Jukes and Cantor Method

Summary

The Jukes and Cantor model is a model which computes probability of substitution from one state (originally the model was for nucleotides, but this can easily be substituted by codons or amino acids) to another.
From this model we can also derive a formula for computing the distance between 2 sequences.

The main idea behind this model is the assumption that probability of changing from one state to a different state is always equal. As well, we assume that the different sites are independent.

In this text, we will relate to states which are nucleotides. One can easily restate the proofs here with any other state (such as amino acids or codons).

The proofs

Therefore, looking at states that are nucleotides, we get:

$$P(t) = \begin{array}{cc} \begin{array}{cccc} \text{A} & \text{T} & \text{G} & \text{C} \end{array} & \\ \begin{pmatrix} 1-3f(t) & f(t) & f(t) & f(t) \\ f(t) & 1-3f(t) & f(t) & f(t) \\ f(t) & f(t) & 1-3f(t) & f(t) \\ f(t) & f(t) & f(t) & 1-3f(t) \end{pmatrix} & \begin{array}{c} A \\ T \\ G \\ C \end{array} \end{array} \qquad (1)$$

- Each item in the matrix represents a substitution probability. For instance, column 2, row 3 represents the probability to change from G to T (or vice-versa) in time t.
- f(t) = the probability to change from one nucleotide to a different one as a function of the time (t) is also notated as $p_{ij}(t)$ for i≠j and $p_{ii}(t\,)=(1\text{-}3f(t))$.
- Note that the values on the diagonal are obtained because we assume that the sum of each row is 1 (since it represents an event).

Calculating f(t)=p_{ij}(t)

$$P'(t) = \begin{pmatrix} -3f'(t) & f'(t) & f'(t) & f'(t) \\ f'(t) & -3f'(t) & f'(t) & f'(t) \\ f'(t) & f'(t) & -3f'(t) & f'(t) \\ f'(t) & f'(t) & f'(t) & -3f'(t) \end{pmatrix} \qquad (2)$$

(We're not getting into why matrix can be differentiated this way – you'll just have to accept it!)

Lemma 1:

$$P'(t) = P(t) \bullet P'(0)$$

Proof:

I. $P(t+\Delta t) = P(t) \bullet P(\Delta t)$ - because of matrix multiplication (it works, write it out!)

II.

$$P'(t) = \lim_{\Delta t \to 0} \frac{P(t+\Delta t) - P(t)}{\Delta t} \underset{\substack{\uparrow \\ \text{following (I)}}}{=} \lim_{\Delta t \to 0} \frac{P(t) \bullet P(\Delta t) - P(t)}{\Delta t} \underset{\substack{\uparrow \\ \text{following (I), since P(t)=P(t+0)}}}{=} P(t) \lim_{\Delta t \to 0} \frac{P(\Delta t) - P(0)}{\Delta t} = P(t) \bullet P'(0)$$

□

Lemma 2:

$$P(t) = e^{Q \bullet t}$$

Proof:

Define Q as:

$$Q = P'(0) = \begin{pmatrix} -3f'(0) & f'(0) & f'(0) & f'(0) \\ f'(0) & -3f'(0) & f'(0) & f'(0) \\ f'(0) & f'(0) & -3f'(0) & f'(0) \\ f'(0) & f'(0) & f'(0) & -3f'(0) \end{pmatrix} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \quad (3)$$

where f'(0) = α

from lemma 1 we get:

$$\frac{dP}{dt} = P(t) \bullet Q$$

Thus, we now have a differential equation. To solve it, we integrate and get:

$$\int \frac{dP}{P(t)} = \int Q dt$$

$$\Rightarrow \quad \ln P(t) = Q \bullet t + c$$

$$\Rightarrow \; P(t) = e^{Q \bullet t} \qquad (4)$$

□

Lemma 3:

From lemma 1 and the definition of Q we get:

$$f(t) = \frac{1}{4} - \frac{1}{4} \bullet e^{-4\alpha t}$$ (where α=f'(0))

Proof:

$$P'(t)=\begin{pmatrix}1-3f(t) & f(t) & f(t) & f(t)\\ f(t) & 1-3f(t) & f(t) & f(t)\\ f(t) & f(t) & 1-3f(t) & f(t)\\ f(t) & f(t) & f(t) & 1-3f(t)\end{pmatrix}\bullet\begin{pmatrix}-3\alpha & \alpha & \alpha & \alpha\\ \alpha & -3\alpha & \alpha & \alpha\\ \alpha & \alpha & -3\alpha & \alpha\\ \alpha & \alpha & \alpha & -3\alpha\end{pmatrix}$$

By multiplying row 1 in the first matrix with column 2 in the second matrix we get:

$$p'_{1,2}(t)=\alpha-3\alpha f(t)-3\alpha f(t)+\alpha f(t)+\alpha f(t)=\alpha-4\alpha f(t)$$

$p'_{1,2}(t)=p'_{i,j}(t)=f'(t)$ for all $i\neq j$ (see equation 2)

Therefore $\dfrac{df}{dt}=\alpha-4\alpha f(t)$

We integrate and get:

$$\int\frac{df}{\alpha-4\alpha f(t)}=\int dt$$

$$\frac{\ln(\alpha-4\alpha f(t))}{-4\alpha}=t+c$$

$$\ln(\alpha-4\alpha f(t))=-4\alpha t+c$$

$$\ln(\alpha-4\alpha f(t))=ce^{-4\alpha t}$$

$$f(t)=\frac{1}{4}-\frac{ce^{-4\alpha t}}{4\alpha}$$

To calculate c:

We assume f(0) = 0. This means that at our starting point, i.e. at time 0, we start with a constant state (for instance at time 0 we start with A, and thus the probability of *changing* from A at time 0 is 0). Therefore:

$$f(0)=0\Rightarrow f(0)=\frac{1}{4}-\frac{c}{4\alpha}=0\Rightarrow c=\alpha$$

$$f(t)=\frac{1}{4}-\frac{e^{-4\alpha t}}{4} \qquad (5)$$

□

Thus we get:

pij(t) = f(t) = $\dfrac{1}{4}-\dfrac{e^{-4\alpha t}}{4}$

pii(t) = 1-3f(t)= $\dfrac{1}{4}+\dfrac{3}{4}e^{-4\alpha t}$

As well:

pij'(t) = f'(t) = $\alpha e^{-4\alpha t}$

pii'(t) = -3f'(t)= $-3\alpha e^{-4\alpha t}$

Calculation of α

$$\sum_{j}\sum_{i\neq j}\Pi_j Q_{ij} = 1 \qquad (6)$$

where j and i are A,T,C.G. Π_j is the probability of starting at state j and is equal to 1/4 . $Q_{i,j}$ is the rate of changing from state i to state j when t =0 (see matrix 3). For each nucleotide we have 3 possibilities of changes, therefore we sum 12 elements and get:
Therefore:

$$\Pi_A Q_{AC} + \Pi_A Q_{AG} + \Pi_A Q_{AT} + \Pi_C Q_{CA} .... = \frac{1}{4} 12 Q_{i,j} = 3\alpha = 1$$

$$\Rightarrow \alpha = \frac{1}{3}$$

Why t is equivalent to the distance between sequences
Since we are using only molecular evidence (i.e. – sequences), we have to define t (which we are used to relating to as time). Here we will define t as the distance between sequences, and this distance will be directly related to the number of observed differences between the sequences.

Finding t: estimating the distance by maximum likelihood
Let $\{x_1,x_2,x_3......x_n\}$ and $\{y_1,y_2,y_3.....y_n\}$ be two sequences of nucleotides with n positions. In order to evaluate the distance between them we use the maximum likelihood method:
The likelihood we are relating to is the likelihood of starting off with one sequence and ending up with the other sequence (this is not exact… actually we we're looking at the likelihood of these 2 sequences being derived from a common ancestor, but it comes out the same. Feel free to ask if you didn't understand the last sentence).
What we're looking for, is the t which gives us the maximum likelihood.
The formula for likelihood of starting with the sequences of x's and ending with the sequences of y's is:

$$L = p(x_1)p(x_1 \to y_1 | t)p(x_2)p(x_2 \to y_2 | t)....p(x_n)p(x_n \to y_n | t) \qquad (7)$$

where $p(x_i)$ is the probability of starting off with nucleotide x_i and $p(x_i \to y_i | t)$ is the probability of changing from x_i to y_i , given t.

The following is derived from equation (7):

$$\ln L = \ln(p(x_1)p(x_1 \to y_1 \mid t)p(x_2)p(x_2 \to y_2 \mid t)....p(x_n)p(x_n \to y_n \mid t))$$

$$\ln L = \ln p(x_1) + \ln p(x_1 \to y_1 \mid t) + \ln p(x_2) + \ln p(x_2 \to y_2 \mid t) + ... + \ln p(x_n) + \ln p(x_n \to y_n \mid t)$$

$$\ln L = \underbrace{\ln p(x_1) + \ln p(x_2) + ... + \ln p(x_n)}_{=const} + \ln p(x_1 \to y_1 \mid t) + \ln p(x_2 \to y_2 \mid t) + ... + \ln p(x_n \to y_n \mid t)$$

$$\ln L = const + \ln p(x_1 \to y_1 \mid t) + \ln p(x_2 \to y_2 \mid t) + ... + \ln p(x_n \to y_n \mid t)$$

The probability of substitution from nucleotide x to y where x ≠ y is equal for all x and y, and the probability of substitution from nucleotide x to x is equal for all x. Therefore, the equation becomes:

$$\ln L = const + m_1 \ln p_{ij}(t) + m_2 \ln p_{ii}(t)$$

where:
$p_{ij}(t) = p(x_k \to y_k)$ if $x_k \neq y_k$, and $p_{ii}(t) = p(x_k \to y_k)$ if $x_k = y_k$
m_1 is the number of positions where the substitution is to a different nucleotide, and
m_2 is the number of positions where the substitution is to the same nucleotide.

In order to find the maximum likelihood for t, we differentiate (ln L) and get:

$$\frac{d(\ln L)}{dt} = \frac{m_1}{p_{ij}(t)} p_{ij}'(t) + \frac{m_2}{p_{ij}(t)} p_{ii}'(t)$$

We equate this to zero and this gives:

$$\frac{m_1}{p_{ij}(t)} p_{ij}'(t) + \frac{m_2}{p_{ij}(t)} p_{ii}'(t) = \frac{m_1}{\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}t}} \bullet \frac{1}{3} e^{-\frac{4}{3}t} + \frac{m_2}{\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}t}} \bullet (-e^{-\frac{4}{3}t}) = 0$$

Let x $= e^{-\frac{4}{3}t}$, therefore we get:

$$\frac{4m_1}{1-x} \bullet \frac{x}{3} - \frac{4m_2}{1+3x} \bullet x = 0$$

$$4m_1 x + 12m_1 x^2 - 12m_2 x + 12m_2 x^2 = 0$$

$$(12m_1 + 12m_2)x^2 + (4m_1 - 12m_2)x = 0$$

$$x = \frac{12m_2 - 4m_1}{12m_1 + 12m_2} = \frac{3m_2 - m_1}{3m_1 + 3m_2}$$

We replace x and get:

$$e^{-\frac{4}{3}t} = \frac{3m_2 - m_1}{3m_1 + 3m_2}$$

$$-\frac{4}{3}t = \ln(\frac{3m_2 - m_1}{3m_1 + 3m_2})$$

$$t = -\frac{3}{4}\ln(\frac{3m_2 + 3m_1 - 3m_1 - m_1}{3m_1 + 3m_2})$$

$$t = -\frac{3}{4}\ln(1 - \frac{4m_1}{3(m_1 + m_2)}) = -\frac{3}{4}\ln(1 - \frac{4m_1}{3n})$$

where n is the number of position in the sequences.

We define p $_{\text{distance}}$ as $\frac{m_1}{n} = \frac{no.of.different}{n}$ and thus get:

$$t = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

THE END!