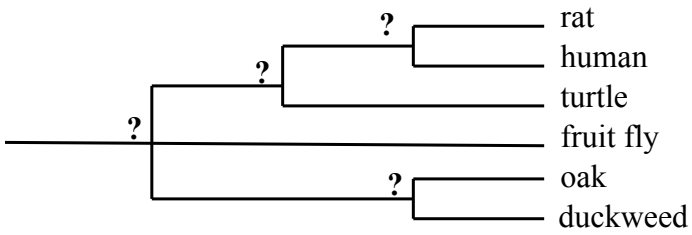
"[M]olecular phylogenists will have failed to find the 'true tree,' not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree."
-- W. F. Doolittle

Trees are *hypotheses* about evolutionary history

So far, we've looked at *understanding* and *formulating* these hypotheses. Now, let's turn our attention to *testing* them.

How confident are we about the inferred phylogeny?



Reliability of Phylogenetic Methods

- Phylogenetic methods can be evaluated in terms of their general performance, particularly their:
  - consistency - approach the truth with more data
  - efficiency - how quickly can they handle how much data
  - robustness - how sensitive to violations of assumptions
- Studies of these properties can be analytical or by simulation

But these are not tests of the data/biology!

Assessing tree reliability

Phylogenetic reconstruction is a problem of statistical inference. One must assess the reliability of the inferred phylogeny and its component parts.

Questions:

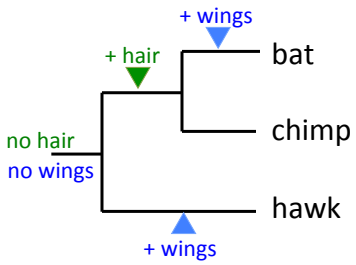
(1) how reliable is the tree?
(2) which parts of the tree are reliable?
(3) is this tree significantly better than another one?

Problems and Errors in Phylogenetic Reconstruction

- Inherent strengths and weaknesses in different tree-making methodologies:
- More is better: Errors in inferred phylogeny may be caused by small data sets and/or limited sampling.
- Unsuitable sequences: those undergoing rapid nucleotide changes or slow to zero changes overtime may skew phylogenetic estimations
- Mutations: Duplications, inversions, insertions, deletions etc. can give inaccurate signals
- Genomic hotspots: small regions of rapid evolution are not easily detected
- Homoplasy: nucleotide changes that are similar but occurred independently in separate lineages are mistakenly assumed as inherited changes
- Sample contamination / mislabeling: always a possibility when working with large data sets

Phylogenetic concepts: Homology and Homoplasy

	Hair?	Wings?
Bat	✓	✓
Chimp	✓	✗
Hawk	✗	✓



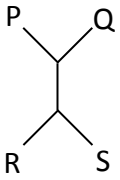
Homology: identity due to shared ancestry (evolutionary signal)

Homoplasy: identity despite separate ancestry (evolutionary noise)

Tree Testing

Let's study the following four sequences:

```
P. A C A T A C G
Q. G T A T A C G
R. G C A C A T G
S. G C A C A C A
```



How can we explain the indicated character?

1. Homology: Changed just once.
2. Homoplasy: Changed twice or more.

Homology more likely, but homoplasy still feasible.

Tree Testing

Now let's look at four other sequences:

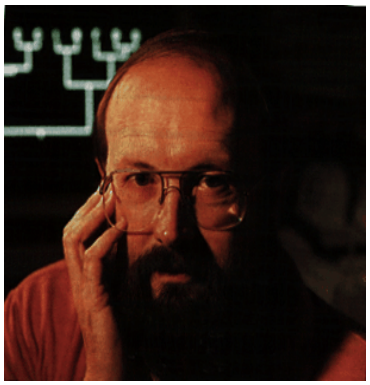
W.	A	C	A	T	G	T	C	A	G	A	C	G
X.	G	T	A	T	G	T	C	A	G	A	C	G
Y.	G	C	A	C	A	C	T	G	A	A	T	G
Z.	G	C	A	C	A	C	T	G	A	A	C	A

Same two explanations possible.
Any changes to their relative likelihood?

Homology ***much*** more likely; homoplasy implausible.

So more evolution leads to stronger signature, but long branch lengths are problematic.

With long branches most methods may yield erroneous trees. For example, the maximum-parsimony method tends to cluster long branches together. This phenomenon is called **long-branch attraction** or the **Felsenstein zone**

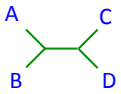


Tree Testing

Basic principle:

Long branches → Strong evolutionary signal, but trees are often harder to construct.

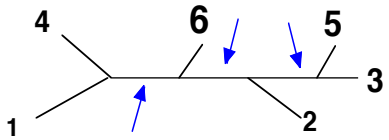
Short branches → Weak evolutionary signal, but trees are often easier to construct correctly.



Zero-length branches → NO evolutionary signal

How much confidence do we have in the branch lengths?

The phylogenetic information expressed by an unrooted tree resides entirely in its internal branches.

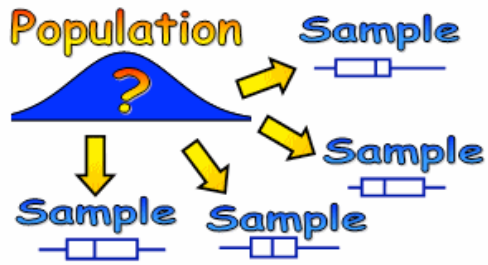


internal branch separating group 1+4 from group 6+2+5+3

- The tree shape can be deduced from the list of its internal branches.
- Testing the reliability of a tree = testing the reliability of each internal branch.

Bootstrapping

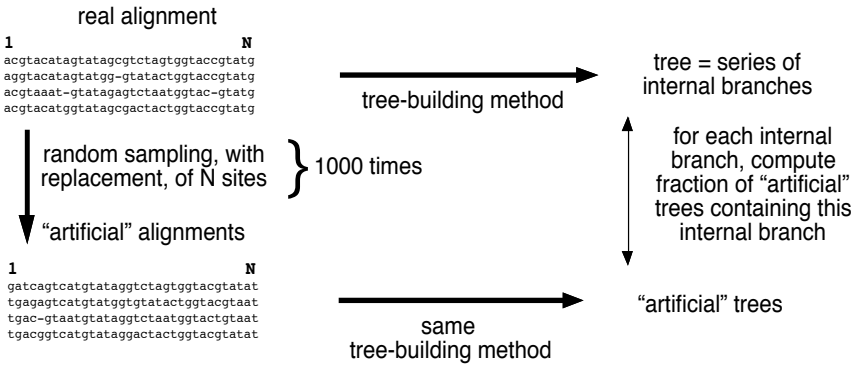
- A statistical technique that uses intensive random resampling of data to estimate a statistic whose underlying distribution is unknown.



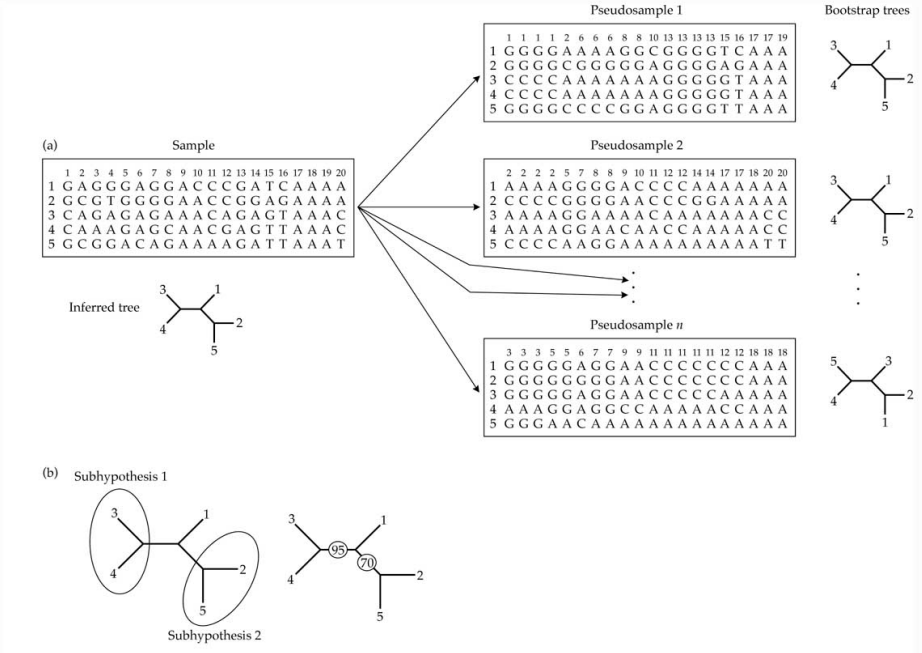
Bootstrapping

- Characters are resampled with replacement to create many bootstrap replicate data sets (pseudosamples)
- Each bootstrap replicate data set is analyzed
- Frequency of occurrence of a group (bootstrap proportions) is a measure of support for the group

Bootstrap procedure

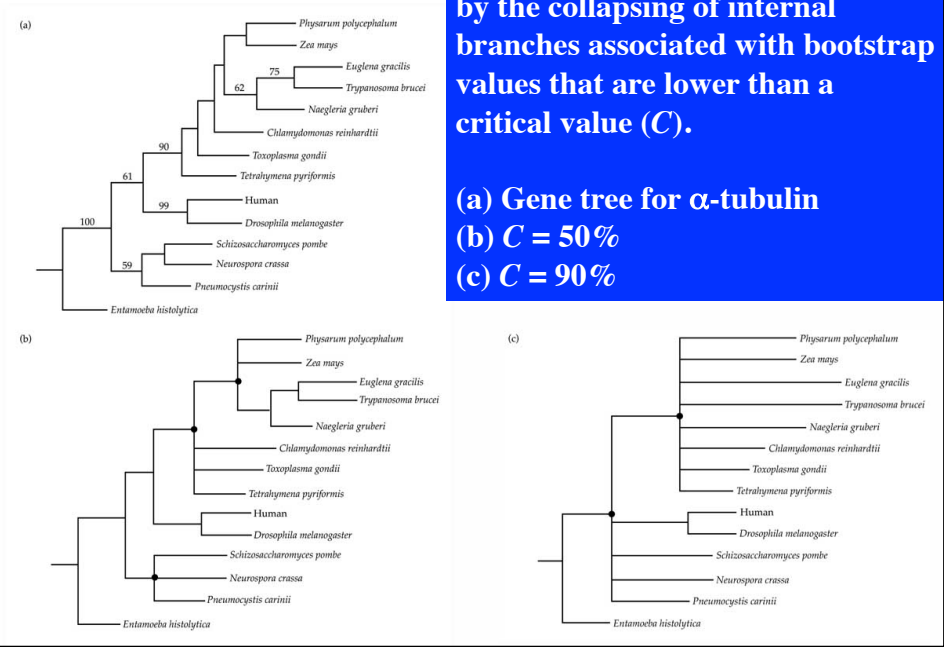


The support of each internal branch is expressed as percent of replicates.



Bootstrap procedure : properties

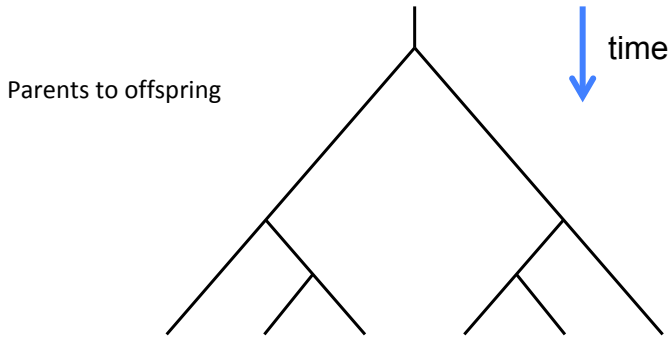
- Internal branches supported by ≥ 90% of replicates are considered as statistically significant. (notice this is another arbitrary cutoff!)
- The bootstrap procedure only detects if sequence length is enough to support a particular node.
- The bootstrap procedure does not help determining if the tree-building method is good. A wrong tree can have 100 % bootstrap support for all its branches!



Reduction of a phylogenetic tree by the collapsing of internal branches associated with bootstrap values that are lower than a critical value (*C*).

(a) Gene tree for α-tubulin
(b) *C* = 50%
(c) *C* = 90%

Homology implies a relationship between sequences, specifically vertical transmission

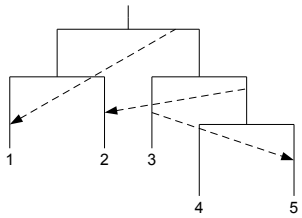


Processes that complicate understanding vertical transmission: horizontal gene transfer, homoplasy and gene duplication

Possible pitfall in reconstruction: Misleading selection of sequences

- Gene/protein sequences can be homologous for several different reasons:

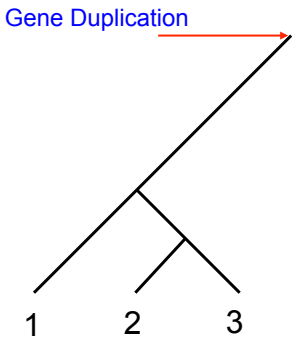
- ☺ **Orthologs** -- sequences diverged after a speciation event
- ☹ **Paralogs** -- sequences diverged after a duplication event (next slides)
- ☹ **Xenologs** -- sequences diverged after a horizontal transfer (e.g., by virus)



Misleading selection of sequences: Using paralogs instead of orthologs

Consider evolutionary tree of three taxa:

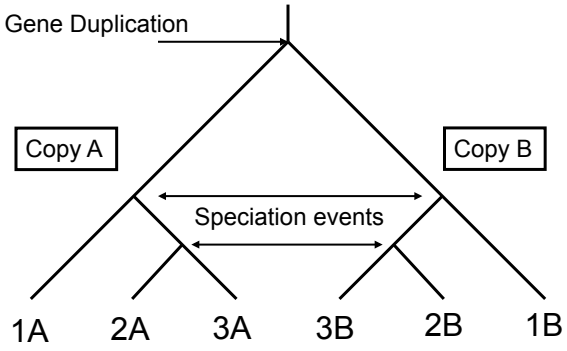
…and assume that at some point in the past a gene duplication event occurred.



21

Paralogs instead of Orthologs

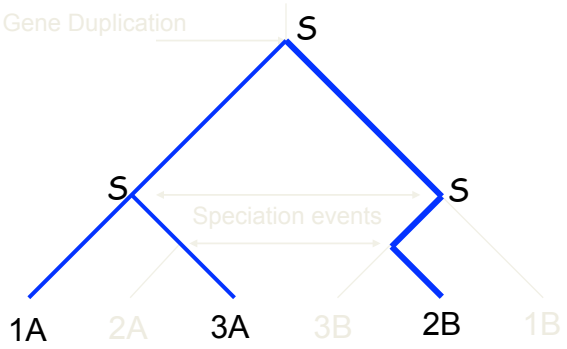
The gene evolution is described by this tree
(1,2,3 are species; *A, B* are the copies of the same gene).



22

Paralogs instead of Orthologs

If we happen to consider genes 1A, 2B, and 3A of species 1,2,3, we get a wrong tree.

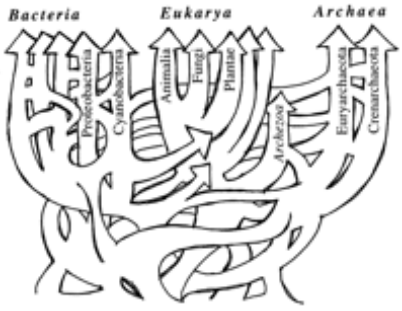


In the sequel we assume all given sequences are orthologs – created from a common ancestor by specification events.

23

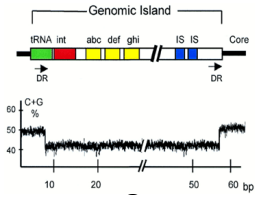
Horizontal Gene Transfer

- The movement of genetic material between two organisms. Once incorporated it is then 'vertically' inherited.

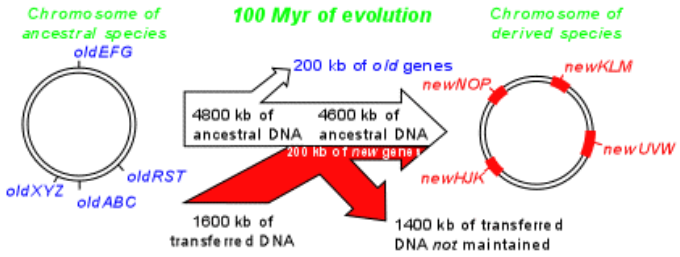


Detecting HGT from genomes: atypical nt composition

- Recent transfers often have unique signature
- "Molecular archaeology" of *E. coli*
  Ochman, 1998)
  - 17.6% HGT, but amelioration since
  - Calculated age distribution based on of divergence

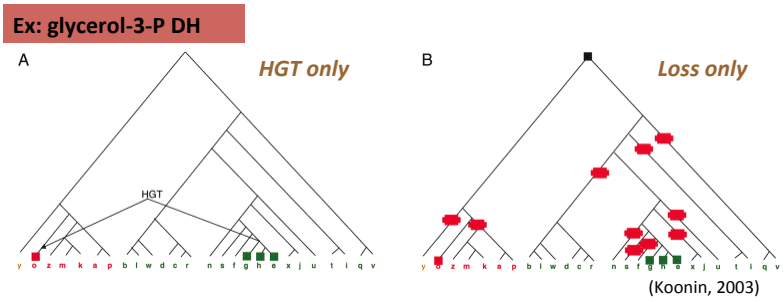


(Hacker & Carniel, 2001)



Detecting HGT: incongruent phylogeny/ synteny

- Any incongruent phylogeny could be explained by HGT or independent gene loss

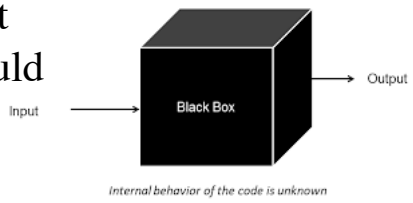


(Koonin, 2003)

General phylogeny recommendations

Avoid the "Black Box"

- Researchers invest considerable resources in producing molecular sequence data.
- They should also invest the time and effort needed to get the most out of their data.
- Modern phylogenetic software makes it easy to produce trees from aligned sequences, but phylogenetic inference should not be treated as a "black box."



Choices are Unavoidable

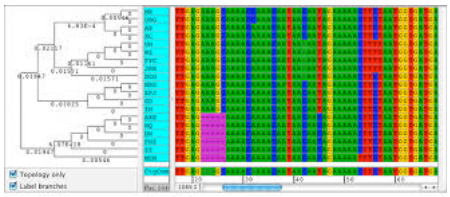
- There are many phylogenetic methods.
- Thus, the investigator is confronted with unavoidable choices.
- Not all methods are equally good for all data.
- An understanding of the basic properties of the various phylogenetic methods is essential for informed choice of method and interpretation of results.

Data are not Perfect

Shelley Long

- Most data includes misleading evidence, and we need to have a cautious attitude to the quality of data and trees.
- Data may have both systematic biases and unbiased noise that affect our chances of getting the correct tree
- Different methods may be more or less sensitive to some problems.

Alignment

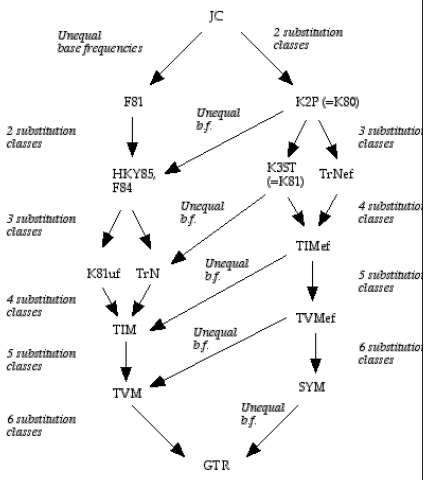


- The alignment is critical to the tree method.
- Be aware of alignment artefacts.
- If using multiple alignment software, explore the sensitivity of the alignment to the parameters used.
- Eliminate regions that cannot be aligned with confidence.

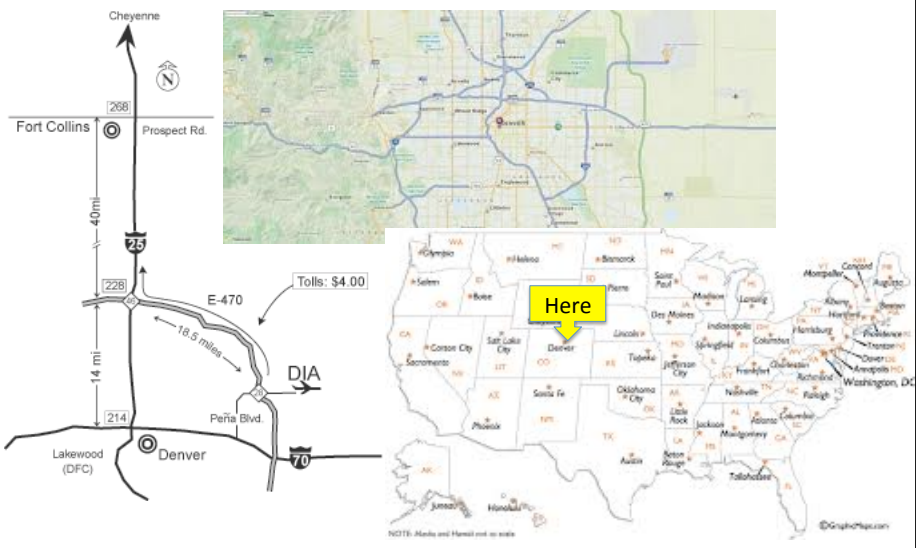
Choice of Models

- Complex models may better approximate the evolution of the sequences and, therefore, might be expected to give more accurate results.
- More complex models require the estimation of more parameters each of which is subject to some error.
- There is a trade-off between more realistic and complex models and their power to discriminate between alternative hypotheses.

$$Q = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix}$$



An analogy


Cheyenne
Fort Collins
Prospect Rd.
268
25
40mi
228
E-470
Tolls: $4.00
18.5 miles
DIA
14 mi
214
Peña Blvd.
Lakewood
(DFC)
Denver
70
Here