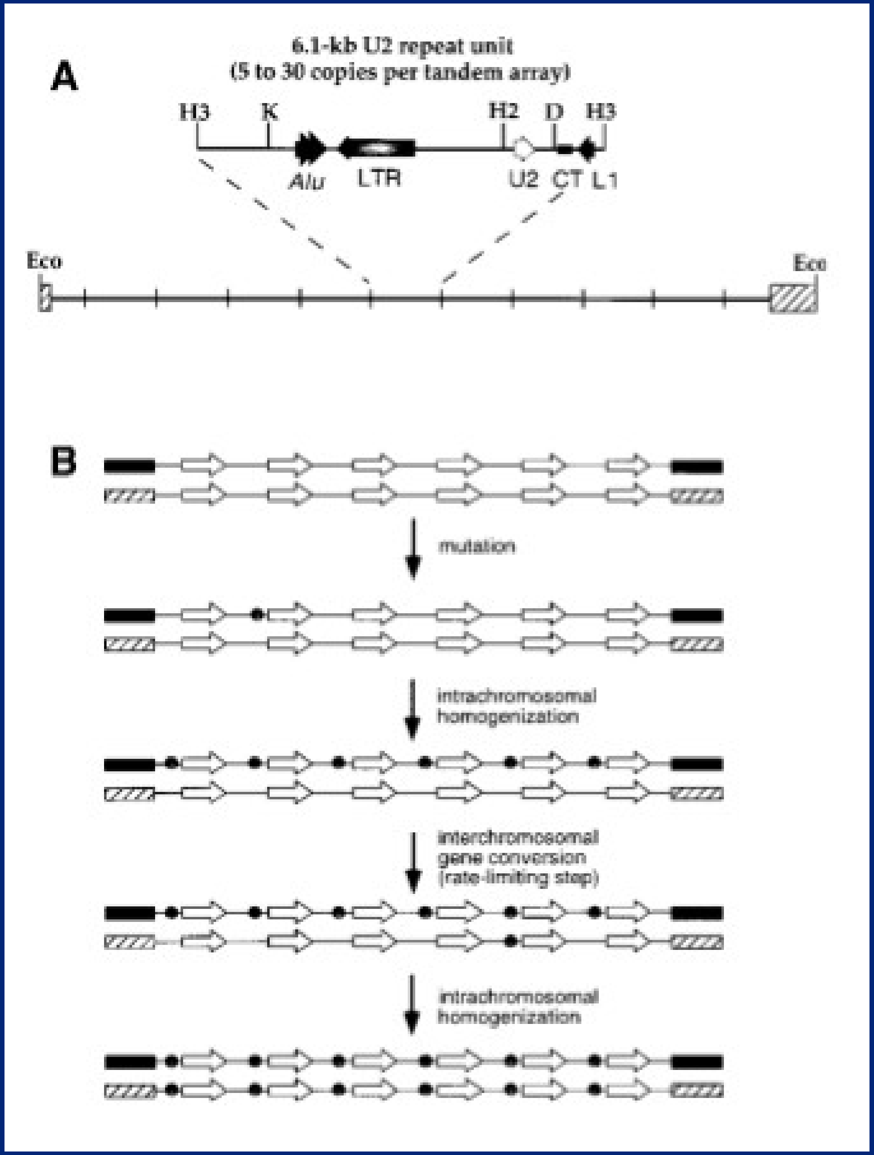
Roadmap

- Left over from Friday:
  - RNU2 locus
  - P element in large vs. small populations
  - Transposition machinery co-opted by cell
- Phylogenetic trees:
  - Interpreting a tree
  - Inferring trees by parsimony
  - Inferring trees by distance methods



From Liao (1999) AJHG

Thought problem

- Cross P into a lab strain and maintain a population in bottles
- Population may live or die
- Which is likely to do better, a large or small population?
- *When this experiment has been done, large populations are more likely to survive*

McClintock’s “genome shock” hypothesis

- Transposons could allow an organism to control its mutation rate:
  - Suppress transposition when well adapted
  - Permit transposition when struggling, “hoping” for a useful mutation
- Alternative hypothesis: transposons are purely selfish
  - Suppress transposition whenever possible
  - Fail to suppress transposition when badly stressed
- Not easy to test these alternatives

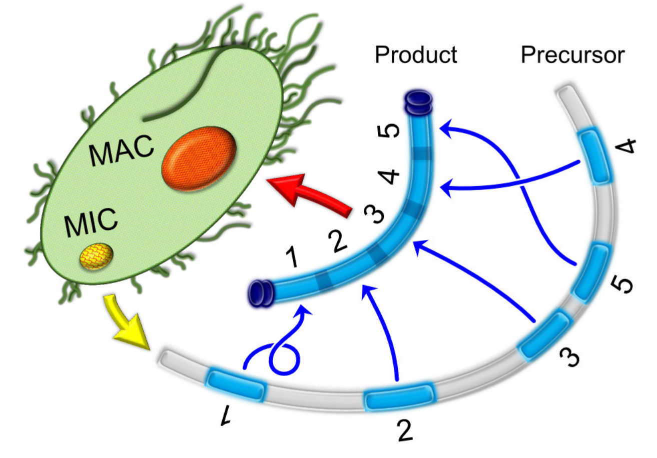
Finding a use for transposons

Oxytricha genome rearrangement

- Two nuclei per cell:
  - Micronucleus used for reproduction, but genes not active
  - Macronucleus expresses genes
- Macronucleus genome is highly rearranged:
  - Cut into around 16,000 tiny chromosomes
  - Usually 1 gene per chromosome
  - Genes are re-assembled from fragments
  - Some are duplicated (dosage control?)
  - 95% of germline genome is destroyed

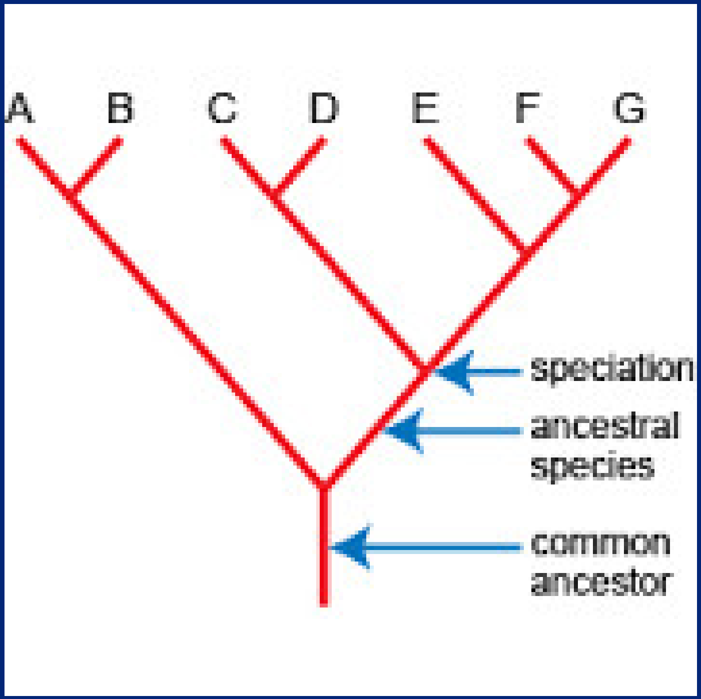
Transposons harnessed to chew up genome

- Germline genome of *Oxytricha* full of transposons
- Macronucleus has none
- If transposases are inactivated, the macronucleus fails to develop properly
- Transposons and transposase probably central in rearrangement process



What is a phylogeny?

- A branching tree showing inferred relationships
- Taxon, taxa: the units at the tips of the tree (species, populations, individuals, genes)
- Clade: all taxa descending from a common ancestor
- Root: the common ancestor of the whole tree



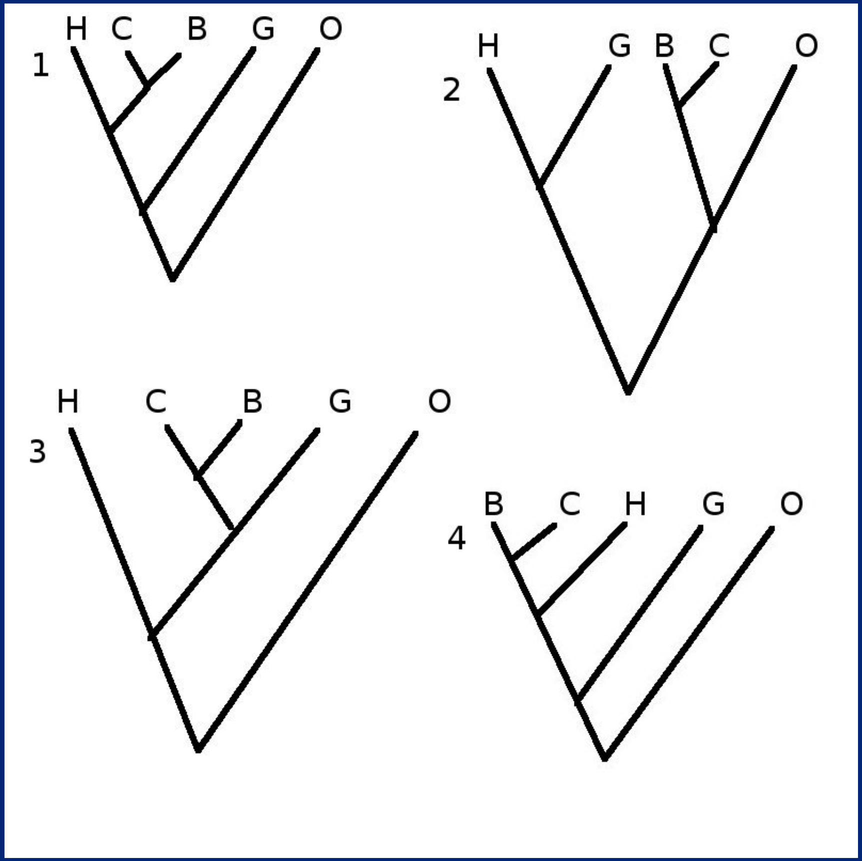
What are phylogenies good for?

- Relationships between organisms, populations, species
- Dates of evolutionary events
- Evolutionary patterns–did some features evolve multiple times?
- Removing influence of phylogeny from ecological analyses ("comparative method")
- Relationships among genes
- Patterns of speciation and diversification

How to look at a phylogeny

- Branching pattern shows pattern of relationships
- Right-left ordering is NOT significant; can be rearranged to emphasize or obscure points!
- Branch lengths may or may not be meaningful
- Biologists draw root at the bottom; math and CS types draw root at the top

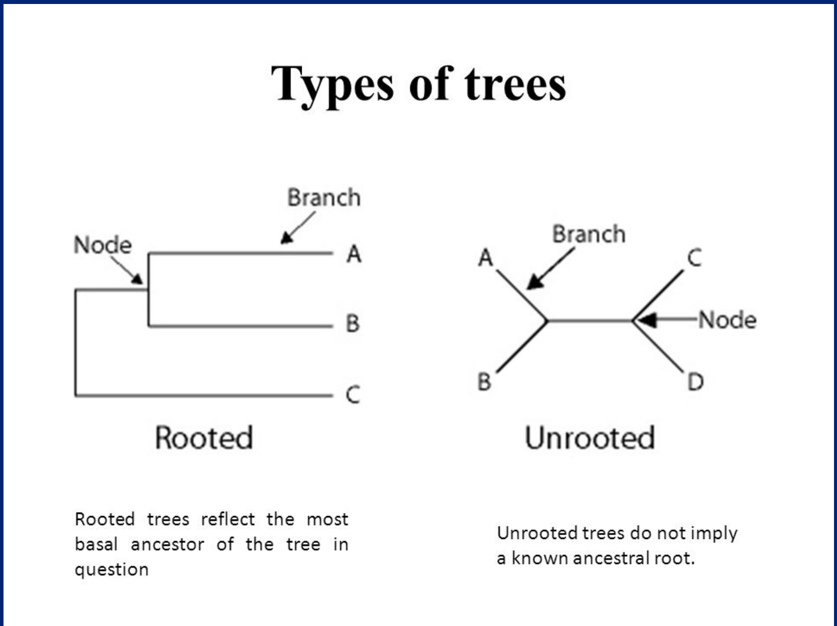
Practice problem



Two of these trees are the same (except for branch lengths). Which two?

Rooted versus unrooted trees

- A rooted tree (phylogeny) has a specific direction of evolution
- The root is the ancestral form from which the others evolved
- This is the most informative type of tree
- Unfortunately, most phylogeny inference methods produce unrooted trees



Types of trees

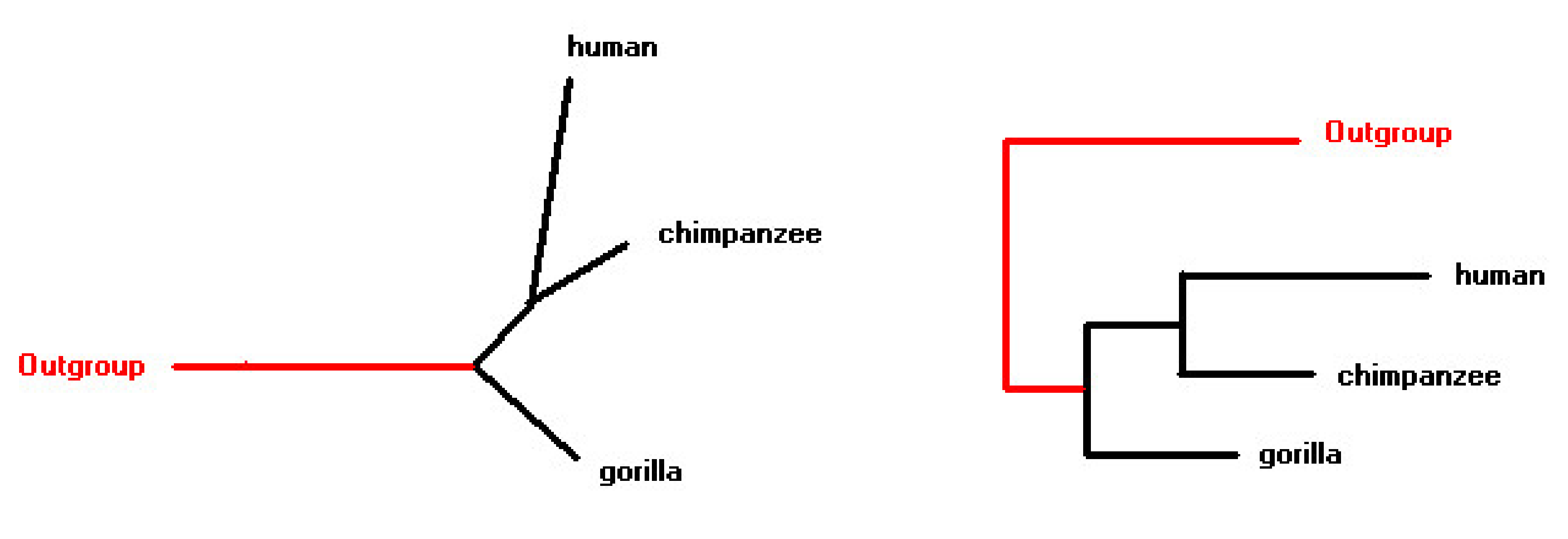
Rooted trees reflect the most basal ancestor of the tree in question

Unrooted trees do not imply a known ancestral root.

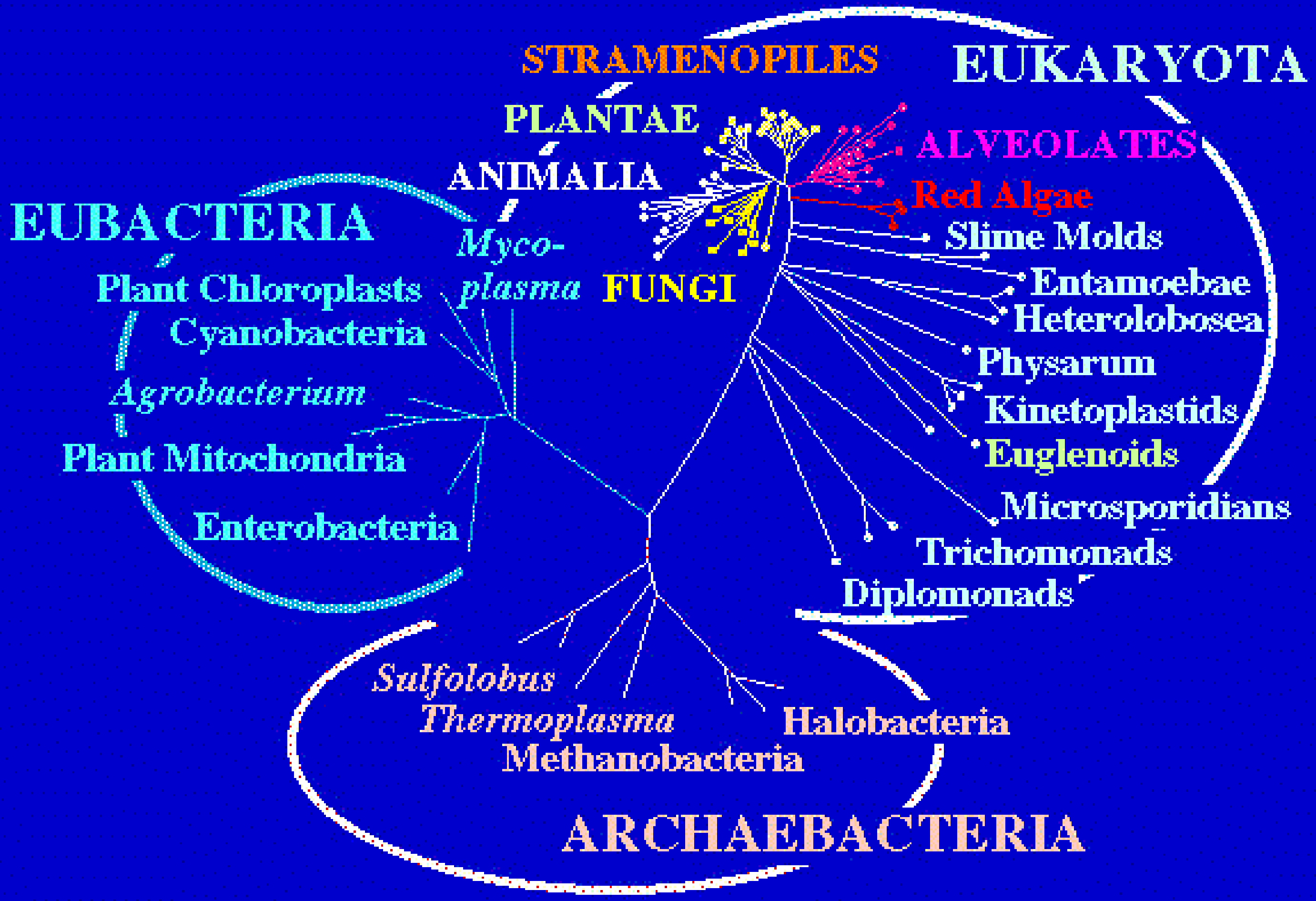
Rooted versus unrooted trees

- An unrooted tree corresponds to a collection of different rooted trees
- We don't know the direction of evolution
- Biological interpretation can be difficult without root
- Ways to root a tree:
  - Outgroup
  - Molecular clock

Outgroup rooting



- Outgroup – species known not to belong to clade
  - Wrong outgroup leads to wrong root
  - Too-distant outgroup leads to noise in data
- Some comparisons have no suitable outgroup


STRAMENOPILES
EUKARYOTA
PLANTAE
ALVEOLATES
ANIMALIA
Red Algae
EUBACTERIA
Myco-
plasma
Slime Molds
FUNGI
Entamoebae
Plant Chloroplasts
Heterolobosea
Cyanobacteria
Physarum
Agrobacterium
Kinetoplastids
Plant Mitochondria
Euglenoids
Microsporidians
Enterobacteria
Trichomonads
Diplomonads
Sulfolobus
Thermoplasma
Halobacteria
Methanobacteria
ARCHAEBACTERIA

Molecular clock

- Can we assume same rate of evolution everywhere (molecular clock)?
- If so:
  - Root is point most distant from all tips
  - Branch length is proportional to time
  - Dating any point on tree dates whole tree
- Clock may not hold:
  - Unequal generation time
  - Different selection constraints
  - Different mutation rates
- Clock assumption safest among closely related species

Appropriate data for phylogenies

- Good phylogenetic data has:
  - Enough variation to show relationships
  - Not so much variation that it randomizes signal
  - Ability to establish homology
  - *Relative freedom from convergent evolution*
  - Mode of evolution relatively well understood
  - If possible, a good clock

- No one type of data works for all problems

Some important dates in history

Origin of the universe	–12[a] ±2
Formation of the solar system	–4.6 ±0.4
First self-replicating system	–3.5 ±0.5
Prokaryotic-eukaryotic divergence	–2.5 ±0.3
Plant-animal divergence	-1.0
Invertebrate-vertebrate divergence	-0.5
Mammalian radiation beginning	-0.1

[a] Billions of years ago

	PAMs[a]/100 res.		
Protein family	/10^8 years	Protein Lookback time[b]	
Pseudogenes	400	45[c]	Primates,Rodents
Fibrinopeptides	90	200	Mammalian Radiation
Lactalbumins	27	670	Vertebrates
Ribonucleases	21	850	Animals
Hemoglobins	12	1.5[d]	Plants/Animals
Acid Proteases	8	2.3	Prokayrotic/Eukarotic
Triosphosphate isomerase	3	6	Archaen
Glutamate dehydrogenase	1	18	?

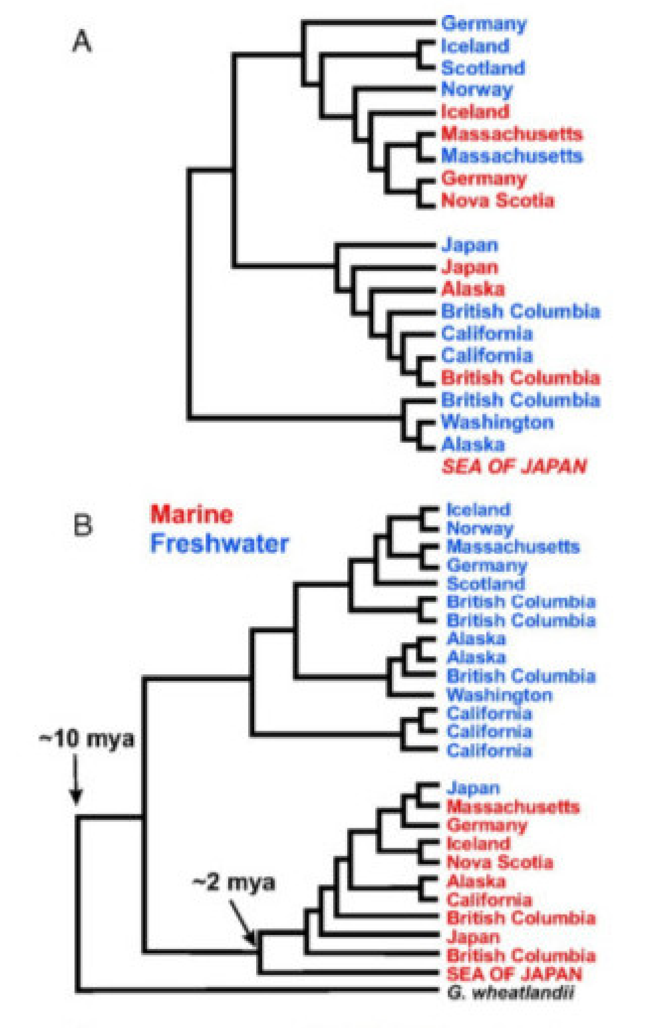
[a] PAMs, point accepted mutations. [b] Useful lookback time, 360 PAMs, 15% identity. [c] Millions of years. [d] Billions of years.

Convergent evolution?

- Why not use loci involved in “exciting” traits of the species?

- Convergent evolution:
  - Two clades are under the same external pressure
  - They independently evolve the same response
  - Not a reliable indicator of relationships

- Upper figure is many random genes; lower is a gene involved in fresh/saltwater adaptation



Why phylogeny inference is hard

Tips	Topologies
3	3
4	18
5	180
6	2700
7	56700
8	1587600
9	57153600
10	2571912000
15	6958057668962400000
20	564480989588730591336960000000
30	43684666131030695124646801986207638914406400000000000000
40	3027333829948007356546303364551457200042939432053862501707888721920000000000
50	3.28632×10^{112}
100	1.37416×10^{284}

Why phylogenies are hard

- In many cases tree search known to be “NP complete”
- No efficient algorithm is known–none may exist but this is unproven
- Solving any NP-complete problem solves ALL OF THEM
- Four consequences of such an algorithm:
  - Reliably find the best phylogeny
  - Win USD 1 million (Millennium Prize) from Clay Institute
  - Crack most/all current codes (business and military)
  - Difficult conversation with the NSA....
- Must use heuristic approximations which will sometimes fail (get the wrong tree)

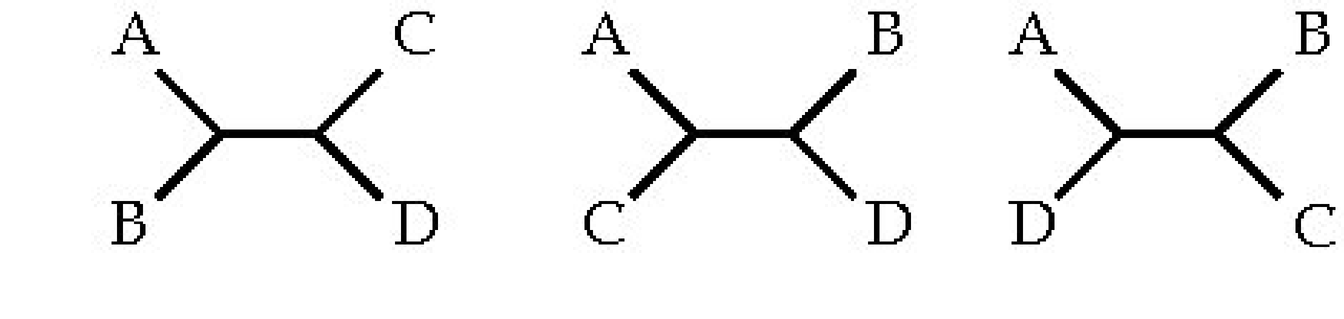
Four major approaches to phylogeny inference

- Prefer the tree which–
  - Parsimony: explains the data with the fewest mutations
  - Distance: minimizes the difference between observed and expected distances between taxa
  - Likelihood: maximizes the probability of the data
  - Bayesian: maximizes the posterior probability of the data given a prior
- The first two are easier: given a correct mutational model the second two are likely more accurate

Parsimony

- Prefer the tree which explains the data with the fewest events (mutations)
  - Does not use a model of the mutation process (so can't use the wrong one)
  - Implicitly assumes changes are rare
  - Applicable to wide range of data:
    - Sequences (DNA, RNA, protein)
    - Genome rearrangements
    - Morphological traits
  - Has issues if some branches are much longer than others

Practice problem–parsimony

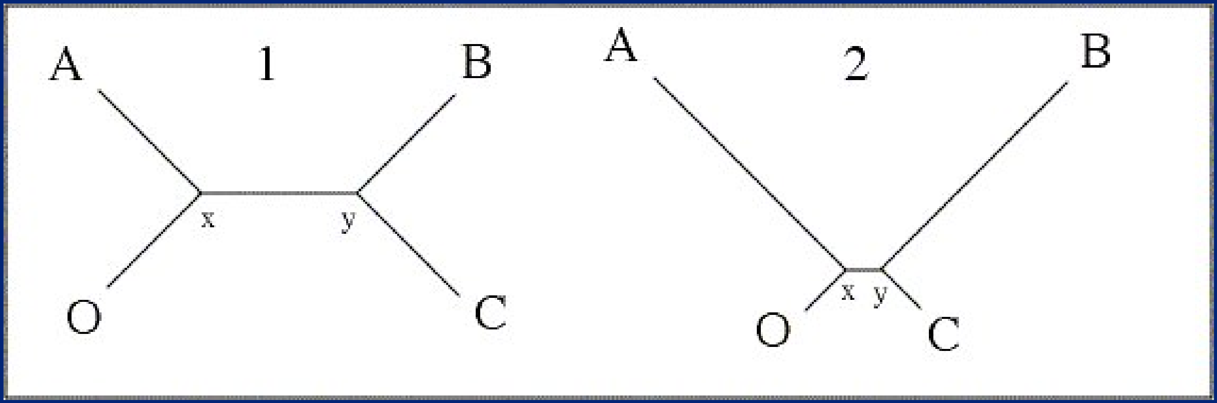


Taxon	1	2	3	4	5
A	A	A	C	G	A
B	T	A	A	T	T
C	T	A	A	G	A
D	A	C	C	G	T

How many changes are needed on each tree topology?

Which topology is preferred by parsimony?

Long branch attraction



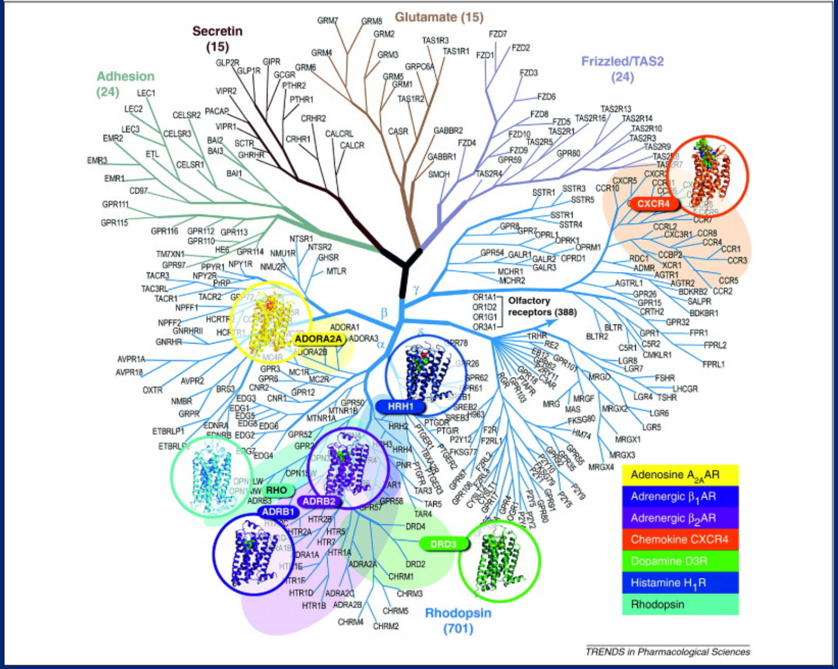
- When 2 changes on a long branch more probable than 1 change on a short branch, parsimony tends to group the long branches together

- This bias gets worse the more data you have (“inconsistency”)

- Discovered by Joe Felsenstein in this department

Betting on your trees

- Ken Rice spent years making parsimony trees of G-protein coupled receptors
  - Maximum likelihood too slow
  - Distance methods didn't perform well
- If new gene groups with:
  - Odor receptors – ignore
  - Neurotransmitter receptors – spend $2K to validate



G-protein coupled receptor genes

Distance methods

- Transform data into a table of pairwise distances
- Find a tree which fits these distances well
- Different distance methods use different fitting criteria

	Human	Bonobo	Chimp	Gorilla	Orang
Human	–	4	5	8	12
Bonobo	4	–	1	9	14
Chimp	5	1	–	8	14
Gorilla	8	9	8	–	13
Orang	12	14	14	13	–

Distance methods

- For very sparse mutations, counting differences may be good enough
- If some sites have mutated multiple times, this will undercount changes on the longer branches
- Use a mutational model to correct the distances
- Various models available:
  - Transition/transversion bias
  - Unequal base frequencies
  - Rate variation
  - Invariant sites

UPGMA

- UPGMA (Unweighted Pair-Group Method of Analysis) is a simple distance method

- Seldom used today:
  - Assumes a molecular clock
  - Behaves badly if clock assumption violated

- Neighbor-joining is a non-clock version that is widely used:
  - Very fast
  - Allows use of a sophisticated mutation model

- UPGMA demonstrates the idea of distance methods in a simple way

UPGMA rules

- Group together the two most similar species
- Divide their distance evenly across the branches leading to them
- Average their distances to all other species
- Rewrite the distance matrix with the new group and distances
- Repeat until tree is finished
- In case of ties, break arbitrarily or draw as three-way split

UPGMA example

	A	B	C	D	E
A	-	5	1	8	9
B	5	-	4	10	11
C	1	4	-	9	9
D	8	10	9	-	2
E	9	11	9	2	-

UPGMA example

	A	B	C	D	E
A	-	5	1	8	9
B	5	-	4	10	11
C	1	4	-	9	9
D	8	10	9	-	2
E	9	11	9	2	-

Group A and C to form AC, with branches of length 0.5

	AC	B	D	E
AC	-	4.5	8.5	9
B	4.5	-	10	11
D	8.5	10	-	2
E	9	11	2	-

UPGMA example

	AC	B	D	E
AC	-	4.5	8.5	9
B	4.5	-	10	11
D	8.5	10	-	2
E	9	11	2	-

Group D and E to form DE, with branches of length 1.0

	AC	B	DE
AC	-	4.5	8.75
B	4.5	-	10.5
DE	8.75	10.5	-

UPGMA example

	AC	B	DE
AC	-	4.5	8.75
B	4.5	-	10.5
DE	8.75	10.5	-

Group B with AC to form ABC, with branches of length 2.25

	ABC	DE
ABC	-	9.625
DE	9.625	-

UPGMA example

	ABC	DE
ABC	-	9.625
DE	9.625	-

Group ABC with DE, with branches of length 4.80

Two hazards of phylogeny

- Garbage in, garbage out:
  - Long pieces of autosomal DNA
  - Misaligned sequences
  - Non-homologous traits
- Gene tree not necessarily the same as the species tree
  - Paralogs
  - Incomplete lineage sorting (ancestral polymorphism)
  - Horizontal gene transfer

Wednesday

- More tree inference:
  - Likelihood methods
  - Bayesian methods
- Hazards of phylogeny inference
- How to validate a phylogeny