A model of epigenome evolution

Jianghan Qu Andrew D. Smith

April 7, 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 factorizes 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 associated with individual positions may change. The types of changes we are most interested in can be more naturally interpreted in terms 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 (e.g. promoters), and the genes themselves have directionality.

We are also interested in modeling sequences that evolve in continuous time, and we use two representations to describe a particular realization of the evolutionary process. These two representations have distinct advantages, both in explaining aspects of our model and in different computations.

Let s be a sequence of N binary states that evolve according to a continuous time Markov process. The state space for s has size 2^N, but we constrain instantaneous transition rates such that each change to s only involves a single site (coordinate) within s. We can summarize the process as a *global path* over time interval $[0, \tau)$ as follows:

$$\begin{aligned} H = (s, Y, V), \text{ where } \quad & s \in \{0,1\}^N, \\ & Y = (y_1, \ldots, y_w), \quad \text{with } 0 \le y_i < y_j < \tau, \quad \text{for } 1 \le i < j \le w, \\ & V = (v_1, \ldots, v_w), \quad \text{with } 1 \le v_i \le N, \qquad \text{for } 1 \le i \le w. \end{aligned} \tag{2}$$

Times Y and positions V are in direct correspondence, but we adopt the notational convention $y_0 = 0$, allowing us to refer to any time $y_i - 1$, without any position corresponding to time 0. For any time $0 \le t < \tau$, we let $s(t)$ denote the state sequence after having applied changes to $s = s(0)$ at positions given by $v_1, v_2, \ldots, v_k$ such that y_k is maximal satisfying $y_k \le t$. From H we can obtain $s(t)$ for any $0 \le t < \tau$ by tracing changes made to s.

We can reorganize the same information to describe the process in terms of *site-specific paths*. In particular, for site $1 \le n \le N$,

$$\begin{aligned} H_n = (s_n, Y_n), \text{ where } \quad & s_n \in \{0,1\}, \\ & Y_n = (y_1, \ldots, y_{w_n}), \quad \text{with } 0 \le y_i < y_j < \tau, \quad \text{for } 1 \le i < j \le w_n. \end{aligned} \tag{3}$$

Our assumption that instantaneous rates are non-zero only for transitions that modify a single position in s defines bijection between the global jump times Y and the union of the site-specific jump times $\cup_n Y_n$. Note that the sequences of site-specific paths $(H_1, \ldots, H_N)$ more concisely describes a realization of our process; obtaining the state sequence $s(t)$ for any time t requires more effort.

The evolution of states at a given site in the state sequence s depends on the states of its neighboring sites at all times during evolution. Consider site n and its two immediate neighbors $n-1$ and $n+1$, and let the contemporaneous states for these sites at time t be $s_{n-1}(t) = i$, $s_n(t) = j$ and $s_{n+1} = j$. The instantaneous rate for a transition at site n from state j to the complementary binary state $\bar{j}$ is $\lambda(i, j, k)$. In other words, the

instantaneous rate $\lambda(i, j, k)$ is a function of the states at the site of interest and its two immediate neighbors. Since we assume the evolutionary process is symmetric with respect to the direction along the genome, we require $\lambda(i, j, k) = \lambda(k, j, i)$. When the states at all positions except for site n remain constant over time interval the state s_n follows a two-state continuous-time Markov process within that 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 state sequence, or evaluate its probability, using the equivalent Markov chain. We can relate the factors in (1) and the transition probabilities of the Markov chain as follows. Define pair-wise potentials $Q(i, j) = \exp(\phi(i, j))$, where $i, j \in \{0, 1\}$ are binary states. The largest eigenvalue of Q is

$$q = \tfrac{1}{2}\left(Q(0,0) + Q(1,1) + \sqrt{\Delta}\right), \text{ where } \Delta = (Q(0,0) - Q(1,1))^2 + 4Q(0,1)Q(1,0).$$

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

$$\frac{h_0}{h_1} = \frac{Q(0,1)}{q - Q(0,0)} = \frac{q - Q(1,1)}{Q(1,0)}.$$

The Markov chain transition matrix is:

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

More specifically,

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

Accordingly, the expected fraction of an epigenome state sequence residing within functional domains (the "features") is:

$$1 - \frac{2Q(0,1)Q(1,0)}{(Q(0,0) - Q(1,1))^2 + 4Q(0,1)Q(1,0) + (Q(1,1) - Q(0,0))\sqrt{\Delta}}.$$

2.3 Relating the substitution model and the stationary distribution

Jensen & Pedersen (2000) gave a sufficient condition for a continuous time evolutionary model with the properties we outlined for λ to have a stationary distribution of the form ϕ. In particular, if

$$\frac{\lambda_{ijk}}{\lambda_{i\bar{j}k}} = \frac{\lambda(i,j,k)}{\lambda(i,\bar{j},k)} = \frac{\exp(\phi(i,\bar{j}) + \phi(\bar{j},k))}{\exp(\phi(i,j) + \phi(j,k))}, \tag{5}$$

then (1) is the stationary distribution for the Markov process with instantaneous rates λ_{ijk}.

Jensen & Pedersen (2000) also provide (Proposition 2) a way of specifying λ from ϕ: substitution rates λ satisfy the relation (6) if and only if they can be written in the form

$$\log(\lambda_{ijk}) = -\psi(i,j,k) + \ell(i,k). \tag{6}$$

This criteria was introduced in the context of an arbitrary number of states for each position in the state sequence, 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 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

$$\log(\lambda_{ijk}) = \ell(i, k) + (\phi(i, \bar{j}) + \phi(\bar{j}, k)), \tag{7}$$

then they satisfy the condition of (6).

2.4 Parameterization and interpretation

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

$$\Lambda = \begin{array}{c} \\ 000 \\ 010 \\ 001 \\ 011 \\ 100 \\ 110 \\ 101 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 010 & 001 & 011 & 100 & 110 & 101 & 111 \end{array} \\ \begin{pmatrix} \cdot & \mathcal{B} & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathcal{D} & \cdot & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & \mathcal{E} & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathcal{C} & \cdot & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdot & \mathcal{E} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathcal{C} & \cdot & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdot & \mathcal{M} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathcal{S} & \cdot \end{pmatrix} \end{array} \tag{8}$$

We may interpret the non-zero entries in Λ 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 Λ, 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 Λ with stationary distribution (1), then the condition in (6) holds, and we have the following constraints for the rates in the above matrix:

$$\mathcal{B}\mathcal{C}^2\mathcal{M} = \mathcal{D}\mathcal{E}^2\mathcal{S}. \tag{9}$$

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

$$\phi(0,0) = \phi(0,1) + \frac{1}{2}\log\left(\frac{\mathcal{D}}{\mathcal{B}}\right), \text{ and } \phi(1,1) = \phi(0,1) + \frac{1}{2}\log\left(\frac{\mathcal{D}\mathcal{E}^2}{\mathcal{B}\mathcal{C}^2}\right). \tag{10}$$

In summary, the transition rate matrix Λ 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 (10).

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

Algorithm 1 Simulating epigenome evolution

Input: Binary state sequence s of length N, time τ and rates Λ.

Output: Simulated global path $H = (s, Y, V)$ for change times Y and positions V.

$t \leftarrow 0$, $V \leftarrow \emptyset$, $Y \leftarrow \emptyset$ and $x \leftarrow s$
while $t < \tau$ **do**
  Sample holding time $y \sim Exp(-R_{ss})$, where $-R_{ss} = \sum_{i,j,k} c_{ijk}(x)\lambda_{ijk}$
  $t \leftarrow t + y$
  if $t < \tau$ **then**
    Sample binary triplet ijk with probability proportional to $c_{ijk}(x)\lambda_{ijk}$
    Sample position n uniformly from the $c_{ijk}(x)$ at center of a ijk triplet in x
    Modify state sequence x to have state $\bar{j}$ at position n
    Append n to V and append t to Y
return $H = (s, Y, V)$

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 two epigenomes x and x' that differ by exactly one position, and further suppose the state at that position in x is j, and $\bar{j}$ in epigenome x'. Neighboring positions in both x and x' have states i and k. The instantaneous rate of a jump between x and x' is λ_{ijk}. These entries can be organized as a transition matrix R with $2^N \times 2^N$ entries, but all positive values corresponding to some λ_{ijk}. The holding time in a state sequence x is exponentially distributed with rate $-R_{xx}$ equal to the sum of instantaneous rates for jumps from x to any state sequence x' that differs from x at exactly one position. Expressed in terms of the triplet rates, for a state sequence x this parameter is

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

where $c_{ijk}(x) = \sum_{n=1}^{N-2} I(x_n = i, x_{n+1} = j, x_{n+2} = k)$ are the *"triplet counts"* counting patterns ijk in x.

Given that a jump has occurred at some fixed time, the probability that the jump changed x at the middle position of triplet ijk is proportional to $c_{ijk}(x)\lambda_{ijk}$. Further, given that a jump occurred with context ijk, our model assumes the jump is equally likely to have changed any position in x having state j with left and right neighbors having states i and k. If the process is stationary, 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, detailed in Algorithm 1.

4 Parameter estimation

4.1 Complete data

First, given a global path H for the evolution of an epigenome over time $[0, \tau)$ with $\tau = 1$, we are interested in estimating the transition rate parameters Λ. We assume a stationary process here, but this assumption is not necessary (see Appendix).

Recall that for global path $H = (s, Y, V)$, the m-th jump occurs at time y_m and position v_m. Let $\mathrm{con}(m)$

denote the *context* of the m-th jump: the triplet ijk of binary states such that

$$\mathrm{con}(m) = ijk \;\; \Rightarrow \;\; s_{v_m-1}(y_{m-1}) = i,\; s_{v_m}(y_{m-1}) = j, \;\text{ and } s_{v_m+1}(y_{m-1}) = k.$$

So $\mathrm{con}(m)$ indicates the triplet that was modified by the m-th jump. Note: $\mathrm{con}(m)$ is defined relative to a specific global path H that we leave implicit in our notation.

The holding time $\Delta_m = v_m - v_{m-1}$ just prior to the m-th jump in the evolving state sequence follows the exponential distribution

$$\Delta_m \sim \mathit{Exp}(\lambda_m), \;\text{ with }\; \lambda_m = -R_{s(v_{m-1})s(v_{m-1})} = \sum_{i,j,k} c_{ijk}(v_{m-1})\lambda_{ijk}.$$

The likelihood function for parameters Λ involves holding times and identities of transitions:

$$L(\Lambda|H) = \prod_{m=1}^{w} \lambda_m \exp(-\lambda_m \Delta_m) \times \left(\tfrac{\lambda_{\mathrm{con}(m)}}{\lambda_m}\right) = \prod_{m=1}^{w} \lambda_{\mathrm{con}(m)} \exp(-\lambda_m \Delta_m). \tag{11}$$

And the log-likelihood function is

$$\begin{aligned} \log L(\Lambda|H) \;\; &= \textstyle\sum_{i,j,k} \left(\sum_{m=1}^{w} \log \lambda_{ijk} \times I\{\mathrm{con}(m) = ijk\} - c_{ijk}(v_{m-1})\lambda_{ijk}\Delta_m \right) \\ &= \textstyle\sum_{i,j,k} \left(J_{ijk} \log \lambda_{ijk} - D_{ijk}\lambda_{ijk} \right), \end{aligned} \tag{12}$$

where

$$J_{ijk} = \sum_{m=1}^{w} I\{\mathrm{con}(m) = ijk\} \;\; \text{ and } \;\; D_{ijk} = \sum_{m=1}^{w} c_{ijk}(v_{m-1})\Delta_m. \tag{13}$$

Here J_{ijk} counts jumps that create a triplet ijk and D_{ijk} measures the evolutionary time spent in context ijk, summed over all sites in the state sequence (see Appendix for gradients used in max-likelihood estimation).

So far we have considered evolution along a single trajectory. In general we are interested in evolution over a tree, and in this setting the relative lengths of branches become important. Let $\mathcal{L} = \{\tau_1, \ldots, \tau_B\}$ denote the lengths of branches in a tree with an assumed topology, with the additional constraint that $\sum_{b=1}^{B} \tau_b = B$. With this view, the branch lengths scale the transition rates. Let H_b denote the global path of jumps along branch b. Then the likelihood is a product over likelihoods for the individual branches:

$$L(\Lambda, \mathcal{L}| \cup_b H(b)) = \prod_{b=1}^{B} L(\Lambda, \tau_b|H(b)) = \prod_{b=1}^{B} L(\tau_b \times \Lambda|H(b)), \tag{14}$$

and

$$\log L(\Lambda, \mathcal{L}| \cup_b H(b)) = \sum_{b=1}^{B} \log L(\Lambda, \mathcal{L}| \cup_b H(b)) = \sum_{b=1}^{B} \log L(\tau_b \times \Lambda|H(b)). \tag{15}$$

Compared with equations (12), the terms on the left above must be adjusted:

$$\log L(\tau_b \times \Lambda|H(b)) = \sum_{i,j,k} J(b)_{ijk} \log(\tau_b \lambda_{ijk}) - D(b)_{ijk} \times (\tau_b \lambda_{ijk}), \tag{16}$$

with $J(b)$ and $D(b)$ defined by equation (13) but specifically in terms of global path $H(b)$ for branch b.

4.2 Inference from incomplete data

In applications we have data, in the form of epigenomic state sequences, associated with leaf nodes. Even at leaf nodes our data may be incomplete. We do not have data corresponding to internal nodes, let alone the full paths of jumps and corresponding positions in continuous time. In addition to estimating model parameters from such incomplete data, we can learn much from estimating state sequences at internal tree nodes.

The method we use follows that outlined by Jensen & Pedersen (2000) and Hobolth (2008). The basic machinery we borrow is most easily explained by first focusing on a single trajectory (single branch tree) with known site-specific paths for all sites except n. We can sample a path H_n according to the distribution

$$\Pr(H_n|H_1,\ldots,H_{n-1},H_{n+1},\ldots,H_N) = \Pr(H_n|H_{\overline{n}})$$

by using MCMC and the likelihood from equation (**??**). When H_n is known partially, for example if the starting state $s_n(0)$ and the ending state $s_n(1)$ are known, MCMC can be applied similarly to sample a path consistent with those end-points. If the data are state sequences at leaf nodes of a tree having known topology, the approach of Hobolth (2008) can be used to sample site-specific paths for all sites and all branches in the tree. Summarizing these samples provides estimates of expected states at all sites for all ancestral nodes. Importantly, such estimates made by sampling site-specific paths are consistent with our distributional assumptions for the global state space. Although this general strategy allows much flexibility, efficient sampling is still a major challenge.

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.

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?

Tree branch partitions imposed by states at neighboring sites. The process an individual site is not homogeneous along a branch, as its transition rates depend on states of neighboring sites at any instant. By partitioning each branch into maximal contiguous intervals with neighboring states unchanged, within each such interval the process is homogeneous. For a site n, define breaks $Z = (z_1 \ldots, z_{w_{n-1}+w_{n+1}})$ as the ordered set of times from Y_{n-1} and Y_{n+1}. During each time interval (z_{i-1}, z_i), the state at site $n-1$ and site $n+1$ remains unchanged. For each branch b in the tree, we can use the set of neighbor-induced breaks $Z(b)$ to define new nodes that each have only one child. Once this is done, we can apply Felsenstein's algorithm to compute the probability of states at each breakpoint on each branch of the tree.

and can be characterized using 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

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

Bottom-up pruning algorithm. We can use Felsenstein's pruning algorithm to compute this probability. Let the phylogenetic tree be G_n consisting of all leaf nodes and internal nodes from G, and all breakpoint nodes at times $Z_n(b)$ on each branch b of G. For any node $u \in G_n$ let $X_n(u)$ be the states of all leaf nodes below u at site n (Figure **??**). The state transition probabilities from the parent of node u to node u are

$$P_u = \exp(Q_{\mathrm{con}(u)} \tau_u).$$

We use the following notation to describe the algorithm:

$$\begin{aligned} p_j(v) &= \Pr(X_n(v)|s_n(u) = j), \quad \text{where } u \text{ is the parent of } v, \\ q_j(u) &= \begin{cases} \Pr(s_n(u) = j), & \text{if } u \text{ is a leaf node,} \\ \prod_{v \in \mathrm{child}(u)} p_j(v), & \text{otherwise.} \end{cases} \end{aligned} \tag{17}$$

If the state $s_n(u)$ is observed (e.g. at a leaf) then $\Pr(s_n(u) = j) \in \{0, 1\}$. If the data are contiuous or given as expected states or posterior probabilities, then $\Pr(s_n(u) = j) \in (0, 1)$. In the bottom-up pass of Felsenstein's pruning algorithm we compute

$$p_j(u) = \textstyle\sum_k P_u(j, k) \times q_k(u). \tag{18}$$

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:

$$\Pr(X_n(r)|H_{\overline{n}}) = \sum_j \Pr(s_n(r) = j|H_{\overline{n}}) q_j(r). \tag{19}$$

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

$$\begin{aligned} \Pr(s_n(r)|H_{\overline{n}}, X_n(r)) =& \frac{\Pr(s_n(r), X_n(r)|H_{\overline{n}})}{\Pr(X_n(r)|H_{\overline{n}})} = \frac{\Pr(X_n(r)|s_n(r), H_{\overline{n}}) \Pr(s_n(r)|H_{\overline{n}})}{\Pr(X_n(r)|H_{\overline{n}})} \\ =& \frac{\Pr(s_n(r)|H_{\overline{n}}) \prod_{v \in \mathrm{child}(r)} p_{s_n(r)}(v)}{\Pr(X_n(r)|H_{\overline{n}})}, \end{aligned} \tag{20}$$

where $\Pr(s_n(r)|H_{\overline{n}})$ 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 $\Pr(X_n(r)|H_{\overline{n}})$ and $\Pr(s_n(r)|H_{\overline{n}}, X_n(r))$.

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

$$\Pr(S|X_n(r), H_{\overline{n}}) = \frac{\Pr(S, X_n(r)|H_{\overline{n}})}{\Pr(X_n(r)|H_{\overline{n}})},$$

where $\Pr(S, X_n(r)|H_{\overline{n}})$ is the forward sampling probability under the given transition probability matrices in individual intervals determined by $H_{\overline{n}}$. 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.

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

$$p_i = \frac{\exp(Q_{ii}\tau)}{P(\tau)_{ii}}. \tag{21}$$

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

Let W be the waiting time until the first jump.

$$\begin{aligned}
&\Pr(W < \tau | s(0) = i, s(\tau) = j) \\
&= \Pr(W < \tau, s(W) = \bar{i} | s(0) = i, s(\tau) = j) \\
&= \frac{\Pr(W < \tau, s(W) = \bar{i}, s(\tau) = j | s(0) = i)}{\Pr(s(\tau) = j | s(0) = i)} \\
&= \int_0^\tau Q_{i\bar{i}} \exp(Q_{ii} w) \frac{P(\tau - w)_{\bar{i}j}}{P(\tau)_{ij}} dw \\
&= \int_0^\tau f_{ij\tau}(w) dw, \text{ where } f_{ij\tau}(t) = Q_{i\bar{i}} \exp(Q_{ii} t) \frac{P(\tau - t)_{\bar{i}j}}{P(\tau)_{ij}}.
\end{aligned} \tag{22}$$

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 (2008) A markov chain monte carlo expectation maximization algorithm for statistical analysis of dna sequence evolution with neighbor-dependent substitution rates. *Journal of Computational and Graphical Statistics* 17:138–162.

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.

A Equations for maximum likelihood estimation

If we do not assume a stationary process, we may also be 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. Define *pair counts* $c_{ij} = \sum_{n=1}^{N-1} I\{s_n(0) = i, s_{n+1}(0) = j\}$ with analogy to the triplet counts defined above. Then

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

The constraints on the transition rates λ_{ijk} as indicated in the matrix (8) and equation (9) are:

$$\begin{aligned} \lambda_{001} &= \lambda_{100} \\ \lambda_{011} &= \lambda_{110} \\ \lambda_{000}\lambda_{110}^2\lambda_{101} &= \lambda_{010}\lambda_{100}^2\lambda_{111} \end{aligned} \tag{23}$$

So for the purpose of estimating parameters, we only require 5 gradients.

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

$$\begin{aligned} \frac{\partial l}{\partial \log \lambda_0} &= J_0 - D_0\lambda_0 + J_7 - D_7\lambda_7 \\ \frac{\partial l}{\partial \log \lambda_2} &= J_2 - D_2\lambda_2 - J_7 + D_7\lambda_7 \\ \frac{\partial l}{\partial \log \lambda_5} &= J_5 - D_5\lambda_5 + J_7 - D_7\lambda_7 \\ \frac{\partial l}{\partial \log \lambda_1} &= J_1 + J_4 - (D_1 + D_4)\lambda_1 - 2J_7 + 2D_7\lambda_7 \\ \frac{\partial l}{\partial \log \lambda_3} &= J_3 + J_6 - (D_3 + D_6)\lambda_3 + 2J_7 - 2D_7\lambda_7 \end{aligned} \tag{24}$$

B Algorithms used for sampling at individual sites

Algorithm 2 Pruning$(u, X_n(u))$

Input: Phylogenetic tree u with associated transition probabilities

Output: Values $p_j(u)$ and $q_j(u)$ for $j \in \{0, 1\}$ as defined in equation (17)

$q_j(u) \leftarrow 1$ iff state $s_n(u) = j$ is compatible with $X_n(u)$, for $j \in \{0, 1\}$
for child v of u **do**
  $\{p_j(v), q_j(v)\} \leftarrow$ Pruning$(v, X_n(v))$
  $q_j(u) \leftarrow q_j(u) \times p_j(v)$ for $j \in \{0, 1\}$
$p_j(u) \leftarrow \sum_{k\in\{0,1\}} P_u(j, k) \times q_k(u)$ for $j \in \{0, 1\}$
return $p_j(u)$ and $q_j(u)$ for $j \in \{0, 1\}$

Algorithm 3 DownwardStateSampling$(u, s(u), X(u))$

Input: Phylogenetic tree node u, binary state $s(u)$ and states $X(u)$ for all leaves below u

Output: Sampled states for all internal nodes v in the tree rooted at u

for each child node v of u **do**
  Sample $s(v) \sim P_v(j, k) \times q_k(v)/p_j(v)$
  Update the direct sampling probability with factor $P_v(j, k)q_k(v)/p_j(v)$
  DownwardStateSampling$(v, s(v), X(v))$
The result is an sample S of states at all internal states, which is sampled from their joint marginal posterior probability distribution $\Pr(S|X_n(r), H_{\overline{n}})$.

Algorithm 4 DirectEndConditionedPathSampling()

Input: a

Output: b

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 (21).
    if $Z = 1$ **then**
      Set $s(t) = i$ for all $0 < t < \tau$, and update $\tau \leftarrow 0$.
  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 $i \leftarrow \bar{i}$, $\tau \leftarrow \tau - W$.