Supplemental Information: A Model of Indel Evolution by Finite-State, Continuous-Time Machines

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

October 2, 2020

Review of statistical models of sequence evolution

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 (Jukes and Cantor, 1969), solved exactly for an ancestor-descendant pair by considering the eigenstructure of the matrix exponential (Kimura, 1980; Hasegawa *et al.*, 1985), and extended to multiple aligned sequences by applying the sum-product algorithm to the phylogenetic tree (Felsenstein, 1981). These results are widely used in bioinformatics. Latent variables can be introduced to model rate heterogeneity (Yang, 1995) or selection (Yang *et al.*, 2000), and the rate parameters estimated efficiently by Expectation-Maximization (Holmes and Rubin, 2002; Hobolth and Jensen, 2005). The site-independent point substitution process 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 (Lunter and Hein, 2004), extending to multiple alignments using variational (mean-field) approaches (Jojic *et al.*, 2004; Wexler and Geiger, 2008).

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 (TKF91) model, which allows only single-residue indels. The TKF model reduces exactly to a linear birth-death process with immigration (Thorne *et al.*, 1991), which allows the joint distribution over ancestor-descendant sequence alignments to be expressed as a Hidden Markov Model (HMM) (Holmes and Bruno, 2001) that can be formally extended to multiple sequences using algebraic composition of automata (Steel and Hein, 2001; Hein, 2001; Holmes, 2003; Westesson *et al.*, 2011; Bouchard-Côté, 2013). This allows a statistical unification of alignment and phylogeny (Lunter *et al.*, 2003; Redelings and Suchard, 2005; Suchard and Redelings, 2006; Novak *et al.*, 2008; Paten *et al.*, 2008; Bouchard-Côté *et al.*, 2009; Westesson *et al.*, 2012a,b; Arunapuram *et al.*, 2013; Herman *et al.*, 2014; Holmes, 2017). However, in practice, the TKF91

model itself is mostly used for inspiration in such applications, since its restriction to single-residue indel events is not consistent with empirical data (Qian and Goldstein, 2001; Chang and Benner, 2004; Strope *et al.*, 2006; Cartwright, 2008) and the consequent over-counting of events causes artefacts in statistical inference of alignments, trees, and rate parameters (Thorne *et al.*, 1992; Hein *et al.*, 2000; Holmes and Bruno, 2001).

Attempts to generalize the TKF91 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 (Miklós and Toroczkai, 2001; Knudsen and Miyamoto, 2003; Miklós *et al.*, 2004; Ezawa, 2016c,b,a; De Maio, 2020), and those that guess closed-form approximations to these probabilities without such *ab initio* justifications (Thorne *et al.*, 1992; Mitchison, 1999; Wang *et al.*, 2006; Redelings and Suchard, 2007; Rivas and Eddy, 2008, 2015; Bouchard-Côté and Jordan, 2013). In this paper we focus on the former type of approach. The latter 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 (De Maio, 2020).

Formal mathematical treatment of the multi-residue indel process begins with Miklós and Toroczkai's analysis of a model that allows long insertions but only single-residue deletions (Miklós and Toroczkai, 2001). 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 (Knudsen and Miyamoto, 2003; Miklós *et al.*, 2004; Ezawa, 2016b), 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.* (2004), formalizing intuition of Knudsen and Miyamoto (2003), 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 (2020). 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 TKF91 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 approximate 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 TKF91 model emerges as a special case and the closed-form solutions to the TKF91 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 (Inc., 2020), and for the manipulation of state machines, for which we used our recently published software Machine Boss (Silvestre-Ryan *et al.*, 2020).

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 TKF91 model as an exact special case, without the need to introduce any additional latent variables such as fragment boundaries.

While our focus is 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, might also be productively analyzed using this approach.

Parameterization of evaluated Pair HMMs

This section gives the mapping that we used between the parameters (λ, μ, x, y) of the GGI model and the transition probabilities of Figure 1, or other model parameters, for the various approximations that we evaluated. In most cases these follow De Maio (2020), though we have extended the mapping to allow for asymmetry between insertions and deletions.

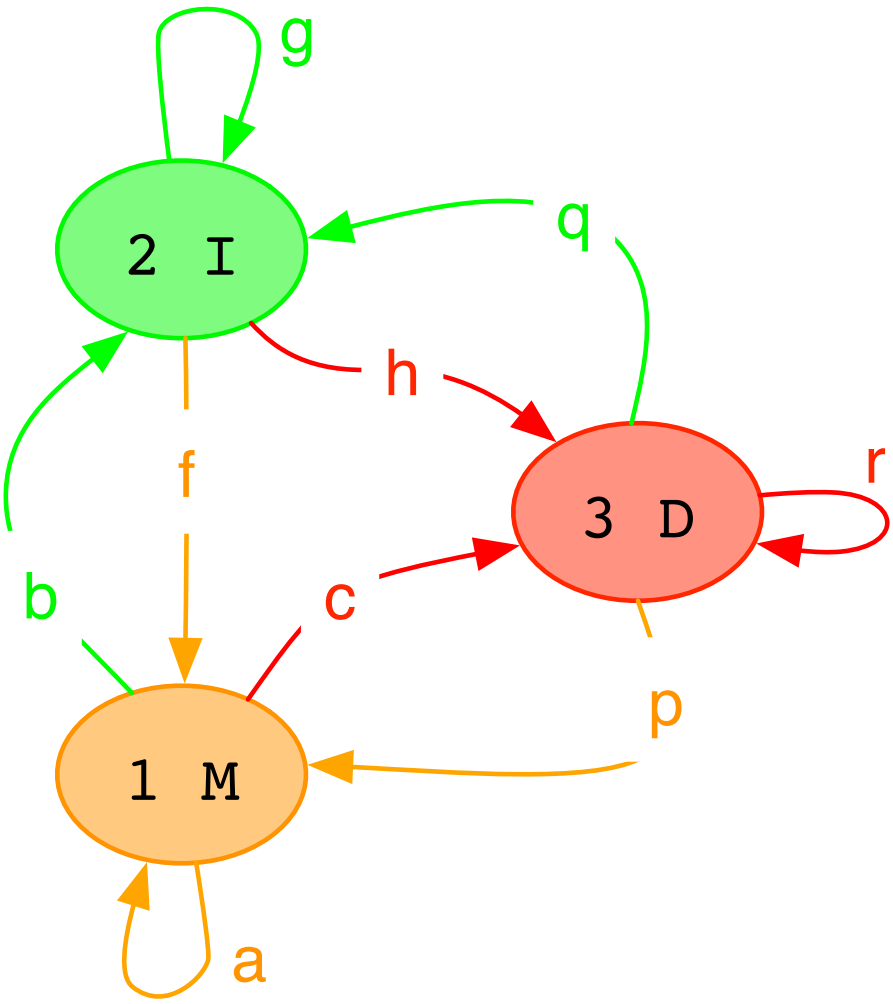

Figure 1: Machine $\mathbb{F}(t)$.

Long indel model

If the subscript LI denotes a parameter of the Long Indel model then

$$\begin{aligned} \lambda_{LI} &= \frac{\lambda}{1-x} \\ \mu_{LI} &= \frac{\mu}{1-y} \end{aligned}$$

Parameters x and y may be used unmodified. (Technically, the Long Indel model is defined in (Miklós *et al.*, 2004) to be reversible, which would constrain x and y by $\lambda_{LI}/\mu_{LI} = x/y$. However, this constraint is not required: the trajectory calculations can be performed for an irreversible model using exactly the same algorithms and equations.)

TKF91

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

where

$$\begin{aligned} \alpha &= \exp(-\mu_0 t) \\ \beta &= \begin{cases} \frac{\lambda_0(\exp(-\lambda_0 t)-\exp(-\mu_0 t))}{\mu_0\exp(-\lambda_0 t)-\lambda\exp(-\mu_0 t)} & \lambda_0 \neq \mu_0 \\ \frac{\lambda_0 t}{1+\lambda_0 t} & \lambda_0 = \mu_0 \end{cases} \\ \gamma &= 1 - \frac{\mu_0\beta}{\lambda_0(1-\alpha)} \\ \lambda_0 &= \frac{\lambda}{1-x} \\ \mu_0 &= \frac{\mu}{1-y} \end{aligned}$$

TKF92

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

where α, β, γ are as defined as in TKF91 and $\epsilon = \frac{1}{2}(x+y)$.

LG05

$$\begin{array}{lllllllll} a(t) & = & \epsilon + (1-\epsilon)(1-2\delta), & b(t) & = & (1-\epsilon)\delta, & c(t) & = & (1-\epsilon)\delta, \\ f(t) & = & (1-\epsilon)(1-2\delta), & g(t) & = & \epsilon + (1-\epsilon)\delta, & h(t) & = & (1-\epsilon)\delta, \\ p(t) & = & (1-\epsilon)(1-2\delta), & q(t) & = & \epsilon + (1-\epsilon)\delta, & r(t) & = & (1-\epsilon)\delta \end{array}$$

where

$$\begin{aligned}\delta &= 1 - \exp\left(-\frac{\rho t}{1-\epsilon}\right)\\ \epsilon &= \frac{1}{2}(x+y)\\ \rho &= \frac{1}{2}(\lambda+\mu)\end{aligned}$$

RS07

$$\begin{array}{lllllllll} a(t) &=& \epsilon + (1-\epsilon)(1-2\delta), & b(t) &=& (1-\epsilon)\delta, & c(t) &=& (1-\epsilon)\delta,\\ f(t) &=& (1-\epsilon)(1-2\delta), & g(t) &=& \epsilon + (1-\epsilon)\delta, & h(t) &=& (1-\epsilon)\delta,\\ p(t) &=& (1-\epsilon)(1-2\delta), & q(t) &=& \epsilon + (1-\epsilon)\delta, & r(t) &=& (1-\epsilon)\delta \end{array}$$

where

$$\begin{aligned}\delta &= \left(1 + \frac{1}{1-\exp\left(-\frac{\rho t}{1-\epsilon}\right)}\right)^{-1}\\ \epsilon &= \frac{1}{2}(x+y)\\ \rho &= \frac{1}{2}(\lambda+\mu)\end{aligned}$$

Parameterization via EM

Sufficient statistics for parameterizing the GGI model are

- S, the number of alignments 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[1]

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

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 $(\lambda_k, \mu_k, x_k, y_k)$ and using them to find a better parameterization (Holmes and Rubin, 2002; Hobolth and Jensen, 2005; Holmes, 2005; Doss *et al.*, 2013)

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

In any state path through the machines $\mathbb{F}$, $\mathbb{G}$, and $\mathbb{FG}$, 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 the matrix and diagrammatic representations in the main paper, and by the rules of algebraic automata composition (Westesson *et al.*, 2011), 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 the individual machine $\mathbb{M}$ makes a transition as part of the compound transition $(i_1, j_1) \to (i_2, j_2)$, and 0 if it does not. Using this, we can write

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

[1]The formula for λ assumes that insertions can occur at the start and end of the sequence, as is usual (Miklós *et al.*, 2004). Strictly, this requires that we specify a start and end state for $\mathbb{G}$, rather than implicitly assuming infinite-length sequences. 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, and from the match state with weight 1. This can be extended with rigor 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 $1-b$. Versions of $\mathbb{G}$ and $\mathbb{F}$ that introduce start and end states in this way are shown in the main paper.

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 $k_1 \in \sigma_{\mathtt{X}}^{\mathbb{F}}$ and $k_2 \in \sigma_{\mathtt{Y}}^{\mathbb{F}}$,

$$\bar{n}^{Z}_{k_1,k_2}\left[\mathbb{F}(t+\Delta t)\right] = \sum_{(i_1 j_1)\in\sigma_{\mathtt{X}}^{\mathbb{FG}}} \sum_{(i_2 j_2)\in\sigma_{\mathtt{Y}}^{\mathbb{FG}}} \frac{E_{\phi|\mathbb{F}(t)\mathbb{G}(\Delta t)}\left[T_{\mathtt{i}_1\mathtt{j}_1,\mathtt{i}_2\mathtt{j}_2}(\phi)\right]}{E_{\phi|\mathbb{F}(t)\mathbb{G}(\Delta t)}\left[T_{\mathtt{XY}}(\phi)\right]} \bar{n}^{Z}_{i_1 j_1, i_2 j_2}\left[\mathbb{F}(t)\mathbb{G}(\Delta t)\right]$$

where $E[T_{\mathtt{ij}}]$ is defined in the same way for transitions between individual states as $E[T_{\mathtt{XY}}]$ is defined for transitions between sets of states (e.g. by defining $\sigma_{\mathtt{i}} \equiv \{i\}$ 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 $\bar{T}_{\mathtt{XY}}(t)$ given in the main paper, 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.

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 the main paper can be used to find $E[S_{\mathtt{I}}]$ and $E[S_{\mathtt{D}}]$ directly, as well as higher moments. Let $\mathbf{X} \in \{\mathbf{I}, \mathbf{D}\}$ be the diagonal matrix indicating membership of $\sigma_{\mathtt{X}}$, so $X_{ij} = \delta(i=j)\delta(i \in \sigma_{\mathtt{X}})$. Then

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

In the case of machine $\mathbb{F}$, the first two of these moments have already been given in the main paper. The others are

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

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

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

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

Discussion

Point substitution models are the foundation of likelihood phylogenetics (Huelsenbeck and Crandall, 1997; Felsenstein, 2003). There is, additionally, a substantial literature combining such models with HMMs (Yang, 1995; Felsenstein and Churchill, 1996; Goldman *et al.*, 1996; Liò and Goldman, 1999; Pedersen and Hein, 2003; Siepel and Haussler, 2004; McCauley and Hein, 2006; Heger *et al.*, 2009; Nguyen Ba *et al.*, 2012; Dhar *et al.*, 2019) and stochastic context-free grammars (SCFGs) (Knudsen and Hein, 2003; Pedersen *et al.*, 2004; Holmes, 2004; Westesson and Holmes, 2012; Sksd *et al.*, 2013) for purposes of sequence

annotation. The development of indel models has been slower, despite evidence that indels are a potentially powerful signal—for example, selection for phase-preserving indels is a highly revealing signature of protein-coding genes (Kellis *et al.*, 2003). This may be, in large part, because integrating alignment and phylogeny is technically and computationally demanding. Multiple sequence alignments are a nuisance variable whose point estimation is a tolerable compromise when considering substitution processes, although several studies report that bias due to alignment error is a significant problem in substitution-founded phylogenetics (Hartmann and Vision, 2008; Sksd *et al.*, 2013; Levy Karin *et al.*, 2014; Md Mukarram Hossain *et al.*, 2015; Bogusz and Whelan, 2016) that must be handled with great care to avoid biasing inference (Jordan and Goldman, 2012; Privman *et al.*, 2012; Sela *et al.*, 2015). When it comes to indel-based analysis, this compromise of conditioning on a single alignment rarely remains tenable, except perhaps in the "big data" limit, e.g. for closely-related sequences at genome scale (Lunter *et al.*, 2006; Rands *et al.*, 2014). So indel-based phylogenetic inference must often co-sample or otherwise marginalize alignments, which is inherently harder (Suchard and Redelings, 2006; Novak *et al.*, 2008; Westesson *et al.*, 2012a; Holmes, 2017). Nevertheless, the inexactitude of existing long-indel approximations may also have been a contributing obstacle to their slow adoption in the bioinformatic tool chain. If so, then the results presented here might help.

Our emphasis on the Generic Geometric Indel model, a continuous-time Markov process defined on sequences of residues, somewhat disadvantages models like TKF92, which technically defines a process on sequences of multi-residue fragments. Our working assumption has been that there is no evidence such indivisible fragments really exist, and so we have instead evaluated TKF92 as an approximation to the GGI model. However, the routine usage of amino acid fragment models to predict protein tertiary structure (Simons *et al.*, 1999) suggests a valid counter-argument that such models may usefully capture some forms of selection. Further, TKF92 can be generalized in other ways, allowing for richer models of fragment mutation; for example to model the evolution of RNA structure (Holmes, 2004). In this context, it is promising that our method recovers TKF91 (and therefore TKF92) as special cases.

It seems possible that our method can be applied to other instantaneous rate models of local evolution where the infinitesimal generator can be represented as an HMM. It is tempting to speculate that a similar approach may also be productively applied to SCFGs (Holmes, 2004; Bradley and Holmes, 2009). Such an approach would be more challenging; for example, elimination of null states from SCFGs is more complicated than for HMMs. One motivating goal would be to describe a realistic evolutionary drift process over RNA structures, with the goal of reconstructing the RNA world (Meyer and Miklós, 2007). It's also conceivable that approaches similar to those described here for biological sequences could be used to analyze phonemes (Bouchard-Côté *et al.*, 2009), literary texts (Barbrook *et al.*, 1998), music (Cochrane and Gatherer, 2020), source code (Miller and Myers, 1985), bird songs (Kershenbaum and Garland, 2015), or other alignable sequences that evolve over time.

References

Arunapuram, P., I. Edvardsson, M. Golden, J. W. Anderson, A. Novak, *et al.*, 2013 StatAlign 2.0: combining statistical alignment with RNA secondary structure prediction. Bioinformatics **29**: 654–655.

Barbrook, A. C., C. J. Howe, N. J. Blake, and P. Robinson, 1998 The phylogeny of the canterbury tales. Nature **394**: 839–839.

Bogusz, M. and S. Whelan, 2016 Phylogenetic Tree Estimation With and Without Alignment: New Distance Methods and Benchmarking. Systematic Biology **66**: 218–231.

Bouchard-Côté, A., 2013 A note on probabilistic models over strings: the linear algebra approach. Bulletin of Mathematical Biology **75**: 2529–2550.

Bouchard-Côté, A. and M. I. Jordan, 2013 Evolutionary inference via the Poisson Indel Process. Proc. Natl. Acad. Sci. U.S.A. **110**: 1160–1166.

Bouchard-Côté, A., D. Klein, and M. I. Jordan, 2009 Efficient Inference in Phylogenetic InDel Trees. In *Advances in Neural Information Processing Systems 21*, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, pp. 177–184, Curran Associates, Inc., Vancouver, British Columbia, Canada.

Bradley, R. K. and I. Holmes, 2009 Evolutionary triplet models of structured RNA. PLoS Computational Biology **5**: e1000483.

Cartwright, R. A., 2008 Problems and Solutions for Estimating Indel Rates and Length Distributions. Molecular Biology and Evolution **26**: 473–480.

Chang, M. S. and S. A. Benner, 2004 Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J. Mol. Biol. **341**: 617–631.

Cochrane, L. J. and D. Gatherer, 2020 Dynamic programming algorithms applied to musical counterpoint in process composition: An example using Henri Pousseurs Scambi.

De Maio, N., 2020 The Cumulative Indel Model: fast and accurate statistical evolutionary alignment. Systematic Biology .

Dhar, A., D. K. Ralph, V. N. Minin, and F. A. M. IV, 2019 A bayesian phylogenetic hidden markov model for b cell receptor sequence analysis.

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

Ezawa, K., 2016a Erratum to: General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable? BMC Bioinformatics **17**: 457.

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

Ezawa, K., 2016c General continuous-time Markov model of sequence evolution via insertions/deletions: local alignment probability computation. BMC Bioinformatics **17**: 397.

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

Felsenstein, J., 2003 *Inferring Phylogenies*. Sinauer Associates, Inc., ISBN 0878931775.

Felsenstein, J. and G. A. Churchill, 1996 A hidden Markov model approach to variation among sites in rate of evolution. Molecular Biology and Evolution **13**: 93–104.

Goldman, N., J. L. Thorne, and D. T. Jones, 1996 Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. Journal of Molecular Biology **263**: 196–208.

Hartmann, S. and T. Vision, 2008 Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC evolutionary biology **8**: 95.

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

Heger, A., C. P. Ponting, and I. Holmes, 2009 Accurate estimation of gene evolutionary rates using XRATE, with an application to transmembrane proteins. Molecular Biology and Evolution **26**: 1715–1721.

Hein, J., 2001 An algorithm for statistical alignment of sequences related by a binary tree. In *Pacific Symposium on Biocomputing*, edited by R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, pp. 179–190, Singapore, World Scientific.

Hein, J., C. Wiuf, B. Knudsen, M. B. Moller, and G. Wibling, 2000 Statistical alignment: computational properties, homology testing and goodness-of-fit. Journal of Molecular Biology **302**: 265–279.

Herman, J. L., C. J. Challis, A. Novak, J. Hein, and S. C. Schmidler, 2014 Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. Mol. Biol. Evol. **31**: 2251–2266.

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

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

Holmes, I., 2004 A probabilistic model for the evolution of RNA structure. BMC Bioinformatics **5**.

Holmes, I., 2005 Using evolutionary Expectation Maximization to estimate indel rates. Bioinformatics **21**: 2294–2300.

Holmes, I., 2017 Historian: Accurate reconstruction of ancestral sequences and evolutionary rates. Bioinformatics (Oxford, England) **33**.

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

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

Huelsenbeck, J. P. and K. A. Crandall, 1997 Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics **28**: 437–466.

Inc., W. R., 2020 Mathematica, Version 12.1. Champaign, IL, 2020.

Jojic, V., N. Jojic, C. Meek, D. Geiger, A. Siepel, *et al.*, 2004 Efficient approximations for learning phylogenetic HMM models from data. Bioinformatics **20 Suppl 1**: i161–168.

Jordan, G. and N. Goldman, 2012 The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Molecular biology and evolution **29**: 1125–1139.

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

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander, 2003 Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature **423**: 241–254.

Kershenbaum, A. and E. C. Garland, 2015 Quantifying similarity in animal vocal sequences: which metric performs best? Methods in Ecology and Evolution **6**: 1452–1461.

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

Knudsen, B. and J. Hein, 2003 Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Research **31**: 3423–3428.

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

Levy Karin, E., E. Susko, and T. Pupko, 2014 Alignment errors strongly impact likelihood-based tests for comparing topologies. Molecular biology and evolution **31**: 3057–3067.

Liò, P. and N. Goldman, 1999 Using protein structural information in evolutionary inference: transmembrane proteins. Molecular Biology and Evolution **16**: 1696–1710.

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

Lunter, G., C. P. Ponting, and J. Hein, 2006 Genome-wide identification of human functional DNA using a neutral indel model. PLoS Computational Biology **2**.

Lunter, G. A., I. Miklós, Y. S. Song, and J. Hein, 2003 An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. Journal of Computational Biology **10**: 869–889.

McCauley, S. and J. Hein, 2006 Using hidden Markov models and observed evolution to annotate viral genomes. Bioinformatics **22**: 1308–1316.

Md Mukarram Hossain, A., B. P. Blackburne, A. Shah, and S. Whelan, 2015 Evidence of Statistical Inconsistency of Phylogenetic Methods in the Presence of Multiple Sequence Alignment Uncertainty. Genome Biology and Evolution **7**: 2102–2116.

Meyer, I. M. and I. Miklós, 2007 SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. PLoS Comput. Biol. **3**: e149.

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

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

Miller, W. and E. W. Myers, 1985 A file comparison program. Software Practice and Experience **15**: 1025–1040.

Mitchison, G. J., 1999 A probabilistic treatment of phylogeny and sequence alignment. Journal of Molecular Evolution **49**: 11–22.

Nguyen Ba, A. N., B. J. Yeh, D. van Dyk, A. R. Davidson, B. J. Andrews, *et al.*, 2012 Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Science Signaling **5**: rs1–rs1.

Novak, A., I. Miklós, R. Lyngsoe, and J. Hein, 2008 StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics **24**: 2403–2404.

Paten, B., J. Herrero, S. Fitzgerald, P. Flicek, I. Holmes, *et al.*, 2008 Genome-wide nucleotide level mammalian ancestor reconstruction. Genome Research **18**: 1829–1843.

Pedersen, J. S. and J. Hein, 2003 Gene finding with a hidden Markov model of genome structure and evolution. Bioinformatics **19**: 219–227.

Pedersen, J. S., I. M. Meyer, R. Forsberg, P. Simmonds, and J. Hein, 2004 A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic Acids Research **32**: 4925–4923.

Privman, E., O. Penn, and T. Pupko, 2012 Improving the performance of positive selection inference by filtering unreliable alignment regions. Molecular biology and Evolution **29**: 1–5.

Qian, B. and R. A. Goldstein, 2001 Distribution of indel lengths. Proteins: Structure, Function, and Bioinformatics **45**: 102–104.

Rands, C. M., S. Meader, C. P. Ponting, and G. Lunter, 2014 8.2% of the human genome is constrained: Variation in rates of turnover across functional element classes in the human lineage. PLOS Genetics **10**: 1–12.

Redelings, B. D. and M. A. Suchard, 2005 Joint Bayesian estimation of alignment and phylogeny. Systematic Biology **54**: 401–418.

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

Rivas, E. and S. Eddy, 2008 Probabilistic phylogenetic inference with insertions and deletions. PLoS Computational Biology **4**: e1000172.

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

Sela, I., H. Ashkenazy, K. Katoh, and T. Pupko, 2015 GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Research **43**: W7–W14.

Siepel, A. and D. Haussler, 2004 Combining phylogenetic and hidden Markov models in biosequence analysis. Journal of Computational Biology **11**: 413–428.

Silvestre-Ryan, J., Y. Wang, M. Sharma, S. Lin, Y. Shen, *et al.*, 2020 Machine Boss: Rapid Prototyping of Bioinformatic Automata. Bioinformatics .

Simons, K. T., R. Bonneau, I. Ruczinski, and D. Baker, 1999 Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins **Suppl 3**: 171–176.

Steel, M. and J. Hein, 2001 Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree. Applied Mathematics Letters **14**: 679–684.

Strope, C. L., S. D. Scott, and E. N. Moriyama, 2006 indel-Seq-Gen: A New Protein Family Simulator Incorporating Domains, Motifs, and Indels. Molecular Biology and Evolution **24**: 640–649.

Suchard, M. A. and B. D. Redelings, 2006 BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics **22**: 2047–2048.

Sksd, Z., B. Knudsen, W. J. J. Anderson, A. Novk, J. Kjems, *et al.*, 2013 Characterising rna secondary structure space using information entropy. BMC Bioinformatics pp. S22–S22.

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

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

Wang, J., P. D. Keightley, and T. Johnson, 2006 MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. BMC Bioinformatics **7**: 292.

Westesson, O., L. Barquist, and I. Holmes, 2012a HandAlign: Bayesian multiple sequence alignment, phylogeny, and ancestral reconstruction. Bioinformatics .

Westesson, O. and I. Holmes, 2012 Developing and applying heterogeneous phylogenetic models with XRate. PLoS ONE **7**: e36898.

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

Westesson, O., G. Lunter, B. Paten, and I. Holmes, 2012b Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. PLoS One **7**: e34572.

Wexler, Y. and D. Geiger, 2008 Variational upper and lower bounds for probabilistic graphical models. J. Comput. Biol. **15**: 721–735.

Yang, Z., 1995 A space-time process model for the evolution of DNA sequences. Genetics **139**: 993–1005.

Yang, Z., R. Nielsen, N. Goldman, and A.-M. Pedersen, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**: 432–449.