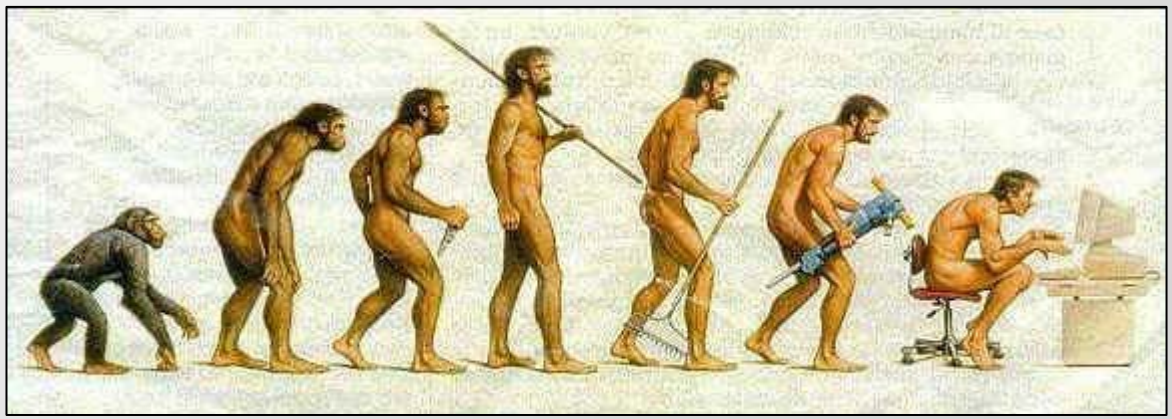
6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

6.047/6.878 - Computational Biology: Genomes, Networks, Evolution

Lecture 11

Oct 7, 2008

Molecular Evolution and Phylogenetics



Patrick Winston's 6.034

Somewhere, something went wrong…

Challenges in Computational Biology

4 Genome Assembly

Regulatory motif discovery

Gene Finding

DNA

Sequence alignment

Comparative Genomics

7 **Evolutionary Theory**

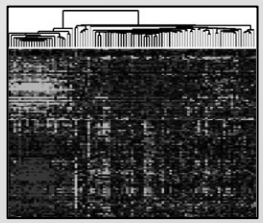
TCATGCTAT
TCGTGATAA
TGAGGATAT
TTATCATAT
TTATGATTT

Database lookup

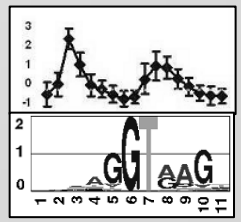
RNA folding

9 Gene expression analysis

RNA transcript

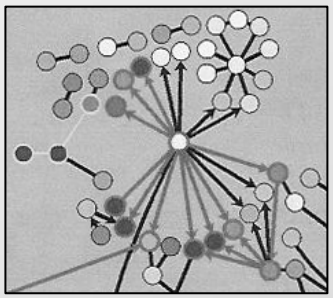


10 Cluster discovery



Gibbs sampling

12 Protein network analysis



13 Regulatory network inference

14 Emerging network properties

Goals for today

- Basics of phylogeny
  - Characters, traits, nodes, branches, lineages
  - Gene trees, species trees
- Modeling sequence evolution
  - Turning sequence data into distances
  - Probabilistic models of nucleotide divergence
  - Jukes-Cantor 1-parameter model, Kimura 2-parameter model
- From distances to trees
  - Ultrametric, Additive, General Distances
  - UPGMA, Neighbor Joining, guarantees and limitations
  - Least-squared error, minimum evolution
- From alignments to trees
  - Parsimony methods: set-based vs. dynamic programming
  - Maximum likelihood methods
  - MCMC and heuristic search

Open questions (?)

- Panda
  - Bear or raccoon?
- Out of Africa
  - mitochondrial evolution story?
- Human evolution
  - Did we ever meet Neanderthal?
- Primate evolution
  - Are we chimp-like or gorilla-like?
- Vertebrate evolution
  - How did complex body plans arise?
- Recent evolution
  - What genes are under selection?

Inferring Phylogenies: Traits and Characters

Trees can be inferred by several criteria:

- Morphology data

Image removed due to copyright restrictions.

- Molecular data

Traits – as many as we have letters in DNA

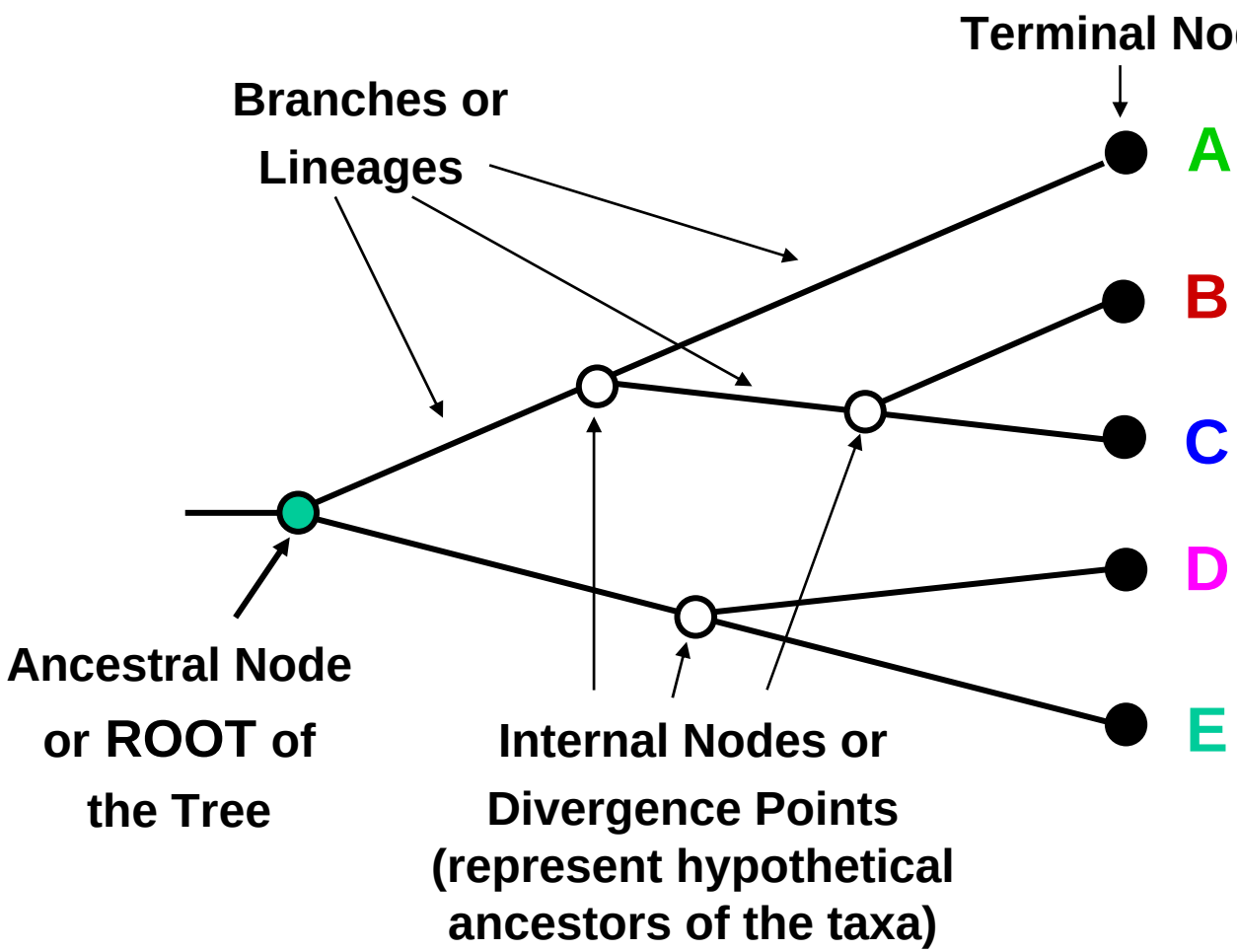
```
YAL042W          -MKRSTLLSLDAFAKTEEDVRVRTRAGGLITLSCILTTLFLLVNEWGQFNSVVTRPQLVV
candida586       MSSRPKLLSFDAFAKTVEDARIKTTSGGIITLICILITLVLIRNEYVDYTTIITRPELVV
cdub17784        MSSRPKLLSFDAFAKTVEDARIKTTSGGIITLICILITLVLIRNEYVDYTTIITRPELVV
cgla72177        -MKKSTLLSFDAFAKTEEDVRIRTRSGGFITLGCLVVTLMLLLSEWRDFNSVVTRPELVI
cgui48535        -MPQPKLLSFDAFAKTVEDARVRTPAGGIITLICVIVVLYLIRNEYLEYTSIINRPELVV
clus15345        MSSRPRLLSLDAFAKTVEDARVKTASGGVITLVCVLIVLFLIRNEYSDYMLVVVRPELVV
ctro67868        MSSRPKLLSFDAFAKTVEDARIKTASGGIITLICVLITLILIRNEYIDYTTIITRPELVV
klac20931        -MKKSPLLSIDAFGKTEEDVRVRTRTGGLITVSCIIITMLLLVSEWKQFSTIVTRPDLVV
                    :. ***:***.** **.*::* :**.**: *:: .: *: .*: ::  :: **:**:

YAL042W          DRDRHAKLELNMDVTFPSMPCDLVNLDIMDDSGEMQLDILDAGFTMSRLNSEG------R
candida586       DRDINKQLDINLDISFINLPCDLISIDLLDVTGDLSLNIIDSGLKKIRLLKNKQGDVIVN
cdub17784        DRDINKQLDINLDISFINLPCDLISIDLLDVTGDLSLNIIDSGLKKIRLLKNKQGDVIVN
cgla72177        DRDRSLRLDLNLDITFPSMPCELLTLDIMDDSGEVQLDIMNAGFEKTRLSKEG------K
cgui48535        DRDINKKLEINLDISFPDIPCDVLTMDILDVSGDLQVDLLLSGFEKFRLLKDG------L
clus15345        NRDVNRQLDINLDITFPDVPCGVMSLDILDMTGDLHLDIVESGFEMFRVLPLG------E
ctro67868        DRDINKQLDINLDISFINLPCDLISVDLLDVTGDQQLDIIDSGLKKVRLLKNKQGDVIIN
klac20931        DRDRHLKLDLNLDVTFPSMPCNVLNLDILDDSGEFQINLLDSGFTKIRISPEG------K
                 :**   :*::*:*::* .:** ::.:*::* :*:  :::: :*:   *:

YAL042W          PVGDATELHVGGNGDGTAPV--NND---PNY-CGPCYGAKDQSQN-ENLAQEEKVCCQDC
candida586       EIEDDEPAFNNDIELSDLAKGLPEGSDENAY-CGSCYGALPQDK--------KQFCCNDC
cdub17784        EIEDDEPAFNNDIELTDLAKGLPEGSDENAY-CGSCYGALPQDK--------KQFCCNDC
cgla72177        VLGTA-DMKIGEAAKKDKEA--QLAKLGANY-CGNCYGARDQGKNNDDTPRDQWVCCQTC
cgui48535        EIRDESPVMSSAGELEERAR----GRAPDGL-CGSCYGALPQDEN-------LDYCCNDC
clus15345        EISDDLPLLSGAKKFEDVCGPLTEDEISRGVPCGPCYGAVDQTD--------NKRCCNTC
ctro67868        EIEDDKPALNSDVSLKELAKGLPEGSDQNAY-CGPCYGALPQDK--------KQFCCNDC
klac20931        ELSKE-KFQVGDKS--SKQS--FNE---EGY-CGPCYGALDQSKN-DELPQDQKVCCQTC
                  :        .                    ** ****  * .           **: *

YAL042W          DAVRSAYLEAGWAFFDGKNIEQCEREGYVSKINEHLN--EGCRIKGSAQINRIQGNLHFA
candida586       NTVRRAYAEKHWSFYDGENIEQCEKEGYVGRLRERINNNEGCRIKGTTKINRVSGTMDFA
cdub17784        NTVRRAYAEKHWSFYDGENIEQCEKEGYVARLRERINNNEGCRIKGTTKINRVSGTMDFA
cgla72177        DDVRQAYFEKNWAFFDGKDIEQCEREGYVQKIADQLQ--EGCRVSGSAQLNRIDGNLHFA
cgui48535        ETVRLAYAQKAWGFFDGENIEQCEREGYVARLNEKINNFEGCRIKGTGKINRISGNLHFA
clus15345        EAVRMAYAVQEWGFFDGSNIEQCEREGYVEKMVSRINNNEGCRIKGSAKINRISGNLHFA
ctro67868        NTVRRAYAEKQWQFFDGENIEQCEKEGYVKRLRERINNNEGCRIKGSTKINRVSGTMDFA
klac20931        DDVRAAYGQKGWAFKDGKGVEQCEREGYVESINARIH--EGCRVQGRAQLNRIQGTIHFG
                 : ** **    * * **..:****:****  :  :::  ****:.*  ::**:.*.:.*.
```

From physiological traits to DNA characters

- Traditional phylogenetics
  - Building species trees
  - Small number of traits
    - Hoofs, nails, teeth, horns
  - Well-behaved traits, each arose once
    - Parsimony principle, Occam's razor

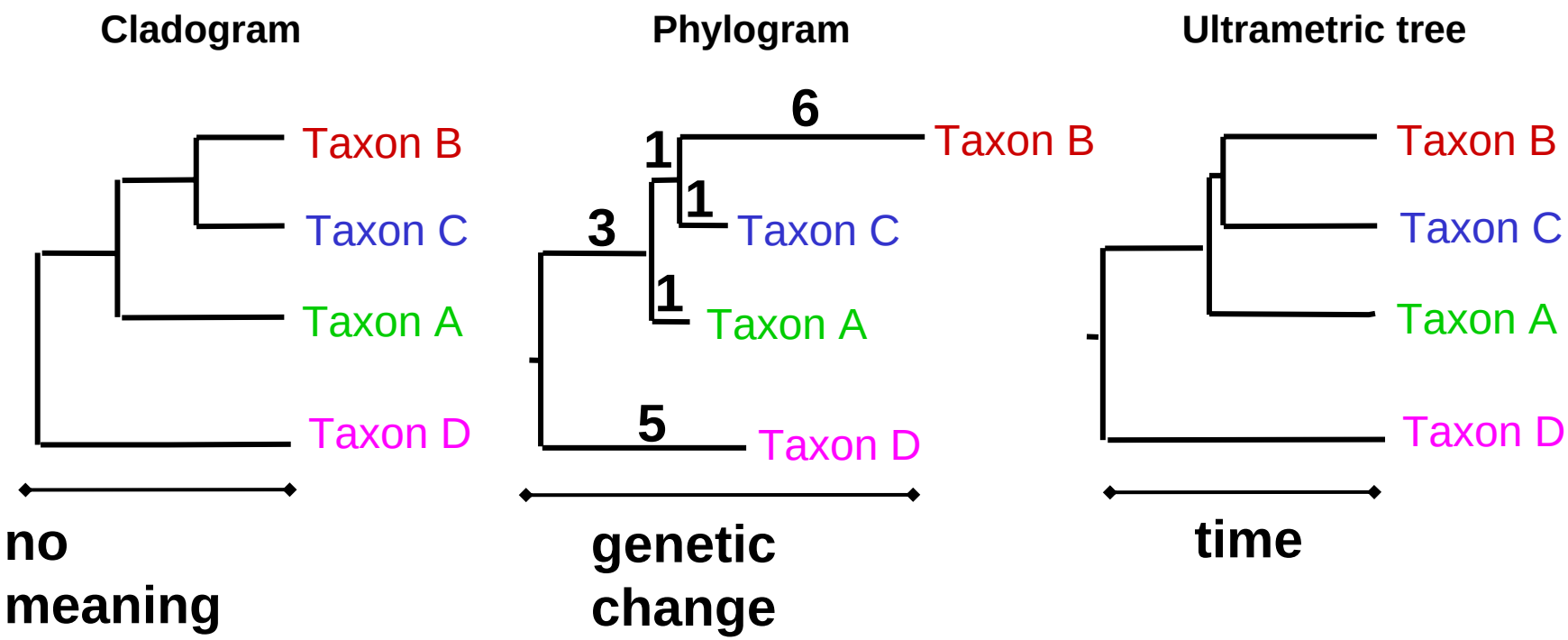
- Modern phylogenetics
  - Building gene trees and species trees
  - Very large number of traits
    - Every DNA base and every protein residue
  - Frequently ill-behaved traits
    - Back-mutations are frequent (convergent evolution)
    - Small number of letters, arise many times independently

Common Phylogenetic Tree Terminology



Represent the TAXA (genes, populations, species, etc.) used to infer the phylogeny

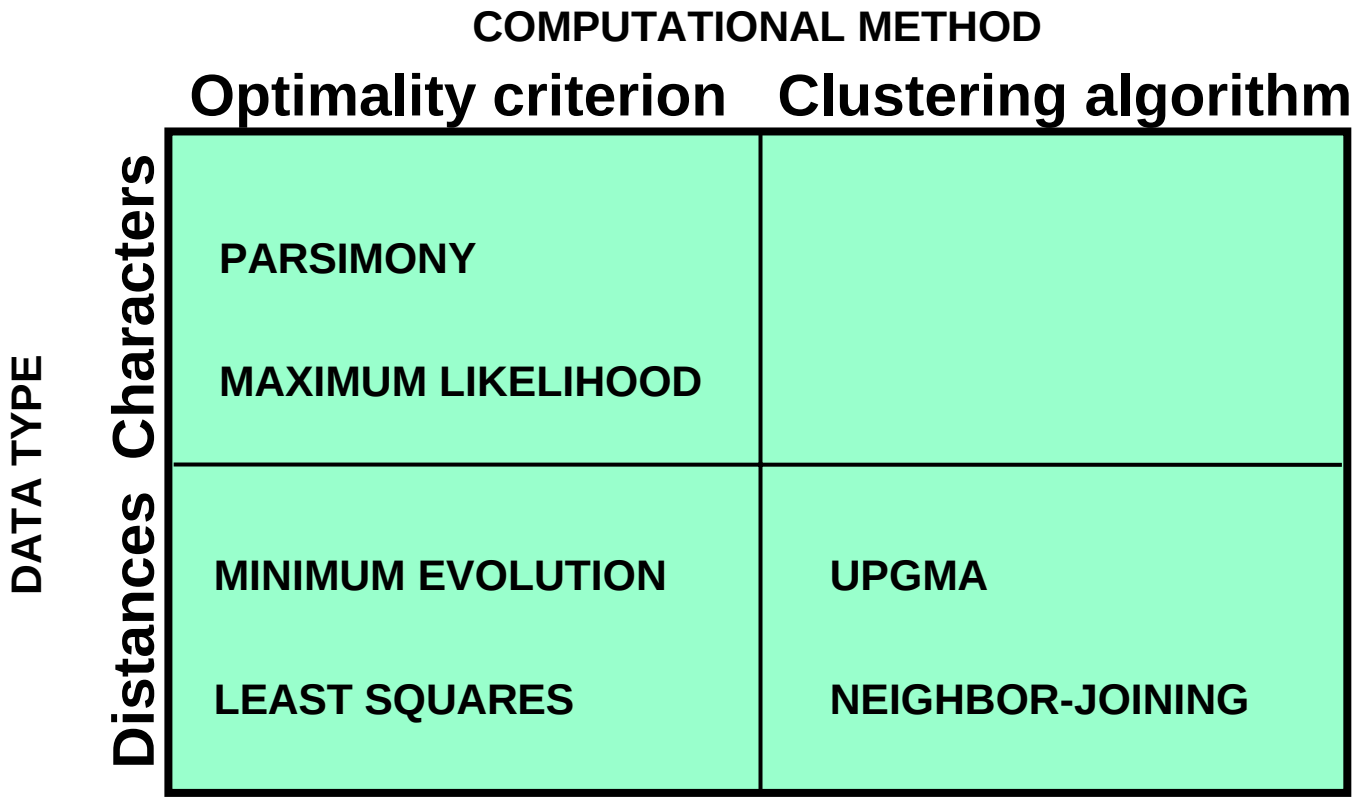
Three types of trees



All show the same evolutionary relationships, or branching orders, between the taxa.

Molecular phylogenetic tree building methods:

Are mathematical and/or statistical methods for inferring the divergence order of taxa, as well as the lengths of the branches that connect them. There are many phylogenetic methods available today, each having strengths and weaknesses. Most can be classified as follows:



2. Modeling evolution

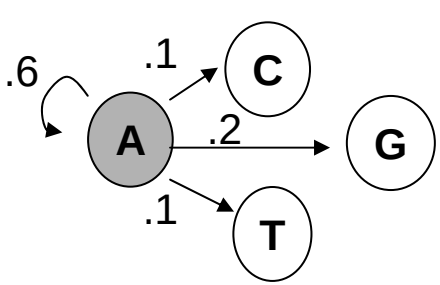
Inferring evolutionary distance

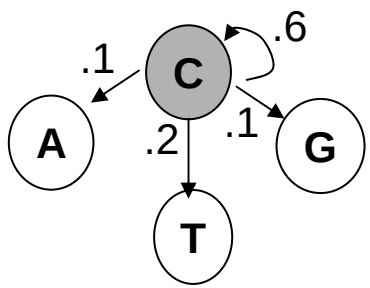
Measuring evolutionary rates

- Nucleotide divergence
  - Uniform rate. Overall percent identity.
- Transitions and transversions
  - Two-parameter model. A-G, C-T more frequent.
- Synonymous and non-synonymous substitutions
  - Ka/Ks rates. Amino-acid changing substitutions

	AGA									UUA					AGC				
	AGG									UUG					AGU				
GCA	CGA						GGA			CUA				CCA	UCA	ACA			GUA	
GCC	CGC						GGC		AUA	CUC				CCC	UCC	ACC			GUC	UAA
GCG	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUC	CUG	AAA		UUC	CCG	UCG	ACG		UAC	GUG	UAG
GCU	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU	GUU	UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

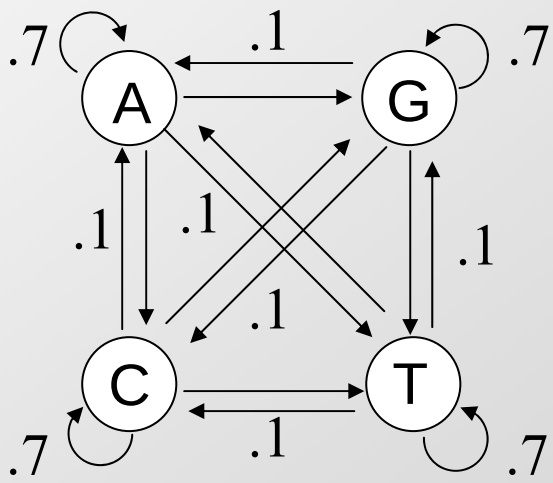
- $N_{substitutions} > N_{mutations}$
  - Some fraction of “conserved” positions mutated twice





‘Evolving’ a nucleotide under random model

- At time step 0, start with letter A
- At time step 1:
  - Remain A with probability 0.7
  - Change to C,G,T with prob. 0.1 each
- At time step 2:
  - In state A with probability 0.52
    - Remain A with probability 0.7 * 0.7
    - Go back to A from C,G,T with 0.1*0.1 each
  - In states C,G,T with prob. 0.16 each



	t=1	t=2	t=3	t=4	t=5
A	1	0.7	0.52	0.412	0.3472
C	0	0.1	0.16	0.196	0.2176
G	0	0.1	0.16	0.196	0.2176
T	0	0.1	0.16	0.196	0.2176

Modeling Nucleotide Evolution

During infinitesimal time Δt, there is not enough time for two substitutions to happen on the same nucleotide

So we can estimate $P(x \mid y, \Delta t)$, for $x, y \in \{A, C, G, T\}$

Then let

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

Modeling Nucleotide Evolution

Reasonable assumption: multiplicative
(implying a stationary Markov process)

$S(t+t') = S(t)S(t')$

That is, $P(x \mid y, t+t') = \Sigma_z P(x \mid z, t)\, P(z \mid y, t')$

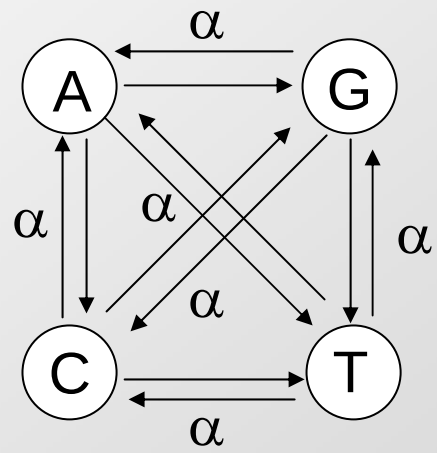
Jukes-Cantor: constant rate of evolution

For short time ε, $S(\varepsilon) =$

$$\begin{pmatrix} 1 - 3\alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & 1 - 3\alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & \alpha\varepsilon & 1 - 3\alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon & 1 - 3\alpha\varepsilon \end{pmatrix}$$

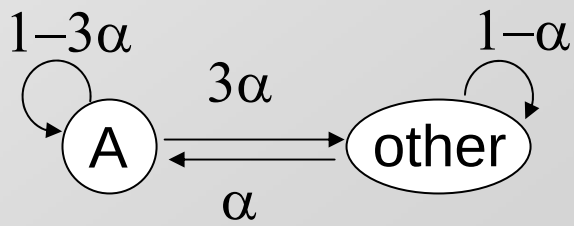
Modeling Nucleotide Evolution

Jukes-Cantor:



For longer times,

$$S(t) = \begin{pmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{pmatrix}$$



Where we can derive:

$$r(t) = \tfrac{1}{4}\,(1 + 3\,e^{-4\alpha t})$$

$$s(t) = \tfrac{1}{4}\,(1 - e^{-4\alpha t})$$

Modeling Nucleotide Evolution

Kimura:

Transitions: A/G, C/T

Transversions: A/T, A/C, G/T, C/G

Transitions (rate α) are much more likely than transversions (rate β)

$$S(t) = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} r(t) & s(t) & u(t) & u(t) \\ s(t) & r(t) & u(t) & u(t) \\ u(t) & u(t) & r(t) & s(t) \\ u(t) & u(t) & s(t) & r(t) \end{pmatrix} \end{array}$$

Where

$$s(t) = \tfrac{1}{4}\left(1 - e^{-4\beta t}\right)$$

$$u(t) = \tfrac{1}{4}\left(1 + e^{-4\beta t} - e^{-2(\alpha+\beta)t}\right)$$

$$r(t) = 1 - 2s(t) - u(t)$$

Distance between two sequences

Given (well-aligned portion of) sequences x^i, x^j,

Define

d_{ij} = distance between the two sequences

One possible definition:

d_{ij} = fraction *f* of sites u where $x^i[u] \neq x^j[u]$

Better model (Jukes-Cantor):

$d_{ij} = - \frac{3}{4} \log(1 - 4f / 3)$

$r(t) = \frac{1}{4} (1 + 3 e^{-4\alpha t})$

$s(t) = \frac{1}{4} (1 - e^{-4\alpha t})$

Observed F = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7])

Actual D = [0.11, 0.23, 0.38, 0.57, 0.82, 1.21, 2.03]

3. From distances to trees

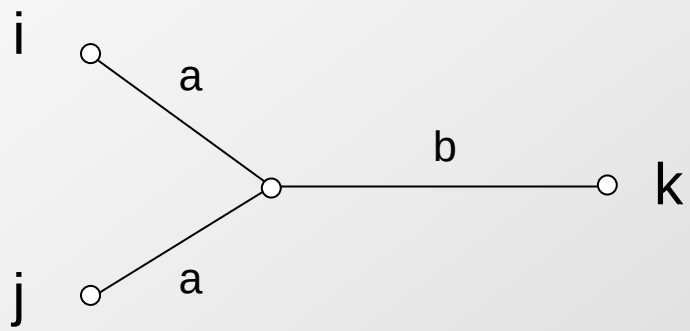
Ultrametric, additive, and general distance matrices

3a. Ultrametric distances

- For all points i, j, k
  - two distances are equal and third is smaller

    d(i,j) <= d(i,k) = d(j,k)

    a+a <= a+b = a+b

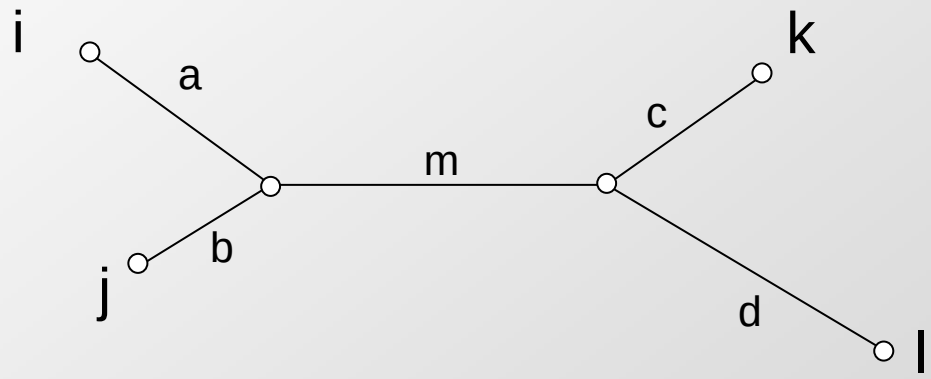


where a <= b

- Result:
  - All paths from labels are equidistant to the root
  - Rooted tree with uniform rates of evolution

3b. Additive distances

- All distances satisfy the four-point condition
  - For all i,j,k,l:
    - d(i,j) + d(k,l) <= d(i,k) + d(j,l) = d(i,l) + d(j,k)
    - (a+b)+(c+d) <= (a+m+c)+(b+m+d) = (a+m+d)+(b+m+c)



- Result:
  - All pairwise distances obtained by traversing a tree

3c. General distances

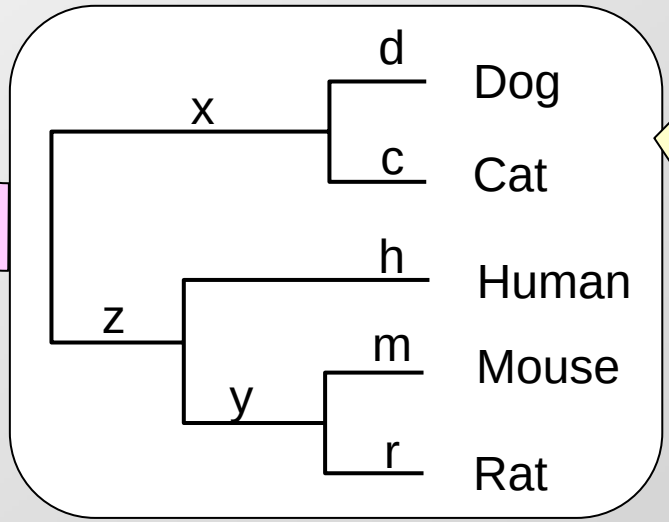
- In practice, a distance matrix is neither ultrametric nor additive
  - Noise
    - Measured distances are not exact
    - Evolutionary model is not exact
  - Fluctuations
    - Regions used to measure distances not representative of the species tree
    - Gene replacement (gene conversion), lateral transfer
    - Varying rates of mutation can lead to discrepancies

- In the general case, tree-building algorithms generate an approximation to the distance matrix
  - Such a tree can be obtained by
    - Enumeration and scoring of all trees (too expensive)
    - Neighbor-Joining (typically gives a good tree)
    - UPGMA (typically gives a poor tree)

Distance matrix ⇔ Phylogenetic tree

	Hum	Mou	Rat	Dog	Cat
Human	0	4	5	7	6
Mouse	h.y.m	0	3	8	5
Rat	h.y.r	m.r	0	9	7
Dog	h.z.x.d	m.y.z.x.d	r.y.z.x.d	0	2
Cat	h.z.x.c	m.y.z.x.c	r.y.z.x.c	d.c	0

Tree implies a distance matrix M_{ij}



Map distances D_{ij} to a tree

$\min \sum_{ij} (D_{ij}-M_{ij})^2$

Goal:

Minimize discrepancy between observed distances and tree-based distances

4. Tree-building algorithms

Mapping a distance matrix to a tree

4a: UPGMA (aka. Hierarchical Clustering)

(Unweighted Pair Group Method with Arithmetic mean)

Initialization:

Assign each x_i into its own cluster C_i

Define one leaf per sequence, height 0

Iteration:

Find two clusters C_i, C_j s.t. d_{ij} is min

Let $C_k = C_i \cup C_j$

Define node connecting C_i, C_j,

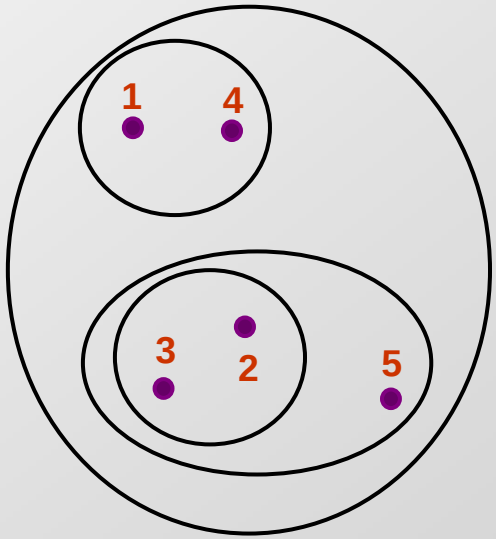
& place it at height $d_{ij}/2$

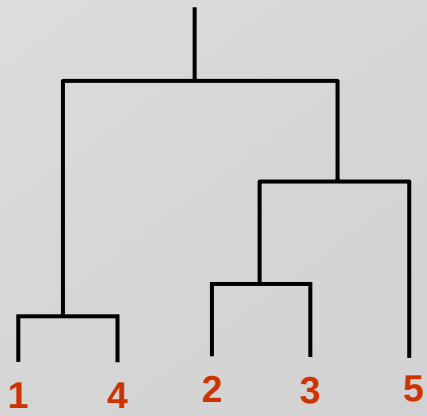
Delete C_i, C_j

Termination:

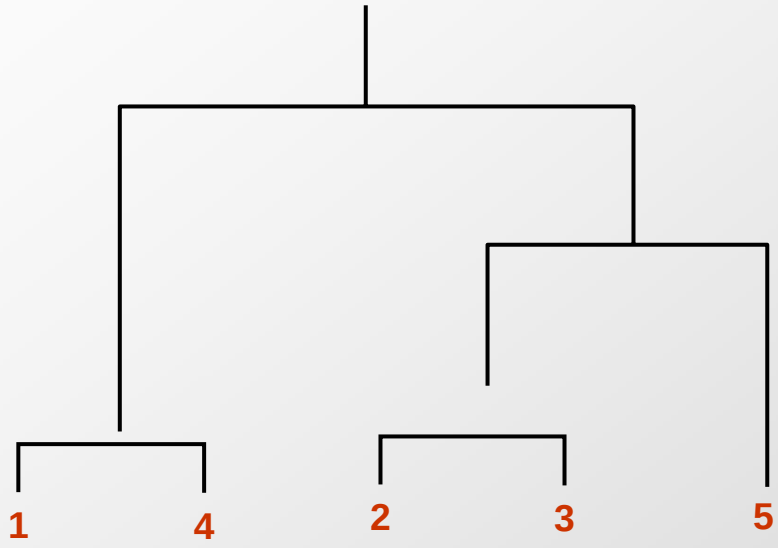
When two clusters i, j remain,

place root at height $d_{ij}/2$





Ultrametric Distances & UPGMA



UPGMA is guaranteed to build the correct tree if distance is ultrametric

Proof:

1. The tree topology is unique, given that the tree is binary
2. UPGMA constructs a tree obeying the pairwise distances

Weakness of UPGMA

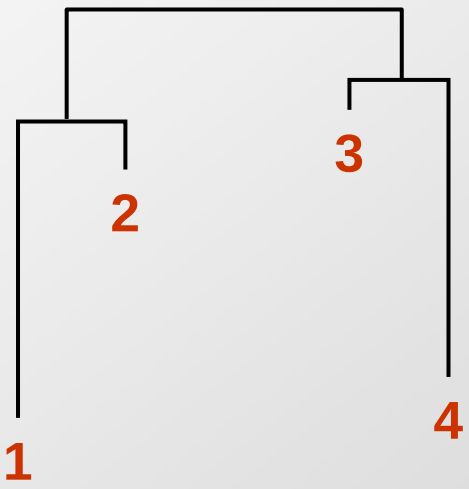
Molecular clock assumption:

implies time is constant for all species

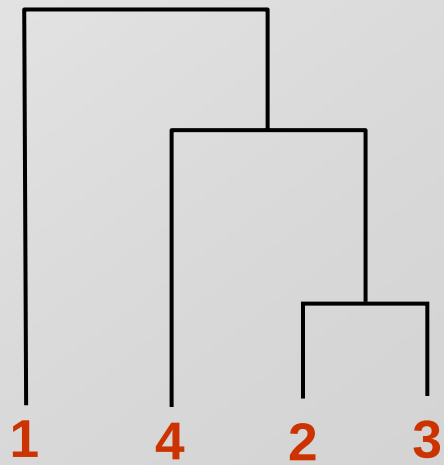
However, certain species (e.g., mouse, rat) evolve much faster

Example where UPGMA messes up:

Correct tree



UPGMA



4b. Neighbor-Joining

- Guaranteed to produce the correct tree if distance is additive
- May produce a good tree even when distance is not additive

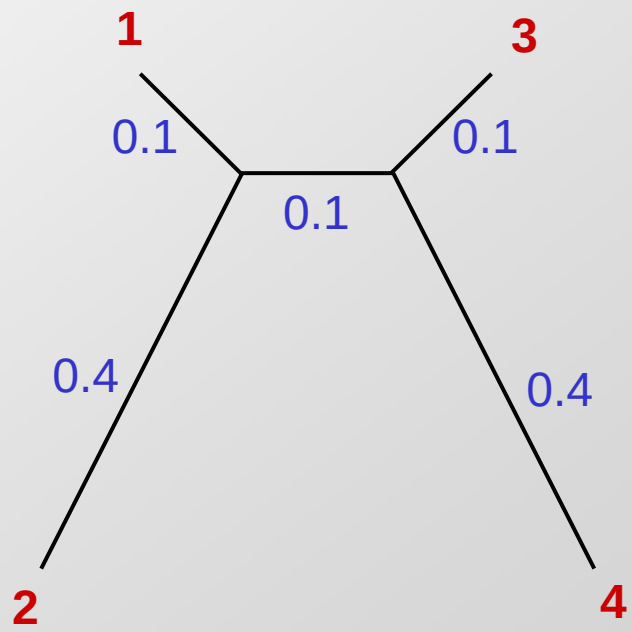
Step 1: Finding neighboring leaves

Define

$$D_{ij} = d_{ij} - (r_i + r_j)$$

Where

$$r_i = \frac{1}{|L| - 2} \Sigma_k d_{ik}$$



Claim: The above "magic trick" ensures that D_{ij} is minimal **iff** i, j are neighbors
Proof: Beyond the scope of this lecture (Durbin book, p. 189)

Algorithm: Neighbor-joining

Initialization:

Define T to be the set of leaf nodes, one per sequence

Let L = T

Iteration:

Pick i, j s.t. D_{ij} is minimal

Define a new node k, and set $d_{km} = ½ (d_{im} + d_{jm} – d_{ij})$ for all $m \in L$

Add k to T, with edges of lengths $d_{ik} = ½ (d_{ij} + r_i – r_j)$

Remove i, j from L;

Add k to L

Termination:

When L consists of two nodes, i, j, and the edge between them of length d_{ij}

5. Alignment-based algorithms

Parsimony (set-based)

Parsimony (Dynamic Programming)

Maximum Likelihood

5a. Parsimony

- One of the most popular methods

Idea:

Find the tree that explains the observed sequences with a minimal number of substitutions

Two computational sub-problems:

1. Find the parsimony cost of a given tree (easy)

2. Search through all tree topologies (hard)

Parsimony Scoring

Given a tree, and an alignment column

Label internal nodes to minimize the number of required substitutions

Initialization:

Set cost C = 0; k = 2N – 1

Iteration:

If k is a leaf, set $R_k = \{ x^k[u] \}$

If k is not a leaf,

Let i, j be the daughter nodes;

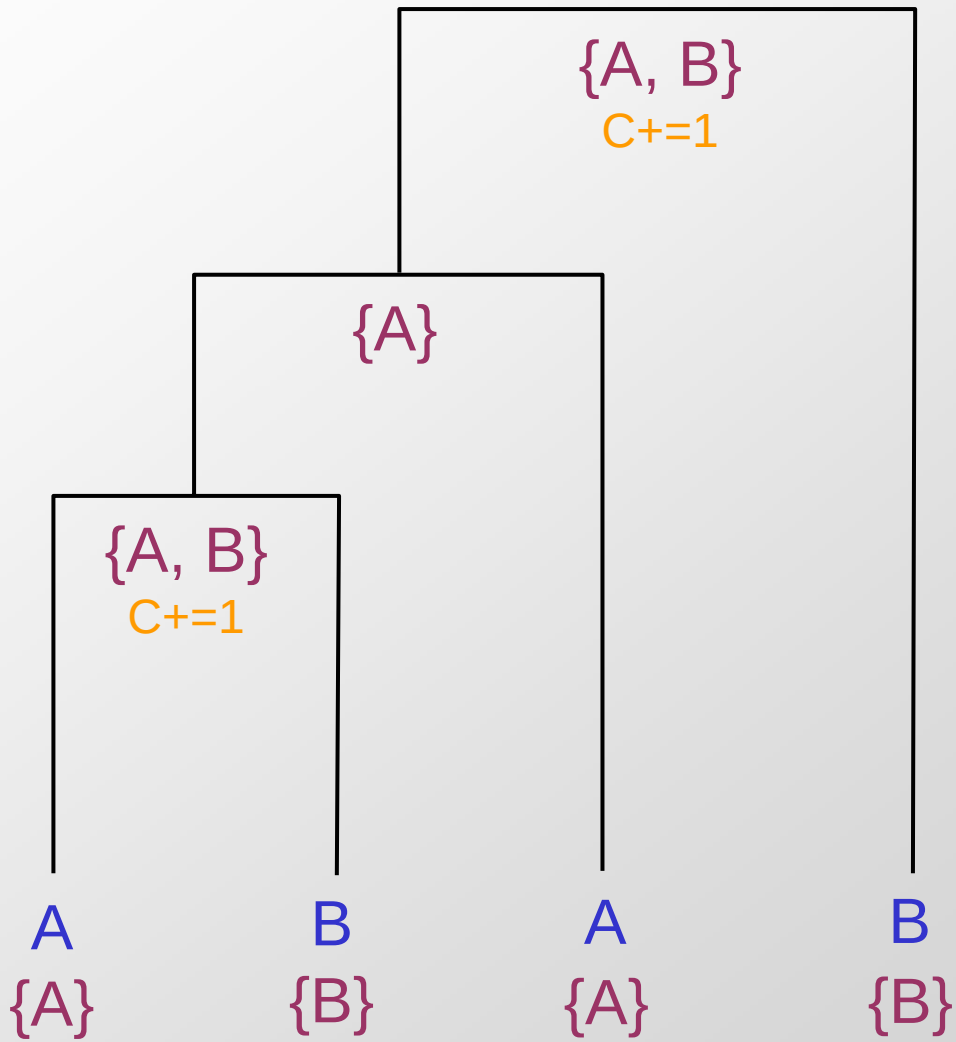
Set $R_k = R_i \cap R_j$ if intersection is nonempty

Set $R_k = R_i \cup Rj$, and C += 1, if intersection is empty

Termination:

Minimal cost of tree for column u, = C

Example



Traceback to find ancestral nucleotides

Traceback:

1. Choose an arbitrary nucleotide from R_{2N-1} for the root

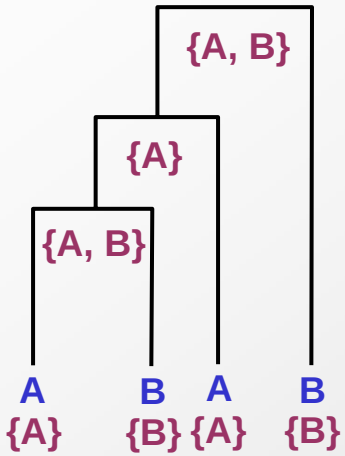
2. Having chosen nucleotide r for parent k,

   If $r \in R_i$ choose r for daughter i

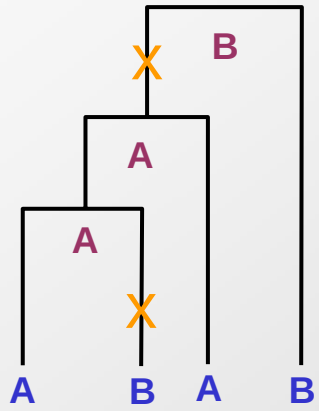
   Else, choose arbitrary nucleotide from R_i

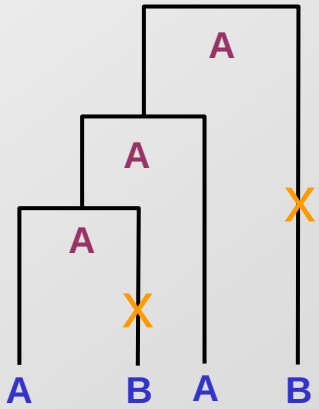
Easy to see that this traceback produces some assignment of cost C

Example

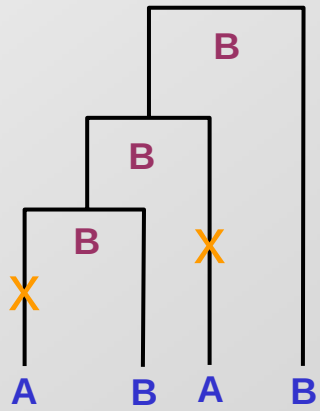


Accessible to traceback





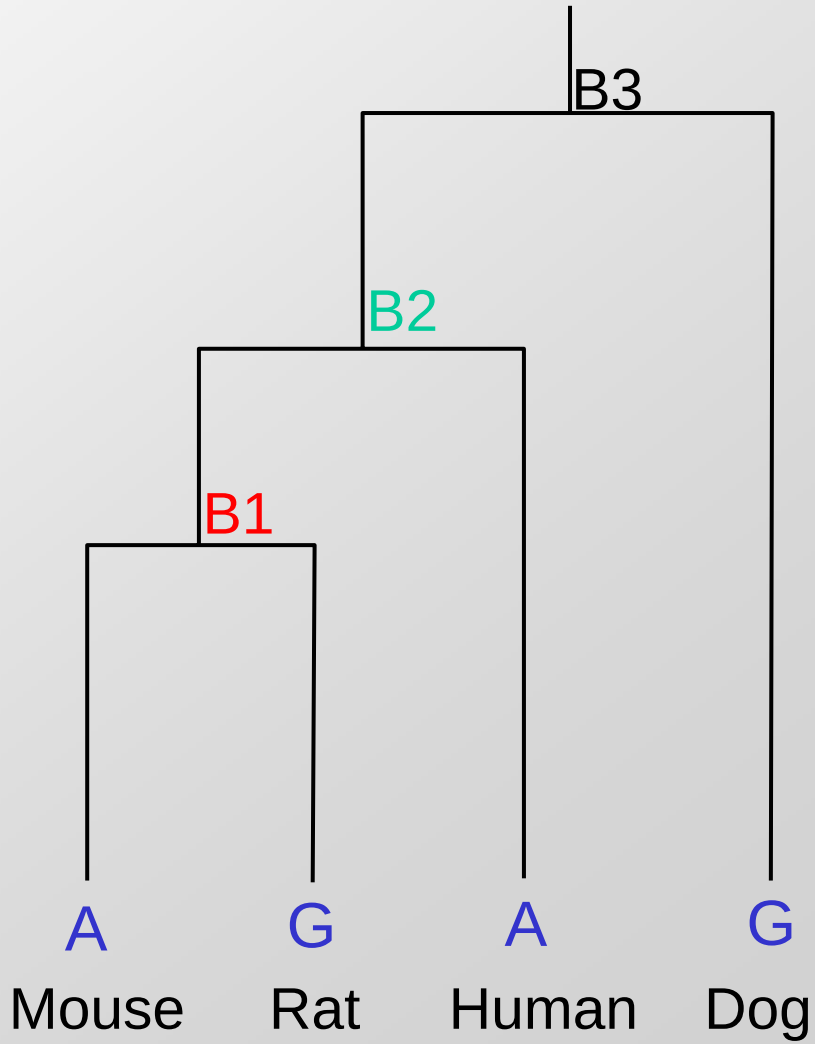
Still optimal, but
not found by traceback



5b. Parsimony with dynamic programming

	M	R	B1	H	B2	D	B3
A	0	1	1	0	1	1	2
C	1	1	2	1	3	1	4
G	1	0	1	1	2	0	2
T	1	1	2	1	3	1	4

- Each cell (N,C) represents the min cost of the subtree rooted at N, if the label at N is C.
- Update table by walking up the tree from the leaves to the root, remembering max choices.
- Traceback from root to leaves to construct a min cost assignment



5c. Maximum Likelihood Methods

Input: Proposed topology T

Output: Prob. that proposed tree gave rise to observed data

Search: Heuristic MCMC search for max likelihood tree.

$B\hat{},T\hat{} = \text{argmax}_{B,T}\ P(D,B,T)$

$= \text{argmax}_{B,T}\ P(D|B,T)\ P(B,T)$

Likelihood P(Data|BranchLengths,Topology)

Prior P(B,T): typically uniform/can use to guide search

Iterate: Iterate over proposed topologies.

- Given current topology T, branch lengths B:
  - Propose many alternative (T',B'), by modifying existing T→T', and inferring branch lengths B' that maximize P(D|B',T')
  - Evaluate P(D|B,T) and P(D|B',T')
  - Select one T' at random based on increase in likelihood
- Heuristics for proposing new topology T'
  - Nearest-neighbor interchange, subtree cut-and-paste, rotations

Advantages/disadvantages of ML methods

- Advantages:
  - Are inherently statistical and evolutionary model-based.
  - Usually the most ‘consistent’ of the methods available.
  - Can be used for character (can infer the exact substitutions) and rate analysis.
  - Can be used to infer the sequences of the extinct (hypothetical) ancestors.
  - Can help account for branch-length effects in unbalanced trees.
  - Can be applied to nucleotide or amino acid sequences, and other types of data.

- Disadvantages:
  - Are not as simple and intuitive as many other methods.
  - Are computationally very intense (Iimits number of taxa and length of sequence).
  - Like parsimony, can be fooled by high levels of homoplasy.
  - Violations of the assumed model can lead to incorrect trees.

Bootstrapping to get the best trees

Main outline of algorithm

1. Select random columns from a multiple alignment – one column can then appear several times

2. Build a phylogenetic tree based on the random sample from (1)

3. Repeat (1), (2) many (say, 1000) times

4. Output the tree that is constructed most frequently

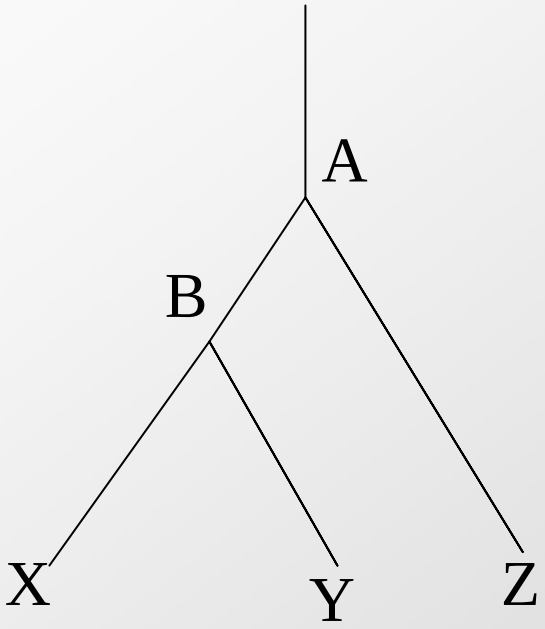
Summary

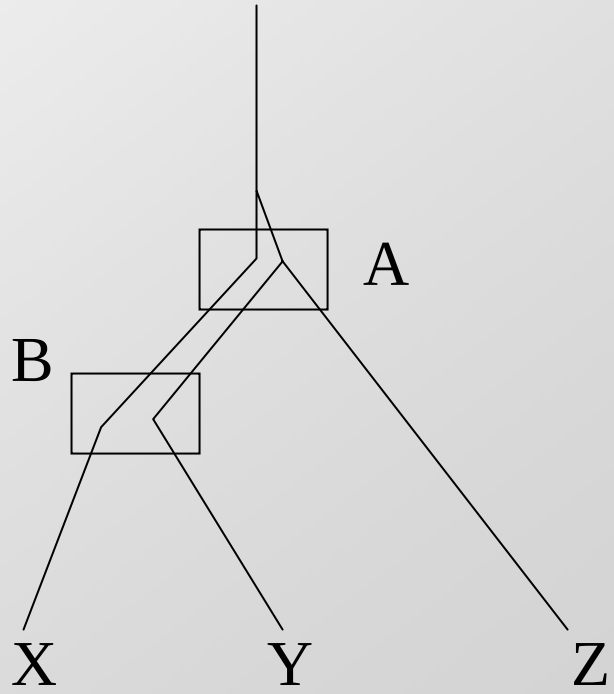
- Basics of phylogeny
  - Characters, traits, nodes, branches, lineages
  - Gene trees, species trees
- Modeling sequence evolution
  - Turning sequence data into distances
  - Probabilistic models of nucleotide divergence
  - Jukes-Cantor 1-parameter model, Kimura 2-parameter model
- From distances to trees
  - Ultrametric, Additive, General Distances
  - UPGMA, Neighbor Joining, guarantees and limitations
  - Least-squared error, minimum evolution
- From alignments to trees
  - Parsimony methods: set-based vs. dynamic programming
  - Maximum likelihood methods
  - MCMC and heuristic search

Extra Time?

Recitation tomorrow: Gene vs. Species evolution

- Genes can start diverging before species separate
  - Genetic polymorphism within population could exist
  - After divergence, forms evolve differently in each species
  - Gene divergence could predate species diverge
  - Gene tree topology could be misleading





- Solution: Use multiple genes to infer a species tree

Phylogenomics

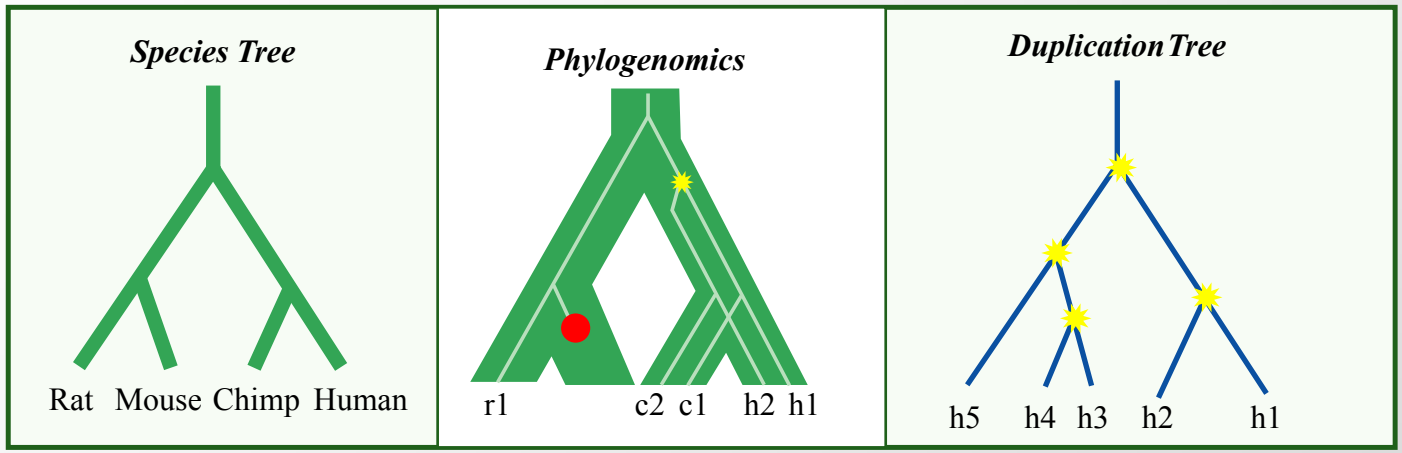


Figure by MIT OpenCourseWare.

Many species One gene → Many species Many genes ← One species Many genes

- Traditional phylogenetics focused on uniform trees
  - Any topology makes a good story
- Phylogenomics imposes additional constraints
  - Gene trees evolve inside species trees
  - Errors imply large-scale duplications and losses

Extending traditional max likelihood methods

- Traditional max likelihood (phylogenetics)

$$\hat{B},\hat{T} = \underset{B,T}{\operatorname{argmax}}\, P(D,B,T)$$
$$\underset{B,T}{\operatorname{argmax}}\, P(D|B,T)\, \cancel{P(B,T)}$$

- Extended likelihood function (phylogenomics)

$$\hat{B},\hat{T} = \underset{B,T}{\operatorname{argmax}}\, P(D,B,T,R|\mathbf{E})$$
$$\underset{B,T}{\operatorname{argmax}}\, P(D|B,T)\, P(B|T,R,\mathbf{E})\cancel{P(R|T,E)P(T|\mathbf{E})}$$

Likelihood of data given proposed branch lengths

Likelihood of proposed branch lengths (given species evolution)

Evaluation: Large increase in accuracy

program	accuracy
SPIML	71.7%
SPIDIR	43.2%
PHYML	23.5%
BIONJ	29.6%
MrBayes	31.4%
DNAPARS	22.2%

Syntenic regions:

Increasing number of species

Diverse lineages

➔ Great increase in accuracy

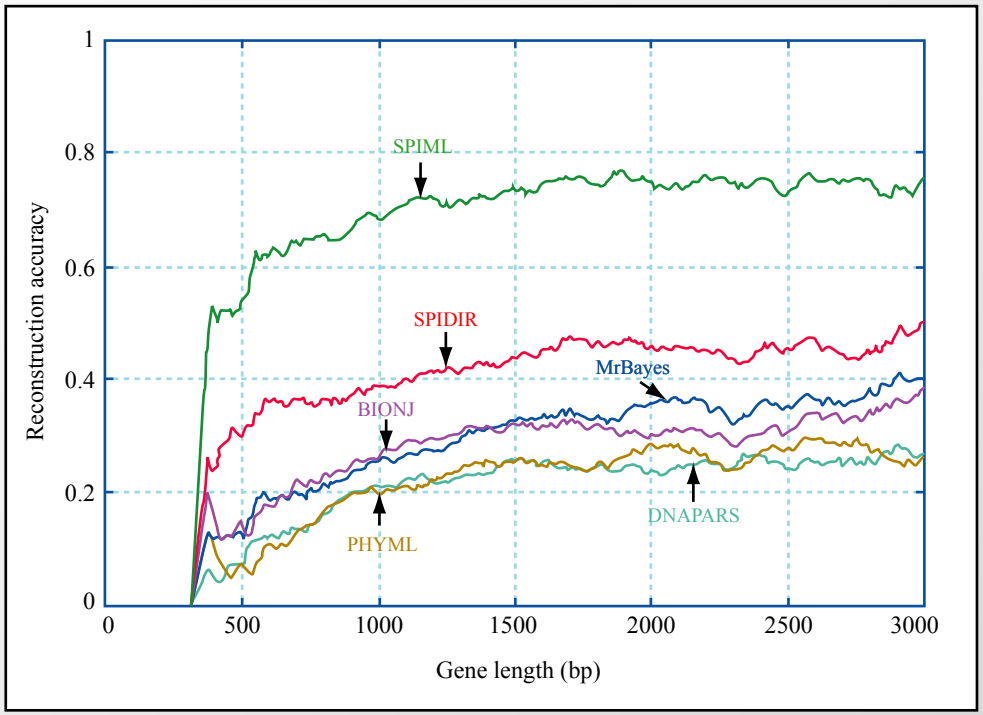


Figure by MIT OpenCourseWare.

- Simulated data:
  - Run (generative) model
  - 1 dup event ⇔ many dup genes
  - Method robust to dup/loss

Length correlation:

Maximally use data

More data → closer to truth

Still depends on gene length

But much higher than other methods