Introduction to Computational Biology
Lecture # 13: Reconstructing phylogenetic trees
Inbal Yanover
9.3.06

1 Brief review

The problem: Given a collection of sequences, we would like to infer the evolutionary process. The methods we have discussed:

1. Distance based methods (like the neighbor-joining).
2. Character based method (like parsimony).

Main problems with parsimony:

- The same weight is given to different changes, no matter how dramatic were those changes.
- The same weight is given to various branch lengths, i.e. no differentiation is made between long branches - where many changes are expected to happen and short branches, where only few changes are expected.

Solution: Add weight to the various changes and to the various branch lengths to produce a better description for the evolution process. And for this we shell construct-

2 Probabilistic model for phylogeny

Let’s start by looking at a single branch, in which we know that the letter A has changed to the letter C in a known time interval t.

We would like to define the probability function: $P(C|A,t)$ as follows:

1. For $t=0$:

$$P(C|A,t) = \begin{cases} 
1 & \text{if } A=C \\
0 & \text{otherwise} 
\end{cases}$$

2. Markov property: The evolution process has no memory, i.e. the only state that matters is the last state, it does not matter how we got there. Markov assumption:

$$P(B | A, t_1 + t_2) = \sum_C P(C | A, t_1) \cdot P(B | C, t_2)$$
3 Continuous time Markov process

In the past, we talked about Markov process which had discrete time steps, and discrete random variables. Today we consider infinite number of variables: 
\[ X = \{ x_t : t \in \mathbb{R} \} \]
A process is the joint probability of this infinite set of variables.

Discrete Markov properties:
1. For \( t_1 > t_2 > t_3 \):
\[ P(x_{t_1} | x_{t_2}, x_{t_3}) = P(x_{t_1} | x_{t_2}) \]
2. Homogeneous:
\[ P(x_{t+d} = a | x_t = b) = P(x_{t'+d} = a | x_{t'} = b) \]
Hence we’ll look only at the changes on a single time step.

We shall construct a matrix \( P(t) \) as follows:
\[ P(t)_{ab} = P(x_t = b | x_0 = a) \]
Note that the Markov property: \( P(t_1 + t_2) = P(t_1) \cdot P(t_2) \) is thus a matrix duplication:
\[ P(t_1 + t_2)_{ab} = \sum_c P(t_1)_{ac} \cdot P(t_2)_{cb} \]

We’ll take the derivative:
\[
\frac{dP(t)}{dt} = \lim_{\epsilon \to 0} \frac{P(t + \epsilon) - P(t)}{\epsilon} = \lim_{\epsilon \to 0} \frac{P(\epsilon)P(t) - P(t)}{\epsilon} = \lim_{\epsilon \to 0} \left[ P(\epsilon) - I \right] \cdot P(t) = \lim_{\epsilon \to 0} \left[ P(\epsilon) - P(0) \right] \cdot P(t) = \frac{dP(0)}{dt} \cdot P(t)
\]
(1)

Note: \( P(0) \) is the unit matrix (I) since the diagonal values are 1, and the values outside it are 0 (see function’s first requirement).

We shall denote \( \frac{dP(0)}{dt} \) as the matrix \( R \). So we got:
\[ \frac{dP(t)}{dt} = R \cdot P(t) \]

\[ P(t) = e^{tR} \]

Since the exponent is a matrix the following equation holds (using Taylor series):
\[ e^A = \sum_{n=0}^{\infty} \frac{1}{n!} \cdot A^n \]

If \( A \) is a symmetric matrix this can be easily computed using \( D \) - the diagonal matrix resembling \( A \):

\[ A = V^T DV \]

hence:
\[ A^n = V^T D^n V \]
4 The rate matrix (R)

The rate matrix defines the rate of the changes in the process.

for $0 < \delta << 1$:

$$P(\delta) \approx P(0) + \delta \cdot R = I + \delta \cdot R$$

Properties of $P(\delta)$:

1. All values are non negative (probabilities).
2. For off diagonal values: $P(\delta) > 0$.
3. Sum of all row values = 1.

Since I contributed 1 to this sum we have:

$$R_{aa} = - \sum_{b \neq a} R_{ab}$$

for $b \neq a$:

$$R_{ab} \geq 0$$

This process can be looked at as a Poisson process. We shall denote:

$$R_a = \sum_{b \neq a} R_{ab}$$

The expected time to stay in the current position $a$ is: $\frac{1}{\pi_a}$

We shall also denote:

$$q_{ab} = \frac{R_{ab}}{R_a}$$

$q$ is the distribution for all the $b$ values when $b \neq a$, i.e. if we change our current state, which is the preferred state. Therefore $0 \leq q_{ab} < 1$.

**Stationary distribution** of the model:

$$\pi_a = \lim_{t \to \infty} P(t)_{ba}$$

$$\pi_a = \sum_b \pi_b \cdot P(t)_{ba}$$

**Reversibility**:

![Figure 1: root location](image-url)
In a reversible process there are few optional locations for the root, as long as the sum of branches length remains constant. If we consider fig 1:

\[ P(z = a) = \pi_a \]
\[ P(x = b | z = a) = P(t_{1ab}) \]

A reversible process implies the condition of detailed balance:

\[ \pi_a \cdot P(t)_{ab} = \pi_b \cdot P(t)_{ba} \]

Deriving \( P(t) \) we get:

\[ \pi_a \cdot R_{ab} = \pi_b \cdot R_{ba} \]
\[ \frac{R_{ab}}{R_{ba}} = \frac{\pi_b}{\pi_a} \]

which forms a real constrain on the rate matrix.

\( R \) is reversible iff there is \( \pi \) and a symmetric matrix \( S \) such that:

\[ \forall a \neq b : R_{ab} = S_{ab} \cdot \pi_b \]

### 4.1 Juke-Cantor matrix

The most simple matrix one can choose (for DNA):

\[
R = \begin{bmatrix}
-3 & 1 & 1 & 1 \\
1 & -3 & 1 & 1 \\
1 & 1 & -3 & 1 \\
1 & 1 & 1 & -3 \\
\end{bmatrix}
\]

In this matrix, the change probabilities are uniformly distributed.

As mentioned earlier, \( P(t) = e^{tR} \). For the above matrix the solution is:

\[
P(t)_{ab} = \begin{cases} 
0.25(1 - e^{-t}) & a \neq b \\
0.25(1 + 3e^{-t}) & a = b 
\end{cases}
\] (2)

When \( t \to \infty \) all states are uniformly distributed. When \( t = 0 \): \( P(a = b) = 1, P(a \neq b) = 0 \). See also fig 2.

If we change 1 to \( \alpha \) and \(-3\) to \(-3\alpha\) the equations will be:

\[
P(t)_{ab} = \begin{cases} 
0.25(1 - e^{-t\alpha}) & a \neq b \\
0.25(1 + 3e^{-t\alpha}) & a = b 
\end{cases}
\]

Which is like changing the time axis (expanding or shrinking it).

We usually define time unit as the time expected for one change to happen. And the time scale is built so this one-change-expectation would be 1.

### 4.2 Kimura 2-param

The Juke-Cantor model does not capture some important biological properties, for example: In DNA, the more common changes are those between A-G and between C-T.

To account for this Kimura proposed a model with this rate matrix:

\[
R = \begin{bmatrix}
-2\alpha - \beta & \alpha & \beta & \alpha \\
\alpha & -2\alpha - \beta & \alpha & \beta \\
\beta & \alpha & -2\alpha - \beta & \alpha \\
\alpha & \beta & \alpha & -2\alpha - \beta 
\end{bmatrix}
\]
If we look at the $P(t)_{ab}$ as follows: (taken from Durbin’s book)

$$P(t)_{ab} = \begin{bmatrix} r & s & u & s \\ s & r & s & u \\ u & s & r & s \\ s & u & s & r \end{bmatrix}$$

Then:

$$s = 0.25(1 - e^{-4\alpha t})$$
$$u = 0.25(1 + e^{-4\alpha t} - 2e^{-(\alpha + \beta)t})$$
$$r = 1 - 2s - u$$

(3)

**JJT** is a rate matrix, 20X20, for amino acids. It was built using experimental data, and assuming changes are reversible.

## 5 How can we estimate the distance between 2 proteins given the matrix R?

If we have few sequences and we align them:

- $a - c$
- $a - a$
- $c - c$

... The likelihood function will be:

$$L(t; D) = P(t)_{ac} \cdot P(t)_{aa} \cdot P(t)_{cg} \cdot \ldots$$

or in general:

$$L(t; D) = \prod_{a,b} P(t)_{ab}^{N[a,b]}$$

The statistic is $N[a, b]$ for every pair a,b. which defines a matrix.

Finding $P(t)$: look for maximum likelihood without derivation, this is called **line search**:

if we could find 3 dots which not all of them are increasing or decreasing then the maximum resides either in the interval between the two maximal points, or between the maximal point and the other two.
Figure 3: Likelihood function