Simulate binary-state epigenome evoluiton

November 14, 2017

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:

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

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.

Simulation scheme We are going to simulate a full history of epigenome evolution for a time interval $[0, t]$.

1. Simulate the starting methylome using a binary-state Markov model
2. Initialize all paths $L_n = \{s_n(0), k_n = 0, T_n = \emptyset, t\}$, for $n = 1, \dots, N$.
3. (For simplicity, fix the paths L_1 and L_N as initialized.) For site $n = 2, \dots, N-1$, simulate L_n given the current paths of L_{n-1} and L_{n+1}:
    - Collect the time intervals from site $n-1$ and site $n+1$, so that within each of the intervals the states of the neighboring sites are unchanged. Let the intervals be represented by a sorted array $\{t_0, t_1, \cdots, t_M\}$.

- Let x be a random variable from an exponential distribution, representing the jumping time of the middle site given that the states of its two neighbors are constant. For a time interval $[t_{m-1}, t_m]$ during which the neighboring sites' states are unchanged, suppose $x \sim \text{Exp}(\lambda)$. The parameter λ is a function of current state and the states of the two neighboring sites.
- Generate a binary observation I with probability $p = \Pr(x < t_m - t_{m-1})$
- If $I = 1$, a jump happens within the time interval $[t_{m-1}, t_m]$.
  (1) Add 1 to the number of jumps $k_n \leftarrow k_n + 1$.
  (2) Generate a random variable from the truncated exponential distribution, for example using importance sampling. Let the sample value be $t' \in (0, t_m - t_{m-1})$. Update $T_n \leftarrow T_n \cup \{t_{m-1} + t'\}$.
  (3) Then update t_{m-1} with $t_{m-1} + t'$, i.e. the new time interval is $[t_{m-1} + t', t_m]$. Update the exponential parameter accordingly, because the current state has changed. Then repeat from the previous step in the outer loop.
- If $I = 0$, there is no jump during this time interval, the middle site keeps its state unchanged till time t_m. If $t_m = t_M$, STOP. Otherwise, we move onto the next time interval $[t_m, t_{m+1}]$ and repeat from the previous step in the outer loop.

4. Repeat last step, until the epigenome summary statistics converge to a stable distribution.

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 **?** gives a sufficient condition:

$$\frac{\gamma(l, s, r)}{\gamma(l, \bar{s}, r)} = \frac{\exp(\phi(l, \bar{s}) + \phi(\bar{s}, r))}{\exp(\phi(l, s) + \phi(s, r))}, \tag{1}$$

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. Let the values of g be as specified in table 1.

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

$$\log(\gamma(l, s, r)) = \ell(l, r) + (\phi(l, \bar{s}) + \phi(\bar{s}, r)), \tag{2}$$

where ℓ is some function independent of s, then the rates satisfy the condition in Equation 1.

	Mutation type in patterns (g parameter)	
γ **level**	$0 \rightarrow 1$	$1 \rightarrow 0$
low	0,0,0 (x_1)	1,1,1 (y_1)
medium	0,0,1 (x_2)	1,1,0 (y_2)
medium	1,0,0 (x_2)	0,1,1 (y_2)
high	1,0,1 (x_3)	0,1,0 (y_3)

Table 1: Level of mutation rates in different patterns