Summer Institute - Phylogenetics

K. S. Dorman

Department of Statistics
Department of Genetics, Development & Cell Biology
Iowa State University

Computational and Systems Biology Summer Institute
2010

Outline

1 The Data

2 Phylogenetic Methods
Likelihood-Based Methods

Notation

Quantity	Meaning
N	number of sequences
L	number of characters
τ	topology
$b = (b_1, b_2, ..., b_{2N-3})$	branchlengths
$T = (\tau, b)$	tree or phylogeny

Data

		L Positions												
		1	2	3	4	5	6	7	8	9	10	11	12	$\cdots$
	1	A	A	C	T	C	C	G	C	G	A	T	A	$\cdots$
N	2	A	A	T	T	C	C	G	G	G	A	T	A	$\cdots$
Taxa	3	C	A	C	T	C	C	G	C	G	A	T	A	$\cdots$
	4	A	A	T	T	C	C	G	C	G	A	T	A	$\cdots$
	5	A	A	T	T	C	C	C	G	G	A	T	A	$\cdots$

Data (Mathematical): The data is summarized as a matrix

$$X = \{X_{ij} : i = 1, 2, \ldots, N; j = 1, 2, \ldots, L; X_{ij} \in \{0, 1, 2, 3\}\},$$

where X_{ij} is the random variable indicating the nucleotide in the jth aligned position of the ith sequence.

Outline

1 The Data

2 Phylogenetic Methods
Likelihood-Based Methods

Maximum Likelihood

- Establish a precise model of evolution that includes *evolutionary parameters* θ and a good dose of assumptions.
- Write likelihood of the data, which is a function of the evolutionary parameters and the phylogenetic tree.

$$L(\theta, \tau, b \mid X)$$

- **Optimality criterion:** Maximize the likelihood over parameters. Our inferred tree is $\hat{T} = (\hat{\tau}, \hat{b})$.

An Evolutionary Model

- Evolution is random and “memoryless”.
- Evolution is reversible.
- *Organisms are related by a tree-like structure.
- Evolution of two sister lineages is independent conditional on their ancestor.
- *Sites are independently and identically distributed (iid).
    - No site-to-site differences in evolution.
    - There is no covariation between sites.

*Assumptions that have been relaxed in more complex models.

A Probability Model : "CTMC"

The continuous time Markov chain model is useful for modeling a random variable $X(t)$ that varies *continuously* in time. Let $X(t)$ be the random variable (nucleotide indicator) at time t at some position in a genome being passed through the generations. The *state* of the chain at time t is an element in the *state space* $S = \{A, C, G, T\}$.

Initial State Distribution

$$P(X_{\text{root},i} = b) := \pi_b, \qquad b \in \{0, 1, 2, 3\}.$$

Transition Probability Matrix

$$\begin{aligned} P(t) &= \{P(X(t) = b \mid X(0) = a) : a, b \in \{0, 1, 2, 3\}\} \\ &= \begin{pmatrix} - & p_{01}(t) & p_{02}(t) & \cdots \\ p_{10}(t) & - & p_{12}(t) & \cdots \\ \vdots & & \ddots & \end{pmatrix} \end{aligned}$$

Jukes-Cantor CTMC

Remapping random variable values $0, 1, 2, 3$ back to more convenient A, C, G, T...

- Assumes all nucleotides are equally likely, i.e. $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$.
- Assumes all mutation rates are equal, so $p_{ab}(t) = p_{cd}(t)$ for all $a \neq b, c \neq d \in \{A, G, C, T\}$.
- There is one free parameter left, call it μ, which is the rate of mutation, any mutation.

Confounded: Mutation Rate and Time

- **Intuition:** There are two ways a branch can be long. Either (1) a lot of time has passed or (2) a lot of mutation happened along that branch.
- **Mathematically:** Time t and mutation rate μ are confounded (functions of each other).
- These two parameters are constrained such that the branch lengths b indicate the expected number of mutations per site along the given branch.
- For example, if $b_3 = 0.1$, then we expect 0.1 mutations per site along the third branch, or *approximately* 1 in 10 sites will have mutated once along this branch.

Transition Probabilities for JC

The Jukes-Cantor(JC) model has simple substitution probabilities (CTMC transition probabilities):

$$p_{ab}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & (a = b) \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & (a \neq b) \end{cases}$$

Note: Only the parameter μt can be estimated, not μ and t separately. In this case $\mu = \frac{1}{3}$ when time t is defined as on the previous slide.

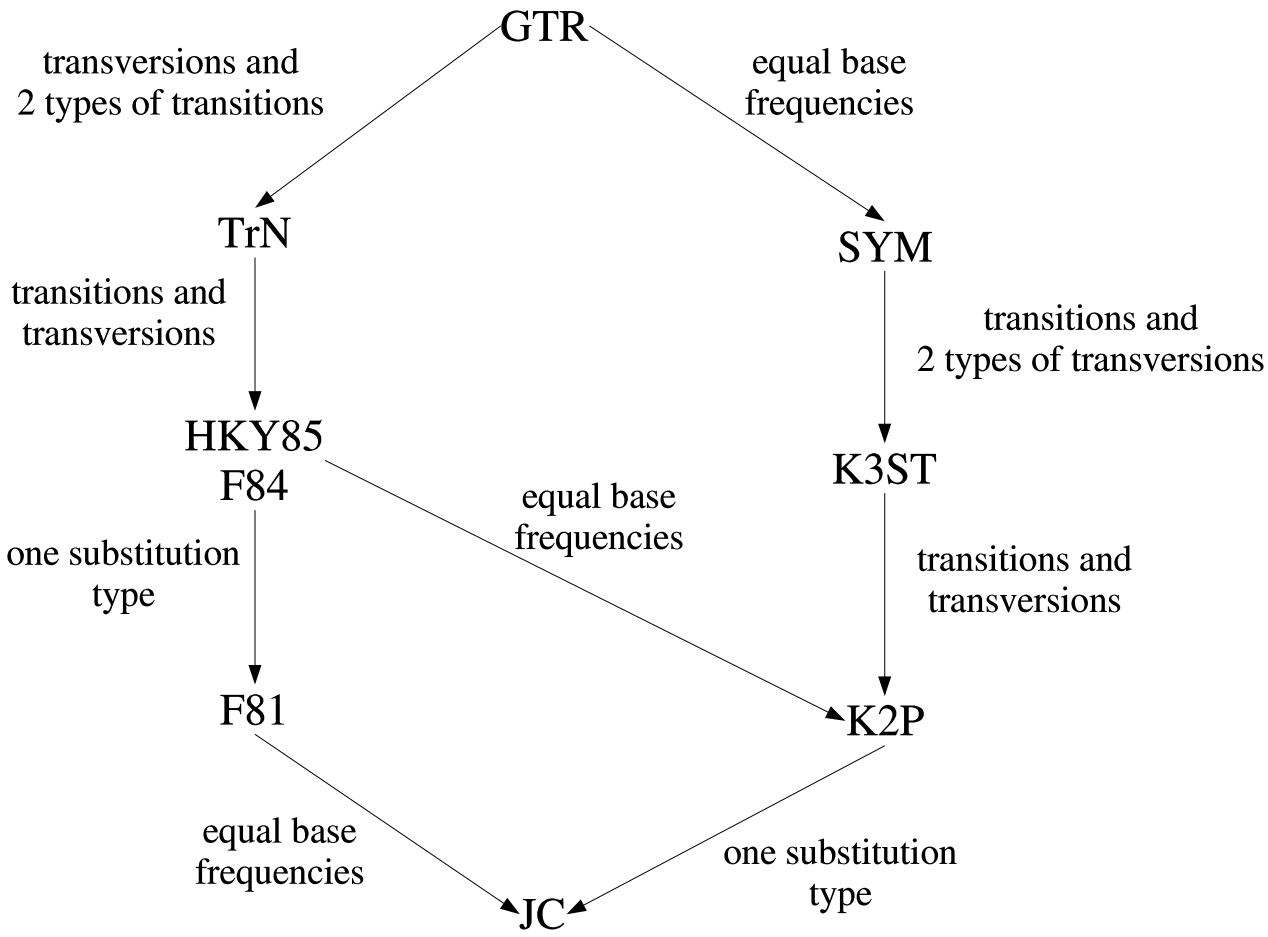
Transition Probabilities for F81

The F81 model assumes equal rates but allows for different equilibrium nucleotide frequencies.

$$p_{ab}(t) = \begin{cases} \pi_b + (1 - \pi_b)e^{-\mu t} & (a = b) \\ \pi_b(1 - e^{-\mu t}) & (a \neq b) \end{cases}$$

where again μ is fixed and t is estimated.

Evolutionary Models



From Chapter 11 of *Molecular Systematics*.

Modeling Other Kinds of Sequences

The only difference is the state space and, most notably, its size.

- **Protein:** 20 amino acids so $s = 20$ and up to $20 * 19/2 = 190$ free parameters!
- **Codons:** 64 possible codons with 3 stop codons so $s = 61$ and up to 1830 free parameters!

The Likelihood

- We seek to obtain an expression for the likelihood function $L(\tau, b, \theta \mid X)$.
- Let $X_{\cdot i}$ represent the ith column of data. Sites are independent, so $P(X) = P(X_{\cdot 1})\, P(X_{\cdot 2}) \cdots P(X_{\cdot L})$. Then,

$$L(\tau, b, \theta \mid X) = \prod_{i=1}^{L} L_i(\tau, b, \theta \mid X_{\cdot i}).$$

- It helps a great deal to use the law of total probability and consider the ancestral nucleotides at the internal nodes as well.

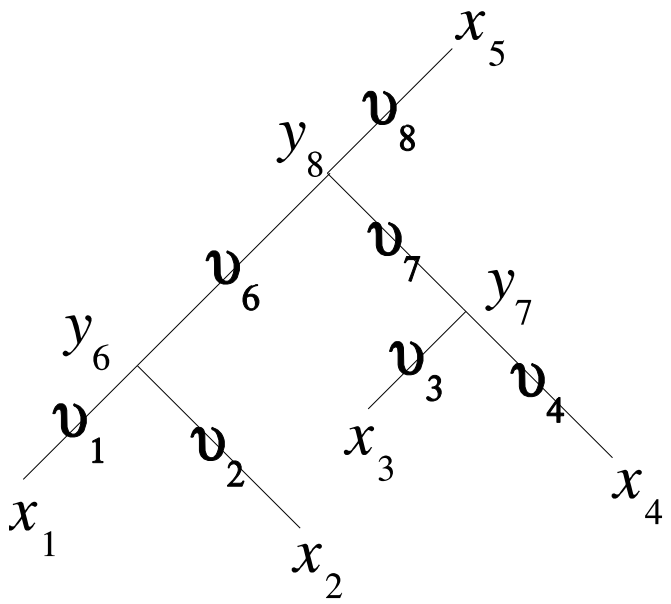
Applying the LTP

- The Law of Total Probability states that $P(B) = P(BA_1) + \cdots + P(BA_n)$ where A_i are nonempty, mutually exclusive events with $P(A_1 \cup \cdots \cup A_n) = 1$.
- X represents the observed data of the *extant* taxa. Let Y represent the unobserved data of the *ancestral* taxa at the internal nodes of the tree.
- Apply LTP: $P(X) = \sum_Y P(XY)$.

Temporarily Simplify Notation

- Let the observed data at a site i be $x_{\cdot i} = (x_{1i}, x_{2i}, \ldots, x_{Ni}) = (x_1, x_2, \ldots, x_N)$.
- Let the unobserved ancestral data at this site be $y = (y_{N+1}, \ldots, y_{2N-2})$.
- Let $N = 5$, so there are 5 elements in vector $x_{\cdot i} = (x_1, x_2, \ldots, x_5)$ and 3 elements in the ancestor vector $y = (y_6, y_7, y_8)$.

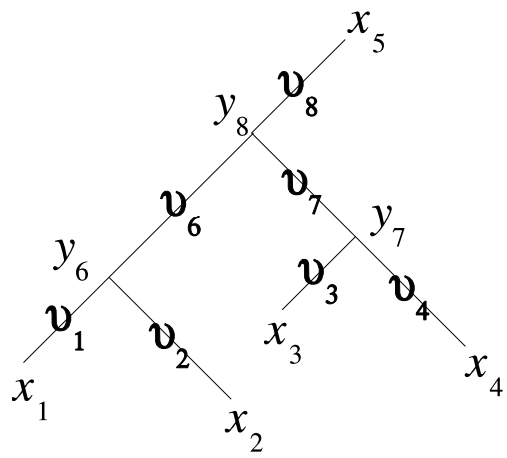
The Ancestors



Conditional on y_6, what is the probability of (x_1, x_2)?

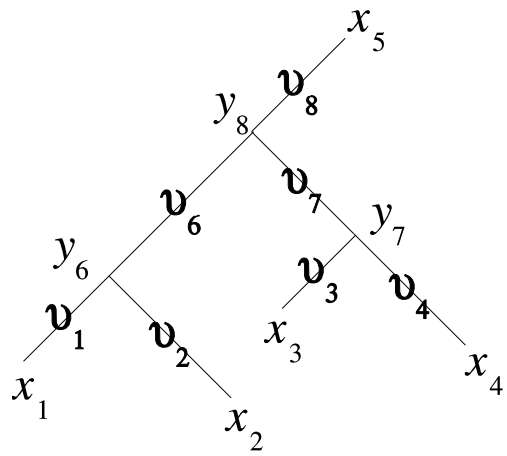
$$P(x_1, x_2 \mid y_6) = ?$$

Data and Ancestor Likelihood



$$P(x_1, x_2, x_3, x_4, x_5, y_6, y_7, y_8) =$$
$$P(x_5, y_8, y_6, y_7, x_1, x_2, x_3, x_4) =$$
$$P(x_5)P(y_8 \mid x_5)P(y_6 \mid y_8)P(y_6 \mid y_8)$$
$$\times P(x_1 \mid y_6)P(x_2 \mid y_6)P(x_3 \mid y_7)P(x_4 \mid y_7) =$$
$$\pi_{x_5} p_{x_5 y_8}(b_8) p_{y_8 y_6}(b_6) p_{y_8 y_7}(b_7) p_{y_6 x_1}(b_1)$$
$$\times p_{y_6 x_2}(b_2) p_{y_7 x_3}(b_3) p_{y_7 x_4}(b_4)$$

Data Likelihood



$$P(x_1, x_2, x_3, x_4, x_5) = \sum_{y_6, y_7, y_8} P(x_1, x_2, x_3, x_4, x_5, y_6, y_7, y_8)$$

ML Issues: Choosing a Model

- You have two alternative models M_1 with parameters θ_1 and M_2 with parameter θ_2.
- Test using the Likelihood Ratio Test Statistic

$$\Lambda = \frac{L(M_1, \hat{\theta}_1 \mid X)}{L(M_2\, \hat{\theta}_2 \mid X)}$$

  and compute $-2 \log \Lambda$.
- If $M_1 \subset M_2$ (meaning M_1 is a nested within and a simplification of M_2), then $-2 \log \Lambda$ is approximately distributed as a Chi-squared random variable with degrees of freedom equal to the difference in the number of free parameters between M_1 and M_2.

ML Issues II

- **Calculating likelihoods:** A sum over all possible ancestor states is a very long sum. Fortunately algorithms exist to speed up the process.
- **Maximizing the likelihood:** Big problem.
  - Select a topology τ.
  - Assuming τ, maximize (using numerical algorithms) the likelihood for (ν, θ).
  - Iterate through your search algorithm.

Bayesian Methods

- Bayes Rule: For any events A and B with positive probability

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- Bayes Rule in evolution context:

$$P(\tau, b, \theta \mid X) = \frac{P(X \mid \tau, b, \theta)P(\tau, b, \theta)}{P(X)}$$

- Optimality criterion: maximize the posterior probability.

Modeling Issues

It is necessary to specify a so-called prior $P(\tau, b, \theta)$ for all model parameters.

- $P(\tau)$ uniform over all possible topologies.
- $P(b) = P(b_1) \cdots P(b_{2N-3})$ with $P(b_i)$ uniform over large range (e.g. $[0, 1000]$) or exponential.
- $P(\theta)$ usually uniform.

Computational Issues

- It is not possible to analytically obtain the posterior probability.
- Solution: simulation via Markov Chain Monte Carlo.
- Output: File of random parameter values sampled from the posterior distribution. For example...

```
iteration tree                                                  kap
1         ((1:0.15,2:0.21):0.08,3:0.06,4:0.10)                  2.3
2         ((1:0.13,3:0.06):0.19,2:0.23,4:0.09)                  2.2
.                          .                                    .
.                          .                                    .
.                          .                                    .
```

Advantages

- Computationally more efficient than ML for large problems and complex models.
- Can integrate over nuisance parameters:

$$\begin{aligned} P(\tau, b \mid X) &= \int_{\theta} P(\tau, b, \theta \mid X) \\ &= \sum_{\theta_i} P(\tau, b, \theta_i \mid X) \end{aligned}$$

- Can even integrate over the model if you don't want to commit to a model.
- The MCMC sample provides a complete summary of the inference, including uncertainty and confidence sets.

Disadvantages

- It is a parametric method; you must intelligently choose a model.
- It is computationally intensive, like maximum likelihood.
- You must specify a prior. Sometimes you have no idea which prior to use. And the choice of prior could impact the results.
- Many difficulties encountered when using MCMC. New worries.