Simulate binary-state epigenome evoluiton

January 10, 2018

Assume the epigenome is a sequence of auto-correlated sequence of binary-state random variables. The auto-correlation reflects the organization of the epigenome as alternating domains bearing a certain modification or not. Let the epigenomic sequence be $S = s_1 s_2 \dots s_N$. As a graphical model, neighboring sites are connected with undirected edges. S evolves over time. Assume the stationary distribution for the epigenome has a Gibbs distribution that factorize over pairs of neighboring sites:

$$\Pr(S) = \frac{1}{Z} \exp\left\{\phi(s_1) + \sum_{n=1}^{N-1} \phi(s_n, s_{n+1}) + \phi(s_N)\right\} \tag{1}$$

The evolution of an individual site is context-dependent. The instantaneous mutation rate from state s to the alternative state $\bar{s}$ is $\gamma(l, s, r)$, where l, s and r are the states of three consecutive sites. For a time interval $[0, t)$, given that the states of l and r are not changed, then the states of site s follows a continouse-time Markov chain, and the holding time thus follows an exponential distribution. An observation of the path can be summarized with

$$L = \left\{s(0), k, \{t_i\}_{i=1}^{k}, t\right\},$$

where $s(0)$ is the state at time 0, k is the total number of jumps, t_i is the time when the i-th jump occurred, and t is the total length of the time interval.

Suppose L_l and L_r are given paths of two neighboring sites, then the union of their jumping times

$$\{0, t\} \cup \{t_{li}\}_{i=1}^{k_l} \cup \{t_{ri}\}_{i=1}^{k_r}$$

defines time intervals, within each of which the states of l and r stayed constant.

Stationary Gibbs measure as Markov chain The stationary distribution in (1) is euqivalent to the distribution of a Markov chain. We can derive the relationship between the factors in (1) and the transition probabilities of the Markov chain. The pair-wise potentials are $Q(a, b) = \exp(\phi(a, b))$, where $a, b \in \{0, 1\}$ are binary states. The largest eigen value of Q is

$$q = \frac{1}{2}\{Q_{00} + Q_{11} + \sqrt{\Delta}\}, \text{ where } \Delta = (Q_{00} - Q_{11})^2 + 4Q_{01}Q_{10}.$$

Let r be a right eigenvector of Q corresponding to q, then we have $\frac{r_0}{r_1} = \frac{Q_{01}}{q - Q_{00}} = \frac{q - Q_{11}}{Q10}$. Then the Markov chain transition matrix is

$$T(a, b) = \frac{Q(a, b)r(b)}{qr(a)}, \text{ where } a, b \in \{0, 1\}.$$

To be more specific,

$$\begin{aligned} T(1,1) &= \frac{2Q_{11}}{Q_{00}+Q_{11}+\sqrt{\Delta}}, \\ T(0,0) &= \frac{2Q_{00}}{Q_{00}+Q_{11}+\sqrt{\Delta}}, \\ T(0,1) &= \frac{4Q_{01}Q_{10}}{(Q_{00}+\sqrt{\Delta})^2-Q_{11}^2}, \\ T(1,0) &= \frac{4Q_{01}Q_{10}}{(Q_{11}+\sqrt{\Delta})^2-Q_{00}^2}. \end{aligned} \tag{2}$$

The expected methylation level is thus $1-\frac{2Q_{01}Q_{10}}{(Q_{00}-Q_{11})^2+4Q_{01}Q_{10}+(Q_{11}-Q_{00})\sqrt{\Delta}}$.

Relationship between mutation rates and stationary distribution What transition rate function γ can lead to a stationary distribution determined by ϕ? The Proposition 1 of Jensen & Pedersen (2000) gives a sufficient condition:

$$\frac{\gamma(l,s,r)}{\gamma(l,\bar{s},r)} = \frac{\exp(\phi(l,\bar{s})+\phi(\bar{s},r))}{\exp(\phi(l,s)+\phi(s,r))}, \tag{3}$$

which is derived from the reversibility property of the stationary distribution.

The proposition 2 and 3 give a way of specifying γ from ϕ. Assume that the log intensities can be written as

$$\log(\gamma(l,s,r)) = -g(l,s,r)+\ell(l,r),$$

and that there exists a function $q(l,r)$ such that

$$g(l,s,r) = g(l,s,*)-g(l,*,*)+g(s,r,*)-g(s,*,*)+q(l,r)$$

Then g bridges γ and ϕ with

$$\phi(l,s) = g(l,s,*)-g(l,*,*),$$

where '*' stands for averaged function value over all values of the indicated operands. So we only need to specify function g, which has 8 possible input configurations. Based on empirical understanding of epigenomes, we want g (and γ) to have left-right symmetry, *i.e.* $g(a,b,c) = g(c,b,a)$. Under this assumption, two pairs of configurations are equivalent, leaving 6 distinct configurations.

However, we can directly verify that if we define mutation rates as follows,

$$\log(\gamma(l,s,r)) = \ell(l,r)+(\phi(l,\bar{s})+\phi(\bar{s},r)), \tag{4}$$

where ℓ is some function independent of s, then the rates satisfy the condition in Equation 3.

We can organzie the mutaion rates into a 8×8 matrix as follows:

$$\Gamma = \begin{array}{c} \\ 000 \\ 010 \\ 001 \\ 011 \\ 100 \\ 110 \\ 101 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 010 & 001 & 011 & 100 & 110 & 101 & 111 \end{array} \\ \begin{pmatrix} . & a & 0 & 0 & 0 & 0 & 0 & 0 \\ b & . & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & . & c & 0 & 0 & 0 & 0 \\ 0 & 0 & d & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & . & c & 0 & 0 \\ 0 & 0 & 0 & 0 & d & . & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & . & e \\ 0 & 0 & 0 & 0 & 0 & 0 & f & . \end{pmatrix} \end{array} \tag{5}$$

If a methylome sequence with mutation rates Γ has stationary distribution (1), then the condition in (3) holds, which leads to the following constraints on the mutation rates:

$$ad^2e = bc^2f. \tag{6}$$

Given the mutation rates a, b, c, d, it is sufficient to derive the relationship between the potentials:

$$\phi(0,0) = \phi(0,1) + \frac{1}{2}\log(\frac{b}{a}), \text{ and } \phi(1,1) = \phi(0,1) + \frac{1}{2}\log(\frac{bc^2}{ad^2}). \tag{7}$$

In summary, the mutation rate matrix can have 5 free parameters. If we add a constraint on the expected number of changes per unit time, then there will be only 4 free parameters. The two ratios $\frac{b}{a}$ and $\frac{bc^2}{ad^2}$ can uniquely determine the stationary distribution (1) through equation (7).

Simulation scheme We model the evolution of the entire sequence that with a continuous time Makov chain that allows instantaneous jump from one sequence to another only if they differ at one position. The jumping rate at each position is dependent on its state and the states of its neighboring sites. We assume that the states of the first and last sites are fixed throughout evolution.

Consider the $2^N \times 2^N$ transition rate matrix M, for any methylome a and methylome b that have a single difference at a position where a has state j and b has state $\bar{j}$, and the neighboring positions have states i and k in both methylomes, the rate of such a jump is $\lambda_{ijk} = \gamma(i,j,k) = \Gamma_{ijk,i\bar{j}k}$.

Given the current methylome a, the holding time

$$X_a \sim Exp(-M_{aa})$$

is an exponential variable. Its rate parameter $-M_{aa}$ is the sum of all instantaneous rates for jumps from a to a methylome that only differs with a at one position:

$$-M_{aa} = \sum_{i,j,k} c_{ijk}(a)\lambda_{ijk},$$

where $c_{ijk}(a) = \sum_{n=1}^{N-2} I(a_n = i, a_{n+1} = j, a_{n+2} = k)$ is the total number of the tripplet pattern ijk in methylome a.

Given that the first jump happened, the probability that the jump occurred in the context of ijk is proportional to $c_{ijk}(a)\lambda_{ijk}$. Given that a jump happend in context ijk, the jump is equally likely among positions with this context. The expected number of changes per site per unit time is $\sum_{ijk} \pi_{ijk}\lambda_{ijk}$, where π_{ijk} is the stationary probability of pattern ijk in the methylome.

In summary, we have the following simulation procedure for the evolution process for a methylome with N sites over time interval $[0, T]$, given that the initial methylome is $a(0)$:

1. Let $t \leftarrow 0$, and initialize all paths $L_n = \{a_n(0), k_n = 0, T_n = \emptyset, t\}$, for $n = 1, \ldots, N$.
2. While $t < T$:
   (a) Generate $x \sim \text{Exp}(-M_{a(t)a(t)})$, where $-M_{a(t)a(t)} = \sum_{i,j,k} c_{ijk}(a(t))\lambda_{ijk}$.

   If $t + x < T$:
   - Choose pattern ijk from $\{ijk : i, j, k \in \{0, 1\}\}$ with probability proportional to $c_{ijk}(a(t))\lambda_{ijk}$.

- Scan methylome $a(t)$, uniformaly choose one position n out of the c_{ijk} positions all with the pattern ijk in $a(t)$.
- Set $a(t+x) \leftarrow a(t)_{1\ldots n-1}\overline{a(t)_n}a(t)_{n+1\ldots N}$.
- Add jump time to the path of position n:

$$k_n \leftarrow k_n + 1,\ \ T_n \leftarrow T_n \cup \{t+x\}.$$

Else $a(T) \leftarrow a(t)$.

(b) $t \leftarrow t + x$

Parameter inference When we're given the complete epigenome evolution path from time 0 to time t, can we effectively recover the initial distribution and mutation parameters and evolutionary time? We are interested in the parameters describing the evolutionary process, which are the mutation rates $\{\lambda_{ijk}\}$. These parameters will be inferred from the state changes in the evolutionary path of the entire epigenome. Meanwhile, we do not require the process to be stationary, so we are also interested in the epigenome properties at time 0, which are characterized by the Markov chain transition probabilities T_0. These transition probabilities are easy to infer given the complete observations at the time-0 epigenome.

Let $c_{ij} = \sum_{n=1}^{N-1} I\{s_n(0) = i, s_{n+1}(0) = j\}$. Then

$$\hat{T}(0,0) = \frac{c_{00}}{\sum_{n=1}^{N-1} I\{s_n(0)=0\}},\ \hat{T}(1,1) = \frac{c_{11}}{\sum_{n=1}^{N-1} I\{s_n(0)=1\}}.$$

Let's first assume that the time span of this complete evolutionary history is known, *i.e.* the value of t is give.

Recall that $L_n = \{s_n(0), K, \{t_k\}_{k=1}^K, t\}$ is a full path at position n in the epigenome. We can pool all the jumping times at all positions in to a sorted sequence $J = \{(t_m, \text{pos}_m, \text{context}_m\}_{m=1}^M$, where pos_m is the position of the m-th jump in the entire evolutionary history of the epigenome, context_m is the 3-tuple context of the mutation.

Let $\Delta_m = t_m - t_{m-1}$ be the holding time before the m-th jump. Then Δ_m is an exponential variable

$$\Delta_m \sim \text{Exp}(\lambda_m), \text{where } \lambda_m = \sum_{i,j,k} c_{ijk}(t_m - \epsilon)\lambda_{ijk},$$

where constant $\epsilon \in (0, \min_{1\le m\le M}\{\Delta_m\})$ so that $c_{ijk}(t_m - \epsilon)$ is the sequence context distribution between the $(m-1)$th jump and the m-th jump.

The likelihood function for parameters $\{\lambda_{ijk}\}$ is thus

$$\begin{aligned} L &= \prod_{m=1}^{M} \lambda_m \exp(-\lambda_m \Delta_m) \times \frac{\lambda_{\text{context}_m}}{\lambda_m} \\ &= \prod_{m=1}^{M} \lambda_{\text{context}_m} \exp(-\lambda_m \Delta_m) \end{aligned} \tag{8}$$

The log-likelihood function is

$$\begin{aligned} l &= \sum_{ijk} \Big(\sum_{m=1}^{M} \log \lambda_{ijk} \times I_{\{\text{context}_m = ijk\}} - c_{ijk}(t_m - \epsilon)\lambda_{ijk}\Delta_m \Big) \\ &= \sum_{ijk} \big(J_{ijk} \log \lambda_{ijk} - D_{ijk}\lambda_{ijk} \big) \end{aligned} \tag{9}$$

where $J_{ijk} = \sum_{m=1}^{M} I_{\{\text{context}_m = ijk\}}$, and $D_{ijk} = \sum_{m=1}^{M} c_{ijk}(t_m - \epsilon)\Delta_m$.

Constraints on the mutation rates $\{\lambda_{ijk}\}$ as indicated in equations (5,6) are

$$\begin{cases} \lambda_{001} = \lambda_{100} \\ \lambda_{011} = \lambda_{110} \\ \lambda_{000}\lambda_{110}^2\lambda_{101} = \lambda_{010}\lambda_{100}^2\lambda_{111} \end{cases} \tag{10}$$

Then the log-likelihood function (9) becomes

$$\begin{aligned} l = & J_{000}\log\lambda_{000} - D_{000}\lambda_{000} + \\ & J_{010}\log\lambda_{010} - D_{010}\lambda_{010} + \\ & J_{101}\log\lambda_{101} - D_{101}\lambda_{101} + \\ & (J_{100} + J_{001})\log\lambda_{001} - (D_{100} + D_{001})\lambda_{001} + \\ & (J_{011} + J_{110})\log\lambda_{011} - (D_{011} + D_{110})\lambda_{011} + \\ & J_{111}\log(\frac{\lambda_{000}\lambda_{011}^2\lambda_{101}}{\lambda_{010}\lambda_{001}^2}) - D_{111}\frac{\lambda_{000}\lambda_{011}^2\lambda_{101}}{\lambda_{010}\lambda_{001}^2}. \end{aligned} \tag{11}$$

When the time span of this complete evolutionary history is unknown, the value of t is also a model parameter to be estimated. We assume that all the jumping times are expressed as a fraction of t. Then the problem of estimating the mutation rates and evolutionary time becomes identifiable. We need an extra constraint – the unit branch length corresponds to 1 expected mutation per site. This is a common constraint in phylogenetic studies.

Here we explain how to formulate this constraint on the $\{\lambda_{ijk}\}$ parameters. Given $\{\lambda_{ijk}\}$, according to (7) $\frac{\lambda_{010}}{\lambda_{000}}$ and $\frac{\lambda_{001}}{\lambda_{011}}$ can uniquely determine the stationary distribution of the epigenome that is described with a Gibbs measure of form (1). The Gibbs measure for the epigenomic sequence, in turn, is equivalent to a Markov chain (2). Given the Markov chain formulation, we can easily compute the expected abundance of triplet patterns $p_{ijk} = \frac{1}{N}\mathbb{E}(c_{ijk})$, where c_{ijk} is the frequency of the triplet pattern in an epigenomic sequence of length N sampled from the Gibbs distribution. Then, we can compute the expected number of changes per position per unit time as $\sum_{ijk} p_{ijk}\lambda_{ijk}$. When the evolutionary time is also unknown, we add the following constraint:

$$\sum_{ijk} p_{ijk}\lambda_{ijk} = 1, \tag{12}$$

where $\{p_{ijk}\}$ as explained above are functions of $\{\lambda_{ijk}\}$. We maximize the log-likelihood (9) over $\{\lambda_{ijk}\} \cup \{t\}$ under the constraints (10) and (12), where the unknown parameter t is buried within $\{D_{ijk}\}$ in (9).

References

Jensen JL, Pedersen AMK (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Advances in Applied Probability* 32:499–517.