MCB 5472

Intro to Trees

Peter Gogarten
Office: *BSP 404*
phone: *860 486-4061,*
Email: *gogarten@uconn.edu*

OLD ASSIGNMENTS

Turn your script that calculates the reverse complement of a sequence into a subroutine

```
use warnings;
#modified from Dustin
MAIN: {
        my $seq='';
        my $revcomp='';
        print "Please enter sequence to be reverse complemented\n";
        chomp($seq=<>);
        $revcomp=&Reverse(&Complement($seq)); #Next line would work just as well
        # $revcomp=Reverse(Complement($seq));
        print "\n";
        print "The reverse complement of \$seq=$seq"." is"." $revcomp\n\n";
exit;}

sub Reverse {
        use strict;
        use warnings;
        my $string = $_[0];
        my $rev= reverse $string;
        return $rev;
}

sub Complement{
        # you can use a local variable with the same name as in the main program e.g.
        my $seq= $_[0];
        $seq =~ tr/ATGCYRatcgyr/TACGRYTAGCRY/;
        return ($seq);
        #$seq in the calling program is not impacted.

        #my$comp= $_[0];
        #$comp =~ tr/ATGCYRatcgyr/TACGRYTAGCRY/;
        #return $comp;
```

OLD ASSIGNMENTS

Write a script that takes all files with the extension .fa (containing a single fasta formated sequence) and writes their contents in a single multiple sequence file.

Solution 1:

```
use warnings;
system ("cat *.fa > all_fa_files.faa");
```

OLD ASSIGNMENTS

Solution 2:

```
use warnings;
open (OUT, ">temp.out") || die "cannot open: $!";#open a file for output -
# as the names of the infiles have not been read yet, I assign this a temporary name to be renamed later
#
while(defined($file=glob("*.fa"))){  #loop opens one file ending in .fa after the other. stores the name in $file
        open (IN, "<$file") || die "cannot open: $!"; #reports error if file cannot be opened
        #and assigns $file to IN
        @filename_parts=split(/\./,$file); # writes first part of filename into $filename_parts[0]
        $outname .= "$filename_parts[0]".'.'; ### concatenated $filename_part[0] into new outname ###
        #repeats this for every cycle
        while (defined($line=<IN>)){print OUT $line;} #writes contents of $file line by line into filehandle OUT
} #no more *.fa files in directory
close OUT; #closes filehandle OUT and temp.out
$outname="$outname".'fa.multiple';#adds useful extension to outfile
system ("mv temp.out $outname"); # uses unix command mv (move)to rename temp.out into new $outname
```

assignments continued (use class 5 as sample)

Assume that you have the following non-aligned multiple sequence files in a directory:

A.fa : vacuolar/archaeal ATPase catalytic subunits ;
B.fa : vacuolar/archaeal ATPase non-catalytic subunits;
alpha.fa : F-ATPases non-catalytic subunits,
beta.fa : F-ATPases catalytic subunits,
F.fa : ATPase involved in the assembly of the bacterial flagella.

Write a perl script that executes muscle or clustalw and
1) aligns the sequences within each file
2) successively calculates profile alignments between all aligned sequences.

Hints:

`system (command);#` **executes "command" as if you had typed** `command` **in the command line**

First loop: Calculate multiple alignments for all *.fa files

```
#!/usr/bin/perl -w

# This program aligns all multiple sequence files with names *.fa
# found in its directory using clustalw.  It then calculates a profile
# alignemnt to which the individual alignments are added succesively.

$counter=0;
# sets the counter to zero, so that the program can do something
# different when running through the loop first (SEE SECOND LOOP BELOW)

while(defined($file=glob("*.fa"))){

    @parts=split(/\./,$file);

#   the last line splits the file name at the period and assigns the parts
#   to an array @parts ($parts[0], $parts[1] ...). The next line assigns only
#   the first part to $file and overwrites the earlier assignment.
#   Not really neccessary, but it avoids an accumulation of file type extensions.

    $file=$parts[0];
    system("clustalw2 -infile=$file.fa -align");
};
```

This results in one *.aln file per *.fa file

Take *.aln files and add them successively to growing alignment

```
while (defined($filename=glob("*.aln"))){

    @parts=split(/\./, $filename);

    $filename=$parts[0];

#   the first time through the loop, we don't run a profile alignemnt,
#   we only assign the first aln file to the file where we store the results
#   (results.out), we use mv to get rid of the aln file.
#   to which the next aln file will be aligned.

    if ($counter==0) {
                system ("mv $filename.aln result.out");
        }
    else {
        system("clustalw2 -profile1=result.out -profile2=$filename.aln -outfile=result.out -profile -align");
                }
        $counter=$counter+1;
#       print "$counter\n";

};

# cleanup:

system ("rm *.dnd");
system ("rm *.aln");
system ("mv result.out result.clusteraln");

exit;
```

challenge:

Often one wants to build families of homologous proteins extracted from genomes. One way to do so is to find reciprocal best hits.
Tools:
The script *blastall.pl* takes the genomes indicted in the first line and calculates all possible genome against genome searches.

This script *simple_rbh_pairs.pl* takes two blastall searches (genome A versus genome B) in -m8 format and listing only the top scoring blast hit for each query) and writes the GI numbers of reciprocal best hits into a table.

The script *run_pairs.pl* runs all possible pairwise extractions of RBHs

Task: write a script that combines the pairwise tables keeping only those families that have a strict reciprocal best blast hit relationship in all genomes.

New assignment

Check chapter U36-U40 in the Unix/Perl primer

Read chapters P17 and P18 on regular expressions
(these are really useful, if you repeatedly run a program that generates complex output, and you want to extract a single value into a table, e.g. a program that calculates theoretical isoelectric points …)

New assignment

Write a script that takes a genome (fna file) and calculates the GC content.

Modify the script to count tetra- and penta-nucleotide frequencies.

Modify the script so that it creates a table that gives the GC, tetra- and penta-nucleotide content in a sliding window moving through the genome.

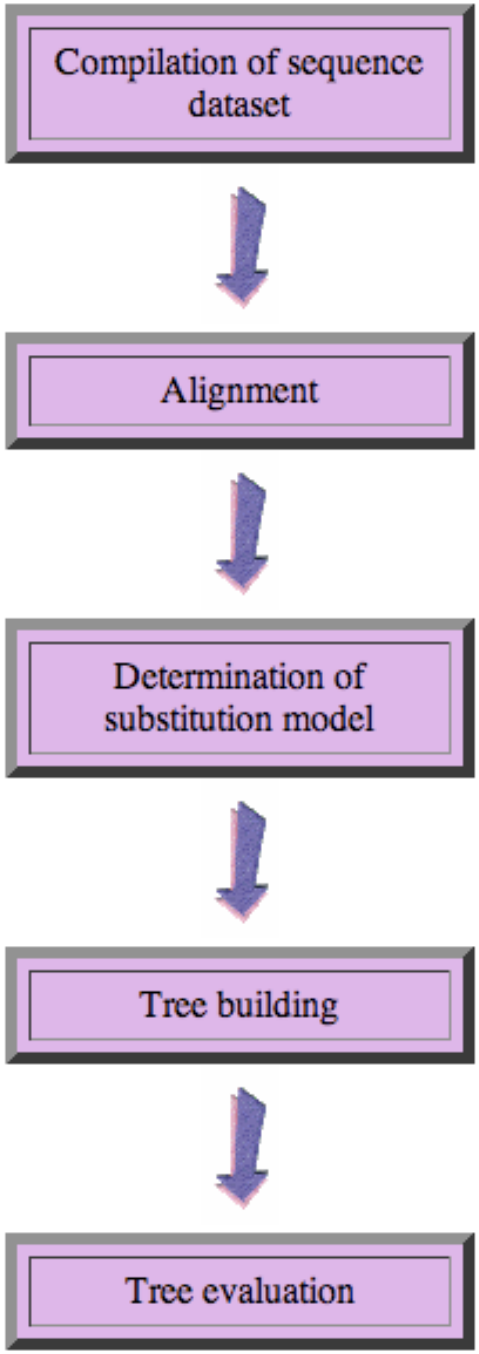
(For which of these programs, and for which problems might you want to consider, or correct for strand bias?)

Steps of the phylogenetic analysis

Phylogenetic analysis is an inference of evolutionary relationships between organisms. Phylogenetics tries to answer the question "How did groups of organisms come into existence?"

Those relationships are usually represented by tree-like diagrams.

Note: the assumption of a tree-like process of evolution is controversial!



more alignment programs: statalign

statalign from Jeff Thorne deserves more attention than it receives. Especially for divergent sequences the initial pairwise alignment usually determines the ultimate result of the phylogenetic reconstruction.

Statalign solves this problem by not calculating a multiple sequence alignment, rather it spends a lot of computational power to calculate pairwise alignments and it extract distances (and their potential error) from these pairwise alignments and then uses these in a distance pased reconstruction. The errors from the individual distances are used to generate bootstrap samples for the distance matrices.

More at *Thorne JL, Kishino H (1992) Freeing phylogenies from artifacts of alignment. Mol Bio Evol 9:1148-1162*

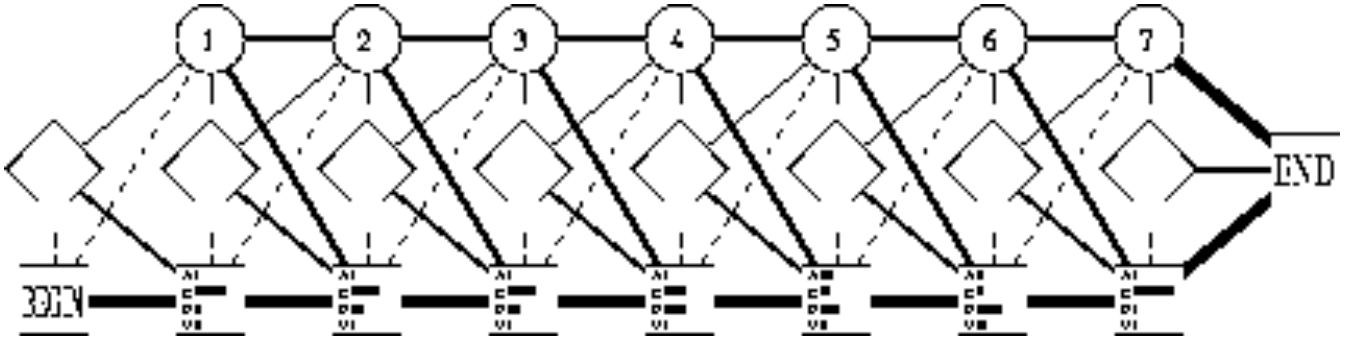
statalign is available in several software archives (e.g. here), the readme file has plenty of information.

more alignment programs: SAM

SAM (sequence alignment and modeling system) by Richard Hughey, Anders Krogh, Christian Barrett, & Leslie Grate at UCSC.

http://www.cse.ucsc.edu/research/compbio/sam.html

The input consists of a multiple sequence file (aligned or not aligned) in FASTA format. The program uses secondary structure predictions, neighboring sites, etc. to place gaps. The program can be accessed through the www and run at UCSC

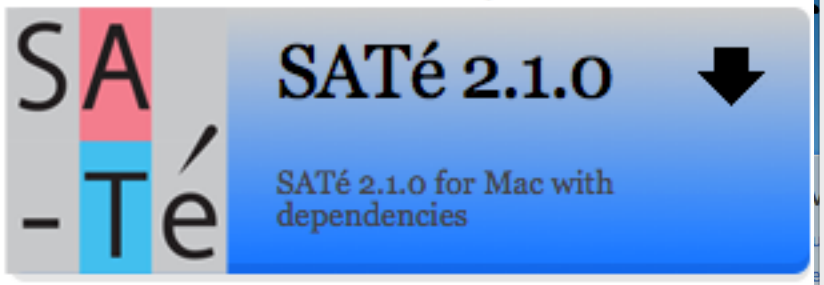


A linear hidden Markov model is a sequence of nodes, each corresponding to a column in a multiple alignment. In our HMMs, each node has a match state (square), insert state (diamond) and delete state (circle). Each sequence uses a series of these states to traverse the model from start to end. Using a match state indicates that the sequence has a character in that column, while using a delete state indicates that the sequence does not. Insert states allow sequences to have additional characters *between* columns. In many ways, these models correspond to profiles.

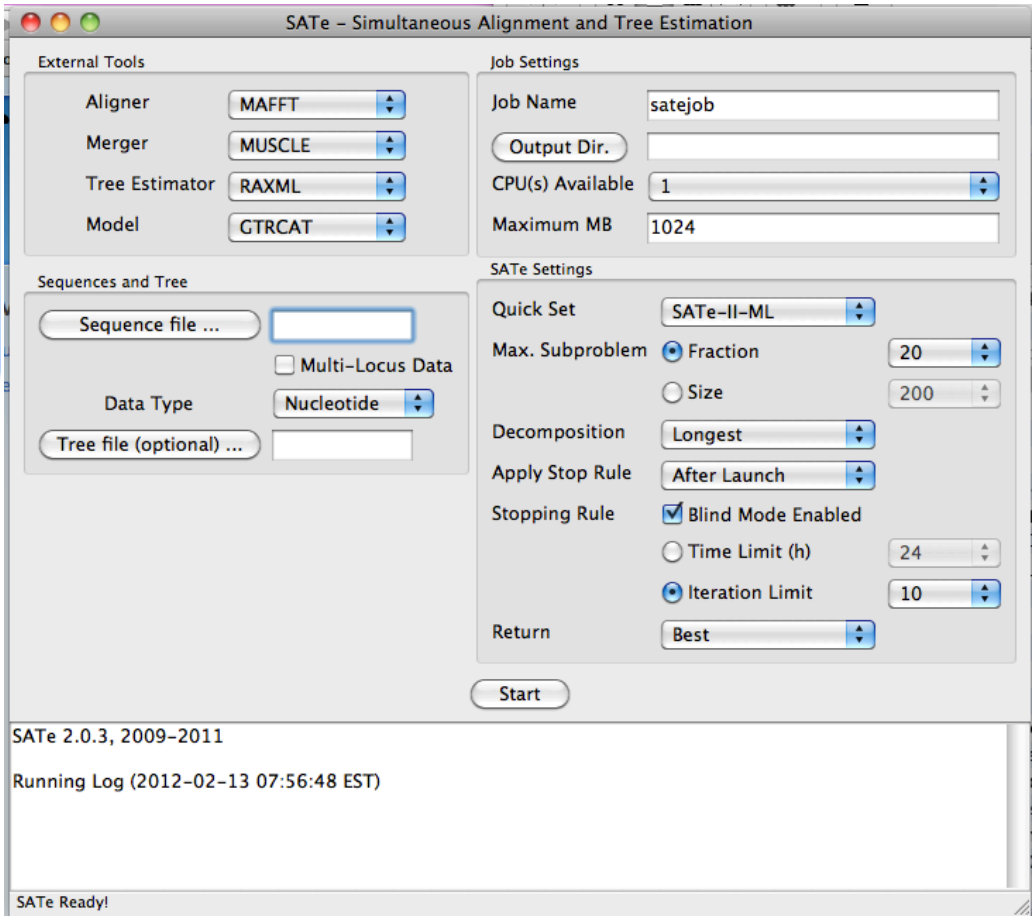
SATé

Simultaneous Alignment and Tree Estimation

http://phylo.bio.ku.edu/software/sate/sate.html



GUI works well on iMacs, but uses only local processors.



Sequence editors and viewers

Jalview Homepage, Description

Jalview is easy to install and run.
Test file is here (ATPase subunits)
(Intro to ATPases: 1bmf in spdbv)
(gif of rotation here)
(Load all.txt into Jalview,
colo**u**r options,
mouse use,
PID tree,
Principle component analysis -> sequence space)
More on sequence space here

Phylogenetic reconstruction - **How**

Distance analyses

calculate pairwise distances
(different distance measures, correction for multiple hits, correction for codon bias)

make distance matrix (table of pairwise corrected distances)

calculate tree from distance matrix

i) using optimality criterion
(e.g.: smallest error between distance matrix
and distances in tree, or use
ii) algorithmic approaches (UPGMA or neighbor joining) B)

Phylogenetic reconstruction - **How**

Parsimony analyses

find that tree that explains sequence data with minimum number of substitutions
(tree includes hypothesis of sequence at each of the nodes)

Maximum Likelihood analyses

given a model for sequence evolution, find the tree that has the highest probability under this model.
This approach can also be used to successively refine the model.

Bayesian statistics use ML analyses to calculate posterior probabilities for trees, clades and evolutionary parameters. Especially MCMC approaches have become very popular in the last year, because they allow to estimate evolutionary parameters (e.g., which site in a virus protein is under positive selection), without assuming that one actually knows the "true" phylogeny.

- For a discussion of Bootstrapping go here http://web.uconn.edu/gogarten/mcb221_2003/class28.html

What is in a tree?

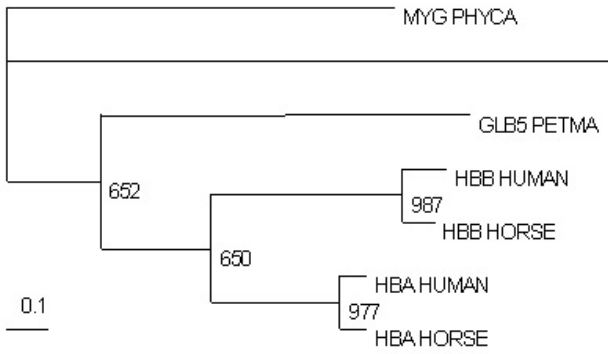
Trees form molecular data are usually calculated as unrooted trees (at least they should be - if they are not this is usually a mistake).
To root a tree you either can assume a molecular clock (substitutions occur at a constant rate, again this assumption is usually not warranted and needs to be tested),
or you can use an **outgroup** (i.e. something that you know forms the deepest branch).

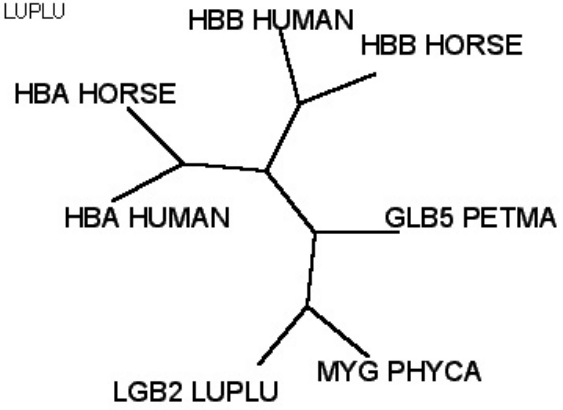
For example, to root a phylogeny of birds, you could use the homologous characters from a reptile as outgroup; to find the root in a tree depicting the relations between different human mitochondria, you could use the mitochondria from chimpanzees or from Neanderthals as an outgroup; to root a phylogeny of alpha hemoglobins you could use a beta hemoglobin sequence, or a myoglobin sequence as outgroup.

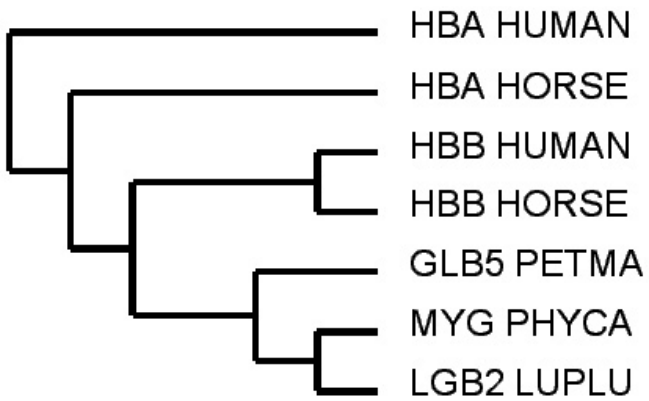
Trees have a branching pattern (also called the **topology**), and **branch lengths**.

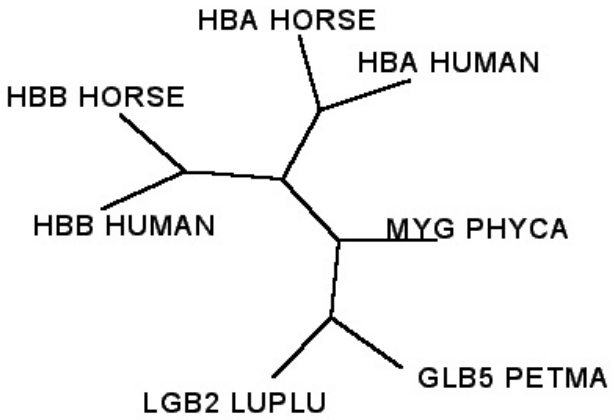
Often the branch lengths are ignored in depicting trees (these trees often are referred to as cladograms - note that cladograms should be considered rooted).
You can swap branches attached to a node, and in an unrooted you can depict the tree as rooted in any branch you like without changing the tree.

Test:Which of these trees is different?









More tests here

Terminology

- Branches, splits, bipartitions
- In a rooted tree: clades (for urooted trees sometimes the term clann is used)
- Mono-, Para-, polyphyletic groups, cladists and a natural taxonomy

The term cladogram refers to a strictly bifurcating diagram, where each clade is defined by a common ancestor that only gives rise to members of this clade. I.e., a clade is monophyletic (derived from one ancestor) as opposed to polyphyletic (derived from many ancestors). (Note: you do need to know where the root is!)

A clade is recognized and defined by shared derived characters (= synapomorphies). Shared primitive characters (= sympleisiomorphies , aternativie spelling is symplesiomorphies) do not define a clade. (see in class example drawing ala Hennig).

To use these terms you need to have polarized characters; for most molecular characters you don't know which state is primitive and which is derived (exceptions:....).

Terminology

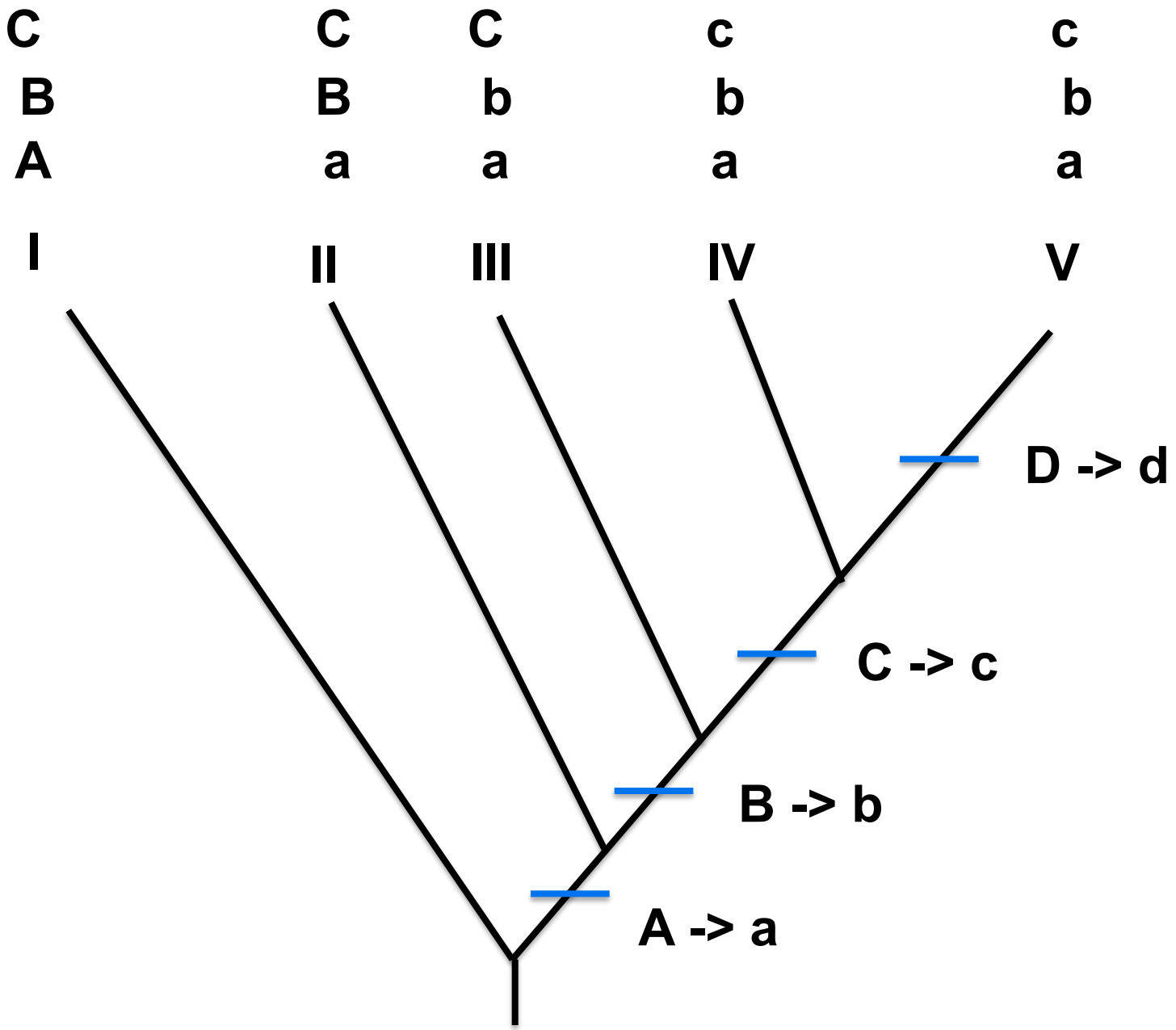
Related terms:

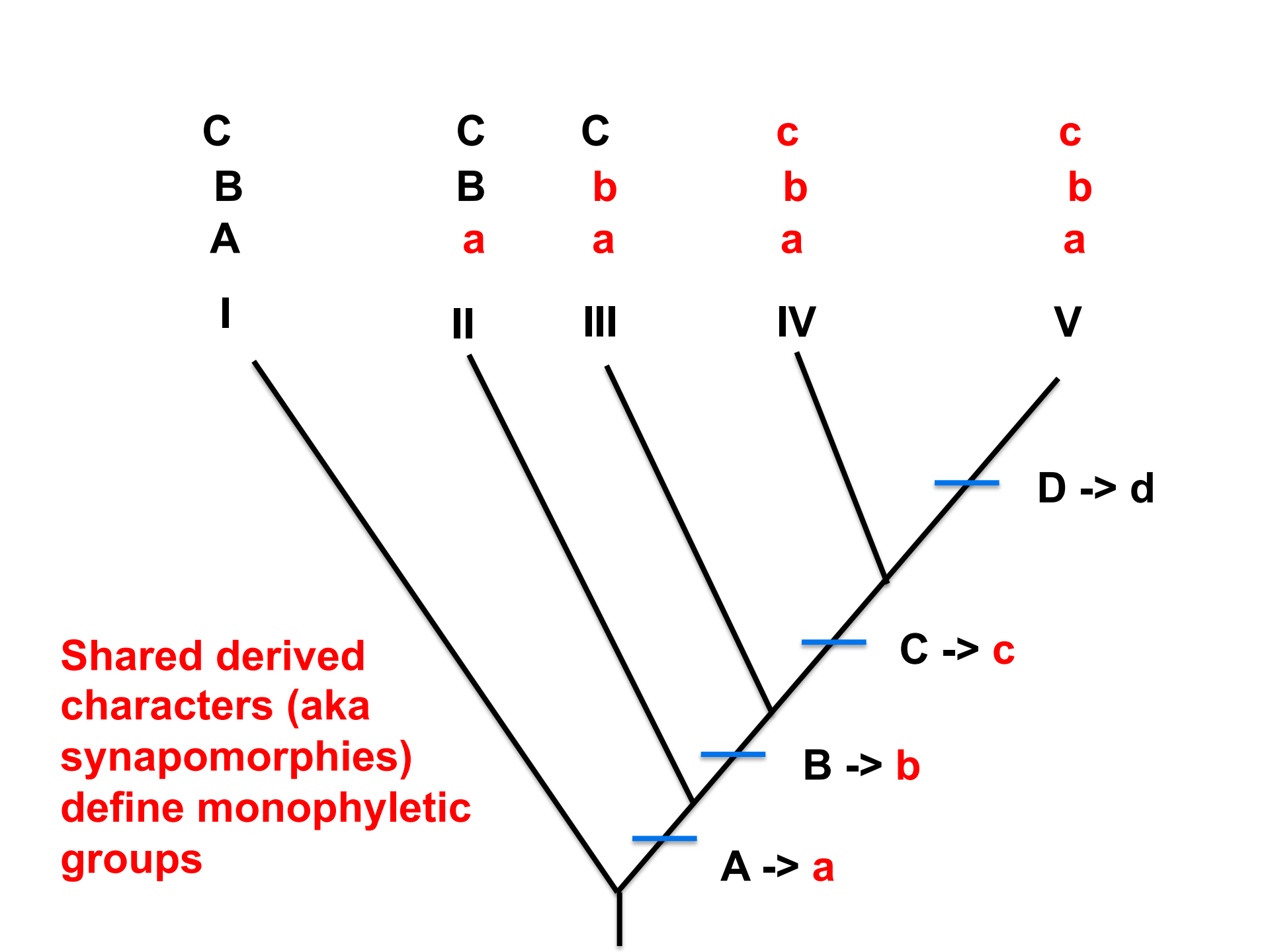
autapomorphy = a derived character that is only present in one group; an autapomorphic character does not tell us anything about the relationship of the group that has this character ot other groups.

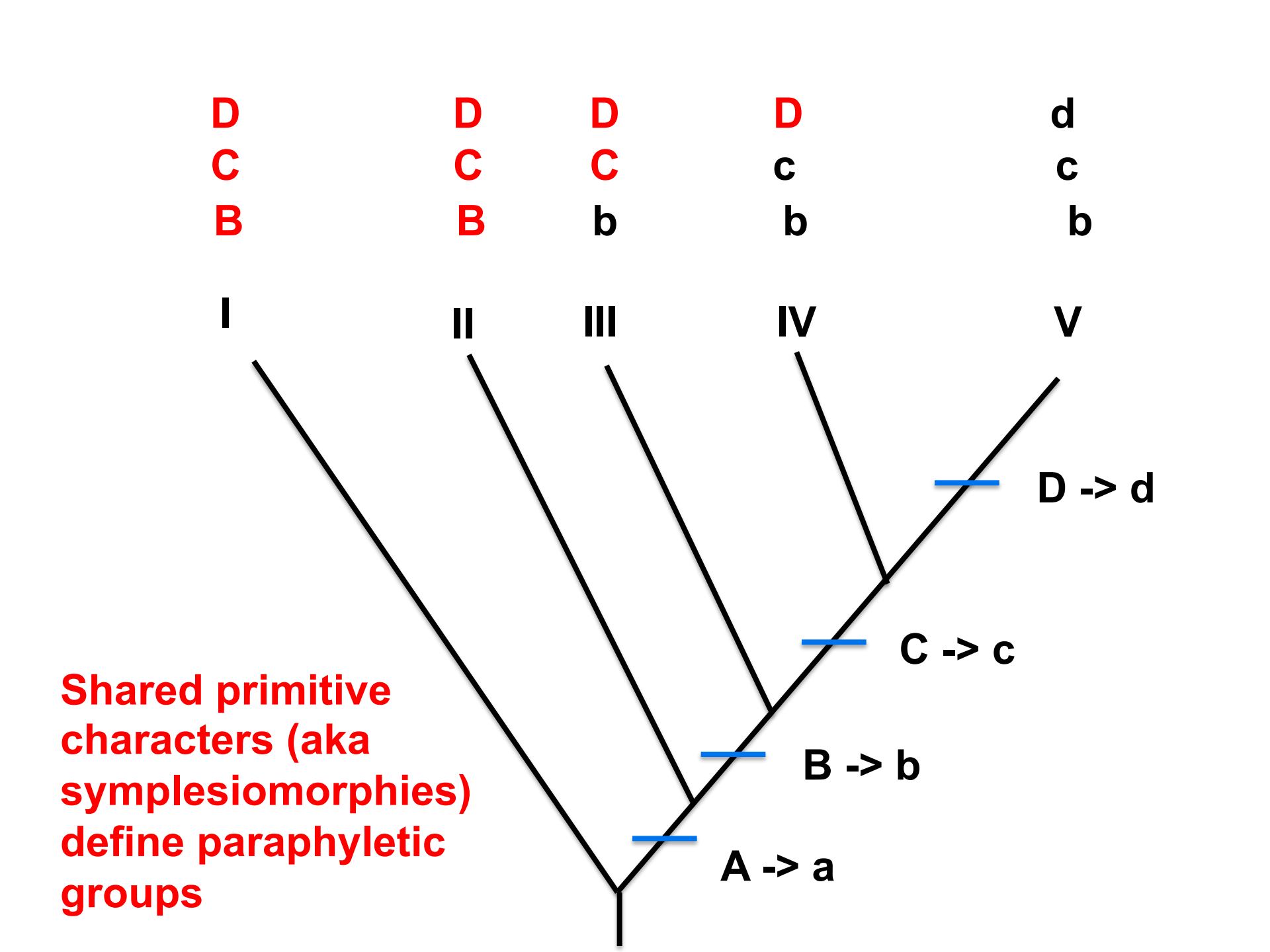
homoplasy = a derived character that was derived twice independently (convergent evolution). *Note that the characters in question might still be homologous (e.g. a position in a sequence alignment, frontlimbs turned into wings in birds and bats).*

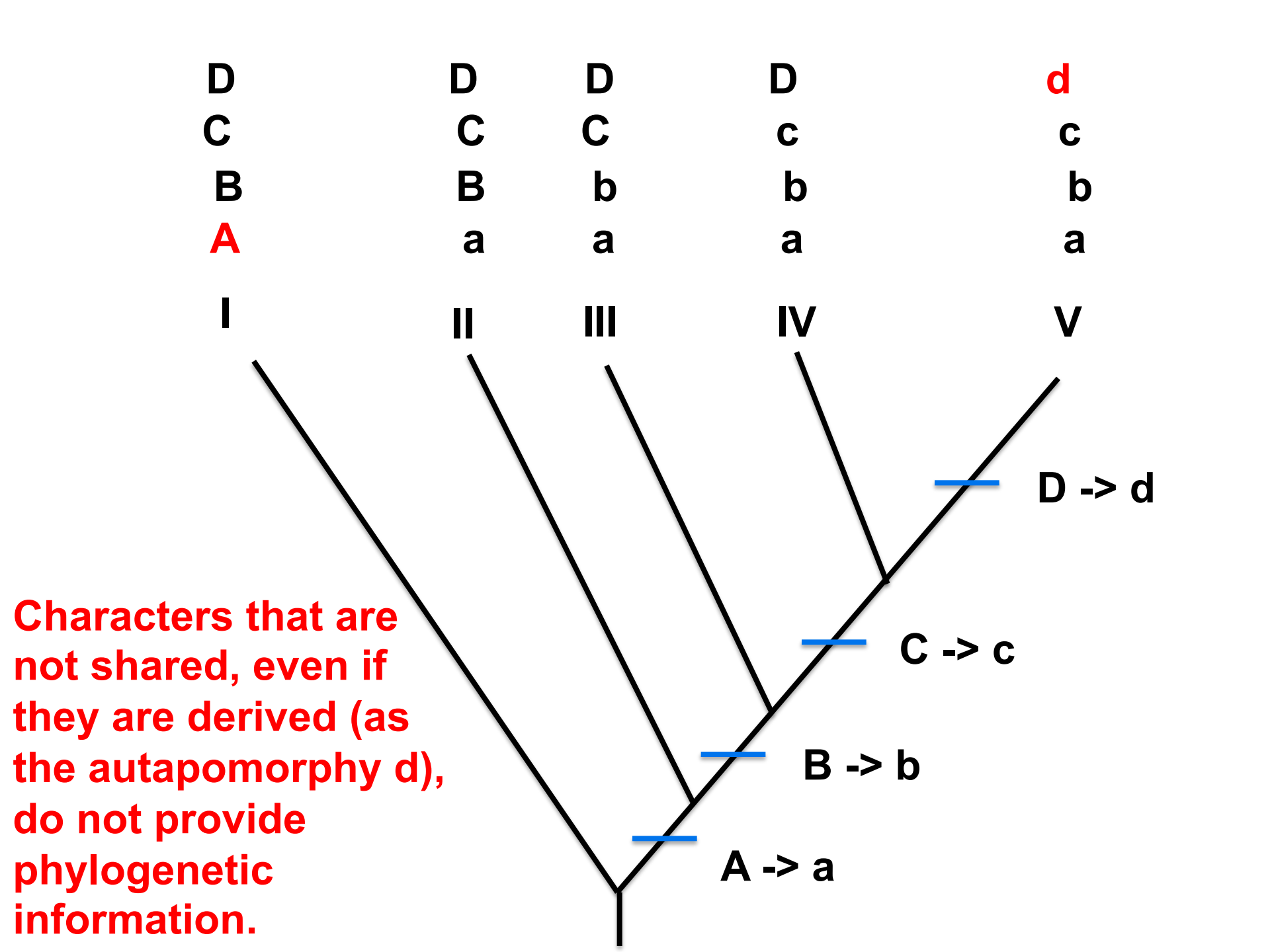
paraphyletic = a taxonomic group that is defined by a common ancestor, however, the common ancestor of this group also has decendants that do not belong to this taxonomic group. Many systematists despise paraphyletic groups (and consider them to be polyphyletic). Examples for paraphyletic groups are reptiles and protists. Many consider the archaea to be paraphyletic as well.

holophyletic = same as above, but the common ancestor gave rise only to members of the group.


C
B
A
I
C
B
a
II
C
b
a
III
c
b
a
IV
c
b
a
V
D -> d
C -> c
B -> b
A -> a


C
B
A
I
C
B
a
II
C
b
a
III
c
b
a
IV
c
b
a
V
D -> d
C -> c
B -> b
A -> a
Shared derived characters (aka synapomorphies) define monophyletic groups


D
C
B
D
C
B
D
C
b
D
c
b
d
c
b
I
II
III
IV
V
D -> d
C -> c
B -> b
A -> a
Shared primitive characters (aka symplesiomorphies) define paraphyletic groups


D
C
B
A
I
D
C
B
a
II
D
C
b
a
III
D
c
b
a
IV
d
c
b
a
V
D -> d
C -> c
B -> b
A -> a
Characters that are not shared, even if they are derived (as the autapomorphy d), do not provide phylogenetic information.

Terminology

Related terms:

autapomorphy = a derived character that is only present in one group; an autapomorphic character does not tell us anything about the relationship of the group that has this character ot other groups.

homoplasy = a derived character that was derived twice independently (convergent evolution). *Note that the characters in question might still be homologous (e.g. a position in a sequence alignment, frontlimbs turned into wings in birds and bats).*

paraphyletic = a taxonomic group that is defined by a common ancestor, however, the common ancestor of this group also has decendants that do not belong to this taxonomic group. Many systematists despise paraphyletic groups (and consider them to be polyphyletic). Examples for paraphyletic groups are reptiles and protists. Many consider the archaea to be paraphyletic as well.

holophyletic = same as above, but the common ancestor gave rise only to members of the group.

homology

Two sequences are homologous, if there existed an ancestral molecule in the past that is ancestral to both of the sequences

Types of Homology

Orthologs: "deepest" bifurcation in molecular tree reflects speciation.
These are the molecules people interested in the taxonomic classification of organisms want to study.

Paralogs: "deepest" bifurcation in molecular tree reflects gene duplication. The study of paralogs and their distribution in genomes provides clues on the way genomes evolved. Gen and genome duplication have emerged as the most important pathway to molecular innovation, including the evolution of developmental pathways.

Xenologs: gene was obtained by organism through horizontal transfer. The classic example for Xenologs are antibiotic resistance genes, but the history of many other molecules also fits into this category: inteins, selfsplicing introns, transposable elements, ion pumps, other transporters,

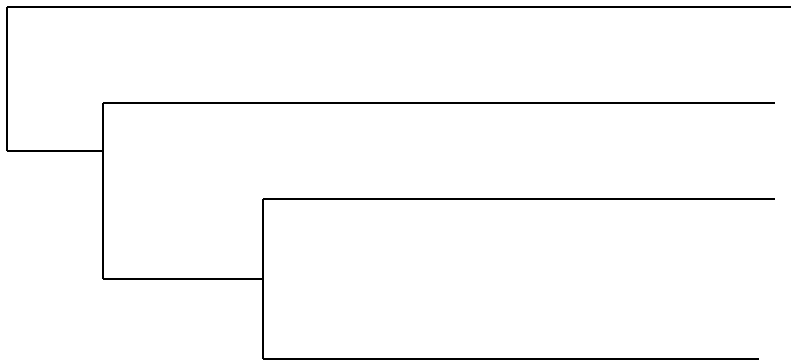
Synologs: genes ended up in one organism through fusion of lineages. The paradigm are genes that were transferred into the eukaryotic cell together with the endosymbionts that evolved into mitochondria and plastids
(the -logs are often spelled with "ue" like in orthologues)

see Fitch's article in TIG 2000 for more discussion.

Trees – what might they mean?

Calculating a tree is comparatively easy, figuring out what it might mean is much more difficult.

If this is the probable organismal tree:



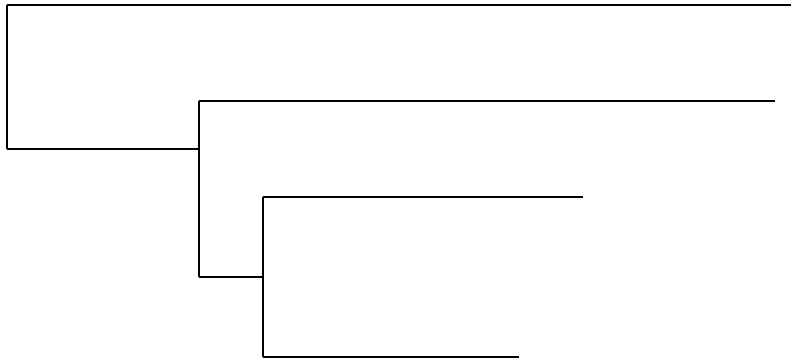
species A

species B

species C

species D

what could be the reason for obtaining this gene tree:



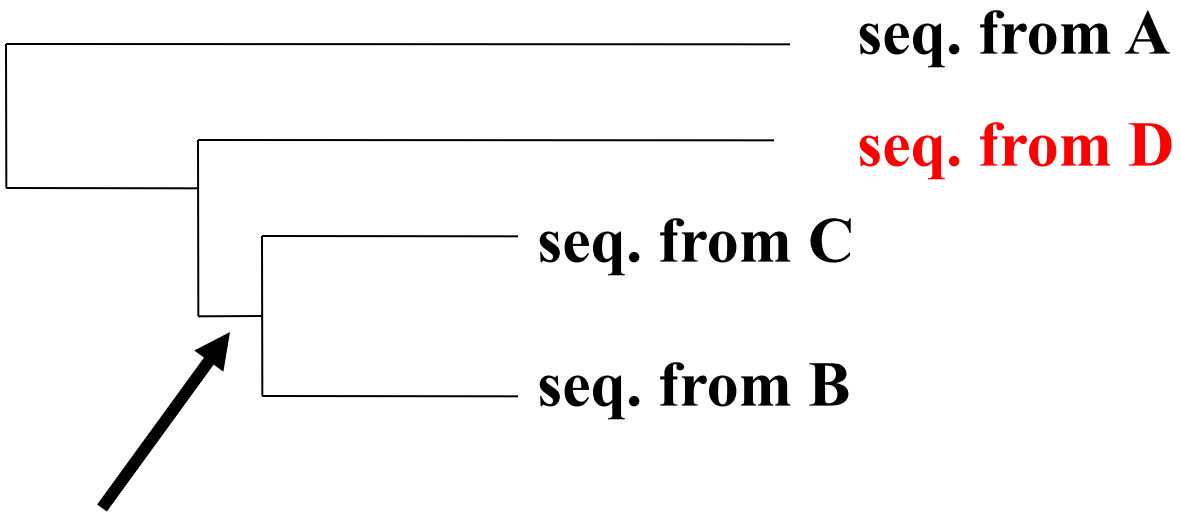
seq. from A

seq. from D

seq. from C

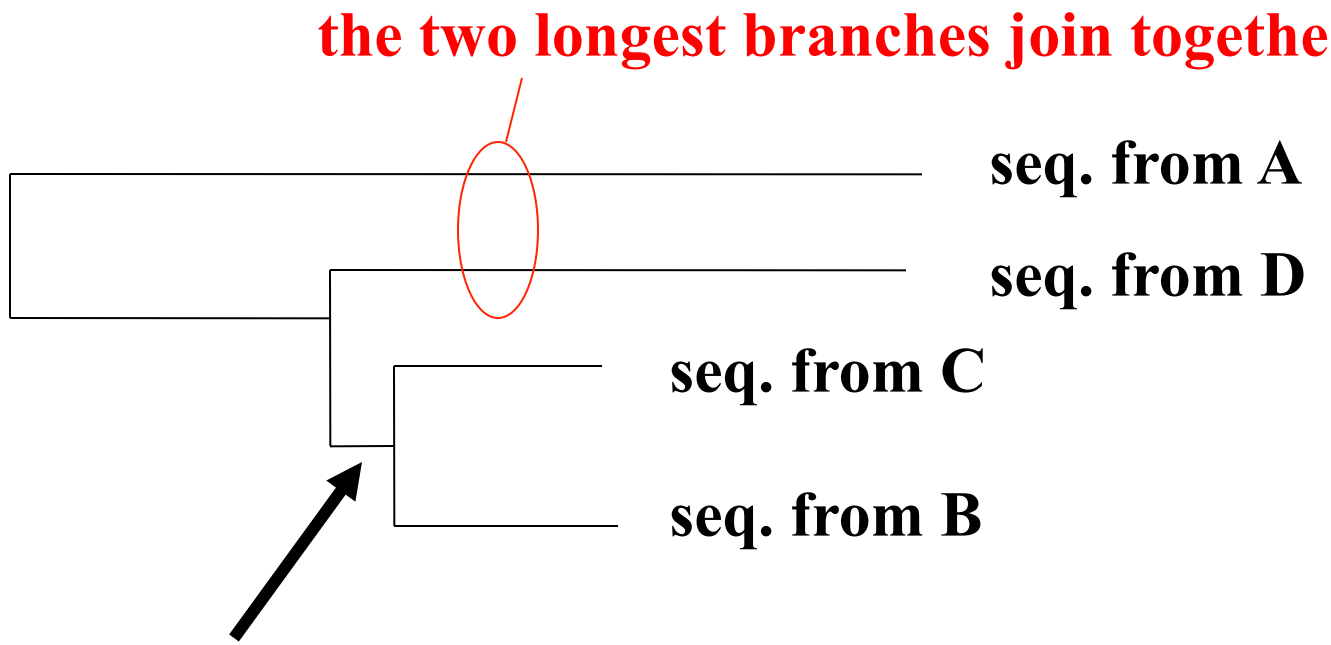
seq. from B

lack of resolution



e.g., 60% bootstrap support for bipartition (AD)(CB)

long branch attraction artifact



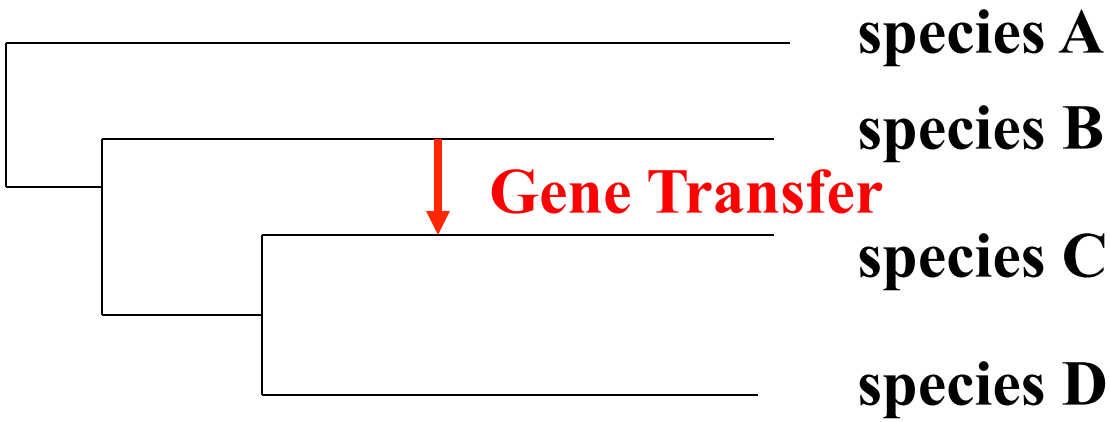
e.g., 100% bootstrap support for bipartition (AD)(CB)

What could you do to investigate if this is a possible explanation?

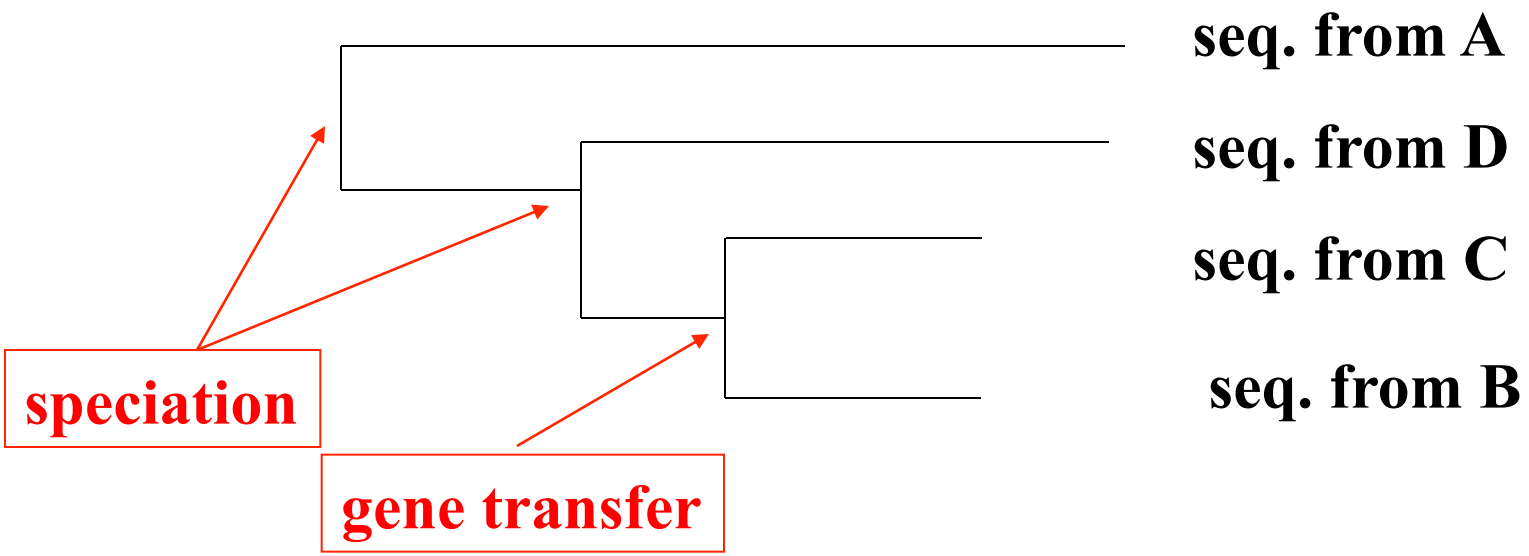
- **use only slow positions,**
- **use an algorithm that corrects for ASRV**
- **add sequences that break-up the long branches**

Gene transfer

Organismal tree:

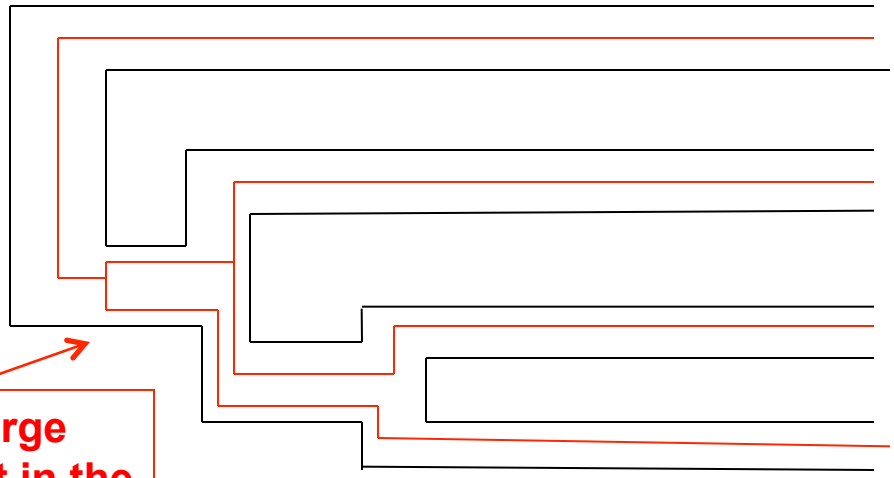


molecular tree:



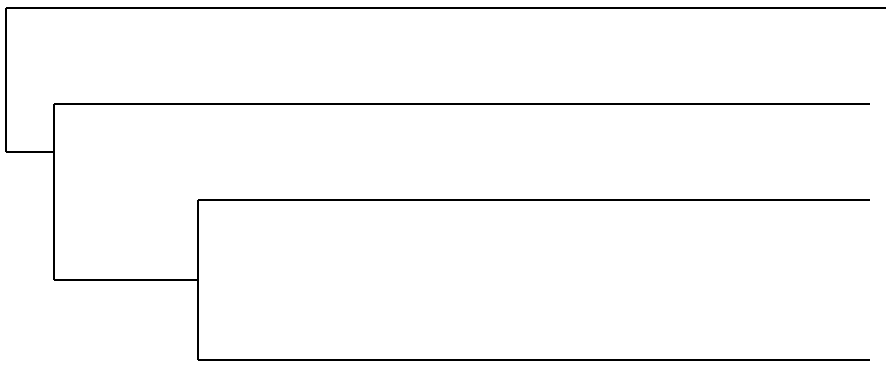
Lineage Sorting

Organismal tree:



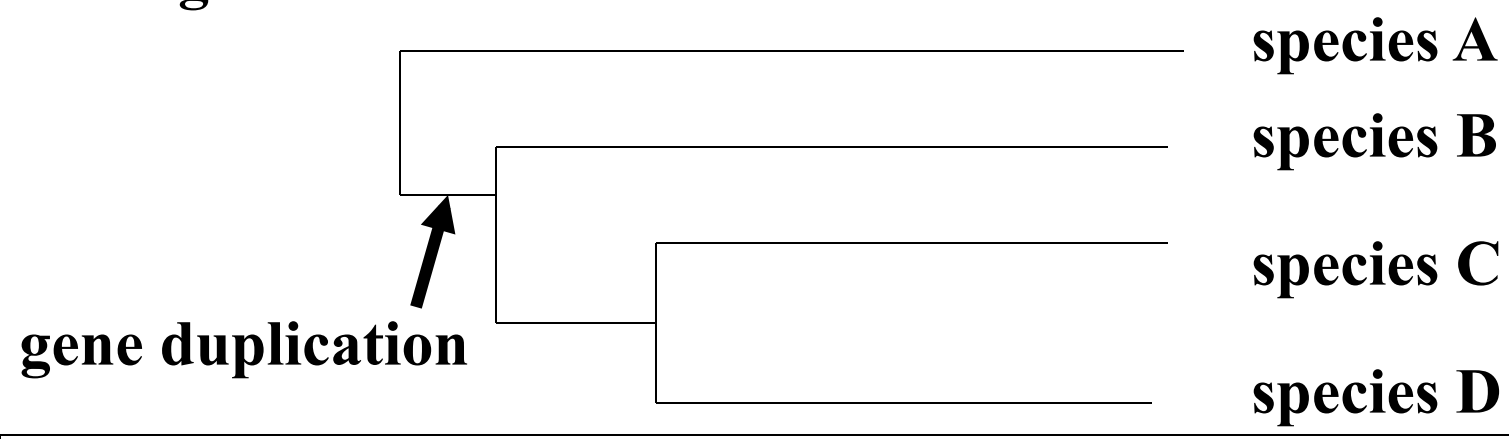
Genes diverge and coexist in the organismal lineage

molecular tree:



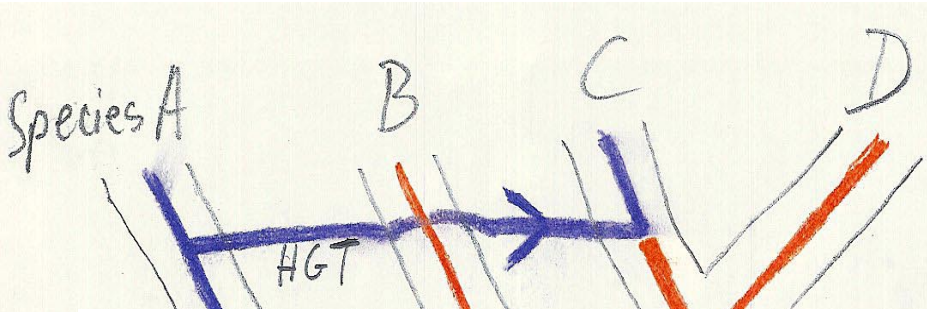
Gene duplication

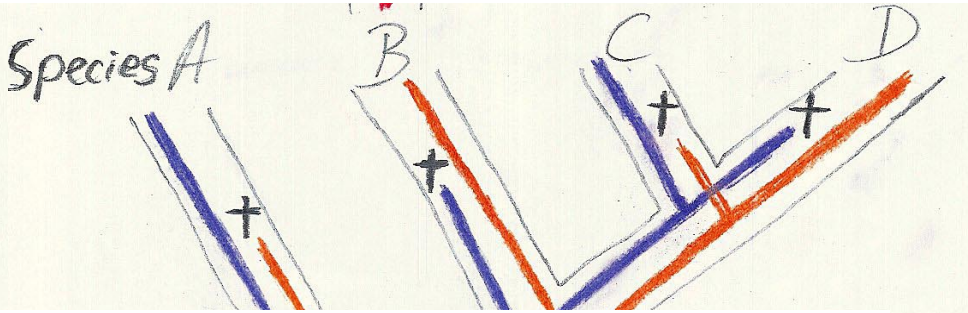
Organismal tree:



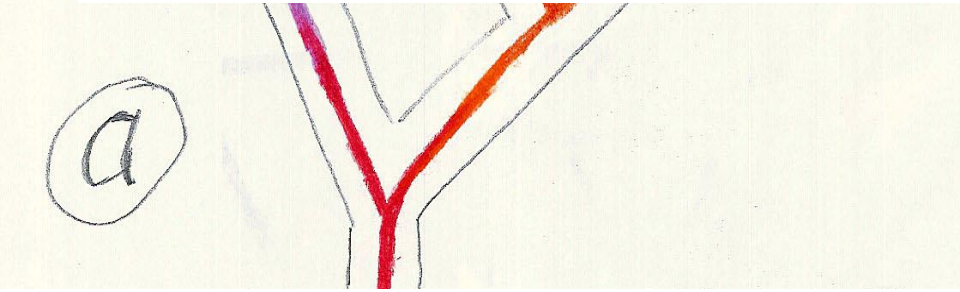
molecular tree:

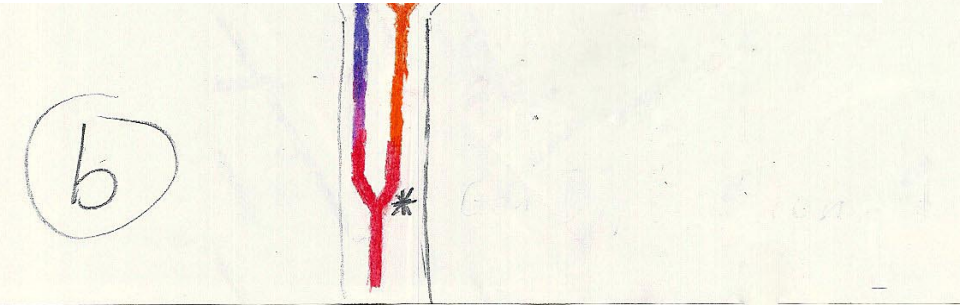
Gene duplication and gene transfer are equivalent explanations.





The more relatives of C are found that do not have the blue type of gene, the less likely is the duplication loss scenario





Horizontal or lateral Gene

Ancient duplication followed by gene loss

Note that scenario B involves many more individual events than A

1 HGT with orthologous replacement

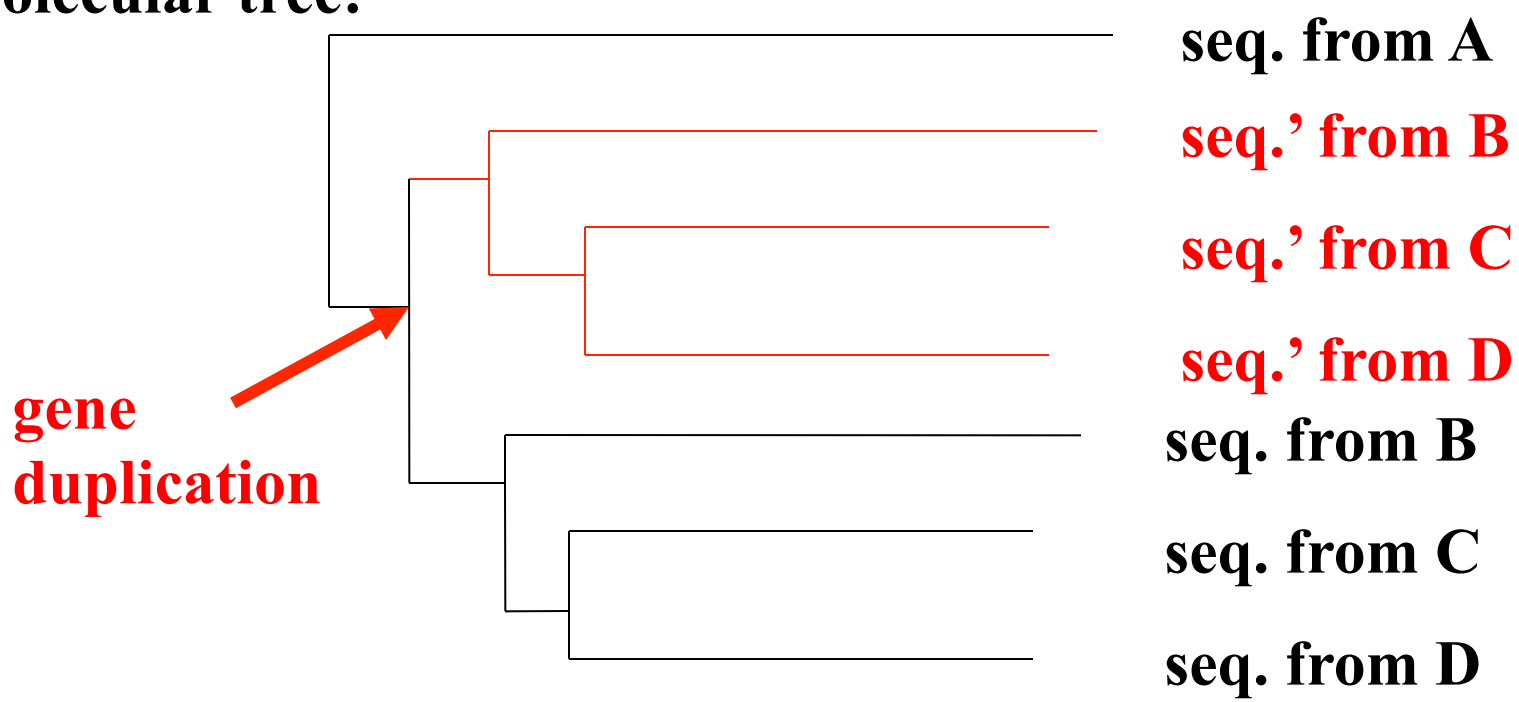
1 gene duplication followed by 4 independent gene loss events

What is it good for?

Gene duplication events can provide an outgroup that allows rooting a molecular phylogeny.
Most famously this principle was applied in case of the tree of life – the only outgroup available in this case are ancient paralogs (see http://gogarten.uconn.edu/cvs/Publ_Pres.htm for more info).
However, the same principle also is applicable to any group of organisms, where a duplication preceded the radiation (example).
Lineage specific duplications also provide insights into which traits were important during evolution of a lineage.

Function, ortho- and paralogy

molecular tree:

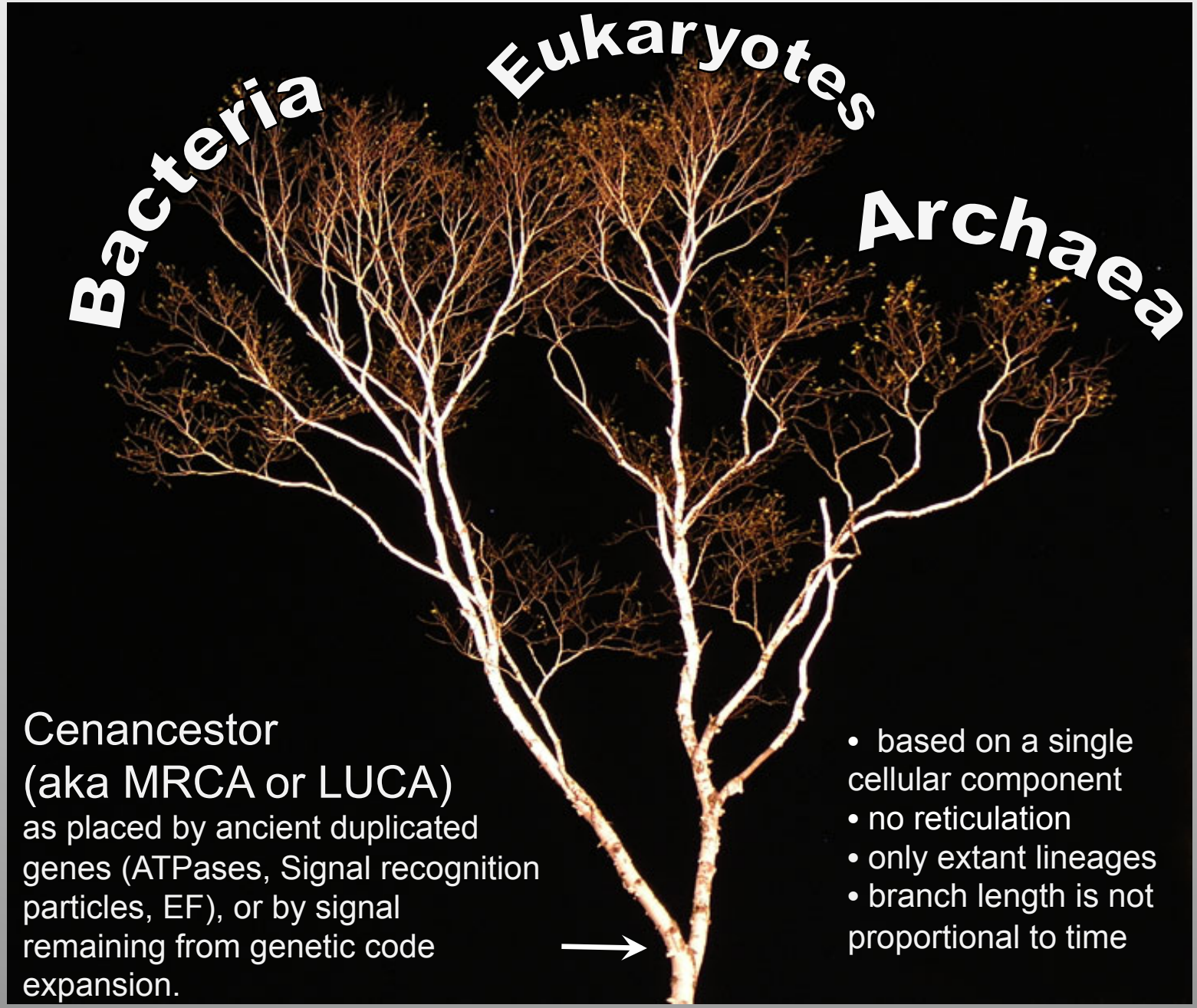


The presence of the duplication is a taxonomic character (shared derived character in species B C D).

The phylogeny suggests that seq' and seq have similar function, and that this function was important in the evolution of the clade BCD.

seq' in B and seq'in C and D are orthologs and probably have the same function, whereas seq and seq' in BCD probably have different function (the difference might be in subfunctionalization of functions that seq had in A. – e.g. organ specific expression)

The Ribosomal “Tree of Life”



SSU-rRNA Tree of Life

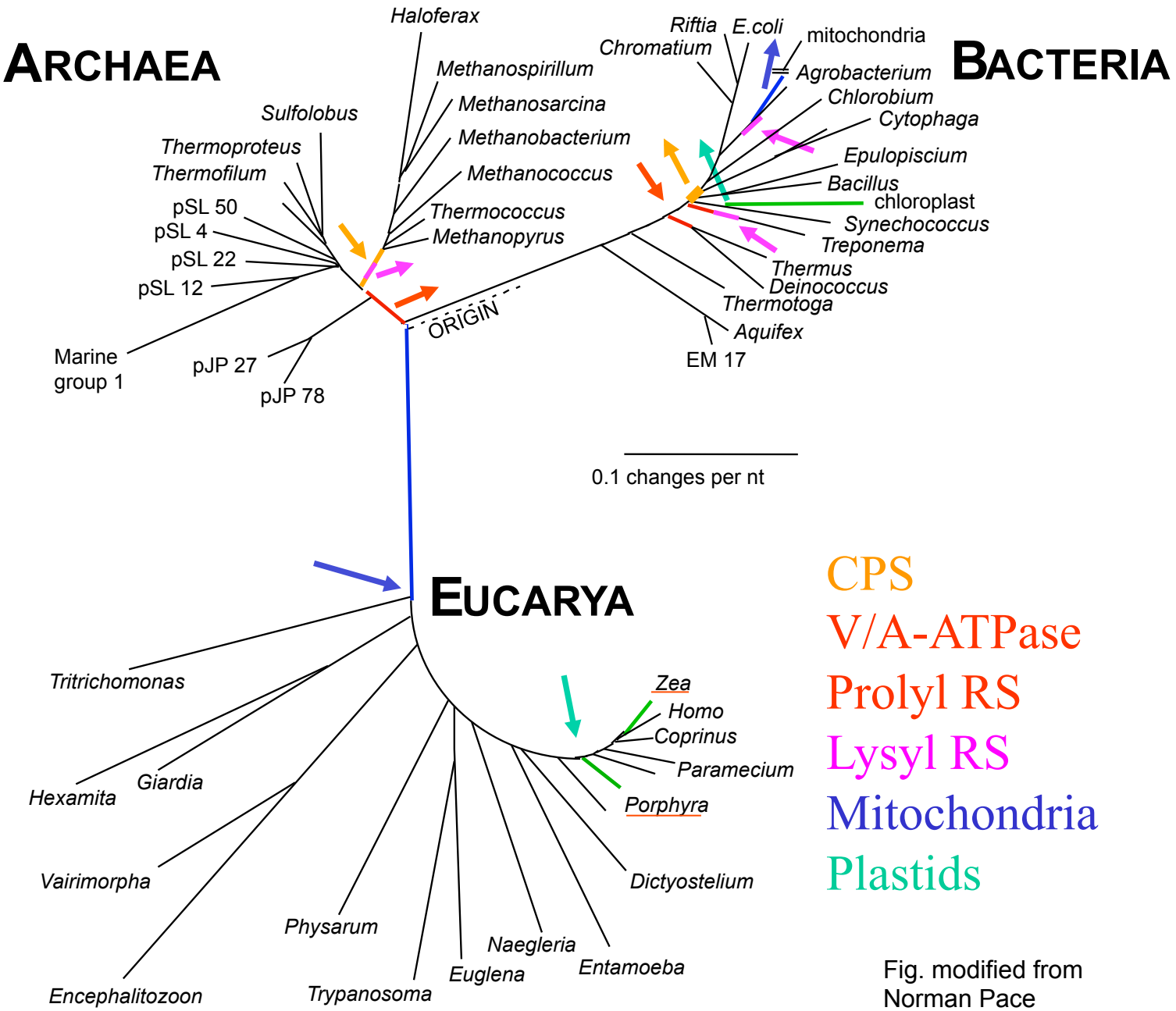
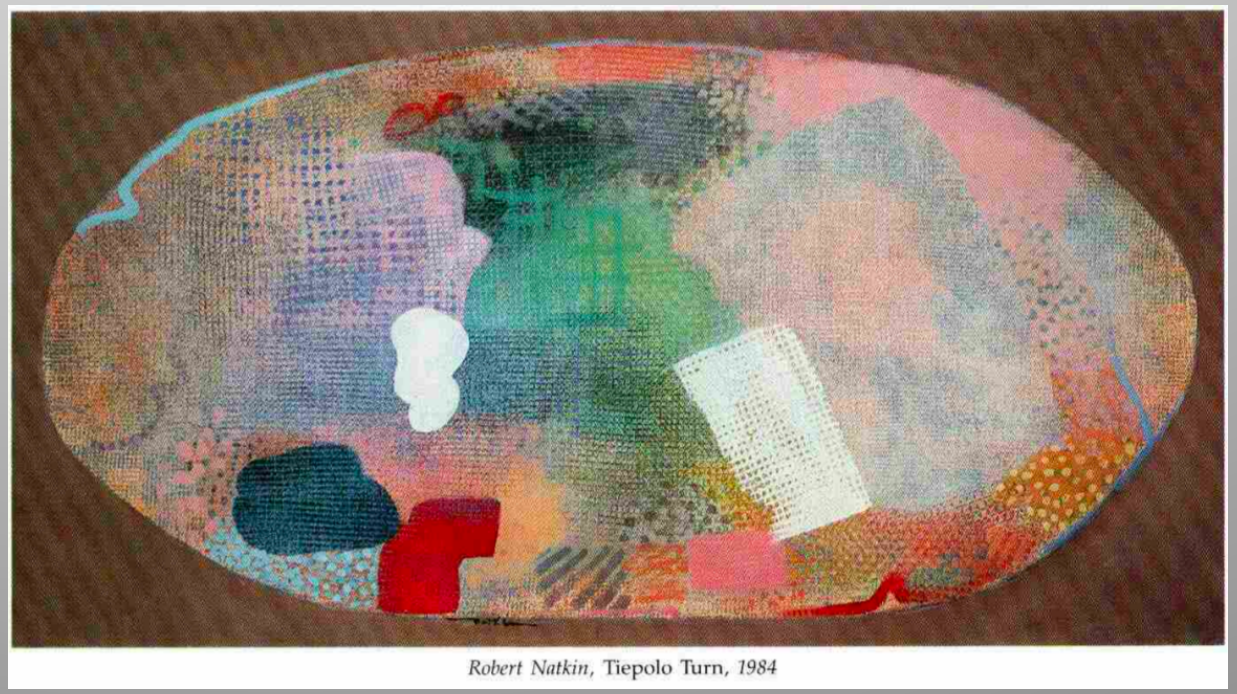


Fig. modified from Norman Pace

Ochiai, K., Yamanaka, T., Kimura, K., and Sawada, O. (1959)
Inheritance of drug resistance (and its tranfer) between *Shigella* strains and Between *Shigella* and *E.coli* strains.
Hihon Iji Shimpor 1861: 34 (in Japanese)

Gray GS, Fitch WM (1983):
Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*.
Mol Biol Evol 1983, 1(1):57-66.

Sorin Sonea (1988):
The global organism:
A new view of bacteria.
The Sciences, 28:38-45.



Robert Natkin, Tiepolo Turn, *1984*

1993:

BioSystems 31 (1993) 111–119

Bio Systems

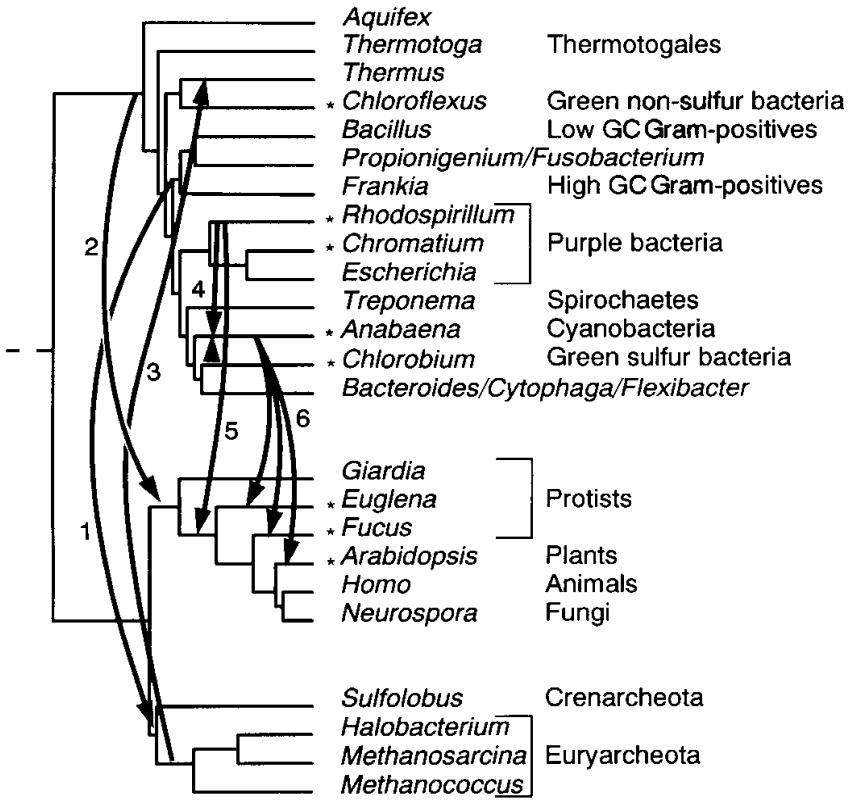
Horizontal transfer of ATPase genes — the tree of life becomes a net of life

Elena Hilario, Johann Peter Gogarten*

Department of Molecular and Cell Biology, University of Connecticut, 75 North Eagleville Rd., Storrs, CT 06269-3044, USA

1995:

GOGARTEN, J.P. (1995):
The early evolution of cellular life,
Trends in Ecology and Evolution, 10,
147-151.

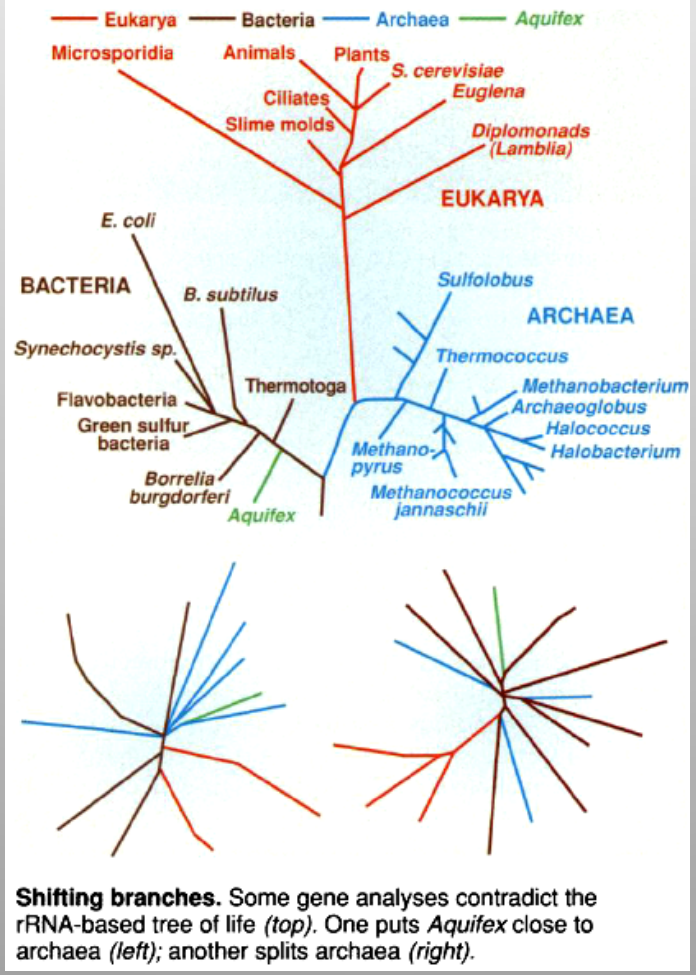


1998:

RESEARCH NEWS

Genome Data Shake Tree of Life

New genome sequences are mystifying evolutionary biologists by revealing unexpected connections between microbes thought to have diverged hundreds of millions of years ago



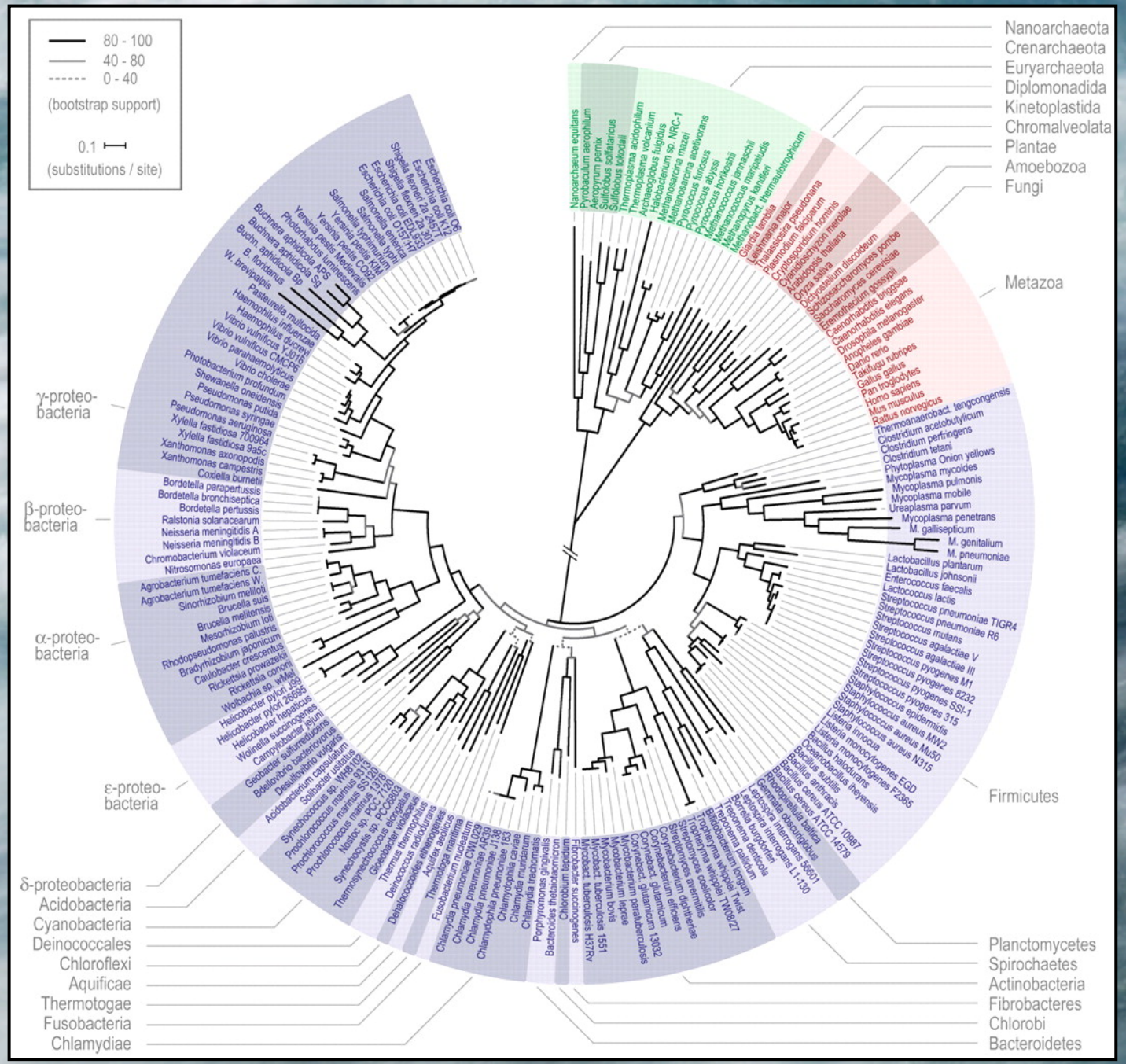
Shifting branches. Some gene analyses contradict the rRNA-based tree of life *(top)*. One puts *Aquifex* close to archaea *(left)*; another splits archaea *(right)*.

Science, **280**, p.672ff (1998)

The tree of one percent*

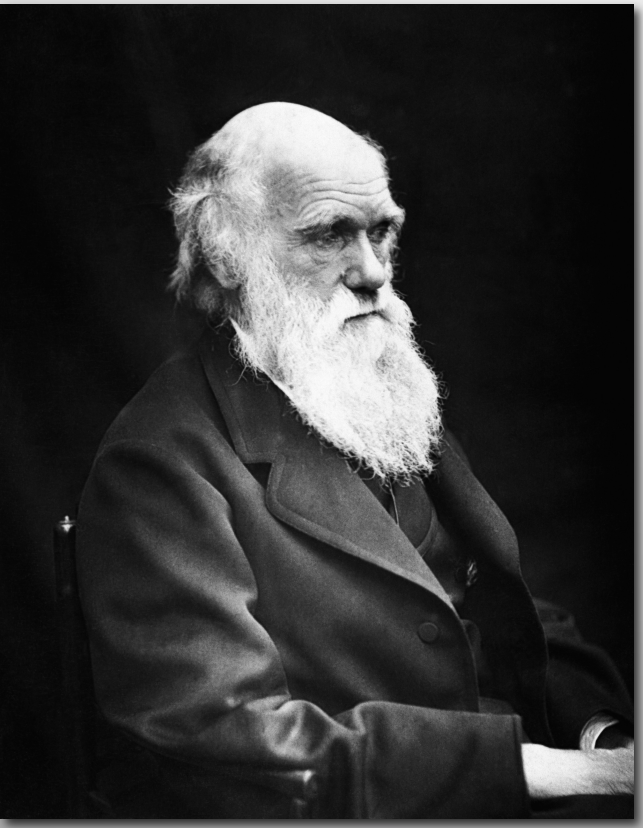
31 orthologous protein sequences present in 191 genomes each, concatenated, 8,089 sites

Ciccarelli *et al.* 2006
Toward automatic reconstruction of a highly resolved tree of life.



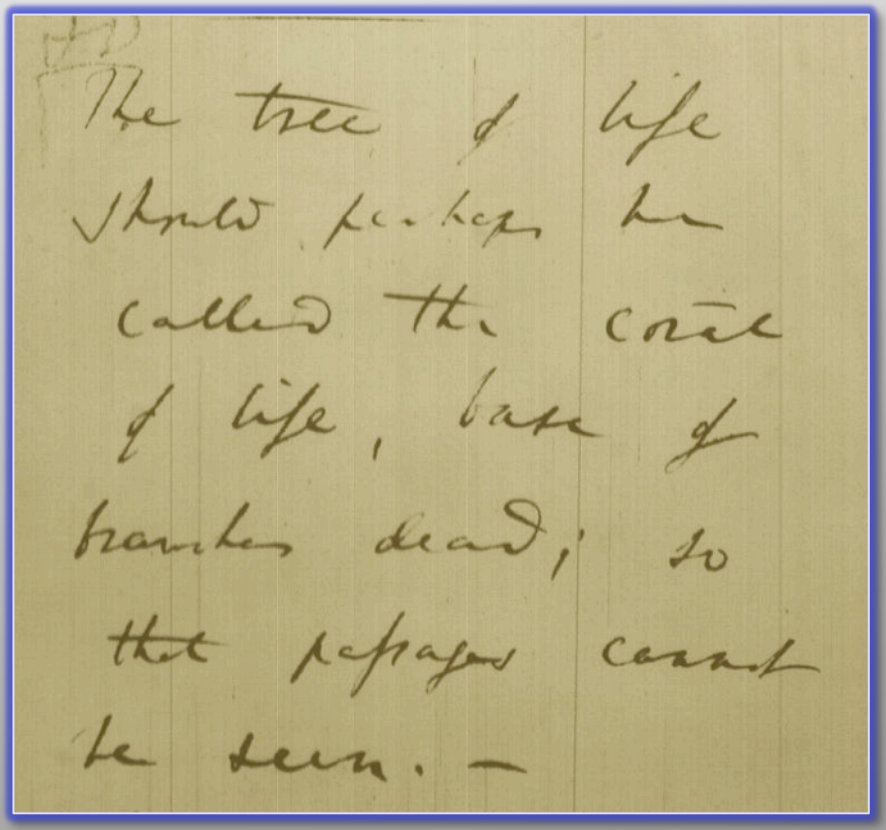
*Dagan and Martin,Genome Biology, 2006

Tree, Web, or Coral of Life?



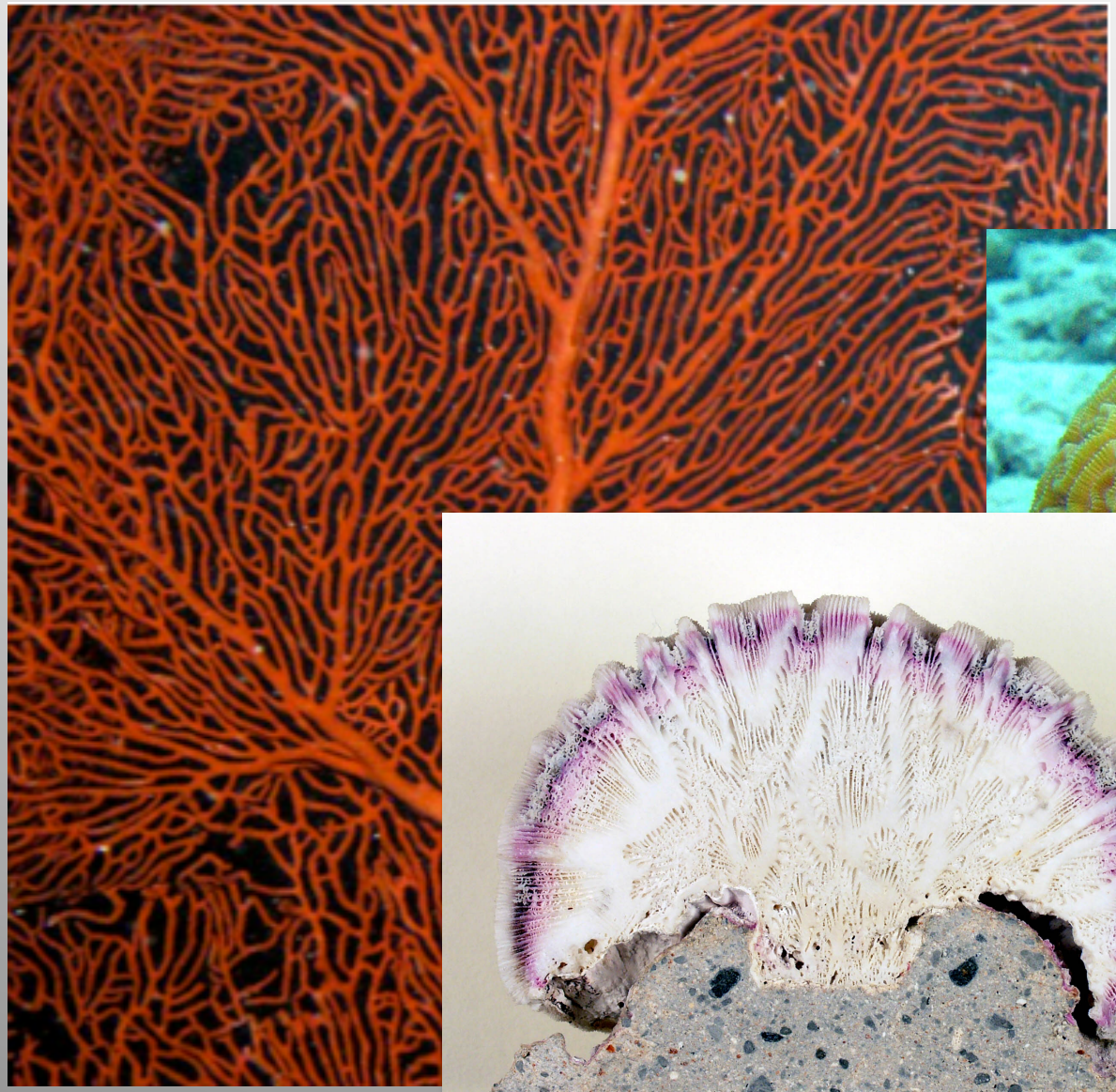
Charles Darwin
Photo by J. Cameron, 1869

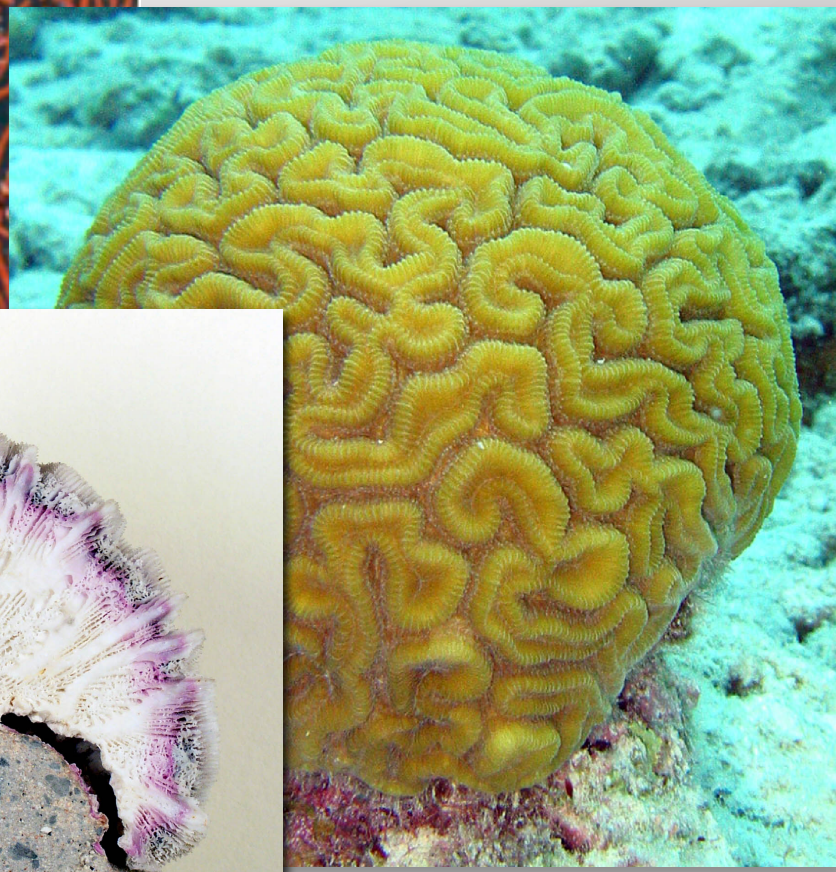
"The tree of life should perhaps be called the coral of life, base of branches dead"



Page B26 from Charles Darwin's (1809-1882) notebook (1837/38)

Which Type of Coral ?





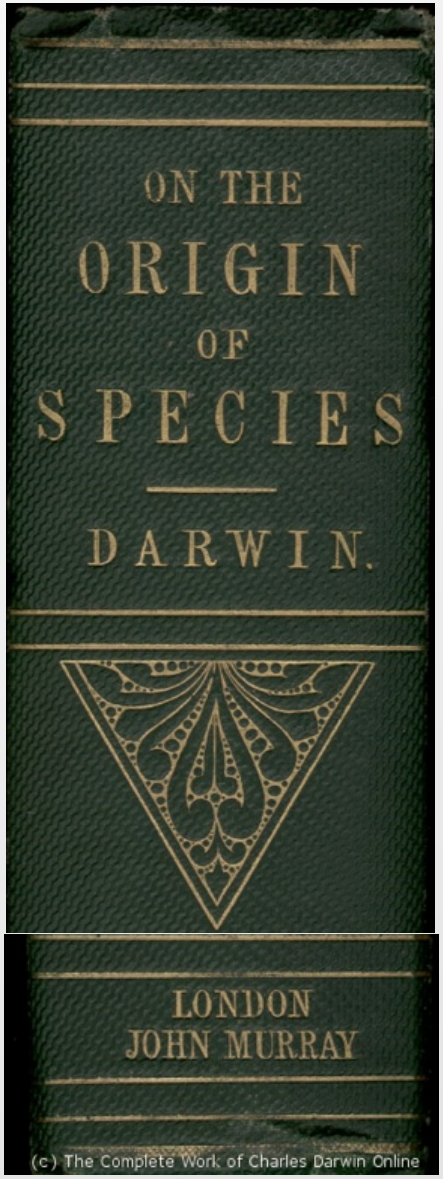
Darwin's Coral a Red Algae?

(Bossea orbignyana)

According to **Horst Bredekamp**, parts of the diagram in Darwin's origin of species (center) may reflect the branching properties of a specimen Darwin collected himself.

From Florian Maderspacher:
"The captivating coral--the origins of early evolutionary imagery."
Current Biology 16: R476-8 2006

“As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever-branching and beautiful ramifications.”



Charles Darwin in
“On the **Origin of Species** by **Means of Natural Selection** or the **Preservation of Favoured Races in the Struggle for Life**”,
p 162 ff, London, John Murray, 1859

THE POTATO OF LIFE

Successive independent crystallization of the three domains of life from a population of pre-cells

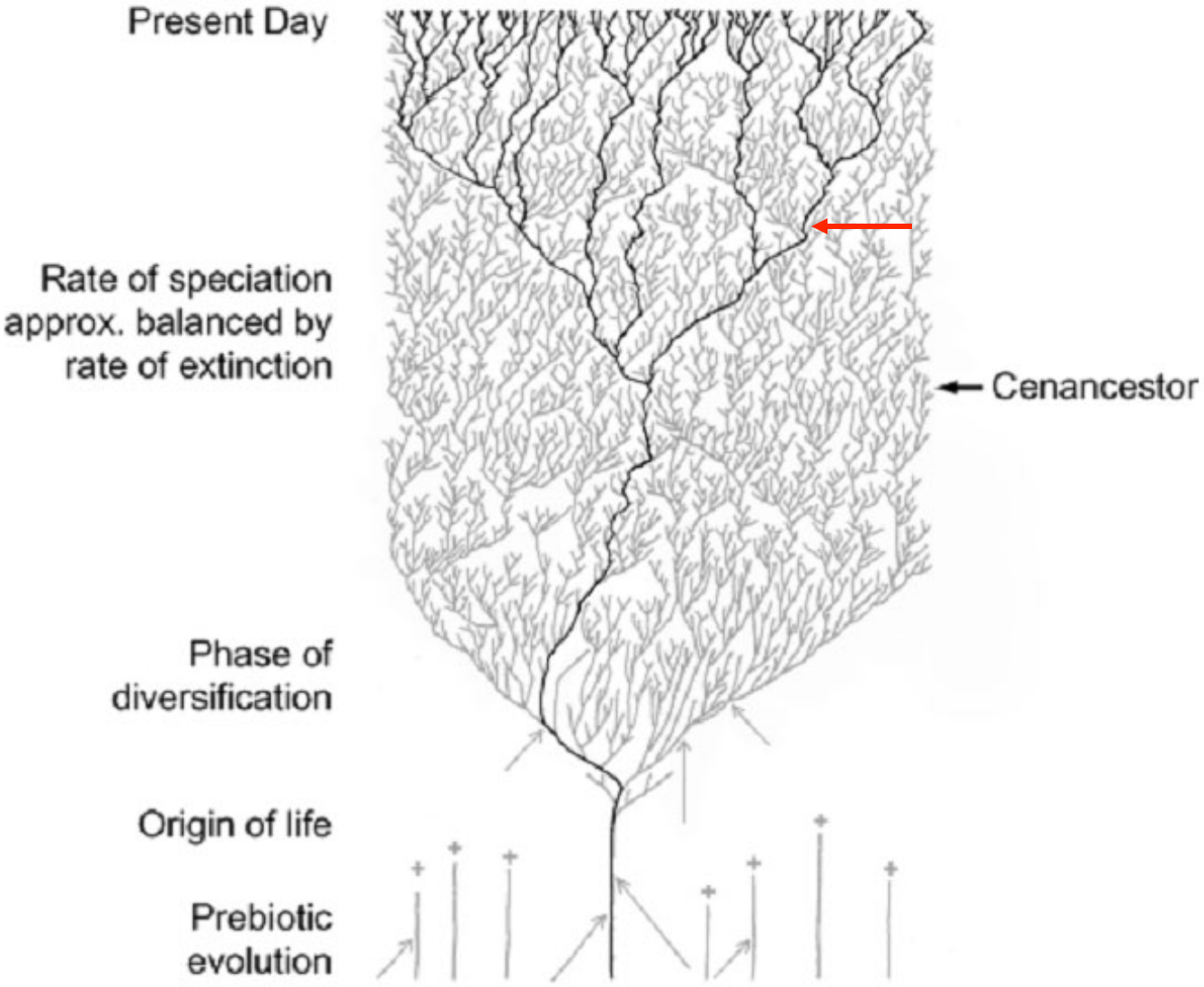
photo from www.sonnentaler.net

Glansdorff N, Xu Y, Labedan B: *The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner.* *Biol Direct* 2008, **3**:29.

Kandler O (1994) *The early diversification of life*. In *Early Life on Earth, Nobel Symposium*, vol. 84, pp. 152-509: Columbia Univ. Press.

Woese CR (2002) *On the evolution of cells*. *PNAS* 99: 8742-7

The Coral of Life (Darwin)



Coalescence – the process of tracing lineages backwards in time to their common ancestors. Every two extant lineages coalesce to their most recent common ancestor. Eventually, all lineages coalesce to the cenancestor.

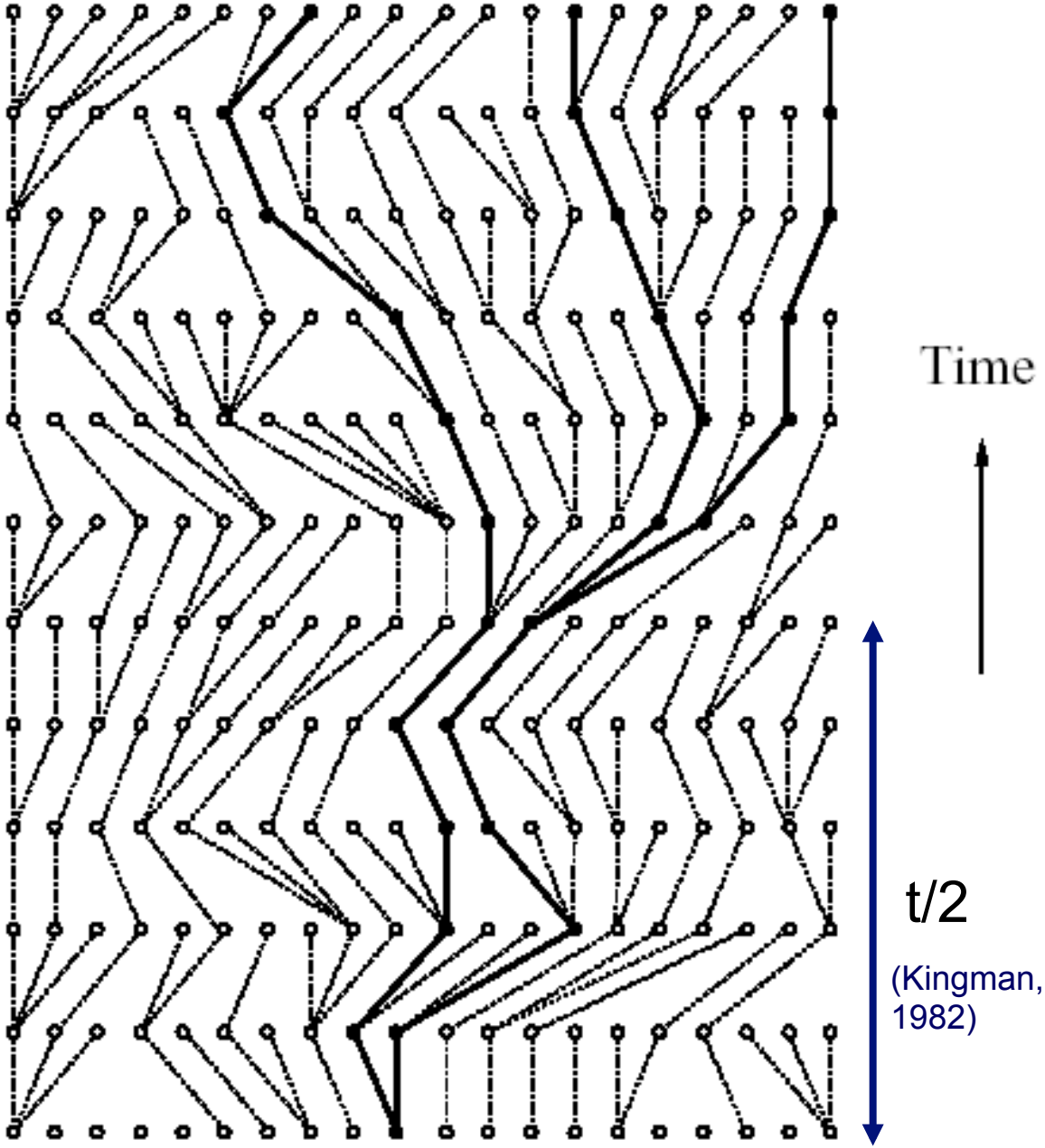
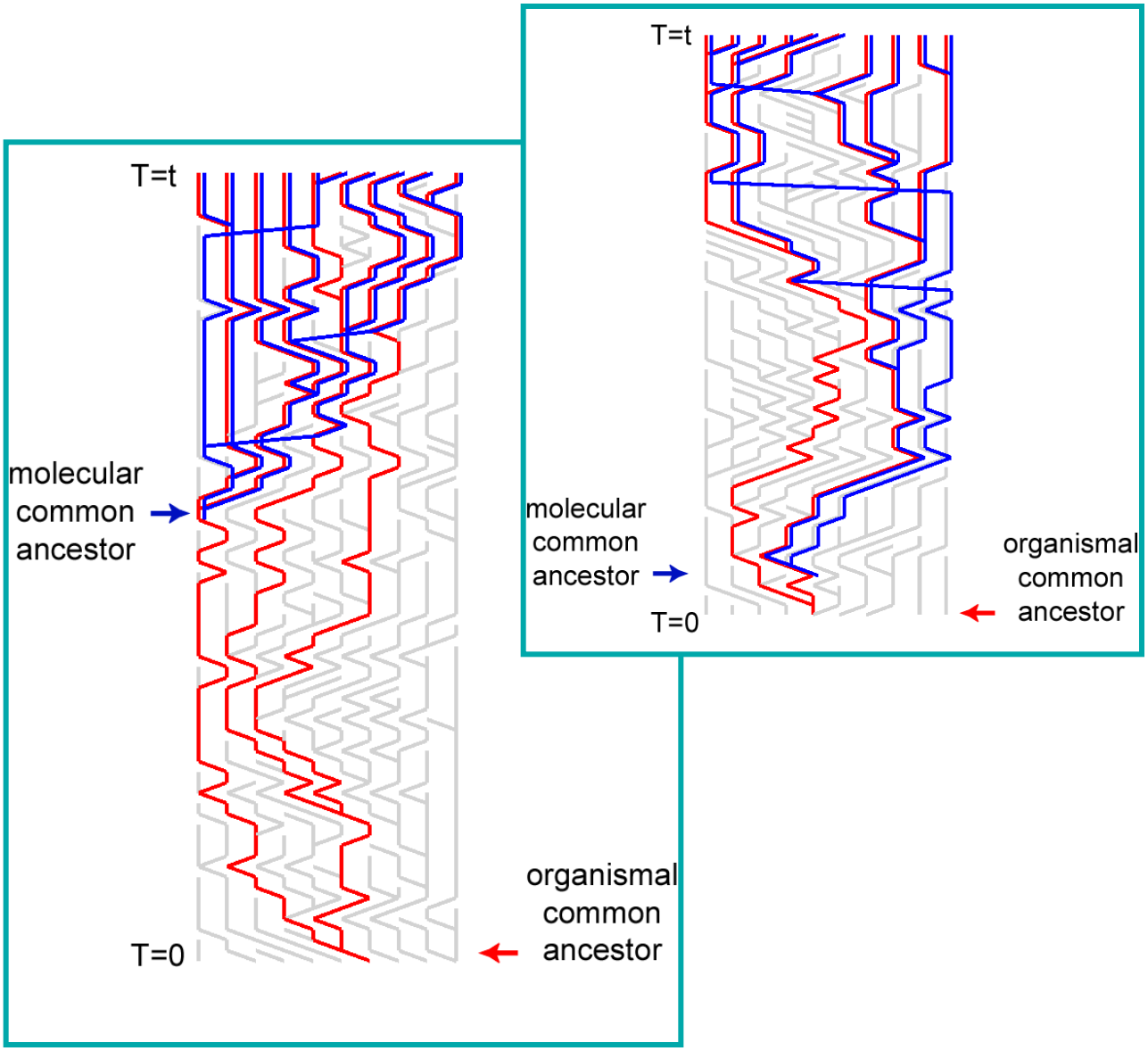


Illustration is from J. Felsenstein, “Inferring Phylogenies”, Sinauer, 2003

Simulations of organismal evolution by coalescence



•One extinction and one speciation event per generation

•Horizontal transfer event once in 10 generations

RED: organismal lineages (no HGT)
BLUE: molecular lineages (with HGT)
GRAY: extinct lineages

RESULTS:
•Most recent common ancestors are different for organismal and molecular phylogenies

•Different coalescence times

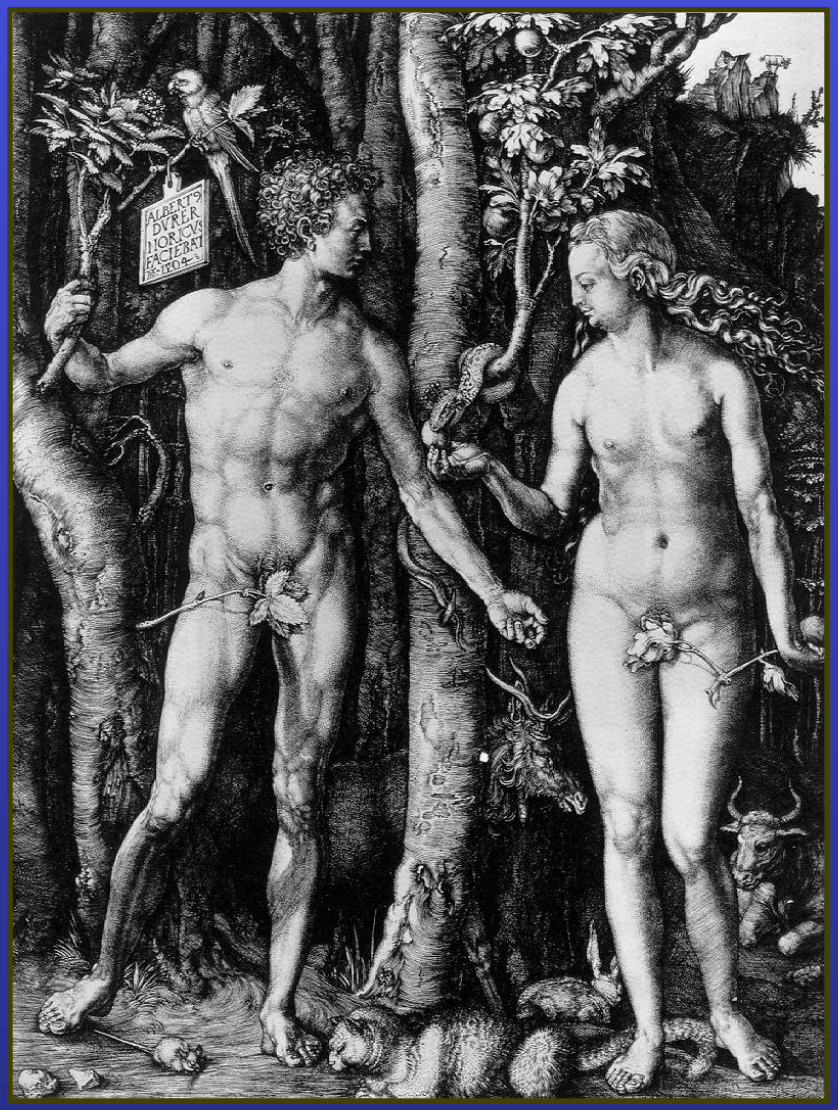
•Long coalescence of the last two lineages

Y chromosome Adam

Lived approximately 50,000 years ago

Thomson, R. *et al.* (2000) *Proc Natl Acad Sci* U S A 97, 7360-5

Underhill, P.A. *et al.* (2000) *Nat Genet* 26, 358-61



Albrecht Dürer, The Fall of Man, 1504

Mitochondrial Eve

Lived 166,000-249,000 years ago

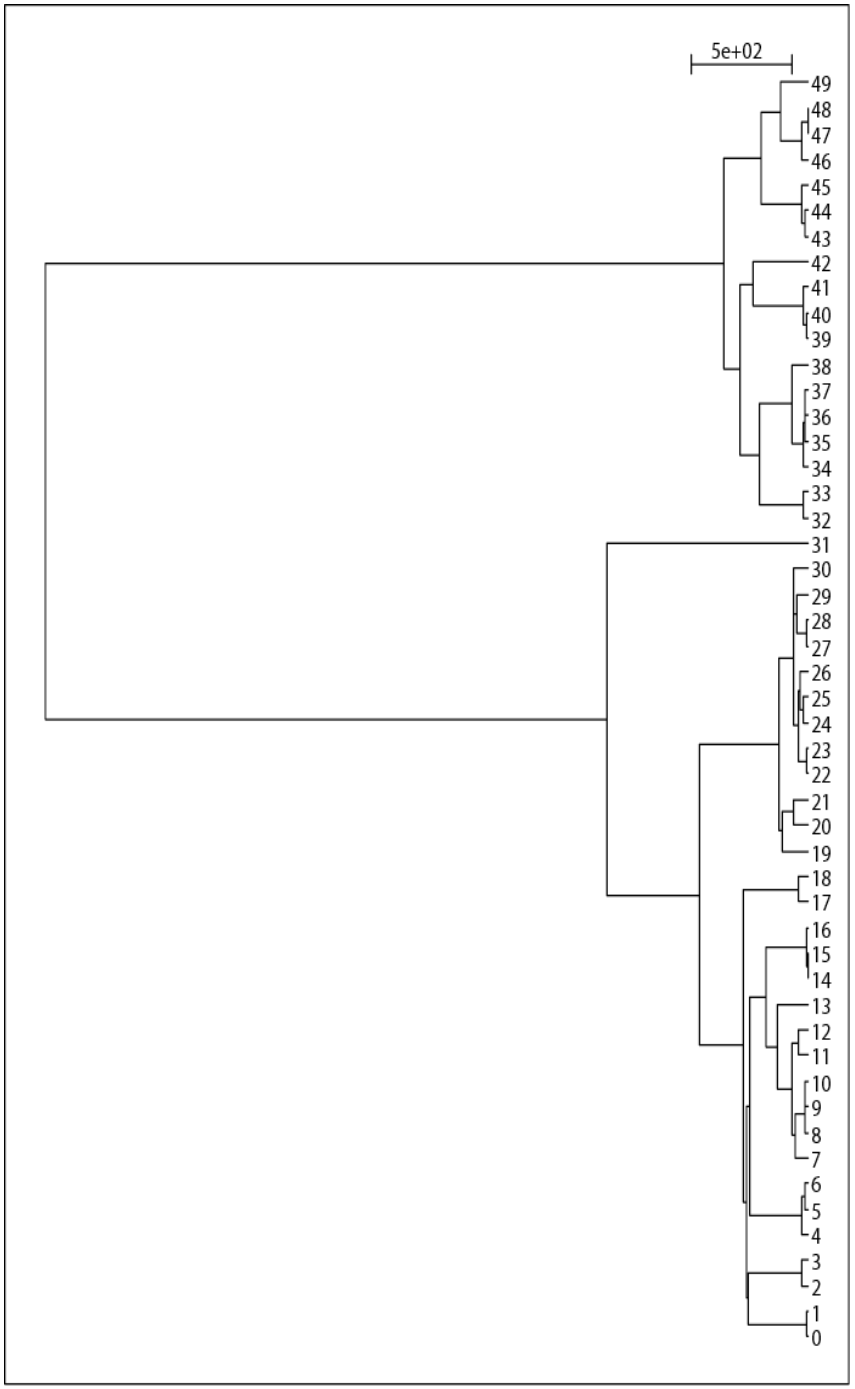
Cann, R.L. *et al.* (1987) *Nature* 325, 31-6

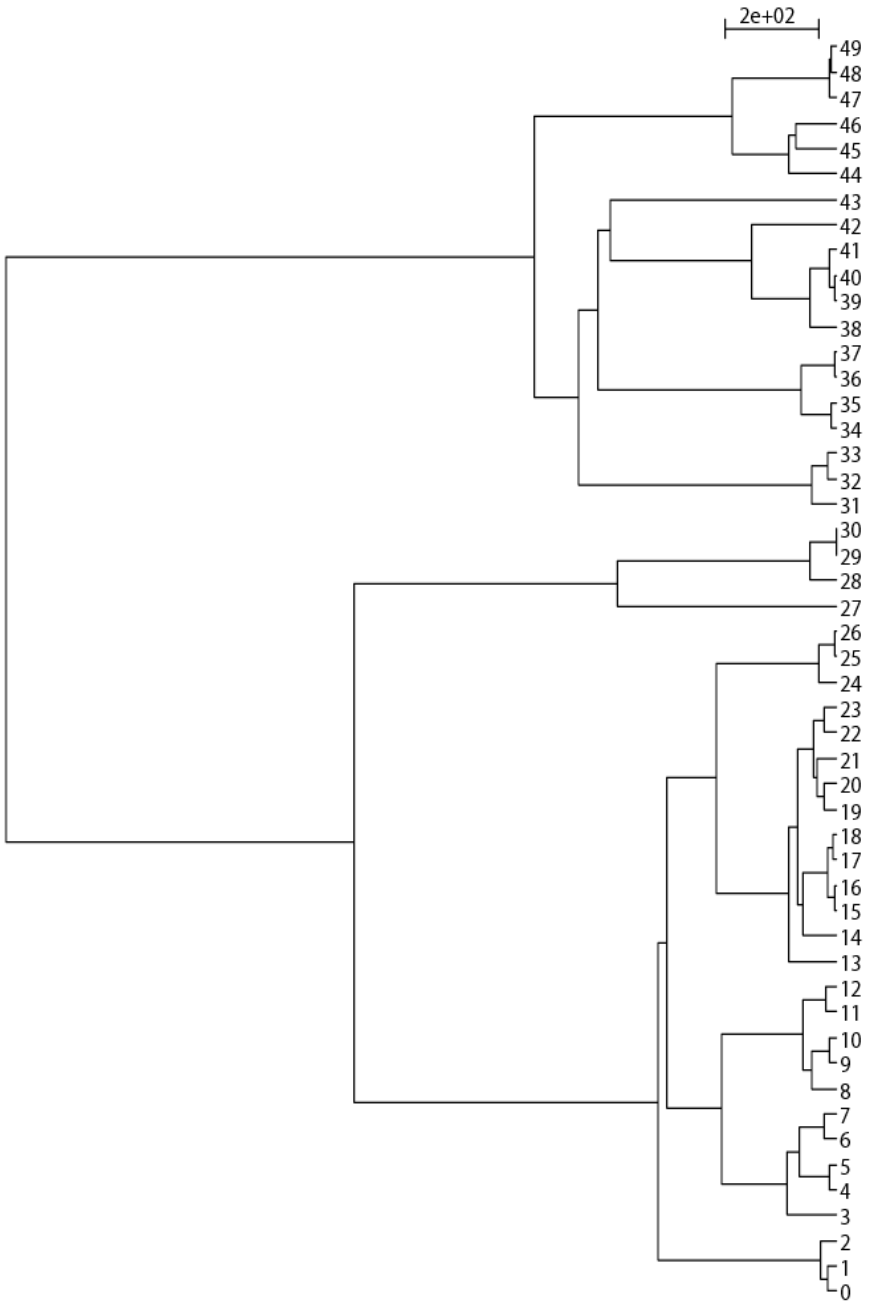
Vigilant, L. *et al.* (1991) *Science* 253, 1503-7

Adam and Eve never met ☹

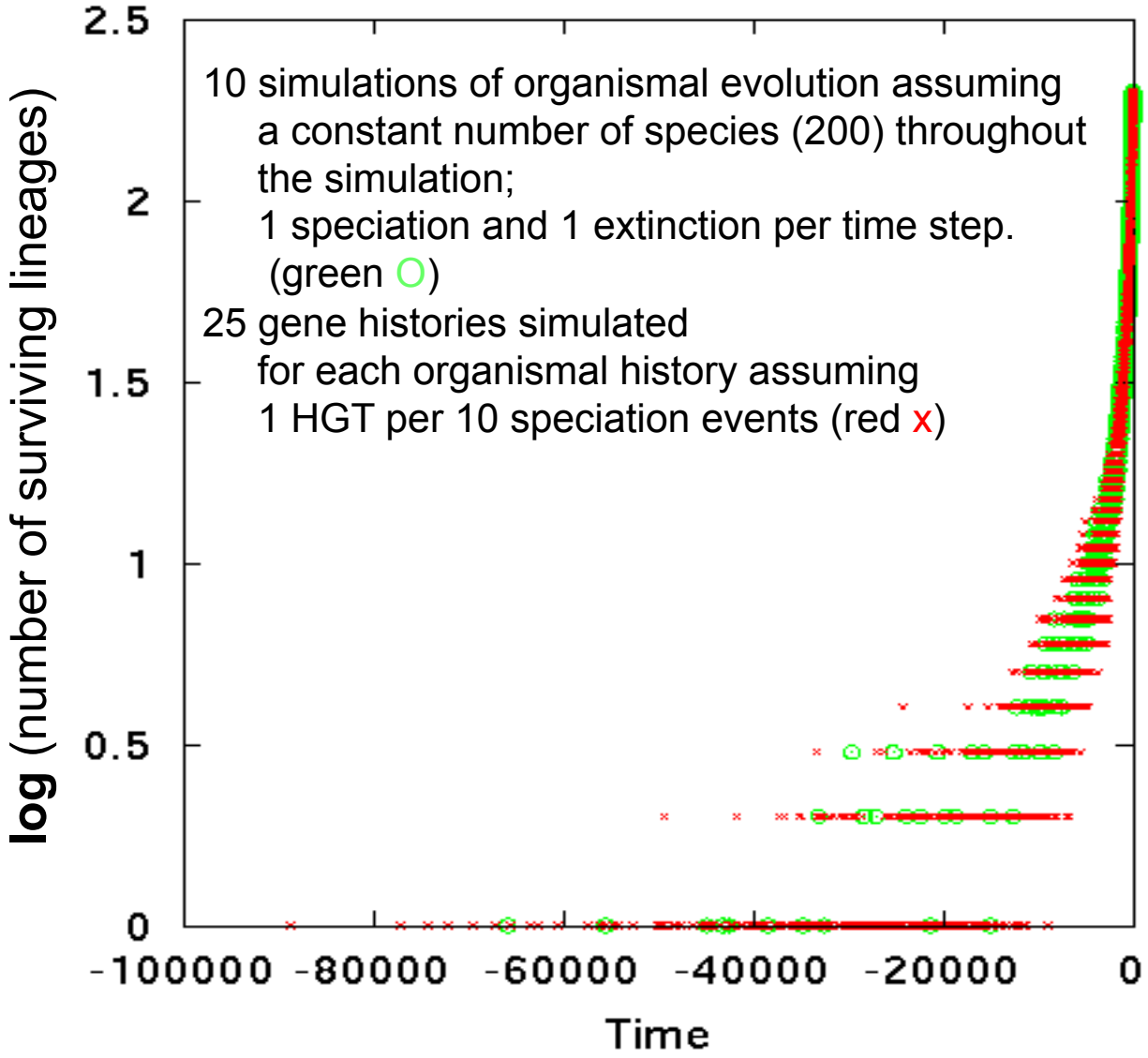
The same is true for ancestral rRNAs, EF, ATPases!

EXTANT LINEAGES FOR THE SIMULATIONS OF 50 LINEAGES

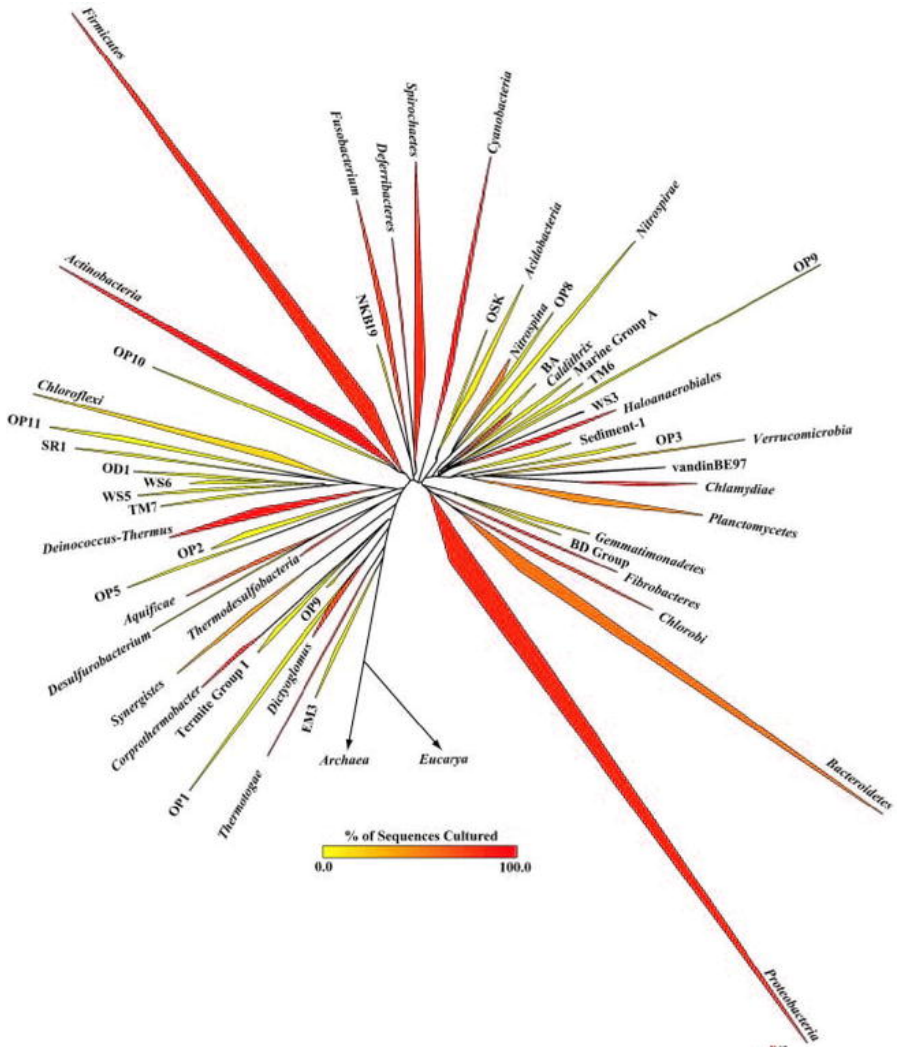




Lineages Through Time Plot



green: organismal lineages ;
red: molecular lineages (with gene transfer)



Bacterial 16SrRNA based phylogeny
(from P. D. Schloss and J. Handelsman, Microbiology and Molecular Biology Reviews, December 2004.)

The deviation from the “long branches at the base” pattern could be due to

- under sampling
- an actual radiation
  - due to an invention that was not transferred
  - following a mass extinction