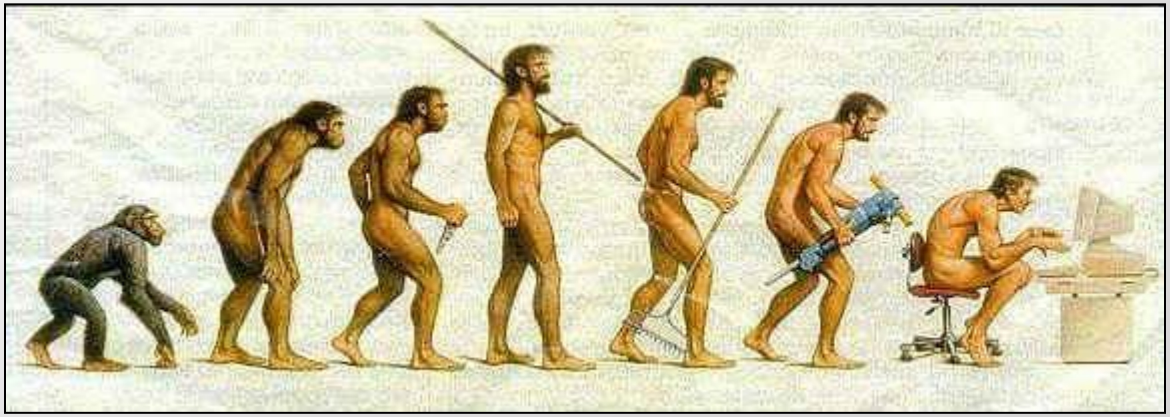
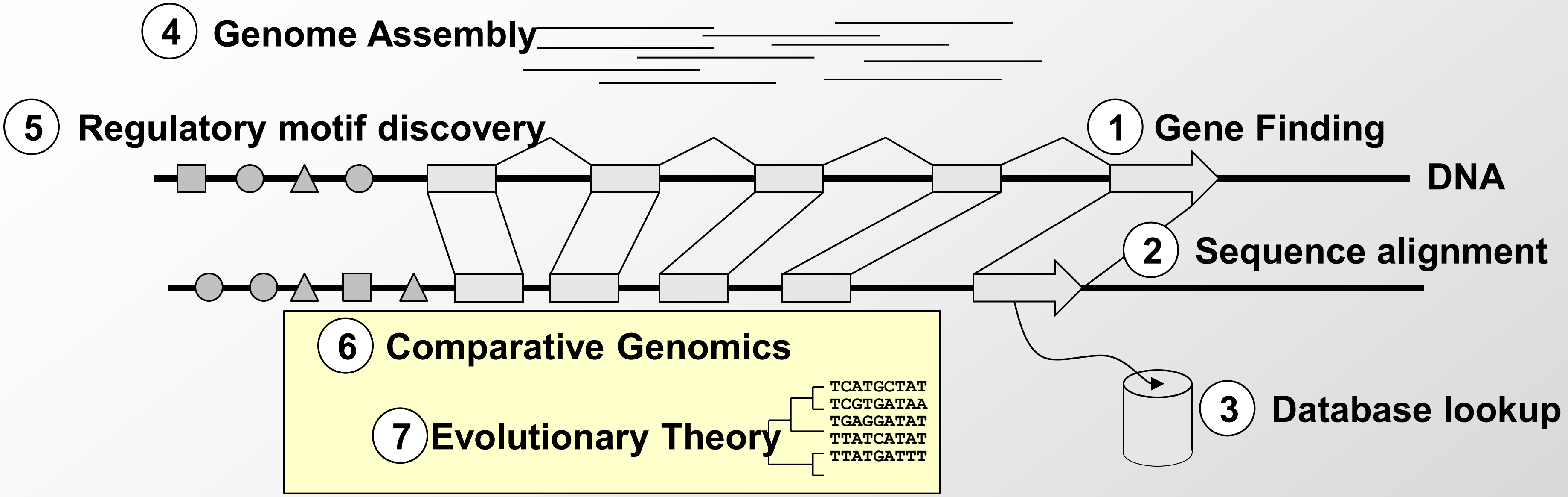
Lecture 18
Molecular Evolution and Phylogenetics



Patrick Winston's 6.034

Somewhere, something went wrong…

Challenges in Computational Biology



4 Genome Assembly

5 Regulatory motif discovery

1 Gene Finding

DNA

2 Sequence alignment

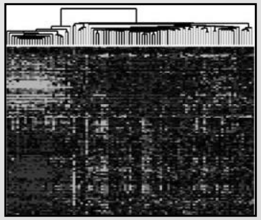
6 Comparative Genomics

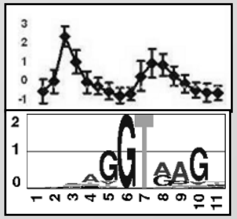
7 Evolutionary Theory

3 Database lookup

8 Gene expression analysis

RNA transcript

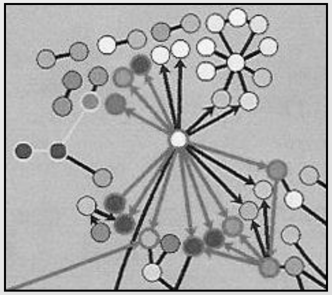




9 Cluster discovery

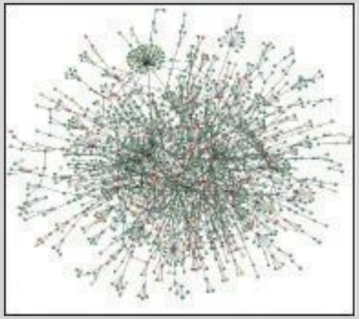
10 Gibbs sampling

11 Protein network analysis

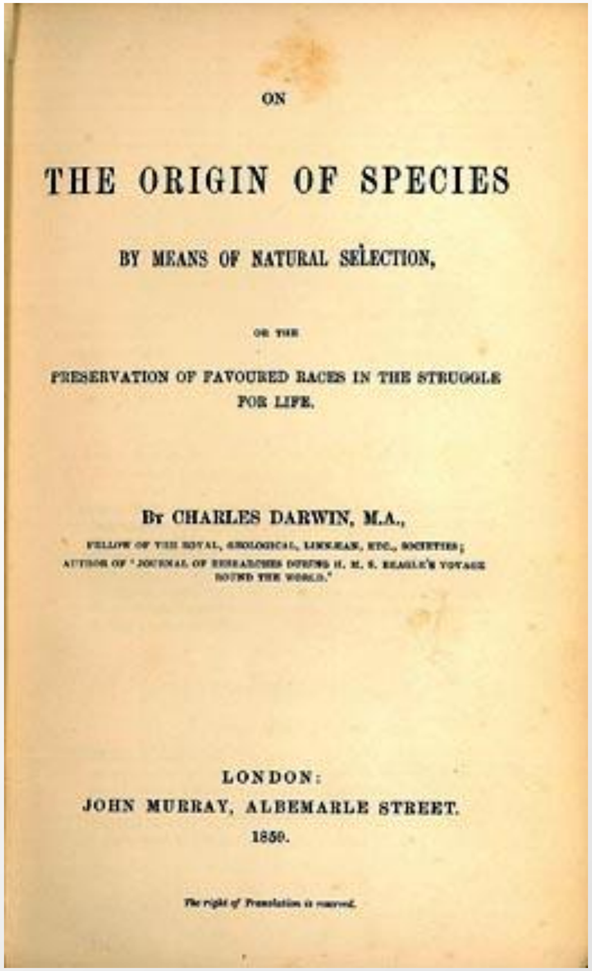


12 Metabolic modelling

13 Emerging network properties



Concepts of Darwinian Evolution



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE
FOR LIFE.

By CHARLES DARWIN, M.A.,

FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF "JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE
ROUND THE WORLD."

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

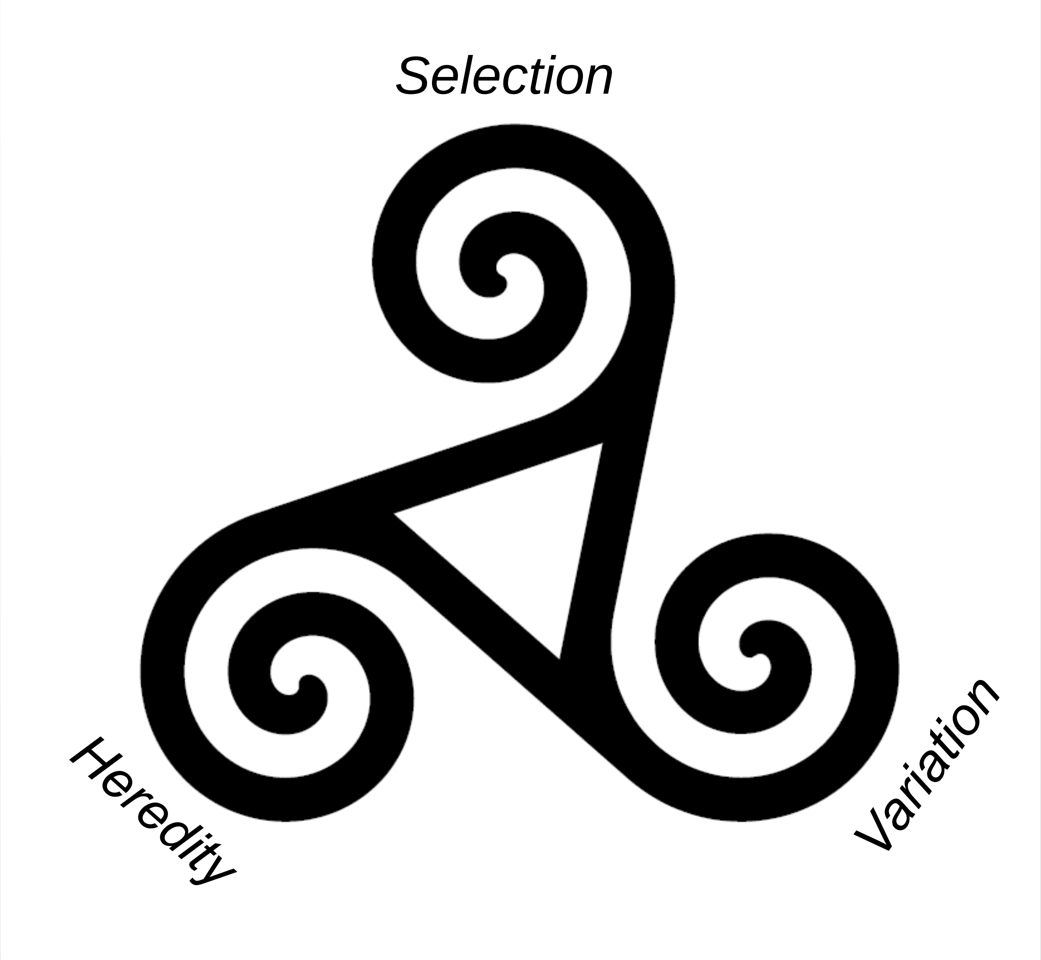


Image in the public domain.

Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Concepts of Darwinian Evolution

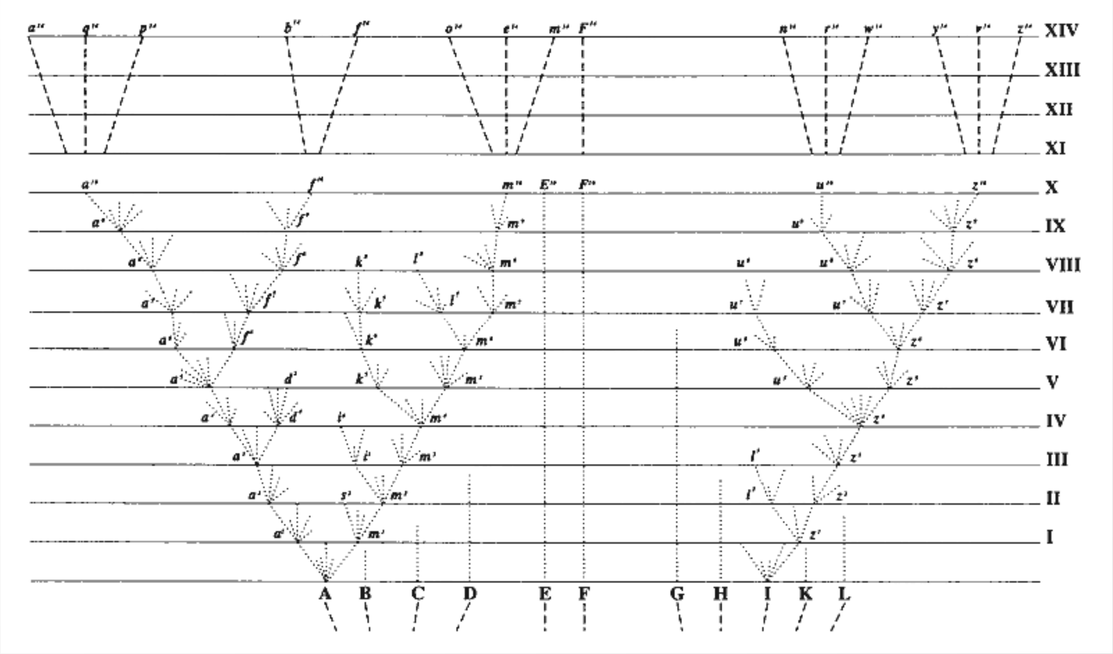

a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
X
IX
VIII
VII
VI
V
IV
III
II
I
A
B
C
D
E
F
G
H
I
K
L

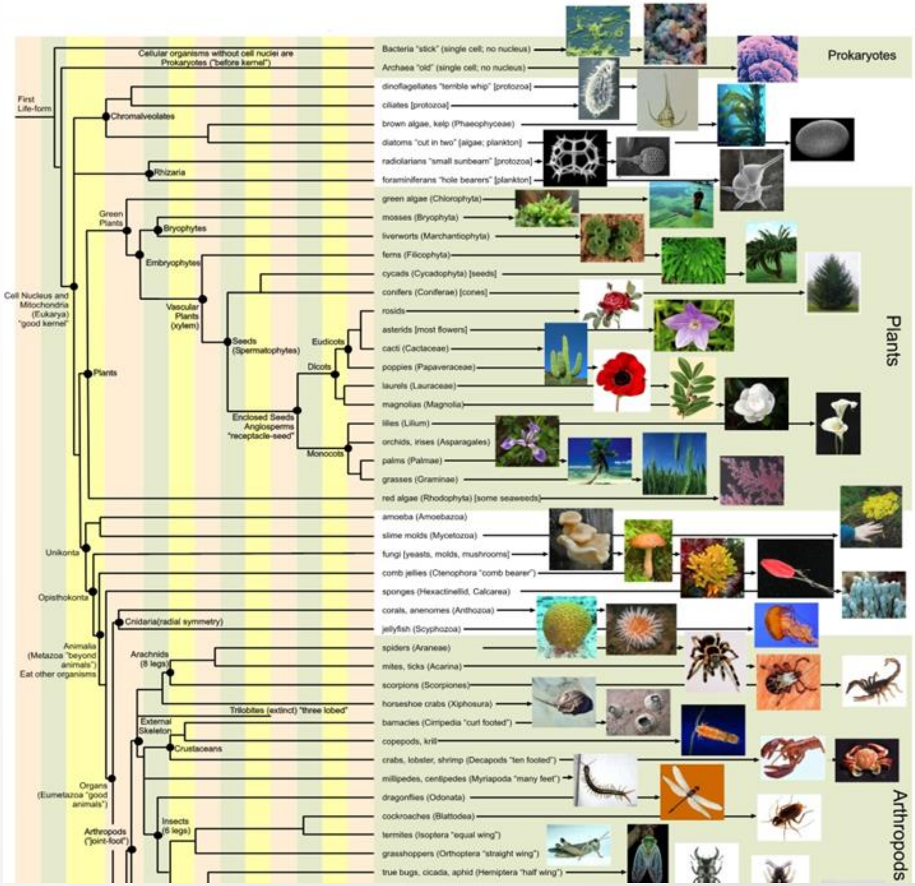
Image in the public domain.

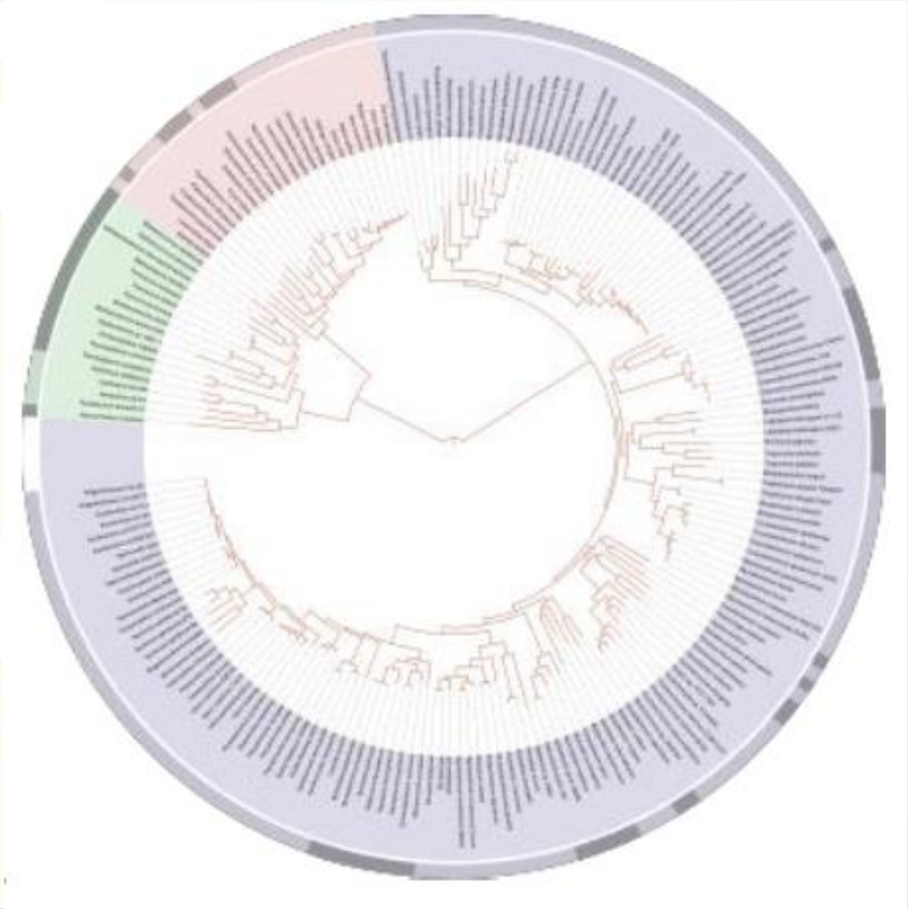
Charles Darwin 1859. *Origin of Species* [one and only illustration]: "descent with modification"

Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Tree of Life


Prokaryotes
Plants
Arthropods
First Life-form
Cellular organisms without cell nuclei are Prokaryotes ("before kernel")
Chromalveolates
Rhizaria
Green Plants
Bryophytes
Embryophytes
Vascular Plants (xylem)
Cell Nucleus and Mitochondria (Eukarya) "good kernel"
Seeds (Spermatophytes)
Eudicots
Dicots
Plants
Enclosed Seeds Angiosperms "receptacle-seed"
Monocots
Unikonta
Opisthokonta
Cnidaria(radial symmetry)
Animalia (Metazoa "beyond animals") Eat other organisms
Arachnids (8 legs)
Trilobites (extinct) "three lobed"
External Skeleton
Crustaceans
Organs (Eumetazoa "good animals")
Insects (6 legs)
Arthropods ("joint-foot")
Bacteria "stick" (single cell; no nucleus)
Archaea "old" (single cell; no nucleus)
dinoflagellates "terrible whip" [protozoa]
ciliates [protozoa]
brown algae, kelp (Phaeophyceae)
diatoms "cut in two" [algae; plankton]
radiolarians "small sunbeam" [protozoa]
foraminiferans "hole bearers" [plankton]
green algae (Chlorophyta)
mosses (Bryophyta)
liverworts (Marchantiophyta)
ferns (Filicophyta)
cycads (Cycadophyta) [seeds]
conifers (Coniferae) [cones]
rosids
asterids [most flowers]
cacti (Cactaceae)
poppies (Papaveraceae)
laurels (Lauraceae)
magnolias (Magnolia)
lilies (Lilium)
orchids, irises (Asparagales)
palms (Palmae)
grasses (Graminae)
red algae (Rhodophyta) [some seaweeds]
amoeba (Amoebazoa)
slime molds (Mycetozoa)
fungi [yeasts, molds, mushrooms]
comb jellies (Ctenophora "comb bearer")
sponges (Hexactinellid, Calcarea)
corals, anenomes (Anthozoa)
jellyfish (Scyphozoa)
spiders (Araneae)
mites, ticks (Acarina)
scorpions (Scorpiones)
horseshoe crabs (Xiphosura)
barnacles (Cirripedia "curl footed")
copepods, krill
crabs, lobster, shrimp (Decapods "ten footed")
millipedes, centipedes (Myriapoda "many feet")
dragonflies (Odonata)
cockroaches (Blattodea)
termites (Isoptera "equal wing")
grasshoppers (Orthoptera "straight wing")
true bugs, cicada, aphid (Hemiptera "half wing")



Goals for today: Phylogenetics

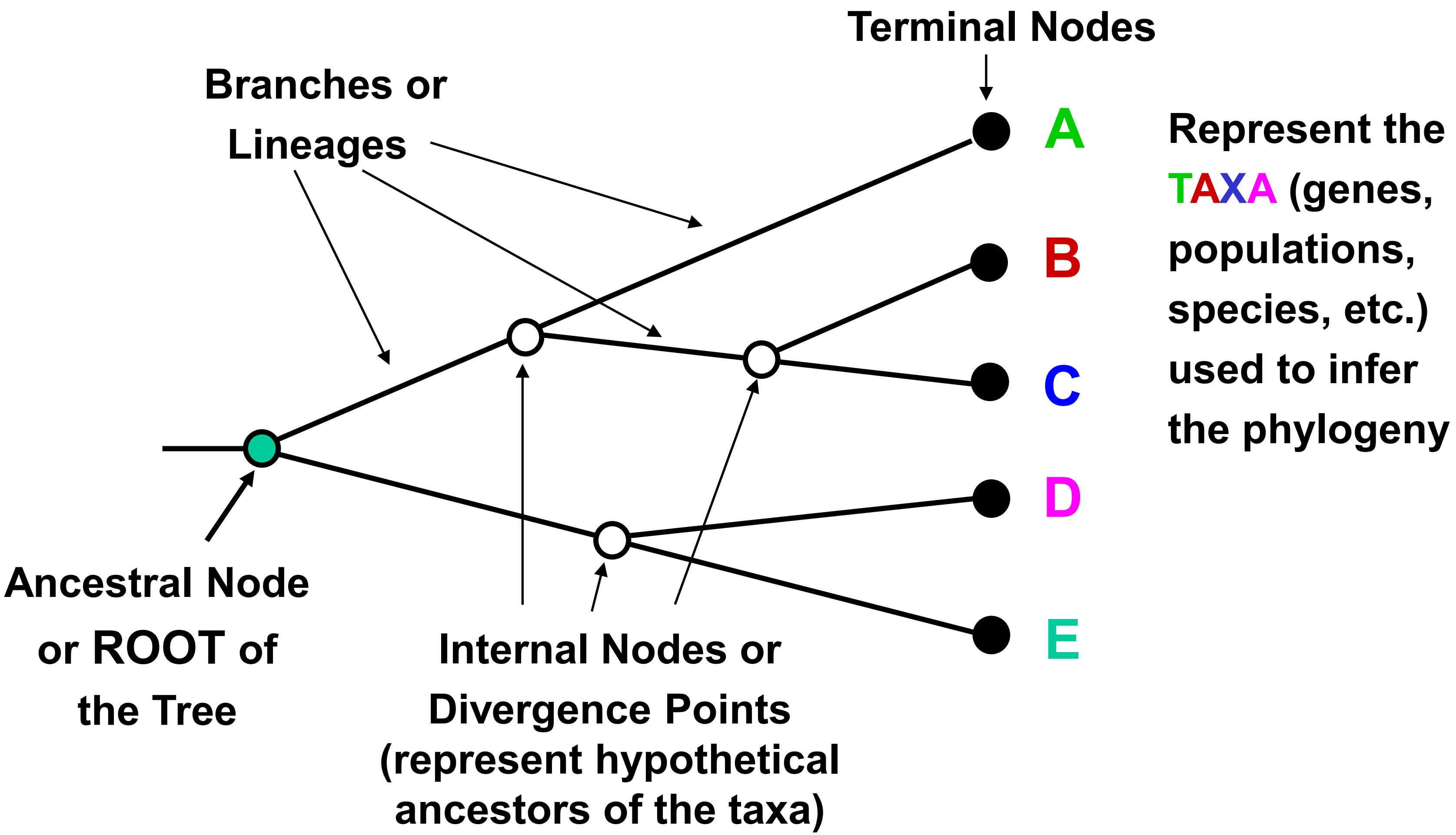
- **Basics of phylogeny: Introduction and definitions**
  - Characters, traits, nodes, branches, lineages, topology, lengths
  - Gene trees, species trees, cladograms, chronograms, phylograms

1. **From alignments to distances: Modeling sequence evolution**
   - Turning pairwise sequence alignment data into pairwise distances
   - Probabilistic models of divergence: Jukes Cantor/Kimura/hierarchy
2. **From distances to trees: Tree-building algorithms**
   - Tree types: Ultrametric, Additive, General Distances
   - Algorithms: UPGMA, Neighbor Joining, guarantees and limitations
   - Optimality: Least-squared error, minimum evolution (require search)
3. **From alignments to trees: Alignment scoring given a tree**
   - Parsimony: greedy (union/intersection) vs. DP (summing cost)
   - ML/MAP (includes back-mutations, lengths): peeling algorithm (DP)
4. Tree of Life in Genomic Era
   - The prokaryotic problem (no real taxa and HGT)
   - Interpreting the forest of life

Introduction: Basics and Definitions

Characters, traits, gene/species trees

Common Phylogenetic Tree Terminology


Terminal Nodes
Branches or Lineages
A
B
C
D
E
Represent the TAXA (genes, populations, species, etc.) used to infer the phylogeny
Ancestral Node or ROOT of the Tree
Internal Nodes or Divergence Points (represent hypothetical ancestors of the taxa)

Extinctions part of life

Phylogenetic tree showing archosaurs, dinosaurs, birds, etc. through geologic time removed due to copyright restrictions.

Phylogenetics

General Problem:
Infer complete ancestry of a set of ‘**objects**’ based on knowledge of their ‘**traits**’

Mammal family tree removed due to copyright restrictions.

‘Objects’ can be: Species, Genes, Cell types, Diseases, Cancers, Languages, Faiths, Cars, Architectural Styles

‘Traits’ can be: Morphological, molecular, gene expression, TF binding, motifs, words…

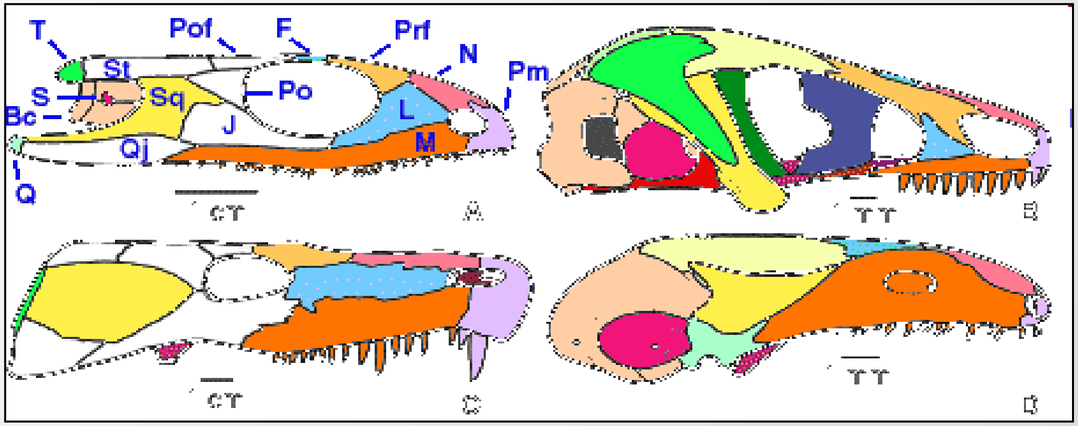
Historical record varies: Fossils, imprints, timing of geological events, ‘living fossils’, sequencing of extinct species, paintings, stories.

Today: Phylogenies using only extant species data
➔ **gene trees** (paralog / ortholog / homolog trees)

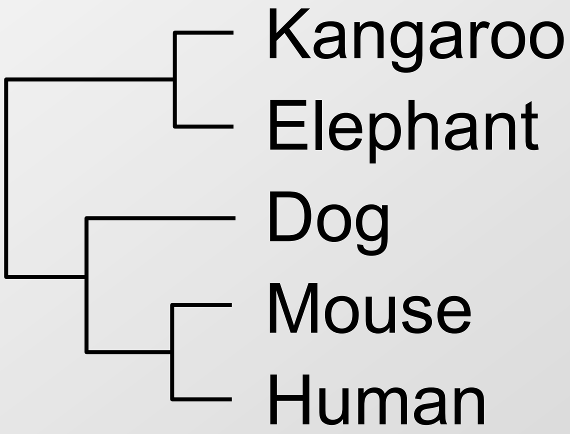
Inferring Phylogenies: Traits and Characters

Trees can be inferred by several criteria:

- Traditional traits: Morphology data



- Modern traits: Molecular data



Kangaroo	ACAGTGACGCCCCAAACGT
Elephant	ACAGTGACGCTACAAACGT
Dog	CCTGTGACGTAACAAACGA
Mouse	CCTGTGACGTAGCAAACGA
Human	CCTGTGACGTAGCAAACGA

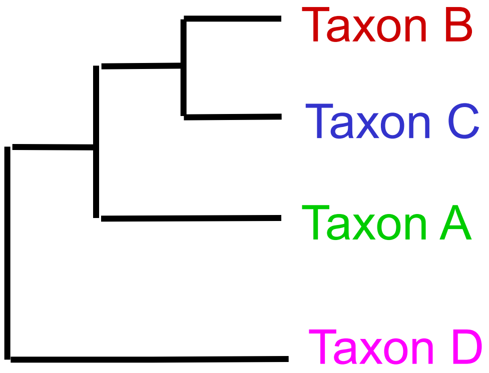
From physiological traits to DNA characters

- 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

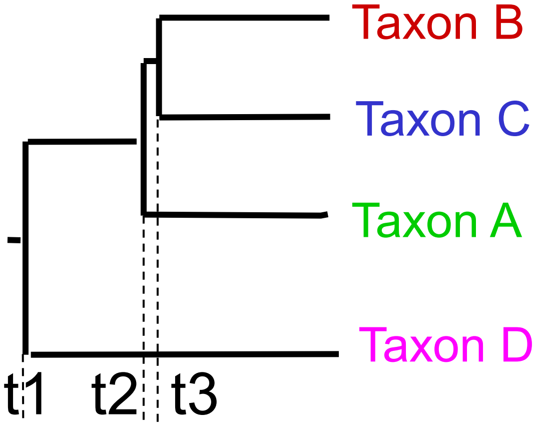
Three types of trees

Cladogram



Topology only

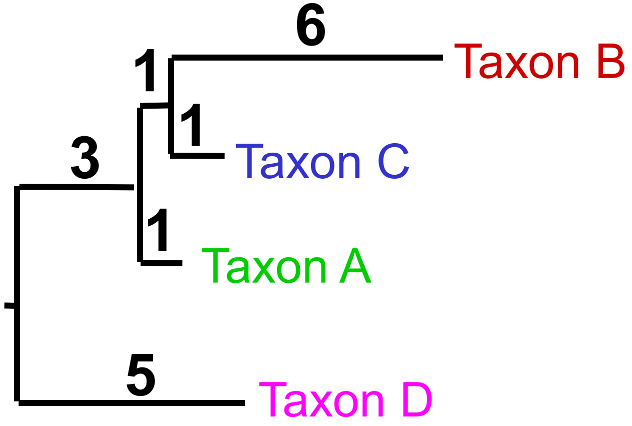
Chronogram



Topology +

Divergence times

Phylogram

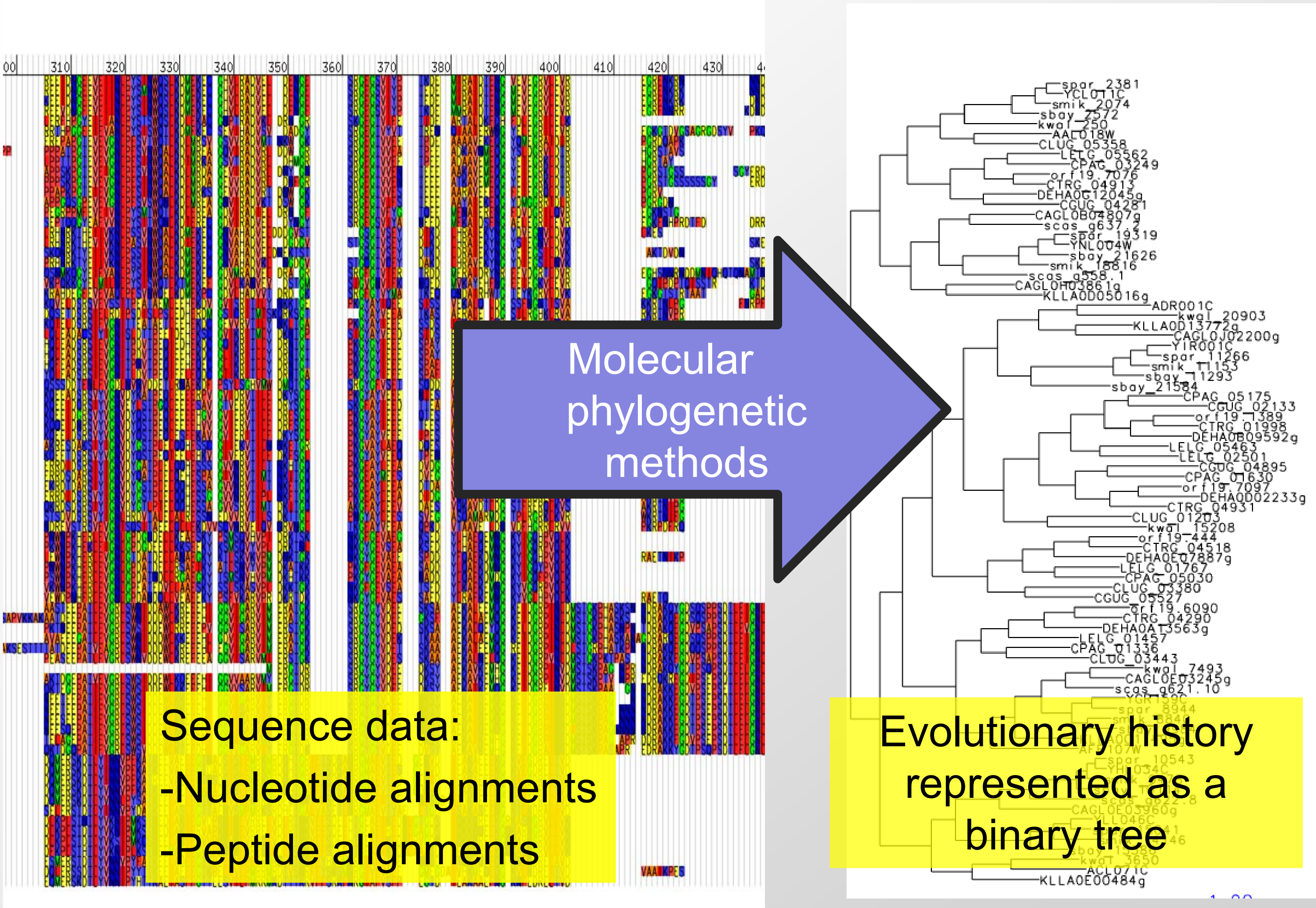


Topology +

Divergence times +

Divergence rates

Inferring a tree from nucleotides/peptides



Molecular phylogenetic methods

Sequence data:

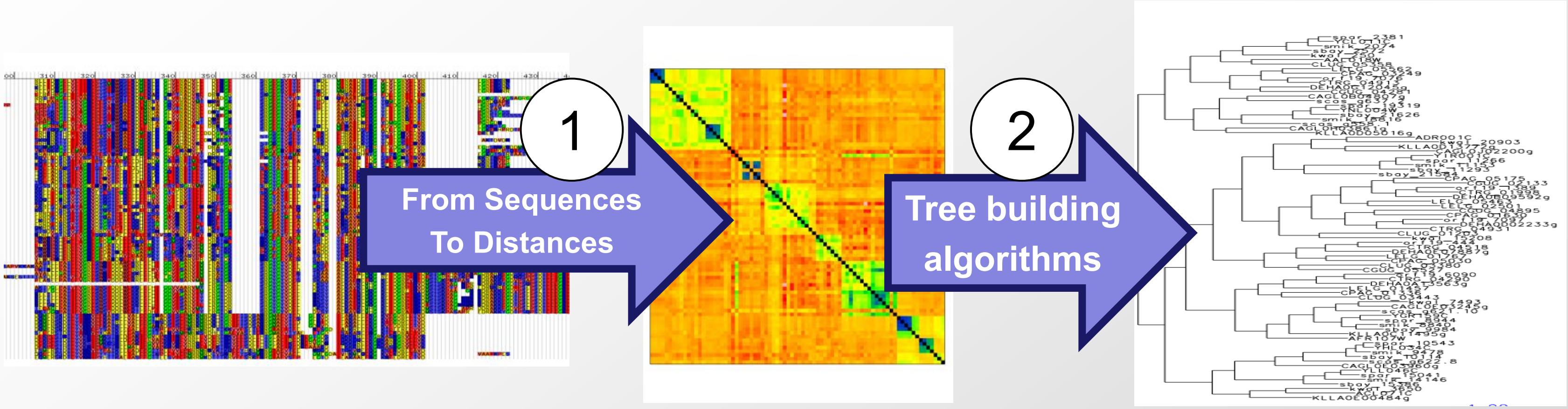
-Nucleotide alignments

-Peptide alignments

Evolutionary history represented as a binary tree

Two basic approaches for phylogenetic inference

Distance based



1

From Sequences To Distances

2

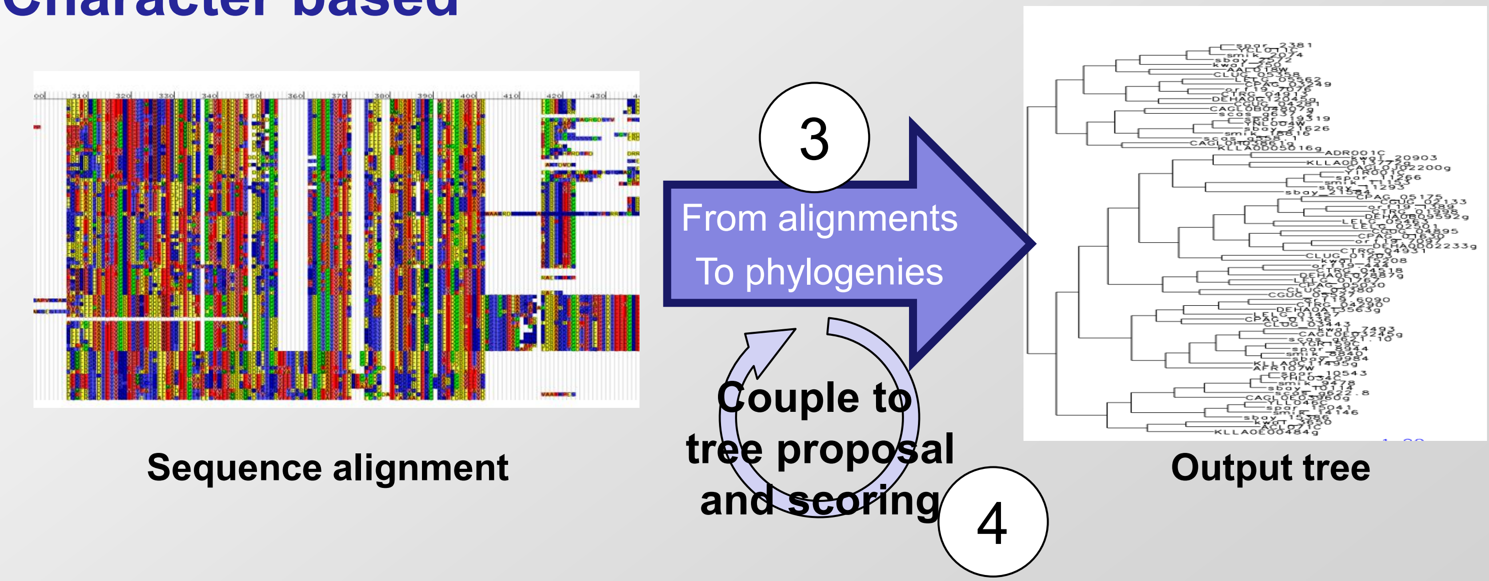
Tree building algorithms

Sequence alignment

Pair-wise distance matrix

Output tree

Character based



3

From alignments To phylogenies

Couple to tree proposal and scoring

4

Sequence alignment

Output tree

Goals for today: Phylogenetics

- **Basics of phylogeny: Introduction and definitions**
  - Characters, traits, nodes, branches, lineages, topology, lengths
  - Gene trees, species trees, cladograms, chronograms, phylograms

1. **From alignments to distances: Modeling sequence evolution**

   1
   - Turning pairwise sequence alignment data into pairwise distances
   - Probabilistic models of divergence: Jukes Cantor/Kimura/hierarchy

2. **From distances to trees: Tree-building algorithms**

   2
   - Tree types: Ultrametric, Additive, General Distances
   - Algorithms: UPGMA, Neighbor Joining, guarantees and limitations
   - Optimality: Least-squared error, minimum evolution (require search)

3. **From alignments to trees: Alignment scoring given a tree**

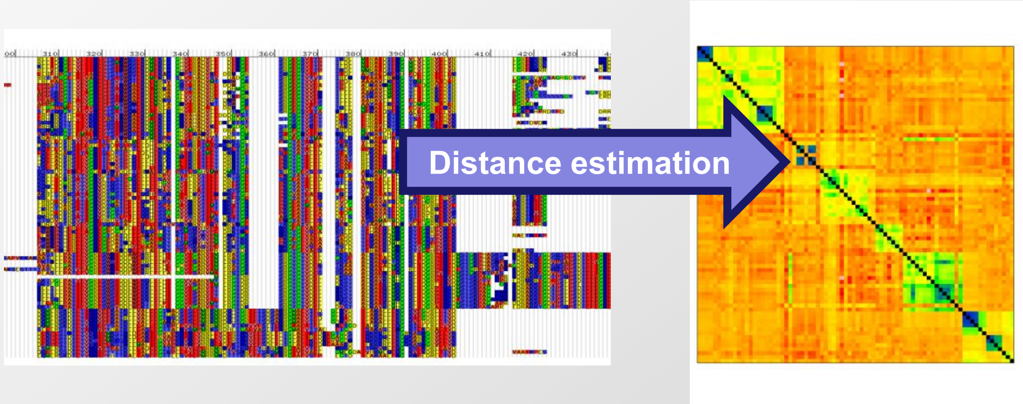
   3
   - Parsimony: greedy (union/intersection) vs. DP (summing cost)
   - ML/MAP (includes back-mutations, lengths): peeling algorithm (DP)

4. Tree of Life in Genomic Era

   4
   - The prokaryotic problem (no real taxa and HGT)
   - Interpreting the forest of life

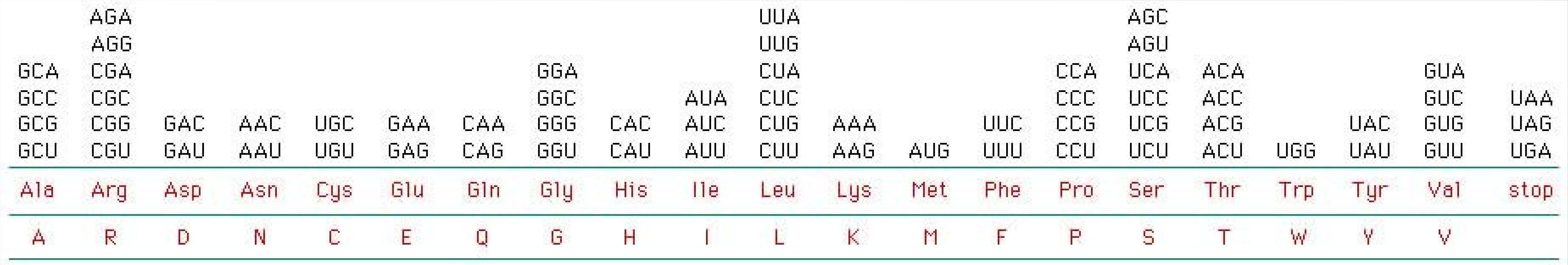
1. From alignments to distances

Modeling evolutionary rates



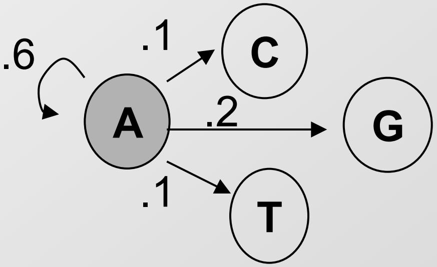
Measuring evolutionary rates

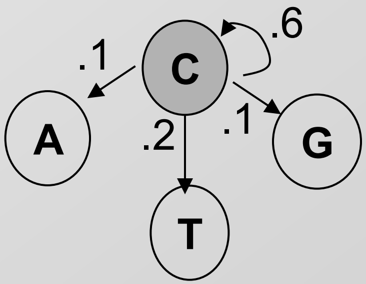
- Nucleotide divergence
  - Uniform rate. Overall percent identity.
- Transitions and transversions
  - Two-parameter model. A-G, C-T more frequent.
- Synonymous and non-synonymous substitutions
  - Ka/Ks rates. Amino-acid changing substitutions



GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG 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	

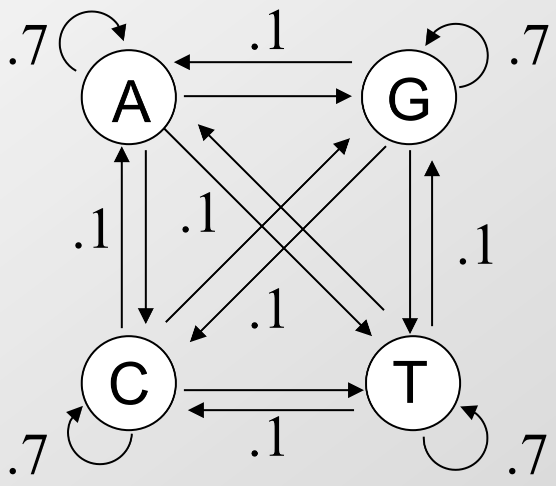
- $N_{\text{actual mutations}} > N_{\text{observed substitutions}}$
  - Some fraction of “conserved” positions mutated twice





‘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)

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

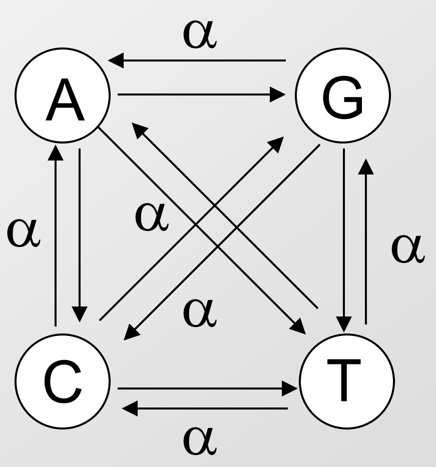
That is, $P(x \mid y, t+t') = \Sigma_z P(x \mid z, t)\, P(z \mid y, t')$

Jukes-Cantor: constant rate of evolution

For short time ε, $S(\varepsilon) = \begin{pmatrix} 1 - 3\alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & 1 - 3\alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & \alpha\varepsilon & 1 - 3\alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon & 1 - 3\alpha\varepsilon \end{pmatrix}$

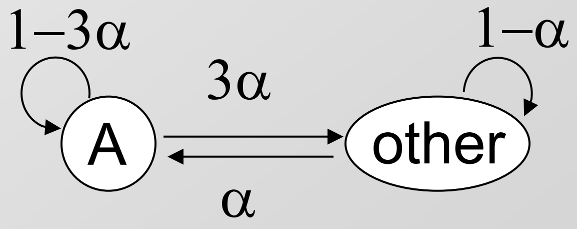
Modeling Nucleotide Evolution

Jukes-Cantor:



For longer times,

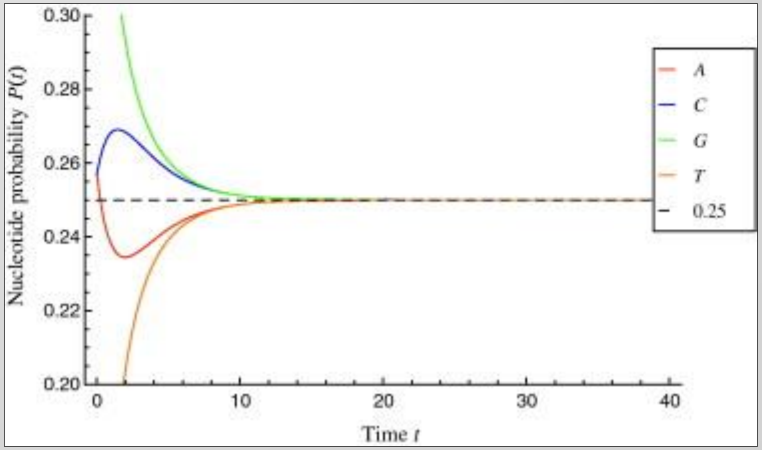
$$S(t) = \begin{pmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{pmatrix}$$



Where we can derive:

$$r(t) = ¼\,(1 + 3\,e^{-4\alpha t})$$

$$s(t) = ¼\,(1 – e^{-4\alpha t})$$



Geometric asymptote to 1/4

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)

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

Where

$$s(t) = \tfrac{1}{4}\,(1 - e^{-4\beta t})$$
$$u(t) = \tfrac{1}{4}\,(1 + e^{-4\beta t} - e^{-2(\alpha+\beta)t})$$
$$r(t) = 1 - 2s(t) - u(t)$$

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):

$d_{ij} = - ¾ \log(1 – 4f / 3)$

$r(t) = ¼ (1 + 3 e^{-4\alpha t})$

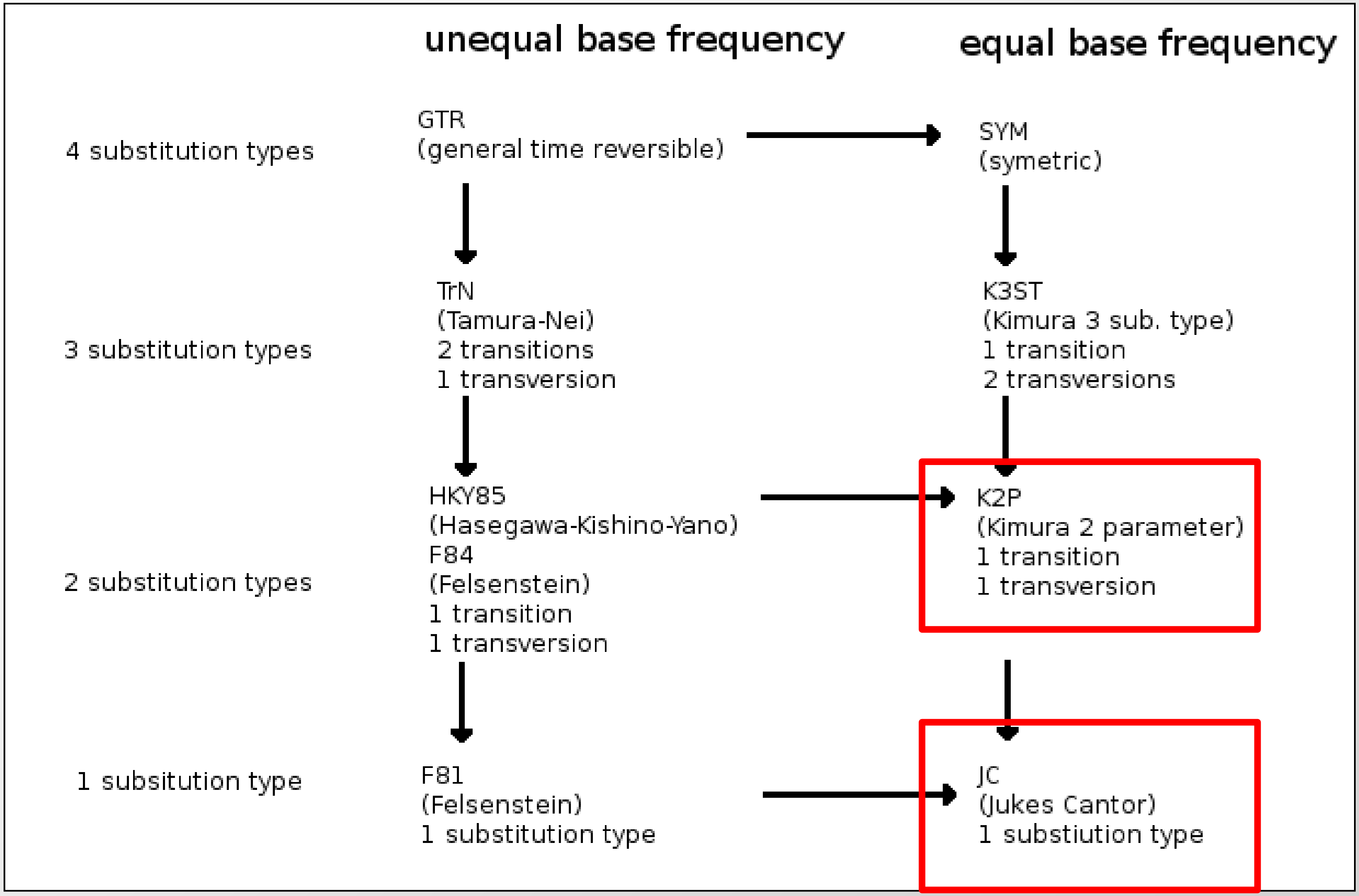
$s(t) = ¼ (1 – e^{-4\alpha t})$

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]

Many nucleotide models have been developed

Varying levels of complexity (parameters)



	unequal base frequency	equal base frequency
4 substitution types	GTR (general time reversible)	SYM (symetric)
3 substitution types	TrN (Tamura-Nei) 2 transitions 1 transversion	K3ST (Kimura 3 sub. type) 1 transition 2 transversions
2 substitution types	HKY85 (Hasegawa-Kishino-Yano) F84 (Felsenstein) 1 transition 1 transversion	K2P (Kimura 2 parameter) 1 transition 1 transversion
1 subsitution type	F81 (Felsenstein) 1 substitution type	JC (Jukes Cantor) 1 substiution type

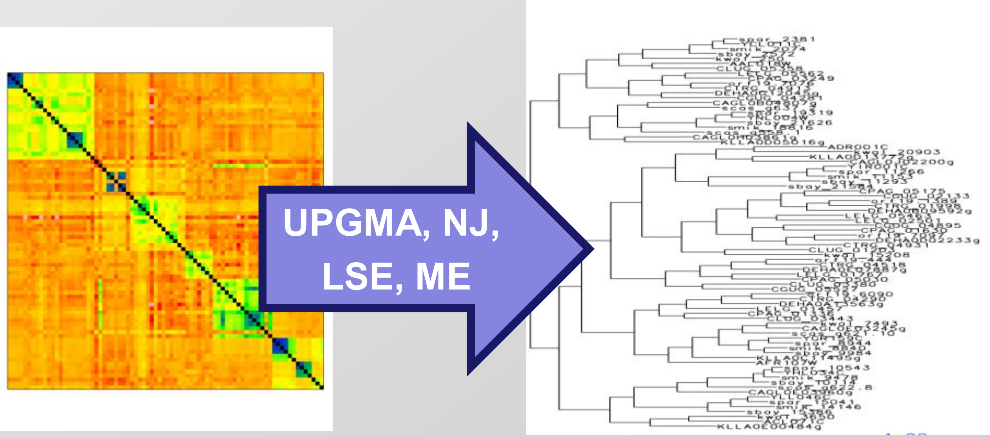
Models also exist for peptides and codons

Goals for today: Phylogenetics

- **Basics of phylogeny: Introduction and definitions**
  - Characters, traits, nodes, branches, lineages, topology, lengths
  - Gene trees, species trees, cladograms, chronograms, phylograms
1. **From alignments to distances: Modeling sequence evolution**
   - Turning pairwise sequence alignment data into pairwise distances
   - Probabilistic models of divergence: Jukes Cantor/Kimura/hierarchy
2. **From distances to trees: Tree-building algorithms**
   - Tree types: Ultrametric, Additive, General Distances
   - Algorithms: UPGMA, Neighbor Joining, guarantees and limitations
   - Optimality: Least-squared error, minimum evolution (require search)
3. **From alignments to trees: Alignment scoring given a tree**
   - Parsimony: greedy (union/intersection) vs. DP (summing cost)
   - ML/MAP (includes back-mutations, lengths): peeling algorithm (DP)
4. Tree of Life in Genomic Era
   - The prokaryotic problem (no real taxa and HGT)
   - Interpreting the forest of life

2. Distance-based tree-building algorithms

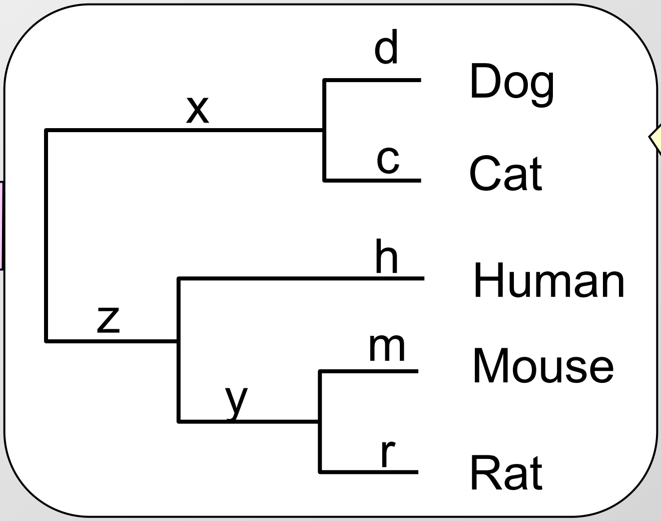
Mapping a distance matrix to a tree



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

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

Goal:

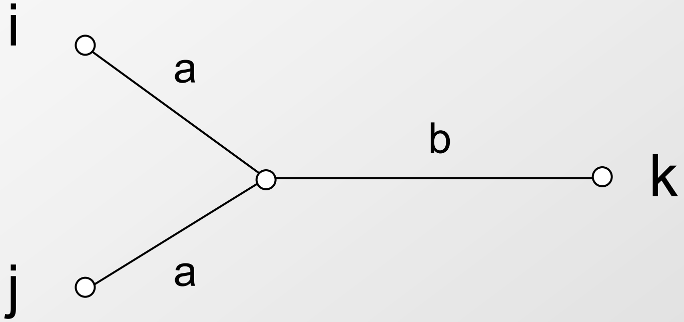
Minimize discrepancy between observed distances and tree-based distances

Ultrametric distances & 3 Point Condition

- 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 leaves are equidistant to the root
  - Rooted tree with uniform rates of evolution

Ultrametric trees

	A	B	C
A	0	6	6
B	6	0	4
C	6	4	0

	A	B	C
A	0	3	3
B	3	0	2
C	3	2	0

For now imagine that these are just the number of substitutions between pairs:

Symmetric 0-diagonal matrix of divergence times

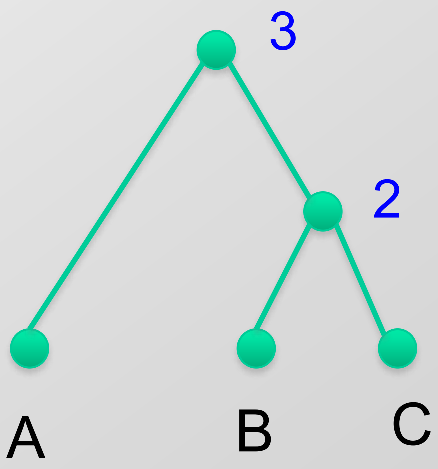
```
A:     GCCCAACTA

B:     GTTTCCCTC
```

Taken from Ran Libeskind-Hadas, Lecture Slides, Fall, 2013

Ultrametric Trees

- Given a symmetric n x n 0-diagonal matrix D, an ultrametric tree T for that matrix is one in which:
  - There are n leaves, one for each row (column) of D
  - Each internal node is labeled by a time in D and has exactly two children
  - Along any path from the root to a leaf, the (divergence) times at the internal nodes strictly decrease
  - For any two leaves i, j of T, the LCA of i, j is labeled with time D(i, j)

	A	B	C
A	0	3	3
B	3	0	2
C	3	2	0



Taken from Ran Libeskind-Hadas, Lecture Slides, Fall, 2013

Ultrametric Matrix Construction

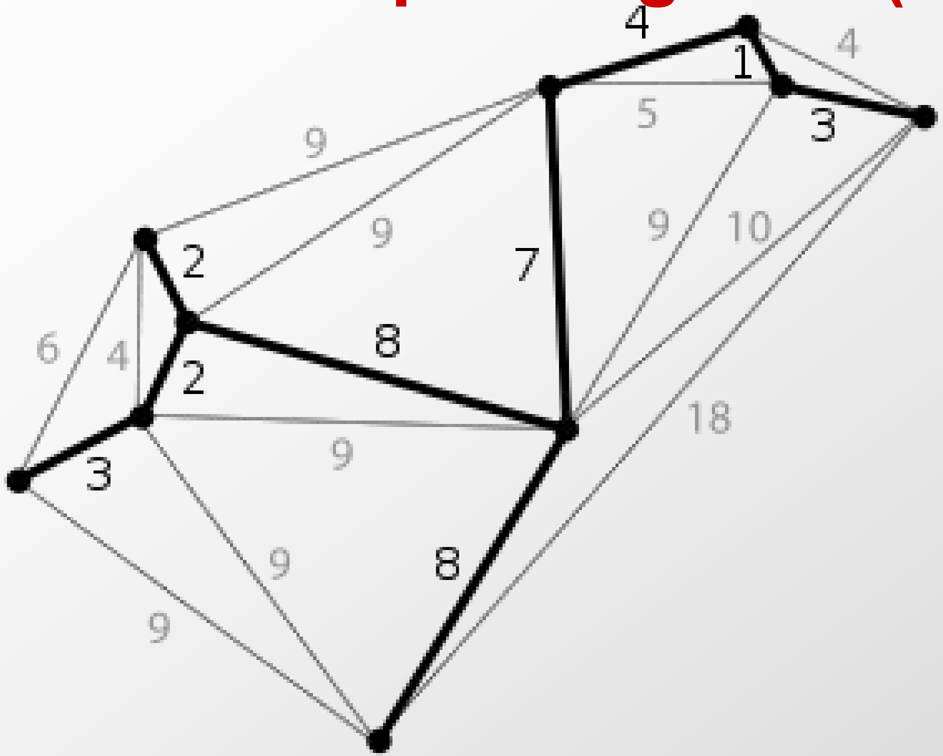
$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 5 & 2 & 5 & 7 \\ B & 5 & 0 & 5 & 3 & 7 \\ C & 2 & 5 & 0 & 5 & 7 \\ D & 5 & 3 & 5 & 0 & 7 \\ E & 7 & 7 & 7 & 7 & 0 \end{array}$$

$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 5 & 2 & 5 & 7 \\ B & 5 & 0 & 4 & 3 & 7 \\ C & 2 & 4 & 0 & 5 & 7 \\ D & 5 & 3 & 5 & 0 & 7 \\ E & 7 & 7 & 7 & 7 & 0 \end{array}$$

- Algorithms exist for “ultrametrifying” matrices.

Taken from Ran Libeskind-Hadas, Lecture Slides, Fall, 2013

Minimum Spanning Tree (MST)



- There is a unique path between any two vertices in a spanning tree
- Adding an edge to a spanning tree creates a cycle
- Any edge on that cycle can be removed and we'll still have a spanning tree
- MST is found using Prim's Algorithm (graph traversal)

Taken from Ran Libeskind-Hadas, Lecture Slides, Fall, 2013

The “Ultrametrification” Algorithm

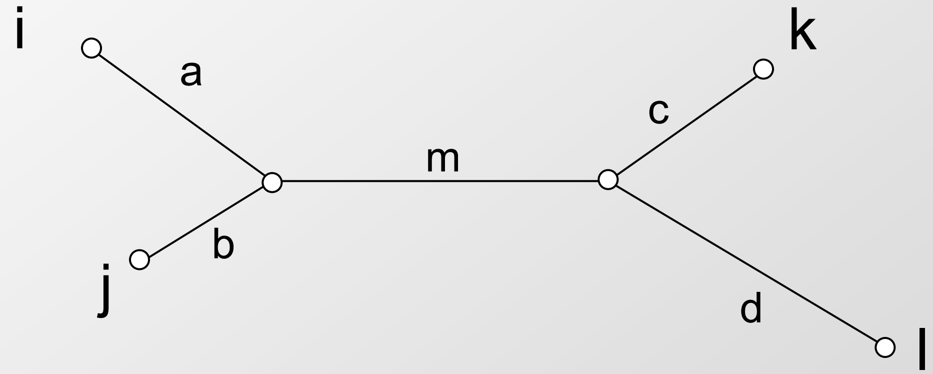
Given n x n symmetric 0-diagonal matrix D that is not ultrametric

1. Construct a completely connected graph with n vertices, one per row of A. The edge weight from vertex i to vertex j is D(i, j).
2. Find a minimum spanning tree (MST) of this graph.
3. Build a new matrix D’ such that D’(i, j) is the ***largest*** weight on the unique path from i to j in the MST.

Taken from Ran Libeskind-Hadas, Lecture Slides, Fall, 2013

Distances: (b) Additive distances

- All distances satisfy the four-point condition
  - Any quartet can be labeled i,j,k,l such that:
    - 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

Distances: (c) 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 must handle noisy distance matrices
  - 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)

Algorithms: (a) 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,

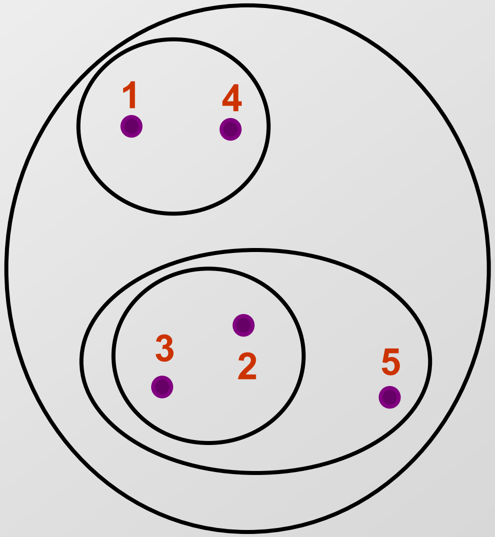
    & place it at height $d_{ij}/2$

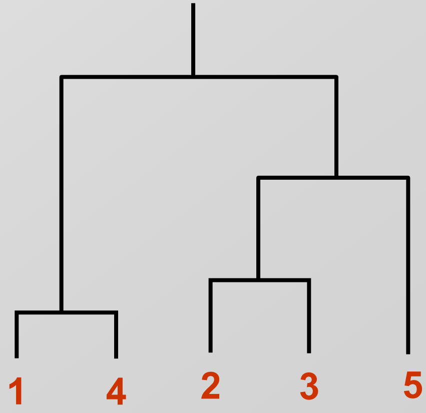
Delete C_i, C_j

Termination:

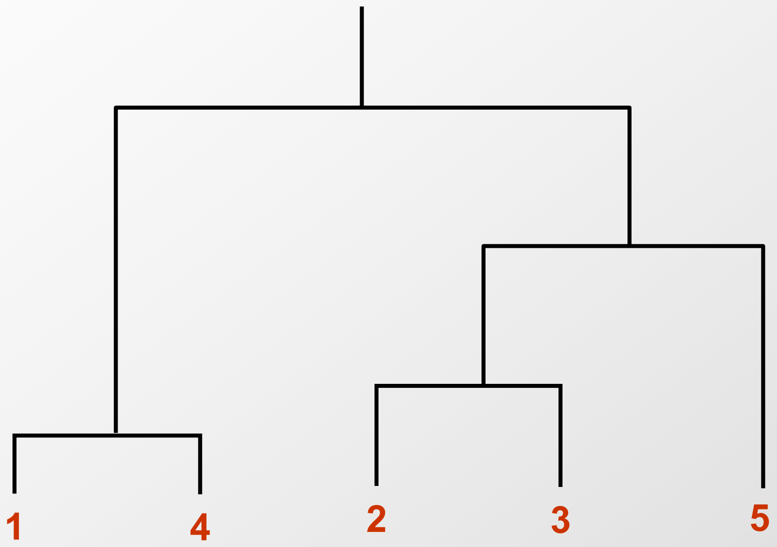
When two clusters i, j remain,

    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

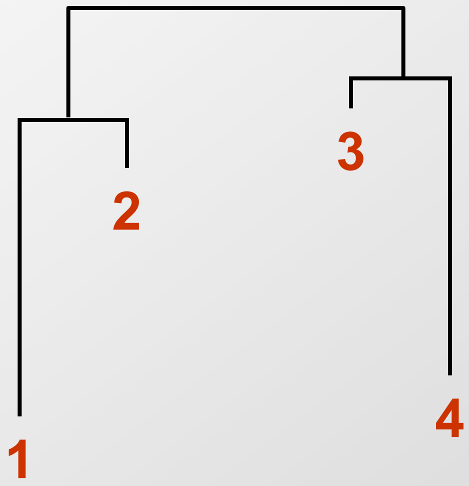
Molecular clock assumption:

implies time is constant for all species

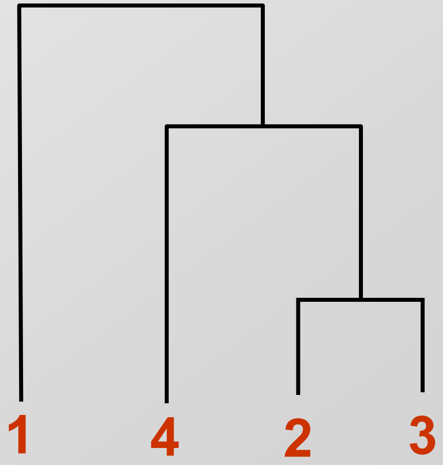
However, certain species (e.g., mouse, rat) evolve much faster

Example where UPGMA messes up:

Correct tree



UPGMA



Algorithms: (b) Neighbor-Joining

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

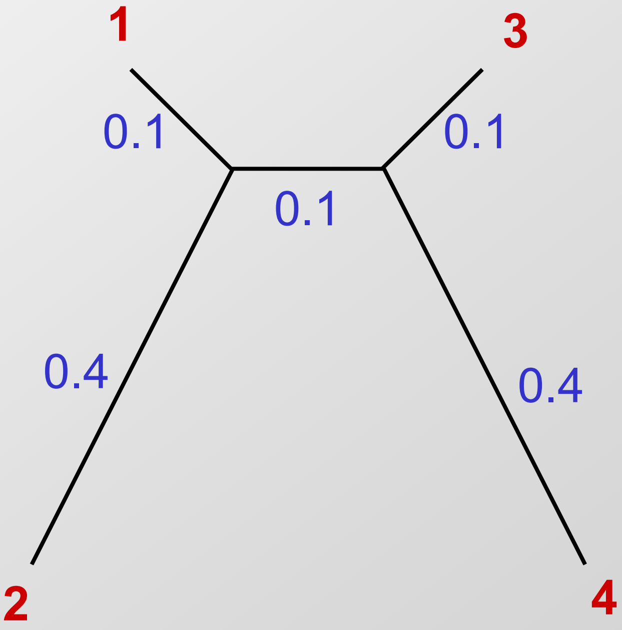
Step 1: Finding neighboring leaves

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}

Algorithms: (c) Distance-fitting algorithms

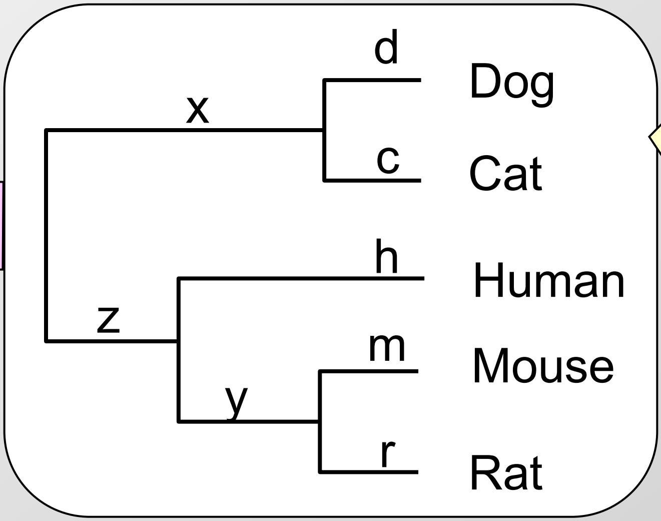
- With distance-based algorithms, we can also aim to directly minimize discrepancy between original distance matrix and tree-based distance matrix

		COMPUTATIONAL METHOD	
		Optimality criterion	Clustering algorithm
DATA TYPE	Characters	PARSIMONY MAXIMUM LIKELIHOOD	
	Distances	MINIMUM EVOLUTION LEAST SQUARES	UPGMA NEIGHBOR-JOINING

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

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

Goal:

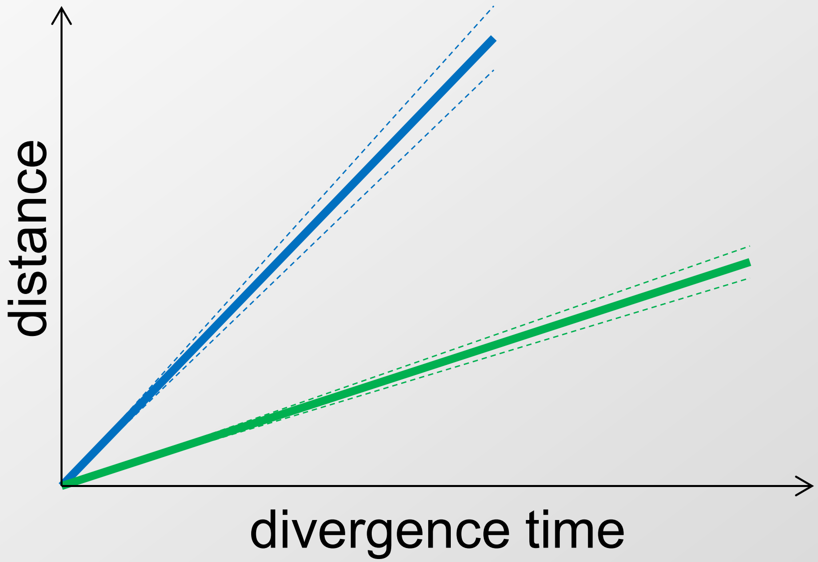
Minimize discrepancy between observed distances and tree-based distances

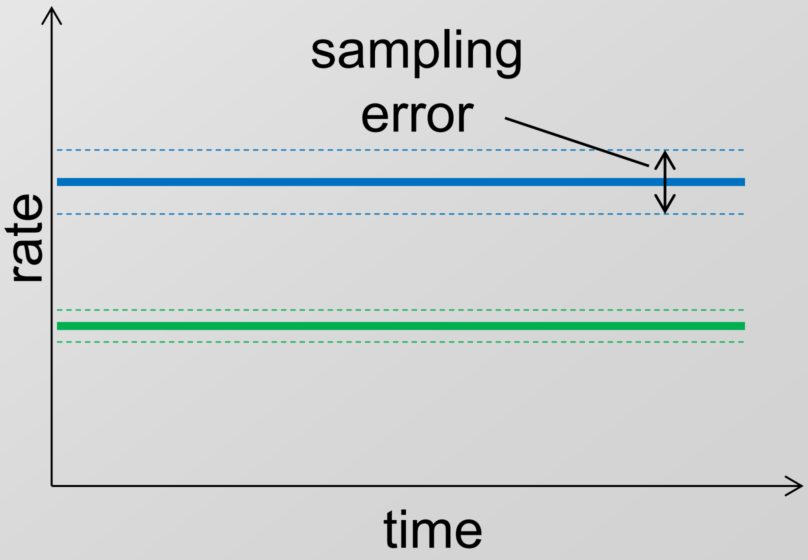
Aside: Alternative to Molecular clock?

Divergence between orthologous sequences is proportional to time separating the species.

Different genes evolve at specific, roughly constant rates.

Zuckerkandl & Pauling 1962





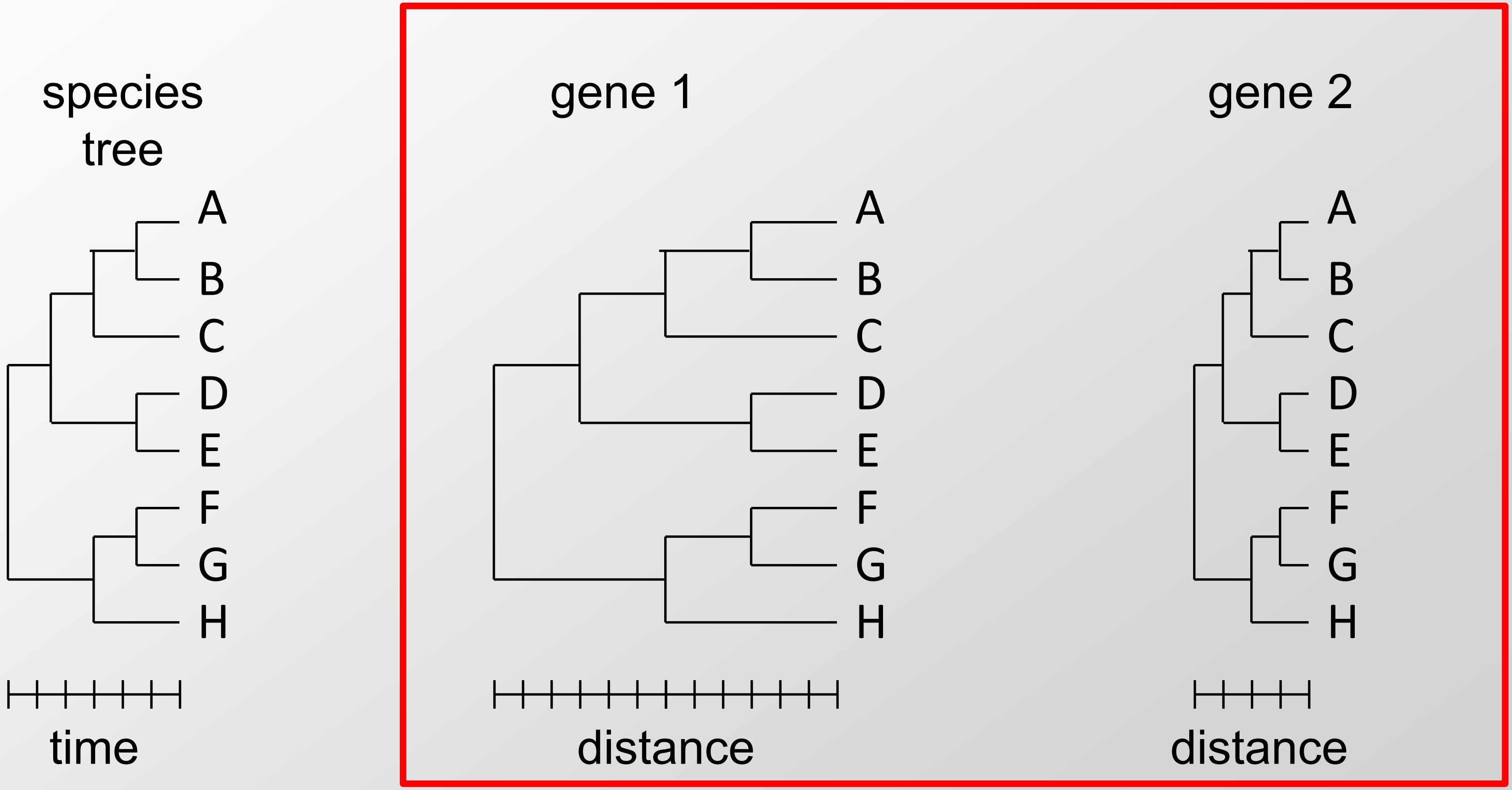
Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Molecular Clock

Under MC all individual gene trees are ultrametric (up to a sampling error) and identical to the species tree up to a scaling factor (evolution rate).

Are these really ultrametric?



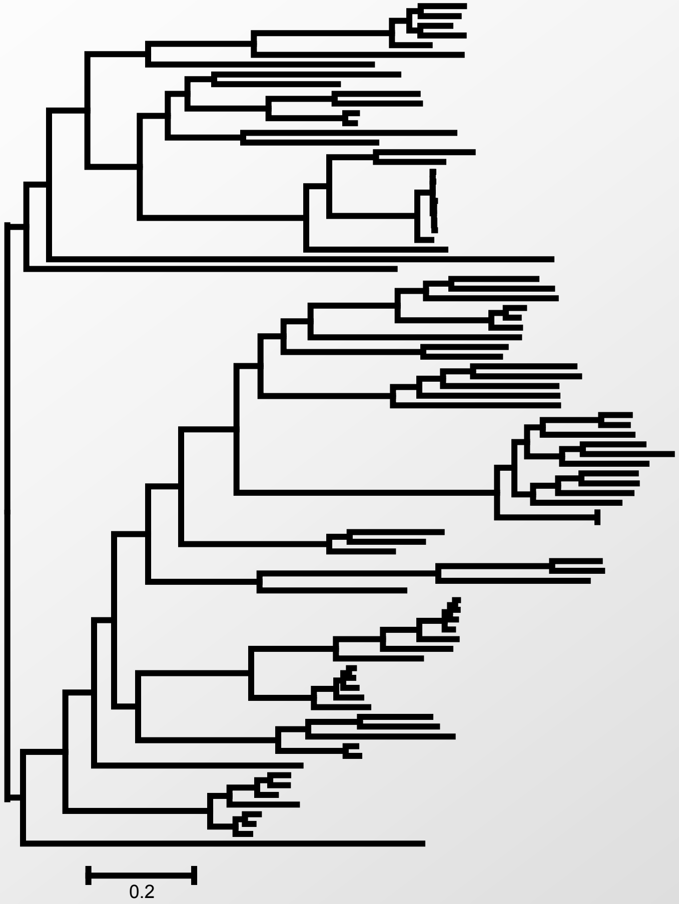
Courtesy of Yuri Wolf; slide in the public domain.

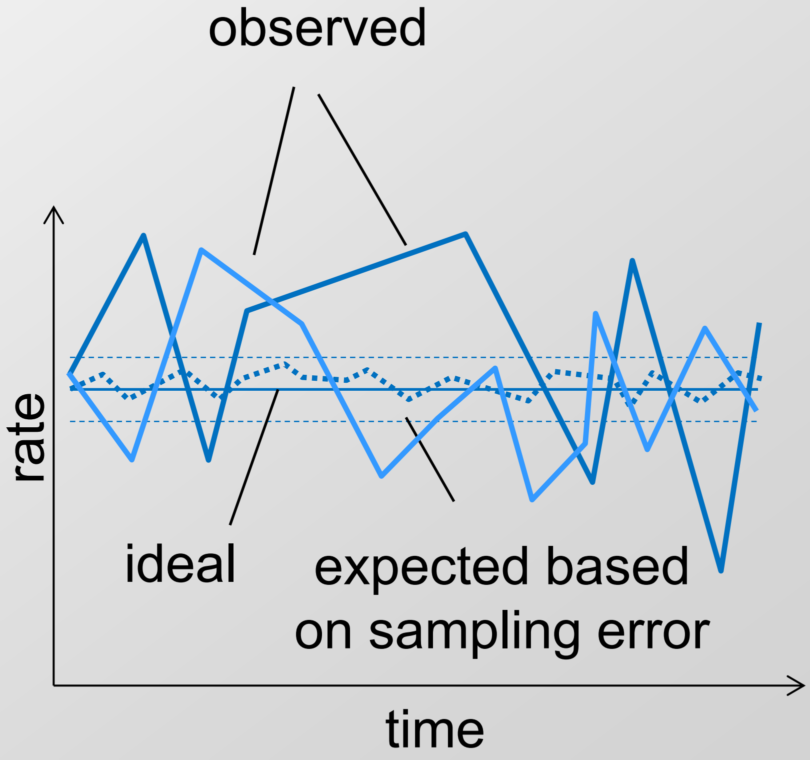
Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Molecular Clock

Most of the real phylogenetic trees are far from being ultrametric.

Molecular clock is substantially overdispersed**.**





Courtesy of Yuri Wolf; slide in the public domain.

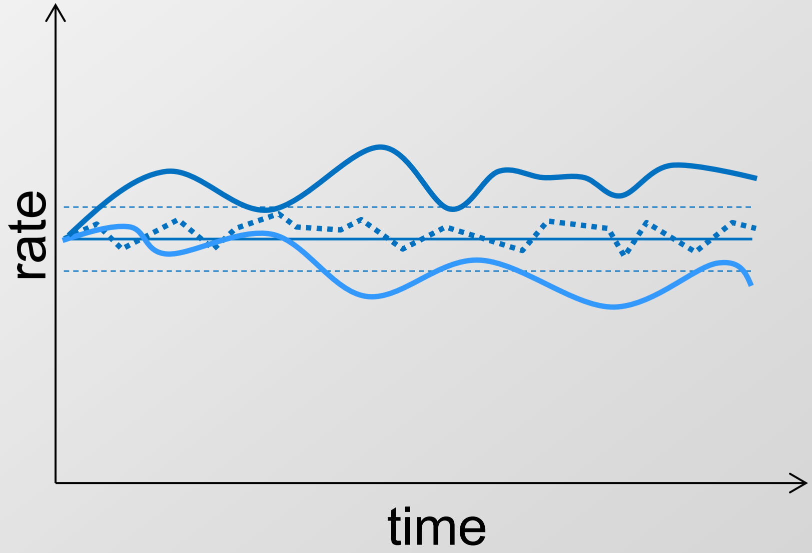
Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Relaxed Molecular Clock

Relaxed molecular clock models allows for rate variation.

Rates are sampled from prior distributions with limited variance, independently or in autocorrelated manner.

Genes are either analyzed individually, or as concatenated alignments (implying evolution as a single unit).



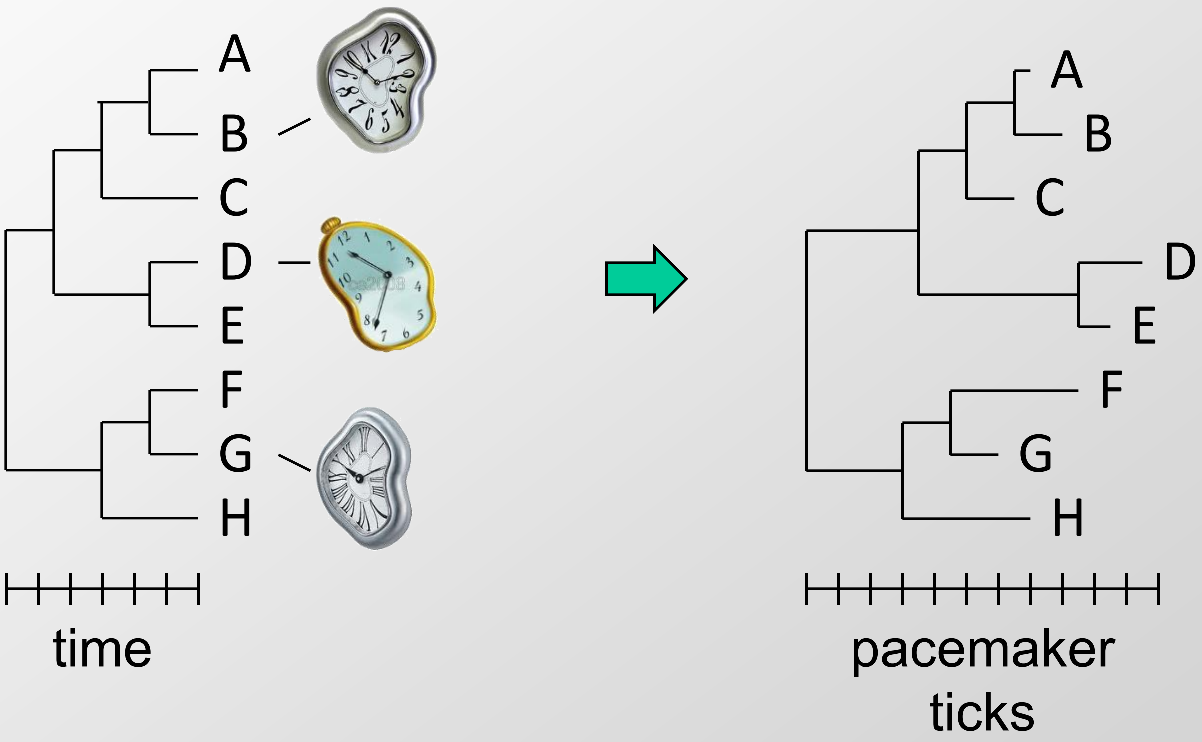
Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Universal Pacemaker

Universal Pacemaker model assumes that evolutionary time runs at different pace in each lineage.

Under the UPM, species trees are intrinsically non-ultrametric.



Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Pacemaker vs Clock

Both overdispersed MC and UPM models predict that individual gene trees would deviate from ultrametricity.

Under MC these deviations are expected to be uncorrelated.

Under UPM these deviations are expected to be correlated, so there exists a non-ultrametric pacemaker tree that can significantly reduce variance of observed rates.

A testable hypothesis!

2,300 trees of 100 prokaryotic species;

7,000 trees of 6 *Drosophila* species

1,000 trees of 9 yeast species

5,700 trees of 8 mammalian species

Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Pacemaker vs Clock

2,300 trees of 100 prokaryotic species;

7,000 trees of 6 *Drosophila* species

1,000 trees of 9 yeast species

5,700 trees of 8 mammalian species

All show an overwhelming support to UPM model.

Snir 2012; work in progress at NCBI (NIH)

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Goals for today: Phylogenetics

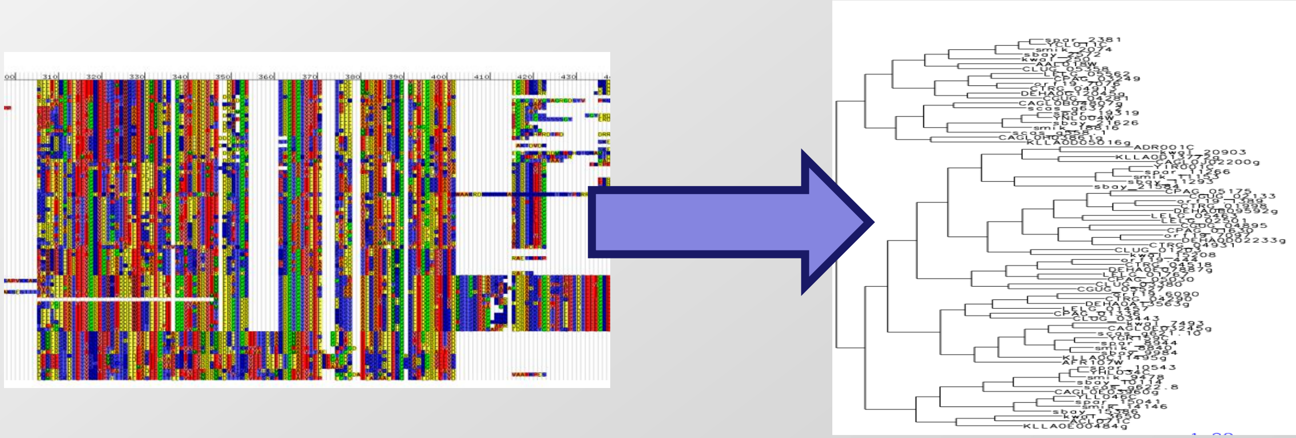
- **Basics of phylogeny: Introduction and definitions**
  - Characters, traits, nodes, branches, lineages, topology, lengths
  - Gene trees, species trees, cladograms, chronograms, phylograms
1. **From alignments to distances: Modeling sequence evolution**
   - Turning pairwise sequence alignment data into pairwise distances
   - Probabilistic models of divergence: Jukes Cantor/Kimura/hierarchy
2. **From distances to trees: Tree-building algorithms**
   - Tree types: Ultrametric, Additive, General Distances
   - Algorithms: UPGMA, Neighbor Joining, guarantees and limitations
   - Optimality: Least-squared error, minimum evolution (require search)
3. **From alignments to trees: Alignment scoring given a tree**
   - Parsimony: greedy (union/intersection) vs. DP (summing cost)
   - ML/MAP (includes back-mutations, lengths): peeling algorithm (DP)
4. Tree of Life in Genomic Era
   - The prokaryotic problem (no real taxa and HGT)
   - Interpreting the forest of life

3. Character-based tree-scoring algorithms

3a: Parsimony (set-based)

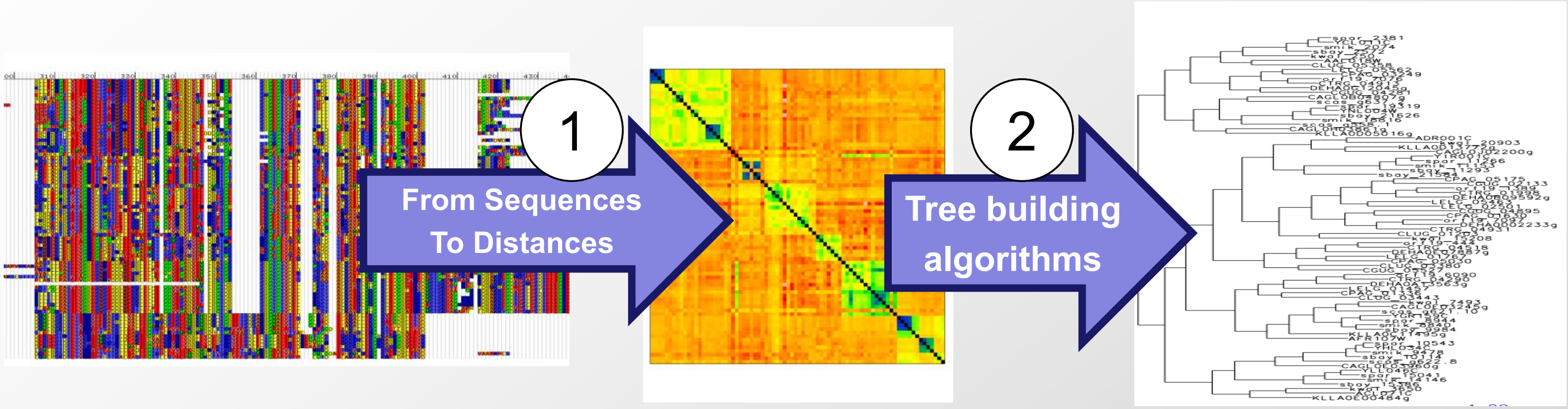
3b: Parsimony (Dyn. Prog.)

3c: Maximum Likelihood



Basic algorithms of phylogenetic methods

Distance based

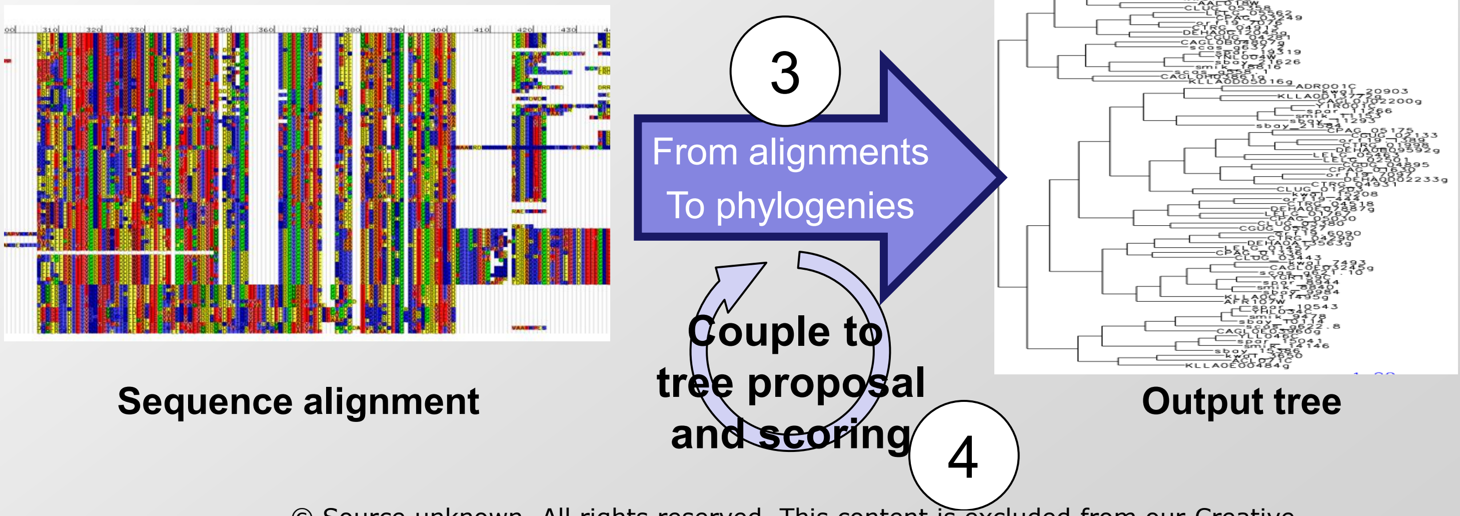


Sequence alignment

Pair-wise distance matrix

Output tree

Character based



Sequence alignment

Output tree

Character-based phylogenetic inference

- Really about tree **scoring** techniques, not tree finding techniques
  - Couple them with tree proposal and update and you have an algorithm (part 4 of the lecture)
- Two approaches exist, all use same architecture:
  - Minimize events: Parsimony (union/intersection)
  - Probabilistic: Max Likelihood / MAP

Parsimony scoring (a): Union and intersection

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,

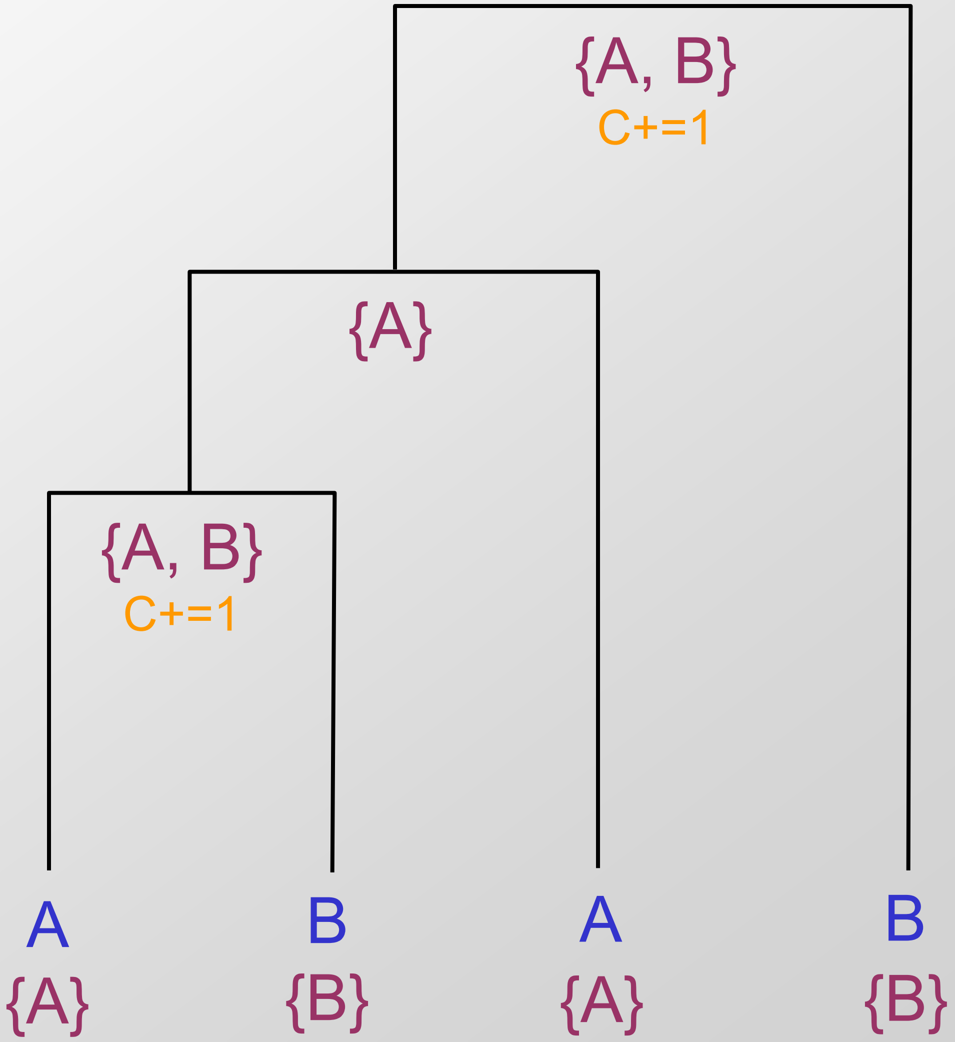
Let i, j be the daughter nodes;

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



Parsimony traceback to find ancestral nucleotides

Traceback:

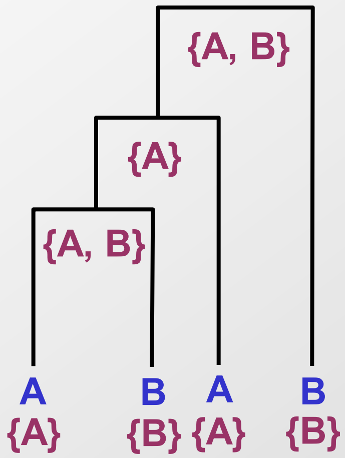
1. Choose an arbitrary nucleotide from R_{2N-1} for the root

2. Having chosen nucleotide r for parent k,

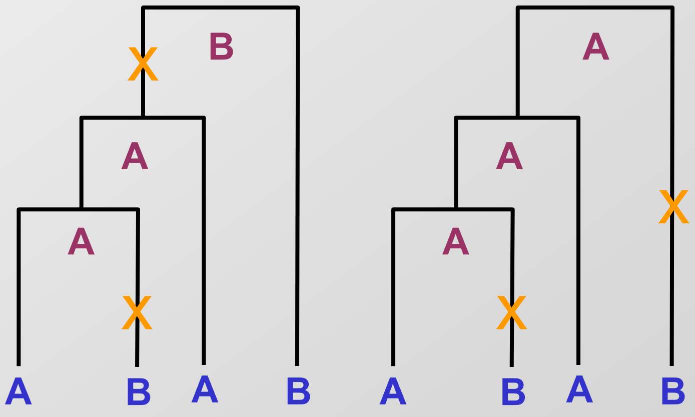
   If $r \in R_i$ choose r for daughter i

   Else, choose arbitrary nucleotide from R_i

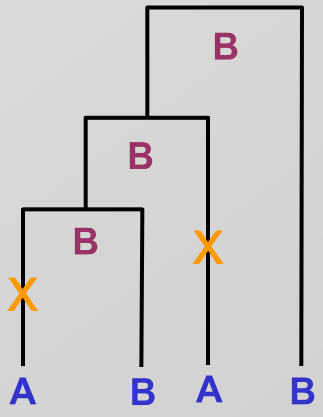
Easy to see that this traceback produces some assignment of cost C



Accessible to traceback

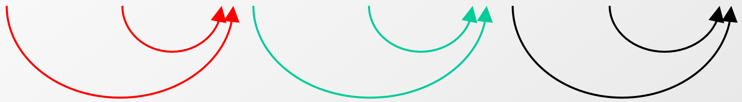


Still optimal, but
not found by traceback

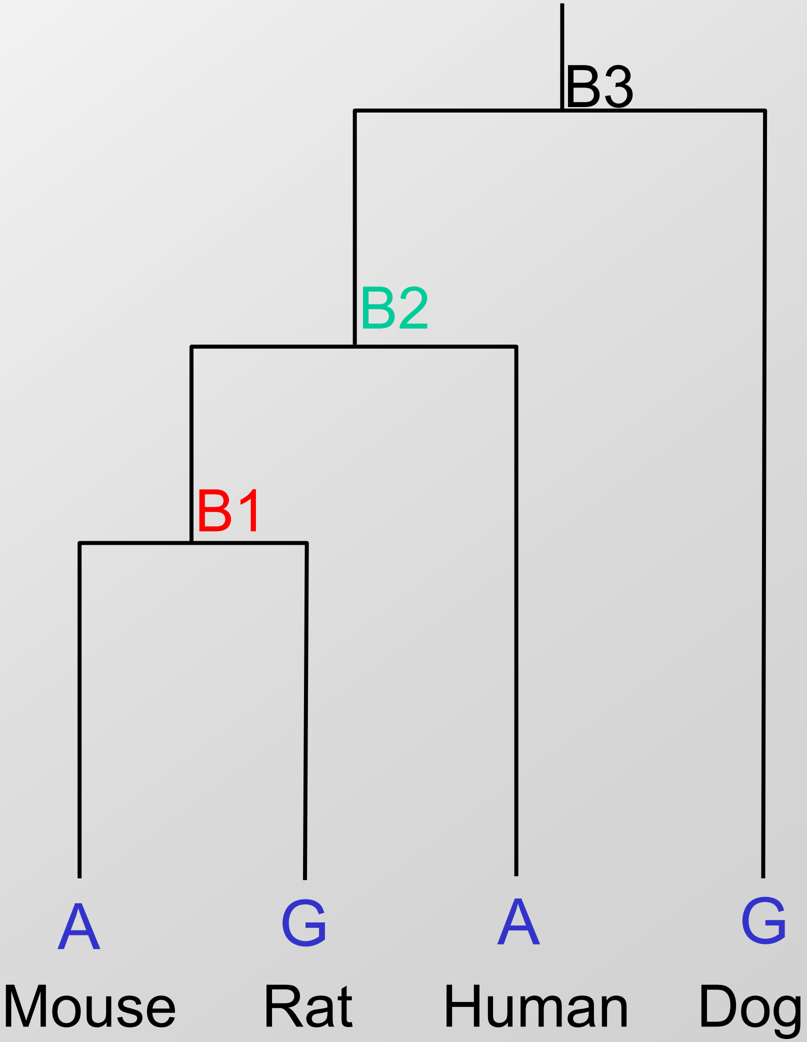


Parsimony Scoring (b): Dynamic programming

	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



Goals for today: Phylogenetics

- **Basics of phylogeny: Introduction and definitions**
  - Characters, traits, nodes, branches, lineages, topology, lengths
  - Gene trees, species trees, cladograms, chronograms, phylograms

1. **From alignments to distances: Modeling sequence evolution**
   - Turning pairwise sequence alignment data into pairwise distances
   - Probabilistic models of divergence: Jukes Cantor/Kimura/hierarchy
2. **From distances to trees: Tree-building algorithms**
   - Tree types: Ultrametric, Additive, General Distances
   - Algorithms: UPGMA, Neighbor Joining, guarantees and limitations
   - Optimality: Least-squared error, minimum evolution (require search)
3. **From alignments to trees: Alignment scoring given a tree**
   - Parsimony: greedy (union/intersection) vs. DP (summing cost)
   - ML/MAP (includes back-mutations, lengths): peeling algorithm (DP)
4. Tree of Life in Genomic Era
   - The prokaryotic problem (no real taxa and HGT)
   - Interpreting the forest of life

Scoring (c) Maximum Likelihood & Max-a-Posteriori

Input: Sequence alignment

Output: tree with maximum likelihood / max a posteriori prob.

Search: Heuristic search for max likelihood tree.

Maximum Likelihood (ML)

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

D = seq. alignment data

B = branch lengths

T = topology

Maximum a Posteriori (MAP)

$$\begin{aligned} B\hat{},T\hat{} &= \text{argmax}_{B,T}\ P(B,T|D) \\ &= \text{argmax}_{B,T}\ P(B,T,D)\ /\ P(D) \\ &= \text{argmax}_{B,T}\ P(B,T,D) \\ &= \text{argmax}_{B,T}\ P(D|B,T)P(B,T) \end{aligned}$$

likelihood likelihood

P(D|B,T) is the likelihood of data given model

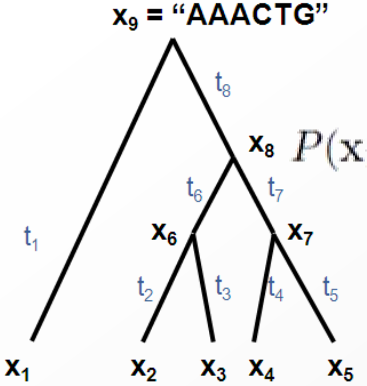
➔ Use seq evolution model: JC,K2P,HKY.

Compute recursively using DP

P(B,T) is a prior on trees/branch lengths

➔ Use Yule process, Birth-Death process to model

‘Peeling’ algorithm for P(D|B,T) term



$$\begin{aligned} P(\mathrm{x}_1, \ldots, \mathrm{x}_{2n-1} | T, \mathrm{t}) &= P(\mathrm{x}_1 | \mathrm{x}_2, \ldots, \mathrm{x}_{2n-1}, T, \mathrm{t}) P(\mathrm{x}_2 | \mathrm{x}_3, \ldots, \mathrm{x}_{2n-1}, T, \mathrm{t}) \ldots P(\mathrm{x}_{2n-1} | T, \mathrm{t}) \\ &= P(\mathrm{x}_1 | \mathrm{x}_{\mathrm{parent}(1)}, t_1) P(\mathrm{x}_2 | \mathrm{x}_{\mathrm{parent}(2)}, t_2) \ldots P(\mathrm{x}_{2n-1}) \\ &= P(\mathrm{x}_{2n-1}) \prod_{i=1}^{2n-2} P(\mathrm{x}_i | \mathrm{x}_{\mathrm{parent}(i)}, t_i) \end{aligned}$$

1. Assume **sites j evolve independently**.

   ➔ Treat each column of the alignment in isolation

2. Assume **branch independence**, conditioned on parent

   ➔ Expand total joint probability into prod of $P(x_i|x_{parent},t_i)$

   ➔ Only $P(x_{2n-1})$ remains, root prior, background nucl. freq.

3. We know how to compute $\mathbf{P(x_i|x_{parent(i)},t_i)}$ for fixed pair

   ➔Defined by our sequence model (JC, K2P, HKY, etc)

   ➔Easily calculate for any given assignment of internal nodes

4. As internal node values are not known → **marginalize**

   ➔Sum over all possible values of all internal/root nodes

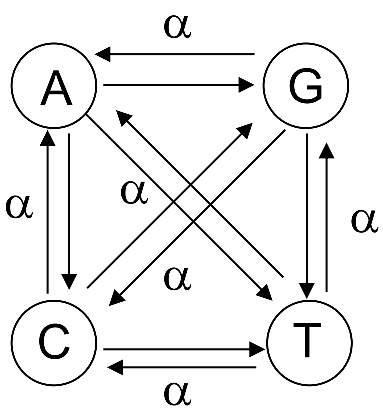
   ➔Let $x_{n+1},\ldots,x_{2n-1}$ represent seqs of n-1 internal nodes

1. Site evolution over single branch

Remember: Jukes-Cantor (JC)

JC is a Continuous-Time Markov Chain (CTMC)

- Defines instantaneous rates of transition between states (bases)

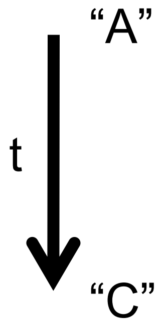


Discrete MC version

- Given time t, we define a discrete MC with transition matrix is S(t), also called a *substitution probability matrix.*
- Gives the probability of seeing base *a* given initial base *b* after duration time *t*.

$$P(a|b,t) = S(t) = \begin{pmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{pmatrix} \qquad \begin{aligned} r_t &= \frac{1}{4}(1+3e^{-4\alpha t}) \\ s_t &= \frac{1}{4}(1-e^{-4\alpha t}). \end{aligned}$$

Use JC to define **single site evolution:**

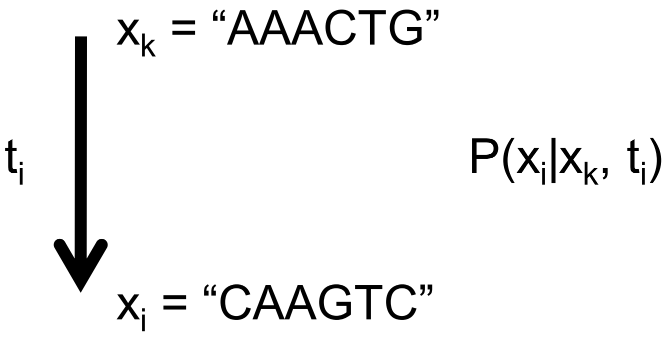


P(a="C"|b="A", t) = $S(t)_{ba}$

2. Sequence evolution over single branch

- Assume site independence
  - $P(x_i | x_k, t_i) = \Pi_j P(b=x_{ij} | a=x_{kj}, t_i)$

Use product to define **sequence evolution:**



x_k = “AAACTG”

t_i

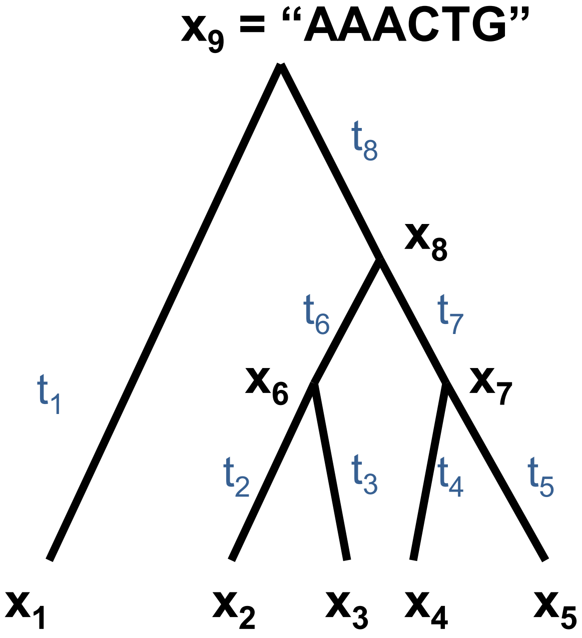
$P(x_i|x_k, t_i)$

x_i = “CAAGTC”

3. Sequence evolution over entire tree

- Assume branch independence
  - $P(x_1, \ldots x_n, \ldots, x_{2n-1} | T, \mathbf{t}) = P(x_{2n-1}) \Pi_i P(x_i | x_{parent(i)}, t_i)$
- Assume prior on root sequence, e.g.
  - $P(x_{2n-1}) = P(x_{2n-1,j}) = (1/4)^m$ for sequence length m

Use product and prior to define **sequence evolution over tree:**

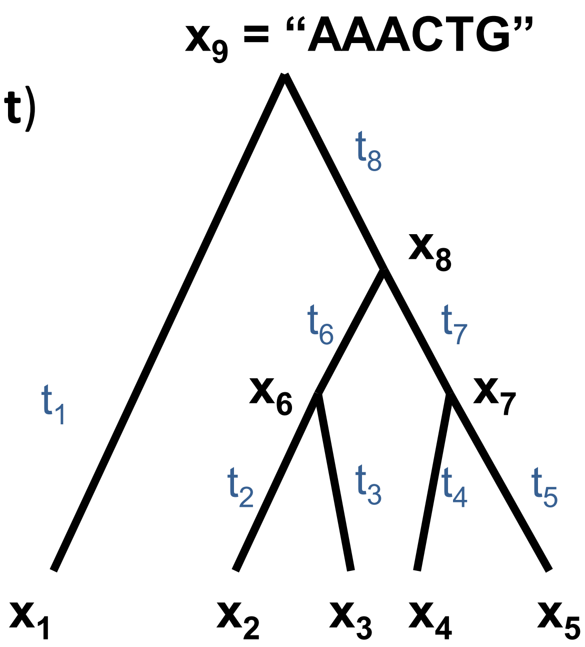


$P(x_1, \ldots x_n, \ldots, x_{2n-1}|T, \mathbf{t})$

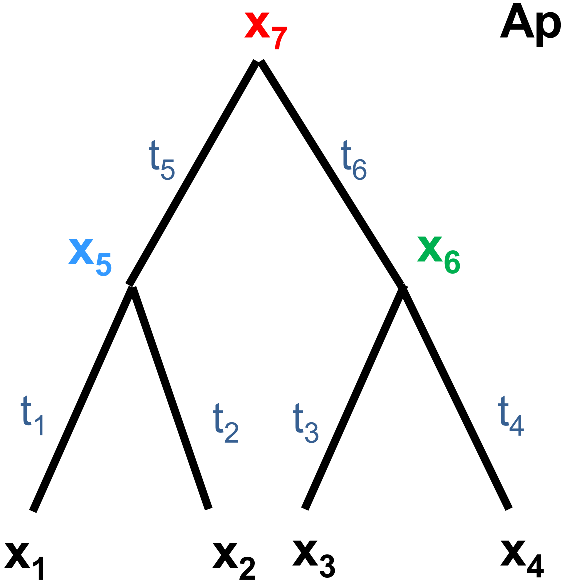
4. Integrate (marginalize) over hidden ancestral seqs!

- Notice, all sequences are needed, both internal nodes and leaves
  - $P(x_1, \ldots x_n, \ldots, x_{2n-1} | T, \mathbf{t})$
- But, only leaves are given: $x_1, \ldots x_n$
- Therefore, need to marginalize (sum) over unknowns: $x_{n+1}, \ldots, x_{2n-1}$

- This looks expensive!
  - $P(x_1, \ldots x_n | T, \mathbf{t}) = \Sigma_{x_{n+1}, \ldots,} \Sigma_{x_{2n-1}} P(x_1, \ldots x_n, \ldots, x_{2n-1} | T, \mathbf{t})$
- Don't worry, dynamic programming can do it efficiently.



Basic trick to efficient marginalization



Apply factorization trick to every internal node in the tree.

$$P(x_1,x_2,x_3,x_4|T, \mathbf{t}) = \Sigma x_5 \Sigma x_6 \Sigma x_7\, P(x_1,x_2,x_3,x_4,x_5,x_6,x_7|T, \mathbf{t})$$

$$= \Sigma_{x_5}\Sigma_{x_6}\Sigma_{x_7}\, P(x1|x5,t1)\, P(x2|x5,t1)$$

$$P(x3|x6,t3)\, P(x4|x6,t4)$$

$$P(x5|x7,t5)\, P(x6|x7,t6)\, P(x7)$$

$$= \Sigma x_7\, P(x7)$$

$$[\Sigma x_5\, P(x5|x7,t5)\, P(x1|x5,t1)\, P(x2|x5,t1)]$$

$$[\Sigma x_6\, P(x6|x7,t6)\, P(x3|x6,t3)\, P(x4|x6,t4)]$$

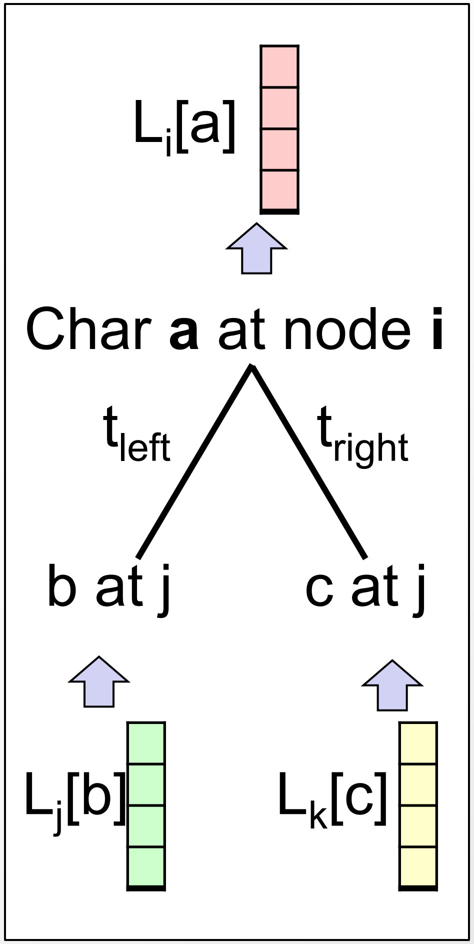
Peeling algorithm

- L(i,j,a) is the DP table.
- Each entry contains the probability of seeing the leaf data below node i, given that node i has base a at site j.
- The leaves of the table are initialized based on the observed sequence. Entries populated in post-order traversal.
- Runtime: O(2n * k^2)

$$P(x_1,\ldots,x_n|T,t) = \prod_j \sum_a L_{2n-1,j,a} P(a)$$

$$L_{i,j,a} = \begin{cases} 1 & \text{if } x_{i,j} = a, i \leq n \\ 0 & \text{if } x_{i,j} \neq a, i \leq n \\ \sum_{b,c} P(b|a,t_{left(i)}) L_{left(i),j,b} & \text{if } i > n \\ P(c|a,t_{right(i)}) L_{right(i),j,c} & \end{cases}$$

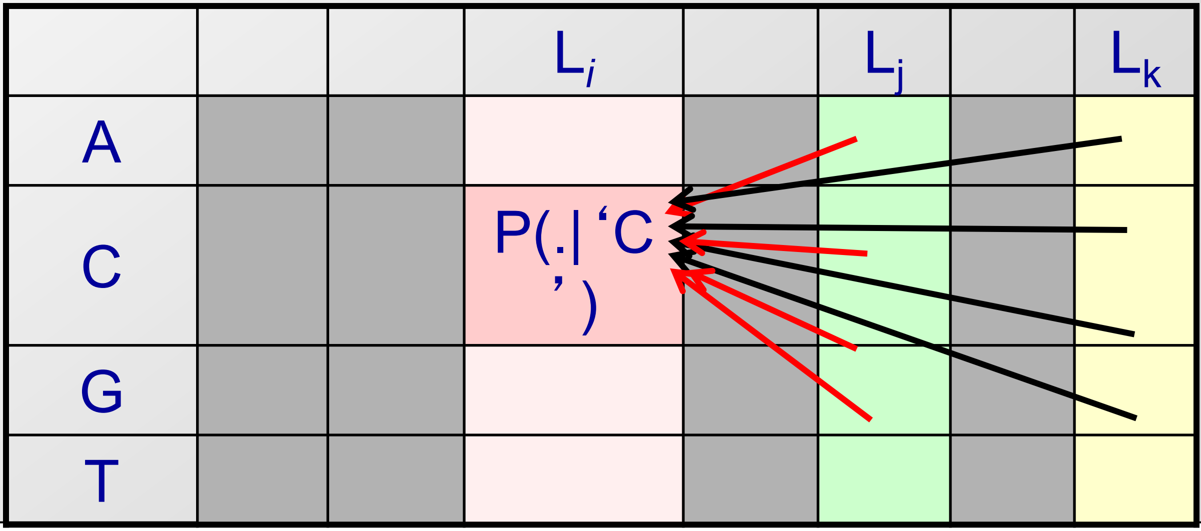
Use DP to compute argmax P(D|B,T) efficiently



- If we know the branch lengths t_{left} & t_{right}.
- And we already have the likelihood tables L_j&L_k of left and right subtrees

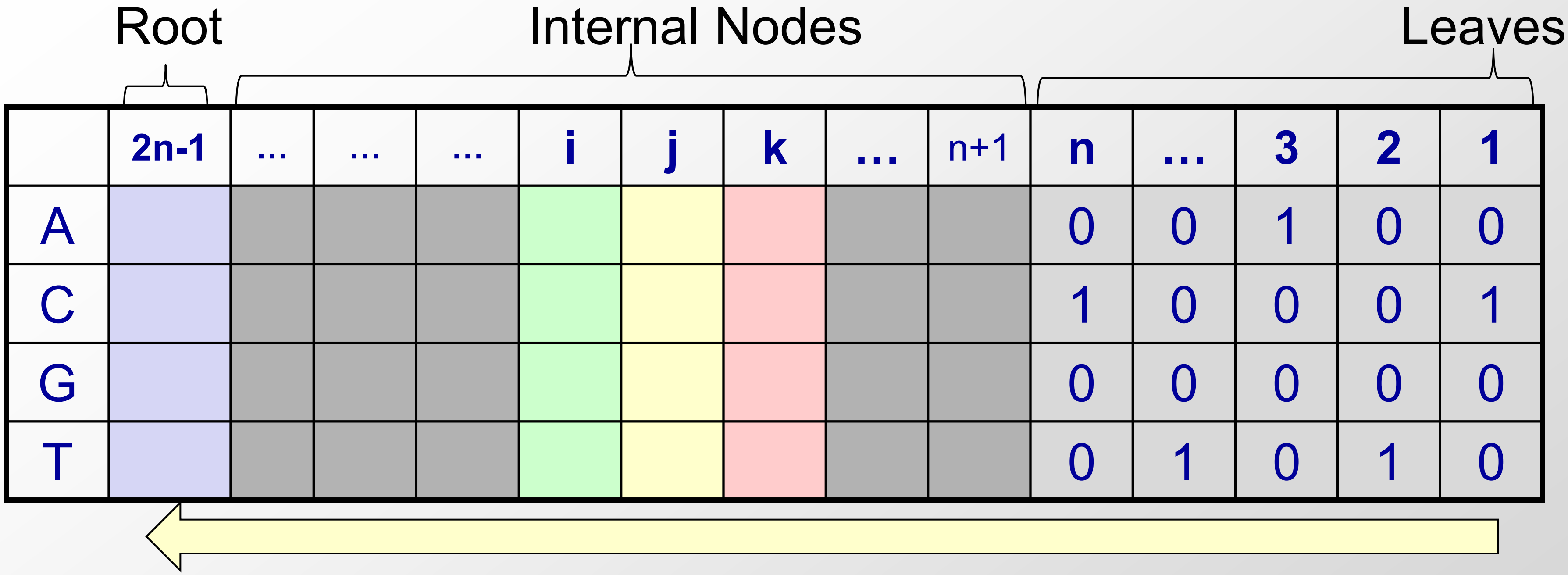
  (for each possible ending character at **b**, **c**)

➔ Fill in likelihood table L_i for each char **a** at i



$$L_i[a] = \sum_{b\in\{ACGT\}} \sum_{c\in\{ACGT\}} \left(\underbrace{\boxed{P(b|a,t_{left})*L_{left}[b]}}_{\textbf{Prob}(a\to b)} * \underbrace{\boxed{P(c|a,t_{right})*L_{right}[c]}}_{\textbf{Prob}(a\to c)} \right)$$

Initialization and Termination



Root

Internal Nodes

Leaves

	2n-1	…	…	…	i	j	k	…	n+1	n	…	3	2	1
A										0	0	1	0	0
C										1	0	0	0	1
G										0	0	0	0	0
T										0	1	0	1	0

- Characters at the leaves are already known
  - Their likelihood is 1 or 0, indicating the known char
- Fill in internal node likelihood vectors iteratively
- Once we reach the root, multiply by the base freqs
- Maximization over Topologies and Lengths

  → Numerical: gradient descent, Newton's method

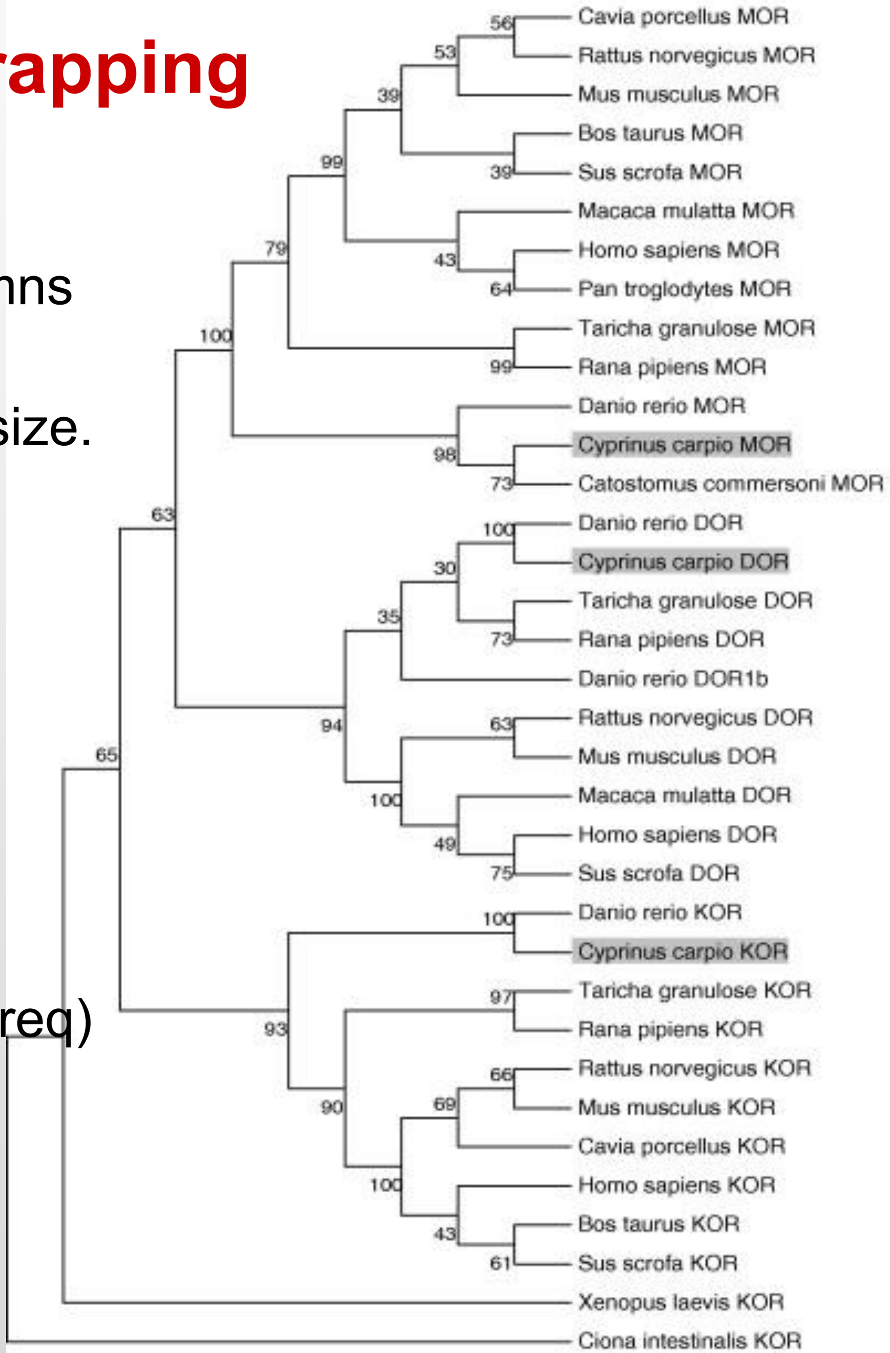
Advantages/disadvantages of ML/MAP methods

- **Advantages:**
  - Inherently statistical and evolutionary model-based.
  - Usually the most ‘consistent’ of the methods available.
  - Used for both character and rate analyses
  - Can be used to infer the sequences of the extinct ancestors.
  - Account for branch-length effects in unbalanced trees.
  - Nucleotide or amino acid sequences, other types of data.

- **Disadvantages:**
  - Not as intuitive as parsimony (e.g. may choose more events if they’re more likely in our probabilistic model)
  - Computationally intense (limits num taxa, sequence length).
  - Like parsimony, can be fooled by high levels of homoplasy.
  - Violations of model assumptions can lead to incorrect trees.

Tree reliability: Bootstrapping

1. Re-sample alignments:
   - Randomly sample alignment columns with replacement
   - Create many alignments of equal size.
2. Build a phylogenetic tree for each sample
3. Repeat (1) and (2) many times
   - 1000s of times
4. Output summary tree
   - Tree constructed most frequently
   - Consensus tree (even if not most freq)
   - Other options
5. Report observation frequency of each branch
   - Each branch is a binary split



Goals for today: Phylogenetics

- **Basics of phylogeny: Introduction and definitions**
  - Characters, traits, nodes, branches, lineages, topology, lengths
  - Gene trees, species trees, cladograms, chronograms, phylograms

1. **From alignments to distances: Modeling sequence evolution**

   1
   - Turning pairwise sequence alignment data into pairwise distances
   - Probabilistic models of divergence: Jukes Cantor/Kimura/hierarchy

2. **From distances to trees: Tree-building algorithms**

   2
   - Tree types: Ultrametric, Additive, General Distances
   - Algorithms: UPGMA, Neighbor Joining, guarantees and limitations
   - Optimality: Least-squared error, minimum evolution (require search)

3. **From alignments to trees: Alignment scoring given a tree**

   3
   - Parsimony: greedy (union/intersection) vs. DP (summing cost)
   - ML/MAP (includes back-mutations, lengths): peeling algorithm (DP)

4. Tree of Life in Genomic Era

   4
   - The prokaryotic problem (no real taxa and HGT)
   - Interpreting the forest of life

Tree of Life in Genomic Era

Genomic era – growing frustration with discrepancies between the trees reconstructed for individual genes and heroic efforts to overcome the noise. Role of horizontal gene transfer in the evolution of prokaryotic genomes is established.

Major lines of approach:

- gene repertoire and gene order
- distribution of distances between orthologs
- concatenated alignments of "non-transferable" gene cores
- consensus trees and supertrees

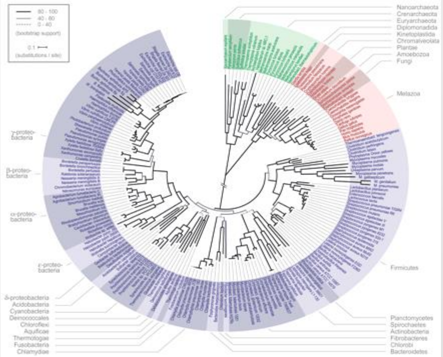


Image in the public domain.

Ciccarelli 2006. *Towards automatic reconstruction of a highly resolved tree of life*. Science 311, 1283-1287 [Figure 2]

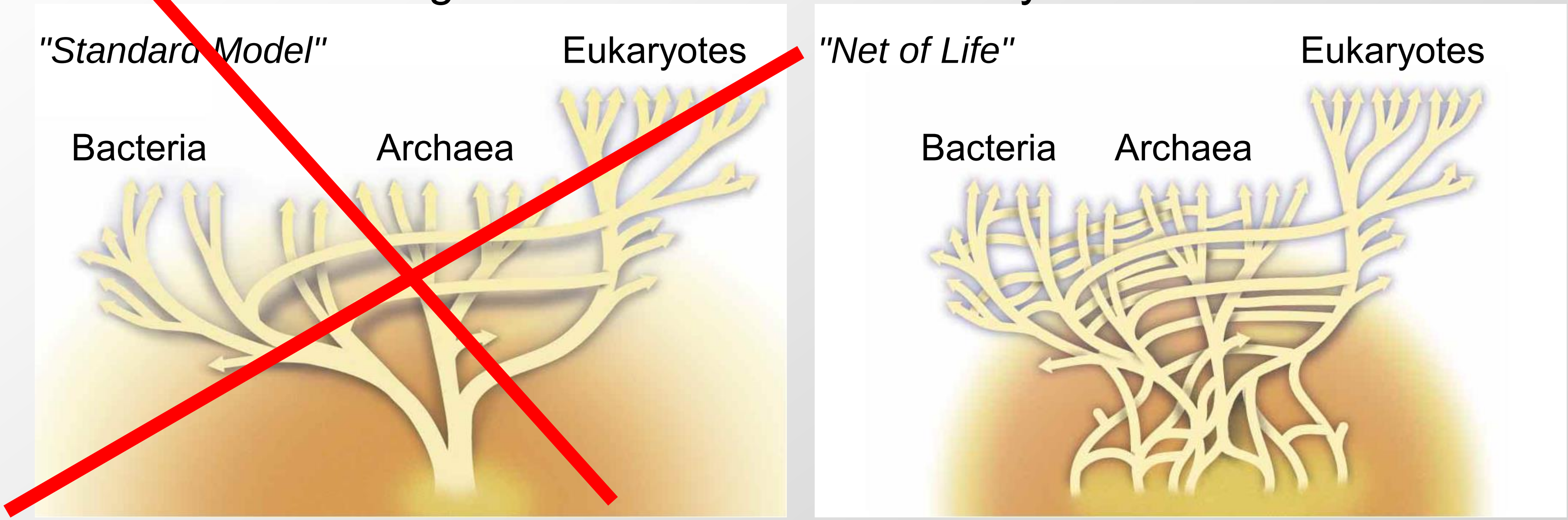
Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Tree of Life, Rejected

Troubled times – "uprooting" of TOL for prokaryotes.

- horizontal gene transfer is rampant; no gene is exempt
- histories of individual genes are non-coherent with each other
- vertical signal is completely lost (or never existed at all)
- there are no species (or other taxa) in prokaryotes
- a consistent signal we observe is created by biases in HGT



Doolittle 2000. *Uprooting the tree of life*. Sci. Am. 282, 90-95 [modified]

Source: Doolittle, W. Ford. "Uprooting the tree of life." Scientific American 282, no. 2 (2000): 90.

Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Forest of Life – Methods

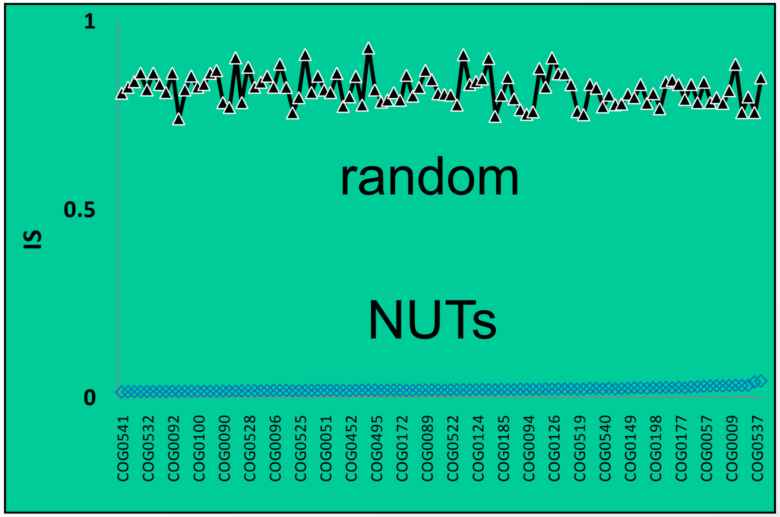
Source data and basic analysis methods:

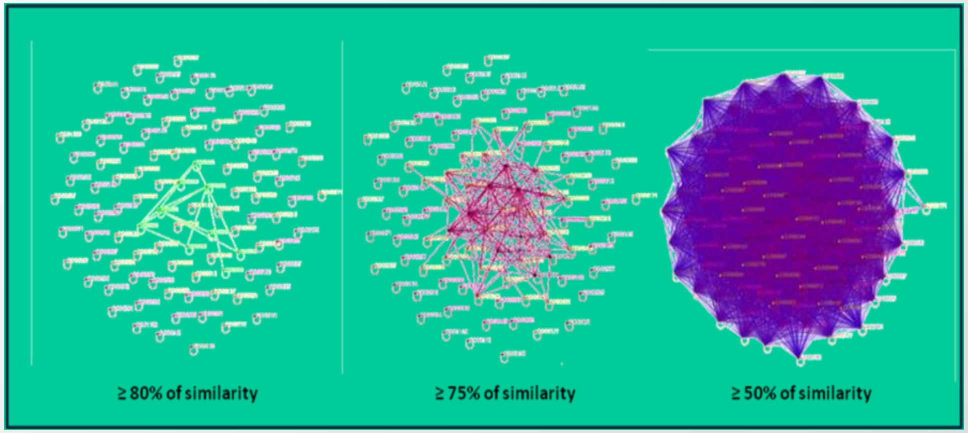
- 100 hand-picked microbial genomes (41 archaea and 59 bacteria) representing a "fair" sample of prokaryote diversity (as known in 2008)
- clusters of orthologous genes (NCBI COGs and EMBL EggNOGs)
- multiple protein sequence alignments → index orthologs → ML phylogenetic trees
- 6901 trees cover 4-100 species; of them 102 cover 90-100 species (Nearly Universal Trees)
- direct tree comparison (distances between trees)
- quartet decomposition; analysis of quartet spectra
- simulation evolutionary models

Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

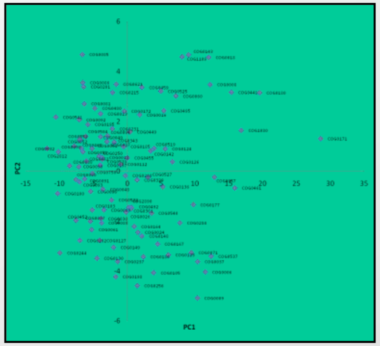
Forest of Life – Analysis

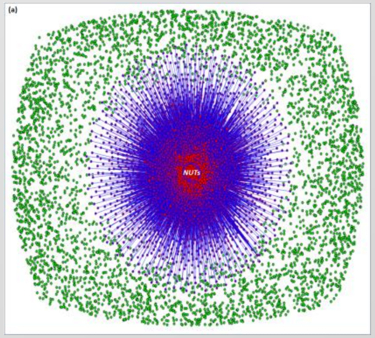




NUTs are much closer to each other than expected by chance

NUTs form a tightly connected network when clustered by similarity





NUTs don’t form clusters (random scatter around center)

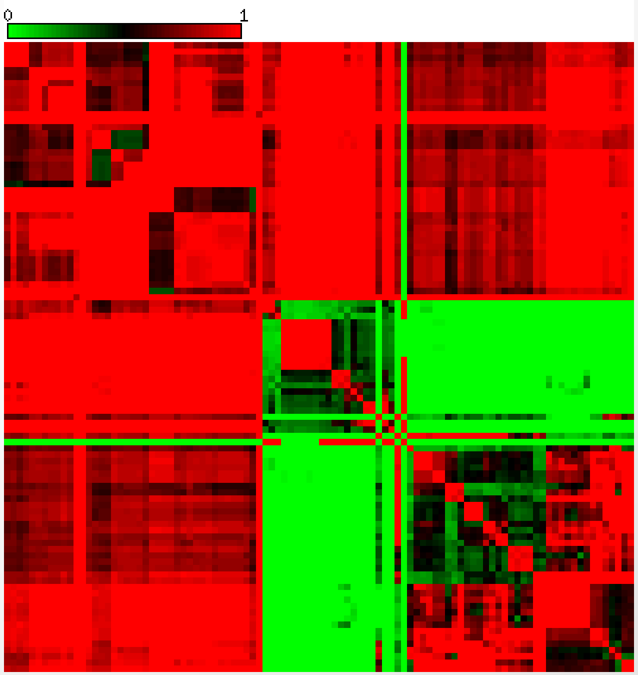
NUTs are connected to the rest of the forest

Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

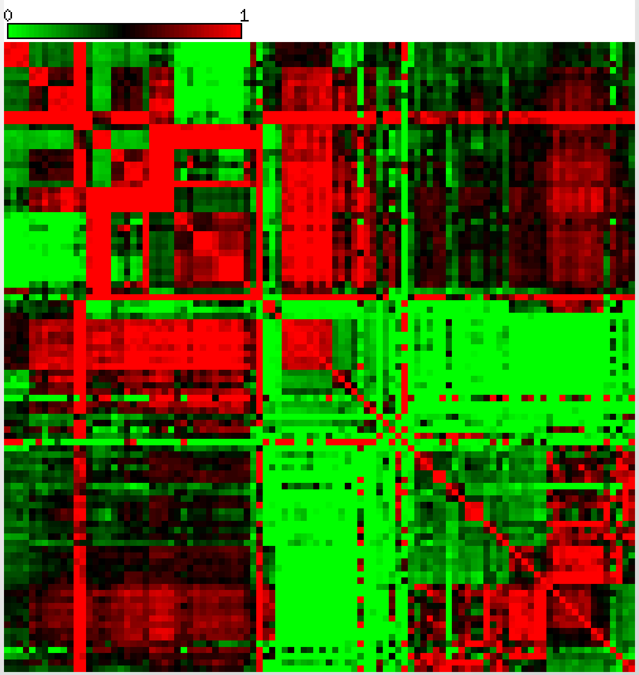
Forest of Life – Analysis

“Tree-like” vs “Net-like” components of the trees (how many quartets agree/disagree with the consensus tree).



NUTs

0.63 +/- 0.35



FOL

0.39 +/- 0.31

NUTs are dominated by tree-like descent

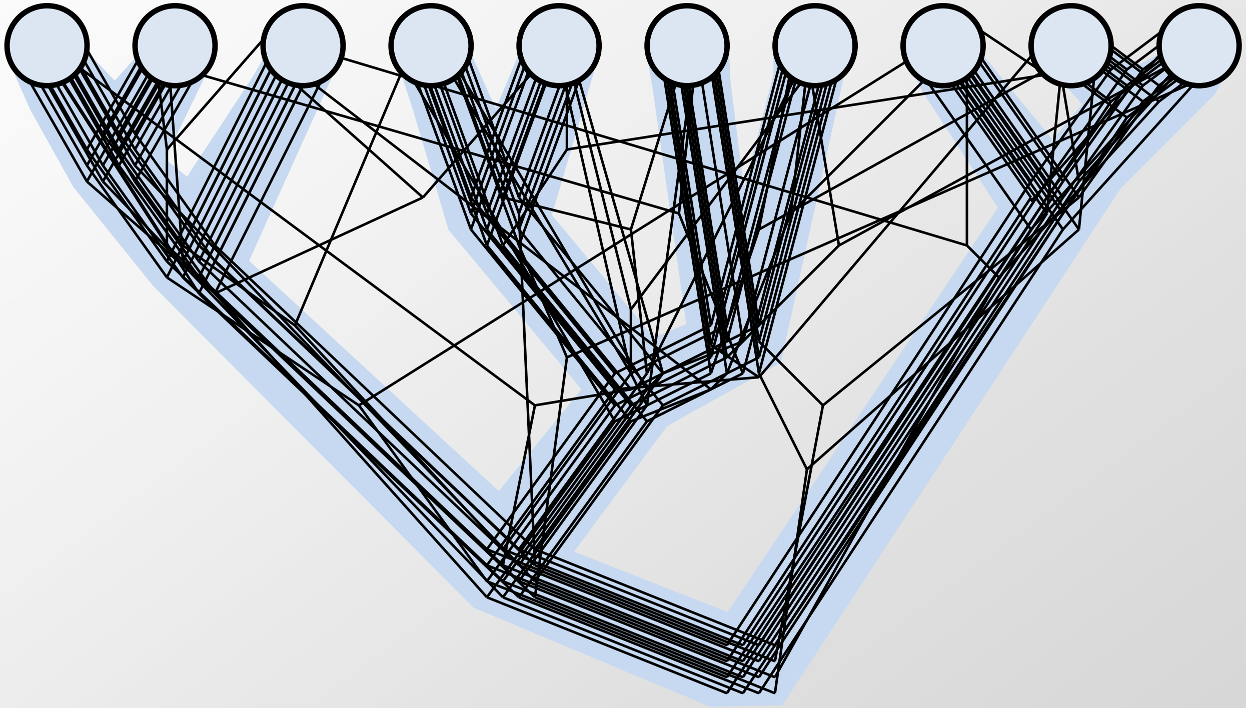
Overall the forest of life is dominated by network-like relationships (HGT)

Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Forest of Life – Analysis

Simulated example of 16 trees for 10 organisms:



No two trees are the same; each contains 2 random deviations from the consensus tree. Common statistical trend is visible.

Courtesy of Yuri Wolf; slide in the public domain.

Taken from Yuri Wolf, Lecture Slides, Feb. 2014

Module V: Evolution/phylogeny/populations

- Phylogenetics / Phylogenomics
  - Phylogenetics: Evolutionary models, Tree building, Phylo inference
  - Phylogenomics: gene/species trees, reconciliation, coalescent, pops
- Population genomics:
  - Learning population history from genetic data
  - Assembling and getting information on genomes
  - Recitation about suffix arrays used in genome mapping and assembly
- Next Pset due on Nov 1st
  - Don't wait until the last week to start it!

6.047 / 6.878 / HST.507 Computational Biology
Fall 2015