Simulate binary-state epigenome evoluiton

December 12, 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:

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

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

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

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

1 Simulation scheme 1

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 $n = 1, \ldots, N$.
3. (For simplicity, fix the paths L_1 and L_N as initialized.) For site $n = 2, \ldots, N-1$, 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 $\{t_0, t_1, \cdots, t_M\}$.
   - For $m = 1, \ldots, M$:
     - 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.
     - Let $t = t_{m-1}$, and a sample value of $X = x$.
     - While $t + x < t_m$:
       (1) Add 1 to the number of jumps $k_n \leftarrow k_n + 1$.
       (2) Update $T_n \leftarrow T_n \cup \{t + x\}$.
       (3) Update $t \leftarrow t + x$,
       (4) Sample another value of $X = x$ from $\text{Exp}(\lambda)$.
4. Repeat step 3, until the epigenome summary statistics converge to a stable distribution.

	Mutation type in patterns (g parameter)	
γ **level**	0→ 1	1→ 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

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

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

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

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

2 Simulation scheme 2

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, and the rate of such jumps are dependent on the states of their neighboring sites. Same as before, we assume that the states of the starting and ending sites are fixed.

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 λ_{ijk}. The rate parameters are symetric for i and k, i.e. $\lambda_{ijk} = \lambda_{kji}$. Thus, we have 6 rate parameters.

Given the current methylome a, the holding time

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

is an exponential variable. The exponential 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. Specifically,

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

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 to have occured at any positions with this context.

To summarize, 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 $a(0)$:

```
• $t \leftarrow 0$.
• While $t < T$:
    1. Generate $x \sim$ 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 $a(t + x) \leftarrow a(t)_{1\ldots n-1}\overline{a(t)_n}a(t)_{n+1\ldots N}$.
       Else: $a(T) \leftarrow a(t)$.
    2. $t \leftarrow t + x$
```

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.