Computational Systems Biology

… Biology X – Lecture 4…

Bud Mishra

Professor of Computer Science, Mathematics, & Cell Biology

Molecular Evolution

Bio-Diversity

- Life is ubiquitous and old.
  - (3.7 billion years old!)
- Living organisms on the Earth have diversified and adapted to almost every environment.

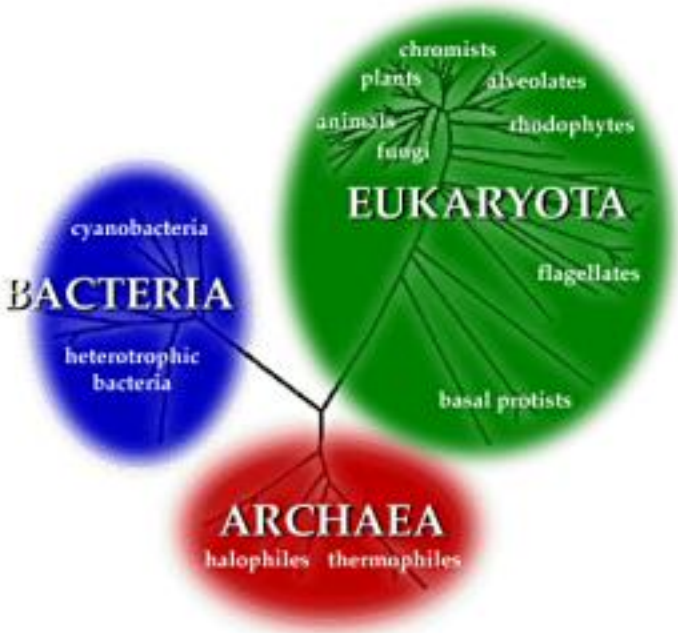
Bio-Diversity

- All living organisms can replicate, and the replicator molecule is DNA.
  - The information stored in DNA is converted into products used to build similar cellular machinery.
  - Comparative study of the DNA can shed light on its function in the cell and the process of evolution.

Five Kingdoms

- All living organisms are divided into five kingdoms:
  1. Protista,
  2. Fungi,
  3. Monera (bacteria),
  4. Plantae, and
  5. Animalia.

Tree of Life



- A different scheme:
  1. **Prokaryotae** (bacteria, etc.)
     1. Bacteria
     2. Archea
  2. **Eukaryotae** (animals, plants, fungi, and protists).
- No one of these groups is ancestral to the others.
- A fourth group of biological entities, the **viruses**, are not organisms...

Human Evolution

- Two Models:
  - Multiregional Model
  - Out of Africa Model
    - Evolution of a tree of hominids originating in Africa. Left Africa about 1 million years ago. Two waves of migration are speculated.
- African human population has the most diversity.

Homo sapiens

- Australopithecus (3.5million years old)
- Homo habilis (2 million yrs)
- Home erectus (1 million yrs),
- **Homo sapiens** (60,000-100,000 yrs)
  - Cro Magnon Man (Our immediate H. sapien ancestor)
  - Neanderthal Man (Became extinct ~30,000 yrs ago.)
- Two distinct species; supported by DNA amplification and sequence alignment (S. Paabo)

Mitochondria and Phylogeny

- **Mitochondrial DNA (mtDNA):** Extra-nuclear DNA, transmitted through maternal lineage. Mitochondria are inherited in a growing mammalian zygote only from the egg.
- 16.5 Kb, contains genes: coding for 13 proteins, 22 tRNA genes, 2 rRNA genes.
- mtDNA has a pointwise mutation substitution rate 10 times faster than nuclear DNA.
- Phylogeny based on human mtDNA can give us molecular (hence accurate!) information about human evolution.

African Eve

- Statistical analysis of mtDNA extracted from placental tissue of 147 women of different races and regions. (Cann, Stoneking, & Wilson, 87).
- Phylogenetic tree (assuming a constant molecular clock) was constructed by Wilson.
- A single rooted tree with the root being closest to the modern African woman.
- **Conclusion**: Modern man emerged from Africa 200,000 years ago. Race differences arose 50,000 years ago. "Mitochondrial Eve Hypothesis"

Mitochondrial Eve's Africanness

- A simple reordering of the data could result in 100 distinct trees all at most 2 steps away---all supporting non-African hypothesis. (Templeton)
- Assuming a non-constant molecular clock results in a least universal common ancestor (Luca) … that is too young.
- In general, mathematical descriptions and algorithms that may lead to "historically correct phylogenetic tree" remain to be developed.

Taxon

- **Taxon (Taxonomical Unit)**: is an entity whose similarity (or dissimilarity) can be numerically measured. E.g., Species, Populations, Genera, Amino Acid Sequences, Nucleotide Sequences, Languages.

Phylogeny

- **Phylogeny** is an organization of the taxa in a rooted tree,
  - with distances assigned to the edges in a such manner that the "tree-distance" between a pair of taxa equals the numerical value measuring their dissimilarity.
- The dissimilarity and the edge lengths of the phylogenic trees can be related to the rate of evolution (perhaps determined by a molecular clock).

Comparing a Pair of Taxa

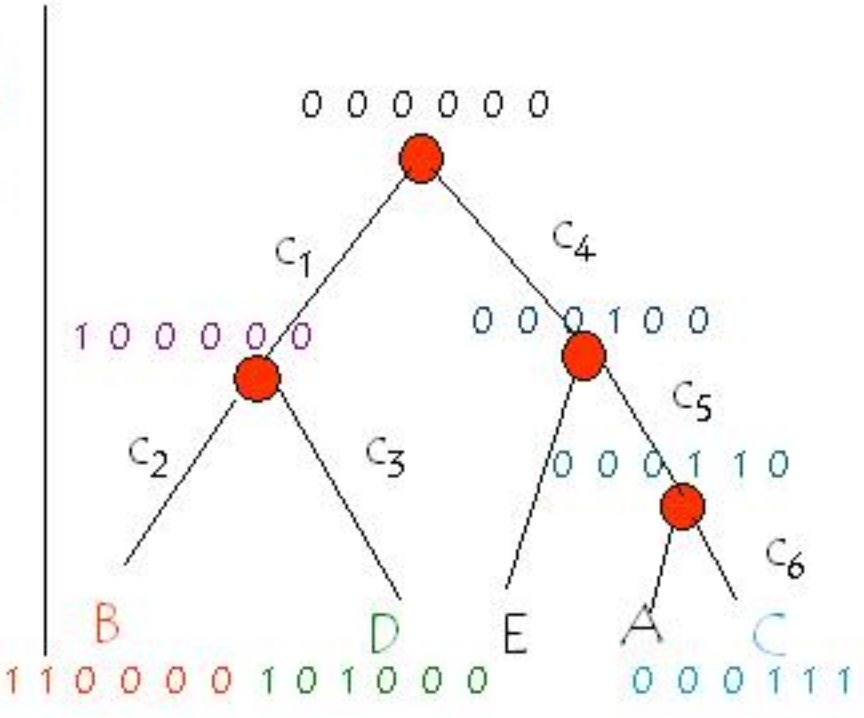
- *Discrete Characters*: Each taxon possesses a collection of characters and each character can be in one of finite number of states. One can describe an n taxa with characters by an $n \times m$ matrix over the state space. Character State Matrix.
- *Comparative Numerical Data:* A *distance* is assigned between every pair of taxa. One can describe the distances between n taxa by an $n \times n$ matrix over $\mathbf{R}_+$. Distance Matrix.

Examples

A character state matrix

Taxon	c_1	c_2	c_3	c_4	c_5	c_6
A	0	0	0	1	1	0
B	1	1	0	0	0	0
C	0	0	0	1	1	1
D	1	0	1	0	0	0
E	0	0	0	1	0	0

Edges where state transition takes place is given by an associated character.



Character States

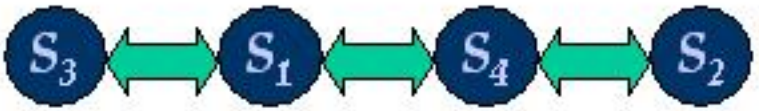
- **Some Assumptions:**
  - The characters are inherited independently from one another.
  - Observed states of a character have evolved from one "original state" of the nearest common ancestor of a taxon.
  - Convergence or parallel evolution are rare. That is the same state of a character rarely evolve in two independent manners.
  - Reversal of a character to an ancestral state is rare.

Classifying Characters

- Characters:
  - **Unordered / Qualitative Character**: All state transitions are possible.
  - **Ordered / Cladistic Character**: Specific rules regarding state transition are assumed.

    – *Linear Ordering*

    – *Partial Ordering* (with a derivation tree).



Perfect Phylogeny

- A **phylogenic tree** T (with edges labeled by state transitions) is called **perfect**, if it does not allow *reversal* or *convergence*--that is, with respect to any character c, and any pair of states w and s at most one edge is labeled

$$w \to s \text{ or } s \to w.$$

- **Example**: Binary characters with two states {0=ancestral, and 1=dervied}: any character c_i labels at most one edge and implies a transition from

$$0 \to 1 \text{ in the } i^{th} \text{ position.}$$

Perfect Phylogeny Problem:

- *Perfect Phylogeny Problem:*
  - Given: A set O with n taxa, a set C of m characters, each character having at most r states.
  - Decide: If O admits a perfect phylogeny.
- A set of defining characters are **compatible**, if a set of objects defined by a character set matrix admits a perfect phylogeny.

Compatibility Criteria

- Allow reversal and convergence properties in the models of evolution.
- *Parsimony Criteria*: Minimize the occurrences of reversal and convergence events in the reconstructed phylogeny tree.
  - Dollo Parsimony Criterion: Minimize reversal while forbidding convergence.
  - Camin-Sokal Parsimony Criterion: Minimize convergence while forbidding reversal.

Compatibility Criteria

- *Compatibility Criteria*: Exclude minimal number of characters under consideration so that the reconstructed phylogeny tree is perfect and does not admit any occurrence of reversal or convergence.

Computational Infeasibility

- **Perfect Phylogeny Problem** for arbitrary (>2) number of unordered characters and arbitrary (> 2) number of states is NP-complete.
- **Optimal Phylogeny Problem under compatibility criteria** is NP-complete.
- **Optimal Phylogeny Problem either under Dollo or Camin-Sokal parsimony criteria** is NP-complete.

Binary Character Set

- Each character has two states = {0, 1}
- If a character is ordered then $0 \rightarrow 1$ (0=ancestral and 1=derived), or converse.
- For binary characters (ordered or unordered), perfect phylogeny problem can be solved efficiently
  - Poly time, for n taxa and m characters, Time = $O(nm)$.
- A two phase algorithm:
  1. **Perfect Phylogeny Decision Problem**
  2. **Perfect Phylogeny Reconstruction Problem**

Compatibility Condition

- T = Perfect Phylogeny for M iff

$$(\forall\, c_i = \text{character})(\exists!\, e = \text{tree-edge})\ \text{label}(e) = \{c_i, 0 \rightarrow 1\}$$

$$\text{root}(T) = (0, 0, 0, \ldots, 0)$$

- A path from root to a taxon t is labeled $(c_{i1}, c_{i2}, \ldots, c_{ij})$

$\Rightarrow$ t has 1's in positions $i_1, i_2, \ldots, i_j$

- **Perfect Phylogeny Condition**
  - $M = n \times m$ Character State Matrix, $j \in \{1..m\}$
  - $O_j = \{i = \text{taxon} : M_{ij} = 1\}$
  - $O_j^c = \{i = \text{taxon} : M_{ij} = 0\}$

Key Lemma

- **Lemma**: *A binary matrix M admits a perfect phylogeny iff*

$$(\forall\ i, j \in \{1, m\})\ O_i \cap O_j = \emptyset \text{ or } O_i \subseteq O_j \text{ or } O_i \supseteq O_j$$

Proof of Lemma

- Proof: ($\Rightarrow$) T_i = subtree containing O_i, T_j = subtree containing O_j, r_i = root(T_i) and r_j = root(T_j)

  r_i is neither an ancestor nor descendant of $r_j \Rightarrow O_i \cap O_j = \emptyset$

  r_j is a descendant of $r_i \Rightarrow O_j \subseteq O_i$

  r_i is an ancestor of $r_j \Rightarrow O_i \supseteq O_j$

- ($\Leftarrow$) By induction, Base case m=1 is trivial. Induction case, $m = k+1$:

  T_k = Tree for k characters. O_{k+1} is contained in a subtree with minimal # taxa rooted at r.

  r must be a leaf node. Either an edge needs to be labeled or the subtree rooted at r has to be split. □

Simple Algorithm based on the Lemma

- Compare every pair of columns for the intersection and inclusion properties. Total of $O(m^2)$ pairs, each comparison can be done in $O(n)$ time.
- Total Time Complexity = $O(nm^2)$
- Can be improved to $O(nm)$ time.

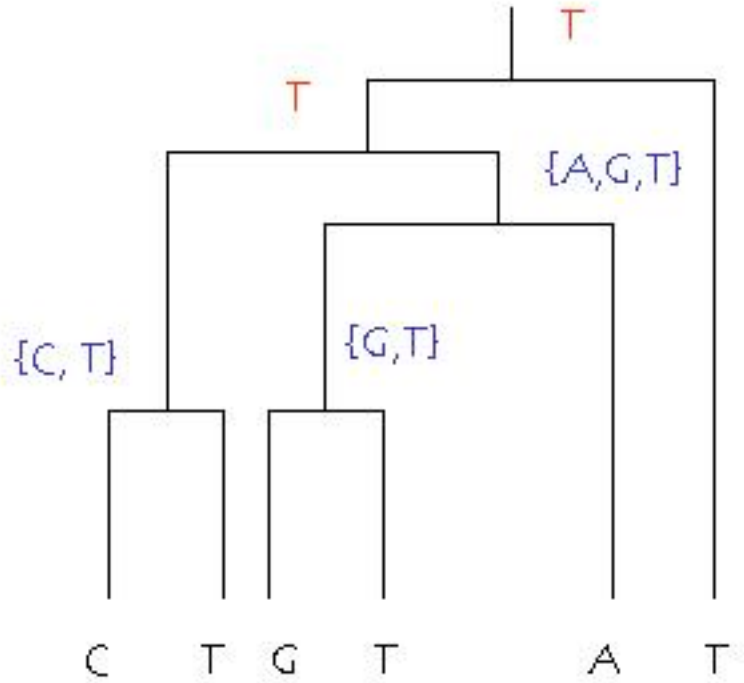
Parsimony with Distance Based Tree

- Fitch's Algorithm
- Finding minimum number of changes for a given tree:
  - Assume any state (e.g., nucleotide, amino acid) can convert to any other state.
  - As before assume that positions are independent.

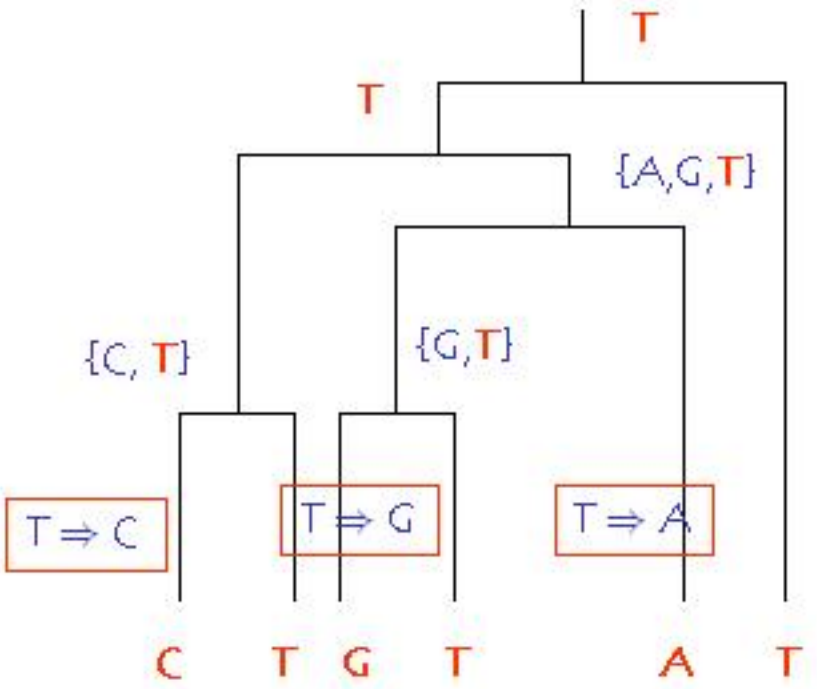
Fitch's Algorithm

- (For Nucleotide Sequences)
- Traverse tree from leaves to root determining set of possible states (e.g. nucleotides) for each internal node.
- Traverse tree from root to leaves picking ancestral states for internal nodes.

Fitch's Algorithm (Example)



Fitch's Algorithm (Example)



Cost of this phylogeny tree for this column of nucleotides = 3

Fitch's Algorithm—

- Step 1: Possible States for Internal Nodes
- Do a post-order (from leaves to root) traversal of tree
- Determine possible states of R_u of internal node u with children i and j:
- R_u =

$$\begin{cases} R_i \cup R_j & \text{if } R_i \cap R_j = \emptyset \\ R_i \cap R_j & \text{otherwise.} \end{cases}$$

Fitch's Algorithm—

- Step 2: Select States for Internal Nodes
- Do a pre-order (from root to leaves) traversal of tree
- Select state r_u of internal node u with parent v:
- $R_u = \begin{cases} r_v, & \text{if } r_v \in R_u \\ \text{Any state} \in R_u, & \text{otherwise.} \end{cases}$

- The cost of the tree is the number of state changes imposed by the tree topology.

Sankoff-Cedergren Algorithm:

- Weighted Version of Fitch's Algorithm
- All state transitions are not necessarily equally likely. Use different costs S(A, B) for different transitions

$$A \Rightarrow B$$

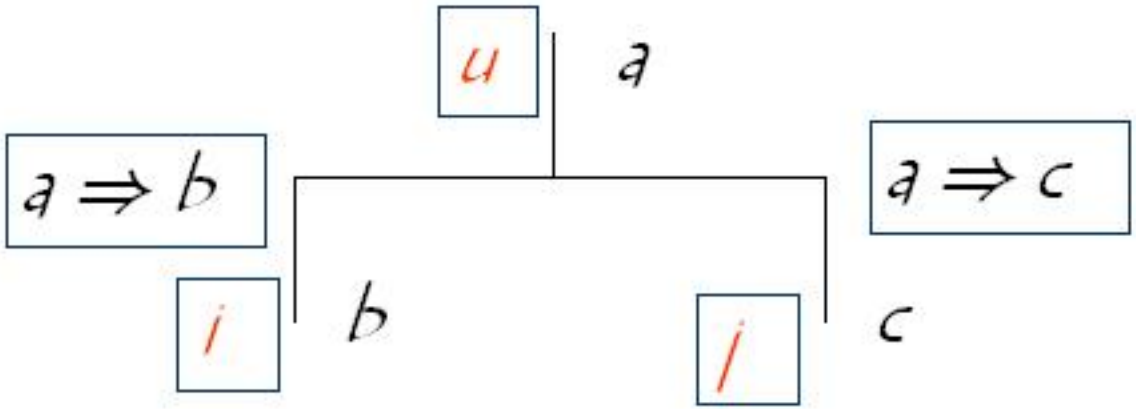
- Modification to Fitch's algorithm: (Step 1) Propagate costs up the tree.
- Base case: Leaf Nodes:
- $R_i(a)$ =

{0, if the leaf is labeled by the character a;

Sankoff-Cedergren Algorithm:

- Internal Nodes:

$$R_u(a) = \min_b [R_i(b) + S(a, b)] + \min_b [R_j(b) + S(a, b)]$$

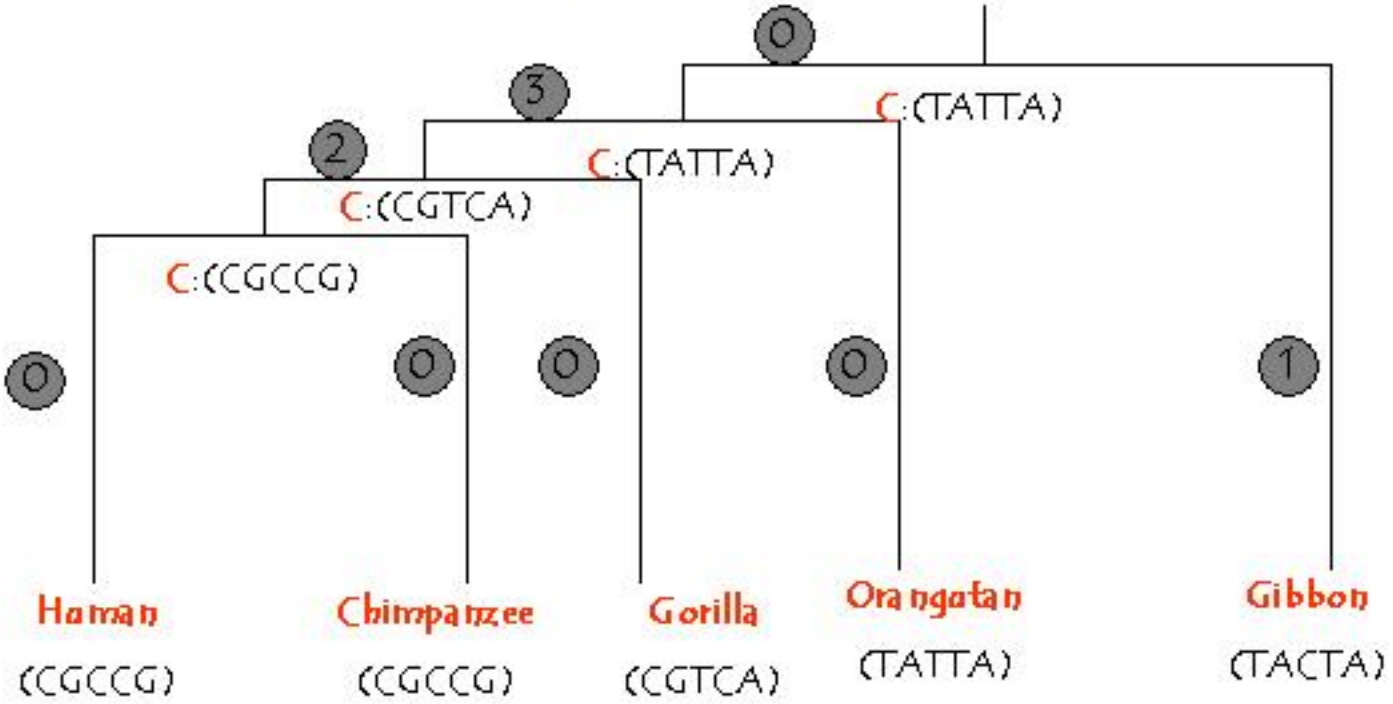


Sankoff-Cedergren Algorithm:

- Step 2:
- Do a pre-order (from root to leaves) traversal of tree
- Select minimal cost character for each internal node.

Maximum Parsimony Tree:

mtDNA data for Primates



Rate of Evolutionary Changes

- Taxa of nucleotide or amino acid sequences.
- Given two taxa s_i and s_j, measure their distance
  - Distance(s_i, s_j), d_{ij} = Edit distance based on pairwise sequence alignment.
- Assumptions about the Molecular Clock (governing rate of evolutionary change):
  - Only independent substitutions
  - No back or parallel mutations
  - Neglect selection pressure.

Distance Based Approaches

- **Given**:

  An $n \times n$ nonnegative-valued distance matrix $M \in R_+^{n \times n}$, where M_{ij} is the distance between objects i and j:

- **Construct**:

  An edge-weighted tree such that the distances between leaves i and j are "close" to M_{ij}

Average Linkage Clustering

- UPGMA
  - (Unweighted Pair-Group Method using an Arithmetic Average).
- Distance between clusters (disjoint sets of taxa) C_i and C_j is
- Distance(C_i, C_j) = d_{ij}

$$= (1/|C_i| \cdot |C_j|) \sum_{p \in C_i,\, q \in C_j} d_{pq}.$$

- This is the average distance between pairs of taxa from each cluster.

UPGMA

- Assign each taxon to is own cluster.
- Define one leaf for each taxon—
  - Place it at height 0.
- While more than two clusters exist
  - Determine two clusters i and j with smallest d_{ij}
  - Define a new cluster $C_k = C_i \cup C_j$
  - Define a node k with children i and j—
    - Place it at height $d_{ij}/2$.
  - Replace clusters C_i and C_j with C_k
- Join the last two clusters (i and j) by root at height $d_{ij}/2$. □

Nucleotide Sequences

- Synonymous or Neutral Substitutions:
  = Nucleotide substitutions with no effect on expressed amino acid sequences
  - RECALL: Genetic code is redundant—Most substitutions to 3rd positions are synonymous.
  - Often a single non-synonymous nucleotide substitution is likely to change one amino acid into a related amino acid (e.g., both hydrophobic).
- Molecular clock is modeled based on non-synonymous substitution rate.

Variability of Nucleotide Mutation Rate

- **Transitional Mutations:**
  - purine-purine, i.e. A ↔ G
  - pyrimidine-pyrimidine, i.e. C ↔ T
- **Transversal Mutations:**
  - purine-pyrimidine: A ↔ T, A ↔ C, G ↔ C, G ↔ T
- Usually transitional mutations are more likely. Mutation into A is more likely.

DNA repair

- Effect of DNA repair mechanism

λ for ...	#per site per year
higher primate	$\approx 1.3 \times 10^{-9}$ /site/yr
sea urchins & rodents	$\approx 6.6 \times 10^{-9}$/site/yr
mammalian mtDNA	$\approx 10^{-8}$/site/yr
plant cpDNA	$\approx 1.1 \times 10^{-9}$/site/yr

Markov Process Model of Mutation

- Evolution is modeled by a stochastic process, $X(t)$ with real-valued time parameter $t \geqq 0$
- A time-homogeneous Markov process
- $(Q, \pi, \mathbf{P}(t))$
- $Q = \{A, C, G, T\}$ = States
- $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ = Initial Distribution
- $\mathbf{P}(t) =$

$$\begin{pmatrix} p_{A,A}(t) & p_{A,C}(t) & p_{A,G}(t) & p_{A,T}(t) \\ p_{C,A}(t) & p_{C,C}(t) & p_{C,G}(t) & p_{C,T}(t) \\ p_{G,A}(t) & p_{G,C}(t) & p_{G,G}(t) & p_{G,T}(t) \\ p_{T,A}(t) & p_{T,C}(t) & p_{T,G}(t) & p_{T,T}(t) \end{pmatrix}$$

Markov Process (Contd.)

- $P_{\sigma,\tau}(t)$
  $= Pr[\sigma \mid \tau, t] = Pr[X(t) = \sigma \mid X(0) = \tau]$
  = Probability that a nucleotide with a value τ at time 0 mutates to a σ by time t
- $P(t+s) = P(t)P(s)$
- $p_i(t) = Pr[X(t) = i]$
  $= \sum_{k \in \{A,C,G,T\}} \pi_k p_{k,i}(t)$
- $\pi^* = \{\pi_A^*, \pi_C^*, \pi_G^*, \pi_T^*\}$ is a stationary distribution for $P(t)$

$$\forall\, t \quad \pi^* P(t) = \pi^*$$

Markov Process (Contd.)

- $P'(t)$
  $= P(t) \lim_{\Delta t \to 0} [P(\Delta t) - P(0)]/[\Delta t]$
  $= P(t)\, \Lambda$
- Solution to the differential equation:
  $P(t) = \exp(\Lambda t) = \sum_{n=0}^{\infty} \Lambda^n t^n / n!$
- Row-sum for Λ is 0:
  $\sum_j \lambda_{i,j} = \lim_{\Delta t \to 0} [\sum p_{i,j} - 1]/[\Delta t] = 0.$

Juke-Cantor Model

- $(\pi_A, \pi_T, \pi_C, \pi_G) = (1/4, 1/4, 1/4, 1/4)$

- $\Lambda = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$

- $\pi = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$

Juke-Cantor Model (Contd.)

- $\Lambda = -4\alpha(1-\pi)$
- $P(t) = e^{-4\alpha(1-\pi)t}$

  $= I\,[\sum_{n=0}^{\infty}(-4\alpha\, t)^n/n!]\;\{\sum_{n=0}^{\infty}\pi^n(4\alpha\, t)^n/n!\}$

  $= I\, e^{-4\alpha t}\{I + \pi(e^{4\alpha t} - 1)\}$

  $= e^{-4\alpha t} I + \pi(1 - e^{-4\alpha t})$
- $p_{i,i}(t) = 1/4(1 + 3e^{-4\alpha t})$
- $p_{i,j}(t) = 1/4(1 - 4e^{-4\alpha t}),\quad i \neq j.$

Example

- (Based on mtDNA Sequences)
- Let q = the proportions of nucleotides that is same in two mtDNA sequences.

$$K \propto t$$

$$q = \frac{1}{4}(1 + 3e^{-K});$$

$$K = \ln(3/(4q-1))$$

- Juke-Cantor distance between a pair of mtDNA sequences is given by

$$K' = (3/4)\ln(3/(4q-1))$$

Example (Contd.)

- Differences in mtDNA sequences

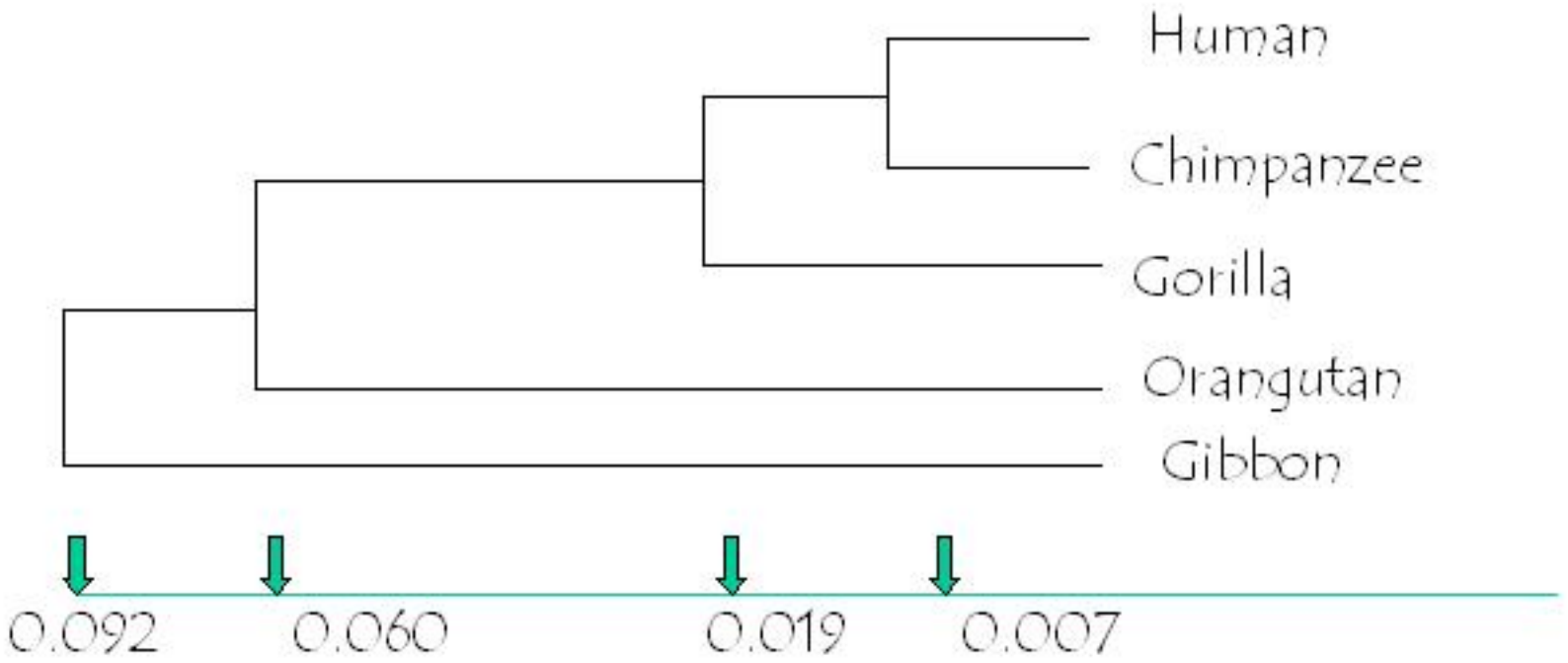
	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	-	1	3	9	12
Chimpanzee		-	2	8	11
Gorilla			-	6	11
Orangutan				-	11
Gibbon					-

Example (Contd.)

- Juke-Cantor distances between primates

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	-	0.015	0.045	0.143	0.198
Chimpanzee		-	0.030	0.126	0.179
Gorilla			-	0.092	0.179
Orangutan				-	0.179
Gibbon					-

UPGMA Phylogeny



UPGMA & The Molecular Clock

- Assumes a *constant molecular clock*:
  - Divergence of sequences is assumed to occur at the same rate at all points in the tree.
- This assumption is in general false
  - Selection pressures vary across time periods, organisms, genes within an organism, regions within a gene.

Ultrametric Trees

- Distance function ρ satisfies the axioms:

  $\rho(i, j) \geq 0$ with equality iff $i = j$;

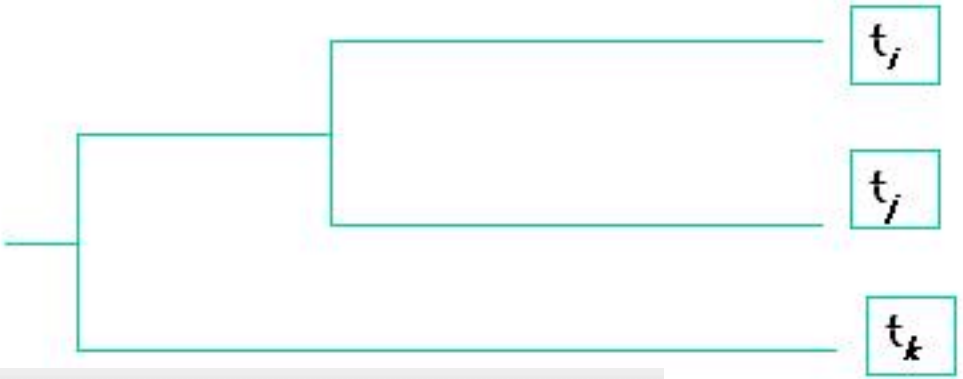
  $\rho(i,j) = \rho(j,i)$ (symmetry);

  $\rho(i,k) \leq \rho(i,j) + \rho(j,k)$ (triangle inequality).

- Path length between i, j of T = Sum of edge weights along the path connecting i and j.
- If $\forall$ i and j, $\rho_{i,j}$ = path length between i, j of T, then ρ is called an additive tree metric.
- If the path length from the root to every leaf is identical then ρ is called an ultrametric.

UPGMA & Ultrametric Data

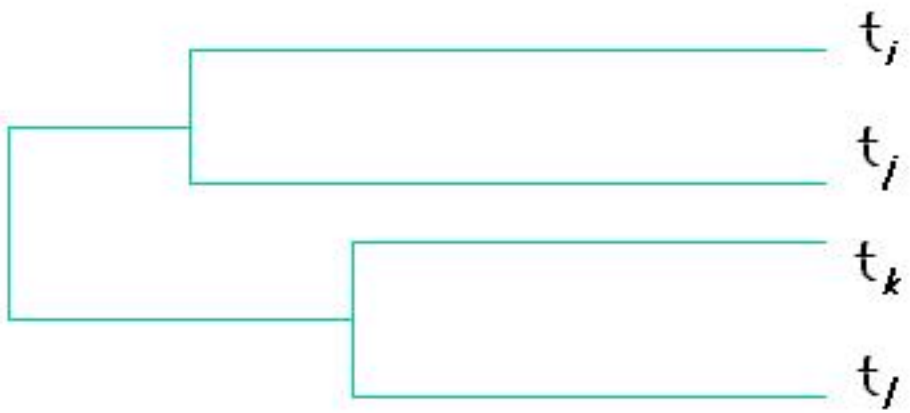
- If the rates of evolution among different lineages are exactly the same, then the data is **ultrametric**.
- Definition (3-Point Condition): For any triplet of sequences, $i \neq j \neq k$, of the three distances d_{ij}, d_{jk}, d_{jk} r two are equal and not less than the third.



$$d_{i,k} = d_{j,k} \geq d_{i,j}$$

Additive Metric

- Every additive metric satisfies the 4-point condition:
- $\forall\ i, j, k, l$, of the three sums $S_1 = d_{i,j} + d_{k,l}$, $S_2 = d_{i,k} + d_{j,l}$ and $S_3 = d_{i,l} + d_{k,j}$ two are equal and not less than the third. E.g. $S_1 \leq S_2 = S_3$.



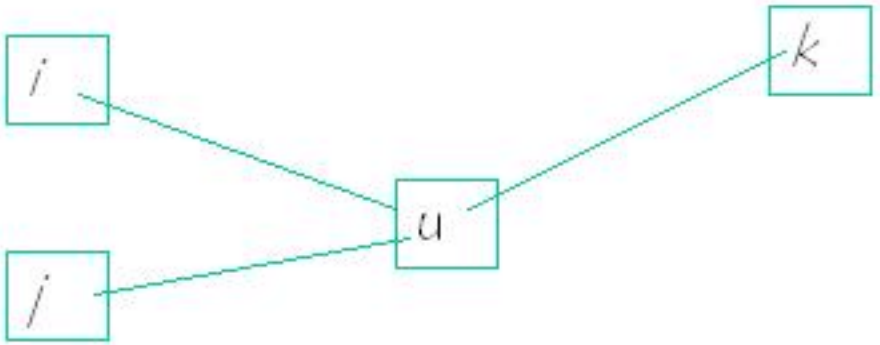
Neighbor Joining

- Like UPGMA constructs a tree by sequentially joining subtrees
- Unlike UPGMA
  - Does not make molecular clock assumption
  - Produces unrooted tree
- Does assume additivity: Distance between a pair of taxa is the path length in the tree.

Distances in Neighbor Joining

- Given a new internal node u, the distance to another node k is given by

$$s_{u,k} = (d_{i,k} + d_{j,k} - d_{i,j})/2$$

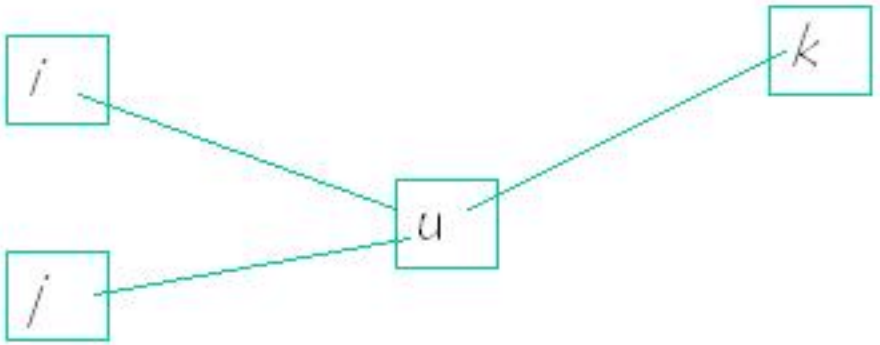


Distances in Neighbor Joining

- Calculate the distance from a leaf to its parent node similarly:

$$s_{i,u} = (d_{i,j} + d_{i,k} - d_{j,k})/2 = d_{i,j}/2 + (d_{i,k} - d_{j,k})/2$$

$$s_{j,u} = d_{i,j} - s_{i,u}$$

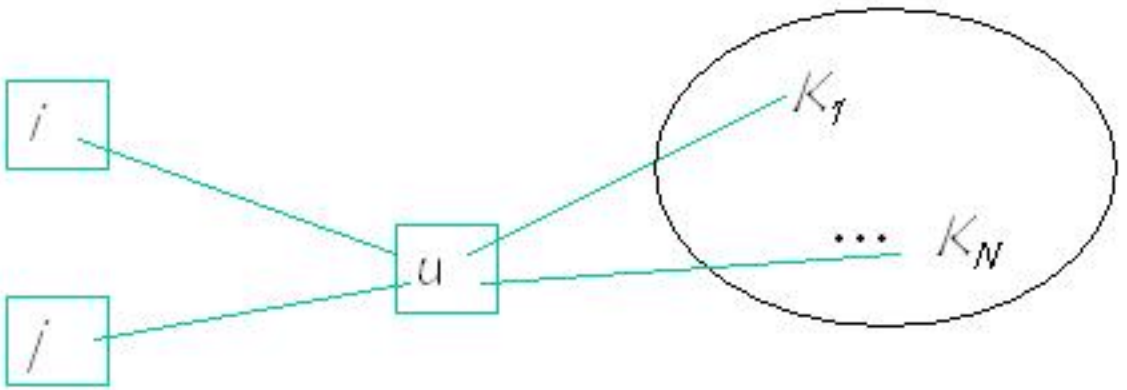


Generalizing the Scheme

- (To more than Three Leaves)
- Define $r_i = \sum_{k=1}^{N} d_{i,k}$
- Rate corrected distance between taxa i and j:

$$m_{i,j} = d_{i,j} - (r_i + r_j)/(N-2)$$

is used to choose the "nearest neighbors" to be joined.

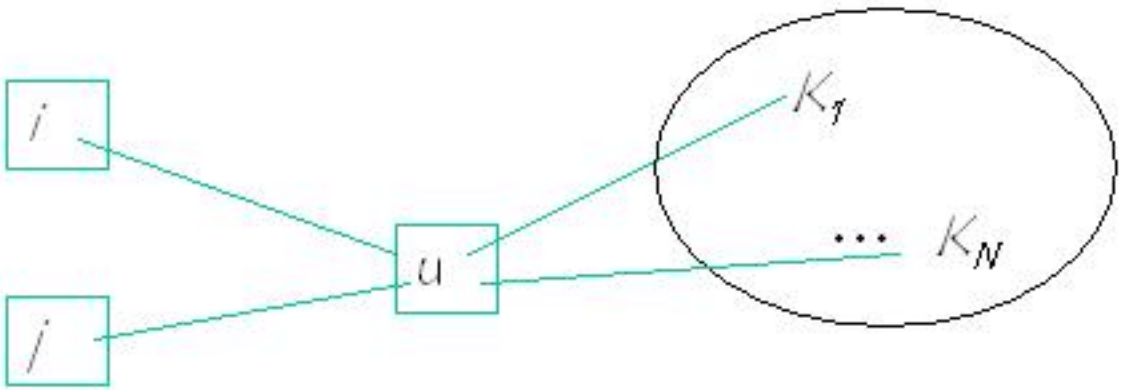


Generalizing Distances in NJ

- Calculate the distance from a leaf to its parent node similarly:

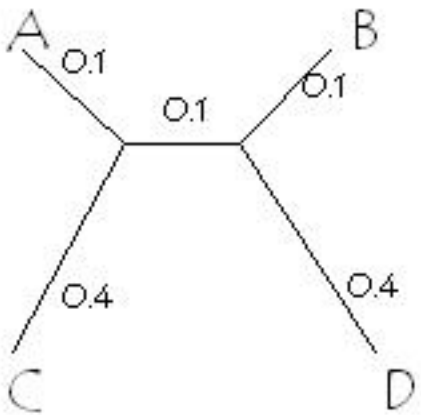
$$s_{i,u} = d_{i,j}/2 + (r_i - r_j)/(2(N-2))$$

$$s_{j,u} = d_{i,j} - s_{i,u}$$



Picking a Pair of Nodes to Join

- At each step, pick a pair of "nearest neighbor" nodes to join- Nearest neighbor is not determined by minimal $d_{i,j}$ but $m_{i,j}$



$d_{A,B} = 0.3$

$d_{A,C} = 0.5$

$r_A = 1.4,\ r_B = 1.4,\ r_C = 2.0$

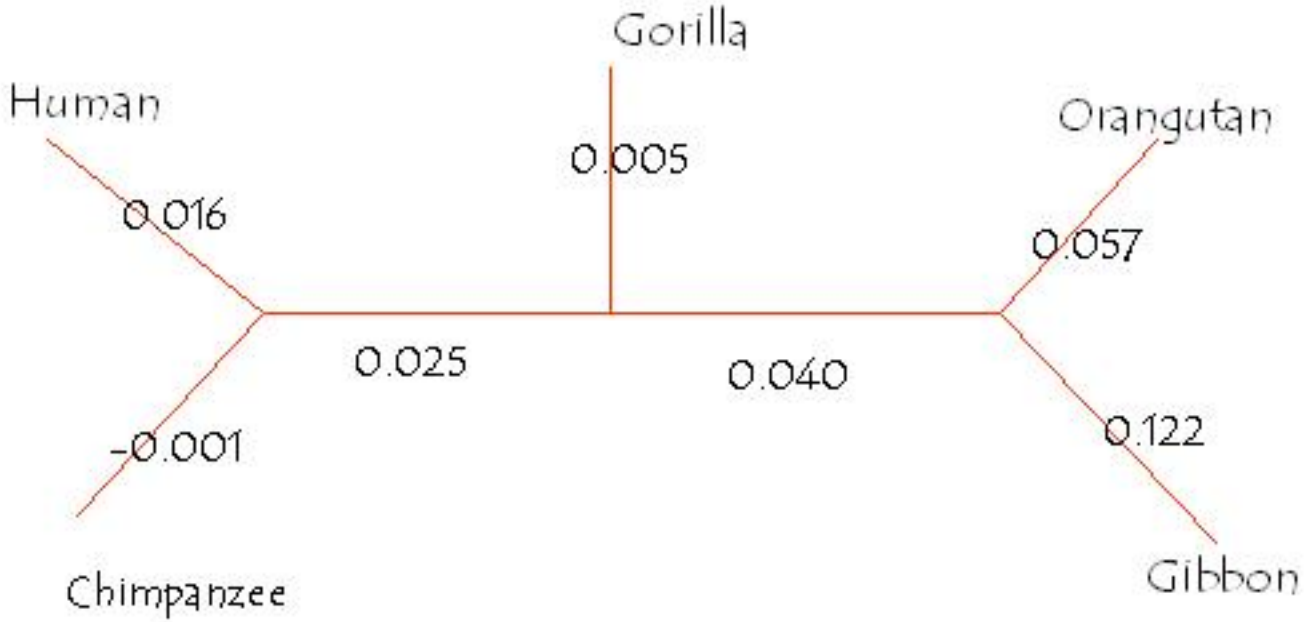
$m_{A,B} = d_{A,B} - (r_A + r_B)/2 = -1.1$

$m_{A,C} = d_{A,C} - (r_A + r_C)/2 = -1.2$

Neighbor Joining Algorithm

- T= Set of leaf nodes
- While more than two subtrees in T
  - Pick a pair i, j in T with minimal m_{ij}
  - Define a new node u joining i and j
  - Remove i and j from T and insert u
- Join the last two remaining subtrees

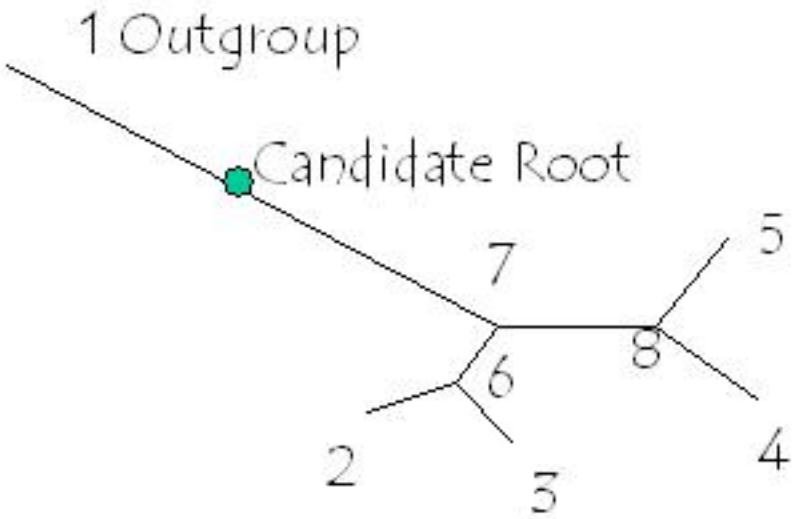
Example



Rooting Trees

- Neighbor joining method creates an unrooted phylogenetic tree.
- A root is assigned to an unrooted tree by finding an *outgroup*.
  - An outgroup is a species known to be more distantly related to remaining species than they are to each other.
  - Point where the outgroup joins the rest of the tree is best candidate for root position.

Rooting Trees



Other Distance Matrix Methods

- Phylogenetic Trees are constructed using:
  - Clustering Method: Identifies groups of close taxa. E.g. UPGMA or Average Linkage Clustering Methods.
    - Sequential
    - Agglomerative
    - Hierarchical
    - Nonoverlapping
  - Pairwise Method: Pairs a taxon (or a group of taxa) with its nearest neighbor. E.g. Additive trees constructed with Fitch-Margolish Algorithm.

To be continued…

. . .