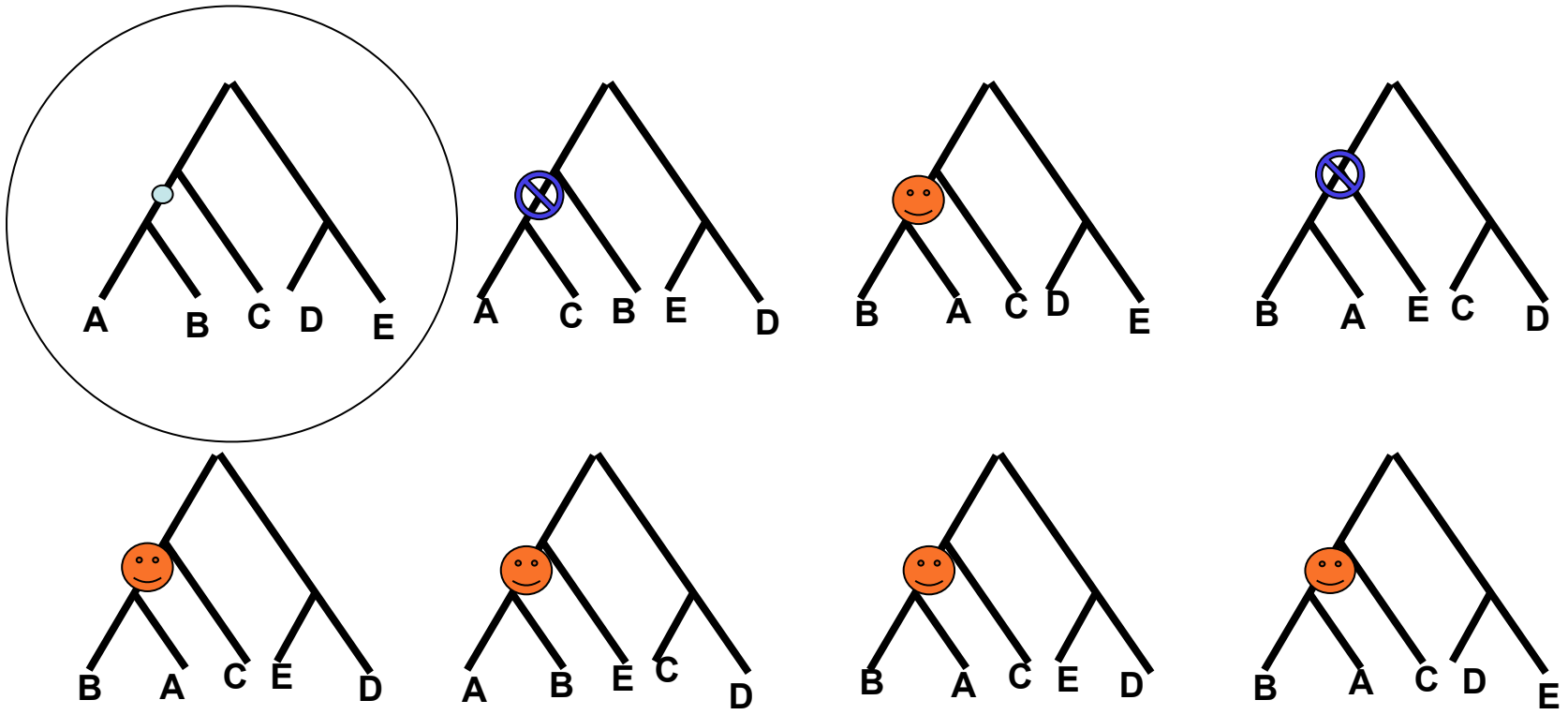








Comparing lineages instead of branchpoints.



- = P(A,B) = 6/8

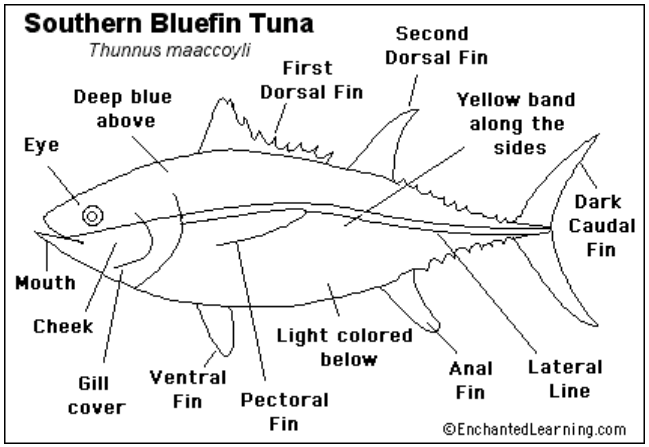
For unrooted trees *or* rooted trees.

- Treat any lineage as the root.
- Ask how often the root branching is conserved.

Atlantic bluefin tuna: *thunnus thynnus*

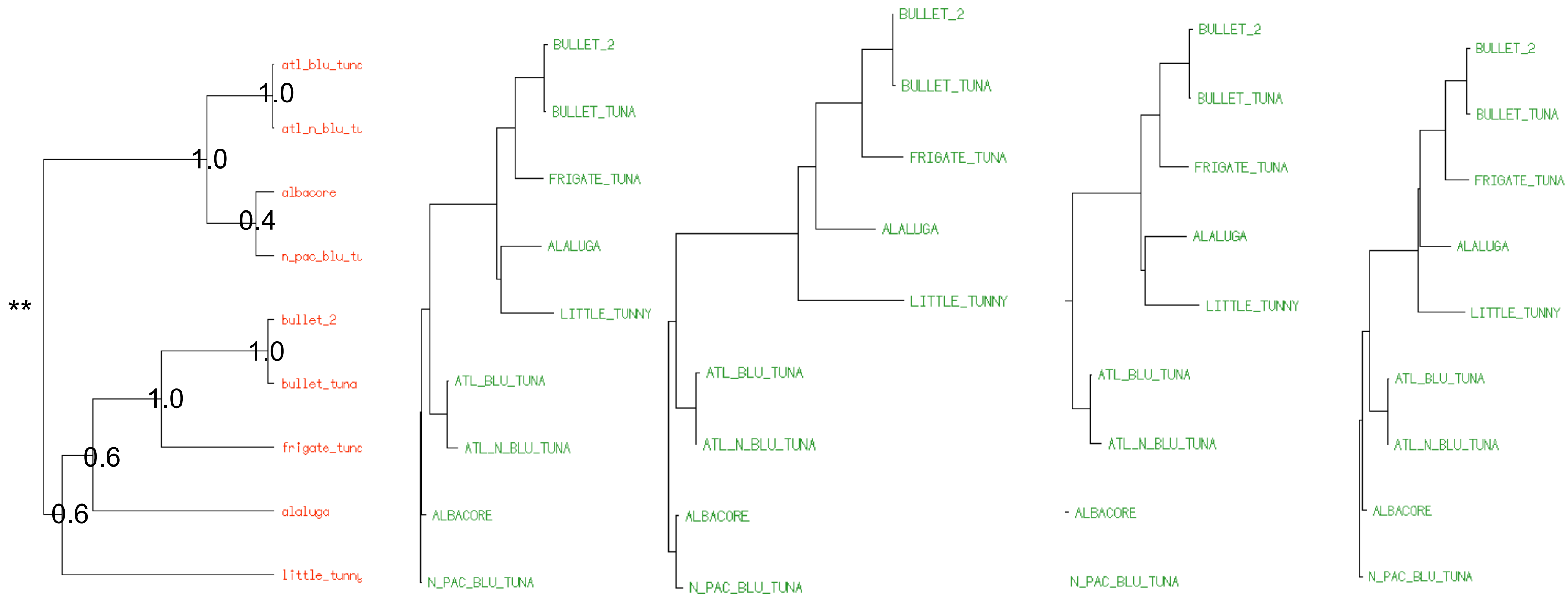
Kingom	Phylum	Class	Order	Family	Genus	Species
ANIMALIA	CHORDATA	ACTINOPTERYGII	PERCIFORMES	SCOMBRIDAE	THUNNUS	THYNNUS
animal	**notochord**	**ray-finned fishes**	**perch-like**	**mackerels**	**tuna**	**bluefin**

Critically Endangered (IUCN 2.3)



http://www.iucnredlist.org/search

Bluefin tuna trees

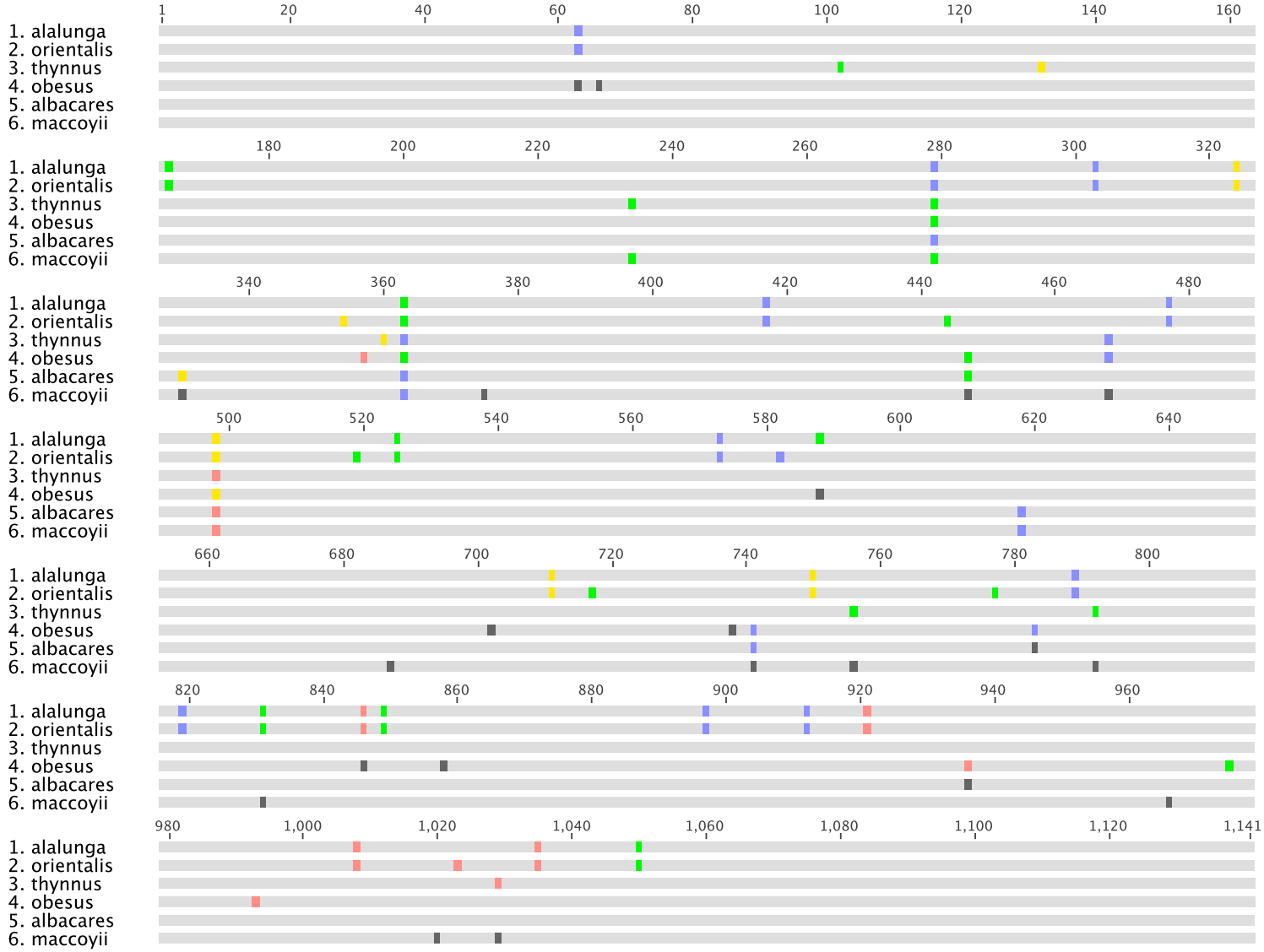


The classification “bluefin” is paraphyletic, since the common ancestor of N. Atlantic Bluefin and Pacific Bluefin includes a non-bluefin tuns, the albacore.

**Unrooted trees. Bootstrap value for root not possible.

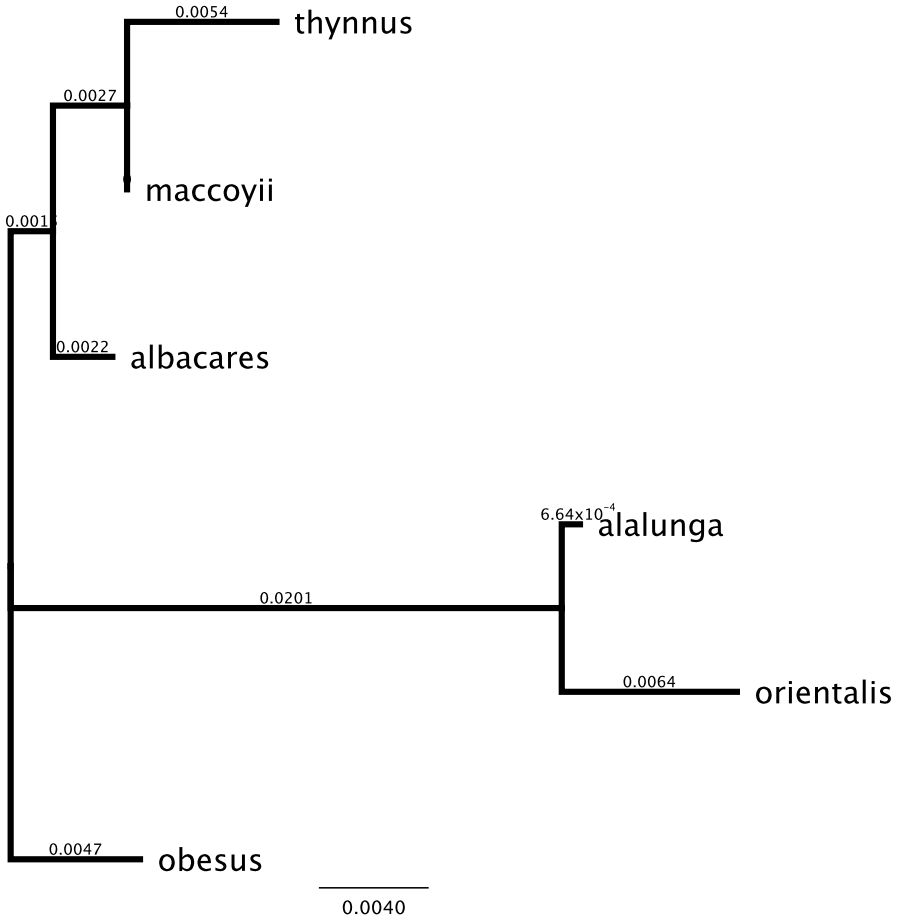
Tuna MSA

Mitochondrial cytochrome B (DNA) -- one of the most widely sequenced genes

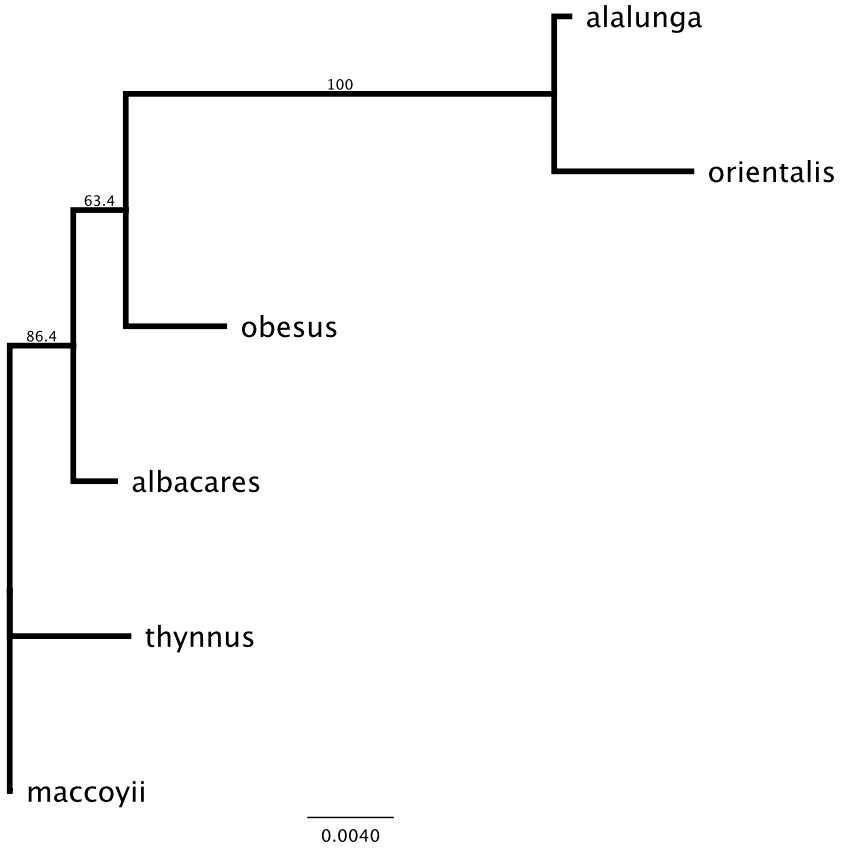


Are there enough differences here?

Phylogram



Jackknife tree (1000 iterations)



Consensus tree

atl_blu_tuna

atl_n_blu_tuna

albacore

n_pac_blu_tuna

1

bullet_2

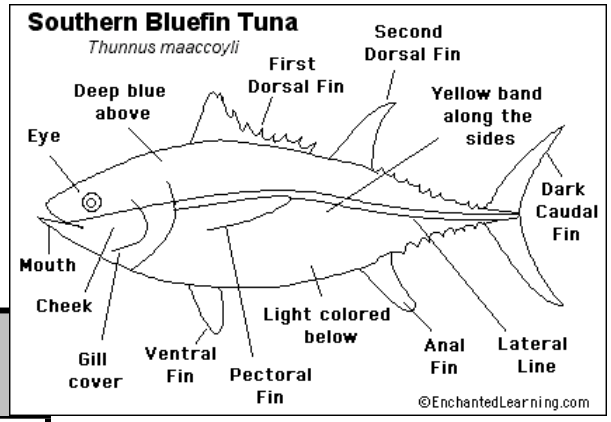
bullet_tuna

frigate_tuna

alaluga

little_tunny

Tuna features



feature	tree result	
• size and shape	monophyletic	1
• tropical versus sub-tropical		
• number of gill rakers		
• number of dorsal spines		
• existence/number of anal spines		
• coloring	paraphyletic	1
• fin shape		
• caudal keels/shape		
• pelagic versus coastal		