Approximation of Indel Evolution by Differential Calculus of Finite State Automata

Holmes Lab, Department of Bioengineering, University of California, Berkeley

June 29, 2020

Abstract

A method of solving insertion-deletion models on sequences is introduced. The method uses automata theory to describe the action of an infinitesimal evolutionary generator on a probability distribution over alignments, both of which can be described using Pair Hidden Markov Models (Pair HMMs). In general, combining HMMs in this way induces a multiplication of their state spaces; however, we introduce a coarse-graining operation to keep the state space at a constant size. This naturally leads to ordinary differential equations for the evolution of the transition probabilities of the approximating Pair HMM, which can be solved numerically. By simulation, we show that the resulting approximate distribution over alignments is a closer fit than previous approximations, and is derived directly from a rigorous infinitesimal limit. Our code and data are available at `https://github.com/ihh/trajectory-likelihood`. We also conjecture an approach to develop similar equations for the sufficient statistics necessary to fit the underlying rate model by Expectation-Maximization.

1 Introduction

In molecular evolution, the equations of motion describe continuous-time Markov processes on discrete sequences. For substitution processes, these equations are reasonably well-understood, but insertion and deletions (indels) have proved elusive.

In models where a sequence is subject only to point substitutions at independently evolving sites, the likelihood can be factorized into a product of small Markov chains [19], solved exactly for an ancestor-descendant pair by considering the eigenstructure of the matrix exponential [20, 10], and extended to multiple aligned sequences by applying the sum-product algorithm to the phylogenetic tree [9]. These results are widely used in bioinformatics. They can be generalized to the case where substitution rates are influenced by neighboring residues—or where substitution events simultaneously affect multiple residues—by expanding the matrix exponential as a Taylor series in neighboring contexts [22], extending to multiple alignments using variational (mean-field) approaches [18, 32].

By comparison, continuous-time Markov chain models of the indel process are tricky. The first—and only—exactly-solved example is the Thorne-Kishino-Felstenstein (TKF) model, which allows only single-residue indels. The TKF model reduces exactly to a linear birth-death process with immigration [28], which allows the joint distribution over ancestor-descendant sequence alignments to be expressed as a Hidden Markov Model (HMM) [15] that can be formally extended to multiple sequences using algebraic composition of automata [12, 30, 1].

Attempts to generalize the TKF model to the more biologically-plausible case of multiple-residue indel events fall into two categories: those that attempt to analyze the process from first principles to arrive at finite-time transition probabilities [24, 21, 23, 8, 7, 6, 3], and those that guess closed-form approximations to these probabilities without such *ab initio* justifications [29, 25, 26]. In this paper we focus on the former type of approach. The latter category of approaches often proceed by breaking the sequence into indivisible multiple-residue fragments—or introducing other latent variables—but lacking any analytic connection of the fragment sizes or other newly-introduced parameters to the infinitesimal mutation rates of the underlying process, their evaluation in a statistical framework must necessarily be somewhat heuristic [3].

Formal mathematical treatment of the multi-residue indel process begins with Miklós and Toroczkai's analysis of a model that allows multi-residue insertions but single-residue deletions [24]. They developed a generating function for the gap length distribution, and used the method of characteristics to solve the associated partial differential equations. Arguably the most important feature of this model is that the alignment likelihood remains factorizable and associated with an HMM (albeit one with infinite states). This remains true for indel processes that allow both insertions and deletions to span multiple residues, under certain assumptions of spatial homogeneity [21, 23, 7], a theoretical result that helps to justify HMM-based approximations. However, calculating the transition probabilities of these HMMs from first principles is still nontrivial. Miklós *et al* [23], formalizing intuition of Knudsen and Miyamoto, [21], obtained reasonable approximations for short evolutionary time intervals by calculating exact likelihoods of short trajectories in the continuous-time Markov process. However, exhaustively enumerating these trajectories is extremely slow, and effectively impossible for trajectories with more than three overlapping indel events, so this approach is of limited use.

A recent breakthrough in this area was made by De Maio [3]. Starting from the approximation that the alignment likelihood can be factored into separate geometric distributions for insertion and deletion lengths, he derived ordinary differential equations (ODEs) for the evolution of the mean lengths of these distributions, yielding transition probabilities for the Pair HMM. De Maio's method produces more accurate approximations to the multi-residue indel process than all previous attempts, though it has limitations: it's restricted to models where the insertion and deletion rates are equal, does not (by design) include covariation between insertion and deletion lengths in the alignment, is inexact for the special case of the TKF model, and requires laborious manual derivation of the underlying ODEs.

In this paper, we build on De Maio's results to develop a systematic differential calculus for finding HMM-based approxmate solutions of continuous-time Markov processes on strings which are "local" in the sense that the infinitesimal generator is an HMM. Our approach addresses the limitations of De Maio's approach, identified in the previous paragraph. It does not require that insertion and deletion rates are equal, or that the process is time-reversible: any geometric distribution over indel lengths is allowed. It does account for covariation between insertion and deletion gap sizes. The TKF model emerges as a special case and the closed-form solutions to the TKF model are exact solutions to our model. Finally, although our equations can be derived without computational assistance, the analysis is greatly simplified by the use of symbolic algebra packages: both for the manipulation of equations, for which we used Mathematica [17], and for the manipulation of state machines, for which we used our recently published software Machine Boss [27].

The central idea of our approach is that the application of the infinitesimal generator to the approximating HMM generates a more complicated HMM that, by a suitable coarse-graining operation, can be mapped back to the simpler structure of the approximating HMM. By matching the expected transition usages of these HMMs, we derive ODEs for the transition probabilities of the approximator. Our approach is justified by improved results in simulations, yielding greater accuracy and generality than all previous approaches to this problem, including De Maio's moment-based method (which can be seen as a version of our method that considers only indel-extending transitions in a symmetric model). Our approach is further justified by the emergence of the TKF model as an exact special case, without the need to introduce any additional latent variables such as fragment boundaries, or arbitrary time-dependence to the TKF formulae.

While we focus here on the multi-residue indel process, the generality of the infinitesimal automata suggests that other local evolutionary models, such as those allowing neighbor-dependent substitution and indel rates, may also be productively analyzed using this approach.

2 Methods

2.1 Expected transition usage

Suppose that we have a hidden Markov model, $\mathbb{M}$, with K states which can be partitioned into match σ_M, insert σ_I, delete σ_D, and null σ_N states, as in [5]. As a shorthand we will write σ_{ID} for $\sigma_I \cup \sigma_D$, and so on. Thus σ_{MIDN} is the complete set of K states.

We'll be considering models with only one match state, which by convention will always be the first state, so $\sigma_M = \{1\}$.

Let $\phi = (\phi_0, \phi_1, \ldots \phi_k)$ denote a state path.

The transition probability matrix is $\mathbf{Q}$ with elements $Q_{ij} = P(\phi_{k+1} = j | \phi_k = i)$.

Let $\Phi(X \to Y \to Z)$ denote the set of state paths with the following properties:

- The path begins in a σ_X-state;

- The path ends in a σ_Z-state;
- For paths with more than two states, the intermediate states are all σ_Y-states.

Let $\mathbf{J}^{(X\to Y)}$ be a matrix that selects transitions from σ_X to σ_Y

$$J_{ij}^{(X\to Y)} = \begin{cases} 1 & i \in \sigma_X, j \in \sigma_Y \\ 0 & \text{otherwise} \end{cases}$$

and let $\mathbf{Q}^{(X\to Y)} \equiv \mathbf{J}^{(X\to Y)} \circ \mathbf{Q}$ where $\circ$ is the Hadamard product.

Consider a random path $\phi \in \Phi(M \to IDN \to M)$ that begins and ends in the match state, passing only through non-match states in between. Let $T_{XY}(\phi) = |\{(i,j) : j > i, (\phi_i \ldots \phi_j) \in \Phi(X \to N \to Y)\}|$ count the number of subpaths of ϕ that go from a state in σ_X, via zero or more null states, to a state in σ_Y. Thus if null states were removed from the path, then T_{XY} would be the number of transitions from X-states to Y-states.

The expectation of T_{XY} is

$$\begin{aligned}
&E_{\phi|\mathbb{M}}\left[T_{XY}(\phi)\right] \\
&= \sum_{\phi\in\Phi(M\to IDN\to M)} P(\phi)T_{XY}(\phi) \\
&= \left(\sum_{i=0}^{\infty}\left(\mathbf{Q}^{(MIDN\to IDN)}\right)^i \left(\mathbf{J}^{(X\to Y)} \circ \sum_{j=0}^{\infty}\left(\mathbf{Q}^{(MIDN\to N)}\right)^j \mathbf{Q}\right) \sum_{k=0}^{\infty}\left(\mathbf{Q}^{(IDN\to MIDN)}\right)^k\right)_{11} \\
&= \left(\mathbf{U}\left(\mathbf{J}^{(X\to Y)} \circ (\mathbf{VQ})\right)\mathbf{W}\right)_{11} \qquad (1)
\end{aligned}$$

with

$$\begin{aligned}
\mathbf{U} &= \left(\mathbf{1} - \mathbf{Q}^{(MIDN\to IDN)}\right)^{-1} \\
\mathbf{V} &= \left(\mathbf{1} - \mathbf{Q}^{(MIDN\to N)}\right)^{-1} \\
\mathbf{W} &= \left(\mathbf{1} - \mathbf{Q}^{(IDN\to MIDN)}\right)^{-1}
\end{aligned}$$

where $\mathbf{1}$ is the $K \times K$ identity matrix.

Let $S_X(\phi)$ be the number of X-states in ϕ, excluding the final state. Thus

$$\begin{aligned}
S_M(\phi) &= 1 \\
S_X(\phi) &= \sum_{Y\in M,I,D,N} T_{XY}(\phi) \\
&= \sum_{Y\in M,I,D,N} T_{YX}(\phi)
\end{aligned}$$

2.2 Three-state HMM

Consider the machine $\mathbb{F}(t)$ shown in Figure 1a with $\sigma_M = \{1\}$, $\sigma_I = \{2\}$, $\sigma_D = \{3\}$, and

$$\mathbf{Q}\left[\mathbb{F}(t)\right] = \begin{pmatrix} a(t) & b(t) & c(t) \\ f(t) & g(t) & h(t) \\ p(t) & q(t) & r(t) \end{pmatrix}$$

with $a+b+c=1$, $f+g+h=1$ and $p+q+r=1$.

Here t will play the role of a time parameter.

Let

$$\begin{aligned}
\bar{T}_{XY}(t) &= E_{\phi|\mathbb{F}(t)}\left[T_{XY}(\phi)\right] \\
\bar{S}_X(t) &= E_{\phi|\mathbb{F}(t)}\left[S_X(\phi)\right] \\
&= \sum_{Y\in M,I,D,N} \bar{T}_{XY}(t) \\
&= \sum_{Y\in M,I,D,N} \bar{T}_{YX}(t) \\
\bar{S}_M(t) &= 1 \qquad (2)
\end{aligned}$$

where the expectations are as defined in (1). (Throughout this paper, such expectations are over $\phi \in \Phi(M \to IDN \to M)$.) Evidently,

$$\begin{array}{lllllllll}
a(t) & = & \bar{T}_{MM}(t), & b(t) & = & \bar{T}_{MI}(t), & c(t) & = & \bar{T}_{MD}(t), \\
f(t) & = & \bar{T}_{IM}(t)/\bar{S}_I(t), & g(t) & = & \bar{T}_{II}(t)/\bar{S}_I(t), & h(t) & = & \bar{T}_{ID}(t)/\bar{S}_I(t), \\
p(t) & = & \bar{T}_{DM}(t)/\bar{S}_D(t), & q(t) & = & \bar{T}_{DI}(t)/\bar{S}_D(t), & r(t) & = & \bar{T}_{DD}(t)/\bar{S}_D(t)
\end{array} \qquad (3)$$

Further,

$$\begin{aligned}
\bar{S}_M(t) &= 1 \\
\bar{S}_I(t) &= \frac{(b(1-r)+cq)(f(1-r)+hp)}{((1-g)(1-r)-hq)^2} \\
\bar{S}_D(t) &= \frac{(c(1-g)+bh)(p(1-g)+fq)}{((1-g)(1-r)-hq)^2} \qquad (4)
\end{aligned}$$

The essence of our approach is to use transducer composition to study infinitesimal increments in $\bar{T}_{XY}(t)$ and thereby obtain differential equations that can be solved to find these parameters.

2.3 Rate of change of expected transition counts

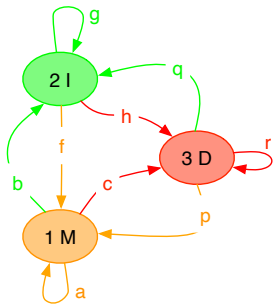
The infinitesimal transducer $\mathbb{G}(\Delta t)$ of Figure 1b has states $\sigma_M = \{1\}$, $\sigma_I = \{2\}$, $\sigma_D = \{3\}$, and transition matrix

$$\mathbf{Q}\left[\mathbb{G}(\Delta t)\right] = \begin{pmatrix} 1-(\lambda+\mu)\Delta t & \lambda\Delta t & \mu\Delta t \\ 1-x & x & 0 \\ 1-y & 0 & y \end{pmatrix}$$

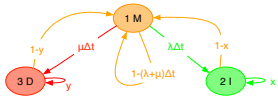
The parameters are the insertion and deletion rates (λ, μ) and extension probabilities (x, y), and the time interval $\Delta t \ll 1$.

Composing $\mathbb{F}(t)$ (Figure 1a) with $\mathbb{G}(\Delta t)$ (Figure 1b) yields $\mathbb{F}(t)\mathbb{G}(\Delta t)$, the machine of Figure 1c. This machine[1] has states $\sigma_M = \{1\}$, $\sigma_I = \{2,3,4\}$, $\sigma_D = \{5,6,7,8\}$, $\sigma_D = \{9\}$, and transition matrix

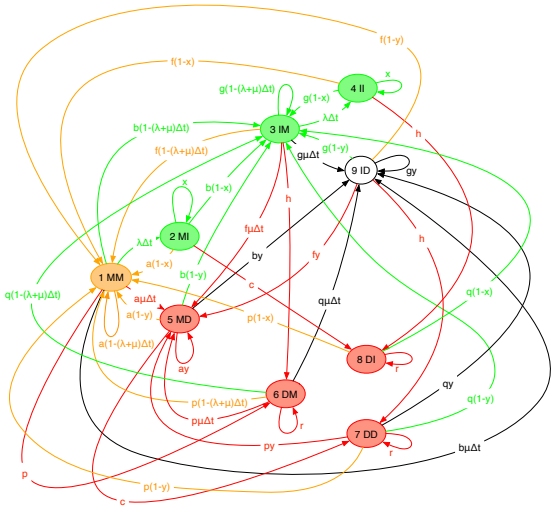
[1]The transducer composition $\mathbb{F}(t) \times \mathbb{G}(\Delta t)$ was performed using the automata algebra program Machine Boss [27]. A general procedure for doing this for any two machines $\mathbb{A}, \mathbb{B}$ involves taking the Cartesian product of the two machines' state spaces and then synchronizing their transitions. This ensures that, if $\mathbb{M}_{XY}$ represents the result of the Forward algorithm for machine $\mathbb{M}$ (with null states eliminated) and sequences X, Y, then $(\mathbb{A}\mathbb{B})_{XZ} = \sum_Y \mathbb{A}_{XY}\mathbb{B}_{YZ}$. More details on these operations can be found in [27, 30, 31].



(a) Machine $\mathbb{F}(t)$.



(b) Machine $\mathbb{G}(\Delta t)$.



(c) Machine $\mathbb{F}(t)\mathbb{G}(\Delta t)$.

Figure 1: Three state machines for modeling indels in alignments. Match states are orange; insert states, green; delete states, red; and null states are uncolored. Transitions are colored by destination state. (1a) Machine $\mathbb{F}(t)$ models alignments at divergence time t. (1b) Machine $\mathbb{G}(\Delta t)$ models the infinitesimal evolution over time Δt. (1c) Machine $\mathbb{F}(t)\mathbb{G}(\Delta t)$, the transducer product of the previous two machines, models alignments at divergence time $t + \Delta t$. Our approach is to approximate this with a machine of the same form as (1a).

$$\mathbf{Q}\left[\mathbb{F}(t)\mathbb{G}(\Delta t)\right] =$$

$$\begin{pmatrix} a(1-(\lambda+\mu)\Delta t) & \lambda\Delta t & b(1-(\lambda+\mu)\Delta t) & 0 & a\mu\Delta t & c & 0 & 0 & b\mu\Delta t \\ a(1-x) & x & b(1-x) & 0 & 0 & 0 & 0 & c & 0 \\ f(1-(\lambda+\mu)\Delta t) & 0 & g(1-(\lambda+\mu)\Delta t) & \lambda\Delta t & f\mu\Delta t & h & 0 & 0 & g\mu\Delta t \\ f(1-x) & 0 & g(1-x) & x & 0 & 0 & 0 & h & 0 \\ a(1-y) & 0 & b(1-y) & 0 & ay & 0 & c & 0 & by \\ p(1-(\lambda+\mu)\Delta t) & 0 & q(1-(\lambda+\mu)\Delta t) & 0 & p\mu\Delta t & r & 0 & 0 & q\mu\Delta t \\ p(1-y) & 0 & q(1-y) & 0 & py & 0 & r & 0 & qy \\ p(1-x) & 0 & q(1-x) & 0 & 0 & 0 & 0 & r & 0 \\ f(1-y) & 0 & g(1-y) & 0 & fy & 0 & h & 0 & gy \end{pmatrix}$$

We now make the approximation $\mathbb{F}(t)\mathbb{G}(\Delta t) \approx \mathbb{F}(t+\Delta t)$ which is to say that the 9-state machine of Figure 1c can be approximated by the 3-state machine of Figure 1a by infinitesimally increasing the time parameter of the simpler machine. This will not, in general, be exact (with the exception of the TKF model, discussed in Section 2.4). However, by mapping states (and hence transitions) of $\mathbb{F}\mathbb{G}$ back to $\mathbb{F}$, and setting $\mathbb{F}$'s transition probabilities proportional to the expected number of times these transitions are used in $\mathbb{F}\mathbb{G}$, we find a maximum-likelihood fit.

The expected transition counts evolve via the coupled differential equations

$$\begin{aligned} \frac{d}{dt}\bar{T}_{XY}(t) &= \lim_{\Delta t\to 0} \frac{E_{\phi|\mathbb{F}(t+\Delta t)}\left[T_{XY}(\phi)\right] - E_{\phi|\mathbb{F}(t)}\left[T_{XY}(\phi)\right]}{\Delta t} \\ &= \lim_{\Delta t\to 0} \frac{E_{\phi|\mathbb{F}(t)\mathbb{G}(\Delta t)}\left[T_{XY}(\phi)\right] - E_{\phi|\mathbb{F}(t)}\left[T_{XY}(\phi)\right]}{\Delta t} \end{aligned}$$

Expanding (1) to first order in Δt and then taking the limit $\Delta t \to 0$, we arrive at the following equations for the expected transition counts[2]

$$\begin{aligned} \frac{d}{dt}\bar{T}_{MM}(t) &= \mu\frac{bf(1-y)}{1-gy} - (\lambda+\mu)a \\ \frac{d}{dt}\bar{T}_{MI}(t) &= -\mu\frac{b(1-g)}{1-gy} + \lambda(1-b) \\ \frac{d}{dt}\bar{T}_{IM}(t) &= \lambda a - \mu\frac{f(1-g)(b(1-r)+cq)}{(1-gy)(f(1-r)+hp)} \\ \frac{d}{dt}\bar{T}_{DI}(t) &= \mu\frac{(1-g)(b(1-r-hq)+cgq)}{(1-gy)(f(1-r)+hp)} \end{aligned} \tag{5}$$

with boundary condition

$$\bar{T}_{XY}(0) = \begin{cases} 1 & X = Y = M \\ 0 & \text{otherwise} \end{cases}$$

The parameters $(a, b, c, f, g, h, p, q, r)$ are defined by (3) for $t > 0$, but must also be specified as part of the boundary condition at $t = 0$, where $a(0) = f(0) = p(0) = 1$ and $b(0) = c(0) = g(0) = h(0) = q(0) = r(0) = 0$.

[2]These calculations were performed using the symbolic algebra program Mathematica [17].

The remaining counts are obtained from (2)

$$\begin{aligned}
\bar{T}_{MD}(t) &= 1 - \bar{T}_{MM}(t) - \bar{T}_{MI}(t) \\
\bar{T}_{II}(t) &= \bar{S}_I(t) - \bar{T}_{MI}(t) - \bar{T}_{DI}(t) \\
\bar{T}_{ID}(t) &= \bar{T}_{MI}(t) + \bar{T}_{DI}(t) - \bar{T}_{IM}(t) \\
\bar{T}_{DM}(t) &= 1 - \bar{T}_{MM}(t) - \bar{T}_{IM}(t) \\
\bar{T}_{DD}(t) &= \bar{S}_D(t) + \bar{T}_{MM}(t) + \bar{T}_{IM}(t) - \bar{T}_{DI}(t) - 1
\end{aligned} \tag{6}$$

The expected state occupancies are governed by the following equations

$$\begin{aligned}
\frac{d}{dt}\bar{S}_I(t) &= \frac{\lambda}{1-x}\left(1 + \bar{S}_I(t)\right) \\
\frac{d}{dt}\bar{S}_D(t) &= \frac{\mu}{1-y}\left(1 + \bar{S}_D(t)\right)
\end{aligned}$$

which, generalizing [3], have the closed-form solution

$$\begin{aligned}
\bar{S}_I(t) &= \exp\left(\frac{\lambda t}{1-x}\right) - 1 \\
\bar{S}_D(t) &= \exp\left(\frac{\mu t}{1-y}\right) - 1
\end{aligned} \tag{7}$$

2.4 TKF model

When $x = y = 0$, our model reduces to the TKF model [28]. It is known [15] that the solution to the TKF model can be expressed as a Pair HMM of the form shown in Figure 1a, with parameters

$$\begin{array}{lllllllll}
a(t) & = & (1-\beta)\alpha, & b(t) & = & \beta, & c(t) & = & (1-\beta)(1-\alpha), \\
f(t) & = & (1-\beta)\alpha, & g(t) & = & \beta, & h(t) & = & (1-\beta)(1-\alpha), \\
p(t) & = & \gamma(1-\alpha), & q(t) & = & \gamma, & r(t) & = & (1-\gamma)(1-\alpha)
\end{array}$$

where

$$\begin{aligned}
\alpha(t;\lambda,\mu) &= \exp(-\mu t) \\
\beta(t;\lambda,\mu) &= \frac{\lambda\left(\exp(-\lambda t) - \exp(-\mu t)\right)}{\mu\exp(-\lambda t) - \lambda\exp(-\mu t)} \\
\gamma(t;\lambda,\mu) &= 1 - \frac{\mu\beta}{\lambda(1-\alpha)}
\end{aligned}$$

It can readily be verified that this is an exact solution to equations (3) through (7) when $x = y = 0$. Thus, our model reduces exactly to the TKF model when indels involve only single residues. In this case, the equivalence $\mathbb{F}(t + \Delta t) = \mathbb{F}(t)\mathbb{G}(\Delta t)$ is exact in the limit $\Delta t \to 0$.

2.5 Parameterization via EM

Sufficient statistics for parameterizing the infinitesimal process of Figure 1b are

- S, the number of sequences in the dataset;
- n^ℓ, the number of sites at which deletions can occur, integrated over time (ℓ/t is the mean sequence length over the time interval);
- n^λ, the number of insertion events that occurred;
- n^μ, the number of deletion events that occurred;
- n^x, the number of insertion extensions ($n^x + 1$ is the total number of inserted residues);
- n^y, the number of deletion extensions ($n^y + 1$ is the total number of deleted residues).

Given these statistics, the maximum likelihood parameterization is[3]

$$\begin{aligned}
\hat{\lambda} &= n^\lambda/(n^\ell + S) \\
\hat{\mu} &= n^\mu/n^\ell \\
\hat{x} &= n^x/(n^x + n^\lambda) \\
\hat{y} &= n^y/(n^x + n^\mu)
\end{aligned}$$

Let ($\bar{n}^\ell$, $\bar{n}^\lambda$, $\bar{n}^\mu$, $\bar{n}^x$, $\bar{n}^y$) denote the expectations of the sufficient statistics over the posterior distribution of histories. The Expectation Maximization algorithm for continuous-time Markov processes alternates between calculating these posterior expectations for some parameterization (λ_k, μ_k, x_k, y_k) and using them to find a better parameterization [16, 11, 14, 4]

$$\begin{aligned}
\lambda_{k+1} &\leftarrow \bar{n}^\lambda/(\bar{n}^\ell + S) \\
\mu_{k+1} &\leftarrow \bar{n}^\mu/\bar{n}^\ell \\
x_{k+1} &\leftarrow \bar{n}^x/(\bar{n}^x + \bar{n}^\lambda) \\
y_{k+1} &\leftarrow \bar{n}^y/(\bar{n}^x + \bar{n}^\mu)
\end{aligned}$$

In any state path through the machines of Figure 1, each transition will make an additive contribution to these statistics. Let $\bar{n}^Z_{ij}[\mathbb{M}]$ denote the contribution to $\bar{n}^Z$ made by transition $i \to j$ of machine $\mathbb{M}$. With reference to Figure 1c, and by the rules of algebraic automata composition [30], each state of $\mathbb{F}(t)\mathbb{G}(\Delta t)$ can be written as a tuple (i, j) of a $\mathbb{F}$-state i and a $\mathbb{G}$-state j, and each transition weight of $\mathbb{F}(t)\mathbb{G}(\Delta t)$ takes the form

$$Q_{i_1 j_1, i_2 j_2}\left[\mathbb{F}(t)\mathbb{G}(\Delta t)\right] = Q_{i_1,i_2}\left[\mathbb{F}(t)\right]^{\tau_{\mathbb{F}}(i_1 j_1, i_2 j_2)} Q_{j_1,j_2}\left[\mathbb{G}(\Delta t)\right]^{\tau_{\mathbb{G}}(i_1 j_1, i_2 j_2)}$$

where $\tau_{\mathbb{M}}(i_1 j_1, i_2 j_2)$ is 1 if machine $\mathbb{M}$ changes state in the transition $(i_1, j_1) \to (i_2, j_2)$, and 0 if it does not. Using this, we can write

$$\bar{n}^Z_{i_1 j_1, i_2 j_2}\left[\mathbb{F}(t)\mathbb{G}(\Delta t)\right] = \tau_{\mathbb{F}}(i_1 j_1, i_2 j_2)\bar{n}^Z_{i_1,i_2}\left[\mathbb{F}(t)\right] + \tau_{\mathbb{G}}(i_1 j_1, i_2 j_2)\bar{n}^Z_{j_1,j_2}\left[\mathbb{G}(\Delta t)\right]$$

[3] The formula for λ assumes that insertions can occur at the start and end of the sequence, as is usual [23]. Strictly, this requires that we specify a start and end state for $\mathbb{G}$, rather than implicitly assuming infinite-length sequences as we have done up to this point. Specifically we start $\mathbb{G}$ in the match state, and add transitions to the end state from the insert state with weight $1 - x$, from the delete state with weight $1 - y$, and from the match state with weight 1. This can be extended with rigour throughout the analysis by also specifying start and end states for $\mathbb{F}$ and deriving differential equations for the transitions involving these states. Since it complicates the presentation to do this, we have omitted it. A heuristic for $\mathbb{F}$ that is probably acceptable for most applications is to start it in the match state, and to allow transitions to the end state from the insert state with weight $1 - g$, from the delete state with weight $1 - q$, and from the match state with weight $1 - b$.

where

$$
\begin{aligned}
\bar{n}^{\ell}\left[\mathbb{G}(\Delta t)\right] &= \begin{pmatrix} \Delta t & \Delta t & \Delta t \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\bar{n}^{\lambda}\left[\mathbb{G}(\Delta t)\right] &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\bar{n}^{\mu}\left[\mathbb{G}(\Delta t)\right] &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\bar{n}^{x}\left[\mathbb{G}(\Delta t)\right] &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\bar{n}^{y}\left[\mathbb{G}(\Delta t)\right] &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}
\end{aligned}
$$

and thus, for $k_1 \in \sigma_X^{\mathbb{F}}$ and $k_2 \in \sigma_Y^{\mathbb{F}}$,

$$
\bar{n}^{Z}_{k_1,k_2}\left[\mathbb{F}(t+\Delta t)\right] = \sum_{(i_1 j_1)\in\sigma_X^{\mathbb{FG}}} \sum_{(i_2 j_2)\in\sigma_Y^{\mathbb{FG}}} \frac{E_{\phi|\mathbb{F}(t)\mathbb{G}(\Delta t)}\left[T_{i_1 j_1, i_2 j_2}(\phi)\right]}{E_{\phi|\mathbb{F}(t)\mathbb{G}(\Delta t)}\left[T_{XY}(\phi)\right]} \bar{n}^{Z}_{i_1 j_1, i_2 j_2}\left[\mathbb{F}(t)\mathbb{G}(\Delta t)\right]
$$

where $E[T_{ij}]$ is defined by (1) in the same way for transitions between individual states as $E[T_{XY}]$ is defined for transitions between sets of states (e.g. by defining $\sigma_i \equiv \{i\}$ for $i \in \{1 \ldots K\}$). **Conjecture.** Expanding these equations to first order in Δt and taking the limit $\Delta t \to 0$ leads to ODEs for $\bar{n}^{Z}_{ij}\left[\mathbb{F}(t)\right]$, analogous to the ODEs for $\bar{T}_{XY}(t)$ given in (5), that can be used to fit the parameters of the infinitesimal generator from unaligned sequence data by weighting the $\bar{n}^{Z}_{ij}$ with the posterior transition usage counts obtained using the Baum-Welch algorithm.

2.6 Higher moments

We here include a few results relating to our model that may be useful, but are not directly needed to derive the differential equations that govern it.

The matrix method of Section 2.1 can be used to find $E[S_I]$ and $E[S_D]$ directly, as well as higher moments. Let $\mathbf{X} \in \{\mathbf{I}, \mathbf{D}\}$ be the diagonal matrix indicating membership of σ_X, so $X_{ij} = \delta(i=j)\delta(i \in \sigma_X)$. Then

$$
\begin{aligned}
E_{\phi|\mathbb{M}}[S_I] &= (\mathbf{UIW})_{11} \\
E_{\phi|\mathbb{M}}[S_D] &= (\mathbf{UDW})_{11} \\
E_{\phi|\mathbb{M}}[S_I^2] &= (\mathbf{UI}\,(\mathbf{2UI}-\mathbf{1})\,\mathbf{W})_{11} \\
E_{\phi|\mathbb{M}}[S_D^2] &= (\mathbf{UD}\,(\mathbf{2UD}-\mathbf{1})\,\mathbf{W})_{11} \\
E_{\phi|\mathbb{M}}[S_I S_D] &= (\mathbf{U}\,(\mathbf{DUI}+\mathbf{IUD})\,\mathbf{W})_{11}
\end{aligned}
$$

In the case of machine $\mathbb{F}$ in Figure 1a, the first two of these moments have already been given in (4). The others are

$$\begin{aligned}
E_{\phi|\mathbb{F}}[S_I^2] &= \frac{(b(1-r)+cq)(f(1-r)+hp)((1+g)(1-r)+hq)}{((1-g)(1-r)-hq)^3} \\
E_{\phi|\mathbb{F}}[S_D^2] &= \frac{(c(1-g)+bh)(p(1-g)+fq)((1-g)(1+r)+hq)}{((1-g)(1-r)-hq)^3} \\
E_{\phi|\mathbb{F}}[S_I S_D] &= \frac{2hq(bf(1-r)+cp(1-g))+(bhp+cfq)((1-g)(1-r)+hq)}{((1-g)(1-r)-hq)^3}
\end{aligned}$$

These results can also be obtained from the moment generating function for the joint distribution $P(S_I, S_D|\mathbb{F})$, which is

$$\begin{aligned}
F(u,v) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} P(S_I = i, S_D = j|\mathbb{F})u^i v^j \\
&= G(H(u;g), H(v;r)) \\
G(u,v) &= a + \frac{uv(bhp+cfq)+bfu+cpv}{1-hquv} \\
H(u;g) &= \frac{u}{1-gu}
\end{aligned}$$

where G is the generating function for $P(T_{\to I}, T_{\to D}|\mathbb{F})$ where $T_{\to I} = T_{MI} + T_{DI}$ and $T_{\to D} = T_{MD} + T_{ID}$, and H is the generating function for a geometric series.

3 Results

We implemented the likelihood calculations of Section 2.3 and those of De Maio (2020) and Miklós, Lunter and Holmes (2004), which are directly comparable, although De Maio requires that $\lambda = \mu$ and the method of Miklós *et al* is only practical for high-throughput evaluation up to gap lengths of about 30 (Figure 2). As a shorthand, we refer to these methods as DM20 [3], MLH04 [23], and H20 for the current method. We also implemented a simulator for the underlying indel process and compared summary statistics across the different methods, including the relative entropy of the joint distribution $P(S_I, S_D)$ (although for MLH04 it was necessary to renormalize this to $P(S_I, S_D|S_I \leq 30, S_D \leq 30)$ to avoid infinities in the relative entropy) and various marginals and moments of this distribution. Our implementations and simulation results are available at `https://github.com/ihh/trajectory-likelihood`.

We performed simulations at parameter settings $\mu \in \{0.05, 0.5\}$, $\lambda \in \{\mu, 0.9\mu\}$, $y \in \{0.5, 0.75, 0.9\}$, $x \in \{\lambda y/\mu, 0.45, 0.8, 0.95\}$, and $t \in \{2^k : -6 \leq k \leq 3\}$. Note that the existence of a reversible equilibrium by detailed balance requires $\lambda < \mu$ [28] and $\mu x = \lambda y$ [23]. When x is larger than this, the sequence length tends to increase without limit, bringing simulations to a halt. The experiments with $x > \lambda y/\mu$ were abandoned at higher values of t for this reason.

Summary statistics for some of these experiments are plotted in Figure 3 and some relative entropies are also given in Table 1. In almost all cases, the new method is a closer approximation than DM20. The MLH04 trajectory-enumerating approximation is often a closer fit at short times (with the proviso that it can only handle short gaps, and takes significantly longer to compute) but it quickly fails when the number of indel events in a typical trajectory exceeds the 3 events that are computationally tractable to enumerate.

Despite its built-in assumption that insertion and deletion length distributions are separable (the random lengths are conditionally independent given that they are nonzero), DM20 does not perform too much worse than MLH04 at predicting covariation between these lengths (Figure 3f), perhaps because the greatest contribution to the covariation arises from the joint probability $P(S_I = S_D = 0)$, which DM20 does explicitly model (Figure 3d).

We mostly report simulation results with $\lambda = \mu$ to allow comparison with DM20, though we did perform some simulations with $\lambda = 0.9\mu$ in order to compare H20 to MLH04 in this asymmetric-but-reversible regime. We note that the performance of H20 is not significantly worse when $\lambda \neq \mu$ (Figure 3c). By contrast, when $\lambda = \mu$ but $x > y$, so the model is symmetric with respect to insertion/deletion rates but irreversible with respect to their lengths, both H20 and DM20 deteriorate; but DM20 more so than H20 (Figure 3b).

4 Discussion

We have shown that an evolutionary model which can be represented infinitesimally as an HMM can be formally connected to a Pair HMM that approximates its finite-time solution. This may be viewed as an automata-theoretic framing of the Chapman-Kolmogorov equation.

Ours is fundamentally a coarse-graining approach. It is generally the case that composing two state machines will yield a more complex machine, since the composite state space is the Cartesian product of the components. Our approach is to approximate this more complex machine with a renormalizing operation that eliminates null states and maps the remaining states of each type back to a single representative state in the approximator.

We have used this approach to derive ordinary differential equations for the transition probabilities of a minimal (three-state) Pair HMM that approximates a continuous-time indel process with geometrically distributed indel lengths. We have implemented numerical solutions to these equations, and demonstrated that they outperform the previous best methods. We have also conjectured similar ODEs for the posterior expectations of the sufficient statistics that would be required to fit this model by Expectation Maximization, though we have not yet tested this approach.

It is tempting to speculate that this method may also be productively applied to stochastic context-free grammars (SCFGs) [13, 2]. Despite the appeal of describing a realistic evolutionary drift process over RNA structures and parse trees, doing so might present some difficulties, since elim-

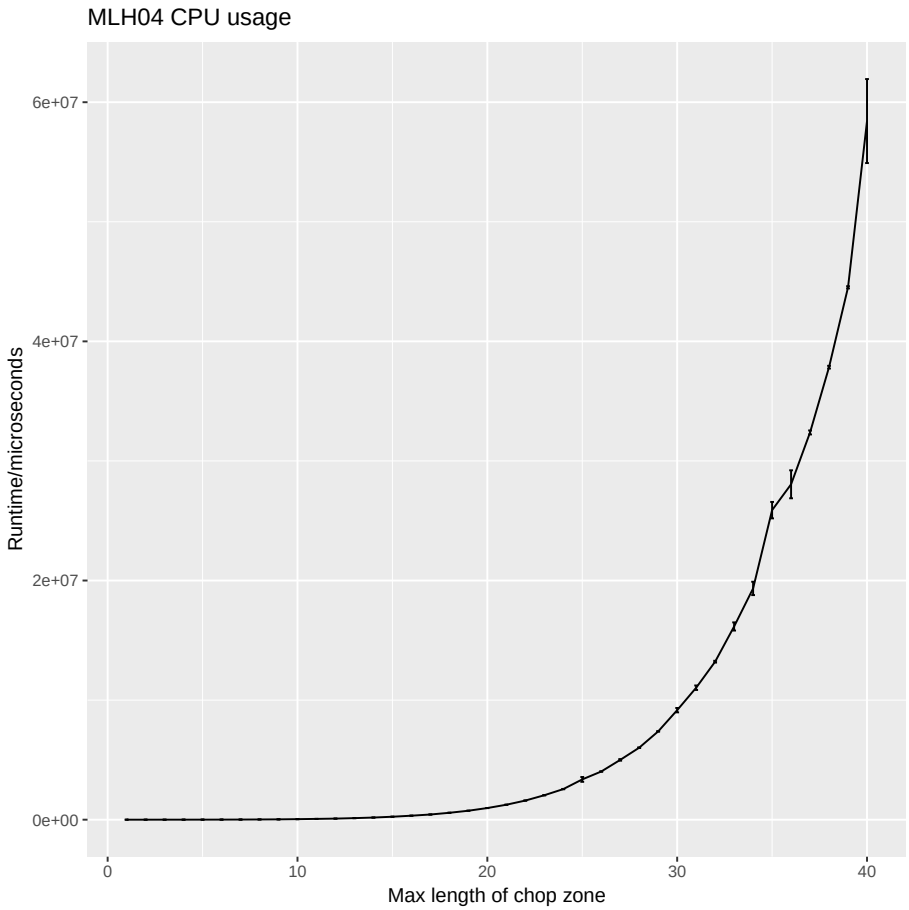


Figure 2: The running time of the MLH04 method is prohibitive and increases steeply for long gaps, since it requires the explicit enumeration of all intermediate indel states in the trajectory [23]. The running times of DM20 and MLH04 are negligible in comparison ($<$ 1ms).

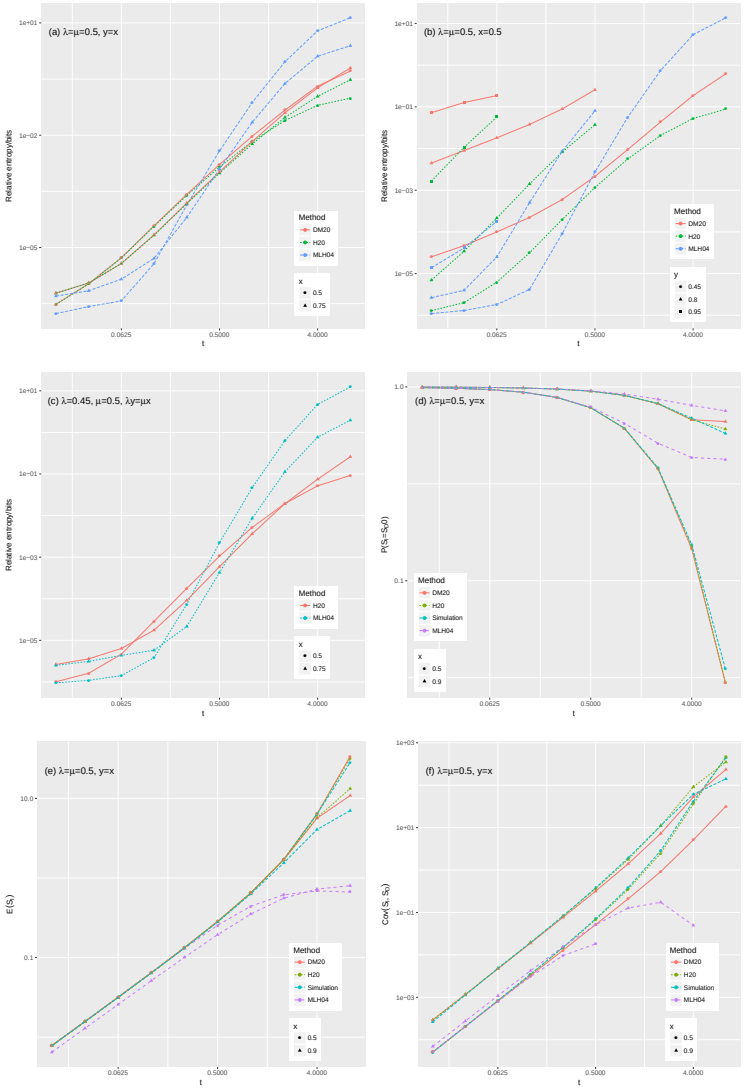


Figure 3: Summary statistics of simulation benchmarks plotted against evolutionary time. Shown are the relative entropies from the simulation distribution to the various approximations for reversible (a), irreversible-with-equal-rates (b), and reversible-with-inequal-rates (c) parameter regimes; the probability of observing no gap (d); the expected number of insertions (e); and the covariance between the number of insertions and deletions (f).

					$D(P(S_I, S_D\|\text{simulated})\|\|P(S_I, S_D\|\text{approx}))$		
μ	x	y	t	Rev?	DM20 [3]	MLH04 [23]	H20
0.05	0.5	0.45	0.125		2.04985×10^{-5}	9.11216×10^{-7}	8.81290×10^{-7}
0.05	0.5	0.45	1		1.68563×10^{-4}	2.69044×10^{-6}	1.69689×10^{-5}
0.05	0.5	0.45	8		5.79716×10^{-3}	2.23302×10^{-2}	3.45910×10^{-3}
0.05	0.5	0.5	0.125	✓	1.22097×10^{-6}	1.17394×10^{-6}	1.22242×10^{-6}
0.05	0.5	0.5	1	✓	2.18927×10^{-5}	3.27733×10^{-6}	2.12916×10^{-5}
0.05	0.5	0.5	8	✓	5.42222×10^{-3}	3.00300×10^{-2}	4.20105×10^{-3}
0.05	0.5	0.8	0.125		3.57032×10^{-3}	2.40144×10^{-6}	4.91577×10^{-6}
0.05	0.75	0.45	0.125		4.26426×10^{-4}	9.50348×10^{-7}	1.11389×10^{-6}
0.05	0.75	0.45	1		3.38901×10^{-3}	2.94778×10^{-6}	8.41352×10^{-6}
0.05	0.75	0.45	8		2.93007×10^{-2}	1.31450×10^{-3}	1.73884×10^{-3}
0.05	0.75	0.75	0.125	✓	3.03606×10^{-6}	2.86492×10^{-6}	3.03568×10^{-6}
0.05	0.75	0.75	1	✓	1.83832×10^{-5}	8.69690×10^{-6}	1.81854×10^{-5}
0.05	0.75	0.75	8	✓	3.75601×10^{-3}	9.11449×10^{-3}	3.32154×10^{-3}
0.05	0.75	0.8	0.125		6.93445×10^{-5}	2.84201×10^{-6}	3.27834×10^{-6}
0.05	0.9	0.45	0.125		4.59847×10^{-4}	7.05138×10^{-7}	1.28355×10^{-6}
0.05	0.9	0.45	1		3.68176×10^{-3}	2.30351×10^{-6}	5.26999×10^{-6}
0.05	0.9	0.45	8		3.01802×10^{-2}	1.62709×10^{-5}	6.49049×10^{-4}
0.05	0.9	0.8	0.125		1.40027×10^{-4}	2.11248×10^{-6}	2.84768×10^{-6}
0.05	0.9	0.9	0.125	✓	5.83560×10^{-6}	4.19581×10^{-6}	5.83597×10^{-6}
0.05	0.9	0.9	1	✓	2.94696×10^{-5}	4.60033×10^{-5}	2.94067×10^{-5}
0.05	0.9	0.9	8	✓	1.59644×10^{-3}	2.88432×10^{-3}	1.45752×10^{-3}
0.5	0.5	0.45	0.125		2.22362×10^{-4}	4.17791×10^{-6}	3.15751×10^{-5}
0.5	0.5	0.45	1		9.41950×10^{-3}	5.63014×10^{-2}	5.59983×10^{-3}
0.5	0.5	0.45	8		6.20265×10^{-1}	1.36809e+01	8.83542×10^{-2}
0.5	0.5	0.5	0.125	✓	3.82629×10^{-5}	3.70476×10^{-6}	3.68489×10^{-5}
0.5	0.5	0.5	1	✓	9.37828×10^{-3}	7.47712×10^{-2}	6.79378×10^{-3}
0.5	0.5	0.5	8	✓	5.16437×10^{-1}	1.37034e+01	9.75359×10^{-2}
0.5	0.5	0.8	0.125		3.78591×10^{-2}	5.03568×10^{-4}	1.41159×10^{-3}
0.5	0.75	0.45	0.125		4.24415×10^{-3}	1.89067×10^{-6}	1.16934×10^{-5}
0.5	0.75	0.45	1		3.77702×10^{-2}	3.42854×10^{-3}	3.13205×10^{-3}
0.5	0.75	0.45	8		8.54601×10^{-1}	9.90949×10^{-1}	2.84669×10^{-1}
0.5	0.75	0.75	0.125	✓	2.17573×10^{-5}	5.14187×10^{-6}	2.13362×10^{-5}
0.5	0.75	0.75	1	✓	6.80341×10^{-3}	2.20785×10^{-2}	5.83371×10^{-3}
0.5	0.75	0.75	8	✓	6.25327×10^{-1}	2.45203e+00	3.02844×10^{-1}
0.5	0.75	0.8	0.125		6.99602×10^{-4}	2.01321×10^{-5}	4.86974×10^{-5}
0.5	0.9	0.45	0.125		4.59746×10^{-3}	2.79371×10^{-6}	8.60162×10^{-6}
0.5	0.9	0.45	1		3.80985×10^{-2}	2.94801×10^{-5}	1.13099×10^{-3}
0.5	0.9	0.45	8		4.67302×10^{-1}	1.44795×10^{-2}	1.96366×10^{-1}
0.5	0.9	0.8	0.125		1.40555×10^{-3}	1.38761×10^{-5}	1.65939×10^{-5}
0.5	0.9	0.9	0.125	✓	3.65783×10^{-5}	6.86812×10^{-5}	3.64414×10^{-5}
0.5	0.9	0.9	1	✓	2.86092×10^{-3}	4.97623×10^{-3}	2.55231×10^{-3}
0.5	0.9	0.9	8	✓	4.92217×10^{-1}	2.11573×10^{-1}	3.28780×10^{-1}

Table 1: Selected relative entropies of simulation benchmark results. In each row the best method is highlighted dark red, and the second-best is highlighted light red.

ination of null states from SCFGs is more complicated than for HMMs.

References

[1] A. Bouchard-Côté. A note on probabilistic models over strings: the linear algebra approach. *Bulletin of Mathematical Biology*, 75(12):2529–2550, Dec 2013.

[2] R. K. Bradley and I. Holmes. Evolutionary triplet models of structured RNA. *PLoS Computational Biology*, 5(8):e1000483, Aug 2009.

[3] N. De Maio. The cumulative indel model: fast and accurate statistical evolutionary alignment. *Systematic Biology*, 2020.

[4] C. R. Doss, M. A. Suchard, I. Holmes, M. Kato-Maeda, and V. N. Minin. Fitting Birth-Death Processes to Panel Data with Applications to Bacterial DNA Fingerprinting. *Ann Appl Stat*, 7(4):2315–2335, 2013.

[5] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[6] K. Ezawa. Erratum to: General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable? *BMC Bioinformatics*, 17(1):457, Nov 2016.

[7] K. Ezawa. General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable? *BMC Bioinformatics*, 17:304, Aug 2016.

[8] K. Ezawa. General continuous-time Markov model of sequence evolution via insertions/deletions: local alignment probability computation. *BMC Bioinformatics*, 17(1):397, Sep 2016.

[9] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[10] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[11] A. Hobolth and J. L. Jensen. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical applications in Genetics and Molecular Biology*, 4(1), 2005.

[12] I. Holmes. Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, 19 Suppl. 1:i147–157, 2003.

[13] I. Holmes. A probabilistic model for the evolution of RNA structure. *BMC Bioinformatics*, 5(166), 2004.

[14] I. Holmes. Using evolutionary Expectation Maximization to estimate indel rates. *Bioinformatics*, 21(10):2294–2300, 2005.

[15] I. Holmes and W. J. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–820, 2001.

[16] I. Holmes and G. M. Rubin. An Expectation Maximization algorithm for training hidden substitution models. *Journal of Molecular Biology*, 317(5):757–768, 2002.

[17] Wolfram Research, Inc. Mathematica, Version 12.1. Champaign, IL, 2020.

[18] V. Jojic, N. Jojic, C. Meek, D. Geiger, A. Siepel, D. Haussler, and D. Heckerman. Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics*, 20 Suppl 1:i161–168, Aug 2004.

[19] T. H. Jukes and C. Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[20] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[21] B. Knudsen and M. Miyamoto. Sequence alignments and pair hidden Markov models using evolutionary history. *Journal of Molecular Biology*, 333(2):453–460, 2003.

[22] G.A. Lunter and J. Hein. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20 Suppl 1:I216–I223, 2004.

[23] I. Miklós, G. Lunter, and I. Holmes. A long indel model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3):529–540, 2004.

[24] I. Miklós and Z. Toroczkai. An improved model for statistical alignment. In *First Workshop on Algorithms in Bioinformatics*, Berlin, Heidelberg, 2001. Springer-Verlag.

[25] B. D. Redelings and M. A. Suchard. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evolutionary Biology*, 7:40, 2007.

[26] E. Rivas and S. R. Eddy. Parameterizing sequence alignment with an explicit evolutionary model. *BMC Bioinformatics*, 16:406, 2015.

[27] J. Silvestre-Ryan, Y. Wang, M. Sharma, S. Lin, Y. Shen, S. Dider, and I. Holmes. Machine Boss: Rapid prototyping of bioinformatic automata. *bioRxiv*, 2020.

[28] J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33:114–124, 1991.

[29] J. L. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, 34:3–16, 1992.

[30] O. Westesson, G. Lunter, B. Paten, and I. Holmes. Phylogenetic automata, pruning, and multiple alignment. *arXiv*, 2011. arXiv:1103.4347.

[31] Oscar Westesson, Gerton Lunter, Benedict Paten, and Ian Holmes. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. *PLoS One*, 7(4):e34572, 20 April 2012.

[32] Y. Wexler and D. Geiger. Variational upper and lower bounds for probabilistic graphical models. *J. Comput. Biol.*, 15(7):721–735, Sep 2008.