RAD MICROBES

BOOT CAMP ON READING, ASSEMBLING, ANALYZING & DECHIPHERING MICROBIAL GENOMES

APRIL 28-29, 2025 | BRC AT RICE UNIVERSITY | HOUSTON, TX

Phylogenetics

An Dinh

UTHealth School of Public Health

Houston Methodist Research Institute

Goals

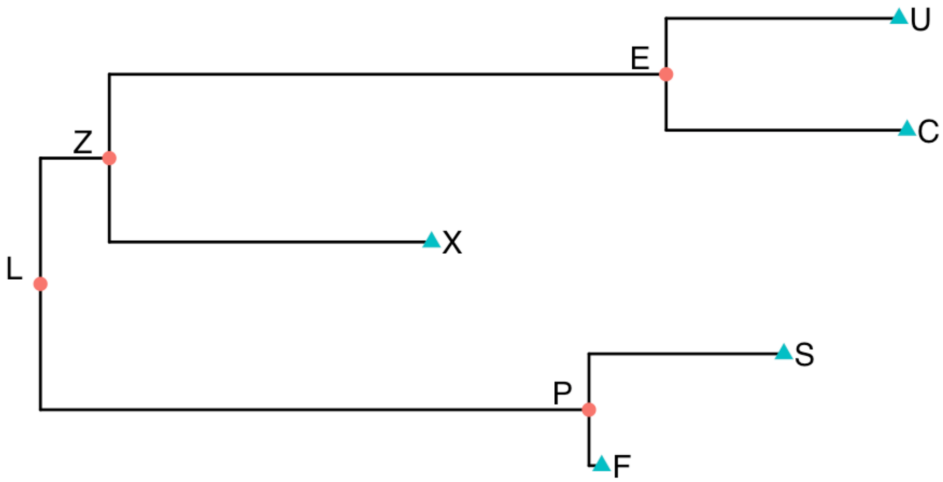
- Introduction to methods and models for phylogenetic inference
- Common language
- Demystification

Outline

- Tree basics
- Building a tree
- Distance matrix methods
- Character-based methods
  - Maximum parsimony
  - Maximum Likelihood
    - Heuristics
    - bootstrapping
  - Bayesian Inference
- Other complexities
- Takeaway

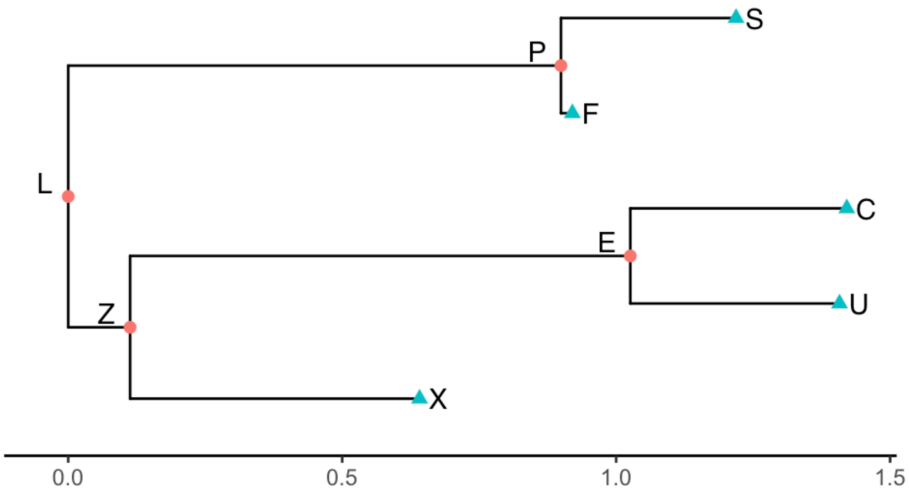
Dissecting a Tree

- 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.
- 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 Hypothetica



Unrooted vs. Rooted

- Unrooted shows evolutionary *relationships*

- Rooted shows evolutionary *history*
  - Allows for directionality
  - Time runs from root to tip
  - Root = MRCA of all the taxa in a tree

- Rooting a tree
  - **Outgroup** – common
  - Time-irreversible models (will cover briefly later)

Generalized workflow to build a tree

- **Data generation & QA/QC** – Session 1 & 2
- **Multiple Sequence Alignment** – Session 2 & 3
- **Tree estimation**
- Garbage in, garbage out!

Constructing Phylogenetic Trees

- Distance Matrix-based
  - Algorithmic
- Character-based
  - Estimate the optimum tree out of a set of trees

Distance matrix-based methods

- Start with an MSA, compute All-to-All distance matrix…

- Unweighted Pair-Group Method with Arithmetic Mean **(UPGMA)**
  - Mutational rate is uniform
  - Branch lengths are equally split
  - **Fitch-Margoliash (FM)** Method modifies this to do least-squares estimate of branch length

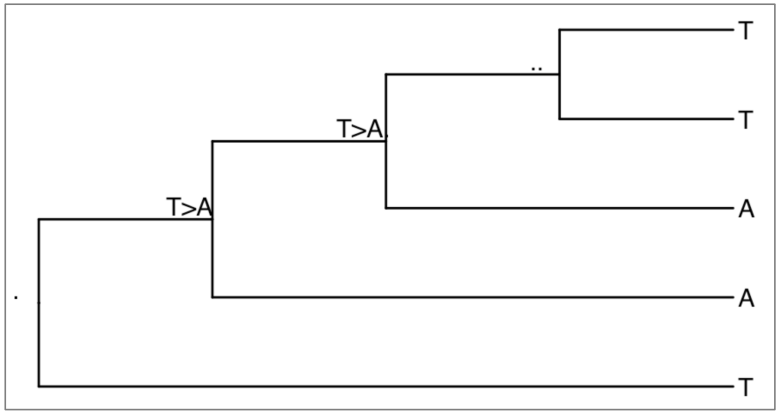
- **Neighbor-Joining (NJ)** - Saitou and Nei, 1987
  - Pairwise distance matrix → overall divergence → *adjusted* distance matrix
  - Not a bad heuristic and sometimes useful for initial starting tree

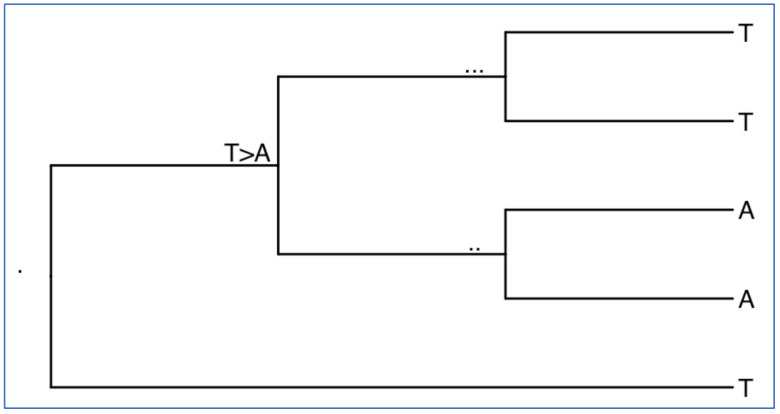
Constructing Phylogenetic Trees – **character-based**

- Maximum Parsimony
- Maximum Likelihood
- Bayesian Inference

Maximum Parsimony Positively misleading

- 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

- **max(Pr(Data|Tree)),** given:
  - A sample set of sequences, and
  - A nucleotide substitution (evolutionary) model

- "Tree" is a model consisting of:
  - Topology (τ)
  - Branch lengths (ν)
  - Substitution model parameters (θ)

- Computationally intensive, will illustrate shortly…

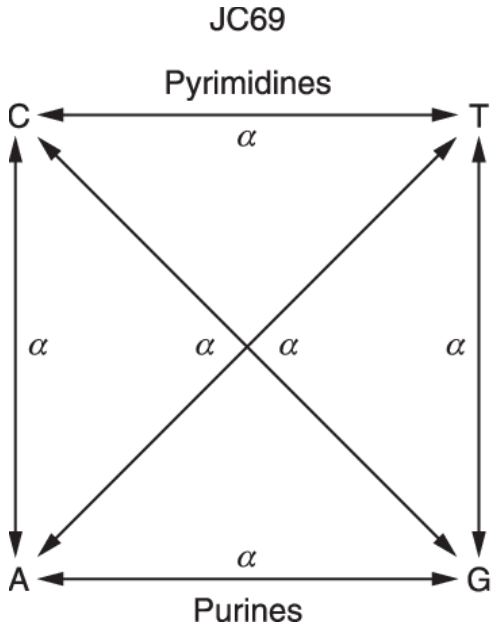
Nucleotide substitution models

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

Basic Markov chain models for substitution

- **Jukes-Cantor** (1969)
  - Rate of nucleotide substitution is the same for all pairs

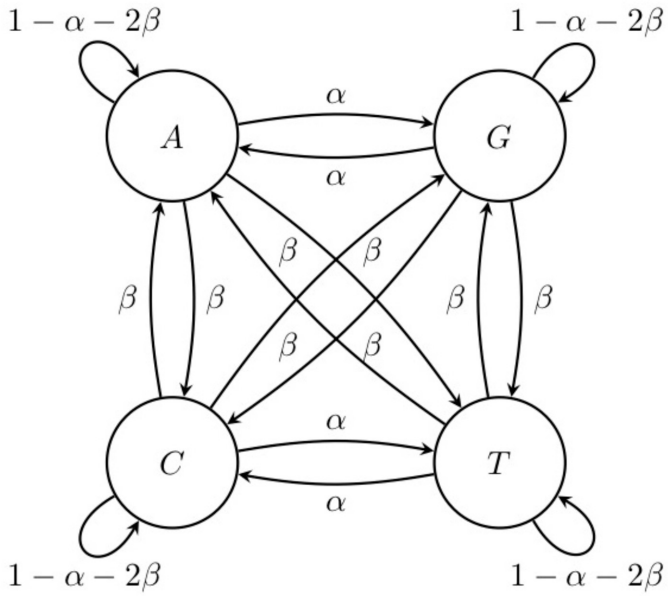
- See also: **Felsenstein** (1981)
  - Like Jukes-Cantor (equal substitution rates), but with unequal base frequency



Egan, Ashley & Crandall, Keith. (2006). Theory of Phylogenetic Estimation. 10.1093/oso/9780195168174.003.0028.

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."

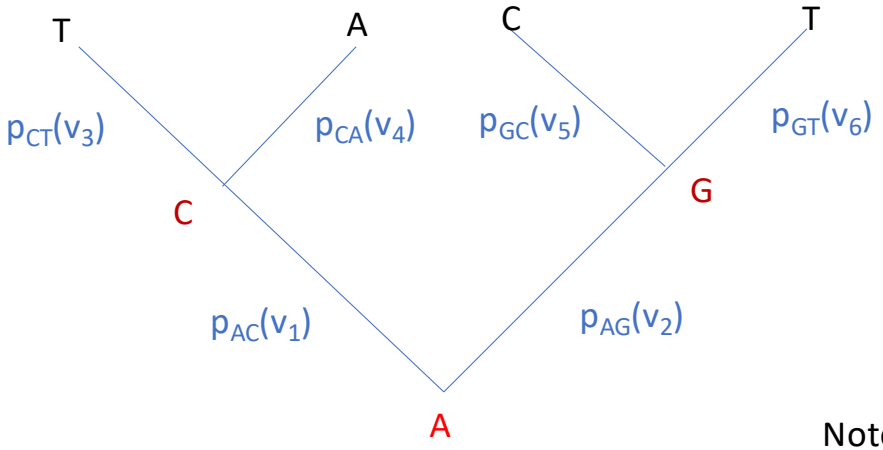
$$Q = \begin{pmatrix} -(a\pi_C + b\pi_G + c\pi_T) & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & -(a\pi_A + d\pi_G + e\pi_T) & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & -(b\pi_A + d\pi_C + f\pi_T) & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & -(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

Model modifications

- **+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
- **+ I** (invariant sites)

Computing the likelihood

- Given a possible tree topology…
  - At a *single* site i…
    - Assume we "know" interior nodes; leaves are observed



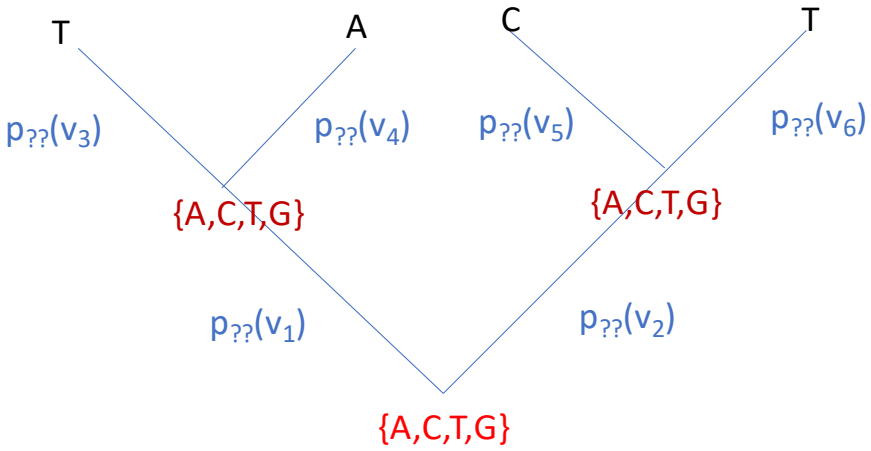
Multiply

$$L(T|D) = \Pr(D|T) = \pi_A \cdot p_{AC}\ (v_1) \cdot p_{AG}\ (v_2) \cdot p_{CT}\ (v_3) \cdot p_{CA}\ (v_4) \cdot p_{GC}\ (v_5) \cdot p_{GT}\ (v_6)$$

Note: we don't actually know what the branch lengths are yet,
And we may not have substitution parameters estimated yet either…

Computing the likelihood (2)

- Given a possible tree topology…
  - At a *single* site i…
    - BUT we do NOT know what the interior nodes are… they could be anything.



$$L(T|D) = \sum\sum\sum L(all\ possible\ inner\ node\ assignments)$$

Sum over

$$4\times4\times4 = 64\ combinations$$

**Fortunately, Felsenstein's Pruning Algorithm brings complexity down to $O(nsc^2)$

But then consider…

- Number of sites

- How many possible trees:
  - $N_{unrooted} = (2n - 5)!!$
  - $N_{rooted} = (2n - 3)!!$

  - Let n = 10 sequences
    - $N_{unrooted} = (2n - 5)!! = (15)!! = 15 * 13 * 11 * 9 * 7 * 5 * 3 * 1 = 2027025$
    - $N_{rooted} = 3.45\text{e}7$

- Exhaustive search infeasible; ML criterion is *NP-hard*

Heuristic (uphill) search algorithms for tractability

- 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) – **RAxML**
- Tree bisection and reconnection (TBR)

Bootstrapping for Branch Support and reliability

- In general, bootstrapping = random subsampling *with replacement* (resampling)
- Estimates the statistical error when sampling distribution is unknown

- In phylogenetic context, estimates the reliability (consistency) of the resultant (majority-rule) **consensus tree**
- Bootstrapped sites will be randomly shuffled, length same as original MSA.

- The final resultant tree might not be the overall maximum-likelihood tree

Bayesian Inferential Methods

- $\Pr(\boldsymbol{Tree}|\boldsymbol{Data}) = \frac{\boldsymbol{Pr}(\boldsymbol{Data}|\boldsymbol{Tree})\cdot\boldsymbol{P}(\boldsymbol{Tree})}{\boldsymbol{P}(\boldsymbol{Data})}$
  - $Pr(D|T)$ = likelihood, but not the same as in ML method due to handling of 'nuisance parameters'
  - $\Pr(T)$ = prior probability distribution: typically assume uniform initially, ie: all trees are equally likely
  - $Pr(D)$ = a normalizing constant, $= \sum(P(D|T)\ x\ P(T_i))$

- Can sometimes be faster than ML but convergence is never certain

Bayesian Inferential Methods

- 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

  - Start with a random tree
  - Attempt a random walk
    - If the next tree is better than the one you're at, move there.
    - If the next tree is not better, then *maybe* move there, depending on how much worse it is.

More complexities to address mutational heterogeneity

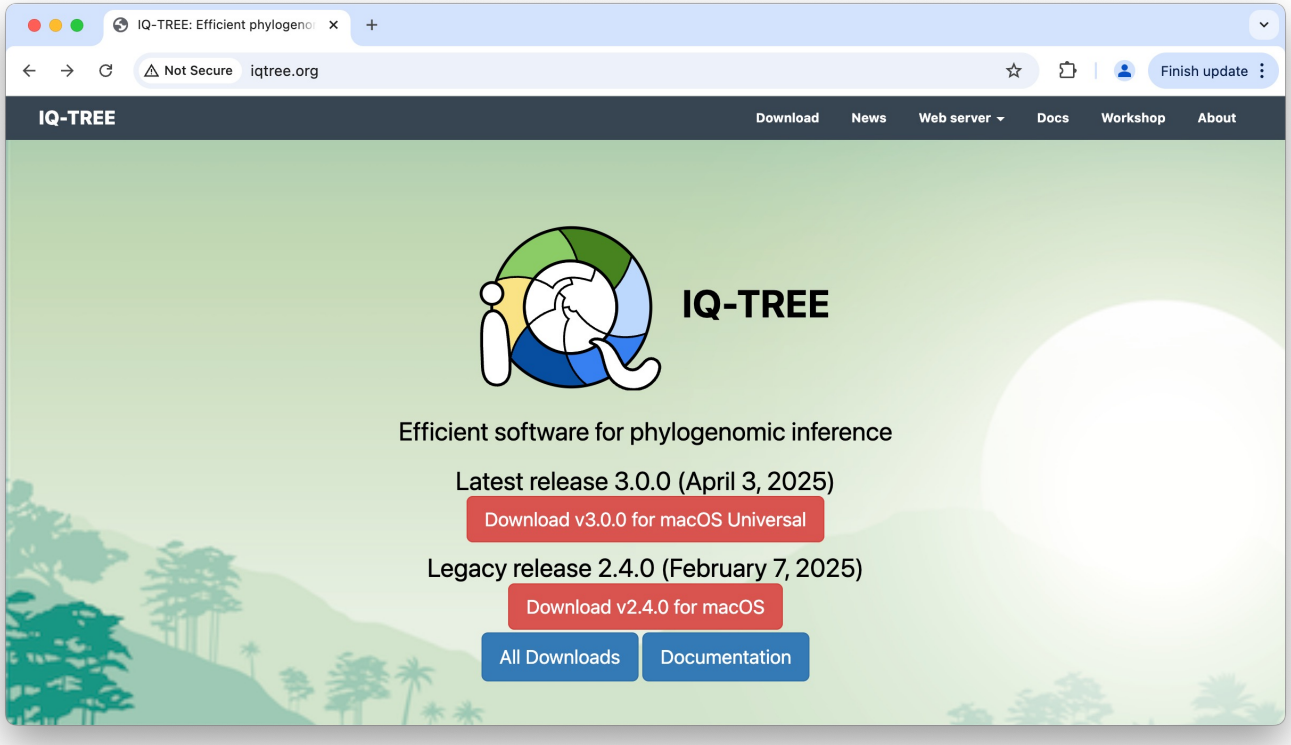
- **Partitioning** (categorization)
  - Apply specific models for a site or regions along a sequence
- **Mixture** models
  - Assign a model (out of a user-specified selection) on a per-site basis
- Some caution may need to be exercised to avoid over-parameterization

Non-reversible models

- **UNREST** (Yang 1994)
  - Completely unrestrained
- "Rootstrap" = bootstrap + root position support
- Relaxes the constant-rate assumption for molecular clock modelling

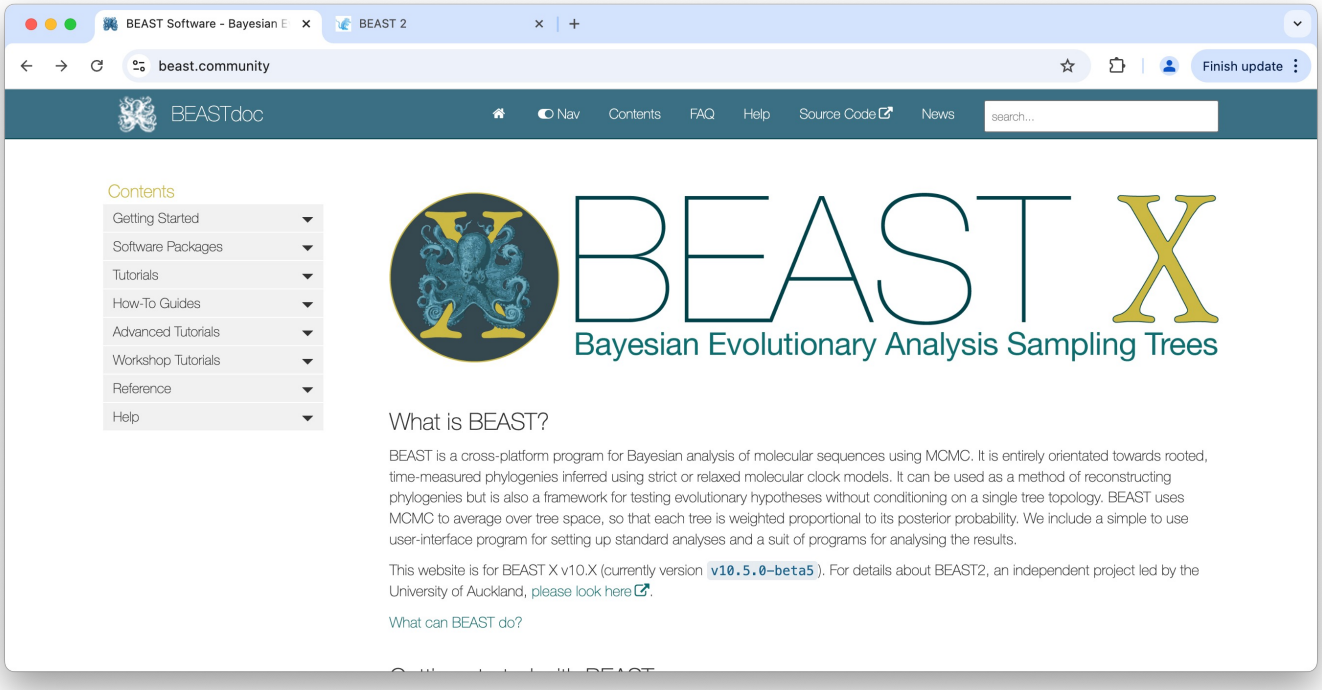
$$\mathbf{Q} = \begin{bmatrix} \cdot & q_{TC} & q_{TA} & q_{TG} \\ q_{CT} & \cdot & q_{CA} & q_{CG} \\ q_{AT} & q_{AC} & \cdot & q_{AG} \\ q_{GT} & q_{GC} & q_{GA} & \cdot \end{bmatrix}$$

IQ-TREE



- "combine elements of:
  - hill-climbing algorithms (NNI)
  - random perturbation of current best (locally optimal) trees,
  - and a broad sampling of initial starting trees"
- Ultrafast bootstrapping
- Feature rich
- Good Documentation

BEAST1/BEASTX ; BEAST2



- BEAST2 more modular, packages include MSC

Takeaways

- Balancing between robustness and efficiency
  - 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

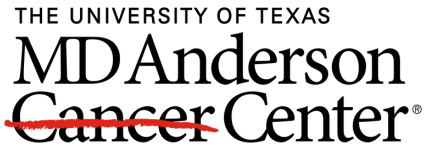
Evolutionary processes leading to non-tree-like structures and Phylogenetic conflicts

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

- What now???
  - Networks?


RICE KEN KENNEDY INSTITUTE
Responsible AI and Computing for Global Impact


RICE ENGINEERING AND COMPUTING
Department of Computer Science


THE UNIVERSITY OF TEXAS
MD Anderson
Cancer Center


PacBio


UTHealth
Houston


HOUSTON
Methodist
LEADING MEDICINE


Baylor
College of
Medicine