Genomics and bioinformatics

BIO-463

Prof. Anne-Florence Bitbol
EPFL

1 Evolutionary distance: the Jukes-Cantor model

Here, we provide some formal details on the Jukes-Cantor model of nucleotide evolution.

Model. Compared to the Hamming distance, the Jukes-Cantor distance is the simplest correction that takes into account multiple substitutions. It is based on an evolutionary model where each site evolves independently of other ones, and where all substitutions are equally likely, which are two simplifying assumptions. Considering DNA sequences, composed of nucleotides A, C, G, T, this means that the substitution rate between any two nucleotides is the same, and we will denote it by λ. This rate is also assumed to be the same at all sites in the sequences considered.

Master equation. Let us focus on a given site and consider (for instance) the probability $P_A(t)$ that the nucleotide present at this site is A. If the state is A at time t, the probabilities that it mutates to C, or G, or T during a small time interval dt are $dp_{A\to C} = dp_{A\to G} = dp_{A\to T} = \lambda dt$. Similarly, if the state is not A at time t, for instance if it is C, then the probability that it mutates to A during a small time interval dt is $dp_{C\to A} = \lambda dt$. Let us now express the probability that the state is A at time $t + dt$, discriminating over the state at time t: we can end up in state A at $t + dt$ either by already being in A at t and not mutating to another state during dt, or by being in another state at t and mutating to A during dt. Assuming that dt is short enough for us to neglect the possibility that more than one event happens during it, we can write

$$\begin{aligned} P_A(t+dt) &= P_A(t)\left[1 - dp_{A\to C} - dp_{A\to G} - dp_{A\to T}\right] + P_C(t)dp_{C\to A} + P_G(t)dp_{G\to A} + P_T(t)dp_{T\to A} \\ &= P_A(t)\left[1 - 3\lambda dt\right] + \left[P_C(t) + P_G(t) + P_T(t)\right]\lambda dt \\ &= P_A(t)\left[1 - 3\lambda dt\right] + \left[1 - P_A(t)\right]\lambda dt\,, \end{aligned} \tag{1}$$

where we used the normalization relation $P_A(t) + P_C(t) + P_G(t) + P_T(t) = 1$ to obtain the last line (at time t, the state of the site we are focusing on has to be either A or C or G or T). Therefore, we have

$$P_A(t+dt) = P_A(t)\left[1 - 4\lambda dt\right] + \lambda dt\,, \tag{2}$$

which gives the following differential equation:

$$\frac{dP_A}{dt}(t) = \lambda\left[1 - 4P_A(t)\right]\,, \tag{3}$$

This is the master equation associated to our evolutionary model.

Solution of the master equation. To solve this inhomogeneous first-order differential equation, let us first solve the homogeneous equation associated to it:

$$\frac{dP_A}{dt}(t) = -4\lambda P_A(t)\,. \tag{4}$$

The solution to this equation is $P_A(t) = Ce^{-4\lambda t}$, where C is a constant. Next, let us find a particular solution of Eq. 3. For this, let us search for a constant solution. It should satisfy $0 = \lambda\left[1 - 4P_A(t)\right]$, which gives $P_A(t) = 1/4$. Thus, the general solution of Eq. 3 can be written as

$$P_A(t) = Ce^{-4\lambda t} + \frac{1}{4}\,. \tag{5}$$

Importantly, given the symmetry of the model, $P_C(t)$, $P_G(t)$ and $P_T(t)$ also take this form, but potentially with different values of C.

Now there are two possibilities: either at $t = 0$ the state of the site of interest was A, or it was another nucleotide. The value of C will be different in each of these cases. In the first case, we have

$$P_A(t) = \frac{3}{4}e^{-4\lambda t} + \frac{1}{4}\,, \tag{6}$$

and

$$P_C(t) = P_G(t) = P_T(t) = -\frac{1}{4}e^{-4\lambda t} + \frac{1}{4}\,. \tag{7}$$

These equations satisfy the normalization relation $P_A(t) + P_C(t) + P_G(t) + P_T(t) = 1$. In the second case, $P_A(t)$ is given by Eq. 7.

Long-time behavior. When $t \to \infty$, we have $P_A(t) \to 1/4$, and the same is true for $P_C(t)$, $P_G(t)$ and $P_T(t)$. This means that under this model, if we wait long enough, we have the same probability (uniform probability distribution) to find any nucleotide at the site of interest. This is consistent with the fact that the model is fully symmetric (all nucleotides are considered equivalent).

Short-time behavior. To understand the short-time behavior, let us consider Eqs. 6 and 7, and perform a first-order Taylor expansion when $t \to 0$. Eq. 6 yields

$$P_A(t) \approx \frac{3}{4}\left[1 - 4\lambda t\right] + \frac{1}{4} = 1 - 3\lambda t\,. \tag{8}$$

This is consistent with our expectations from the model, since $dp_{A\to C} = dp_{A\to G} = dp_{A\to T} = \lambda dt$. Meanwhile Eq. 7 yields

$$P_A(t) \approx -\frac{1}{4}\left[1 - 4\lambda t\right] + \frac{1}{4} = \lambda t\,. \tag{9}$$

This is also consistent with our expectations from the model, since in this case, we start at $t = 0$ from a given other nucleotide, e.g. C, and the relevant rate is then $dp_{C\to A} = \lambda dt$.

Evolutionary distance. Evolutionary distance represents the number of mutations that occurred between two sequences. In the Jukes-Cantor model, the total rate of substitution to a different nucleotide is 3λ. Thus, for sequences separated by an evolutionary time t (for instance an ancestral sequence at time 0 and a sequence descending from it observed at t), evolutionary distance is $d_{JC} = \int_0^t 3\lambda dt = 3\lambda t$ in the Jukes-Cantor model.

Consider a site where the ancestral state (at time $t = 0$) is A. Then, at time t, the probability that the state at this site is different from A reads

$$P_d(t) = P_C(t) + P_G(t) + P_T(t) = -\frac{3}{4}e^{-4\lambda t} + \frac{3}{4}\,, \tag{10}$$

where we have used Eq. 7. In particular, for short times t, we have $P_d(t) \approx 3\lambda t = d_{JC}$.

Consider the Hamming distance $d_H = x/L$ between the ancestral sequence at $t = 0$ and the observed sequence at t, where x are the number of sites that differ between them, while L is the number of nucleotide sites in the sequences of interest (i.e. their length). Recall that we assume for simplicity that each site in a sequence evolves independently from others and with the same substitution rate λ. Thus, the Hamming distance $d_H = x/L$ is an empirical estimate of $P_d(t)$. Equating them gives:

$$d_H = -\frac{3}{4}e^{-4\lambda t} + \frac{3}{4} \quad \text{i.e.} \quad e^{-4\lambda t} = 1 - \frac{4}{3}d_H\,, \tag{11}$$

Thus, we obtain an estimate of the evolutionary distance $d_{JC} = 3\lambda t$:

$$d_{JC} = 3\lambda t = -\frac{3}{4}\log\left[1 - \frac{4}{3}d_H\right]\,, \tag{12}$$

where log is the natural logarithm. Eq. 12 allows to estimate the evolutionary distance d_{JC} (under the Jukes-Cantor model) from the observed Hamming distance d_H. In particular, for short times t,

we have $d_{JC} \approx d_H$. This makes sense: for short times, multiple substitutions are negligible and the Hamming distance is a good estimate of evolutionary distance.

So far, we focused on the evolutionary distance between an ancestral protein and a current protein. However, given the symmetries of the Jukes-Cantor model, the evolutionary distance between two current proteins that descend from the same ancestral protein can be obtained in the exact same way, using Eq. 12. In this case, which is more relevant in practice, t is the total time the two sequences have diverged for, i.e. it is equal to twice the time they diverged from their last common ancestor.

Remark: very distant sequences. Eq. 12 is undefined if $d_H > 3/4$. This means that this approach cannot be used for very distant sequences. If the Hamming distance tends to $3/4$ while being smaller, the estimated Jukes-Cantor evolutionary distance tends to infinity. This makes sense since the limit of P_d for $t \to \infty$ is $3/4$: under the Jukes-Cantor model, we expect that maximally diverged sequences have a Hamming distance of $3/4$.