Estimating Phylogenies (Evolutionary Trees) I

Biol4230 Tues, Feb 27, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Goals of today's lecture:

- Why estimate phylogenies?
  - Origin of man (woman)
  - Origin of diseases (HIV)
  - Origin of life
  - Origin of functions (orthology)
- What are phylogenies?
  - Types of trees
  - How many trees
- How to estimate?
  - What is the goal? How do we judge?
  - Parsimony, Distance
  - Statistical (Model based) approaches

To learn more:

- Pevsner Bioinformatics Chapter 6 pp 179–212
- Graur and Li (2010) "Fundamentals of Molecular Evolution" Sinauer Associates
- Nei (1987) "Molecular Evolutionary Genetics" Columbia Univ. Press
- Hillis, Moritz, and Mable (1996) "Molecular Systematics" Sinauer
- Felsenstein (2003) "Inferring Phylogenies" Sinauer

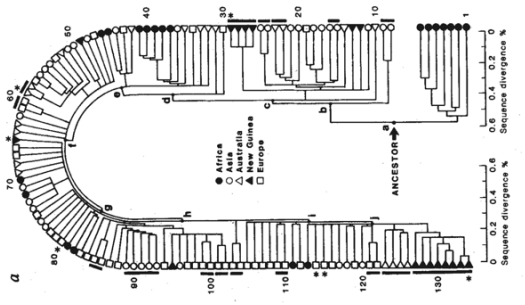
Why Build Phylogenies I
The Origin of Man (Woman)

Mitochondrial DNA and human evolution

Rebecca L. Cann*, Mark Stoneking & Allan C. Wilson

Department of Biochemistry, University of California, Berkeley, California 94720, USA

Mitochondrial DNAs from 147 people, drawn from five geographic populations have been analysed by restriction mapping. All these mitochondrial DNAs stem from one woman who is postulated to have lived about 200,000 years ago, probably in Africa. All the populations examined except the African population have multiple origins, implying that each area was colonised repeatedly.

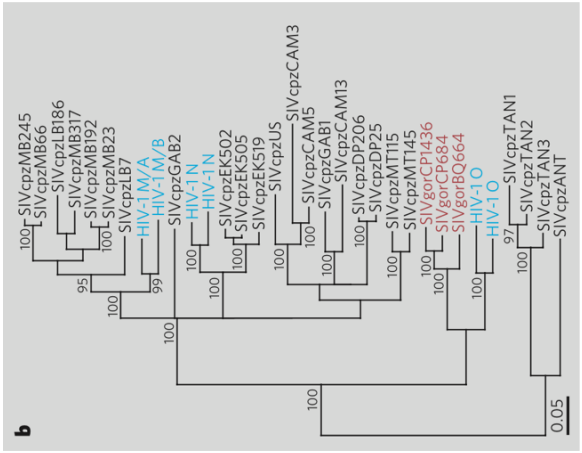


Nature **325,** 31–36 (1987).

3

Why Build Phylogenies IIa
The Origin of Disease (HIV)

Van Heuverswyn, F. *et al.* Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature* **444,** 164–164 (2006).



fasta.bioch.virginia.edu/biol4230

4

Why Build Phylogenies IIb
Tracing the Origin of specific HIV infections

Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences

Diane I. Scaduto[a,b], Jeremy M. Brown[c,1], Wade C. Haaland[a,b], Derrick J. Zwickl[c,2], David M. Hillis[c,3], and Michael L. Metzker[a,b,d]

Phylogenetic analysis has been widely used to test the a priori hypothesis of epidemiological clustering in suspected transmission chains of HIV-1. Among studies showing strong support for relatedness between HIV samples obtained from infected individuals, evidence for the direction of transmission between epidemiologically related pairs has been lacking. During transmission of HIV, a genetic bottleneck occurs, resulting in the paraphyly of source viruses with respect to those of the recipient. This paraphyly establishes the direction of transmission, from which the source can then be inferred. Here, we present methods and results from two criminal cases, … which provided evidence that direction can be established from blinded case samples. The observed paraphyly from each case study led to the identification of an inferred source (i.e., index case), whose identity was revealed at trial to be that of the defendant.

Proc Natl Acad Sci USA
107, 21242–21247 (2010).

Why Build Phylogenies IIb
Tracing the Origin of specific HIV infections

Fig 1. Pol protein (no paraphyly)

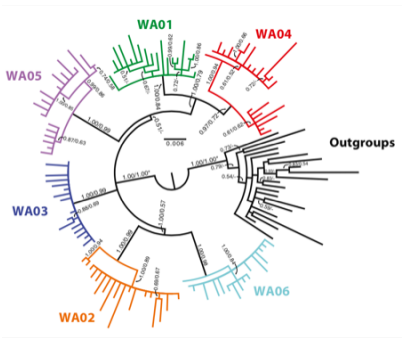
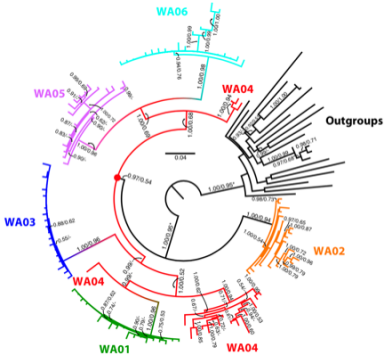


Fig 2. Env protein (paraphyly)



… the direction of transmission (source → recipient) would further strengthen the a priori hypothesis under investigation. This is possible if a paraphyletic relationship (i.e., a subset of source viral sequences is more closely related to all recipient sequences than to other source sequences) is observed in the phylogenetic tree.

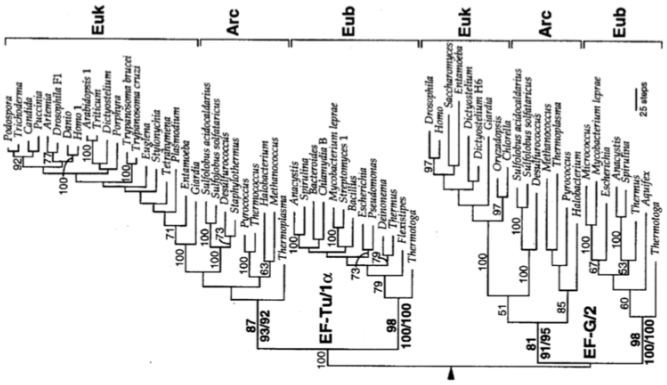
Proc Natl Acad Sci USA
107, 21242–21247 (2010).

Why Build Phylogenies III
Placing the Origin of Life

Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 7749–7754, July 1996
Evolution

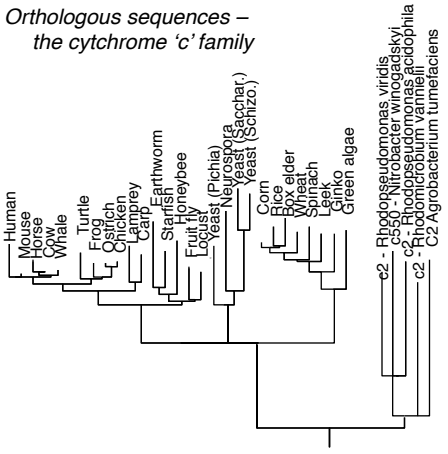
The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny

SANDRA L. BALDAUF*†, JEFFREY D. PALMER‡, AND W. FORD DOOLITTLE*

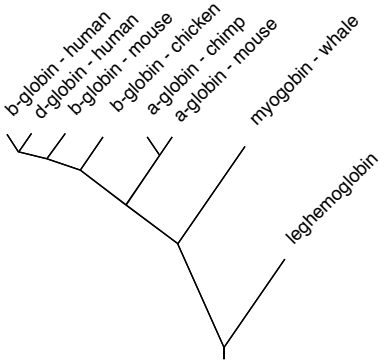


Why Build Phylogenies IV
Inferring Function: Orthology and Paralogy

Orthologous sequences – the cytchrome 'c' family



Paralogous genes – globins



What are phylogenies?

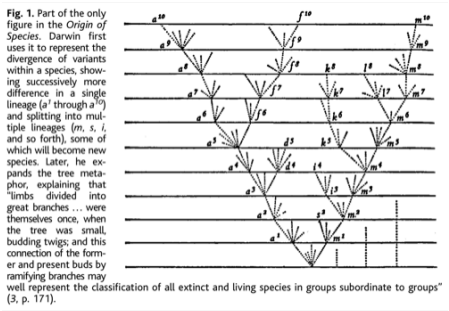


Fig. 1. Part of the only figure in the *Origin of Species*. Darwin first uses it to represent the divergence of variants within a species, showing successively more difference in a single lineage (a^1 through a^{10}) and splitting into multiple lineages (*m*, *s*, *i*, and so forth), some of which will become new species. Later, he expands the tree metaphor, explaining that "limbs divided into great branches ... were themselves once, when the tree was small, budding twigs; and this connection of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups" (3, p. 171).

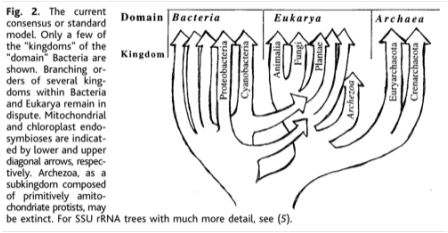
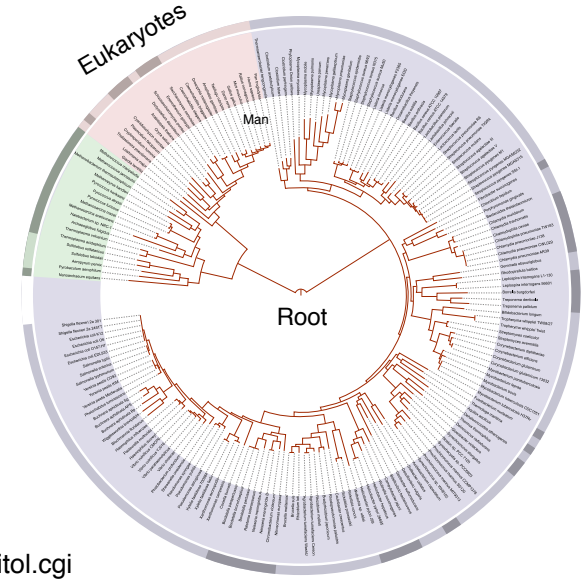


Fig. 2. The current consensus or standard model. Only a few of the "kingdoms" of the "domain" Bacteria are shown. Branching orders of several kingdoms within Bacteria and Eukarya remain in dispute. Mitochondrial and chloroplast endosymbioses are indicated by lower and upper diagonal arrows, respectively. Archezoa, as a subkingdom composed of primitively amitochondriate protists, may be extinct. For SSU rRNA trees with much more detail, see (5).

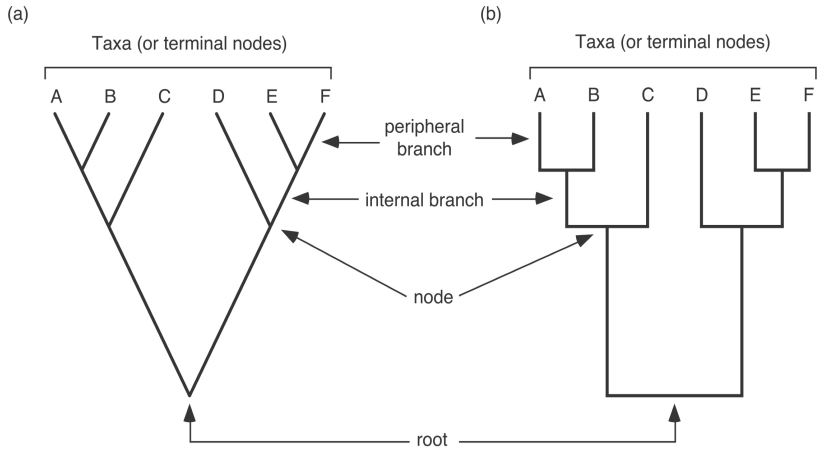
Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284,** 2124–2129 (1999).

What are phylogenies?



itol.embl.de/itol.cgi

What are phylogenies? Terminologies for trees

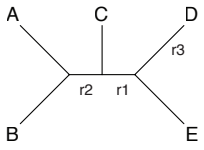


Chap 2, Figure 1

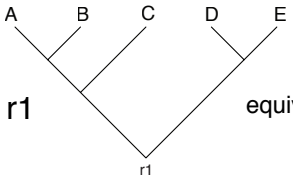
From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Trees – Rooted and UnRooted

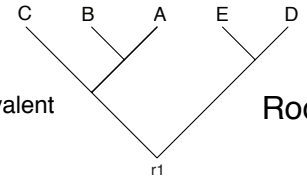
Unrooted tree



Rooted r1

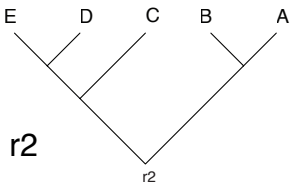


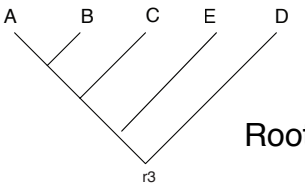
equivalent



Rooted r1

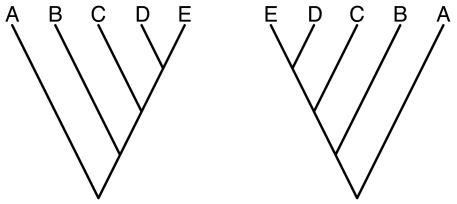
Rooted r2



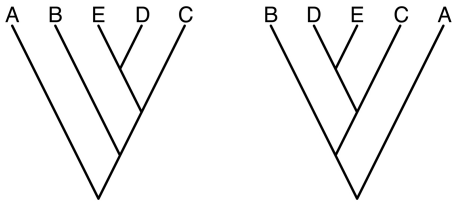


Rooted r3

Representing trees

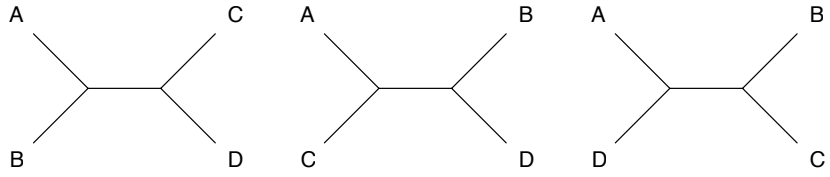


(A, (B, (C, (D,E)))) Newick format

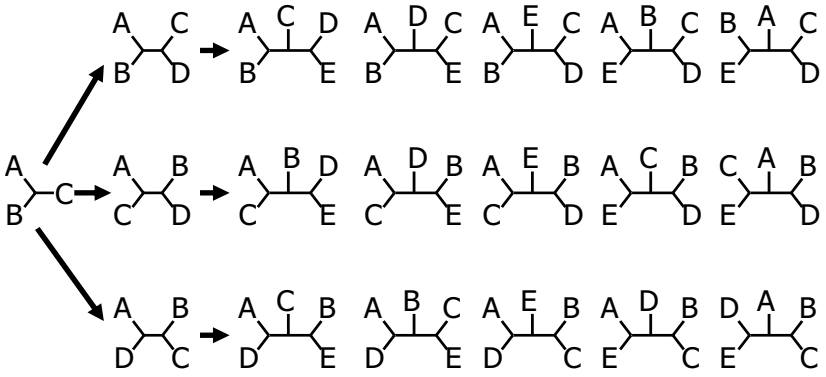


Chap 2, Figure 3 From Hillis lecture: www.doublehelixranch.com/WoodsHoleMole2014.pdf

Unrooted 4-Taxa Trees



Exhaustive Evaluation



How many trees are there?

An unrooted, bifurcating tree with T terminal nodes has T – 2 internal nodes. It also has T – 3 internal branches and T peripheral branches, for a total of 2T – 3 branches. Adding a root to the tree also adds a branch, since the root divides one branch into two.

The number of labeled, unrooted binary trees is:

$$N_u = \prod_{i=3}^{T} (2i - 5)$$

which expands to: $(2 \bullet 3 - 5)(2 \bullet 4 - 5)(2 \bullet 5 - 5)...(2 \bullet T - 5)$

The number of labeled, rooted binary trees is:

$$N_u = \prod_{i=3}^{T} (2i - 3)$$

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Number of trees

Taxa	Unrooted	Rooted
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	3E7
15	7E12	2E14
20	2E20	8E21
50	3E74	
100	2E182	
1000	2E2860	

Finding the best / Estimating trees

- Most strategies to reconstruct evolutionary trees optimize some measure of "goodness"
  - Parsimony methods minimize the number of mutations
  - Distance methods produce trees that match the global distances between the sequences
  - Maximum likelihood methods seek the tree that best fits the data
- What is the "best" method?
  - produces accurate trees with the least data?
  - converges to the correct tree as data increases?
- We cannot know the "correct" tree

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Finding the best / Estimating trees

- An optimality criterion defines how we measure the fit of the data to a given solution
  - parsimony / distance / Maximum likelihood
- Tree searching is a separate step; this is how we search through possible solutions (which we then evaluate with the chosen optimality criterion)
  - Except for Neighbor-Joining and UPGMA, which produce a result based on the search strategy

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Optimality criteria:

- Nonparametric methods:
  - parsimony and related approaches
- Semi-parametric methods:
  - pairwise distance approaches
- Parametric methods:
  - Likelihood and Bayesian approaches

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Advantages

- Parsimony:
  - Widely applicable to many discrete data types (often used to combine analyses of different data types)
  - Requires no explicit model of evolutionary change
  - Computationally relatively fast
  - Relatively easy interpretation of character change
  - Performs well with many data sets
- Distance:
  - Can be used with pairwise distance data (e.g., non-discrete characters)
  - Can incorporate an explicit model of evolution in estimation of pairwise distances
  - Computationally relatively fast (especially for single-point estimates)
- Likelihood/Bayesian:
  - Fully based on explicit model of evolution
  - Most efficient method under widest set of conditions
  - Consistent (converges on correct answer with increasing data, as long as assumptions are met)
  - Most straight-forward statistical assessment of results; probabilistic assessment of ancestral character states

From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Disadvantages:

- Parsimony methods:
  - No explicit model of evolution; often less efficient
  - Nonparametric statistical approaches for assessing results often have poorly understood properties
  - Can provide misleading results under some fairly common conditions
  - Do not provide probabilistic assessment of alternative solutions
- Distance methods:
  - Model of evolution applied locally (to pairs of taxa), rather than globally
  - Statistical interpretation not straight-forward
  - Can provide misleading results under some fairly common conditions (but not as sensitive as parsimony)
  - Do not provide probablistic assessment of alternative solutions
- Likelihood/Bayesian:
  - Requires an explicit model of evolution, which may not be realistic or available for some data types
  - Computationally most intense

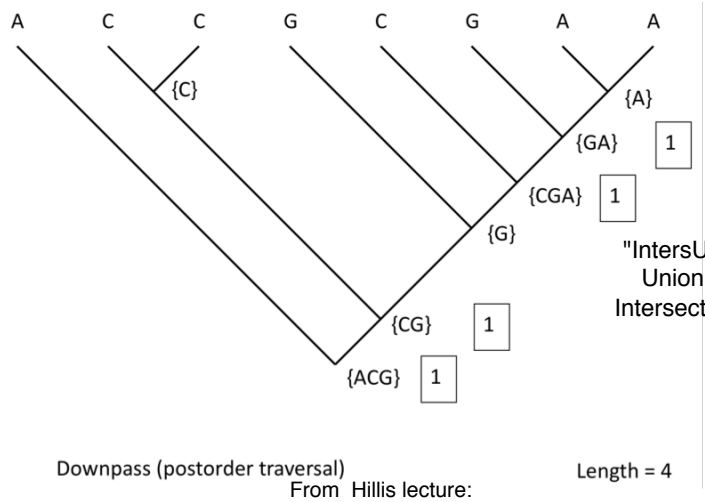
From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

The Parsimony Criterion:

- Under the parsimony criterion, the optimal tree (the shortest or minimum length tree) is the one that minimizes the sum of the lengths of all characters in terms of evolutionary steps (a step is a change from one character-state to another).
- For a given tree, find the length of each character, and sum these lengths; this is the tree length.
- The tree with the minimum length is the most parsimonious tree.
- The most parsimonious tree provides the **best fit** of the data set under the parsimony criterion.

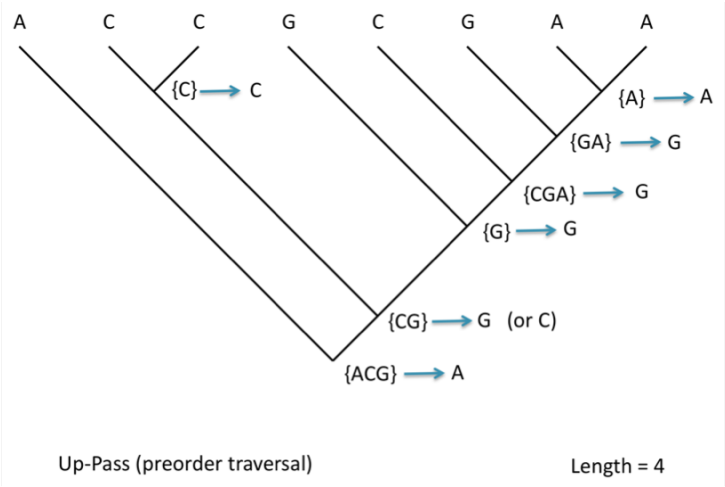
From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Parismony: counting changes



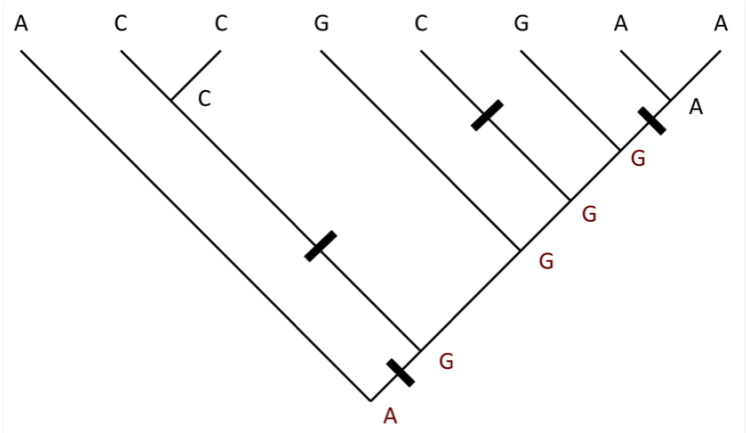
From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Parismony: ancestral states



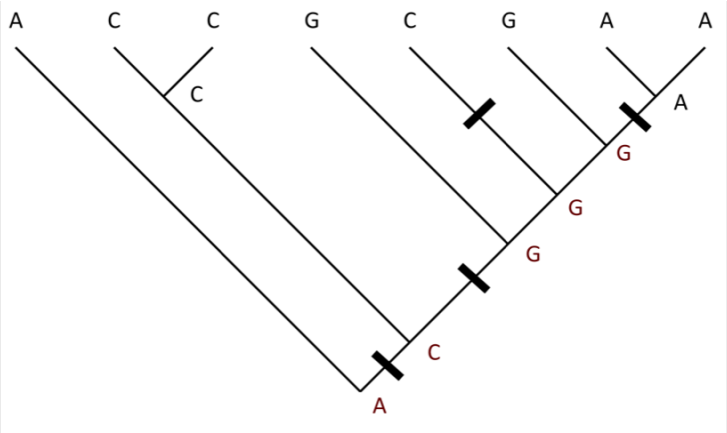
From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Parismony: ancestral states



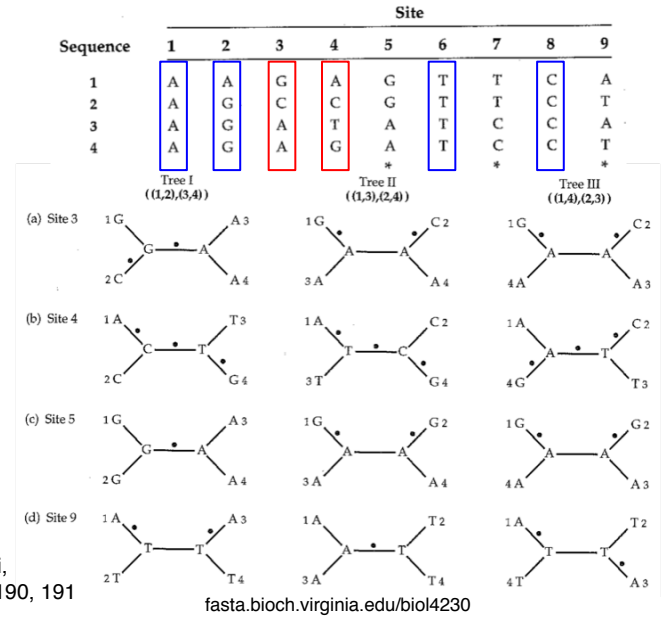
From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Parismony: ancestral states



From Hillis lecture:
www.doublehelixranch.com/WoodsHoleMole2014.pdf

Parsimony – Informative sites



Graur and Li,
Chap 5, pp 190, 191

Parsimony – Informative sites

Paup analysis of 3000 sites from primate mitochondrial D-loop

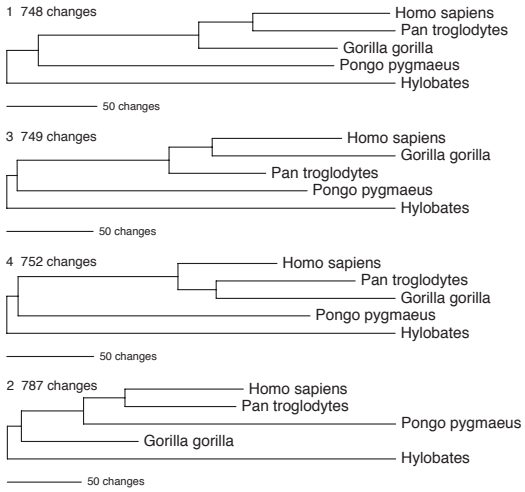
```
 Character-status summary:
    13203 characters are excluded  (selected 1-3000)
    Of the remaining 3000 included characters:
      All characters are of type 'unord'
      All characters have equal weight
      2397 characters are constant
      431 variable characters are parsimony-uninformative
      Number of (included) parsimony-informative characters = 172
  Gaps are treated as "missing"
  Multistate taxa interpreted as uncertainty

Tree #    1   2   3   4   5   6   7   8   9  10
Length  748 787 749 752 792 787 792 789 789 789
```

172/3000 = 5.7% of data used to build tree

Parsimony – Informative sites

Paup analysis of 3000 sites from primate mitochondrial D-loop



Parsimony – Informative sites

```
 Character-status summary:
    13203 characters are excluded  (selected 1-3000)
    Of the remaining 3000 included characters:
      All characters are of type 'unord'
      All characters have equal weight
      2397 characters are constant
      431 variable characters are parsimony-uninformative
      Number of (included) parsimony-informative characters = 172
  Gaps are treated as "missing"
  Multistate taxa interpreted as uncertainty

Tree #    1   2   3   4   5   6   7   8   9  10
Length  748 787 749 752 792 787 792 789 789 789
```

172/3000 = 5.7% of data used to build tree

94.3% of data "not informative"
95% identical??
25% identical??

Estimating Phylogenies

- Why estimate phylogenies?
  - Origin of man (woman)
  - Origin of diseases (HIV)
  - Origin of life
  - Origin of functions (orthology)
- What are phylogenies?
  - Types of trees
  - How many trees
- How to estimate?
  - What is the goal? How do we judge?
  - Parsimony, Distance
  - Statistical (Model based) approaches