A model of epigenome evolution

Jianghan Qu Andrew D. Smith

April 4, 2018

Abstract

Epigenetic marks along the mammalian genome are organized into alternating genomic domains bearing and lacking the mark. The location and size of domains enriched for an epigenetic mark are indicative of the presence, function and activity of regulatory elements and the chromatin states. Comparative epigenomic studies aim to resolve the evolutionary history of regulatory elements by comparing epigenomic profiles in multiple species. However, computational methods for comparing epigenetic marks at high resolution, inferring evolution rates along different phylogenetic lineages and reconstructing the evolutionary history are still limited. In this study, we aim to establish a simulation, sampling and inference framework for studying the evolution of the genomic distribution of an epigenetic from the profiles of multiple extant species. We model the profile of an epigenetic mark in a species with a two-state Markov chain, and model the evolution of an epigenomic sequence with a continuous-time Markov chain, where instantaneous transition rates at a site is dependent on the contemporary states of its neighboring sites. We use a MCMC sampling method for estimating the context-dependent transition rates and inferring the evolutionary history that lead to diverse profiles in extant species from a common ancestral epigenome. We show with applications to DNA methylation and histone modification profiles that our methods can reveal both genome-wide evolutionary features through estimates of the model parameters and high-resolution evolutionary patterns in local regions through posterior sampling of the evolutionary history.

1 Introduction

The epigenome of a mammalian cell reflects much of the complexity we associate with cell phenotype and behaviors (). Individual epigenomic marks, for example a histone modification, may be viewed from a simpler perspective as contiguous genomic intervals where the presence or absense of that mark is associated with genomic function. Intervals of the genome that have a high density of H3K9me3 are often associated with condensed chromatin state and silencing of genes within those intervals (Nakayama et al., 2001). Genomic intervals with high density of H3K4me3, on the other hand, are associated with accessibility by transcription factors and are associated with gene promoters (Santos-Rosa et al., 2002). So despite the complexity often ascribed to the mammalian epigenome (Bernstein et al., 2007), studies focusing on individual epigenomic modifications have been highly successful in elucidating transcriptional regulation in a variety of systems (Martin & Zhang, 2005).

One challenge in modeling epigenome evolution is that desirable models should account for the inherent auto-correlation of epigenomic state along the genome. The most well known models of molecular evolution, applied to amino acids or nucleotides, treat each site as evolving independently – a simplifying assumption that has proven very useful. When models allow for dependencies between sites, we additionally hope that those dependencies can be interpreted.

Pedersen et al. (1998) examined the problem of modeling evolution at the codon level and designed

an approach capable of describing CpG depression across codon boundaries, a form of interdependence between adjacent codons. With similar motivation, Jensen & Pedersen (2000) examined the properties of evoluationary models for which the stationary distribution on sequences naturally exhibits particular frequencies of dinucleotides. The result was an approach to model the evolution of an individual nucleotide as a function of that nucleotide's neighbors in a way that induces a Markov process on the stationary distribution.

Our goal of modeling the evolution of an epigenome can be viewed in analogy with phylogenetic hidden Markov models (). The phastCons algorithm () is the best known variant of phylogenetic HMM, and has dramatically impacting the field of comparative genomics by providing a general approach to model "conserved" genomic intervals for a set of species. The states of a simple phylogenetic HMM correspond to alternating conserved and non-conserved intervals of aligned genomes. The generalization associates a binary state label with each nucleotide in each species and presents algorithmic challenges when viewed generally as a chain graph (Koller & Friedman, 2009).

The remainder of this paper is organized as follows. We first describe our model for epigenome evolution as an adaptation of the principles introduced by Jensen & Pedersen (2000) and use simulation fo explore model parameterization. Then we explain how to inferences are made in the context of this model, using simulation to demonstrate the accuracy of our procedures. Finally, we apply this model on an existing data set.

2 The model

2.1 Biological assumptions and assumed representation for the epigenome

We assume that the epigenome is a sequence of binary states super-imposed on the genome. This assumption is restrictive, but it allows us to consider the epigenome from either the perspective of an individual epigenomic modification, or as reflecting a particular type of functional interval. One example of former is a sequence of binary variables corresponding to the presence or absence of a H3K27me3. Another example is a binary variable to indicate accessibility, as determined by a particular assay. An example of a functionally-defined binary variable may be "accessible" or "enhancer," both of which can be associated with different epigenomic modifications, but are known to be organized as contiguous intervals. Epigenomic state is correlated along the genome as a reflection of the organization of the epigenome into contiguous intervals.

Most types of data that inform us about the epigenome are based on sequencing, and usually based on density of mapped reads (e.g. from ChIP-seq or ATAC-seq). Often these data are summarized in non-overlapping "bins" through the genome. When we refer to a position in the epigenome, we assume that such position corresponds in a meaningful way to either individual nucleotide positions in the genome, or appropriate bins. We only require that the neighboring relations are preserved.

Let epigenome s be the sequence $s = s_1 s_2 \cdots s_N$ with s_i denoting the state at position i. As a graphical model, neighboring sites are connected with undirected edges. An epigenome s evolves over time, but we assume the stationary distribution for the epigenome has a Gibbs distribution that factorize over pairs of neighboring sites:

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

As the epigenome evolves the state at each position may change. Although we have operationally defined the epigenome by associating a state with each position along the genome, the types of changes we hope to model are more naturally interpreted in terms of sets of contiguous intervals. We use the term epigenomic

"feature" to refer to a consecutive interval having the "1" state. The main types of changes we are interested in are the following:

- *Birth and death:* During evolution, some new epigenomic feature may appear, or a feature that existed in an ancestor may disappear.
- *Expansion and contraction:* As the epigenome evolves, epigenomic segments can become wider or more narrow, and this may happen in either direction.
- *Merging and separating:* Two epigenomic features that are nearby in the ancestral genome may merge into a single interval. Conversely, a single epigenomic feature in the ancestral epigenome may separate into two intervals.

We will describe a model that treats the expansion and contraction of features symmetrically in both directions along the genome. This choice in modeling is often reasonable, but not always. For example, certain epigenomic modifications are frequently associated with parts of genes, which have directionality.

For a site n and an evolutionary time interval $[0, \tau)$, the initial state and all the time points of state transitions within this time interval constitute a complete evolutionary history for this site. This is because we have assumed binary states, and an evolutionary path is consecutive time intervals with alternating states. In other words, an observation of the evolutionary path at site n can be summarized as

$$L_n = \big\{ s_n(0), J_n = \{t_1, \ldots, t_{m_n}\}, \tau \big\},$$

where $s_n(t) \in \{0, 1\}$ is the state of site n at time t, τ is the total length of the time interval, and J_n is an ordered sequence of m_n jummping times at site n within this time interval, *i.e.* $0 < t_1 < \cdots < t_{m_n} < \tau$.

The evolution of states at a given site depends on the states of its neighboring sites during evolution. Consider site n and its two immediate neighbors $n-1$ and $n+1$, and let the contermporaneous states at these sites be $s_{n-1}(t) = i$, $s_n(t) = j$ and $s_{n+1} = j$. The instantaneous rate for a transition from state j to its alternative state $\bar{j}$ at site n is $\gamma(i, j, k)$. In other words, the instantaneous rates $\gamma(i, j, k)$ is a function of the states at the site of interest and its two immediate neighbors. In addition, since we assume the evolutionary process is symmetric with respect to the two directions along the genome, we require $\gamma(i, j, k) = \gamma(k, j, i)$. When the states at all positions except for site n stay constant for a time interval $[0, \tau)$, the state at site n, s_n, follows a two-state continuous-time Markov process within this time interval.

2.2 Stationary Gibbs measure as Markov chain

The stationary distribution in (1) is equivalent to the distribution of a Markov chain. Therefore, when the epigenome is modeled with a Gibbs measure, we can sample an instance of the sequence epigenomic states or evaluate its probability with the equivalent Markov chain. We can derive the relationship between the factors in (1) and the transition probabilities of a Markov chain. The pair-wise potentials are $Q_{ij} = \exp(\phi(i, j))$, where $i, j \in \{0, 1\}$ are binary states. The largest eigenvalue of Q is

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

Let h be a right eigenvector of Q corresponding to q, then we have

$$\frac{h_0}{h_1} = \frac{Q_{01}}{q - Q_{00}} = \frac{q - Q_{11}}{Q_{10}}$$

Then we have the Markov chain transition matrix:

$$T(i,j) = \frac{Q_{ij}h_j}{qh_i}, \text{ where } i,j \in \{0,1\}.$$

More specifically,

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

The expected fraction of an epigenome residing within functional domains is thus:

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

2.3 Relating the substitution model and the stationary distribution

Jensen & Pedersen (2000) gave a sufficient condition for the continuous time evolutionary model γ to have a stationary distribution determined by ϕ. In particular, if

$$\frac{\gamma(i,j,k)}{\gamma(i,\bar{j},k)} = \frac{\exp(\phi(i,\bar{j}) + \phi(\bar{j},k))}{\exp(\phi(i,j) + \phi(j,k))}, \tag{3}$$

then (1) is the stationary distribution for the Markov process with intensities $\gamma(i,j,k)$.

Proposition 2 of Jensen & Pedersen (2000) provides a way of specifying γ from ϕ: substitution rates γ satisfy the relation (4) if and only if they can be written in the form

$$\log(\gamma(i,j,k)) = -\psi(i,j,k) + \ell(i,k). \tag{4}$$

This criteria was introduced in the context of an arbitrary number of states, and the ℓ function can be understood as $\ell(j,j';i,k)$, for any two different states j and j', which is symmetric in (j,j'). In our setting of modeling epigenomic features, the states are binary and ℓ is only a function of the two neighboring states (i,k). Moreover, we can directly verify that if we define substitution rates as

$$\log(\gamma(i,j,k)) = \ell(i,k) + (\phi(i,\bar{j}) + \phi(\bar{j},k)), \tag{5}$$

then they satisfy the condition of (4).

2.4 Parameterization and interpretation

We can organize the neighbor-dependent transition rates of defined above according to an 8×8 matrix:

$$\Gamma = \begin{array}{c} \\ 000 \\ 010 \\ 001 \\ 011 \\ 100 \\ 110 \\ 101 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 010 & 001 & 011 & 100 & 110 & 101 & 111 \end{array} \\ \begin{pmatrix} \cdot & \mathcal{B} & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathcal{D} & \cdot & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & \mathcal{E} & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathcal{C} & \cdot & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdot & \mathcal{E} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathcal{C} & \cdot & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdot & \mathcal{M} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathcal{S} & \cdot \end{pmatrix} \end{array} \tag{6}$$

We may interpret the non-zero entries in Γ as corresponding to biological events outlined in Section 2.1. The values $\mathcal{B}$ and $\mathcal{D}$ are the rates of "birth" and "death," respectively, for epigenomic features. The value $\mathcal{M}$ corresponds to the merging of two features into a single contiguous interval ($101 \to 111$). Conversely, the value $\mathcal{S}$ corresponds to epigenomic features "splitting" and becoming two separate intervals ($111 \to 101$). The remaining non-zero values, $\mathcal{E}$ and $\mathcal{C}$, correspond to the expansion (widening) and contraction (narrowing), of epigenomic features. Both of these parameters appear twice in Γ, reflecting our assumption that the rates governing any widening or narrowing of intervals do not depend on direction, as explained in Section 2.1.

If an epigenome evolves according to substitution rates Γ with stationary distribution (1), then the condition in (4) holds, and we have the following constraints for the rates in the above matrix:

$$\mathcal{B}\mathcal{C}^2\mathcal{M} = \mathcal{D}\mathcal{E}^2\mathcal{S}. \tag{7}$$

So given substitution rates $\mathcal{B}, \mathcal{D}, \mathcal{E}, \mathcal{C}$, we have the following relationships between horizontal potentials:

$$\phi(0,0) = \phi(0,1) + \frac{1}{2}\log\left(\frac{\mathcal{D}}{\mathcal{B}}\right), \text{ and } \phi(1,1) = \phi(0,1) + \frac{1}{2}\log\left(\frac{\mathcal{D}\mathcal{E}^2}{\mathcal{B}\mathcal{C}^2}\right). \tag{8}$$

In summary, the transition rate matrix Γ has 5 free parameters. If we add an additional constraint on the expected number of changes per unit time, then the model will have only 4 free parameters. The two ratios $\mathcal{D}/\mathcal{B}$ and $\mathcal{D}\mathcal{E}^2/(\mathcal{B}\mathcal{C}^2)$ then uniquely determine the stationary distribution (1) through equation (8).

3 Simulate epigenomic changes during evolution

We model the evolution of the entire epigenome (as a sequence of states) using a continuous time Markov process that only allows instantaneous transitions from one sequence to another if the two differ at a single position. The jumping rate at each position in the epigenome is dependent on the state at that position and on the states of the left and right neighboring positions. We assume for convenience 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 pair of epigenomes x and y that differ at exactly one position. The state at that position is j in epigenome x and $\bar{j}$ in epigenome y. Neighboring positions in both x and y have states i and k. The instantaneous rate of a jump between x and y is

$$\lambda_{ijk} = \gamma(i,j,k) = \Gamma(ijk, i\bar{j}k).$$

If the current epigenome is denoted x then the holding time is an exponential variable:

$$X_x \sim Exp(-M_{xx}).$$

The rate parameter $-M_{xx}$ is the sum of instantaneous rates for jumps from x to any other epigenome that only differs from x at one position:

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

where $c_{ijk}(x) = \sum_{n=1}^{N-2} I(x_n = i, x_{n+1} = j, x_{n+2} = k)$ is the total number of times the pattern ijk appears as consecutive triplets of positions in x.

Given that a jump has occurred, the probability that the jump changed x at the middle position of a triple ijk is proportional to $c_{ijk}(x)\lambda_{ijk}$. Further, given that a jump happend with context ijk, we assume the jump

Algorithm 1 Simulating epigenome evolution

$t \leftarrow 0$
Initialize paths $L_n = \{x_n(0), k_n = 0, T_n = \emptyset, t\}$, for $n = 1, \ldots, N$
while $t < T$ **do**
  Generate $y \sim Exp(-M_{x(t)x(t)})$, where $-M_{x(t)x(t)} = \sum_{i,j,k} c_{ijk}(x(t))\lambda_{ijk}$.
  if $t + y < T$ **then**
    Choose triple $ijk \in \{0,1\}^3$ with probability proportional to $c_{ijk}(x(t))\lambda_{ijk}$.
    Sample position n uniformly from the $c_{ijk}(x(t))$ positions having pattern ijk
    $x(t+y) \leftarrow x(t)[1..n-1]\overline{x(t)_n}x(t)[n+1..N]$.
    $k_n \leftarrow k_n + 1$
    $T_n \leftarrow T_n \cup \{t + y\}$
  else
    $x(T) \leftarrow x(t)$
  end if
  $t \leftarrow t + y$
end while

is equally likely to have changed any position in x having state j with left and right neighbors having states i and k. The expected number of changes per site, per unit time, is $\sum_{ijk} \pi_{ijk}\lambda_{ijk}$, where π_{ijk} is the stationary distribution for the pattern ijk in the epigenome. These assumptions suggest a simulation procedure for the evolutionary process followed by an epigenome of N sites over a time interval $[0, T]$, starting from an initial sequence $x(0)$.

To implement this simulation scheme requires maintaining the variables c_{ijk} at each time point, and updating them by removing

4 Parameter estimation

If we are 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 transition rates $\{\lambda_{ijk}\}$. These parameters will be inferred from the state changes in the evolutionary path for the entire epigenome. Meanwhile, we do not require the process to be stationary, so we may also interested in the properties of the epigenome 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

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

We first assume that the time span of this complete evolutionary history is known, *i.e.* the value of t is given.

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 as an ordered sequence of timepoints $J = \{(t_m, \mathrm{pos}_m, \mathrm{con}_m\}_{m=1}^M$, where pos_m is the position of the m-th jump in the entire evolutionary history of the epigenome, con_m is the 3-tuple context of the change that occurred at the m-th jump.

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

$$\Delta_m \sim Exp(\lambda_{(m)}), \text{ with } \lambda_{(m)} = \sum_{i,j,k} c_{ijk}(t_m^-)\lambda_{ijk}.$$

The time point t_m^- indicates the instant before the m-th jump, so that $c_{ijk}(t_m^-)$ is the sequence context distribution between the $(m-1)$-th jump and the m-th jump.

4.1 Likelihood expressions

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

$$L = \prod_{m=1}^{M} \lambda_{(m)} \exp(-\lambda_{(m)}\Delta_m) \times \frac{\lambda_{\text{con}_m}}{\lambda_{(m)}} = \prod_{m=1}^{M} \lambda_{\text{con}_m} \exp(-\lambda_{(m)}\Delta_m). \tag{9}$$

And the log-likelihood function is

$$\begin{aligned} l &= \sum_{i,j,k} \left(\sum_{m=1}^{M} \log \lambda_{ijk} \times I_{\{\text{con}_m = ijk\}} - c_{ijk}(t_m^-)\lambda_{ijk}\Delta_m \right) \\ &= \sum_{ijk} \left(J_{ijk} \log \lambda_{ijk} - D_{ijk}\lambda_{ijk} \right) \end{aligned} \tag{10}$$

where $J_{ijk} = \sum_{m=1}^{M} I_{\{\text{con}_m = ijk\}}$, and

$$\begin{aligned} D_{ijk} &= \sum_{m=1}^{M} c_{ijk}(t_m^-)\Delta_m \\ &= \sum_{m=1}^{M} \left(\sum_{n=1}^{N} I_{\{\text{con}(n;t_m^-)=ijk\}} \right) \times (t_m - t_{m-1}) \\ &= \sum_{n=1}^{N} \int_0^{t_M} I_{\{\text{con}(n;t)=ijk\}} dt \\ &= \text{Total time in context } ijk \text{ aggregated over all sites} \end{aligned} \tag{11}$$

where $\text{con}(n;t)$ is the sequence context of site n at time t.

The constraints on the transition rates λ_{ijk} as indicated in the matrix (6) and equation (7) are:

$$\begin{aligned} \lambda_{001} &= \lambda_{100} \\ \lambda_{011} &= \lambda_{110} \\ \lambda_{000}\lambda_{110}^2\lambda_{101} &= \lambda_{010}\lambda_{100}^2\lambda_{111} \end{aligned} \tag{12}$$

So the log-likelihood function (10) becomes

$$\begin{aligned} l = & \left(J_{000}\log\lambda_{000} - D_{000}\lambda_{000}\right) + \left(J_{010}\log\lambda_{010} - D_{010}\lambda_{010}\right) + \left(J_{101}\log\lambda_{101} - D_{101}\lambda_{101}\right) + \\ & \left((J_{100}+J_{001})\log\lambda_{001} - (D_{100}+D_{001})\lambda_{001}\right) + \left((J_{011}+J_{110})\log\lambda_{011} - (D_{011}+D_{110})\lambda_{011}\right) + \\ & J_{111}\log\left(\frac{\lambda_{000}\lambda_{011}^2\lambda_{101}}{\lambda_{010}\lambda_{001}^2}\right) - D_{111}\frac{\lambda_{000}\lambda_{011}^2\lambda_{101}}{\lambda_{010}\lambda_{001}^2} \\ = & \sum_{c=0,2,5}\left(J_c\log\lambda_c - D_c\lambda_c\right) + \sum_{c=1,3}\left((J_c+J_{c'})\log\lambda_c - (D_c+D_{c'})\lambda_c\right) + \\ & \left(J_7\log(\frac{\lambda_0\lambda_3^2\lambda_5}{\lambda_2\lambda_1^2}) - D_7\cdot\frac{\lambda_0\lambda_3^2\lambda_5}{\lambda_2\lambda_1^2}\right), \end{aligned} \tag{13}$$

where c' is the symmetric context pattern of c, *i.e.* if c is 001, then c' is 100.

4.1.1 Gradients of log-likelihood

We treat $\{\log\lambda_c : c = 0, 1, 2, 3, 5\}$ as free parameters as they can take any real value.

$$\begin{aligned} \frac{\partial l}{\partial\log\lambda_0} &= J_0 - D_0\lambda_0 + J_7 - D_7\lambda_7 \\ \frac{\partial l}{\partial\log\lambda_2} &= J_2 - D_2\lambda_2 - J_7 + D_7\lambda_7 \\ \frac{\partial l}{\partial\log\lambda_5} &= J_5 - D_5\lambda_5 + J_7 - D_7\lambda_7 \\ \frac{\partial l}{\partial\log\lambda_1} &= J_1 + J_4 - (D_1+D_4)\lambda_1 - 2J_7 + 2D_7\lambda_7 \\ \frac{\partial l}{\partial\log\lambda_3} &= J_3 + J_6 - (D_3+D_6)\lambda_3 + 2J_7 - 2D_7\lambda_7 \end{aligned} \tag{14}$$

4.1.2 Scaling parameters

After the estimates are made, we can scale the transition rates and branch lengths to have unit branch length corresponding to 1 expected transition per site. Given λ_c according to (8), the ratios $\frac{\lambda_{010}}{\lambda_{000}}$ and $\frac{\lambda_{001}}{\lambda_{011}}$ uniquely determine the stationary distribution for the epigenome described by a Gibbs measure of the form (1). The Gibbs measure for the epigenome sequence, in turn, is equivalent to the Markov chain (2). Given the Markov chain formulation, we can compute the expected frequency of triplet patterns

$$p_{ijk} = \pi_i T[i,j]T[j,k].$$

Then the expected number of changes per position per unit time is $\mu = \sum_{ijk} p_{ijk}\lambda_{ijk}$. We can scale transition rates and branch lengths as follows:

$$\begin{aligned} \lambda_{ijk} &\leftarrow \lambda_{ijk}\times\frac{1}{\mu} \\ t_b &\leftarrow \sum_{m=1}^{M_b}(t_{b,m} - t_{b,m-1})\times\mu. \end{aligned} \tag{15}$$

4.1.3 Estimates when evolutionary time is a parameter

In the context of phylogenetic inference, we are given the observations at extant species, *i.e.* leaf nodes of the phylogenetic tree, and will rely on EM procedures to both estimate model parameters (M-step) and making inferences about evolutionary paths form the common ancestor (E-step). In the E-step, we compute the expected values of the complete-data sufficient statistics Js and Ds along individual branches conditional on the observed leaf data and our current model parameters (branch lengths and transition rates). In M-step, we find the set of parameters that give the best likelihood with the expected values of sufficient statistics. For an Exponential random variable X:

$$X \sim Exp(\lambda) \Leftrightarrow X/t \sim Exp(\lambda t), \text{ for any constant } t > 0.$$

Let our current estimates of branch lengths be $\{\ell_b\}$, the new branch lengths be $\{\ell'_b\}$, and the new rates be $\{\lambda'_c\}$. Then for a jumping interval Δ_{bm} on branch b observed when the branch length is ℓ_b, the scaled interval length is $\Delta_{bm} \times \frac{\ell'_b}{\ell_b}$ under the new branch length. Therefore, under the new model parameter set $\{\ell'_b\}$ and $\{\lambda'_c\}$, we have

$$\Delta_{mb} \times \frac{\ell'_b}{\ell_b} \sim Exp(\lambda'_{(bm)}) \Rightarrow \Delta_{mb} \sim Exp(\lambda'_{(bm)} \times \frac{\ell'_b}{\ell_b}).$$

Let $\tau_b = \frac{\ell'_b}{\ell_b}$ be the scaling factor for the new branch lengths relative to the old branch lengths. Now we can write the likelihood of observing all of the given jumping intervals $\{\Delta_{b,m}\}$ under the new parameter set as :

$$\begin{aligned} L(\{\Delta_{b,m}\}; \{\ell'_b\}, \{\lambda'_c\}) &= \prod_{b=1}^{B} \left(\prod_{m=1}^{M_b-1} \lambda'_{(b,m)} \tau_b \exp(-\lambda'_{(b,m)} \tau_b \Delta_m) \times \frac{\lambda'_{\mathrm{con}_m}}{\lambda'_{(m)}} \right) \times \exp(-\lambda'_{(b,M_b)} \tau_b \Delta_{M_b}) \\ &= \prod_{b=1}^{B} \left(\prod_{m=1}^{M_b-1} \lambda'_{\mathrm{con}_m} \tau_b \exp(-\lambda'_{(m)} \tau_b \Delta_m) \right) \times \exp(-\lambda'_{(b,M_b)} \tau_b \Delta_{M_b}). \end{aligned} \tag{16}$$

The log likelihood becomes

$$\begin{aligned} l &= \sum_{b=1}^{B} \sum_{m=1}^{M_b-1} \log(\lambda'_{c_m} \tau_b) - \sum_{b=1}^{B} \sum_{m=1}^{M} \lambda'_{(m)} \Delta_{bm} \tau_b \\ &= \sum_{b=1}^{B} \sum_{c} J_{bc} \log(\lambda'_c \tau_b) - \sum_{b=1}^{B} \sum_{c} D_{b,c} \lambda'_c \tau_b, \end{aligned} \tag{17}$$

where $J_{b,c} = \sum_{m=1}^{M_b-1} I_{\{c_{b,m}=c\}}$, $D_{bc} = \sum_{m=1}^{M_b} I_{\{c_{b,m}=c\}} \Delta_{b,m}$, $c_{b,m}$ is the context of the m-th jump on branch b, and Δ_{bm} is the holding time before the m-th jump on branch b of length ℓ_b. Then, we can optimize (17) over $\{\tau_b\}$ and $\{\lambda'_c\}$, and set new branch lengths $\{\ell'_b = \tau_b \ell_b\}$.

4.2 Posterior distribution of a path given two neighboring paths

Here we consider inference on the path for an single position in an epigenome that is evolving for some specified time. We assume the paths for all other positions are known for the entire evolutionary time span. In particular, we will make inferences about site n and have access to the jump times $J = \{(t_m, \mathrm{pos}_m, \mathrm{con}_m)\}_{m=1}^{M}$

are known for all positions except position n. We also assume that the model parameters $\{\lambda_{ijk}\}$ and total evolutionary time t are known, along with the the initial state of the epigenome. We seek inferences related to the posterior distribution of the path L_n:

$$\Pr(L_n|L_{-n}, \{\lambda_{ijk}\}) \propto \Pr(L_n \cup L_{-n}),$$

where L_{-n} denotes all paths for sites other than n. The approach we adopt is using MCMC to sample from the posterior distribution of L_n.

Method 1: First, sample a starting state s_0 from a Bernoulli distribution with probabilities (π_0, π_1). Then, propose a number K of jumps from a Poisson distribution with rate parameter $\lambda = \sum \lambda_{ijk}/8$, which is chosen to approximate average mutation rate among different contexts. Given K, sample jump times uniformly on the time interval $(0, t)$; In other words, from the Dirichlet distribution with concentration parameters all equal to 1. Therefore, the probability (density) of proposing a specific path $L' = \{s_0, K, \{t_k\}_{k=1}^K, t\}$ is

$$q(L') = \pi_{s_0} \frac{\lambda^K \exp(-\lambda)}{K!} \frac{1}{(K-1)!}.$$

This is known as an *independence sampler*(). If the current path at position n is L_n, then we accept the proposed path L' with probability

$$\alpha(L') = \min\left\{ \frac{\pi(L', L_{-n})/q(L')}{\pi(L_n, L_{-n})/q(L_n)}, 1 \right\}, \tag{18}$$

where π is the complete data likelihood function (9).

Method 2: The method outlined above uses a proposal distribution that is approximately uniform. This can be highly inefficient, *i.e.* lead to a low rate of acceptance, although Jensen & Pedersen (2000) used a sampling procedure in the same spirit. If we use more information from the paths at neighboring sites, we may be able to improve sampling efficiency. The neighboring paths L_{n-1} and L_{n+1} partition the evolutionary time interval $(0, t)$ into time segments during which the states at positions $n-1$ and $n+1$ do not change.

- 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 $\{t_0, t_1, \cdots, t_M\}$.
- Initialize proposal probability $p \leftarrow 1$.
- 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 $X \sim Exp(\lambda_{ijk})$, where i and k are the states of the two neighboring sites in this time interval, and j is the starting state of position n. Let f_λ be the p.d.f. of Exponential distribution $Exp(\lambda)$.
  - Let $t = t_{m-1}$, and a sample value of $X = x$.
  - While $t + x < t_m$:

(1) Update proposal probability $p \leftarrow p \times f_{\lambda_{ijk}}(x)$.
(2) Update the state of the center site $j \leftarrow \bar{j}$.
(3) Add 1 to the number of jumps $k_n \leftarrow k_n + 1$.
(4) Update $T_n \leftarrow T_n \cup \{t + x\}$.
(5) Update $t \leftarrow t + x$,
(6) Sample another value of $X = x$ from $Exp(\lambda_{ijk})$.

– Update proposal probability $p \leftarrow p \times \Pr_{\lambda_{ijk}}(X > t_m - t)$.

The proposal probability density function $q()$ is thus the product of the appropriate Exponential distribution probability densities for the holding times at position n, and is calculated as p. The proposal distribution is independent of any current guess of the path. Therefore, this is also an *independence sampler*. The rejection rule stays the same, as in (18).

Note: For two versions of complete evolutionary history that only differ at one position, their difference in likelihood is determined by the paths at that position, and two positions to each side, *i.e.* 5 positions in total. Because the total time spent in each triplet context (11) is altered at the center position and its two neighboring positions. Let n be the center position. Let $D_{n,c}$ be the total time spent in triplet context c at position n in the current instance of evolutionary paths, and let $D'_{n,c}$ be that of the new proposed evolutionary path. Similarly let $J_{n,c}$ and $J'_{n,c}$ be the total number of jumps out of context c at position n in the current and proposed paths. Thus, the ratio of complete data likelihood in acceptance probability (18) can be computed as

$$\frac{\pi(L', L_{-n})}{\pi(L_n, L_{-n})} = \prod_c \prod_{s=n-1}^{n+1} \lambda_c^{J'_{s,c} - J_{s,c}} \exp(-(D'_{s,c} - D_{s,c})\lambda_c).$$

5 Inferences in the context of a tree structure with fixed leaf data

Now we assume the epigenome has evolved according to a tree structure, and we wish to make inferences about a bifurcating evolutionary process. The paths associated with a given site must satisfy additional constraints to be "consistent." We first consider a tree with 3 nodes, a root u with a single child node v and two leaf nodes below v, denoted w_1 and w_2. The branches of the tree are (u, v), (v, w_1) and (v, w_2). We assume we are given the epigenome at each node in this tree, with the exception of the state at position n of node v. We are also given all paths for sites other than site n. We want to make inferences about the state at position n of v, and also about the paths for edges incident on $v(n)$. We require a method for sampling paths at site n that share a common state at their common end-point.

Hobolth & Stone (2009) reviewed and compared three approaches to sample paths of discrete-state continuous-time Markov chain conditional on end-point states, namely the rejection sampling, direct sampling and uniformization. Method 1 in the previous section is similar to uniformization, since the neighboring sites may have state changes, and the rate of jumps depends on the contemporary states of the neighboring sites. We proposed to use an averaged rate to for the Poisson distribution, and let the acceptance probability to do most of the work of correcting the proposal distribution towards the posterior distribution. Method 2 is forward sampling, which is the basis of rejection sampling.

Rejection sampling for the proposal: This approach first does forward sampling as described in Method 2 above for position n along the branch (u, v) and (v, w_1), which can be viewed as a single path that proceeds

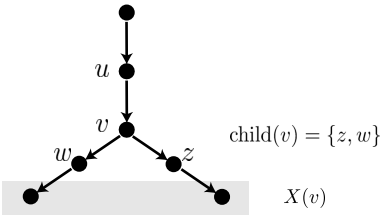

Figure 1: Tree with break points as nodes

along two consecutive edges. The sampled path provides a state x corresponding to node v. Next, a path is sampled for edge (v, w_2) assuming state x at node v. If this path finishes with the known state at position n in w_2, we retain the 3 paths $L_n^{(u,v)}$, $L_n^{(v,w_1)}$ and $L_n^{(v,w_2)}$ as a proposal. The proposal is evaluated using the same acceptance rule given above in equation (18).

A couple of performance statistics should be measured for this proposed sampling method. These include (1) the success rate for generating valid paths that satisfy the end conditions; (2) the acceptance rate of proposed valid paths in MCMC sampling.

5.1 Posterior sampling of a path on the entire phylogenetic tree

For a phylogenetic tree with known branch lengths and known evolutionary context-dependent transition rates, how can we sample the evolutionary path (on the entire tree) for one site from its posterior distribution given observed evolutionary paths at all other sites and the observed states at all leaf nodes for this site?

States at break points. The evolution process for this site is not homogeneous along the tree, its transition rate matrix depends on the concurrent states of neighboring sites. We can, however, partition each branch into intervals, within each of which the states at the two neighboring sites stay unchanged. Thus the evolutionary process for the site of interest is homogeneous within each interval, and can be characterized with a 2×2 transition rate matrix Q_{bm} for the m-th interval on branch b. The states of this site of interest at the start and the end of a single interval, $s_{b,m-1}$ and $s_{b,m}$ are connected through a transition probability matrix $P_{bm} = \exp(Q_{bm} t_{bm})$. The likelihood of observing leaf node states at this site $\{s_n(v)\}$, given the evolutionary paths at other sites is

$$\Pr(\{s_n(v)\}|L_{-n}) = \sum_{\{s_{bm}\}} \prod_b \prod_m P_{bm}(s_{b,m-1}, s_{b,m})$$

Bottom-up pruning algorithm. We can use Felsenstein's pruning algorithm to compute this probability. Let the phylogenetic tree be τ consisting of all leaf nodes, internal nodes, and all break points between intervals on individual branches as nodes. Let v be a node in the phylogenetic tree, and $X_n(v)$ be the states of all leaf nodes that are descendants of v (Figure 5.1) at site n in the genome. Let u be the parent of v, and w, z be the children of v. Since the sequence contexts are known for the current site at all intervals, the transition rate matrix for the state of the current site is known and denoted with $Q_{\mathrm{con}(v)}$. Thus the transition probability matrix between any parent-child pair (u, v) is known, which we denote with

$$P_v = \exp(Q_{\mathrm{con}(v)} \ell_v).$$

For node v, given that its parent u has state j, let the conditional probability of observing states $X_n(v)$ at terminal descendants of node v be $p_j(v)$, *i.e.*

$$p_j(v) = \Pr(X_n(v)|s_u = j).$$

For notational convenience we define

$$q_k(v) = \begin{cases} \Pr(s_v = k) & \text{if } v \text{ is a leaf node,} \\ \prod_{c\in\text{child}(v)} p_k(c) & \text{otherwise,} \end{cases}$$

where $\Pr(s_v = k) \in \{0, 1\}$ when a state is observed, and $\Pr(s_v = k) \in (0, 1)$ when the observed data are continuous levels representing a probability distribution over the state space. Then, we can use Felsenstein's pruning algorithm to iteratively compute this probability from bottom of the tree upwards to the root node:

$$p_j(v) = \sum_k \Big(P_v[j, k] \times q_k(v) \Big). \tag{19}$$

In this way, we can compute the marginal posterior probability of observing the states $X_n(v)$ at leaf nodes at site n, given the evolutionary paths at all other sites:

$$\Pr(X_n(r)|L_{-n}) = \sum_j \Pr(s_r = j|L_{-n}) q_j(r) \tag{20}$$

At the root node, the marginal posterior distribution of the state is

$$\begin{aligned} \Pr(s_r|L_{-n}, X_n(r)) =& \frac{\Pr(s_r, X_n(r)|L_{-n})}{\Pr(X_n(r)|L_{-n})} = \frac{\Pr(X_n(r)|s_r, L_{-n}) \Pr(s_r|L_{-n})}{\Pr(X_n(r)|L_{-n})} \\ =& \frac{\big(\prod_{c\in\text{child}(r)} p_{s_r}(c)\big) \Pr(s_r|L_{-n})}{\Pr(X_n(r)|L_{-n})}, \end{aligned} \tag{21}$$

where $\Pr(s_r|L_{-n})$ is the initial distribution of the root node state given its sequence context, which we can compute from the stationary distribution of the root epigenomic sequence characterized by either a Gibbs measure, or a Markov chain.

In summary, during the bottom-up Pruning algorithm 2, we compute $p_j(v), q_k(v)$, and ultimately the marginal probabilities $\Pr(X_n(r)|L_{-n})$ and $\Pr(s_r|L_{-n}, X_n(r))$.

Top-down posterior sampling at breakpoints. Let S be the states at all internal nodes. The posterior probability of all internal nodes taking the states S, given the leaf node states $X_n(r)$ and the evolutionary paths at all other sites, is

$$\Pr(S|X_n(r), L_{-n}) = \frac{\Pr(S, X_n(r)|L_{-n})}{\Pr(X_n(r)|L_{-n})},$$

where $\Pr(S, X_n(r)|L_{-n})$ is the forward sampling probability under the given transition probability matrices in individual intervals determined by L_{-n}. Therefore, we can do exact posterior sampling of the joint states at the break points between the intervals on the phylogenetic tree through direct sampling following Algorithm 3.

Algorithm 2 Pruning algorithm with observed leaf states $X_n(r)$

for node v in post-order traversal **do**
  if $v \neq r$ **then**
    if v is a leaf node **then**
      Assign $q_k(v) \leftarrow I\{\Pr(s_v = k)\}$ for $k = 0, 1$.
    else
      Compute $q_k(v) \leftarrow \prod_{c \in \text{child}(v)} p_k(c)$
    end if
    Compute $p_j(v) \leftarrow \sum_k \left(P_v[j,k] \times q_k(v) \right)$, for $j = 0, 1$.
  else
    Compute $\Pr(X_n(r)|L_{-n})$ as in (20), and $\Pr(s_r|L_{-n}, X_n(r))$ as in (21)
  end if
end for

Algorithm 3 Posterior sampling at breakpoints

Sample the root state from distribution $\Pr(s_r|L_{-n}, X_n(r))$ as in (21).
for internal node v in pre-order traversal **do**
  Suppose its parent node u has state j
  Sample state k from the distribution

$$k \sim \frac{1}{p_j(v)} \times P_v[j,k] \times q_k(v)$$

  Update the direct sampling probability with factor $\frac{P_v[j,k] q_k(v)}{p_j(v)}$
end for
The result is an sample S of states at all internal states, which is sampled from their joint marginal posterior probability distribution $\Pr(S|X_n(r), L_{-n})$.

Direct sampling of path between breakpoints with fixed states. Next, we sample the state transitions at this position within each single time interval, during which its two neighbors have unchanged states, and the initial and final states for this time interval are fixed. Focusing on one interval, let 0 and τ be the end time points. Let an instance of the path within this interval be $L_n = \{t_m\}_{m=0}^M$, which is a set of jumping times and the start and end time points ($t_0 = 0$ and $t_M = \tau$). Let $\Delta_m = t_m - t_{m-1}$ be the holding time between jumps. Let the homogeneous transition rate matrix be Q for this site in this time interval, and $P(t)$ be the transition probability matrix for states at two time points separated by time t. We use the direct sampling procedure (Hobolth & Stone, 2009) for endpoint-conditioned path sampling.

If the endpoints have the same state i, the probability that there are no change during this time interval is

$$p_i = \frac{exp(Q_{ii}\tau)}{P(\tau)_{ii}}. \tag{22}$$

Then with probability $1 - p_i$, at least one (actually two) state change occurs.

Algorithm 4 Direct sampling of end-conditioned path

Suppose interval has length τ, and start and end states i, j:
while $\tau > 0$ **do**
  if $i = j$ **then**
    Sample $Z \sim$ Bernoulli(p_i), where p_i is given in (22).
    if $Z = 1$ **then**
      Set $s(t) = i$ for all $0 < t < \tau$, and update $\tau \leftarrow 0$.
    end if
  end if
  if $i \neq j$ or $Z = 0$ **then**
    Sample the waiting time W in state i according to the continuous density $f_{ij\tau}(w)$, $0 < w < \tau$, set $s(t) = i$ for $0 < t < W$
    Update $i \leftarrow \bar{i}$, $\tau \leftarrow \tau - W$.
  end if
end while

Let W be the waiting time until the first jump.

$$\begin{aligned}
&\Pr(W < \tau | s(0) = i, s(\tau) = j) \\
&= \Pr(W < \tau, s(W) = \bar{i} | s(0) = i, s(\tau) = j) \\
&= \frac{\Pr(W < \tau, s(W) = \bar{i}, s(\tau) = j | s(0) = i)}{\Pr(s(\tau) = j | s(0) = i)} \\
&= \int_0^\tau Q_{i\bar{i}} exp(Q_{ii} w) \frac{P(\tau - w)_{\bar{i}j}}{P(\tau)_{ij}} dw \\
&= \int_0^\tau f_{ij\tau}(w) dw, \text{ where } f_{ij\tau}(t) = Q_{i\bar{i}} exp(Q_{ii} t) \frac{P(\tau - t)_{\bar{i}j}}{P(\tau)_{ij}}.
\end{aligned} \tag{23}$$

The direct sampling procedure for a homogeneous continuous-time Markov path within an end-conditioned interval is detailed in Algorithm 4.

6 Discussion

As noted earlier, the generalizations previously described for phylo-HMMs have many similarities with the model we proposed. In particular, these generalizations associate a binary variable with each nucleotide in each sequence, including ancestral sequences. The most notable difference is our emphasis on interpretability of the horizontal relationships, specifically that an individual epigenome (extant or ancestral) follows a Markov process. Another difference is our adoption of a homogeneous continuous time process, which is both built into the model, and exploited by our inference procedure. Phylogenetic HMMs are applied to determine states along aligned genomes where those genome sequence evolve according to a particular substitution model. Our notion of the epigenome corresponds more closely to the idea of "conserved" and "non-conserved" states, but the details of our approaches are more similar to the substitution process of nucleotides, rather than general time-dependent graphical models.

References

Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. *Cell* 128:669–681.

Hobolth A, Stone EA (2009) Simulation from endpoint-conditioned, continuous-time markov chains on a finite state space, with applications to molecular evolution. *The annals of applied statistics* 3:1204.

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

Koller D, Friedman N (2009) *Probabilistic graphical models: principles and techniques* MIT press.

Martin C, Zhang Y (2005) The diverse functions of histone lysine methylation. *Nature reviews Molecular cell biology* 6:838.

Nakayama Ji, Rice JC, Strahl BD, Allis CD, Grewal SI (2001) Role of histone h3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* 292:110–113.

Pedersen A, Wiuf C, Christiansen FB (1998) A codon-based model designed to describe lentiviral evolution. *Molecular biology and evolution* 15:1069–1081.

Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NT, Schreiber SL, Mellor J, Kouzarides T (2002) Active genes are tri-methylated at k4 of histone h3. *Nature* 419:407.