BIBMS: Phylogenetic reconstruction (I). Overview and trees, trees, trees.

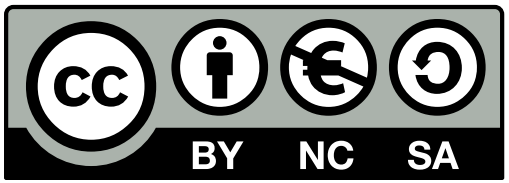
Ramón Díaz-Uriarte

Dept. Bioquímica
Universidad Autónoma de Madrid
Madrid, Spain
ramon.diaz@iib.uam.es
http://ligarto.org/rdiaz

October 2018

(Rev: 6b99b1d)

License and copyright



Outline

Overview

- Phylogenies and tree-thinking
- Basic (math) models of substitution (DNA and proteins)
- Phylogenetic reconstruction

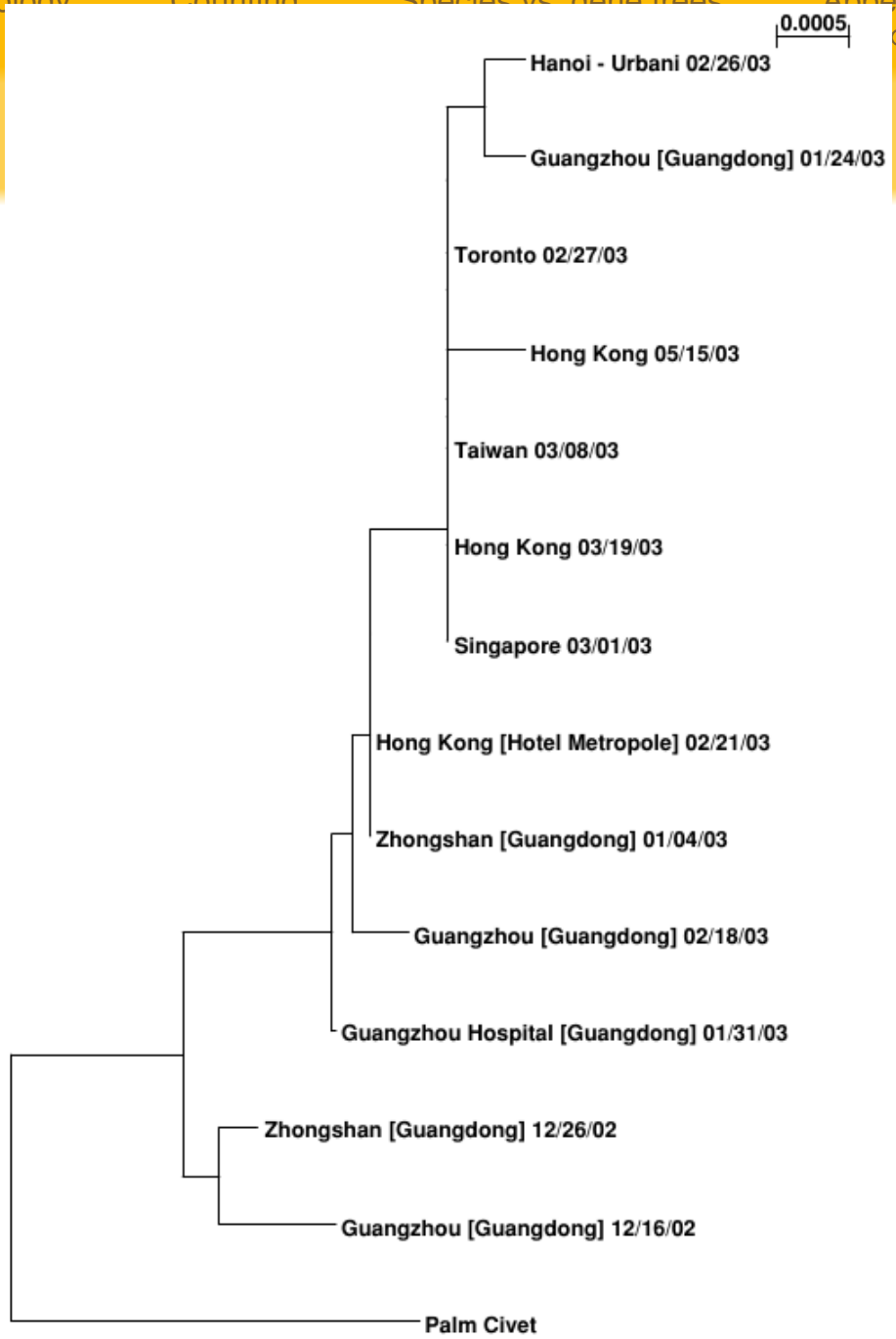
Why do phylogenies matter?

- Dobzhansky's "Nothing in biology makes sense except in the light of evolution".
- Alignment and scores.
- The major groups of organisms, the history of life, our place in all the mess, etc.

Why do phylogenies matter?

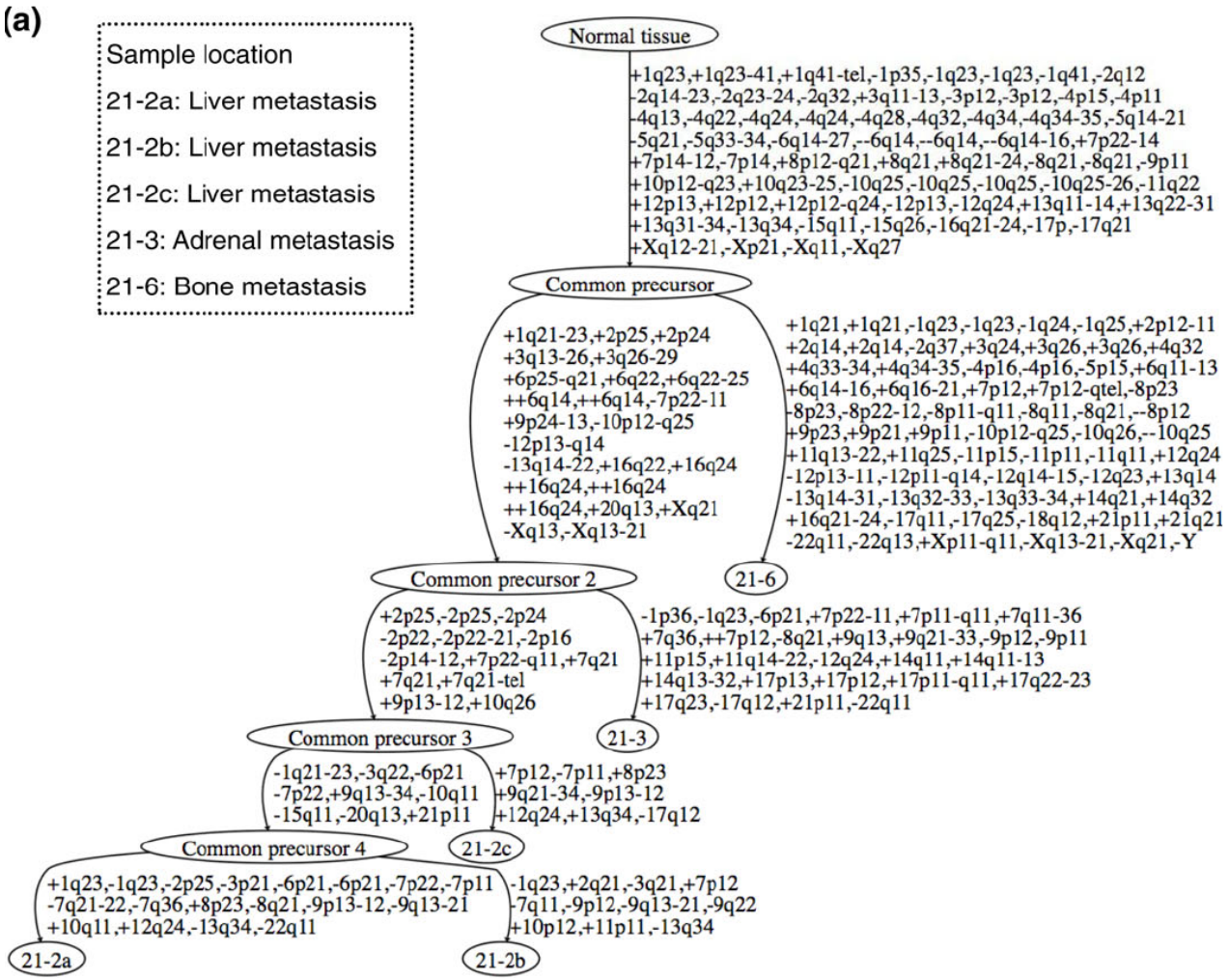
- HIV, SARS, Ebola, etc: phylogenies used to:
  - identify source of virus (geographical source);
  - date the onset of epidemic;
  - detect recombination;
  - track viral evolution within patient;
  - identify modes of transmission;
  - key mutations for spreading;
  - original viral host;

SARS history



From Cristianini and Hahn, 2006

Intrapatient tumor phylogeny



From Letouze et al., 2010

Reconstructing ancestral states: "molecular archaeology"

OPEN ACCESS Freely available online

PLOS BIOLOGY

Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication

Karin Voordeckers[1,2☯], Chris A. Brown[1,2,3,4,5☯], Kevin Vanneste[6,7], Elisa van der Zande[1,2], Arnout Voet[8], Steven Maere[6,7*], Kevin J. Verstrepen[1,2*]

1 VIB Laboratory for Systems Biology, Leuven, Belgium, **2** CMPG Laboratory for Genetics and Genomics, KU Leuven, Leuven, Belgium, **3** Fathom Information Design, Boston, Massachusetts, United States of America, **4** Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge, Massachusetts, United States of America, **5** Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, United States of America, **6** VIB Department of Plant Systems Biology, Gent, Belgium, **7** Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent, Belgium, **8** Laboratory for Molecular en Structural Biology, KU Leuven, Leuven, Belgium

Abstract

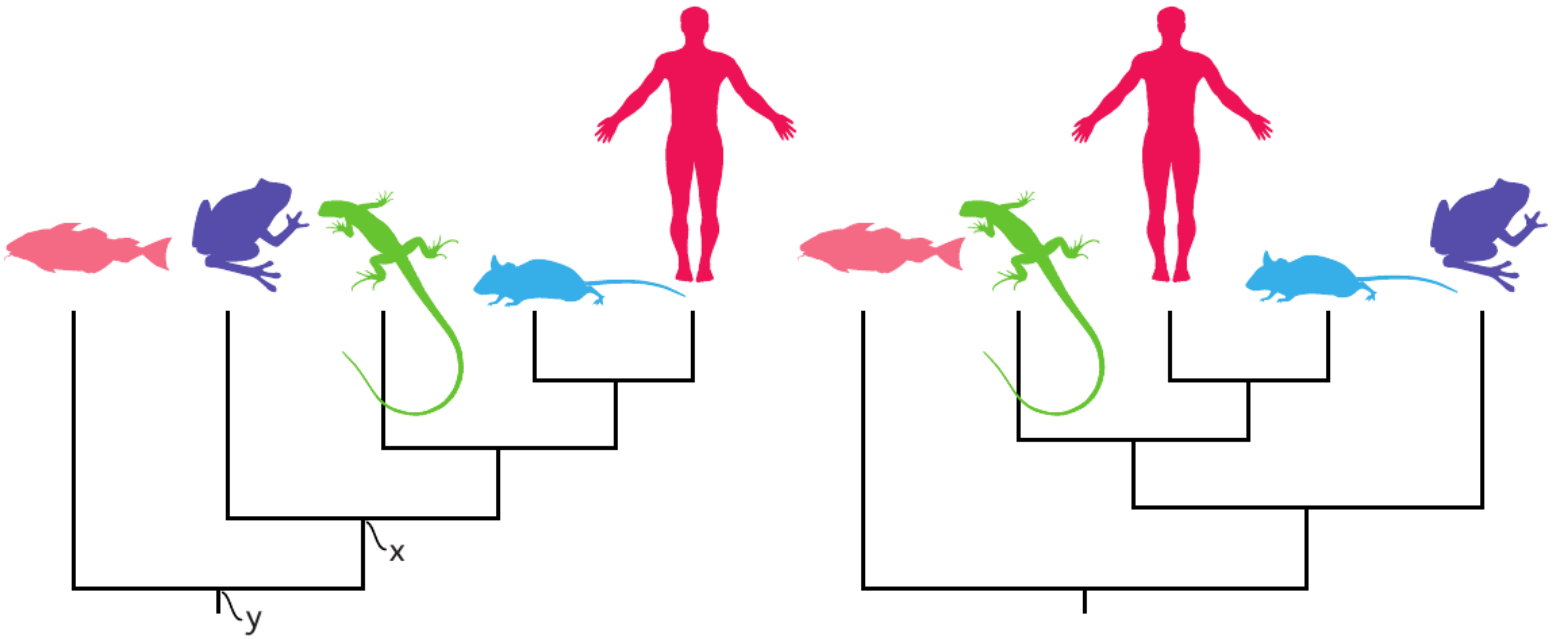
Gene duplications are believed to facilitate evolutionary innovation. However, the mechanisms shaping the fate of duplicated genes remain heavily debated because the molecular processes and evolutionary forces involved are difficult to reconstruct. Here, we study a large family of fungal glucosidase genes that underwent several duplication events. We reconstruct all key ancestral enzymes and show that the very first preduplication enzyme was primarily active on maltose-like substrates, with trace activity for isomaltose-like sugars. Structural analysis and activity measurements on resurrected and present-day enzymes suggest that both activities cannot be fully optimized in a single enzyme. However, gene duplications repeatedly spawned daughter genes in which mutations optimized either isomaltase or maltase activity

Reasons of why we cover what we cover

- Because it is important
- Because it is beautiful
- To introduce and connect with other ideas
  - Just algorithms vs. probabilistic models
  - It takes soooooooo long: Computational complexity or how many trees are there?
  - Maximum likelihood
  - Bayesian approaches
  - Assessing confidence: the bootstrap

Interpreting trees

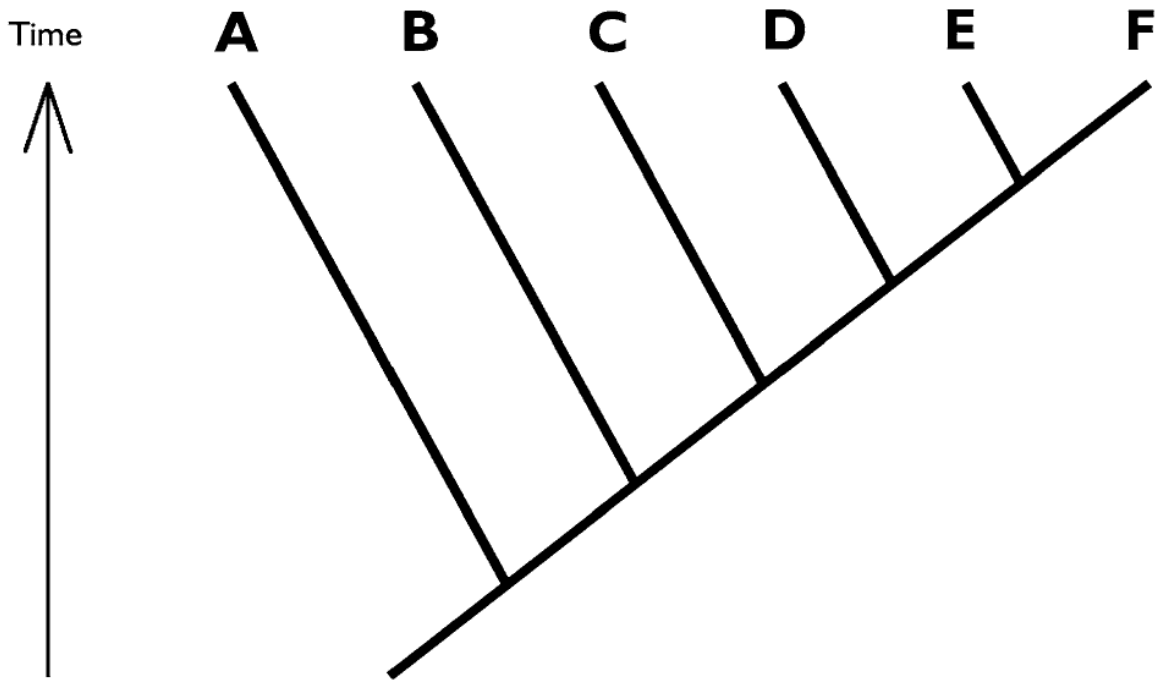
Which phylogeny is correct? And using the left, is the frog more closely related to the fish or the human?



From Baum et al., 2005.

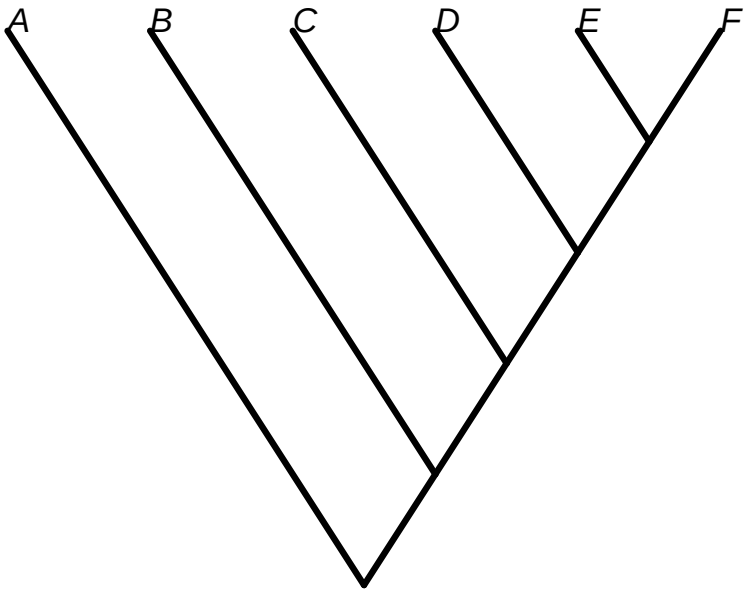
Interpreting trees

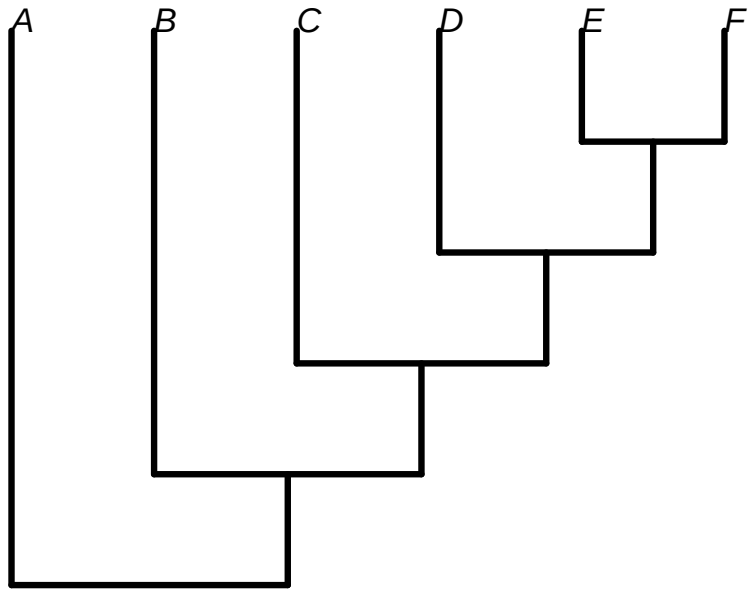
Which is (are) the species/sequences most closely related to B?



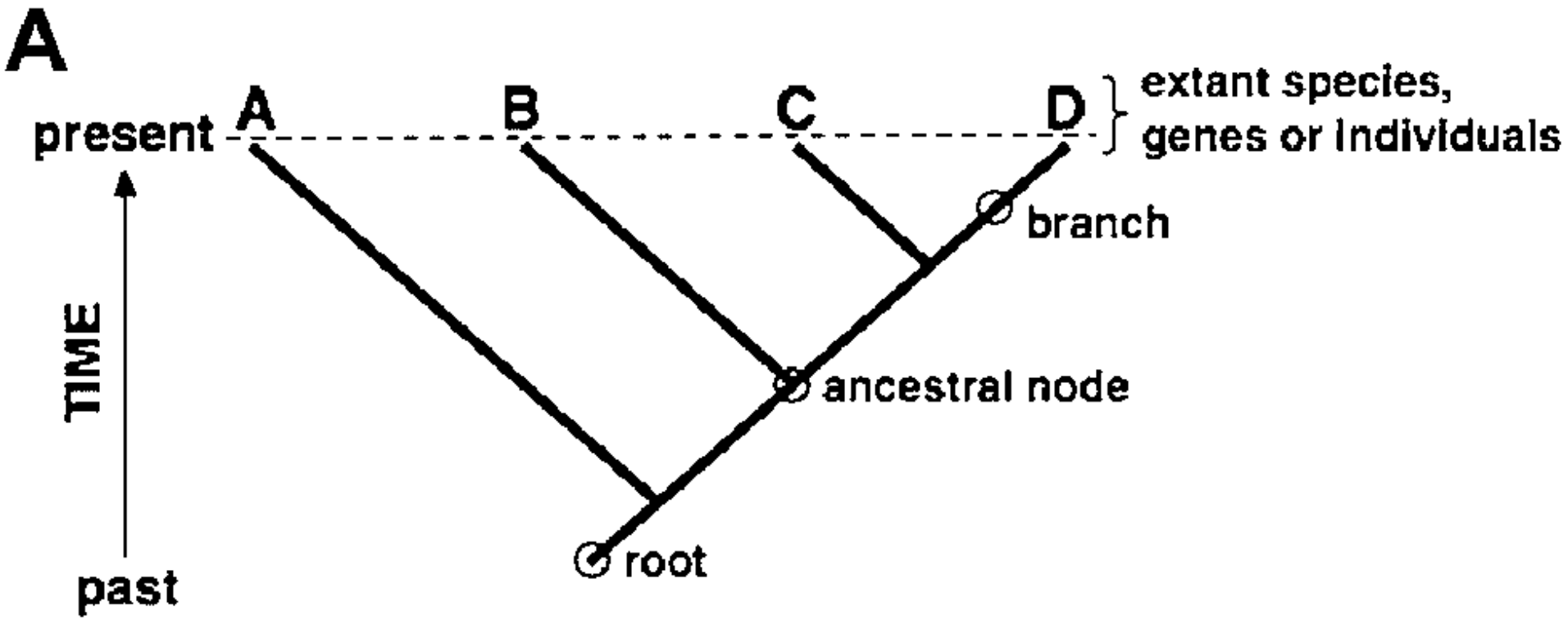
From Sandvik, 2008.

Maybe simpler?





Some terminology

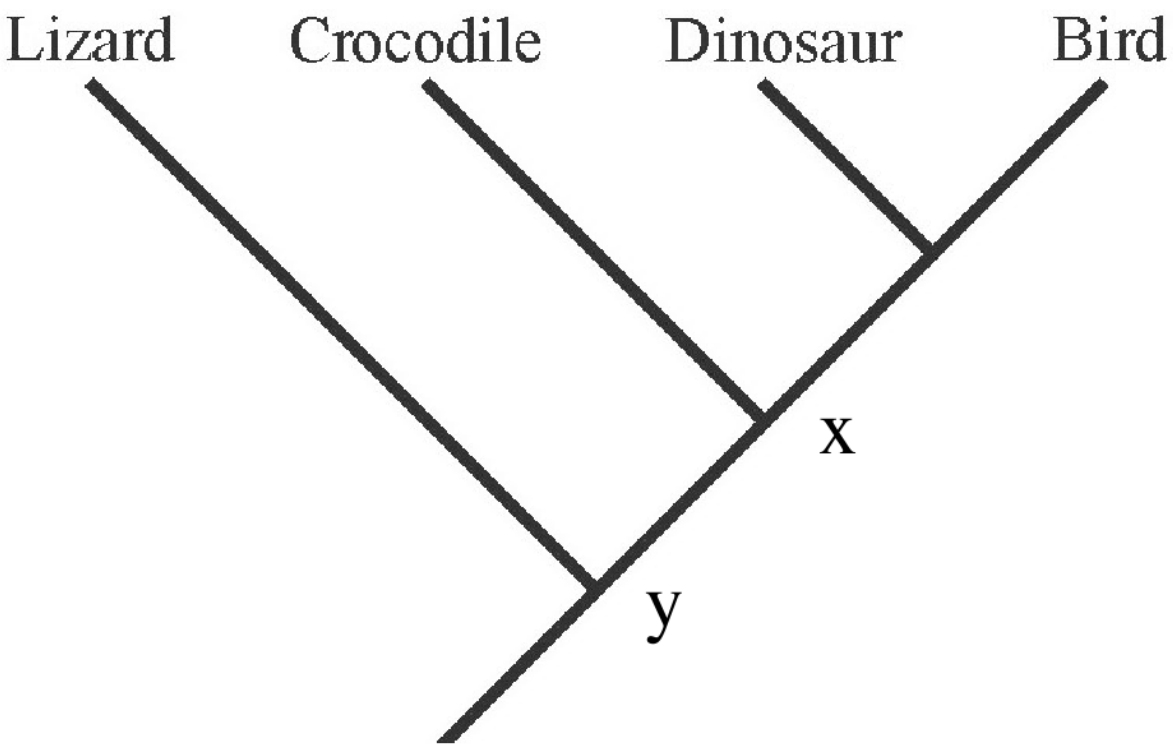


From Omland et al., 2005.

Usual terms

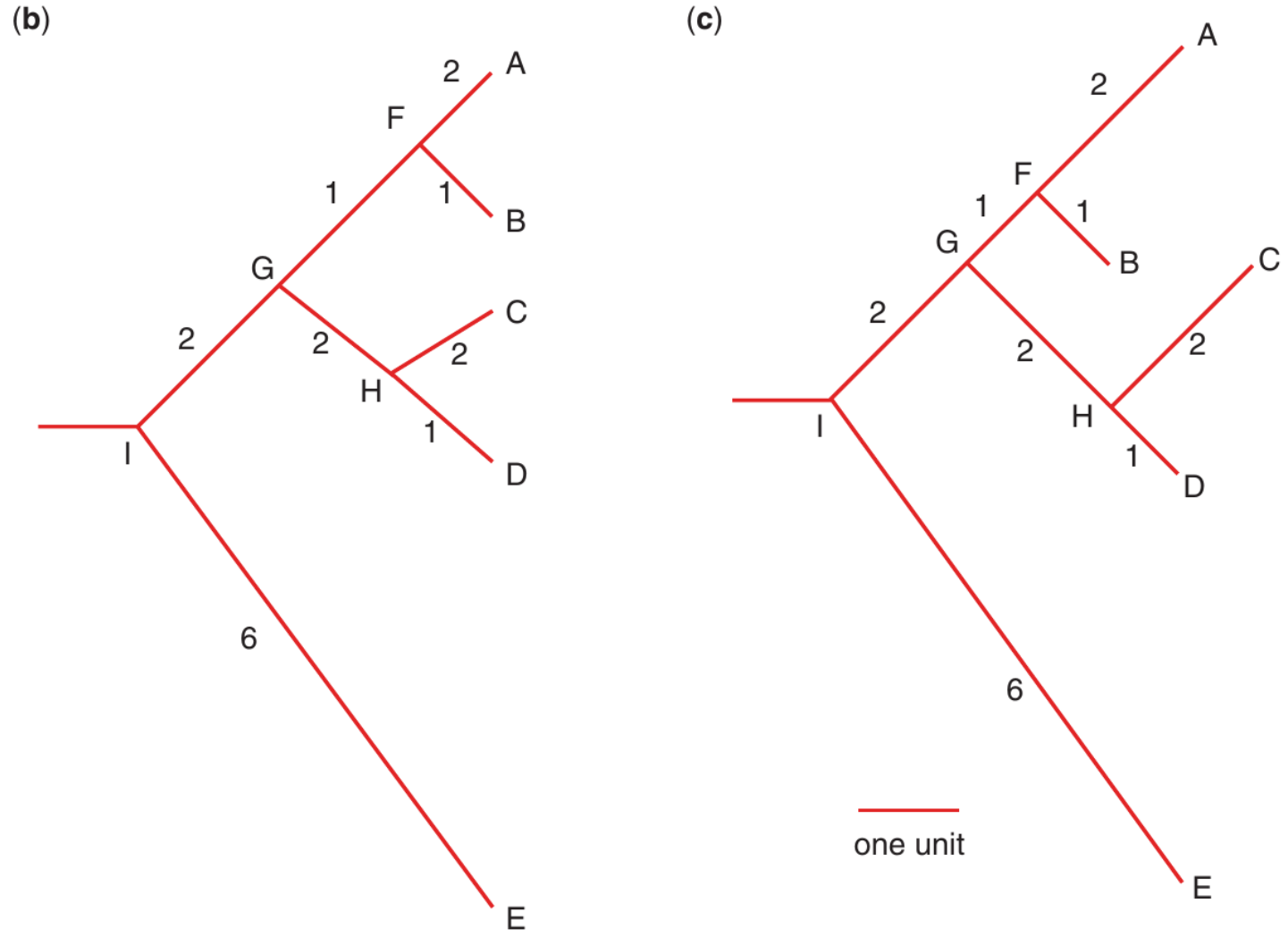
- Leave, tip, terminal node, taxon/taxa, sequences
- Internal nodes
- Branches, edges
- Clade, monophyletic group
- Polytomy, multifurcation

A classical non-monophyletic group



A croc: is it more closely related to a bird or to a lizard?
A dinosaur: is it more closely related to a bird or to a croc or a lizard?

Unscaled vs. scaled



From Pevsner, 2009 (p. 232).

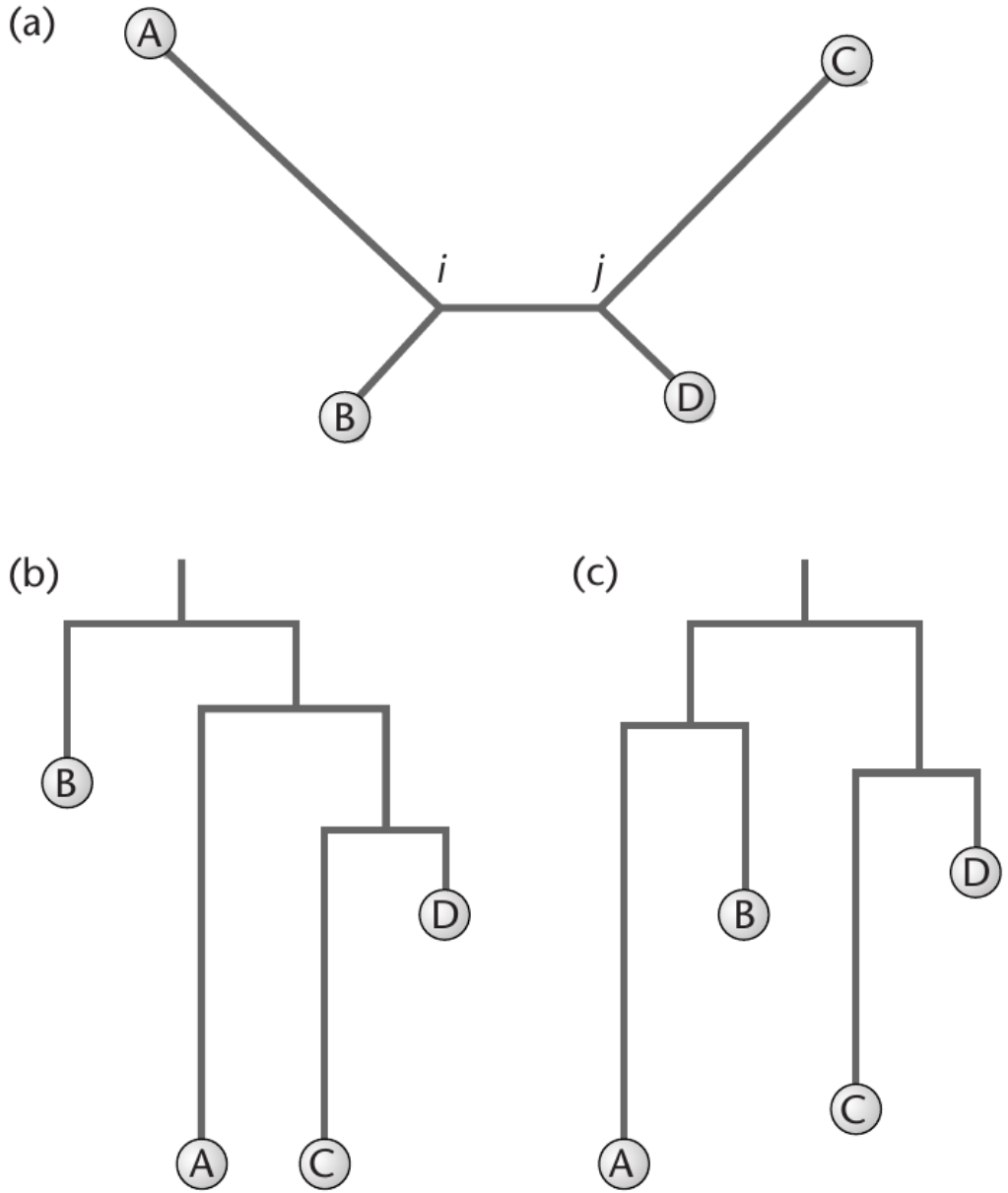
Scaled?

- Is that time?

Scaled?

- Is that time?
- Amount of change (substitutions or whatever)
- Amount of change not necessarily $\propto$ time. Why?

Rooted and unrooted



(Higgs & Attwood, 2005 (p. 160).)

How are they rooted?

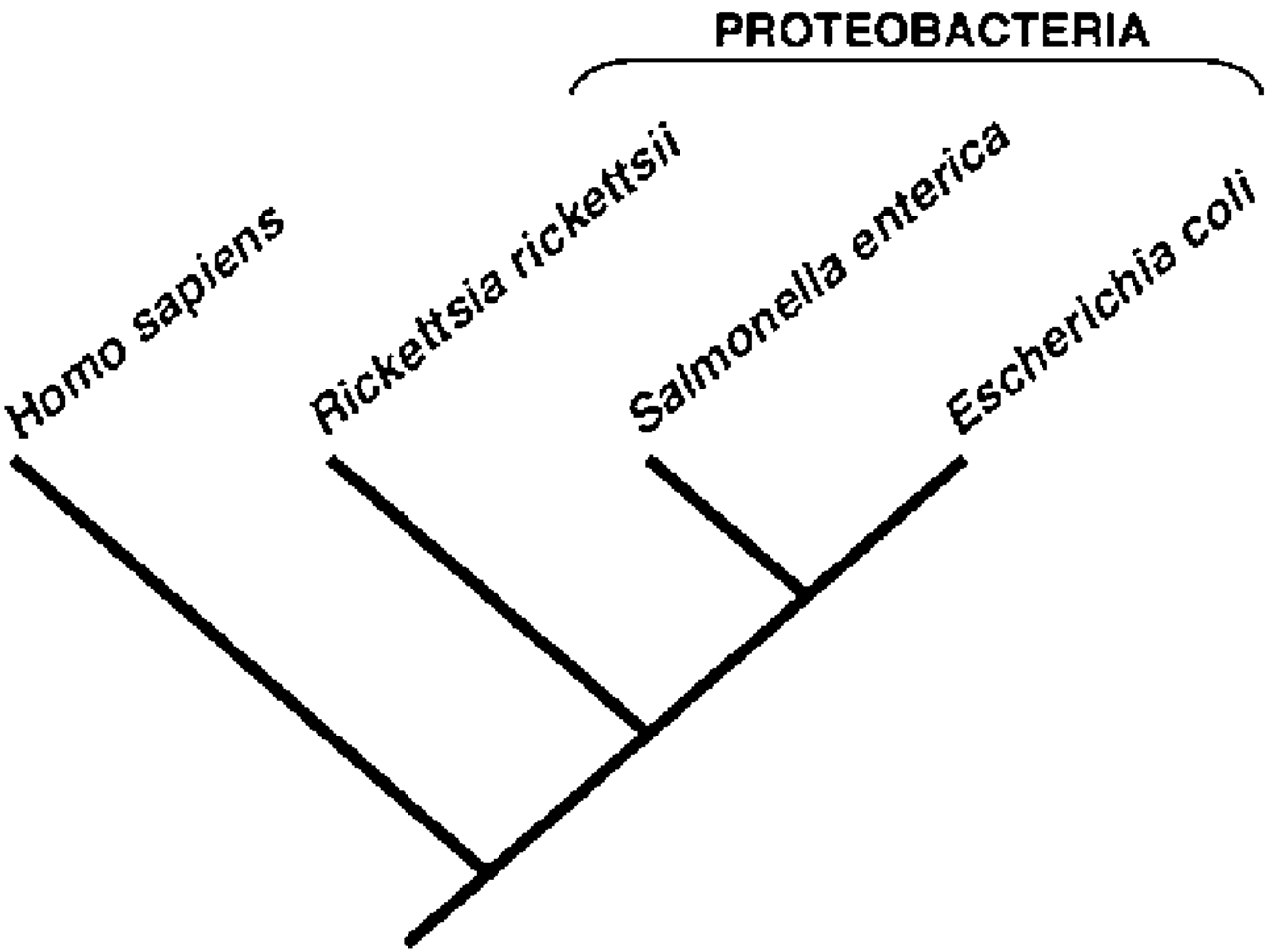
- Some methods return rooted trees. Most don't.
- Simple heuristics.
- Outgroups ("which amounts to knowing the truth").

Beware of representation!!!

- Most methods return unrooted trees.
- Many programs, by default, represent them as rooted.
- MEGA does that.

Time, age, and the left

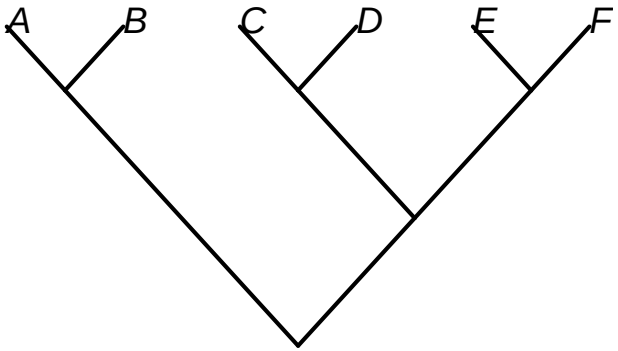
Anything strange?

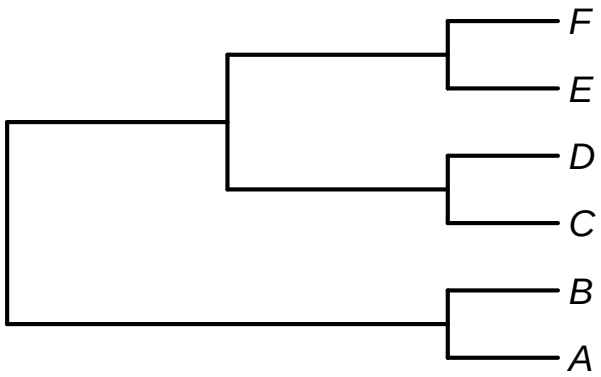


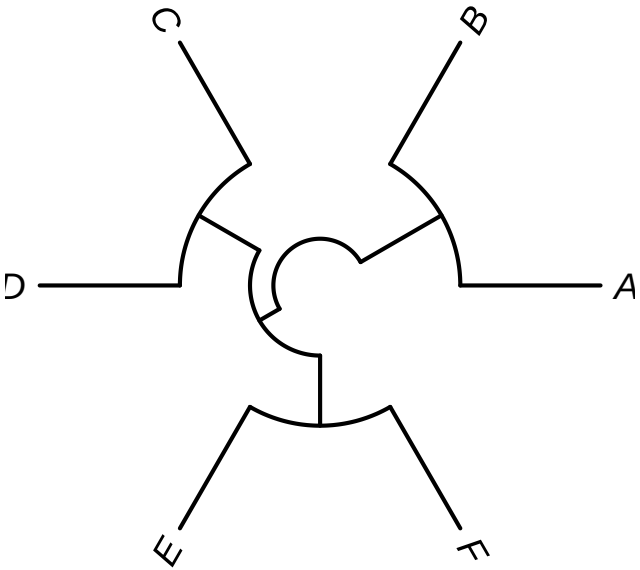
From Omland et al., 2005.

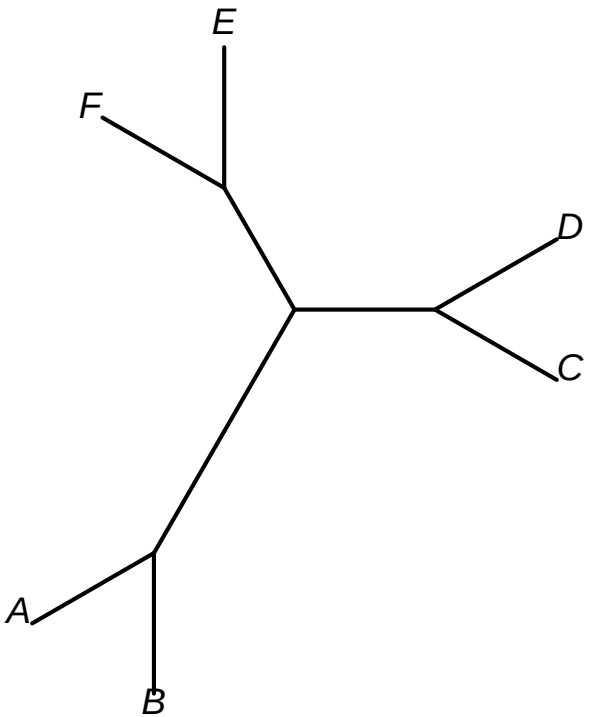
- Left does not mean old
- Outgroups (and outgroups need not be primitive)
- Cannot say which is oldest/youngest/most derived/most complex

Just different representations (if unrooted? if rooted?)





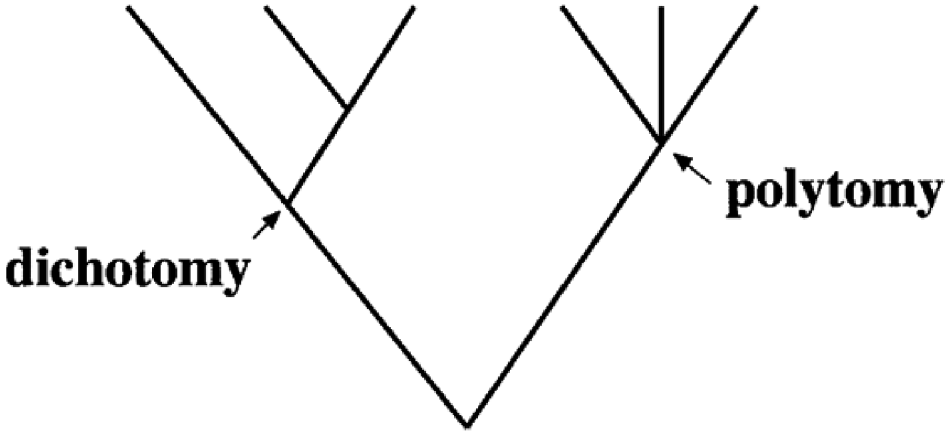


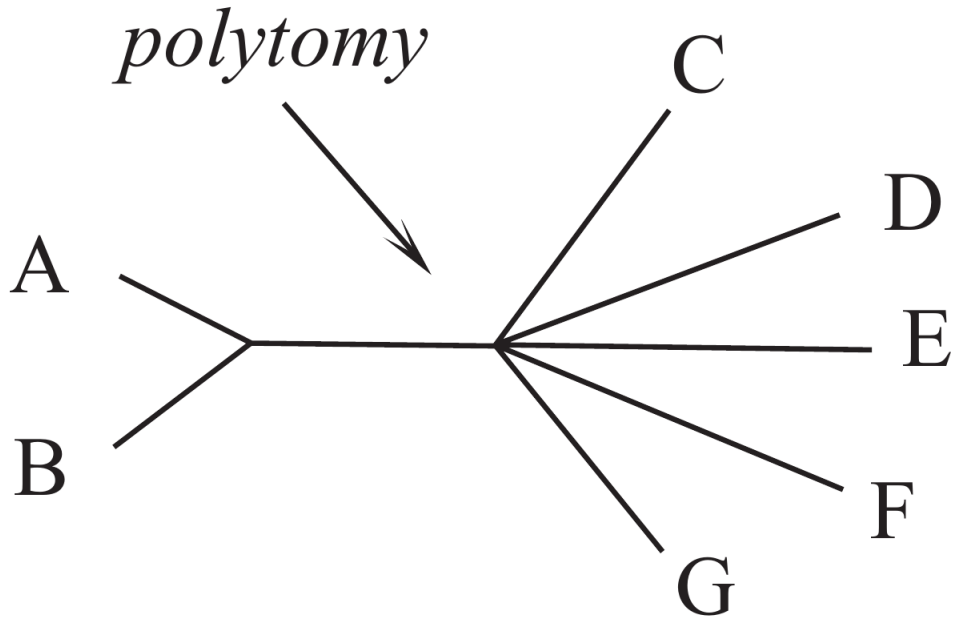


Representations: do it with MEGA at home

- Open MEGA.
- Build a tree with Drosophila data set.
- Change the root.

Polytomies





Left figure from Xiong, 2006. Right figure from Vandamme, in Lemmey et al., 2009.

Polytomies

- How are multifurcations interpreted.
- How are multifurcations represented: binary splits with 0 length.
  - Beware of representations!

Exercises

Open the paper by Capra and Kotska about DNA methylation. Locate the figure where they have something like a tree, and identify if:

- It is rooted or unrooted (and why)
- It is scaled or unscaled
- Are there any polytomies

Counting trees: why?

A simple request:

- You have 50 sequences.
- You want to find the best phylogeny.
- Build/construct all phylogenies and compare them.
- So . . . how many trees do we need to consider?

Guess: how many rooted bifurcating trees?

For 50 sequences/species

1. 10^3 to 10^5
2. 10^7 to 10^{10}
3. 10^{15} to 10^{50}
4. 10^{70} to 10^{90}
5. 10^{100} to 10^{150}

Counting: key results

(Details in Appendix)
For n taxa/leaves/terminal nodes:

- If unrooted tree
  - $(n-2)$ internal nodes
  - $(2n-2)$ total nodes
  - $(2n-3)$ branches
  - $(2n-5)!!$ trees
- If rooted tree
  - $(n-1)$ internal nodes
  - $(2n-1)$ total nodes
  - $(2n-2)$ branches
  - $(2n-3)!!$ trees

Counting: exercise (Do it now)

You have downloaded a small data set of 12 protein sequences from NCBI and you want to reconstruct their phylogenetic history. What is the total number of possible . . .

- . . . unrooted trees
- . . . rooted trees

The moral of counting

- Counting is important.
- We need an idea of the size of our problems before jumping into them.

Can we reconstruct large phylogenies?

Yes, definitely.

- Some methods quickly obtain a phylogeny without looking through existing alternatives.
- Other methods do not examine ALL possible alternatives.

What are we reconstructing the history of?

- Species?
- Genes?

Species vs. gene phylogenies

- What is the difference?

  Species trees Branching points represent speciation events.

  Gene trees Branching points: divergence of the gene sequence.
  Branching points might also represent gene duplication events (not necessarily).

- Might, or might not, coincide.
- For reconstructing speciation events, we want to use orthologous genes.

Gene trees: recall these facts

"Genes have gene trees because of gene replication. As a gene copy at a locus in the genome replicates and its copies are passed on to (...) offspring, branching points are generated" (Maddison, 1997).

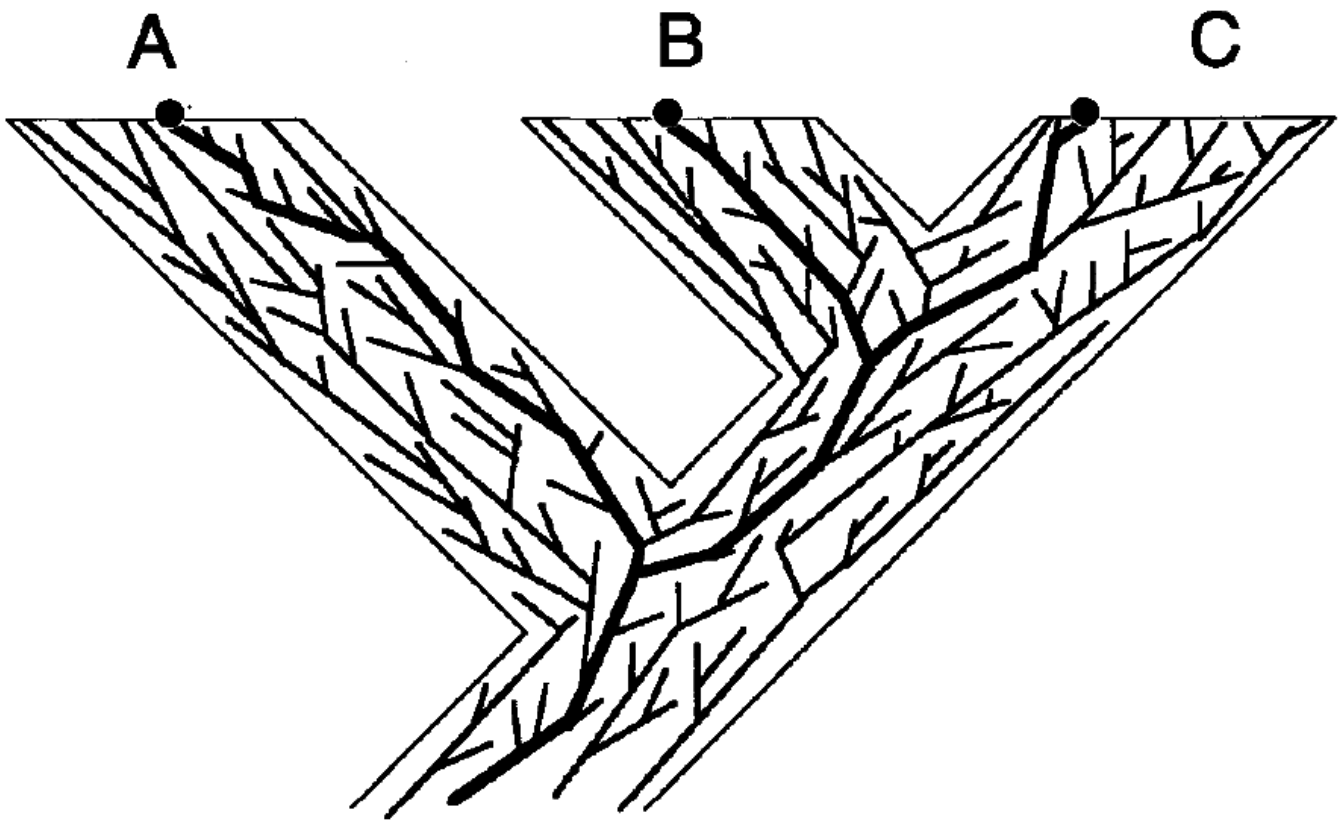
"When dealing with a gene that has polymorphic sites in the parent and daughter species, the nodes never really reflect the speciation event, but merely separation between different alleles." (Vandamme, in Lemey et al., 2006)

Why the difference between species and gene trees?

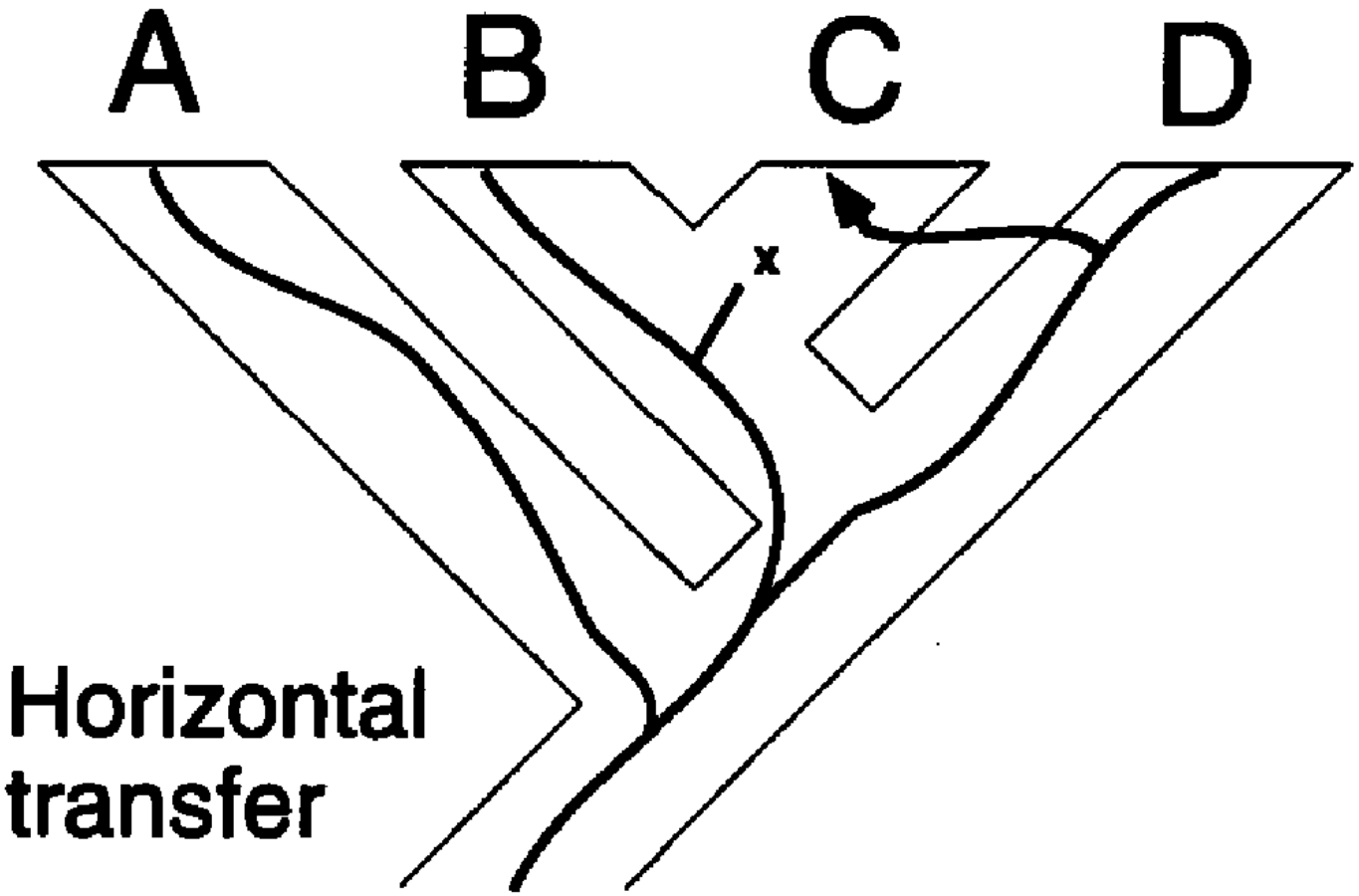
- Horizontal gene transfer
- Deep coalescence or lineage sorting: ancestral polymorphisms that persist through speciation events.
- Gene duplication (and extinction).
- And we reconstruct from samples (e.g., the sequence of hemoglobin of one **specific** cow or cows).

All of the figures for this section from Maddison, 1997, *Systematic Biology*, 46

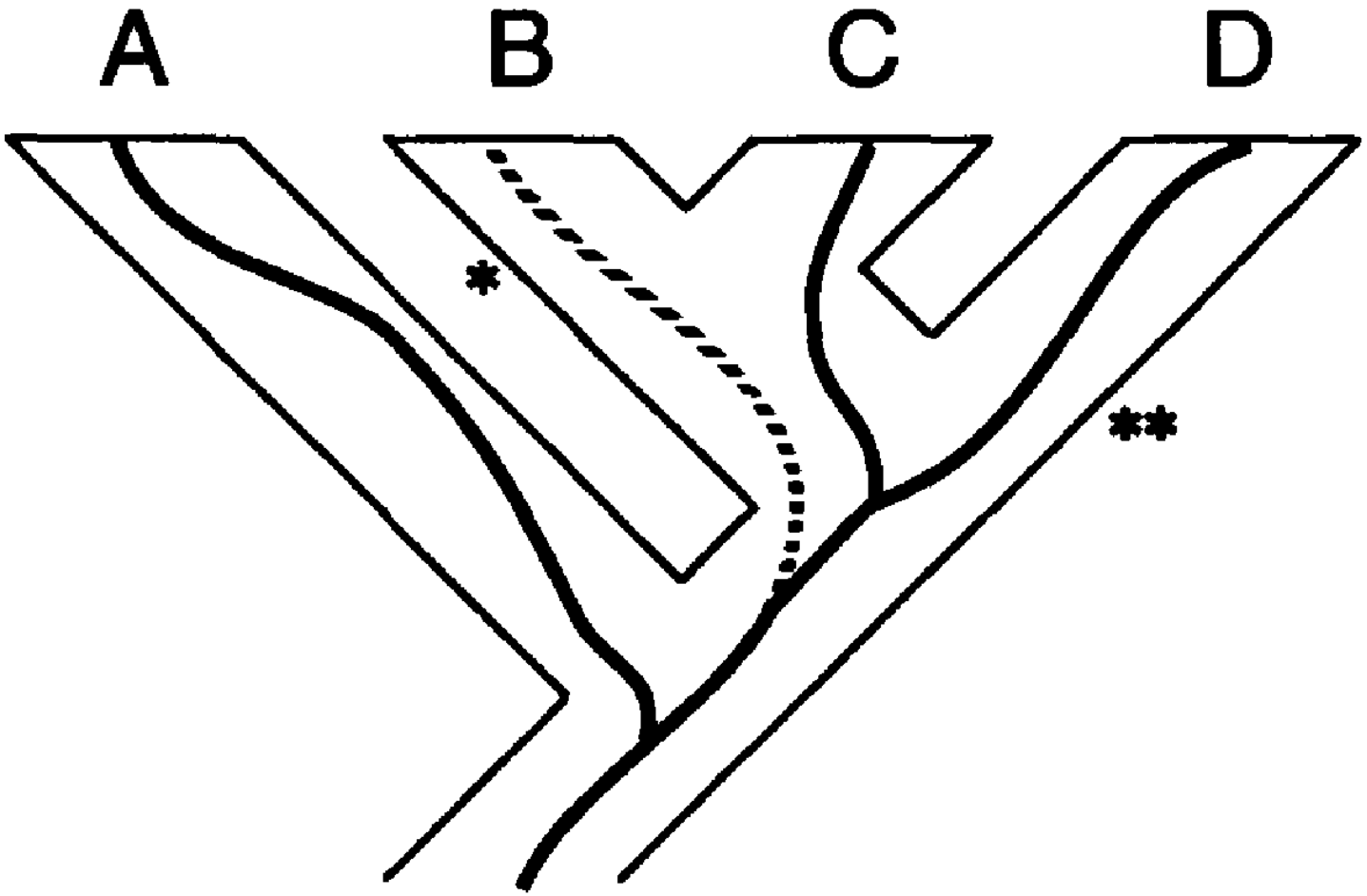
No problem here



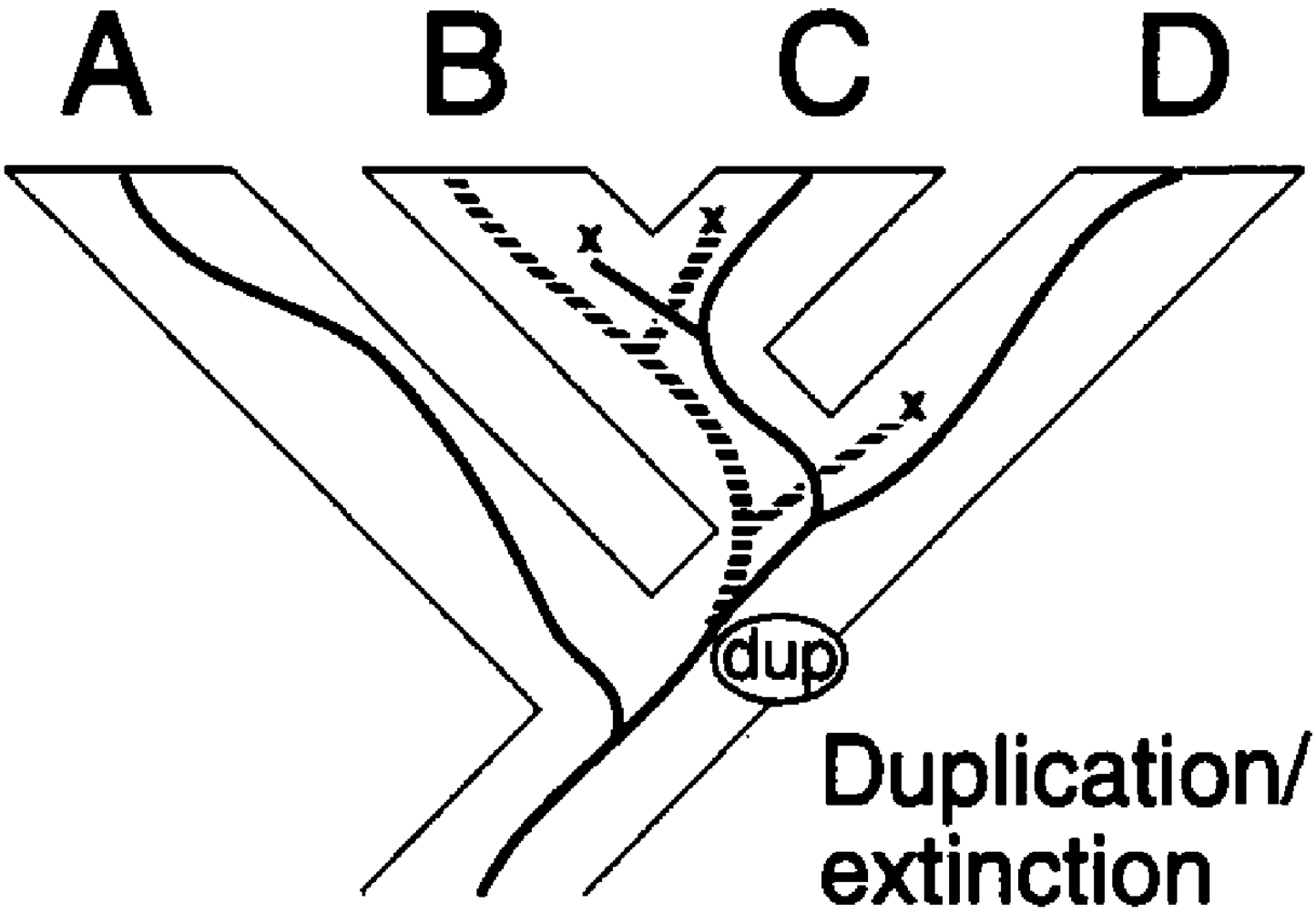
Horizontal gene transfer



Deep coalescence/species sorting



Gene duplication and extinction (“paralogous sampling”)



How likely?

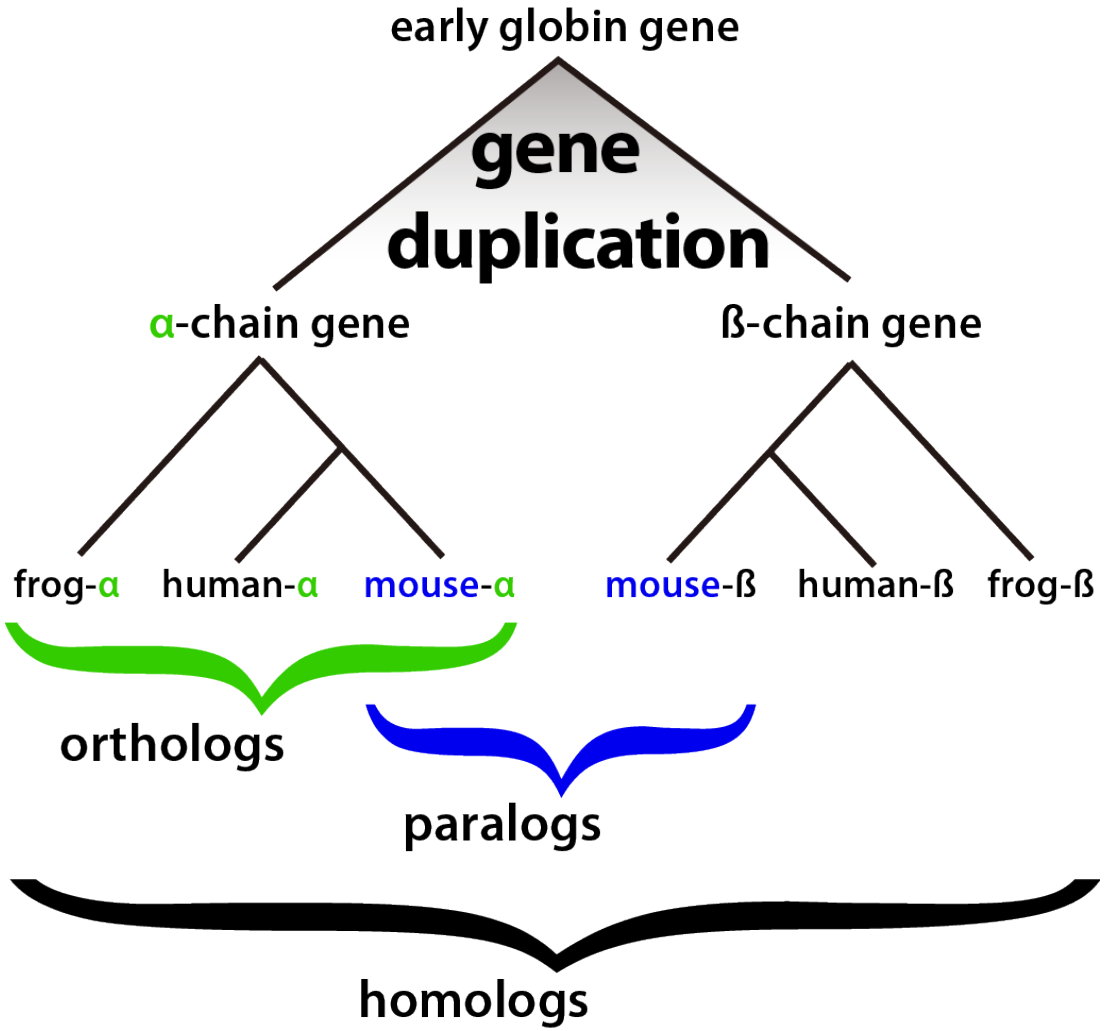
Horizontal transfer Type of organism and how closely related.

Deep coalescence Depends on speciation speed and population size.

What should we do?

- If you care about a gene, reconstruct the gene tree.
- If you care about species/speciation:
  - use several genes (yes, but how?)
  - try to avoid and/or disentangle possible causes (lineage sorting, paralogous sampling and gene duplication, etc)

Exercise: Is this a species or a gene tree?



From http://commons.wikimedia.org/wiki/File:Homology.png
http://commons.wikimedia.org/wiki/File%3AHomology.png

Counting trees: key elements of arguments

- Find out number of internal nodes for a given number, n, of species/leaves/terminal nodes/taxa.
- Find out number of branches/edges.
- Express number of trees for n species as "something * number of trees for $(n-1)$ species." (And then use a recursive argument down to $n = 3$ for unrooted trees).
- Realize that number of rooted trees (for n species) is "something else * number of unrooted trees for n species."

As extra help

- Get a piece of paper and draw them.
- Start with unrooted trees. $n = 1, 2, 3, 4$ species.

Number of species	Number of unrooted trees
1	something
2	something
3	something
4	something
...	...

Let's count: Edges and nodes (bifurcating trees)

(Taken from Durbin et al., 1998 and Felsenstein, 2004)

Suppose n terminal/extant species/sequences. (Take a piece of paper. Set $n = 3$ and then $n = 4$).

- How many nodes if tree is rooted?
  - n terminal nodes (or taxa, or leaves).
  - $n - 1$ internal nodes: why?
  - Thus: $(2n - 1)$ nodes and $(2n - 2)$ edges/branches.
- If unrooted?
  - One fewer of each:
  - $(2n - 2)$ nodes and $(2n - 3)$ edges/branches.

Counting: Unrooted to rooted.

- How do we turn an unrooted tree into a rooted one? Add a root.
- Where?
- To any branch.
- Since there are $(2n - 3)$ edges in an unrooted tree, an unrooted tree with n leaves/taxa produces $(2n - 3)$ rooted trees.

Counting: number of unrooted (is something * number of unrooted of one fewer species)

- Let's add a new species, not a root, to an unrooted tree.
- An unrooted tree with n species can have a new species added at any of $(2n - 3)$ places.
- We are done!

Counting rooted: really, are we done?

- Suppose $n = 4$.
- How many unrooted trees do I get if I add a fifth species? In an unrooted tree with 4 species I can add a fifth species in any of the internal branches, so at any one of $(2 * 4 - 3) = 5$ places. Thus, the number of unrooted trees for five species is 5 times the number of trees I have with 4 species.
- How many do I have with 4 species? I can add a fourth species at $(2 * 3 - 3) = 3$ places. Thus, I have 3 times the number of trees I have with 3 species.
- How many do I have with 3? 1. (Draw it!)

Counting rooted: really, are we done?

- Suppose $n = 4$.
- How many unrooted trees do I get if I add a fifth species? In an unrooted tree with 4 species I can add a fifth species in any of the internal branches, so at any one of $(2 * 4 - 3) = 5$ places. Thus, the number of unrooted trees for five species is 5 times the number of trees I have with 4 species.
- How many do I have with 4 species? I can add a fourth species at $(2 * 3 - 3) = 3$ places. Thus, I have 3 times the number of trees I have with 3 species.
- How many do I have with 3? 1. (Draw it!)
- $1_{\text{three species}} * 3_{\text{four species}} * 5_{\text{five species}}$

Counting rooted: really, are we done?

- So ...
- If we add species number n, we have $(2 * (n-1) - 3) = 2n - 5$ as many unrooted trees as for species $n - 1$.
- Number of unrooted trees for n species: $(2n - 5)!!$
- The "!!" is like a factorial, skipping numbers. E.g.: $9!! = 9 * 7 * 5 * 3 * 1$.

And rooted trees?

- I can add a root at any of the $2n-3$ edges. So I have $2n-3$ as:
  many rooted trees as unrooted trees.
- Number of rooted trees: $(2n-3)!!$.

By the way that is a recursive relationship

- We express the number of trees with n species as: *something* $*$ number of trees with $(n-1)$ species.
- (But we can compute it iteratively)

BIBMS: Phylogenetic reconstruction (II). Method overview and models of substitution

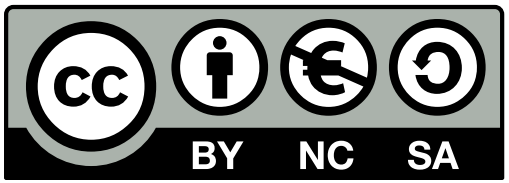
Ramón Díaz-Uriarte

Dept. Bioquímica
Universidad Autónoma de Madrid
Madrid, Spain
ramon.diaz@iib.uam.es
http://ligarto.org/rdiaz

October 2018

(Rev: 1fa718b)

License and copyright



Outline

1. Overview: key steps, main methods, relationships
   - Main steps
   - DNA or proteins?
   - Models of substitution: their role
2. Models of DNA and amino acid substitution
   - Distances (first attempt)
   - Substitution matrices
   - Jukes-Cantor
   - Other models
   - Models for aminoacid substitution
   - Choosing model and parameters
3. Why do we care?

1. Overview: key steps, main methods, relationships
   - Main steps
   - DNA or proteins?
   - Models of substitution: their role
2. Models of DNA and amino acid substitution
3. Why do we care?

Main steps to build a tree

1. Select sequences
2. Align them
3. Decide on a model of substitution for nucleotides or AAs.
4. Build tree(s): find the best one(s)
5. Evaluate tree(s): how reliable is/are the tree(s)

DNA or proteins for constructing phylogenetic trees?

- Closely related organisms: DNA often better (faster evolution).
- For not-so-closely related: DNA might have changed too much (be saturated).
- High quality multiple alignment: easier with proteins.
- ML (Maximum likelihood) and Bayesian methods often too slow with proteins. (But then, if that is what you want . . .)

DNA or proteins for constructing phylogenetic trees? (II)

Additional considerations.

- If using proteins, cannot differentiate and use info from silent substitutions.
- Nucleotides: third codon often a different rate; must be modeled. (No need to worry about this with proteins).
- DNA allows study of synonymous vs. non-synonymous substitution rates: selection.
- With DNA can use non-coding regions: sometimes these can vary greatly in rates; some can have neutral rates.
- With DNA we can use pseudogenes.

1. Overview: key steps, main methods, relationships

2. Models of DNA and amino acid substitution
   - Distances (first attempt)
   - Substitution matrices
   - Jukes-Cantor
   - Other models
   - Models for aminoacid substitution
   - Choosing model and parameters

3. Why do we care?

Main steps to build a tree

1. Select sequences
2. Align them
3. **Decide on a model of substitution for nucleotides or AAs.**
4. Build tree(s): find the best one(s)
5. Evaluate tree(s): how reliable is/are the tree(s)

The matrix of distances (Number of substitutions)

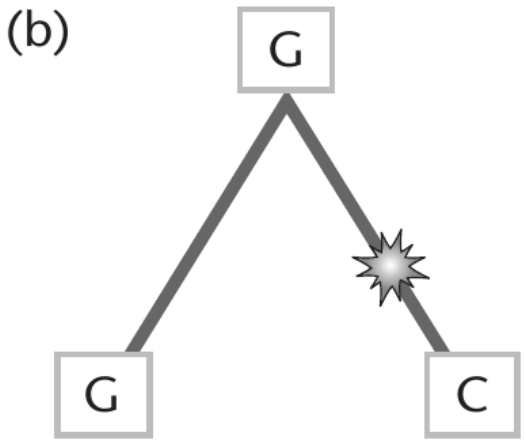
$$
\begin{array}{c} \\ H.s. \\ Frog \\ Chicken \\ M.m. \end{array}
\begin{array}{c}
\begin{array}{cccc} H.s. & Frog & Chicken & M.m. \end{array} \\
\begin{pmatrix} 0 & 5 & 5 & 3 \\ 5 & 0 & 6 & 6 \\ 5 & 6 & 0 & 4 \\ 3 & 6 & 4 & 0 \end{pmatrix}
\end{array}
$$

p-distance

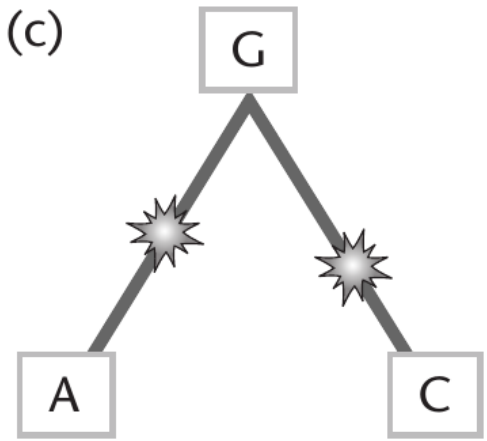
A minor thing:

- Instead of number of changes we will often want to use the **p-distance**: proportion of nucleotide sites that differ.
- (Automatically normalizes by number of comparisons made)
- Divide the matrix by the number of comparisons.

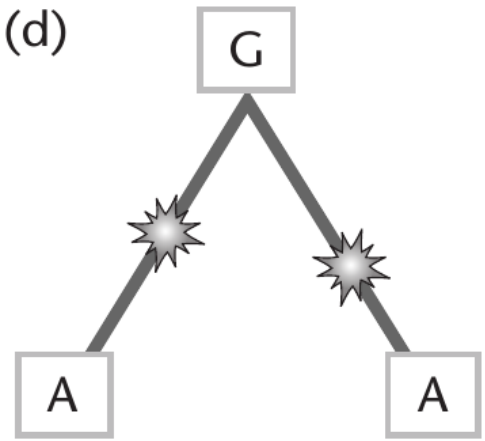
Number of changes underestimates . . .



One substitution happened – one is visible



Two substitutions happened – only one is visible



Two substitutions happened – nothing visible

(Higgs & Attwood, 2005.)

We need a model

- Multiple alignment $\rightarrow$ amount change
- The distances we use to reconstruct trees are supposed to reflect amount of change.

Number/proportion of changes are not good enough

- Underestimation of true number of changes
- (If number of changes are not good enough, neither are **p-distances**.)

We depend on the model

- No such thing as “model-free” phylogenetic reconstruction
- There is no such thing as “model-free” inference (here or anywhere else).
- No model, no inference. (e.g, E. Sober, 1998, “Reconstructing the past”).

Substitution matrices

$$S(t) = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} P(A|A,t) & P(C|A,t) & P(G|A,t) & P(T|A,t) \\ P(A|C,t) & P(C|C,t) & P(G|C,t) & P(T|C,t) \\ P(A|G,t) & P(C|G,t) & P(G|G,t) & P(T|G,t) \\ P(A|T,t) & P(C|T,t) & P(G|T,t) & P(T|T,t) \end{pmatrix} \end{array}$$

(ugly those column and row names, thus often)

$$S(t) = \begin{pmatrix} P(A|A,t) & P(C|A,t) & P(G|A,t) & P(T|A,t) \\ P(A|C,t) & P(C|C,t) & P(G|C,t) & P(T|C,t) \\ P(A|G,t) & P(C|G,t) & P(G|G,t) & P(T|G,t) \\ P(A|T,t) & P(C|T,t) & P(G|T,t) & P(T|T,t) \end{pmatrix}$$

Substitution matrices and distances

Evolutionary distance $\rightarrow$ prob. change $\rightarrow$ p-distance.

- If we know probabilities of change (the previous matrix)
  - We can obtain the probability that a given site differs between two sequences after some time
  - We can obtain the expected number of sites with change, or the p-distance corresponding to a given evolutionary distance
- If we measure p-distance
  - We can infer the evolutionary distance

A model gives the relationship: p-distance $\leftrightarrow$ evolutionary distance
(a formula to go from one to the other)

(Yes, you use formulas like that all the time)

- At supermarket x, one can of lentils costs 79 cents.
- I have 20 cans of lentils in my cart; I will pay . . .
- I have paid 7.9 euros; there were . . . cans of lentils in my cart.

(Image from

https://st1.tudespensa.com/rep/14b9/imagenes/41526/109/las-lentejas-de-la-abuela-litoral-lata-440-gr.jpg)

Jukes-Cantor

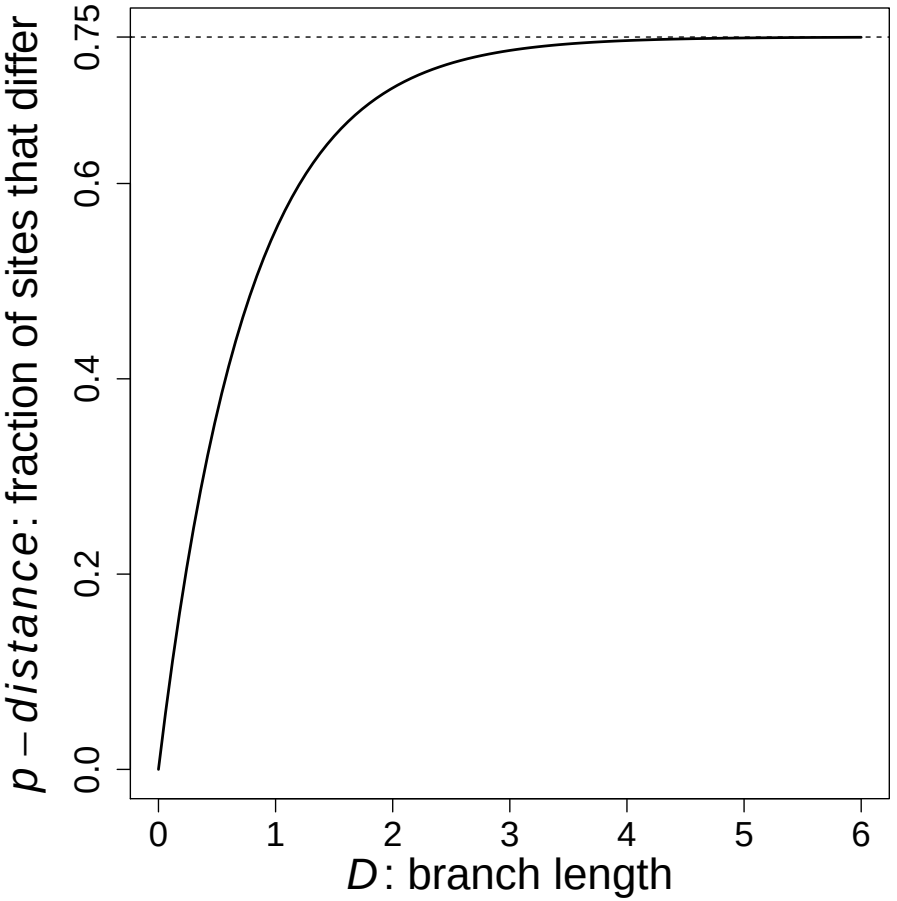
$$D = -\frac{3}{4}\ln(1 - \frac{4p}{3})$$

where

- D: true distance (true number of nucleotide substitutions per site; some books use K or other terms)
- p (some books use d, D, or f): fraction of sites that differ, **p distance**

(So instead of price per can of lentils you have p-distance and instead of total amount per cart you have true evolutionary distance.)

Jukes-Cantor: a figure



$$p = \frac{3}{4}(1 - \exp^{-\frac{4D}{3}})$$

Jukes-Cantor: assumptions

- All nucleotides undergo transitions at same rate α.

Jukes-Cantor: assumptions and details

All nucleotides undergo transitions at same rate α.

- This is the rate matrix: total rate of change is 3α:

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \end{array}$$

- The equilibirum frequency of all nucleotides is the same: $q_A = q_C = q_G = q_T = 0.25$ [this is really a consequence].

Kimura's 1980 model

And if transitions more common than transversions? Kimura's model. Rate matrix:

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \beta \\ \beta & \alpha & \beta & -2\beta-\alpha \end{pmatrix} \end{array}$$

- When $t = \infty$, also $q_A = q_C = q_G = q_T = \frac{1}{4}$.

(Now, your cart includes cans of lentils and eggs, and eggs and lentils have different prices. But you can still figure out the total amount you will pay.)

Other models?

- Yes, a bunch of others, commonly used.
- F84 (Felsenstein 84) and HKY (Hasegawa, Kishino, Yano): like Kimura, but arbitrary base frequencies.
- Tamura's adjusts for GC content.
- GTR (general time reversible)
- ...

And can rates vary among sites?

- YES!!!
- We model the distribution of the rates (usually a Gamma distribution).

What do these models give us

- A way of multiple alignment $\rightarrow$ evolutionary distance
- A way of making probabilistic statements about each position in alignment:
  - How likely is it that we get, say, A from C in t time?
  - How likely is it that C in sequence 1 and G in sequence 23 have the same common ancestor?

Substitution models for proteins

- Not 4x4 but 20x20.
- Most empirically derived.
- PAM
- PAM in particular can be easily turned into something that looks similar to J-C, Kimura, etc, matrices.
- **JTT** (Jones-Taylor-Thornton)
- **Poisson** model to correct for multiple substitutions:
  - Uses the number of changes.
  - Adds a correction term ($D = -\ln(1 - p)$)
  - Also applicable to nucleotides.
- ...

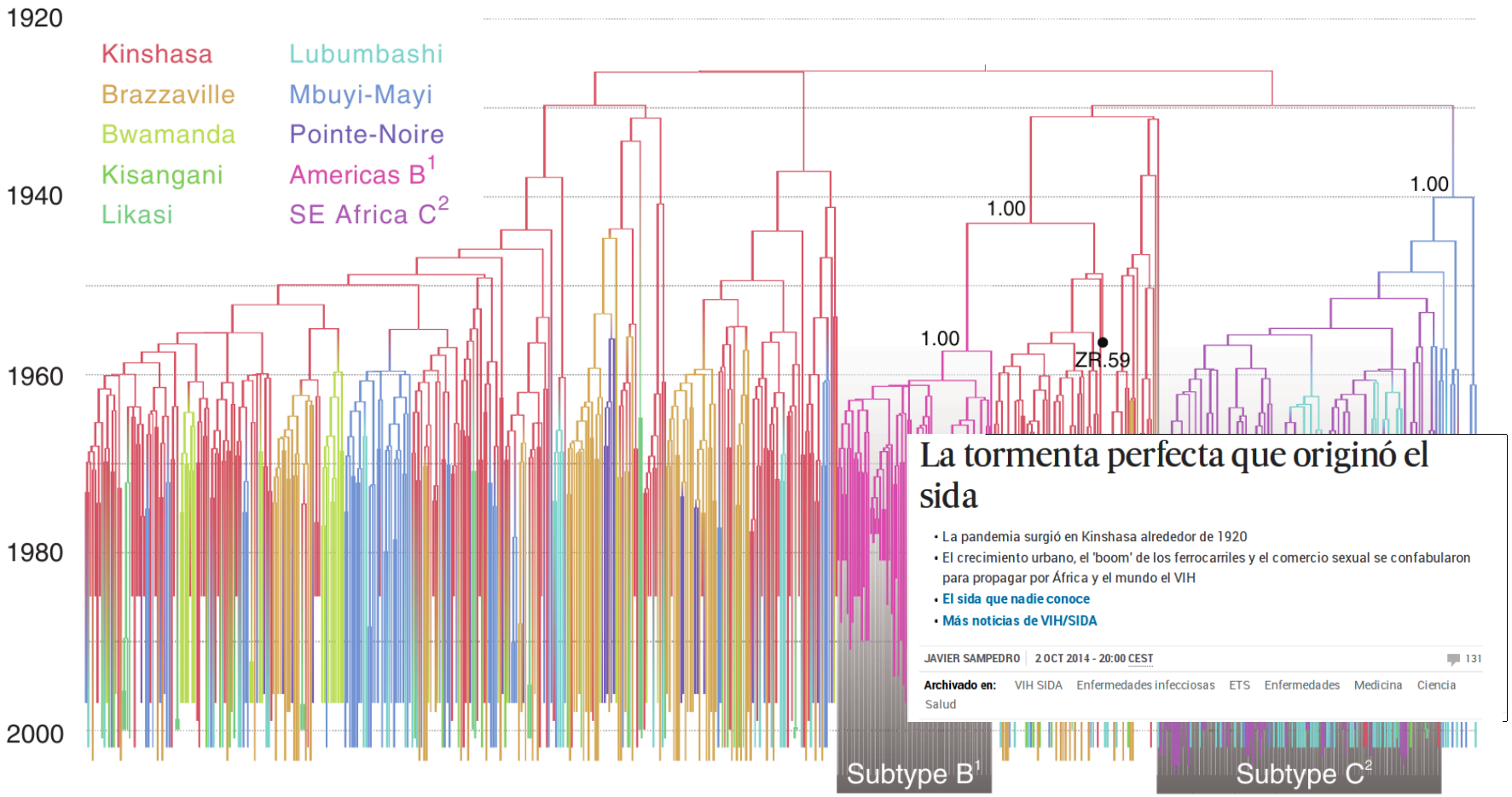
Choosing models and parameters

- Parameters: They can be estimated while/before we carry out the tree-building.
- Model: we can assess fit, and choose best fitting one (or use a mixture).
- MEGA: under “Models”.
- JModelTest (Posada, Crandall, et al.)
- etc
- Do not use uncorrected distances.

Exercise

- Open MEGA. Find, in the Help, where the models of substitution are discussed (hint: it is in “Part IV: Evolutionary analysis”).
- Find (and look over quickly) the Jukes-Cantor and the Kimura 2-parameter explanation.
- How many other models are discussed?
- And what about models for amino acid sequences?

HIV



La tormenta perfecta que originó el sida

- La pandemia surgió en Kinshasa alrededor de 1920
- El crecimiento urbano, el 'boom' de los ferrocarriles y el comercio sexual se confabularon para propagar por África y el mundo el VIH
- El sida que nadie conoce
- Más noticias de VIH/SIDA

JAVIER SAMPEDRO | 2 OCT 2014 - 20:00 CEST 131

Archivado en: VIH SIDA Enfermedades infecciosas ETS Enfermedades Medicina Ciencia Salud

Fig. 1. Time-scaled phylogeographic history of pandemic HIV-1. Branch colors represent the most probable location of the parental node of each branch. The respective colors for each location are shown in the upper left. U.S./Haiti/Trinidad subtype B and southeast African subtype C lineages are highlighted by boxes with a gradient shading, along with the posterior probabilities for their ancestral nodes. The tip for the ZR59 sequence is highlighted with a black circle.

From Faria et al., 2014. *Science*, 346 (3-October-2014): 56–61 (and from *El Pais* http://elpais.com/elpais/2014/10/02/ciencia/1412260639_097968.html)

Ebola

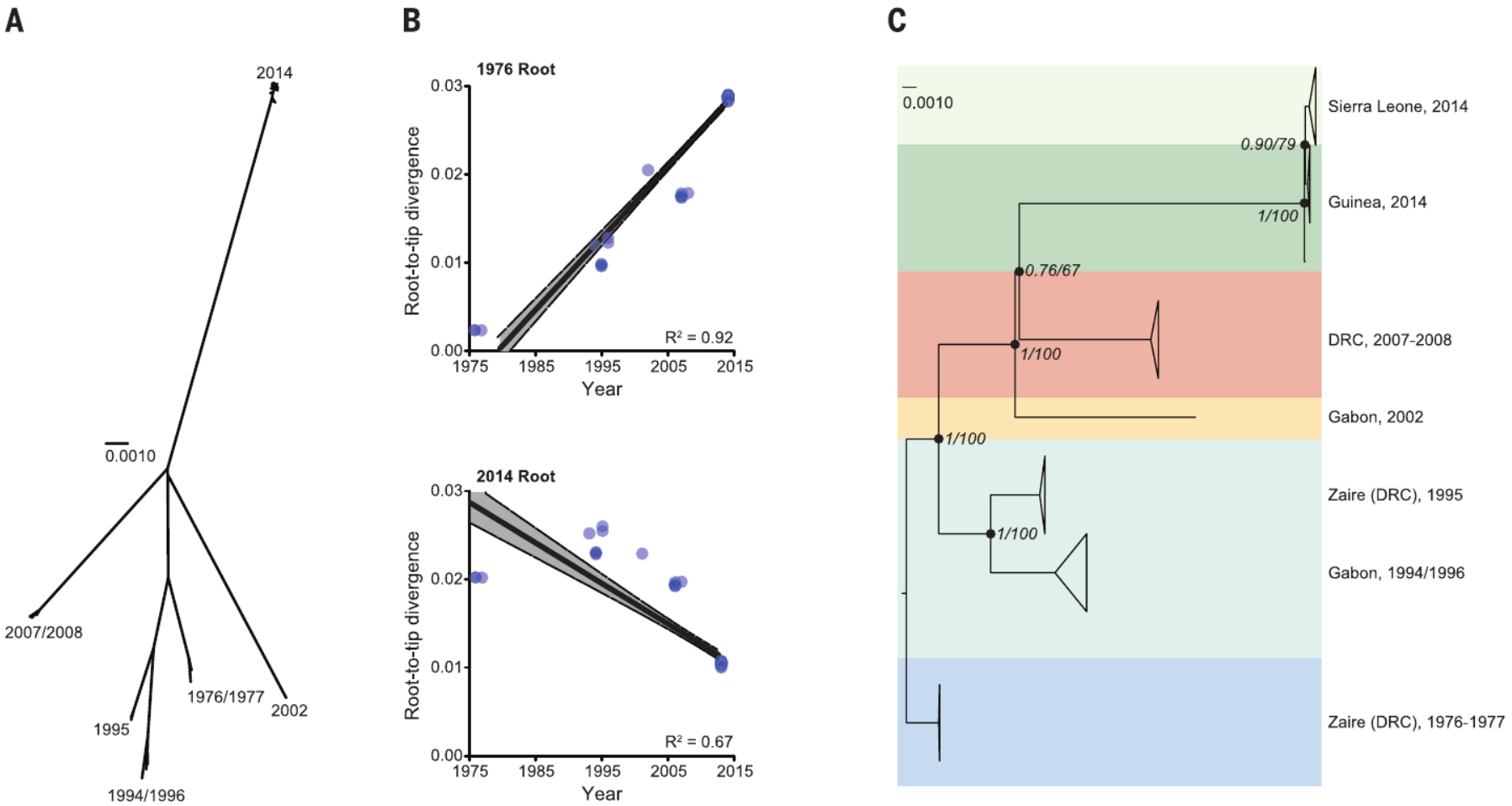


Fig. 2. Relationship between outbreaks. (**A**) Unrooted phylogenetic tree of EBOV samples; each major clade corresponds to a distinct outbreak (scale bar = nucleotide substitutions per site). (**B**) Root-to-tip distance correlates better with sample date when rooting on the 1976 branch (R^2 = 0.92, top) than on the 2014 branch (R^2 = 0.67, bottom). (**C**) Temporally rooted tree from (A).

From Gire et al., 2014. "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak", *Science*, 345 (12-September)

BIBMS: Phylogenetic reconstruction (III). Methods and odds and ends

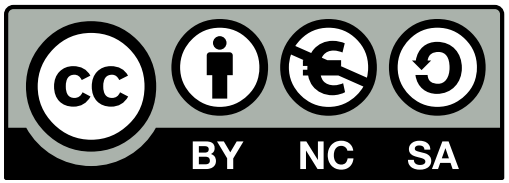
Ramón Díaz-Uriarte

Dept. Bioquímica
Universidad Autónoma de Madrid
Madrid, Spain
ramon.diaz@iib.uam.es
http://ligarto.org/rdiaz

October 2018

(Rev: 1fa718b)

License and copyright



Outline

So we have a model for substitutions . . .

. . . now what?

- We can get a matrix of distances that reflect amount of evolutionary change
- We can compute probability of a given substitution

Main methods (I)

Distance-based methods Work with distances.

- From alignment to a distance
- (Summarize the alignment in a single number: evolutionary distance.)
- Tree that fits that distance

Character-based methods Use the alignment directly.

- Use sequence of characters directly.
- Find tree for that set of characters.
- Tree found/chosen with a model for the characters

Main methods (II)

Algorithmic Estimate a single tree from the data with an algorithm.

- Many distance based.
- Single tree: good and bad.

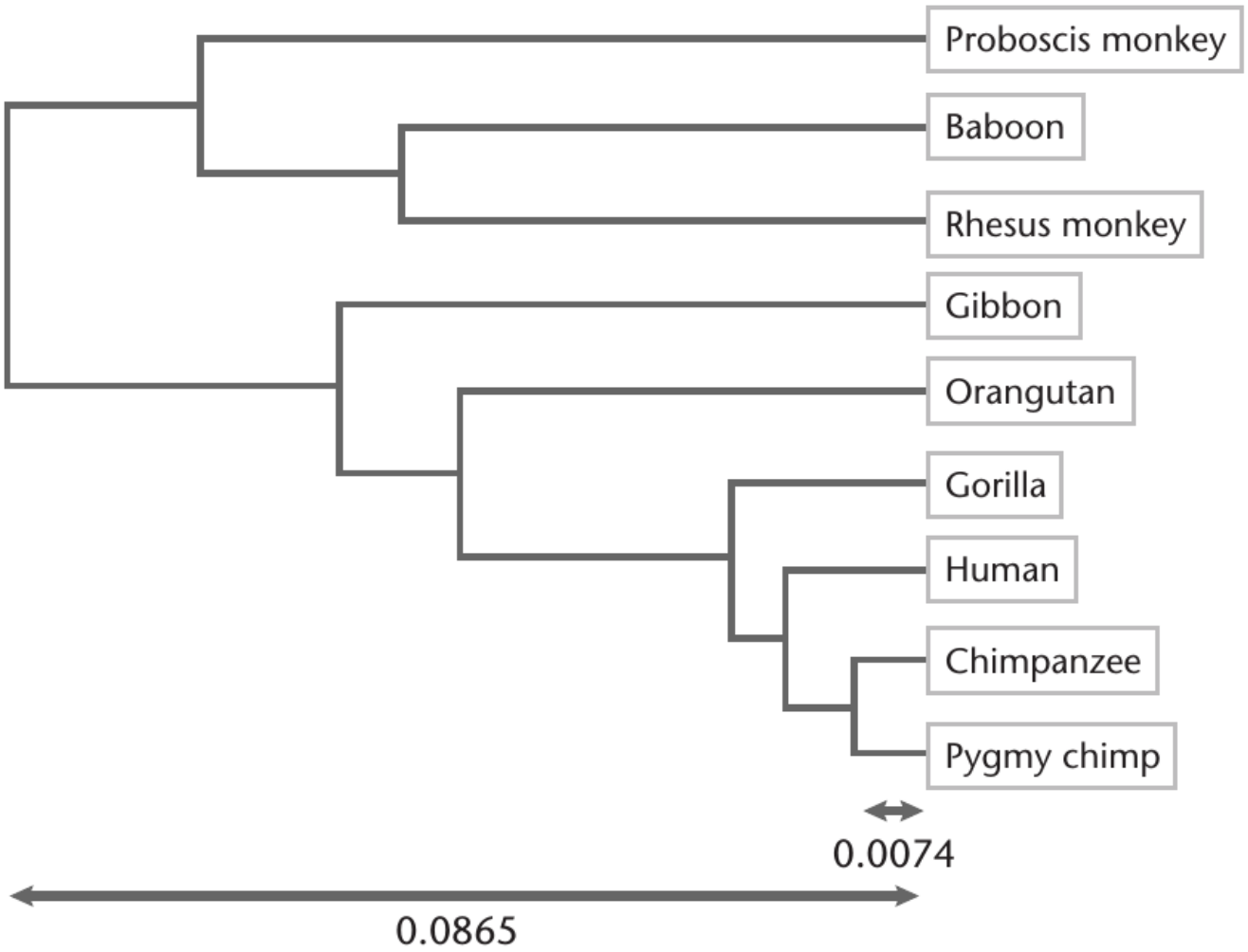
Tree-searching Build many trees, compare them, keep the best one(s).

- Character-based, some distance-based.
- Many trees: good and bad.
- Slower and how do we move in the space of trees?

A catalog of some methods

Method	Distance/ Character	Algorithmic/ Tree search
UPGMA	Distance	Algorithmic
NJ (Neighbor joining) et al. (BIONJ, . . .)	Distance	Algorithmic
Minimum Evolution	Distance	Tree search
Least squares (e.g., Fitch-Margoliash)	Distance	Tree search
Parsimony	Character	Tree search
ML (Max. likelihood)	Character	Tree search
Bayesian	Character	Tree search

UPGMA: a figure



From Higgs and Attwood, 2005.

UPGMA: main characteristics

- Uses distances.
- Returns a single rooted tree.
- All leaves equally distant from root $\Rightarrow$ molecular clock with constant rate.
- (Assumes distances are **ultrametric**; details in "Appendix")
- If **forces** the distances in the tree to fit a particular model, **even if the original distances do not fit that at all**.
- By forcing distances to fit a very restrictive model, no only is the figure distorted; **ancestor-descendant relationships can be seriously wrong**.
- Do not use this method for real in phylogenetic reconstruction
- (What about other uses?)

UPGMA: key elements of the algorithm

(Only if you want the details; skip otherwise!)

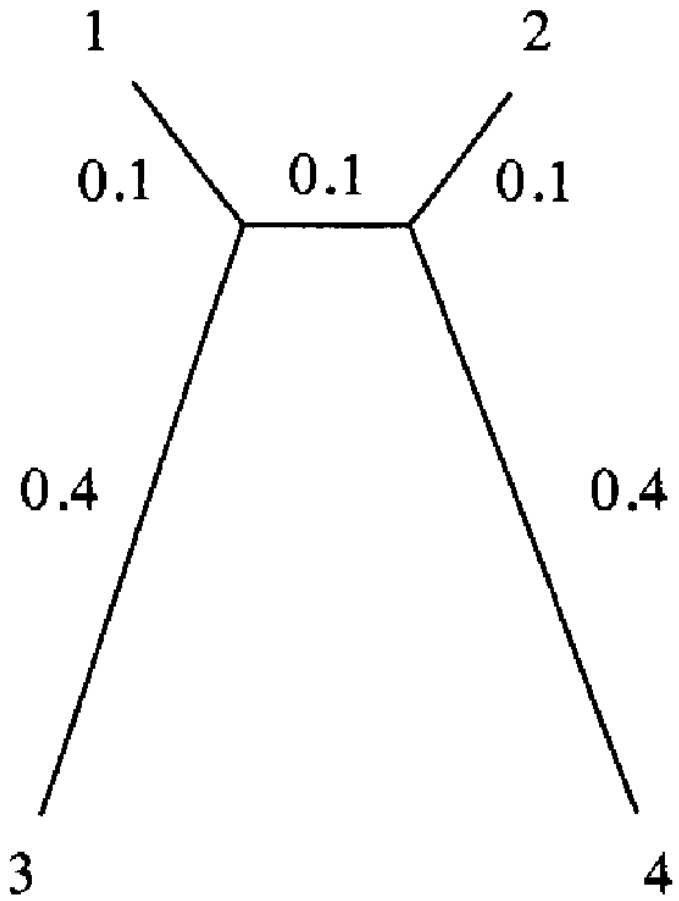
1. Start from the tips (“move up”).
2. Find pair of taxa with smallest distance.
3. Height of new node: place parent node at midpoint of branch.
4. Distance of any other node to new cluster: average of distance between “other” and members of new cluster.
5. Repeat until done.

(Details in ▸ “Appendix”)

Why do we even talk about this?

- So that you do not use it for phylogenetic tree building.
- Because widely used for clustering.
- To highlight differences between “other types of clustering” and phylogenetic tree building.
- To see what NJ does.

NJ: a picture



From Durbin et al, 1998. Notice the branch lengths!!

Neighbor Joining (NJ): key features of algorithm

- Uses distances.
- Returns a single unrooted tree.
- Start from the tips. Does not construct clusters (clades) but directly calculate distances to internal nodes
  1. Compute the average distance of each taxon, i, to each other taxa. “Net divergence”, “how far from the rest”.
  2. Correct pairwise distances by the net divergence. We get D_{ij}.
  3. Taxa with minimal D_{ij} put together in an internal node.
  4. Compute distance between the new node and its daughter taxa. Daughter taxa need not be equally distant from parent!
  5. Compute distances between new node and remaining taxa.
  6. Repeat until only two taxa left.
- Details in ▸ “Appendix”

Neighbor Joining (NJ): assumptions

- Assumes **additivity**: distances between any two nodes sum of lengths of all branches between them. No molecular clock assumption.
- Can we use NJ if distances deviate from additivity? Yes, but correct tree no longer guaranteed.

NJ: some features

- Fast.
- Often a very good tree.
- Use on its own, or as starting point for other more computationally intensive methods (parsimony, ML, Bayesian).
- There are other variants (see “Appendix”).

Exercise

Open the paper by Sottoriva et al.

- In less than 10 seconds: go to page 4 and answer if that figure is a phylogeny? How can you tell?
- In less than 10 seconds, go to p. 5, and answer if figure 4 B is showing a phylogeny. Do you think these are scaled or unscaled?
- In 20 seconds, give a more complete answer to the previous question: where do they specify how they built the phylogeny? What characters did they use?
- In less than 40 seconds: Go to p. 12. Are those phylogenies or something else? Find (Supplementary material) where they say how those were built. And what characters did they use?

Exercise

Open the paper by Wang et al.

- In less than 5 seconds, go to p. 4, and say if figure 3 d is a pylogeny, if it is rooted (and if so, how), and the method used.
- In another 15 seconds, find where, in the paper, are the details given (hint, go to the Methods, that for *Nature* tends to be “supplementary”, starting here on p. 7).
- Do you think this is a good or a bad idea?
- Do you see anything similar/different with what Sottoriva et al. do?

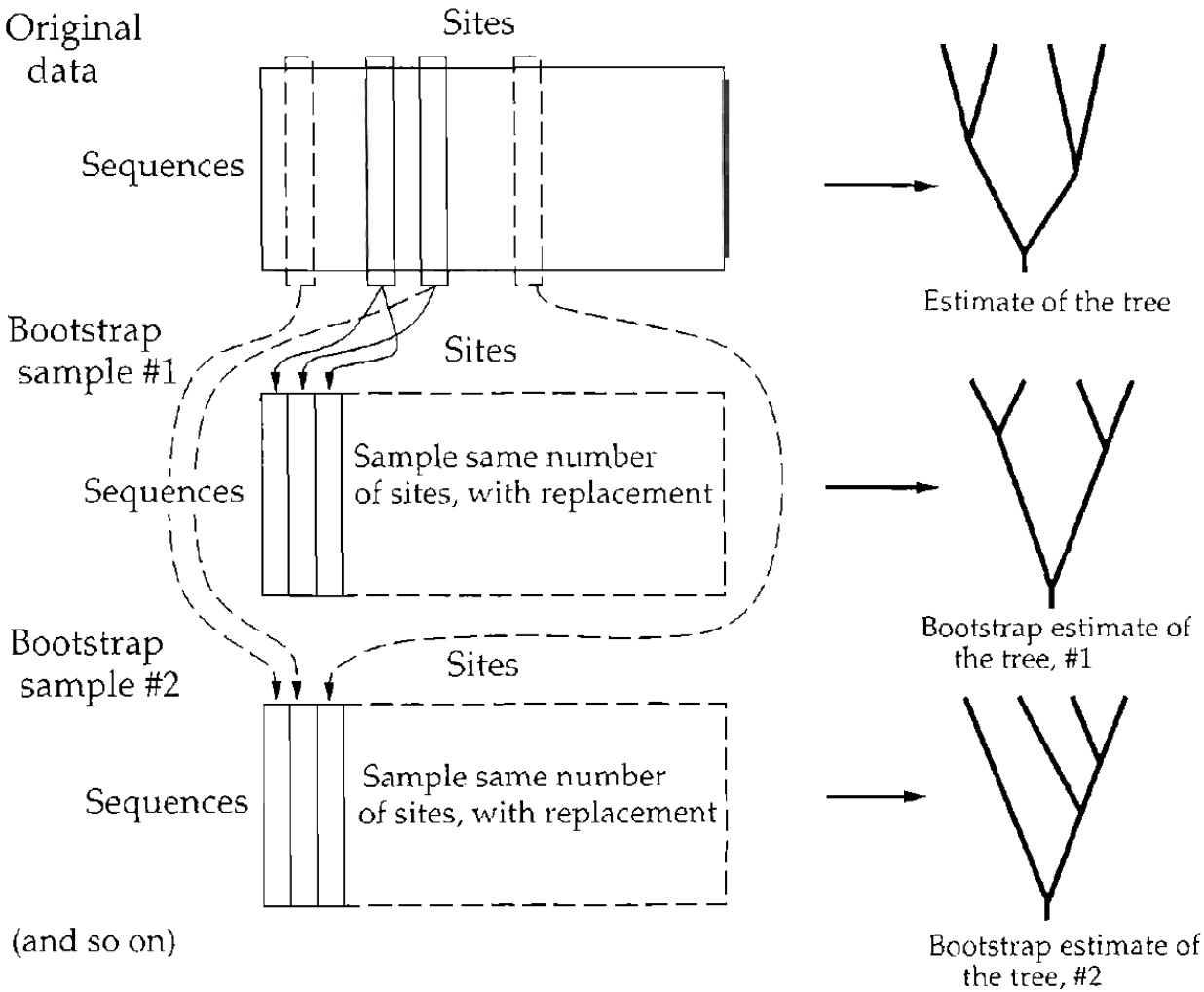
Moral

- Not every tree is a phylogeny, not every phylogeny looks like a dendrogram.
- You can use different types of characters.
- Details DO matter a lot.

How reliable is the tree?

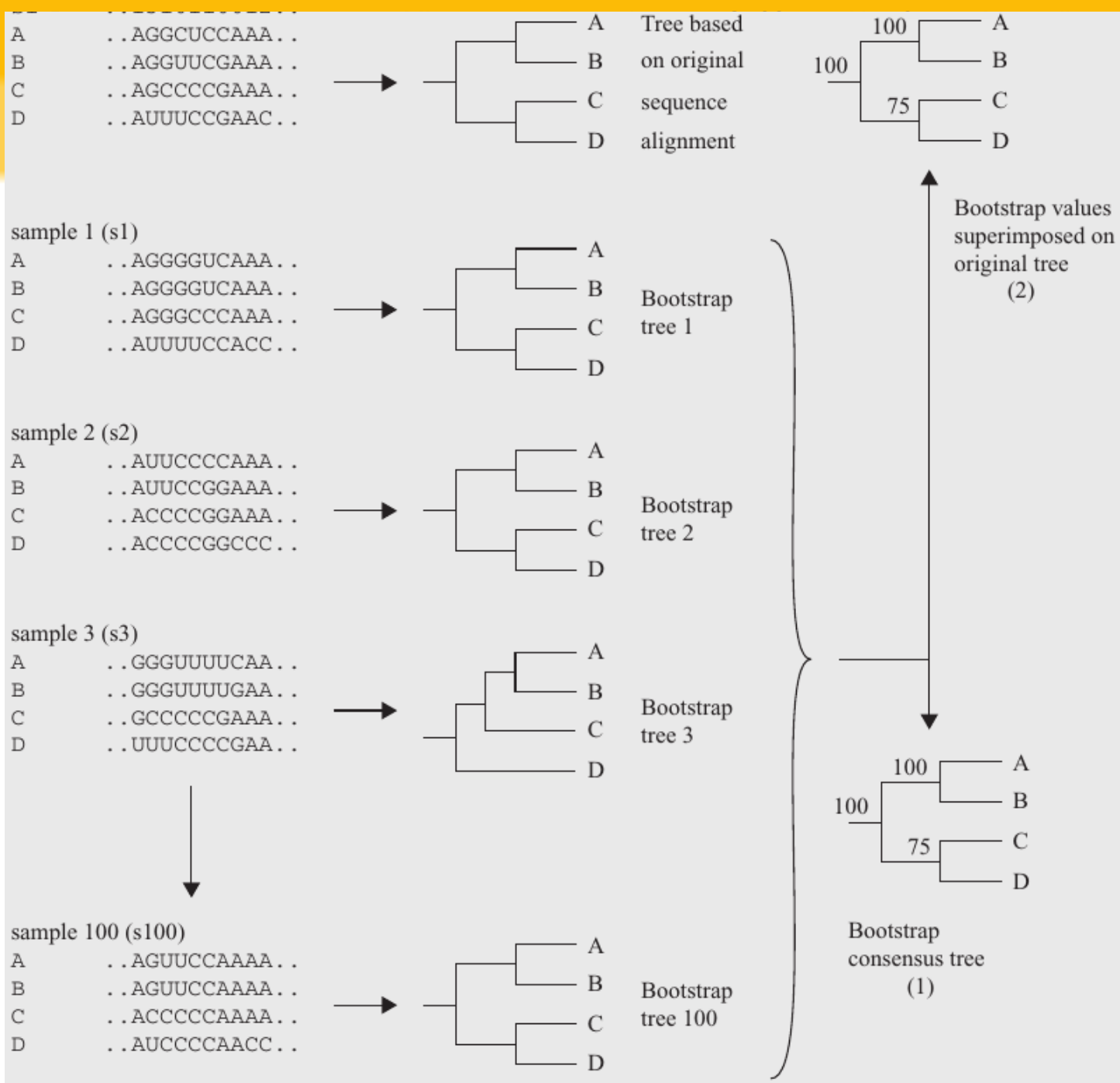
- **Reliability of group membership**: Are members of a group really members of that group? (emphasis on branches that split groups, not distances).
- (Interior branches, not clades).
- Resample (with replacement) the aligment and build trees.

Bootstrap



From Felsenstein, 2004.

Bootstrap



From Van de Peer, in Lemey et al, 2009. Beware of numbers on roots!

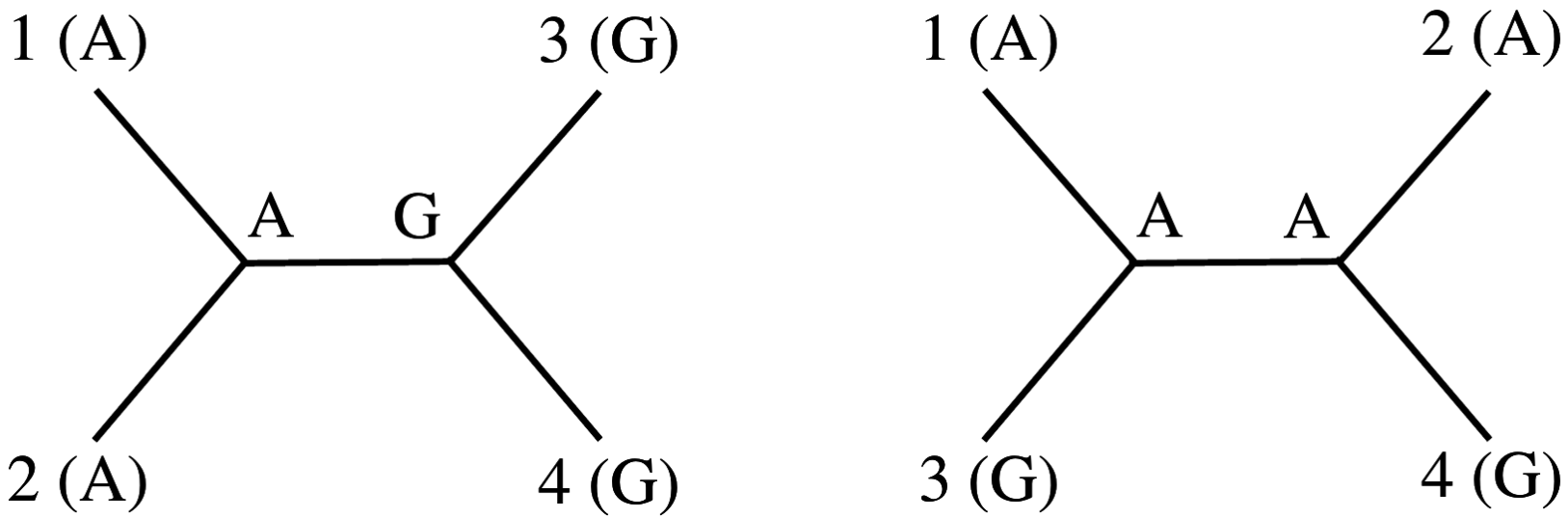
Bootstrap: miscell

- General statistical technique (and we will use it with ML and parsimony too).
- Number of replicates: please, nothing less than 200.
- Original tree need not be the same as *Bootstrap consensus tree*

Parsimony

Find the tree(s) that requires the smallest number of substitutions to explain the data.

(Only the leaves are observed!! Ancestors are hypothesized states)



We prefer the left one.

From Yang, 2006.

Parsimony: two things we need

- **A way of scoring a phylogenetic tree** or how to say if a tree is better than another tree: smaller number of substitutions is better. (details of an algorithm in ▸ “Appendix”).
- **A way of exploring tree space** to search for better trees. (This is not specific to parsimony).

Exploring space of trees

- Exhaustive only feasible for few taxa.
- Heuristic search methods (no guarantees we will hit the best).
- (More in ▸ “Appendix”)

Parsimony: assessing reliability

- Bootstrap.
- And there might be several equally good trees with original data: consensus tree from the original data.

Parsimony: issues

- Fast.
- Simple to understand.
- Robust to inter-site rate variation.
- What model is that anyway? Occam's razor?
  - Ch. 10 in Felsenstein
  - E. Sober, 1998, "Reconstructing the past"
- No principled way of exploring alternative weights/models (compare haphazard weighted parsimony with model comparison).
- Fast evolution and lots of reversals: problematic.
- Long branch attraction (because branch length is disregarded): statistically not consistent.

Maximum likelihood

- One coin.
- Toss it ten times.
- Get heads 6 times.
- What is your estimate of probability of heads $\hat{p}$?

Maximum likelihood

- One coin.
- Toss it ten times.
- Get heads 6 times.
- What is your estimate of probability of heads $\hat{p}$?
- $\hat{p} = 0.6$ is the maximum likelihood estimate: No other p will make the observed data more likely.
- $p^{ML} = \underset{p}{\text{argmax}}\ P(Data|p)$

ML for phylogenetic inference

- Find the tree (topology and branch lengths) that make the observed data most likely.
- If *Data* had a single column in the alignment:

$$Tree^{ML} = \underset{Tree}{\text{argmax}}\ P(Data|Tree)$$

- If we have more than one column, each position in the alignment usually taken as independent:

$$P(D_1, D_2, \ldots, D_n|Tree) = \prod_{i=1}^{i=n} P(D_i|Tree)$$

  and find the tree that makes the above the largest.
- (Often you'll see logs: so as to turn products into sums.)

ML: ingredients

- A way to find $\prod P(D_i|Tree)$
- A way to move around (explore) the space of trees. We've seen this already. ▸ "Exploring space of trees"

ML: ingredients

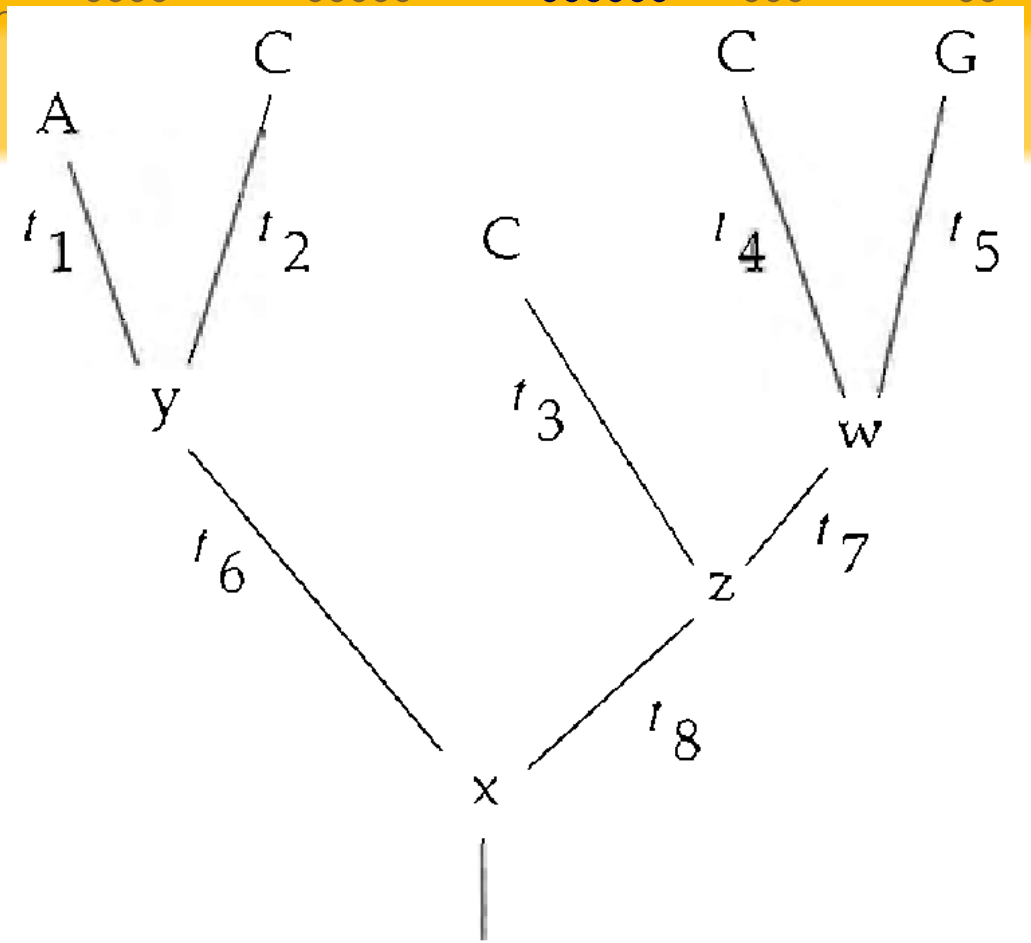
- A way to find $\prod P(D_i|Tree)$
- A way to move around (explore) the space of trees. We've seen this already. ▸ "Exploring space of trees"
- This *modus operandi* seen before with parsimony.
  - With parsimony we want to minimize number of changes
  - With ML we want to maximize the likelihood
- And then we search for the max (or the min, in parsimony).

Iterate over those steps (draw it in the blackboard).

How do we find the probability?

- The evolutionary model!
- Recall J-C: we can obtain the probability of, say, getting a C from a T in 10 units of time
- Just need to be careful and go over the (unknown) internal nodes.
- Place the root somewhere, and cover whole tree.

ML: a figure



$$P(A, C, C, C, G, x, y, z, w|T) = P(x)P(y|x, t_6)P(A|y, t_1)P(C|y, t_2)$$
$$P(z|x, t_8)P(C|z, t_3)P(w|z, t_7)P(C|w, t_4)P(G|w, t_5)$$

(Then sum over all possible $P(x), P(y), P(z), P(w)$)

From Felsenstein, 2004.

How to assess the tree

- Bootstrap: ▸ “How reliable is the tree”

The Bayesian idea

- ML gives us the parameters that make the data most likely.
- Bayesian methods give as the parameters that are most likely, given the data.

Bayes rule with trees

- Bayes rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- $P(Tree|Data) = \frac{P(Data|Tree)P(Tree)}{P(Data)}$
- On the left: **the posterior**
- Likelihood: $P(Data|Tree)$
- Bayesians also need $P(Tree)$: the *prior*.
- $P(Tree|Data) \propto P(Data|Tree)P(Tree)$
- ($P(Data)$: we will not care much about it; just a normalization constant. Often we can ignore it)

The prior

- Flat priors, non-informative priors, issues of scale, how to come up for priors for trees, etc.
- If you have enough data, the prior is completely swamped by the likelihood. Little effect.
- Still, the prior can be a (very) contentious issue.

Bayesian: no need for bootstrap

- We get probability estimates directly
- Easier to interpret than bootstrap (if we trust the prior and models)

Bayesian: Miscell

- Can be faster than ML
- Might be (in practice) more flexible than Maximum Likelihood
- Appropriate usage of Bayesian approaches might require more skill than with other methods.

Which method to use?

- One ordering: Bayesian slightly better than ML slightly better than Parsimony slightly better than NJ.
- Caveats about parsimony (might not be statistically consistent, Felsenstein zone, hides the model, etc).
- Caveats about Bayesian (priors).
- Time constraints.
- Available software.
- Difficulty of using it well.
  - A great method might be great if used by a skilled user but terrible if used by inexperienced users.
  - An average method might perform better if used by a not-so-skilled user.
- Other possible uses (ancestral reconstructions)

A bit of history and philosophy

- Ch. 10 in Felsenstein
- David Hull's "Science as a process"
- Elliot Sober's "Reconstructing the past"

What we haven't covered

A lot!

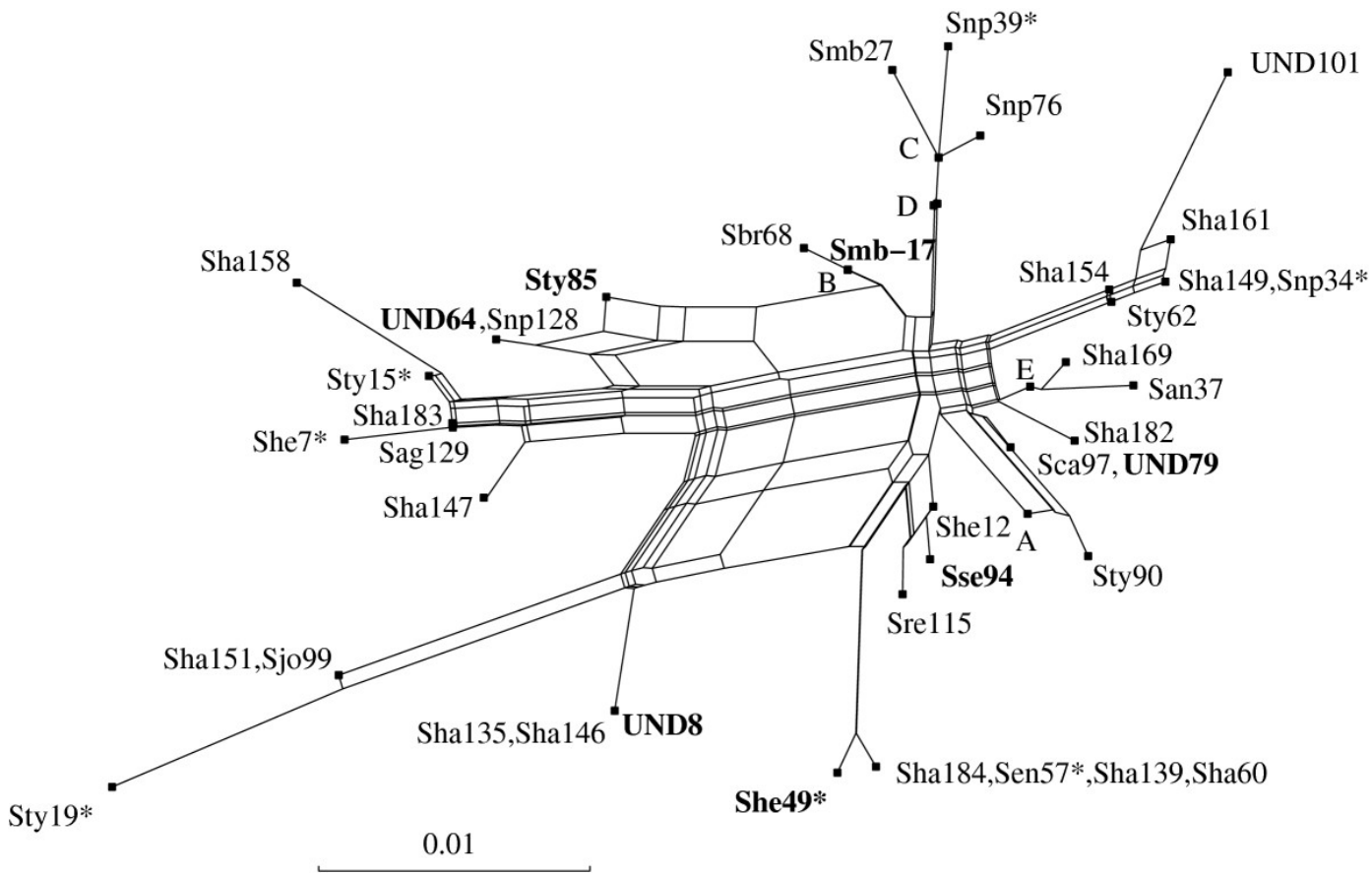
- Phylogenetic networks
- Reconstructing ancestral states ("molecular paleontology")
- Combining information
- Detecting adaptive evolution (dN/dS ratios)

Detecting adaptive evolution

- dN/dS ratios.
- > 1: positive selection
- < 1: purifying selection
- $= 1$: neutral.
- How exactly to do this? See Nei and Kumar, 2000.

Phylogenetic networks

Gene transfer, recombination, hybridization:



From Bryant et al., 2007, *Algorithms in Molecular Biology*. Image from

http://www.almob.org/content/2/1/8/figure/F1?highres=y

OPEN ACCESS Freely available online

PLOS BIOLOGY

Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication

Karin Voordeckers[1,2☯], Chris A. Brown[1,2,3,4,5☯], Kevin Vanneste[6,7], Elisa van der Zande[1,2], Arnout Voet[8], Steven Maere[6,7*], Kevin J. Verstrepen[1,2*]

1 VIB Laboratory for Systems Biology, Leuven, Belgium, **2** CMPG Laboratory for Genetics and Genomics, KU Leuven, Leuven, Belgium, **3** Fathom Information Design, Boston, Massachusetts, United States of America, **4** Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge, Massachusetts, United States of America, **5** Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, United States of America, **6** VIB Department of Plant Systems Biology, Gent, Belgium, **7** Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent, Belgium, **8** Laboratory for Molecular en Structural Biology, KU Leuven, Leuven, Belgium

Abstract

Gene duplications are believed to facilitate evolutionary innovation. However, the mechanisms shaping the fate of duplicated genes remain heavily debated because the molecular processes and evolutionary forces involved are difficult to reconstruct. Here, we study a large family of fungal glucosidase genes that underwent several duplication events. We reconstruct all key ancestral enzymes and show that the very first preduplication enzyme was primarily active on maltose-like substrates, with trace activity for isomaltose-like sugars. Structural analysis and activity measurements on resurrected and present-day enzymes suggest that both activities cannot be fully optimized in a single enzyme. However, gene duplications repeatedly spawned daughter genes in which mutations optimized either isomaltase or maltase activity

- Try with MEGA
- Definitely read ch. 13 in Hall, 2011.

Combining information from multiple genes

- How should we incorporate multiple different sequences which might require possibly different model parameters?

What next?

- Are you ready to prepare publication-quality phylogenetic trees?
- Almost

What next? (II)

You should look at

- Hall, 2011, “Phylogenetic trees made easy”
- Yang, 2014, “Molecular evolution: a statistical approach”.
- Lemey et al., 2009, “The phylogenetic handbook” (some chapters, as needed).

What next? (II)

You should look at

- Hall, 2011, “Phylogenetic trees made easy”
- Yang, 2014, “Molecular evolution: a statistical approach”.
- Lemey et al., 2009, “The phylogenetic handbook” (some chapters, as needed).
- Probably take a look at:
- Nei and Kumar, 2000, “Molecular evolution and phylogenetics”.
- Graur and Li, 2000, “Fundamentals of molecular evolution” (ch. 5)
- Felsenstein, 2004, “Inferring phylogenies”.
- Huson et al. 2011 “Phylogenetic networks”. (if you deal with this)

Software

- Exhaustive (huge!) list at: http://evolution.genetics.washington.edu/phylip/software.html.
- MEGA.
- PHYLIP: probably most widely distributed phylogeny package. Command line and a Java interface. Parsimony, distance, ML. Free software.
- MrBayes for Bayesian. Free software.
- R. Free sofware.
- For serious parsimony: probably want PAUP* or Phylip. (MEGA seems a little limited). PAUP* is NOT free.

Software (II)

- Web servers
  - LIRMM: http://www.phylogeny.fr/. “Robust for the non specialist”.
  - Pasteur Institute: http://mobyle.pasteur.fr/cgi-bin/portal.py
  - University of Oslo: http://www.bioportal.uio.no/ (requires getting a free account).

10 Appendix

- Further details about algorithms
- More about alignments
- Bayesian approaches: MCMC

UPGMA: the algorithm

1. Put each taxon (or sequence) in its own cluster. (So we start from the bottom up).
2. Find pair of clusters with smallest distance. Suppose these are i, j.
3. Create a new cluster, find its height, recompute distances:
   a. Put i, j are put into a cluster. Let's call it IJ. i and j are removed from the distance matrix (but not the new cluster IJ).
   b. Height of node $IJ = \frac{1}{2} d_{ij}$. (So this is the same as placing parent node, IJ at midpoint of branch)
   c. Recompute distance matrix: distance of any other taxa, k, to IJ is average of distance between k and i and j (i.e., average of d_{ki}, d_{kj}).
4. Repeat 2. and 3. until done.

UPGMA: ultrametricity

- Assumes **ultrametricity**. Ultrametric distances: for any three taxa, i, j, k, distances d_i, d_j, d_k either all equal, or two equal and the third is smaller. Check the tree to understand this!
- Ultrametricity OK if molecular clock. Not otherwise.
- UPGMA forces the tree to be ultrametric (even if original distances are not).

UPGMA: oooops!

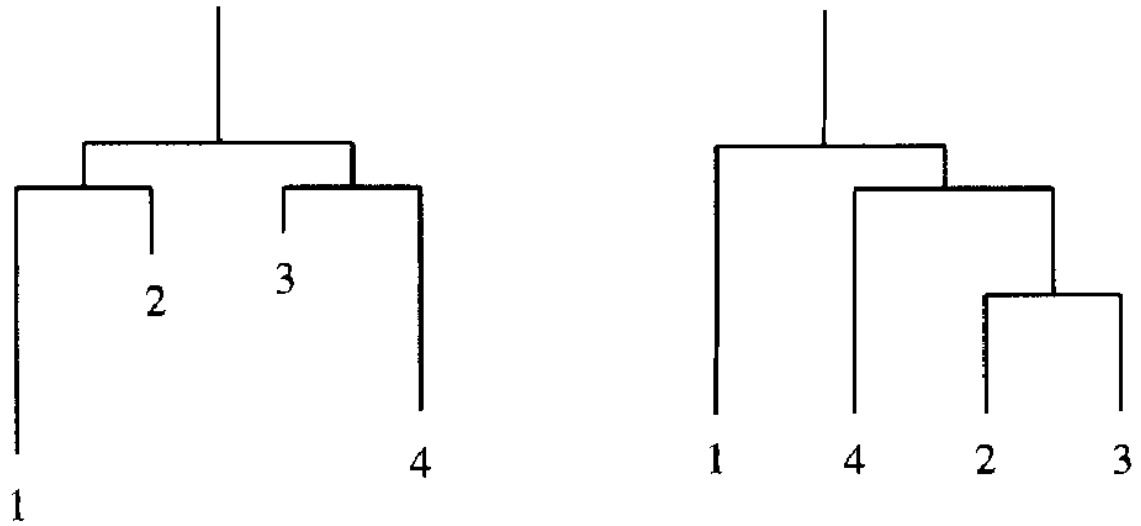


Figure 7.5 *A tree (left) that is reconstructed incorrectly by UPGMA (right).*

From Durbin et al, 1998.

Neighbor Joining, key features of algorithm: formulas

1. Compute the average distance of each taxon, i, to each other taxa: r_i.
2. Correct pairwise distances: $D_{ij} = d_{ij} - (r_i + r_j)$.
3. Find min in D_{ij}. Call k the new taxon.
4. Compute distance between the new node, k and its daughter taxa: d_{ik}, d_{jk}. $d_{ik} = \frac{1}{2}(d_{ij} + r_i - r_j)$. d_{ik} need not be equal to d_{jk}.
5. Compute distance between k and remaining taxa. For all m in the remaining taxa: $d_{km} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$.

Neighbor Joining (NJ): assumptions

- Returns a single unrooted tree.
- Assumes **aditivity**.
- A tree with additive distances: distances between any two nodes sum of lengths of all branches between them.
- NJ will take a distance matrix and return an (unrooted) tree with additive distances.
- (We can check if a distance matrix is additive: **the four point condition**.)
- Can we use NJ if distances deviate from additivity? Yes, but correct tree no longer guaranteed.
- No method can guarantee *the* correct tree in real life.

Other distance-based

- Variants of NJ: e.g., BIONJ.
- Try to find the best fitting tree.
- What is best? E.g.:
  - Minimum evolution over all tree (total branch lengths of reconstructed tree).
  - Least-squares methods (minimize deviations of distances in tree from distances in original distance matrix). Several types.
- These methods give a criterion for choosing among trees.
- These methods do not give an algorithm for building the tree!

Parsimony: one algorithm for scoring

We have a tree and a set of sequences. What is the score of the tree?

Unweighted parsimony, main steps:

- Each character is treated independently.
- Go up (from leaves to root)
- If daughters share the state, set a pseudo-ancestral state (minimal cost residues) to the shared state (and do not penalize).
- If daughters do not share state, set pseudo-ancestral as the union, and increment homoplasy count.
- Can go down if need the reconstruct ancestral states, but can miss solutions. More sophisticated ways.
- Most “for real” implementations use other approaches (e.g., Sankoff’s).

And the root? It does not matter where it is placed.

Exploring space of trees. Exhaustive search

Exhaustive only feasible for few taxa.

- Start with three taxa, and keep adding.
- Can use **branch-and-bound** (ramificación y poda?).
  - Suppose we have a tree with 10 taxa and cost 4.
  - We are now in tree with 5 taxa and cost 5. No need to continue adding taxa to this tree (we get rid of a whole family of trees).

Heuristic search methods

- Get a tree. Modify it. Is it any better? Can it be improved by minor modifications?
- “Shake the system” to explore the parameter space.
- Popular moves:
  - Exchange neighbors (**nearest neighbor interchange**)
  - Move subtrees (**subtree prunning and regrafting**)
  - Cut the tree and reconnect in one random branch (**tree bisection and reconnection**)
  - There are others.
  - (A figure in “Appendix”)

Tree movements: a figure

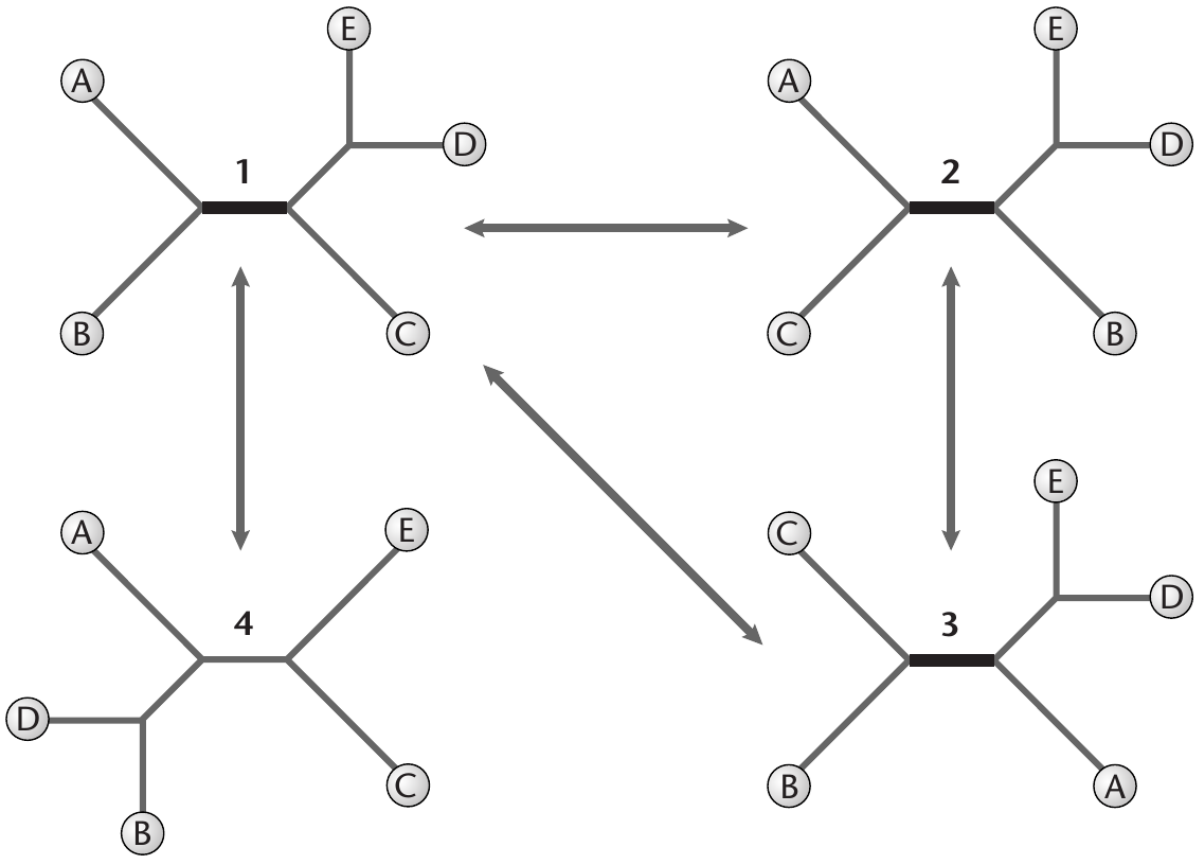


Fig. 8.10 Examples of changes in tree topology. Trees 1, 2, and 3 all differ from each other by a single nearest neighbor interchange. Tree 4 differs from tree 1 by a subtree pruning and regrafting operation.

From Higgs and Attwood, 2005. In 4, we the subtree was “D”.

Alignments

- We take them as given
- In real life
  - Examine them carefully
  - Possibly not include certain parts of the alignment

Consequences of alignment problems

- Phylogenetic tree building can be robust to minor problems in alignment.
- At least two tasks can be very sensitive:
  - Reconstructing ancestral states
  - Detecting adaptive evolution

Alignments: What can we do?

- Know your alignment software well and use good ones.
- Look at alignments and possibly edit them.
- Some tools available:
  - GUIDANCE (see Hall, 2011, ch. 12)
  - ALTAVIST (see ch. 3 in Lemey et al., 2009)
- **Definitely** read ch. 4 and 12 of Hall, 2011.

When alignments are not used

- In some cases we do not use, as such, multiple alignment.

When alignments are not used

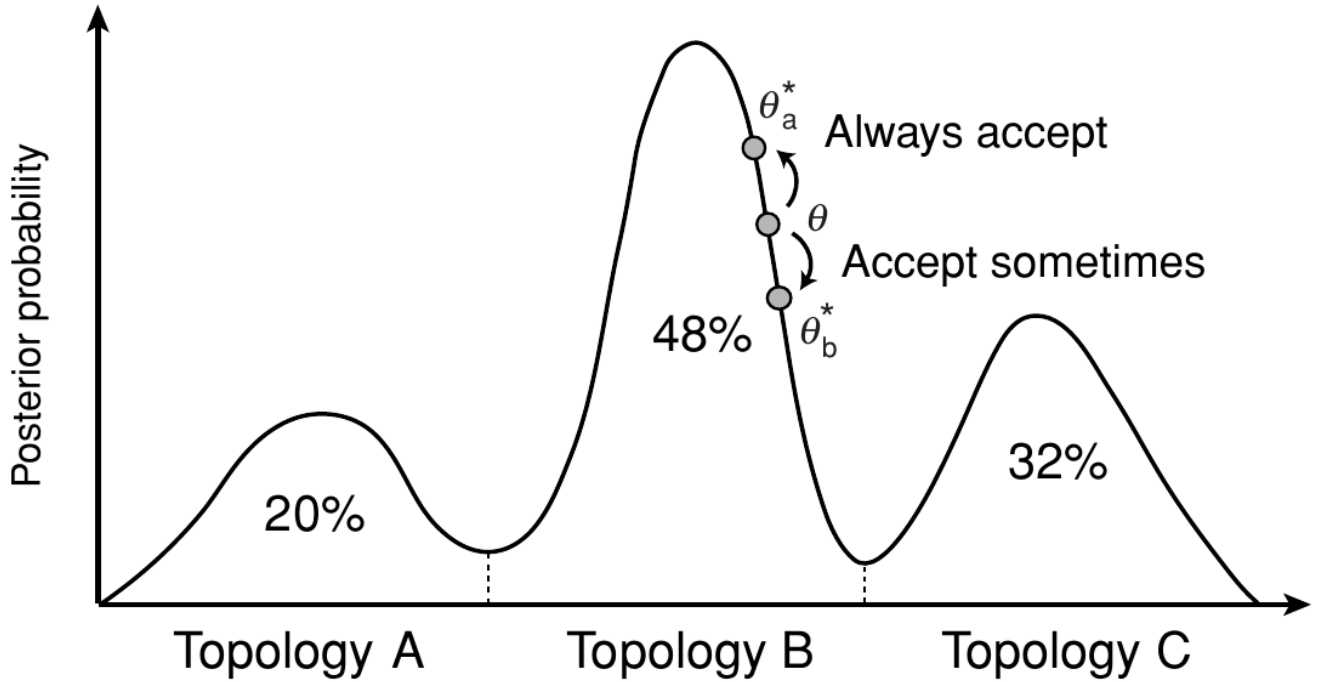
- In some cases we do not use, as such, multiple alignment.
- Morphological characters
- Phylogenies from CNVs
- ...
- Principles the same:
  - Get a distance matrix from original data and build phylogeny
  - Use a model and build phylogeny from original data

Computation: MCMC

Markov Chain Monte Carlo

- We want to get the posterior: $P(Tree|Data)$
- We cannot get it analytically.
- But we might be able to numerically calculate P.
  - Set up a Markov Chain to jump between parameter states (tree states), so that the posterior is the stationary distribution.
  - Sample from the posterior.
  - Discard first samples, as not reached stationarity (burn-in).

Computation: MCMC



From Ronquist et al., in Lemey et al, 2006.