Stochastic Models for Horizontal Gene Transfer

Dajiang Liu
Department of Statistics

Main Reference

- Marc A. Suchard:
  - Stochastic Models for Horizontal Gene Transfer: Taking a Random Walk through Tree Space *Genetics 2005*

Introduction

- Horizontal gene transfer (HGT) plays a critical role in all domains of life,
  - in particular for prokaryotes
    - Prokaryotes agile at adapting to new environment
      - Ability obtained more from HGT than mutation

Three Mechanisms for HGT

- Transformation in which free DNA sequences absorbed from the environment
- Conjugation between two different prokaryotic species
- Transduction of genetic material through viruses
- Example:
  - HGT among bacterial pathogens of antibiotic genes -> multi-drug resistant bacteria

Three General Methods for Examining HGT

- I. Single genome approach:
  - Examine the nucleotide base composition, and codon usage patterns
- Comparative Studies across species:
  - II. Similarity approaches using gene content, propose average genome and species trees
  - III. **Using phylogenetic reconstruction using orthologous genes**
    - **Compare species tree and genes tree**
      - **HGT offers a possible explanation of inconsistency**

Difficulties for Using Phylogenetic Tree

- The true species tree is usually unknown:
- Two solutions:
  - Fix the species tree using specific gene:
    - E.g. 16S rRNA
  - Simultaneously estimate the species tree and gene trees given a biological model relating them

A Set of Hierarchical Models

- Bayesian model is used to reconstruct each gene tree from its multiple alignment
- Conditioning on the genes trees, HGT models imposes a second probabilistic models.
- Jointly models
  - the gene trees given the unknown species tree, and
  - An unknown number of HGT leading from that unknown species tree to each gene tree

With-in Gene Reconstruction Model I

- Data:

$$Y = (Y_1, \cdots Y_k)$$

- Y_i represent aligned sequences of length L_k for a specific genes
- Y_i could be further divided into ordered homologues sites

$$Y_{il} = (Y_{il1}, \cdots Y_{ilN})$$

- Sites with in a partition is independently and identically distributed Y_{il}
  - The sites are jointly distributed as a multinomial with 4^N outcomes
  - The probability of each outcome depends on τ_k
    - the unknown tree
    - Branch length
    - Nucleotide mutation rate

With-in Gene Reconstruction Model II

- Nucleotide substitution model:
  - Transition:Transversion rate ratio:
    - α_k between A and G
    - γ_k between C and T
  - Stationary distribution

$$\pi_k = \left(\pi_{kA}, \pi_{kG}, \pi_{kC}, \pi_{kT}\right)$$

  - Tree branch length follows exponential distribution $t_k \sim Exp(\mu_k)$

Across Gene Hierarchical Model

- Joint distribution on the parameters

$$\begin{pmatrix} \log \alpha_k \\ \log \gamma_k \\ \log \mu_k \end{pmatrix} \sim \mathrm{N}(V, \Sigma)$$

$$\pi_k \sim Dirichlet(\mathrm{N}_\Pi \times \Pi)$$

- Parameters in different partitions share the same hyper-parameter
  - Enables simultaneous estimation across partitions
  - Non-informative priors are used on V, Σ, Π

Horizontal Gene Transfer Model

- Tree space:
  - There are M different tree topology relating N extant taxa

$$M = (2N-5)!/2^{N-3}(N-3)!$$

  - Graph constructed out of the tree space

$$G = (\upsilon, \varepsilon)$$

    - Degrees of vertex $d(v)$
    - Size of neighborhood

$$\left|\Gamma(v)\right| = d(v)$$

    - Simple graph
      - Pairs of vertices are connected by a single edge
      - No vertex is connected with itself by a looping edge

Two Models:

- Discrete Time Markov Chain(DTMC)
  - E.g. Subtree-Pruning-Regrafting Model (SPR)

- Continuous Time Markov Chain(CTMC)
  - Complete graph:
    - E.g. Generalized Jukes-Cantor model(GJC)
      Generalized Kimura model(GK)

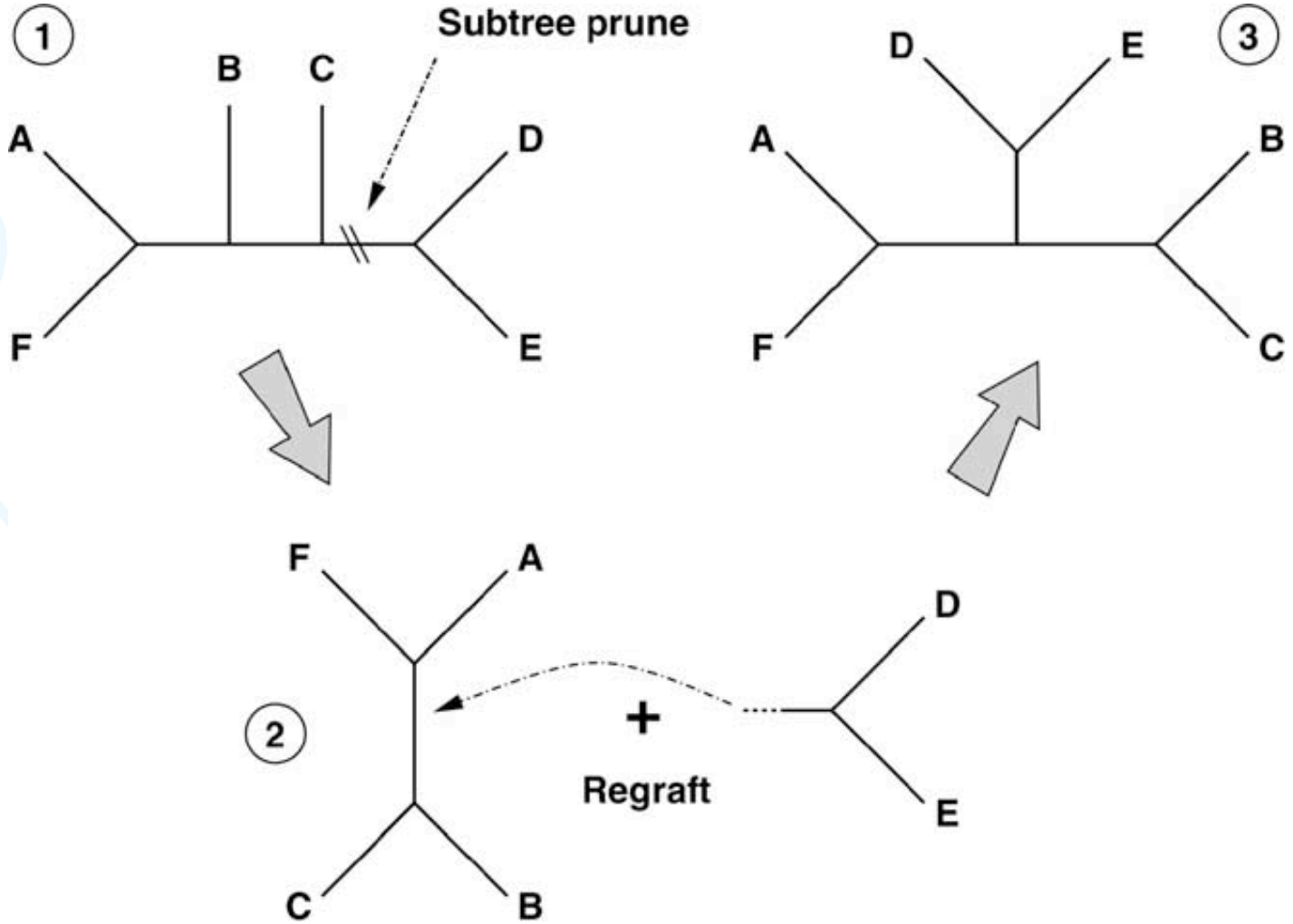
Discrete Time Markov Chain

Subtree-Pruning-Regrafting Model

How to define the neighbor

- Each one of M different tree topologies will be a node of the tree
- The neighbors of each possible tree stems from subtree transfer operator
  - The collection of trees one operation away from node v becomes its neighborhood $\Gamma(v)$
  - One such operator is subtree-prune-regraft (SPR) operator

Subtree-Prune-Regraph-based model



Stochastic Process on the Simple Graph I

- Un-weighted random walk on the simple graph
  - Given current node, the random walk will choose one of its neighbors uniformly with transition matrix given by

$$(A)_{uv} = \begin{cases} \frac{1}{d(u)} & \text{if u and v are neighbors} \\ 0 & \text{otherwise} \end{cases}$$

- Assume:
  - Unknown species tree Y exists, and is taken as the initial state of the random walk
  - Each gene tree τ_k is taken as the final state of k^{th} chain
  - The chain k has E_k transitions
  - Each chain is independent conditional on Y and A
    - Horizontal gene transfer for each gene tree is independent

Stochastic Process on the Simple Graph II

- The probability of path k is obtained through iterating matrix A

$$q(\tau_k = v \mid Y = u, E_k) = (A^{E_k})_{uv}$$

- Choice of priors

$$Y \sim Multi(z)$$

  - One possible choice of z is each gene tree (node) has equal probability of being the species tree $E_k \sim Poisson(\Lambda_k)$
  - Expected number of HGT for gene k follows a Poisson distribution with parameter Λ_k

Stochastic Process on the Simple Graph III

- The probability of each tree topology could be given by

$$q(\tau_k = v \mid Y = u, \Lambda_k) = \sum q(\tau_k = v \mid Y = u, E_k) q(E_k \mid \Lambda_k)$$

- The nuisance parameter E_k could be integrated out

- Multi-step transition probability matrix could be given by

$$P = \sum A^{E_k} q(E_k \mid \Lambda_k) = \sum A^{E_k} \exp(-\Lambda_k) \frac{\Lambda_k^{E_k}}{E_k!} = \exp(\Lambda_k (A - I))$$

Continuous Time Markov Chain

- DTMC computationally hard
- Some CTMC models have analytical solutions
- Random walk on a complete graph
  - Generalized Jukes Cantor model GJC
  - Kimura model GC
    - Allows rate of HGT differs between groups of taxa

Modeling Differences across Gene Classes

- Differences in the number of HGT across K genes could be modeled
  - E.g If there are C=2 gene classes

$$\Lambda_k = \begin{cases} \exp(\lambda_1) & \text{if gene k is in class 1} \\ \exp(\lambda_1 + \lambda_2) & \text{if gene k is in class 2} \end{cases}$$

Statistical Framework

- Use Bayes factor:
  - Bayesian version of LRT

$$BF = \frac{m(Y \mid M_1)}{m(Y \mid M_0)} = \frac{p(M_1 \mid Y)P(Y)/P(M_1)}{p(M_0 \mid Y)P(Y)/q(M_0)} = \frac{p(M_1 \mid Y)/P(M_1)}{p(M_0 \mid Y)/q(M_0)}$$

- Example:
  - Want to test if the mean number of HGTs in different genes differ:

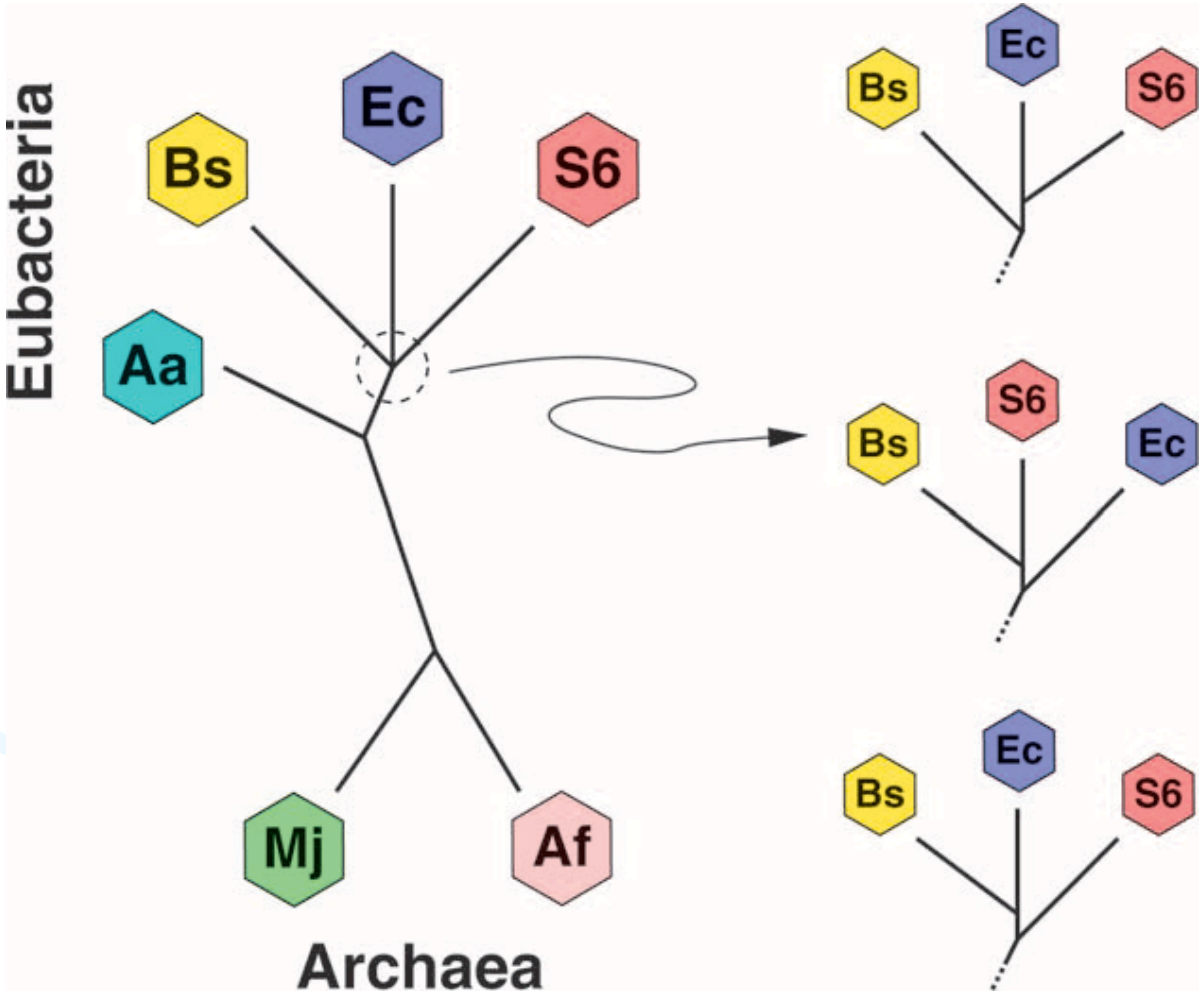
$$M_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_k$$

$$M_1 : \text{Unconstained model}$$

Example

- Data set:
  - 144 separate gene alignment
    - Each alignment consists of a single gene from six prokaryotes
    - For comparison reason,
      - Assume two different gene class
      - Exclude third codon position

Species Tree



Applications:

- Selection of Stochastic Model:
  - Compare $BF_{SPR,GJC}$, $BF_{GK,GJC}$
- Estimating the species tree
  - Compare $BF_{BS\text{-}S6}$, $BF_{GK,GJC}$
- Estimates of evolutionary pressure
- Tests of varying rates of HGT

Limitations

- The model does not consider the changes in branch length
  - All trees with the same topology are considered the same
    - All HGT between nearest neighbors are unidentified in the model
  - Possible solution:
    - Incorporate tree branch length into the model
- The conditional independence assumption of discrete time MC model is unrealistic
  - It means the evolution of all genes are independent given the species tree Y
  - Possible solutions:
    - Allow for linked genes

Advantages

- Jointly model species tree topology and HGT
  - Avoided using a fixed tree topology while knowledge about it is unclear
- A probabilistic model avoids the downward bias in the parsimonious approach
- Allows for formal statistic testing framework through Bayes factors