

Systematics

An Dinh

UTHealth School of Public Health

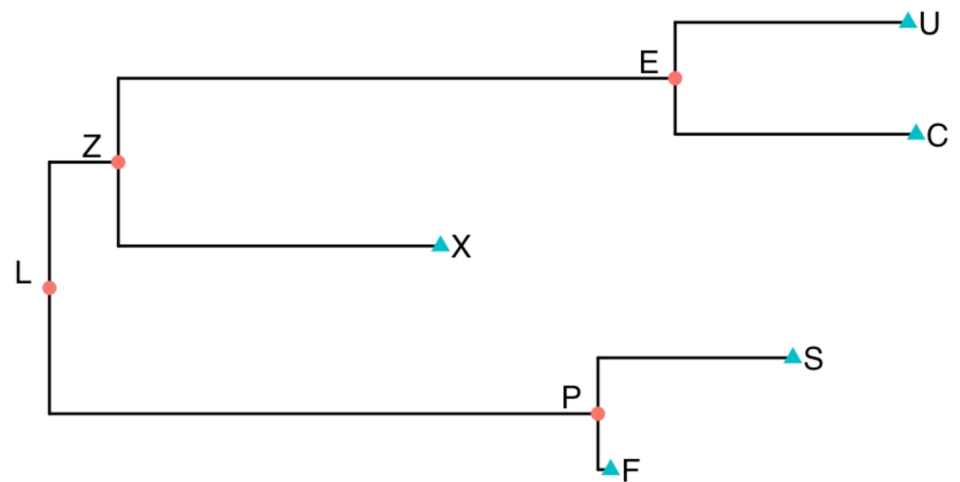
Houston Methodist Research Institute

Goals

- Introduction to methods and models for phylogenetic inference
- Parameter selection
- Common language

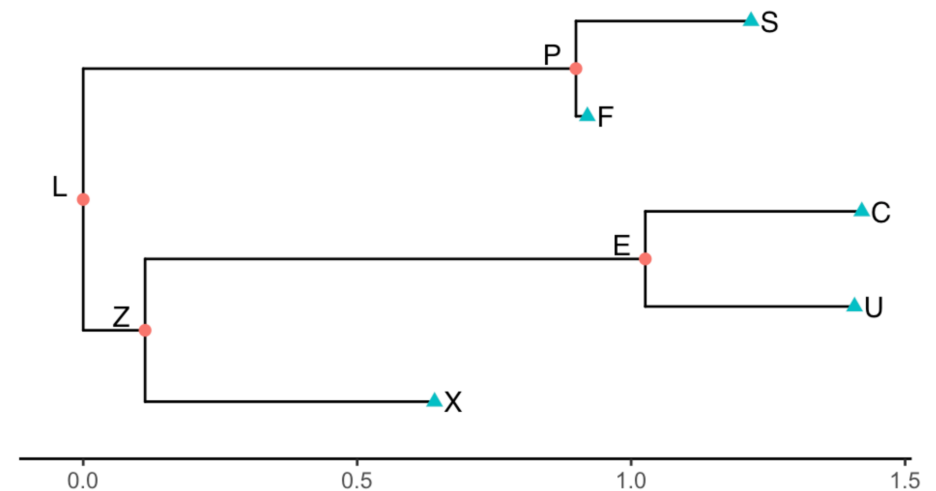
Dissecting a Tree

- Node (vertex) = divergence point
 - Internal nodes = nodes with at least one child branch
 - External nodes = nodes without child branches. Aka: “leaves”, “tips”,
- Branch (edge) = connection
 - Branch length represents the amount of distance
 - Can be external (terminating at a leaf), or internal



More on Phylogenetic Trees

- Observed (sequenced) samples are on branch tips.
- Time runs from root to tip
- The order of the leaves don't really mean anything
- Most Recent Common Ancestor (MRCA)
 - Or, Last Common Ancestor (LCA)
 - Located on internal node and Hypothetical!



Constructing Phylogenetic Trees – distance matrix-based

- **Algorithmic**
- Unweighted Pair-Group Method with Arithmetic Mean (UPGMA)
 - Branch lengths are equally split
 - Fitch-Margoliash (FM) Method modifies this to do least-squares estimate of branch length
- Neighbor-Joining (NJ) - Saitou and Nei, 1987
 - Pairwise distance matrix → overall divergence → adjusted distance matrix
 - Not a bad heuristic and useful for initial starting tree

Phylogenetic Inferences directly from sequences

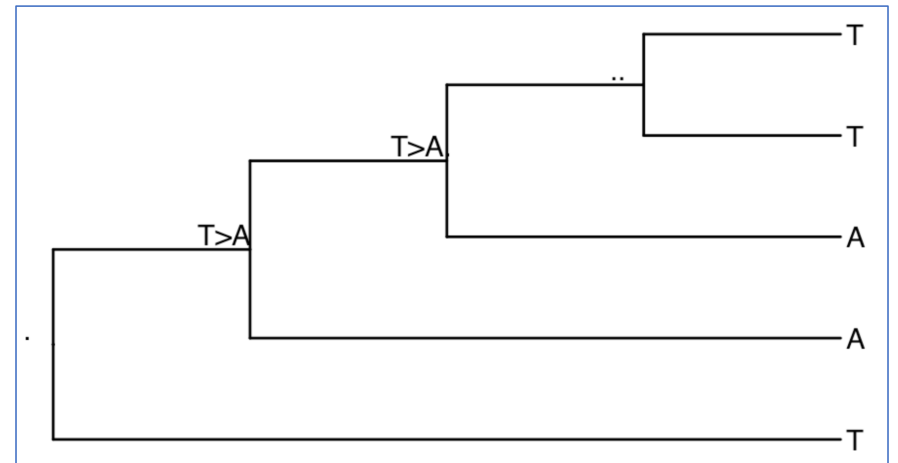
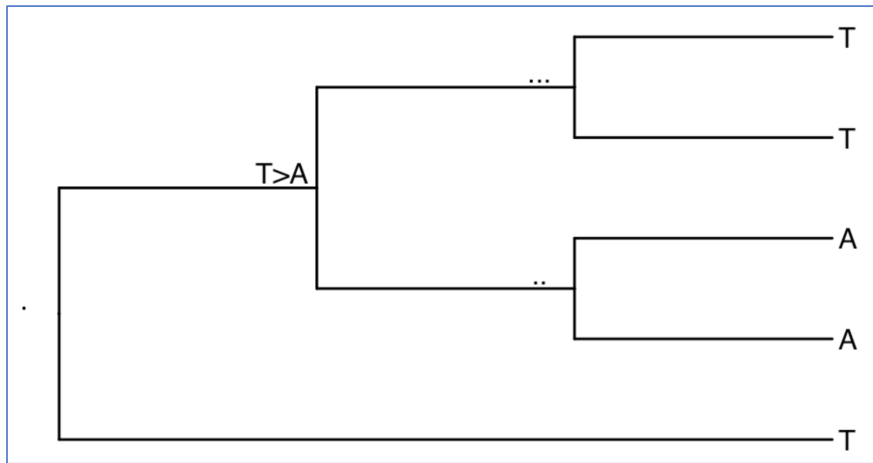
Maximum Parsimony

- Out of all possible trees, the best tree will have the least number of evolutionary (substitution) events
- Assumptions
 - Homologous similarity – identical alleles came from a common ancestor
 - Homoplasy (such as convergent evolution or reversal) is rare

Maximum Likelihood

- $\text{Max}(\text{Pr}(\text{data} | \text{tree}))$, given:
 - A sample set of sequences, and
 - A nucleotide substitution (evolutionary) model
- Assumptions
 - Evolutionary model
- Very computationally intensive!

Maximum Parsimony



Maximum Likelihood

- Topology
 - Nearest-neighbor interchange (NNI)
 - Given a topology, local arrangements are assessed for likelihood by successively deleting interior branches and testing the other possible reconstructions
 - Needs branch length (sequence distances) to compute
 - Subtree prune and regraft (SPR)
 - Tree bisection and reconnection (TBR)
- Branch length
 - Sum of log-likelihood for each nucleotide at each site (position), given a topology

Branch Support

- **Bootstrapping (Felsenstein 1985)**
 - In general, bootstrapping = random subsampling with replacement
 - Estimates the statistical error when sampling distribution is unknown
 - In phylogenetic context, estimates the reliability (consistency) of the resultant (majority-rule) consensus tree
 - The final resultant tree might not be the overall maximum-likelihood tree

Lots of models available

Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like HKY but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like TN but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like K81 but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like TPM2 but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like TPM3 but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIME	3	Like TIM but equal base freq.	012230
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like TIM2 but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like TIM3 but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like TVM but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

- Where to start...?

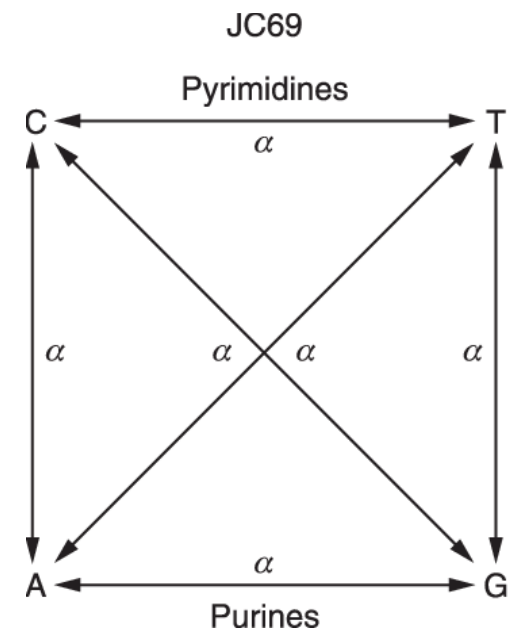
Basic Markov chain models for substitution

- **Jukes-Cantor (1969)**

- Rate of nucleotide substitution is the same for all pairs

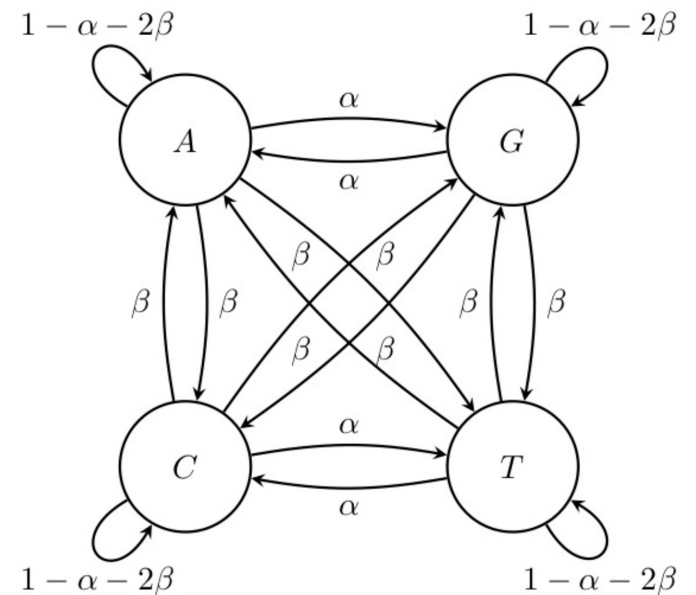
- See also: **Felsenstein (1981)**

- Like Jukes-Cantor (equal substitution rates), but with unequal base frequency



More basic models for substitution

- **Kimura (1980)**
 - Transitions (intragroup changes) have a different rate than Transversions (intergroup changes)
- See also: **Hasegawa-Kishino-Yano (1985)**
 - Like Kimura (unequal transition vs. transversion rates), but with unequal base frequency



More complex (less constrained) substitution models - GTR

- **General Time-Reversible substitution model - (Tavare 1986)**

- Like HKY, but each possible nucleotide pair has it's own "exchangeability rate" along with unequal base frequencies
- "Generalised time reversible (GTR) is the most general neutral, independent, finite-sites, time-reversible model possible."

$$Q = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

Model modifications - gamma

- **+ Γ (Yang 1994)**
 - Removes the assumption of per-site independence
 - Models rate heterogeneity in mutational rates across sites (ie: not all sites have equivalent nor independent mutational rates)
 - Applies gamma distribution with a shape parameter (alpha), itself usually estimated from the sample data

More complexities to address mutational heterogeneity

- Partitioning (categorization)
 - Apply specific models for a site or regions along a sequence
- Mixture models
 - Assign a model (out of a user-specified selection) on a per-site basis
- Some caution may need to be exercised to avoid over-parameterization

Why so many models and which to choose?

- Balancing between robustness and efficiency
 - ML criterion is **NP-hard**
 - Computational time, memory consumption...
- Inferential methods (ML, Bayesian) are statistically consistent
- GTR is probably the most popular model
- At least some form of rate heterogeneity modelling should be employed
- Statistical testing between models – Likelihood Ratio Tests
- Software packages and their features can help!

Bayesian Inferential Methods

- Allows for direct $\Pr(\text{Tree} \mid \text{Data})$, but requires a prior probability distribution over all of the possible tree topologies...
- Uses Markov Chain Monte Carlo (MCMC) search methods
 - Metropolis coupled MCMC
 - Stochastic algorithms which should (in theory) help avoid getting trapped in sub-optimal solutions
- Can sometimes be faster than ML, but convergence is never certain

Optimizations

- Heuristics and Approximations
- Hybrids and Combinatorial approaches
 - FastTree
- Technical optimizations

Evolutionary processes leading to non-tree-like structures and Phylogenetic conflicts

- Horizontal Gene Transfer
 - Transformation
 - Conjugation
 - Transduction
 - Gene transfer agents (ie: defective prophages)
- Gene Duplication and Loss
- Transposition of Mobile Genetic Elements
- What now???