An adjacency co-evolutionary model

Juan Wang

1 Introduction to related concepts

Definition 1. Given two adjacencies a={a$_1$,a$_2$} and b={b$_1$,b$_2$}, an adjacency pair is defined as a combination of two adjacencies.

For example: $<\{a_1,a_2\},\{b_1,b_2\}>$ is an adjacency pair.

Definition 2. In genome G, we define a certain adjacency pair $< a, b >$ to have five states.

For example: $< 1, 1 >$ indicates that both adjacencies are present in genome G, $< 0, 0 >$ indicates that neither adjacency is present in genome G, $< 1, 0 >$ and $< 0, 1 >$ indicate the cases where a is present and b is absent and a is absent and b is present, respectively. Particularly, if a and b are both present in genome G but located in different chromosomes, we use $< 1, 1 >'$.

2 Description of an adjacency Co-evolutionary Model

2.1 Calculation of the adjacency-pairs probability

To obtain each state of each adjacency pair of a particular node, it is sufficient to be given a description of the structure of the evolutionary tree.

2.1.1 Probabilistic Models

Given a transition rate matrix Q describing the birth-death process.

$$Q = \begin{pmatrix} -3a-b & a & a & a & b \\ a & -3a-b & a & a & b \\ a & a & -3a-b & a & b \\ a & a & a & -3a-b & b \\ b & b & b & b & -4b \end{pmatrix} \tag{1}$$

Where a denotes the transition rate between two adjacency-pairs in the same chromosome and b denotes the transition rate between two adjacency-pairs in different chromosomes. Let the matrix $P(t) = P_{ij}(t)$ be a 5-dimensional matrix, where $P_{ij}(t)$ denotes the probability that state i transitions to state j after time t. $P(t + dt)$ is the state transition matrix at time $t + dt$. We can derive from the rate matrix Q that $P(t + dt) = P(t) + QP(t)dt$. Then we can obtain the differential equation $P'(t) = QP(t)$, and by solving this equation we can further obtain $P(t) = P(0) \times e^{Qt}$, for $P(0) = 1$, we have $P(t) = e^{Qt}$. Then we diagonalize the Q matrix to obtain: $Q = VDU$. where V,D,U are:

$$V = \begin{pmatrix} 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & -4 & 0 & 0 & 0 \end{pmatrix} \tag{2}$$

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -5b & 0 & 0 & 0 \\ 0 & 0 & -4a-b & 0 & 0 \\ 0 & 0 & 0 & -4a-b & 0 \\ 0 & 0 & 0 & 0 & -4a-b \end{pmatrix} \tag{3}$$

$$V = \begin{pmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/20 & 1/20 & 1/20 & 1/20 & -1/5 \\ -1/4 & -1/4 & -1/4 & -1/4 & 0 \\ -1/4 & -1/4 & 3/4 & -1/4 & 0 \\ -1/4 & -1/4 & -1/4 & 3/4 & 0 \end{pmatrix} \tag{4}$$

Next, we perform a Taylor expansion of the formula $P(t) = e^{Qt}$ to obtain the following transition probability matrix: $p_{ij}(t) = \sum_{k=1}^{5} V_{ik} U_{kj} exp(td_k)$. where $V_{i.}$ is the value of the ith row of the V matrix. $U_{.j}$ is the jth column value of the U matrix. d_k is the kth eigenvalue of the D matrix. So the transition probability matrix is as follows:

$$P = \begin{pmatrix} p_1 & p_2 & p_2 & p_2 & p_3 \\ p_2 & p_1 & p_2 & p_2 & p_3 \\ p_2 & p_2 & p_1 & p_2 & p_3 \\ p_2 & p_2 & p_2 & p_1 & p_3 \\ p_3 & p_3 & p_3 & p_3 & p_4 \end{pmatrix} \tag{5}$$

Where:
$p_1 = \frac{1}{5} + \frac{1}{20}e^{-5bt} - \frac{3}{4}e^{-4at-bt}$
$p_2 = \frac{1}{5} + \frac{1}{20}e^{-5bt} - \frac{1}{4}e^{-4at-bt}$
$p_3 = \frac{1}{5} - \frac{1}{5}e^{-5bt}$
$p_4 = \frac{1}{5} + \frac{4}{5}e^{-5bt}$

2.1.2 Parameter estimation

From the value of p_i we can derive: $p_4 - p_3 = e^{-5bt}$ and $p_1 - p_2 = e^{-4at-bt}$. Accordingly, we can make the following parameter estimates: $bt = -\frac{1}{5}\ln(p_4 - p_3)$, $at = \frac{1}{20}ln(p_4 - p_3) - \frac{1}{4}ln(p_1 - p_2)$.

2.1.3 Calculation of the adjacency pairs probability

For an evolutionary tree T with all leaf node genomes known, we complete the collection of adjacency-pairs according to the method in Section 3.2 and perform the screening of adjacency-pairs according to the method in Section 3.3. Based on the filtered list of adjacency-pairs, we encoded five states for each adjacency-pair of each genome separately. Suppose a is an internal node in an evolutionary tree T and a site s is one of the sites in the list of adjacent-pairs. D_a is the observed data for all leaves of the evolutionary tree at site s. We calculate the conditional probability of each state on each site s of the internal node a using the following Bayesian formula:

$$p(s_a|D_a) = \frac{P(s_a)P(D_a|s_a)}{P(D_a)} = \frac{P(s_a)P(D_a|s_a)}{\sum_{s_a} P(s_a)P(D_a|s_a)} = \frac{\pi_{s_a} L_a(s_a)}{\sum_{s_a} \pi_{s_a} L_a(s_a)} \tag{6}$$

$P(s_a)$ is the prior probability estimated from π_{s_a}, while π_{s_a} is the frequency of state s_a in the leaf genome. In the equation, $P(D_a)$ is calculated by the total probability of all states of the site s. $P(D_a|s_a)$ is expressed through $L_a(s_a)$, which is the conditional probability of the state s_a of the site s in the descendant of the internal node a. It is known that $L_a(s_a)$ can be computed recursively by taking a post-order traversal [26], and we compute the value of $L_a(s_a)$ according to the following equation:

$$L_a(s_a) = \begin{cases} 1 & if\, a\, is\, a\, leaf\, node\, with\, state = s_a\, at\, the\, site \\ 0 & if\, a\, is\, a\, leaf\, node\, with\, state \neq s_a\, at\, the\, site \\ \sum_{s_l} ps_a s_l L_l(s_l) \cdot \sum_{s_r} p_{s_a s_r} L_r(s_r) & otherwise \end{cases} \tag{7}$$

where $p_{s_a s_l}$ is the transition probability of state s_a to state s_l after evolution time t. After this step, we will obtain the probability of each state for each adjacency-pair of each internal node.