Maximum likelihood Phylogenetic model

Jenny Qu

October 5, 2015

1 Model and notation

A methylome is a sequence of CpG sites (or bins of CpG sites) with each site displaying a methylation state from the set $\{0, 1\}$. During evolution, two processes jointly govern the methylation state inheritance across generations, and the correlation between neighboring sites within each generation.

The inheritance process is described with a continuous-time Markov process over the state space $\{0, 1\}$. Let the initial distribution be $\pi = (\pi_0, \pi_i)$. Let the transition rate matrix matrix be

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \eta & -\eta \end{bmatrix}.$$

In continuous-time Markov process, the transition probability between two time points separated by time t is determined by two terms $(t(\lambda + \eta), \lambda/\eta)$. Therefore, we let $\eta = 1 - \lambda$, $\lambda \in (0, 1)$. The transition probability matrix becomes

$$P(t) = exp(Qt) = \begin{bmatrix} 1 - \lambda T & \lambda T \\ (1 - \lambda)T & 1 - (1 - \lambda)T \end{bmatrix},$$

where $T = 1 - exp(-t) \in (0, 1)$ for $t > 0$.

The autocorrelation within a methylome is described with a discrete-time (corresponding to discrete CpG sites) Markov chain over the state space $\{0, 1\}$. The initial distribution is π (only relevant in the root species). The transition probability matrix,

$$G = \begin{bmatrix} g_0 & 1 - g_0 \\ 1 - g_1 & g_1 \end{bmatrix}$$

is assumed to be homogeneous in all species (this assumption can be relaxed).

For a given set of extant species and their phylogenetic relationship, we use $\tau = \{V, E\}$ to denote the phylogenetic tree, including the set of nodes, and the set of edges E. We assume that the tree structure is given, and the branch lengths are unknown parameters.

The model parameter space is thus

$$\Theta = \{E, \lambda, g_0, g_1, \pi_0\},$$

where E is the collection of branch lengths, and λ is the transition rate from 0 to 1 in the continuous-time Markov process. π_0 is the hypomethylation probability in the root node.

The two Markov processes are combined in the following way. The direction of inheritance process is obvious. For the discrete Markov process, let the first CpG site from the 5' end of the '+' strand to be the start position of a chain. The root methylome has no ancestor, so it is modeled only with the discrete time Markov chain described by $\{\pi, G\}$. For CpG sites at start position of a DNA fragment, the evolution of methylation states on this site is modeled only with the inheritance process described with $\{\pi, Q, \tau\}$. For CpG sites from internal nodes and leaf nodes that are not the start of a DNA fragment, we will combine the two processes to model its evolution. Let one such site be position j in node v. Let $i = j - 1$ be the previous site in the DNA fragment. Let node u be v's parent in the phylogenetic tree. Use the notation u_j to denote the methylation state of node u at position j, $u_j \in \{0, 1\}$. Let $t_v = e(u, v)$ be the branch length between nodes u and v. We define the following transition probability:

$$\Pr(v_j|v_i, u_j) = \frac{G_{v_i v_j} P(t_v)_{u_j v_j}}{\sum_{s=0,1} G_{v_i s} P(t_v)_{u_j s}} \tag{1}$$

Assume hypomethylation probabilities at N sites are observed in the methylome at each leaf node of a phylogenetic tree. Let O_i be the observed hypomethylation probabilities at location i in all extant species, $i = 1, \ldots, N$. Let $O = O_1 O_2 \cdots O_N$ be the total observed methylation states at leaf nodes. Let $S = S_1 \cdots S_N$ be the unobserved methylation states at internal nodes.

2 Incomplete data

We have observations of bisulfite read counts, or precomputed hypomethylation probability at leaf nodes. There is no prior information about the methylation states at internal nodes.

2.1 Data Likelihood

The above description defines a hidden Markov model over the space

$$H = \{h \in \{0, 1\}^{|V|} : \text{all possible combination of methylation states over the phylogenetic tree}\}.$$

Each state specifies the history of the methylation evolution at CpG site, and thus we call it History of Methylation Evolution (HME). The transition probability between two HMEs can be calculated using transition probabilities in the previous section. Let h_i, h_j be the HMEs at two neighboring sites in the genome. Let the root node be r. Let v be an internal or leaf node in the phylogenetic tree, and let its parent node be u. Let r_i be the methylation state at node r as specified by HME h_i. Let the HME transition probability matrix be $A = \{a_{ij}\}$, where

$$a_{ij} = \Pr(h_j|h_i) = G_{r_i r_j} \prod_{(u,v) \in E} P(v_j|v_i, u_j).$$

The data likelihood calculation can be written as

$$L = P(O|\Theta) = \sum_{H_1, \ldots, H_N \in H} \Pr(H_1) \Pr(O_1|H_1) \prod_{i=1}^{N} \Pr(H_i|H_{i-1}, \Theta) \Pr(O_i|H_i), \tag{2}$$

where H_t is a HME random variable at position t. H is the HME space.

Forward algorithm

Alternatively, we can use the forward algorithm for likelihood calculation. Define the forward variables as below:

$$\begin{aligned} \alpha_1(i) &= \Pr(O_1, H_1 = h_i|\Theta) = \Pr(h_i|\Theta)\Pr(O_1|h_i) \\ \alpha_{n+1}(j) &= \sum_{i=1,\ldots,|H|} \alpha_n(i)a_{ij} \cdot \Pr(O_{n+1}|h_i) \end{aligned} \tag{3}$$

The data likelihood is thus

$$L = P(O|\Theta) = \sum_{i=1,\ldots,|H|} \alpha_N(i).$$

And the log-likelihood is

$$l = \log L = \log \sum_{i=1,\ldots,|H|} \alpha_N(i).$$

2.2 First order derivative

Our parameter space is $\Theta = \{E, \lambda, g_0, g_1, \pi_0\}$. Let $E' = \{e' = 1 - exp(-e) : \forall e \in E\}$. There is a 1-1 mapping between E and E', so the model parameter space is equivalently $\Theta = \{E', \lambda, g_0, g_1, \pi_0\}$. For simplicity, we use the same notation Θ. Let x be a parameter of interest.

$$\frac{\partial l}{\partial x} = \frac{1}{L} \sum_{i=1,\ldots,|H|} \frac{\partial}{\partial x}\alpha_N(i)$$

Recursion by forward algorithm Using the recursive definition of forward variables,

$$\frac{\partial}{\partial x}\alpha_{n+1}(j) = \sum_{i=1,\ldots,|H|} \big[\frac{\partial}{\partial x}\alpha_n(i) \cdot a_{ij} + \alpha_n(i) \cdot \frac{\partial}{\partial x}a_{ij}\big] \cdot \Pr(O_{n+1}|h_j),$$

the first order derivative computation comes down to computing derivatives for HME transition probabilities a_{ij}, and the base case forward variables $\alpha_1(i)$, where $i, j \in \{1, \ldots, |H|\}$.

Some Preparation The continuous-time Markov process' transition probability matrix corresponding to a given branch (u, v) contains two parameters: λ, and $T_u = 1 - exp(-t_u)$. The derivatives with respect to these two parameters are

$$\begin{aligned} \frac{\partial}{\partial \lambda}P(t) &= T \cdot \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \\ \frac{\partial}{\partial T}P(t) &= Q \end{aligned} \tag{4}$$

where $T = 1 - exp(-t)$.

The discrete-time Markov chain involves two parameters, g_0 and g_1.

$$\begin{aligned} \frac{\partial}{\partial g_0}G_{ij} &= \delta_{\{i=0\}} \cdot (2 * \delta_{\{j=0\}} - 1) \\ \frac{\partial}{\partial g_1}G_{ij} &= \delta_{\{i=1\}} \cdot (2 * \delta_{\{j=1\}} - 1) \end{aligned} \tag{5}$$

The combined transition probability $\Pr(v_j|v_i, u_j)$ contains parameter T_v, g_{v_i}, and λ.

$$\begin{aligned}
\frac{\partial}{\partial T_v}\Pr(v_j|v_i,u_j) &= \frac{G_{v_iv_j}\cdot\frac{\partial}{\partial T_v}P(t_v)_{u_jv_j}}{[\sum_{s=0,1}G_{v_is}P(t_v)_{u_js}]} - \frac{G_{v_iv_j}P(t_v)_{u_jv_j}\cdot[\sum_{s=0,1}G_{v_is}\frac{\partial}{\partial T_v}P(t_v)_{u_js}]}{[\sum_{s=0,1}G_{v_is}P(t_v)_{u_js}]^2} \\
\frac{\partial}{\partial g_{v_i}}\Pr(v_j|v_i,u_j) &= \frac{\frac{\partial}{\partial g_{v_i}}G_{v_iv_j}\cdot P(t_v)_{u_jv_j}}{[\sum_{s=0,1}G_{v_is}P(t_v)_{u_js}]} - \frac{G_{v_iv_j}P(t_v)_{u_jv_j}\cdot[\sum_{s=0,1}\frac{\partial}{\partial g_{v_i}}G_{v_is}P(t_v)_{u_js}]}{[\sum_{s=0,1}G_{v_is}P(t_v)_{u_js}]^2} \\
\frac{\partial}{\partial \lambda}\Pr(v_j|v_i,u_j) &= \frac{G_{v_iv_j}\cdot\frac{\partial}{\partial \lambda}P(t_v)_{u_jv_j}}{[\sum_{s=0,1}G_{v_is}P(t_v)_{u_js}]} - \frac{G_{v_iv_j}P(t_v)_{u_jv_j}\cdot[\sum_{s=0,1}G_{v_is}\frac{\partial}{\partial \lambda}P(t_v)_{u_js}]}{[\sum_{s=0,1}G_{v_is}P(t_v)_{u_js}]^2} \\
&= \frac{G_{v_iv_j}T_v\cdot(2\delta_{\{v_j=1\}}-1)}{[\sum_{s=0,1}G_{v_is}P(t_v)_{u_js}]} - \frac{G_{v_iv_j}P(t_v)_{u_jv_j}T_v[\sum_{s=0,1}G_{v_is}(2\delta_{\{s=1\}}-1)]}{[\sum_{s=0,1}G_{v_is}P(t_v)_{u_js}]^2}
\end{aligned} \tag{6}$$

Derivative of λ

$$\frac{\partial}{\partial\lambda}\alpha_1(i) = \alpha_1(i)\times\big[\sum_{(u,v)\in E}\frac{T_v\cdot(2\delta_{\{v_i=1\}}-1)}{P(t_v)_{u_iv_i}}\big]$$

$$\frac{\partial}{\partial\lambda}a_{ij} = a_{ij}\times\big[\sum_{(u,v)\in E}\frac{\frac{\partial}{\partial\lambda}P(v_j|v_i,u_j)}{P(v_j|v_i,u_j)}\big]$$

Derivative of branch length T_v

$$\frac{\partial}{\partial T_v}\alpha_1(i) = \alpha_1(i)\times\frac{Q_{u_iv_i}}{P(t_v)_{u_iv_i}}$$

$$\frac{\partial}{\partial T_v}a_{ij} = a_{ij}\times\frac{\frac{\partial}{\partial T_v}\Pr(v_j|v_i,u_j)}{\Pr(v_j|v_i,u_j)}$$

Derivative of transition probability g_s

$$\frac{\partial}{\partial g_s}a_{ij} = a_{ij}\times\big[\delta_{\{r_i=s\}}\times\frac{(2\delta_{\{v_j=s\}}-1)}{G_{r_ir_j}} + \sum_{(u,v)\in E}\delta_{\{s=v_i\}}\times\frac{\frac{\partial}{\partial g_{v_i}}P(v_j|v_i,u_j)}{P(v_j|v_i,u_j)}\big]$$

Derivative of initial distribution π_0

$$\frac{\partial}{\partial\pi_0}\alpha_1(i) = \alpha_1(i)\times\frac{2\delta_{\{r_i=0\}}-1}{\pi_{r_i}}$$

All other partial derivatives of $\alpha_n(i)$ and a_{ij} that are not specified above have value 0.

2.3 Reduce transition complexity

The total number of possible transitions between HMEs is $2^{2|V|}$. We introduce constraints on transition between HMEs to reduce the computational complexity.

Neighbor HME Let $N(h) = \{h' \in H : P(H_{n+1} = h' | H_n = h) > 0\}$, be the neighbor set of an HME h. We always allow self-transition, *i.e.* $h \in N(h), \forall h \in H$. For any two HMEs, they are neighbors if one HME can be converted to the other HME by picking a subtree and forcing all nodes in the subtree to have a same methylation state.

By this definition, the complete hypomethylated or hypermethylated HME is neighbor with all HMEs. Different HMEs might have different number of neighbors. The neighbor definition implies a symmetric relationship, *i.e.* if $h \in N(h')$ then $h' \in N(h)$. The restricted transition probabilities are defined as below:

$$a'_{ij} = 1_{h_j \in N(h_i)} \frac{a_{ij}}{\sum_{k \in N(i)} a_{ik}}$$

. The relevant partial derivatives should be recalculated as:

$$\frac{\partial}{\partial x} a'_{ij} = 1_{h_j \in N(h_i)} \left[\frac{\frac{\partial}{\partial x} a_{ij}}{\sum_{k \in N(i)} a_{ik}} - \frac{\sum_{k \in N(i)} \frac{\partial}{\partial x} a_{ik}}{\left(\sum_{k \in N(i)} a_{ik} \right)^2} \right]$$

where parameter $x \in \{\lambda, g_0, g_1\} \bigcup \{T_v : v \in V\}$.

3 Approximation method

In the incomplete data situation, all possible transitions between HMEs have to be considered, which give rise to a complexity of $O(c^n)$ where n is the number of nodes in the phylogenetic tree.

We would like to get around the computational complexity by approximating the transition probabilities between HME with a product of weighted transition probabilities along each branch, with the weight being a function of the posterior hypomethylation probabilities at relevant nodes. The work flow is described below:

1. Initialize hypomethylation probability at internal nodes
2. Compute likelihood and gradient
3. Optimize parameter
4. Update posterior hypomethylation probability at internal nodes using Markov blanket and the updated model parameters
5. Repeat step 2-4 until convergence of parameter.

3.1 Likelihood and gradient

Let $s_u^{(i)} \in \{0, 1\}$ be the methylation sate of node u on the phylogenetic tree at position i in the methylome. Let $p_{s_u}^{(i)}$ be the observed or posterior probability of node u having methylation state s_u at position i.

Likelihood and gradient for the first site in the genomic fragment is

$$L_0 = \sum_{s_r} p_{s_r} \pi_{s_r} \prod_{(u,v) \in E} \left[\sum_{s_u, s_v} p_{s_u} p_{s_v} \Pr(T_{u,v})_{s_u s_v} \right]$$

$$\frac{\partial \log L_0}{\partial x} = \frac{1}{L_0} \sum_{s_r} \left(p_{s_r} \pi_{s_r} \prod_{(u,v) \in E} \left[\sum_{s_u, s_v} p_{s_u} p_{s_v} \Pr(T_{u,v})_{s_u s_v} \right] \right) \times \left(\frac{\frac{\partial \pi_{s_r}}{\partial x}}{\pi_{s_r}} + \sum_{(u,v) \in E} \frac{\sum_{s_u, s_v} p_{s_u} p_{s_v} \frac{\partial \Pr(T_{u,v})_{s_u s_v}}{\partial x}}{\sum_{s_u, s_v} p_{s_u} p_{s_v} \Pr(T_{u,v})_{s_u s_v}} \right)$$

Transition probability between site i and $i+1$ is

$$L_{i,i+1} = \sum_{s_r^{(i-1)} s_r^{(i)}} p_{s_r^{(i)}} p_{s_r^{(i+1)}} G_{s_r^{(i)} s_r^{(i+1)}} \prod_{(u,v)\in E} \Big[\sum_{s_v^{(i)} s_u^{(i+1)} s_v^{(i+1)}} p_v^{(i)} p_u^{(i+1)} p_v^{(i+1)} \Pr(s_v^{(i+1)} | s_v^{(i)}, s_u^{(i+1)}) \Big]$$

$$\frac{\partial}{\partial x} \log L_{i,i+1} = \frac{1}{L_{i,i+1}} \sum_{s_r^{(i)} s_r^{(i+1)}} \left(p_{s_r^{(i)}} p_{s_r^{(i+1)}} G_{s_r^{(i)} s_r^{(i+1)}} \prod_{(u,v)\in E} \Big[\sum_{s_v^{(i)} s_u^{(i+1)} s_v^{(i+1)}} p_v^{(i)} p_u^{(i+1)} p_v^{(i+1)} \Pr(s_v^{(i+1)} | s_v^{(i)}, s_u^{(i+1)}) \Big] \right) \times$$
$$\left(\frac{\frac{\partial}{\partial x} G_{s_r^{(i)} s_r^{(i+1)}}}{G_{s_r^{(i)} s_r^{(i+1)}}} + \sum_{(u,v)\in E} \frac{\sum_{s_v^{(i)} s_u^{(i+1)} s_v^{(i+1)}} p_v^{(i)} p_u^{(i+1)} p_v^{(i+1)} \frac{\partial}{\partial x} \Pr(s_v^{(i+1)} | s_v^{(i)}, s_u^{(i+1)})}{\sum_{s_v^{(i)} s_u^{(i+1)} s_v^{(i+1)}} p_v^{(i)} p_u^{(i+1)} p_v^{(i+1)} \Pr(s_v^{(i+1)} | s_v^{(i)}, s_u^{(i+1)})} \right)$$

The overall log-likelihood and gradients are

$$\log L = \log L_0 + \sum_{i=0}^{N-1} \log L_{i,i+1}$$

$$\frac{\partial}{\partial x} \log L = \frac{\partial}{\partial x} \log L_0 + \sum_{i=0}^{N-1} \frac{\partial}{\partial x} \log L_{i,i+1}$$

3.2 Update posterior

All CpGs in all the species in a genomic region form a directed graph. Each branch either links an ancestral CpG to a descendent CpG, or links one CpG to the next CpG in the same species. Ideally, we would like to have the joint posterior distribution for the methylation states at all CpGs, or the marginal posterior distribution at each node.

We update the marginal posterior iteratively through all the nodes, using the Markov blanket of each node. There are 9 different cases. The joint distribution of nodes in the Markov blanket is approximated with the product of their marginal distributions.

Case 1: node v_{cur} is root in the start position. Nodes in its Markov blanket are the three children $l_{cur}, m_{cur} r_{cur}$, and the v_{next}. Let B the set of joint states for nodes in v_{cur}'s Markov blanket.

$$p_b = p_{l_{cur}} p_{m_{cur}} p_{r_{cur}} p_{v_{next}}, \forall\, b \in B.$$

For each combination of Markov blanket state, the normalized conditional hypomethylation probability is calculated by

$$\begin{aligned} &\Pr(v_{cur} = u | b) \\ \propto &\Pr(l_{cur} | v_{cur} = u) \Pr(m_{cur} | v_{cur} = u) \Pr(r_{cur} | v_{cur} = u) \Pr(v_{next} | v_{cur} = u) \end{aligned}$$

The updated posterior hypomethylation probability is $p'_{v_{cur}}$

$$p'_{v_{cur}} = \sum_b p_b \Pr(v_{cur} = u | b)$$

Case 2: node v_{cur} is root and in the end position. Its Markov blanket includes l_{cur}, m_{cur}, r_{cur}, l_{prev}, m_{prev}, r_{prev} and v_{prev}.

$$\Pr(v_{cur} = u | b) \propto \Pr(v_{cur} = u | v_{prev}) \prod_{c \in children(v)} \Pr(c_{cur} | v_{cur} = u, c_{prev})$$

Case 3: node v_{cur} is a root node in the middle. Its Markov blanket includes c_{cur}, c_{prev}, where $c \in children(v)$, v_{prev} and v_{next}.

$$\Pr(v_{cur} = u|b) \propto \Pr(v_{cur} = u|v_{prev}) \Pr(v_{next}|v_{cur} = u) \prod_{c \in children(v)} \Pr(c_{cur}|v_{cur} = u, c_{prev})$$

Case 4: node v_{cur} is a leaf node in the start position. Its Markov blanket includes v_{next}, and m_{cur}, m_{next}, where m is the parent of node v.

$$\Pr(v_{cur} = u|b) \propto \Pr(v_{cur} = u|m_{cur}) \Pr(v_{next}|v_{cur} = u, m_{next})$$

Case 5: node v_{cur} is a leaf node in the end position. Its Markov blanket includes v_{prev} and m_{cur}, where m is the parent of node v.

$$\Pr(v_{cur} = u|b) \propto \Pr(v_{cur} = u|v_{prev}, m_{cur})$$

Case 6: node v_{cur} is a leaf node in the middle. Its Markov blanket includes v_{prev}, v_{next}, m_{cur} and m_{next}.

$$\Pr(v_{cur} = u|b) \propto \Pr(v_{cur} = u|v_{prev}, m_{cur}) \Pr(v_{next}|v_{cur}, m_{next})$$

Case 7: node v_{cur} is an internal node in the start position. Its Markov blanket includes $c_{cur}(c \in \text{children}(v))$, v_{next}, m_{cur} and m_{next}.

$$\Pr(v_{cur} = u|b) \propto \Pr(v_{cur} = u|m_{cur}) \Big(\prod_{c \in \text{children}(v)} \Pr(c_{cur} = u|v_{cur} = u) \Big) \Pr(v_{next}|v_{cur}, m_{next})$$

Case 8: node v_{cur} is an internal node in the start position. Its Markov blanket includes $c_{prev}, c_{cur}(c \in \text{children}(v))$, v_{prev} and m_{cur}.

$$\Pr(v_{cur} = u|b) \propto \Pr(v_{cur} = u|m_{cur}, v_{prev}) \Big(\prod_{c \in \text{children}(v)} \Pr(c_{cur} = u|v_{cur} = u, c_{prev}) \Big)$$

Case 9: node v_{cur} is an internal node in the middle. Its Markov blanket includes $v_{prev}, c_{prev}, c_{cur}(c \in \text{children}(v))$, v_{next}, m_{cur} and m_{next}.

$$\Pr(v_{cur} = u|b) \propto \Pr(v_{cur} = u|m_{cur}, v_{prev}) \Big(\prod_{c \in \text{children}(v)} \Pr(c_{cur}|v_{cur} = u, c_{prev}) \Big) \Pr(v_{next}|v_{cur} = u, m_{next})$$