

Summer Institute - Phylogenetics

K. S. Dorman

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

Computational and Systems Biology Summer Institute
2009

Outline

1 The Data

2 Phylogenetic Methods Likelihood-Based Methods

Notation

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

L Positions

		1	2	3	4	5	6	7	8	9	10	11	12	...
<i>N</i> Taxa	1	A	A	C	T	C	C	G	C	G	A	T	A	...
	2	A	A	T	T	C	C	G	G	G	A	T	A	...
	3	C	A	C	T	C	C	G	C	G	A	T	A	...
	4	A	A	T	T	C	C	G	C	G	A	T	A	...
	5	A	A	T	T	C	C	C	G	G	A	T	A	...

Data (Mathematical)

The data is summarized as a matrix

$$X = \{x_{ij}, i = 1, 2, \dots, N, j = 1, 2, \dots, L\},$$

where x_{ij} is the nucleotide in the j th aligned position of the i th 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.

Ubiquitous CTMC Assumptions

We shall model evolution using a continuous time Markov chain model (CTMC). We will describe a number of different models, but certain assumptions will rein throughout.

- *The CTMC is time homogeneous.
- Evolution is at equilibrium.

Markov Chain Equilibrium

- A Markov chain (discrete or continuous) that runs long enough will reach a kind of equilibrium.

... ACCATGGAGGTGGTACGGGATTTCAAGAGAGA ...

- After equilibrium is achieved, sample a subsection of the chain and observe

$$P(X_t = A) = \pi_A$$

$$P(X_t = C) = \pi_C$$

$$P(X_t = G) = \pi_G$$

$$P(X_t = T) = \pi_T$$

Continuous Time Markov Process

The process $\{X(t), t \geq 0\}$ with countable state space $S = \{0, 1, 2, \dots\}$ is a continuous-time Markov chain (CTMC) if for all $0 \leq u < s, t \geq 0$ and states $i, j, x(u) \in S$

$$\begin{aligned} P[X(t+s) = j \mid X(s) = i, X(u) = x(u), 0 \leq u < s] \\ &= P[X(t+s) = j \mid X(s) = i] \\ &= p_{ij}(s, t+s) \end{aligned}$$

$p_{ij}(s, t+s)$ is the CTMC transition probability.
Time homogeneity yields

$$p_{ij}(s, t+s) = p_{ij}(t)$$

Specifying a CTMC

$$Q = \begin{pmatrix} - & q_{01} & q_{02} & \cdots \\ q_{10} & - & q_{12} & \cdots \\ \vdots & & \ddots & \end{pmatrix}$$

Instantaneous
transition matrix
($s(s - 1)$ free
parameters)

$$p_0(i) = P[X(0) = i]$$

Initial state
distribution
($s - 1$ free
parameters)

Transition Probability Matrix

$$P(t) = \begin{pmatrix} - & p_{01}(t) & p_{02}(t) & \cdots \\ p_{10}(t) & - & p_{12}(t) & \cdots \\ \vdots & & \ddots & \end{pmatrix}$$

$$P(t) = e^{Qt}$$

where Q is the instantaneous transition matrix.

Matrix Exponential

The matrix exponential is *defined as* the infinite sum

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots$$

In some special cases this infinite sum can be computed analytically. Otherwise, there are numerical approximations.

Evolutionary Instantaneous Rate Matrix

For nucleotide sequences,

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

Applying the Evolutionary Assumptions

- Equilibrium assumption

$$p_0(i) = \pi_i$$

No need to specify initial distribution.

- Time reversibility assumption

$$\pi_i q_{ij} = \pi_j q_{ji}$$

Leads to $s(s - 1)/2$ fewer parameters (6 maximally for nucleotide sequences)

Jukes-Cantor Evolutionary CTMC

- Assumes all nucleotides are equally likely, i.e.
 $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$.
- Assumes all rates $p_{ij}, i < j$ are equal.
- There is one free parameter left, call it μ , which is the rate of mutation, any mutation.

Confounding 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).
- Q matrices are parameterized 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):

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & (i = j) \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & (i \neq j) \end{cases}$$

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

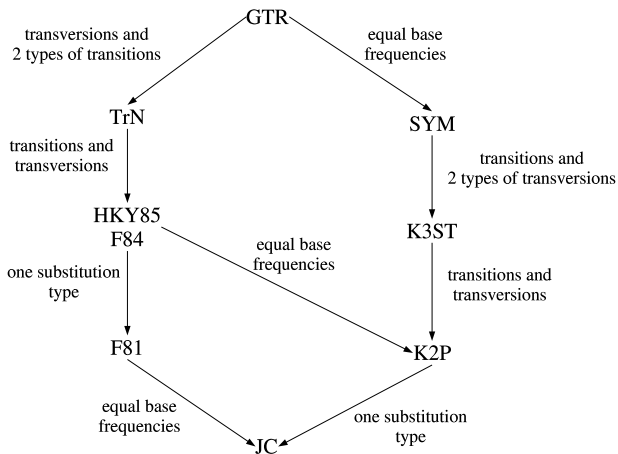
Transition Probabilities for F81

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

$$p_{ij}(t) = \begin{cases} \pi_j + (1 - \pi_j)e^{-\mu t} & (i = j) \\ \pi_j(1 - e^{-\mu t}) & (i \neq j) \end{cases}$$

where again μ is fixed and t is estimated.

Evolutionary Models



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 $L(\tau, b, \theta | X)$.
- Let $x_{.i}$ represent the i th column of data. Sites are independent, so $P(X) = P(x_{.1}) P(x_{.2}) \cdots P(x_{.L})$. Then,

$$L(\tau, b, \theta | X) = \prod_{i=1}^L L_i(\tau, b, \theta | x_{.i}).$$

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

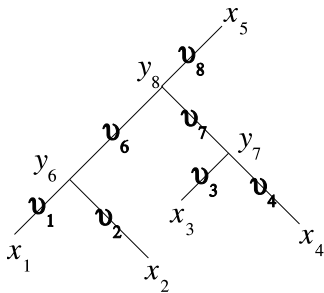
Applying the LTP

- The Law of Total Probability states that $P(B) = P(BA_1) + \dots + P(BA_n)$ where A_i are nonempty, mutually exclusive events with $P(A_1 \cup \dots \cup A_n) = 1$.
- 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 data at a site i be $x_{.i} = (x_{1i}, x_{2i}, \dots, x_{Ni}) = (x_1, x_2, \dots, x_N)$.
- Let the unobserved ancestral data at this site be $y = (y_{N+1}, \dots, y_{2N-2})$.
- Let $N = 5$, so there are 5 elements in vector $x_{.i}$ and 3 elements in the ancestor vector $y = (y_6, y_7, y_8)$.

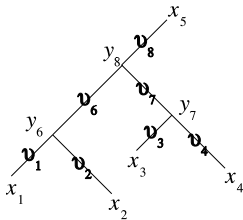
The Ancestors



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

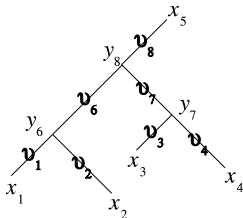
$$P(x_1, x_2 \mid y_6) = ?$$

Data and Ancestor Likelihood



$$\begin{aligned}
 &P(x_1, x_2, x_3, x_4, x_5, y_6, y_7, y_8) = \\
 &P(x_5, y_8, y_6, y_7, x_1, x_2, x_3, x_4) = \\
 &P(x_5)P(y_8 | x_5)P(y_6 | y_8)P(y_7 | y_8) \\
 &\times P(x_1 | y_6)P(x_2 | y_6)P(x_3 | y_7)P(x_4 | y_7) = \\
 &\pi_{x_5} p_{x_5 y_8}(b_8) p_{y_8 y_6}(b_6) p_{y_8 y_7}(b_7) p_{y_6 x_1}(b_1) \\
 &\quad \times p_{y_6 x_2}(b_2) p_{y_7 x_3}(b_3) p_{y_7 x_4}(b_4)
 \end{aligned}$$

Data Likelihood



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

Site-by-Site Rate Heterogeneity

- ML is inconsistent when rate heterogeneity across sites is extreme.
- Option 1: Categorize sites *a priori* into several rate classes and estimate branchlengths separately for each class.
- Option 2: Assume the rate of a site is a discrete random variable R with probability distribution $P(R = r)$. If the rate of site j is r , then the effective branch length is br . But we don't know site j 's rate category, so the likelihood is:

$$\begin{aligned}L(\tau, \mathbf{b}, \theta \mid x_{.j}) &= P(x_{.j} \mid \tau, \mathbf{b}, \theta) \\ &= \sum_{i=1}^{s_r} P(x_{.j} \mid \tau, \mathbf{b}, \theta, r_i) P(R = r_i)\end{aligned}$$

where s_r is the number of rate categories.

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

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

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

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

- Bayes Rule in evolution context:

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

- 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:

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

- 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.