Assignment 2.1 derivation / BB_ II / 2013

Assignment 2.1

Modelling nucleotide substitutions is an elementary part of (for example) phylogeny reconstructions (with the exception of parsimony methods). Jukes-Cantor model assumes that all substitutions occur with equal probabilities. Derivation is given in lecture slides and, for convenience, here anew.

Derive the two-parameter model including separate parameters for transitions and transversions. In case you consider this as too demanding, it is enough that you construct the strating scheme and the first equations, i.e. the part corresponding equations (1) – (3) in Jukes-Cantor one-parameter model.

Derivation is here after Jukes-Cantor derivation

Assumption: all nucleotide substitutions occur with equal probabilites, α, Jukes-Cantor model (1969)

- The rate of substitution for each nucleotide is 3α per unit time

	A	T	C	G
A		α	α	α
T	α		α	α
C	α	α		α
G	α	α	α	

- At time 0: Assumption that at a certain nucleotide site there is A, $P_{A(0)} = 1$
- Question: probability that this site is occupied by A at time t , $P_{A(t)}$?
- At time 1, probability of still having A at this site is

$$P_{A(1)} = 1 - 3\alpha \qquad (1)$$

- 3α is the probability of A changing to T, C, or G
- The probability of the site having A at time 2 is

$$P_{A(2)} = (1 - 3\alpha)P_{A(1)} + \alpha\,[1 - P_{A(1)}] \qquad (2)$$

This includes two possible courses of events from time points $t=0$ --> $t = 1$ --> $t = 2$

$t = 0$	----------→	$t = 1$	----------→	$t = 2$
A	no substitution	A	no substitution	A
A	substitution	T or C or G	substitution	A

- The following recurrence equation holds for any t

$$P_{A(t+1)} = (1 - 3\alpha)P_{A(t)} + \alpha[1 - P_{A(t)}] \quad (3)$$

Note that this holds also for $t = 0$, because $P_{A(0)} = 1$ and thus

$$P_{A(0+1)} = (1 - 3\alpha) P_{A(0)} + \alpha [1 - P_{A(0)}] = 1 - 3\alpha$$

which is identical with equation (1).

- The amount of change in $P_{A(t)}$ per unit time, rewriting equation (3):

$$\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)} = - 3\alpha P_{A(t)} + \alpha[1 - P_{A(t)}] = - 4\alpha P_{A(t)} + \alpha \quad (4)$$

- Approximating the previous discrete-time model by a continuous-time model, by regarding $\Delta P_{A(t)}$ as the rate of change at time t. With this approximation equation (4) is rewritten as

$$dP_{A(t)} / dt = - 4\alpha P_{A(t)} + \alpha \quad (5)$$

- The solution of this first-order linear differential equation is

$$P_{A(t)} = ¼ + (P_{A(0)} - ¼)e^{-4\alpha t} \quad (6)$$

- The starting condition was A at the given site, $P_{A(0)} = 1$, consequently

$$P_{A(t)} = ¼ + ¾ e^{-4\alpha t} \quad (7)$$

- Equation (6) holds regardless of the initial conditions, for example if the initial nucleotide is not A, then $P_{A(0)} = 0$, and the probability of having A at time t

$$P_{A(t)} = ¼ + ¼ e^{-4\alpha t} \quad (8)$$

- Equations (7) and (8) describe the substitution process. If the initial nucleotide is A, then $P_{A(t)}$ decreases exponentially from 1 to ¼ . If the initial nucleotide is not A, then $P_{A(t)}$ will increase monotonically from 0 to ¼ .

- Under this simple model, after reaching equilibrium, $P_{A(t)}=P_{T(t)}=P_{C(t)}=P_{G(t)}$ for all subsequent times.

- Equation (7) can be rewritten in a more explicit form to take into account that the initial nucleotide is A and the nucleotide at time t is also A

$$P_{AA(t)} = ¼ + ¾ e^{-4\alpha t} \quad (9)$$

- If the initial nucleotide is G instead of A, from equation (8)

$$P_{GA(t)} = ¼ + ¼ e^{-4\alpha t} \quad (10)$$

Since all the nucleotides are equivalent under the Jukes-Cantor model, the general probability, $P_{ij(t)}$, that a nucleotide will become *j* at time *t*, given that it was *i* at time 0, equations (9) and (10) give the general probabilities $P_{ii(t)}$ and $P_{ij(t)}$, where $i \neq j$.

$$P_{ii(t)} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \quad \text{and} \quad P_{ij(t)} = \frac{1}{4} + \frac{1}{4} e^{-4\alpha t} \qquad (11)$$

Number of substitutions, *nucleotide divergence*, between two sequences

- We assume that all sites in sequence evolve at the same rate and follow the same substitution scheme. The number of sites compared between two sequences is denoted by *L*.

- Consider the probability that a nucleotide at a given site at time *t* is the same in both sequences. Suppose that the nucleotide at a given site was A at time point 0. At time *t*, the probability that a descendant sequence will have A at this site is $P_{AA(t)}$, and consequently the probability that two descendant sequences have A at this site is $P^2_{AA(t)}$. Similarly, the probabilities that both sequences have T, C or G at this site are $P^2_{AT(t)}$, $P^2_{AC(t)}$, and $P^2_{AG(t)}$

- The probability that the nucleotide at a given site at time *t* is the same in both sequences is

$$I_{(t)} = P^2_{AA(t)} + P^2_{AT(t)} + P^2_{AC(t)} + P^2_{AG(t)} \qquad (12)$$

- From equations (11) we obtain

$$I_{(t)} = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t} \qquad (13)$$

- Equation (13) also holds for T, C or G. Therefore, regardless of the initial nucleotide at a given site, $I_{(t)}$ represents the proportion of *identical* nucleotides between two sequences that diverged *t* time units ago. The probability that the two sequences are *different* at a site at time *t* is $p = 1 - I_{(t)}$. Thus

$$p = \frac{3}{4}(1 - e^{-8\alpha t}) \quad \text{or} \quad 8\alpha t = \ln(1 - (4/3)\, p) \qquad (14)$$

- The time of divergence between two sequences is usually not known, and thus estimation of α is not possible. Instead, it possible to calculate ***K*, which is the number of substitutions per site since the time of divergence between the two sequences**. In the case of the one-parameter model, $K = 2(3\alpha t)$, where $3\alpha t$ is the number of substitutions per site in a single lineage.

$$K = 6\alpha t = -\frac{3}{4} \ln(1 - (4/3)\, p) \qquad (15)$$

where *p* is the observed proportion of different nucleotides between the two sequences.

An example. Page 3 (book chapter page 143) in *Phylogeny methods based on distance matrices* (see course webpage, week 1) shows how Jukes-Cantor model serves like a *correction* to sequence diverge calculation.

Assumption: transitions occur with probability α, transversions occur with probability β, Kimura (1980) two-parameter model

● Motoo Kimura´s motivation to derive this model was the fact that accumulating observations at that time showed that transitions (changes between A and G or between C and T) seemed to be more frequent than transversions (A – T, T – A, C – G, G – C, A – C, C – A, G – T, T – G)

● In Jukes-Cantor model, two possible courses of events from time points $t = 0 \rightarrow t = 1 \rightarrow t = 2$ were considered.

● Now we have to consider a more complex situation with four courses of events. The questions posed are the same as above in Jukes-Cantor model derivation.

$t = 0$	----------→	$t = 1$	----------→	$t = 2$
A	no substitution	A	no substitution	A
A	transition	G	transition	A
A	transversion	C	transversion	A
A	transversion	T	transversion	A

● Considering the the probability that a site that has A at time 0 will have A at time t. After one time unit, the probability of A changing into G is α and into C or T is 2β. Thus, the probability of A remaining unchanged after one time unit is

$$P_{AA(1)} = 1 - \alpha - 2\beta \qquad (16)$$

● At time 2, the probability of having A at this site is given by the sum of four different courses of possible events (see the scheme above)

$$P_{AA(2)} = (1 - \alpha - 2\beta)P_{AA(1)} + \beta P_{TA(1)} + \beta P_{CA(1)} + \alpha P_{GA(1)} \qquad (17)$$

● By extension (cf. above in Jukes-Cantor (3)) we obtain the following recurrence equation for the general case

$$P_{AA(t+1)} = (1 - \alpha - 2\beta)P_{AA(t)} + \beta P_{TA(t)} + \beta P_{CA(t)} + \alpha P_{GA(t)} \qquad (18)$$

● Rewriting thios equation as the amount of change in $P_{AA(t)}$ per unit time, and after approximating the discrete-time model by the continuous-time model, we obtain the following differential equation (cf. above in Jukes-Cantor (5))

$$dP_{AA(t)} / dt = -(\alpha - 2\beta)P_{AA(t)} + \beta P_{TA(t)} + \beta P_{CA(t)} + \alpha P_{GA(t)} \qquad (19)$$

● Similarly, equations for for $P_{TA(t)}$, $P_{CA(t)}$, and $P_{GA(t)}$, and from this set of of four equations, the following solution (details not shown)

$$P_{AA(t)} = \tfrac{1}{4} + \tfrac{1}{4}e^{-4\beta t} + \tfrac{1}{2}\, e^{-2(\alpha + \beta)t} \qquad (20)$$

● Recall from equation (11) that in Jukes –Cantor model, the probability that the nucleotide at a site at time t is identical to that at time 0 is the same for all nucleotides: $P_{AA(t)} = P_{GG(t)} = P_{CC(t)} = P_{TT(t)}$. Because of symmetry of the substitution scheme, this equality also holds for Kimura´s two-parameter model.

● We denote this probability by $X_{(t)}$. Therefore equation (20) is

$$X_{(t)} = \tfrac{1}{4} + \tfrac{1}{4}e^{-4\beta t} + \tfrac{1}{2}\, e^{-2(\alpha + \beta)t} \qquad (21)$$

● At equilibrium, i.e. at $t = \infty$, equation (21) reduces to $X_{(\infty)} = \tfrac{1}{4}$. Thus, like in Jukes-Cantor model, the equilibrium frequencies of the four nucleotides are ¼.

● Under the Jukes-Cantor model, equations (11) hold regardless of whatever the change from nucleotide i to nucleotides j is transition or a transversion. In two-parameter model we need to make a distinction between transitions and transversion. We denote by $Y_{(t)}$ the probability that the initial nucleotide and the nucleotide at time t differ from each other by a transition. because of the symmetry of the substitution scheme, $Y_{(t)} = P_{AG(t)} = P_{GA(t)} = P_{TC(t)} = P_{CT(t)}$. It can be shown that

$$Y_{(t)} = \tfrac{1}{4} + \tfrac{1}{4} e^{-4\beta t} - \tfrac{1}{2} e^{-2(\alpha + \beta)t} \qquad (22)$$

● The probability, $Z_{(t)}$, that the nucleotide at time t and the initial nucleotide differ by a specific type of transversion is given by

$$Z_{(t)} = \tfrac{1}{4} - \tfrac{1}{4} e^{-4\beta t} \qquad (23)$$

● Note that each nucleotide is subject to two types of transversion, but only one type of transition. For example, if the initial nucleotide is A, then the two possible transversional changes are A → C and A → T. Therefore, the probability that the initial nucleotide and the nucleotide at time t differ by one of the two types of transversion is twice the probability given in equation (23). Note also that $X_{(t)} + Y_{(t)} + 2Z_{(t)} = 1$.

Recall from Jukes-Cantor: number of substitutions, *nucleotide divergence*, between two sequences, i.e. usage of the model for data. Now the same by using Kimura´s two-parameter model.

● K, the number of substitutions per site since the divergence of two sequences (cf. equation (15)). Let P and Q be the proportions of transitional and transversional differences between the two sequences, respectively. Then the number of nucleotide substitutions per site between the two sequences, K, is estimated by

$$K = \tfrac{1}{2} \ln [\, 1 / (1 - 2P - Q)] + \tfrac{1}{4} \ln [\, 1 / (1 - 2Q)] \qquad (24)$$

● Note that this equation reduces to equation (15) if transitions and transversion are not distinguished separately.

An example. Two sequences of length 200 nucleotides differ from each other by 20 transitions and 4 tranversions. Thus $P = 20/200 = 0.10$ and $Q = 4/200 = 0.02$. According to two-parameter model, $K \approx 0.13$. The total number of substitutions (i.e. what is behind the observed numbers) can be obtained by multiplying the number of substitutions per site, K, by the number of sites, L (the length, here 200 nucleotides). We obtain an estimate of about 26 substitutions, resulting in 24 observed differences between the two sequences. According to one-parameter model, $p = 24/200 = 0.12$ and $K \approx 0.13$. In this example the two models give essentially the same estimate.

Another example. Two sequences of length 200, differ from each other by 50 transitions and 16 transversions. $P = 50/200 = 0.25$ and $Q = 16/200 = 0.08$. Two-parameter model: $K \approx 0.48$, one parameter model $p = 66/200 = 0.33$ and $K \approx 0.43$ which is 10% smaller than by using two-parameter model.