A finite-automata model of molecular sequence evolution

Miklós Csűrös*

September 27, 2021

1 Introduction

1.1 Problem statement

A **phylogeny** is a rooted binary tree with nodes numbered $u \in [R] = \{1, 2, \ldots, R\}$. Every node either has two non-null child nodes, or is a terminal node (a **leaf**) with no children. For ease of notation, we assume that the nodes are indexed respecting postfix order, with every child's index being less than the parent's, so that the last one is the root. The tree is identified by its root R and its edges $T \subset [R] \times [R]$ directed from parent to child. The edges in the subtree rooted at a node u are denoted by T_u, including $T_R = T$. The set of leaves is denoted by $\mathcal{L}$, and the leaf set for T_u by $\mathcal{L}_u$; in particular, $\mathcal{L} = \mathcal{L}_R$. For simplicity, start the indices with the leaves respecting the postfix order, so that $\mathcal{L} = [L]$ and every subset $\mathcal{L}_u$ comprises consecutive integers.

Consider the problem of **homolog sequence evolution**: each node u has an associated random variable, its **sequence** $X_u = X_{u,1} \cdots X_{u,\ell}$ of some length $\ell = \xi_u$, which is itself a random variable (and even $\xi_u = 0$ is allowed for an empty sequence). The sequence characters $X_{u,i}$ (or **residues**) are atomic, taking values over some finite alphabet $[A] = \{1, \ldots, A\}$. The joint distribution is determined by the dependencies along the phylogeny:

$$\mathbb{P}\{X_1 = x_1, \ldots, X_R = x_R\} = \mathbb{P}\{X_R = x_R\} \prod_{uv \in T} \underbrace{\mathbb{P}\{X_v = x_v \mid X_u = x_u\}}_{\text{change on edge } uv} \quad (1)$$

The leaf variables are observable, and non-leaf nodes are (hypothetical) ancestors with unobserved sequences. The **sequence inference** problem is that of estimating $\{X_u\}_{u \notin \mathcal{L}}$ for ancestral nodes, knowing the distribution of Eq. (1), and the leaf sequences $\{X_v\}_{v \in \mathcal{L}}$. The **phylogeny inference** problem is that of deducing T given the leaf sequences. The **homology inference** problem is to

*Department of Computer Science and Operations Research, Université de Montréal; C.P. 6128 succursale Centre-Ville, Montréal, Québec H3C 3J7, Canada

partition the leaf sequence set into homologous sequence sets connected along separate phylogenies.

1.2 The TKF91 model

A stochastic model of sequence evolution was proposed by Thorne, Kishino and Felsenstein in 1991. The TKF91 model explains the different lengths of homologous sequences by a continuous-time Markov process that acts along each edge uv for some time $0 \le t_{uv}$ (the **edge length**), mutating, inserting, and deleting residues. The sequence length transition probabilities $\mathbb{P}\{\xi_v = m \mid \xi_u = n\}$ follow from a birth-death process $\{\xi(t) \colon 0 \le t \le t_{uv}\}$ with constant instantaneous rates for deletion $\mu > 0$ and insertion $\lambda \ge 0$, so that $n \to (n-1)$ *death* events arrive with a rate of μn, and $n \to (n+1)$ *birth* events arrive with a rate of $\lambda(n+1)$. The Kolmogorov backward equations for the sequence length $p_n(t) = \mathbb{P}\{\xi(t) = n\}$ are

$$p'_n(t) = \{n > 0\}\lambda n p_{n-1}(t) + \mu(n+1)p_{n+1}(t) - (\lambda(n+1) + \mu n)p_n(t) \qquad (2)$$

with $p'_n(t) = \frac{\partial p_n(t)}{\partial t}$.

Residue substitutions occur by a continuous-time Markov process [2], defined by the $A \times A$ instantaneous rate matrix $\mathbf{Q}_{uv}$. Specifically, let $\{\ell \lozenge_v^u i\}$ denote residue homology between $X_{u,\ell}$ and $X_{v,i}$

$$\mathbb{P}\{X_{v,i} = x' \mid \ell \lozenge_v^u i; X_{u,\ell} = x\} = \mathbf{M}_{uv}(x, x')$$

with the $A \times A$ stochastic matrix of substitution probabilities $\mathbf{M} = \exp(\mathbf{Q}_{uv} t_{uv})$. Inserted residues are picked by a distribution π (which is usually chosen as the stationary distribution with $\pi\mathbf{Q} = 0$):

$$\mathbb{P}\{X_{v,i} = x \mid \emptyset \lozenge_v^u i\} = \pi_{uv}(x),$$

where $\{\emptyset \lozenge_v^u i\}$ denotes the lack of homology at the ancestor. In order to track sequence lengths, we include two meta-characters. The start character $\circ$ is in position 0 for all sequences. The end character $\bullet$ can be placed anywhere, but only once (including the first position if empty).

Residues evolve independently in the TKF91 model. At any time point t, the TKF91 process defines a segmentation by the independent fates of the ancestor residues:

$$\begin{bmatrix} \circ & & & \\ \circ & X_{v,1} & \dots & X_{v,b_1-1} \end{bmatrix} \underbrace{\begin{bmatrix} X_{u,i} & & \\ X_{v,b_i} & \dots & X_{v,b_{i+1}-1} \end{bmatrix}}_{\text{for } i = 1, \dots, n_u} \cdots \begin{bmatrix} \bullet \\ \bullet \end{bmatrix} \qquad (3)$$

with $b_0 = 0 < b_1 \le b_2 \le \cdots \le b_{n_u+1} = n_v + 1$. Block 0 comprises the residues inserted at the start; every other block $i = 1, \dots, n = \xi_u$ consists of the homolog of $X_{u,i}$ if it survives, and inserted residues. The block lengths along X_v define the random variables $\zeta_0(t), \zeta_1(t), \dots, \zeta_n(t)$ with $\zeta_i = b_{i+1} - b_i$

for $0 \leq i \leq n = n_u$. Let $h_n(t)$ denote the distribution of ζ_0, and let $g_n(t)$ denote the common distribution for ζ_i in blocks $0 < i \leq n$. The stochastic differential equation of (2) has a closed-form solution for the block sizes [1]:

$$h_n(t) = \mathbb{P}\{\zeta_0 = n\} = (1-q)q^n \tag{4a}$$

$$g_n(t) = \mathbb{P}\{\zeta_i(t) = n\} = \begin{cases} p & \{n = 0\} \\ (1-p)(1-q)q^{n-1} & \{n > 0\} \end{cases} \quad \{i > 0\} \tag{4b}$$

with the parameters

$$p = \frac{\mu - \mu e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}} \tag{5a}$$

$$q = \frac{\lambda - \lambda e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}} \tag{5b}$$

assuming $\lambda \neq \mu$; or if $\lambda = \mu$,

$$p = q = \frac{\mu t}{1 + \mu t}. \tag{5c}$$

Summing the basic transition probabilities (4) across the blocks gives the sequence length transitions [3]:

$$\mathbb{P}\Big\{\xi(t) = m \;\Big|\; \xi(0) = n\Big\} = \sum_{s=0}^{\min\{n,m\}} \binom{m}{m-s}(1-q)^{1+s}q^{m-s}\binom{n}{s}p^{n-s}(1-p)^s \tag{6}$$

with the parameters p, q defined in Eqs. (5).

The block for any ancestor residue $X_{u,i}$ may be empty (lost ancestor residue, no insertions either). In a non-empty block, the first residue X_{v,b_i} is not necessarily homologous to $X_{u,i}$. The probability of such **nonhomologus replacement** is [1]

$$r = \mathbb{P}\{\emptyset \diamondsuit_v^u b_i \mid b_i > b_{i-1}\} = \frac{1 - p_v - e^{-\mu_{uv}t_{uv}}}{1 - p_v}. \tag{7}$$

If the ancestor residue is replaced, then its replacement is chosen just like the other inserted residues:

$$\mathbb{P}\{X_{v,b_i} = x \mid \emptyset \diamondsuit_v b_i\} = \pi_{uv}(x).$$

1.3 Homology structures for conserved sequences

Homology is a binary relation defined by common ancestry, and our principal goal is to track various homologies created by the random sequence evolution model. The TKF91 mutation model defines the probabilities for the residues

given the homologies between ancestor sequences. A residue in the ancestor sequence at u generates a block of length 0 (lost, and no insertions) on the edge uv with probability p_v, and a block of length $n > 0$ with probability $(1-p_v)(1-q_v)q_v^n$. In the latter case, the first residue may be a nonhomologous replacement with probability r_v. The distinction does not matter for sequence lengths. We adapt thus a notion of positional homology, grouping the compensatory loss-mutation with the substitution model. The first *position* in a non-empty block is homologous to the ancestor residue's position. The positional homology differs from the residue's biological homology because the slot may be filled with an inserted residue that has no relation to the original. The position, however, is conserved because it can be traced back to an ancestral residue.

Since we cannot infer residues without known positional homologs at an ancestral node u (we return to this question later), we let from now on X_u denote the sequence of *conserved* residues: those that have homologous positions in at least one descendant within $\mathcal{L}_u$. If $\tilde{\xi}_u = 0$, then X_u is the empty sequence; and if u is a leaf, then X_u is observed. On every edge $uv \in T$, let Y_v and Z_v denote the random sequence of conserved residues from the ancestor, so that Y_v contains the conserved residues in X_v and Z_v contains the conserved residues at the parent X_u. Define $\tilde{\xi}_u$ as the length of X_u, and let $\tilde{\eta}_v$ be the common length of Y_v and Z_v. In other words, $\tilde{\eta}$ and $\tilde{\eta}$ count only the progenitors of residues at the leaves. Note that the *ancestral* residue counts $\tilde{\xi}, \tilde{\eta}$ refer to ancestors of residues in extant (leaf) sequences, as opposed to the *ancestors'* residue counts ξ, η that include all the ancestors' sequences. In order to track the homologies through the automata, define the **insert homology** relation $\triangle_v$ between positions in Y_v and X_v, the **difference homology** relation $\triangledown_v$ between positions in X_u and Z_v, and the **mutation homology** $\cong_v$ between positions in Z_v and Y_v. The shorthand notations $\{\ell \triangledown_v \emptyset\}$ and $\{\emptyset \triangle_v \ell\}$ mean that there are no $s \triangle_v \ell$, and no $s \triangledown_v \ell$ with any choice of s, i.e., that $X_{u,\ell}$ was lost and that $X_{v,\ell}$ was inserted on edge uv. Lack of homology, or non-homologous replacement between $Z_{v,s}$ and $Y_{v,s}$ is denoted by $s \ncong_v s$. See Figure 1 for an illustration.

As binary relations, $\triangle_v \subset [\tilde{\eta}_v] \times [\tilde{\xi}_v]$ and $\triangledown_v \subset [\tilde{\xi}_u] \times [\tilde{\eta}_v]$ give the mappings from parent to child sequence positions.

Definition 1. *An* **alignment** *between any two sequences of lengths n, m is defined as a binary relation $A \subseteq [n] \times [m]$ on their positions that*

(i) is a one-to-one (partial) mapping of positions between the sequences: for all i, j', j' if $\{i\, A\, j\}$ and $\{i\, A\, j'\}$ then $j = j'$, and for all i', i', j if $\{i\, A\, j\}$ and $\{i'\, A\, j\}$ then $i = i'$; and

(ii) is monotonic: for all $i < i'$ and j, j', if $\{i\, A\, j\}$ and $\{i'\, A\, j'\}$ then $j < j'$;

A **global alignment** *includes the homologies for the sequence endpoints, i.e., position 0, and the first non-occupied position for sequence ends.*

The two homology relations $\triangle$ and $\triangledown$ are alignments, and so are $\cong$ and $\ncong$. It is easy to see that the composition of two alignments is also an alignment, because monotonocity and one-to-one correspondance are preserved. The composition of the indel relations produces the **position homology** $\bowtie_v = \triangledown_v \cdot \triangle_v$

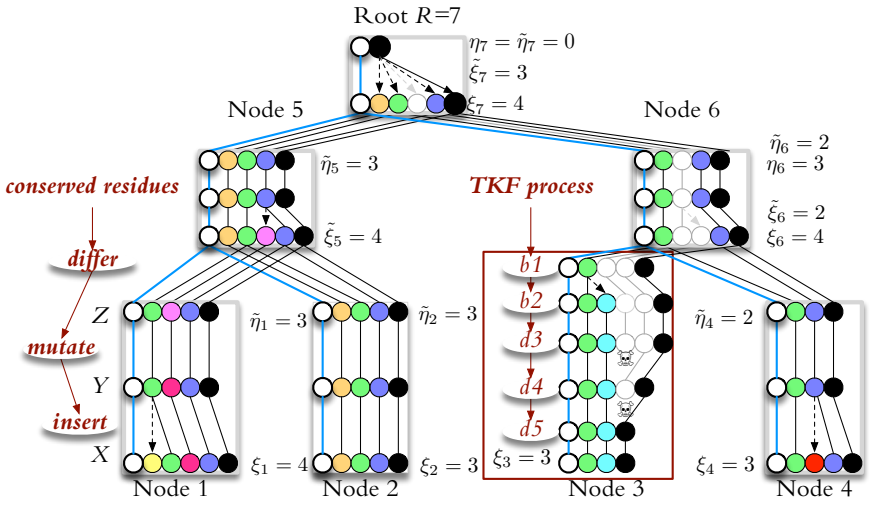

Figure 1: Sequence evolution on a 4-leaf full binary phylogeny. Circles denote residues with lines connecting (positional) homologs among them. The black circles mark sequence ends. The random variables ξ and η count all residues and inherited residues from the parent. On the edge to leaf 3, the TKF91 process is illustrated with 5 events at random time intervals: births of residues (b1 after the immortal link and b2 after the first mortal link), and deaths of residues (d3, d4, d5). The *ancestral* sequence lengths $\tilde{\xi}$ and $\tilde{\eta}$ count only the positions that have homologs in at least one descendant. The full TKF91 history includes the residues that are deleted in all descendants (white circles). An equivalent probability distribution is generated by a three-step manufacturing of conserved residues that have homologies at the leaves. The ancestral sequences X, Y, Z comprise the conserved residues only. The dependancies between the random residues are modeled by finite-state machines that copy their input (one of X, Y, Z) to their output (also one of the ancestral sequences) with "errors." The difference machine forwards the conserved residues to the children, implementing retainment (*match*) and lineage-specific deletion. The mutator machine implements substitution and non-homologous replacement. The insert machine adds lineage-specific conserved residues.

between the ancestral and child sequence positions on the edge:

$$\ell \bowtie_v i \leftrightarrow \exists s\colon \quad \underbrace{(\ell \bigtriangledown_v s) \wedge (s \bigtriangleup_v i)}_{\textit{linked through position } s \textit{ in } Y_v} \quad . \tag{8}$$

Residue homology is transferred through $\cong$:

$$\ell \lozenge_v i \leftrightarrow \exists s\colon \underbrace{(\ell \bigtriangledown_v s) \wedge (s \cong_v s) \wedge (s \bigtriangleup_v i)}_{\textit{linked through } Z_{v,s}}. \tag{9}$$

Further compositions extend the relations from ancestors to all nodes in their subtrees by recursion. In particular, for a node u with a child v, and all leaves $w \in T_u$,

$$\bowtie_w^u = \bowtie_v \cdot \bowtie_w^v \qquad \bigtriangleup_w^u = \bigtriangleup_v \cdot \bowtie_w^v \qquad \lozenge_w^u = \lozenge_v \cdot \lozenge_w^v, \tag{10}$$

At all nodes u, the self-homologies are identity relation

$$\bowtie_u^u = \lozenge_u^u = \{(i,i)\}.$$

2 Results and discussion

2.1 An inductive model for conserved sequence evolution

The TKF91 indel process parameters p, q from Equation (5) determine the joint distribution for position homologies. First we establish the basic parameters for conserved sequence evolution.

Theorem 1. *Let v be any non-root node, and let u be its parent. Let $\tilde{p}_v$ be the probability that a residue at parent u goes extinct towards v, and ϵ_v the probability of extinction in T_u:*

$$\tilde{p}_v = \mathbb{P}\{\forall w \in \mathcal{L}_v\colon \ell \bowtie_w^u \emptyset\} = \mathbb{P}\{\forall w \in \mathcal{L}_v\colon \ell \bigtriangledown_w^u \emptyset\}$$
$$\epsilon_u = \mathbb{P}\{\forall w \in \mathcal{L}_u\colon \ell \bowtie_w^u \emptyset\} = \mathbb{P}\{\forall w \in \mathcal{L}_u\colon \ell \bigtriangledown_w^u \emptyset\}$$

for some position $0 < \ell \le \xi_u$ in the original, *complete ancestor sequence, given some $0 < \xi_u$. (Since residues are lost independently, the position ℓ does not matter.) Then*

$$\epsilon_u = 0 \qquad \textit{if } u \textit{ is a leaf,}$$
$$\epsilon_u = \tilde{p}_v \tilde{p}_w \qquad \textit{if } u \textit{ ancestral with } uv, uw \in T\textit{;}$$
$$\tilde{p}_u = \frac{p_u(1-\epsilon_u) + \epsilon_u(1-q_u)}{1 - q_u\epsilon_u} \qquad \textit{at all nodes } u$$

At all nodes u, define

$$\tilde{q}_u = q_u \frac{1-\epsilon_u}{1-q_u\epsilon_u} \textit{ with } 1-\tilde{q}_u = \frac{1-q_u}{1-q_u\epsilon_u}.$$

After each retained residue, insertions follow a geometric distribution with $\tilde{q}$:

$$\mathbb{P}\{\ell + 1 \triangle_v s + n \mid \ell \triangle_v s\} = (1-\tilde{q}_v)(\tilde{q}_v)^{n-1}.$$

Let v be a non-root node. Then

$$\tilde{r}_u = \mathbb{P}\{\emptyset \cong_v s \mid i \bigtriangledown_v s\} = \frac{1-\tilde{p}_v - e^{-\mu t}(1-\epsilon)}{1-\tilde{p}_v}.$$

over homologies between $Z_{v,s}$ and $Y_{v,s}$, independently. (If $p \neq q$, then $(\mu t) = \ln((1-q_v)/(1-p_v))/(1-q_v/p_v)$, and if $p = q$ then $(\mu t) = p/(1-p)$.)

Proof. Since $\bigtriangledown_w^u = \bigtriangledown_v \cdot \bowtie_w^v$,

$$\tilde{p}_v = \underset{\ell \bigtriangledown_v \emptyset}{p_v} + \underbrace{(1-p_v)\sum_{n=1}^{\infty}(1-q_v)q_v^n(\epsilon_v)^n}_{\{\ell \bigtriangledown_v s\} \wedge \{s \bowtie_w^v \emptyset\}} = p_v + (1-p_v)\frac{1-q_v}{1-q_v\epsilon_v}.$$

By the same argument, the number of inserted and retained residues follows a geometric distribution with parameter $\tilde{q}_v$.

For the nonhomologous replacement parameters, separate the event when the homolog is kept, and does not get lost either, which happens with probability $e^{-\mu t}(1-\epsilon)$. Hence, $(1-\tilde{p}_v)\tilde{r}_v = (1-\tilde{p}_v) - e^{-\mu t}(1-\epsilon)$ giving the result for $\tilde{r}$. □

Since ancestral copies are lost independently, for $0 \le \ell \le n$, $\mathbb{P}\{\tilde{\xi}_u = \ell \mid \xi_u = n\} = \binom{n}{\ell}(1-\epsilon_u)^\ell(\epsilon_u)^{n-\ell}$ and, for all $0 \le s \le t$, $\mathbb{P}\{\tilde{\eta}_u = s \mid \eta_u = t\} = \binom{t}{s}(1-\epsilon_u)^s(\epsilon_u)^{t-s}$ gives the conditional distributions for the ancestral sequence lengths.

Theorem 1 implies that TKF91 evolution for conserved sequences can be decomposed into a three-step procedure of descent with modification, as shown in see Figure 1. The three steps correspond to the different modifications at the residue level: differential loss, substitution or replacement, and insertion. Each step is performed by a two-state transducer that writes a randomly modified version of the input sequence. Every machine has an active state (○) and a finished state (●), one input tape, and one or two output tapes. The machines read and write the sequences one character at a time. They are activated by the start character ○, which is copied to their output, and resets their reading and writing positions to 0. The automata stop after writing the end character ●.

The **difference machine** at an ancestral node u with children $uv, uw \in T$ reads X_u and writes the sequences Z_v, Z_w. It uses a read position counter ℓ and two write position counters s, t. Its transitions generate the inheritance of conserved residues. They can be lost in either but never in both children:

state	$X_{u,\ell}$	probability	$Z_{v,s}$	$Z_{w,t}$	next	transition
○	x	$\frac{(1-\tilde{p}_v)\tilde{p}_w}{1-\tilde{p}_v\tilde{p}_w}$	x		○	*loss in w*
○	x	$\frac{(1-\tilde{p}_v)(1-\tilde{p}_w)}{1-\tilde{p}_v\tilde{p}_w}$	x	x	○	*no loss*
○	x	$\frac{\tilde{p}_v(1-\tilde{p}_w)}{1-\tilde{p}_v\tilde{p}_w}$		x	○	*loss in v*
○	●	1	●	●	●	*stop*

(D)

The **mutator machine** reads Z_v and writes Y_v on an edge $uv \in T$, substituting or replacing residues. It uses the same position counter s for reading and writing. Substitutions occur by the mutation probabilities $\mathbf{M}_{uv}$, and replacements are picked by π_{uv}:

state	$Z_{v,s}$	probability	$Y_{v,s}$	next	transition
$\circ$	x	$(1-\tilde{r}_v)\mathbf{M}_{uv}(x,x')$	x'	$\circ$	*substitute*
$\circ$	x	$\tilde{r}_v \pi_{uv}(x')$	x'	$\circ$	*replace*
$\circ$	$\bullet$	1	$\bullet$	$\bullet$	*stop*

(M)

The **insert machine** copies Y_v on an edge uv and writes X_v, while randomly inserting residues by the distribution π_{uv}. It uses a position counter for reading (s) and another for writing (ℓ).

state	$Y_{v,s}$	probability	$X_{v,\ell}$	next	transition
$\circ$		$\tilde{q}_v \pi_{uv}(x)$	x	$\circ$	*insert*
$\circ$	x	$1-\tilde{q}_v$	x	$\circ$	*copy*
$\circ$	$\bullet$	$1-\tilde{q}_v$	$\bullet$	$\bullet$	*stop*

(I)

The three types of automata generate and destroy the three basic homologies: $\triangledown$, $\cong$ and $\triangle$

In order to follow the generation of random residues and their origins, define the **dependency relation** $\xrightarrow[W]{R}$ between positions in the input sequence R and the output sequence W at each machine:

machine	transition	homology	dependence
D	*loss in v*	$\ell \triangledown^u_v s; \ell \triangledown^u_w \emptyset$	$\ell \xrightarrow[Z_v]{X_u} s$
	no loss, stop	$\ell \triangledown^u_v s; \ell \triangledown^u_w t$	$\ell \xrightarrow[Z_v]{X_u} s; \ell \xrightarrow[Z_w]{X_u} t$
	loss in w	$\ell \triangledown^u_v \emptyset; \ell \triangledown^u_w t$	$\ell \xrightarrow[Z_w]{X_u} t$
M	*substitute*	$s \cong_v s$	$s \xrightarrow[Y_v]{Z_v} s$
	replace	$s \ncong_v s$	$s \xrightarrow[Y_v]{Z_v} s$
I	*insert*	$\emptyset \triangle_v \ell$	$s \xrightarrow[X_v]{Y_v} \ell$
	copy	$s \triangle_v \ell$	$s \xrightarrow[X_v]{Y_v} \ell$

(11)

Note that the insert machine assigns $\xrightarrow[X_v]{Y_v}$ for inserted characters to the *next* input symbol to be copied, which is different from the traditional TKF91 block segmentation of (3), but completely equivalent to it regarding the random sequence distributions. The dependency of the start characters follows the underlying phylogeny

$$0 \xrightarrow[Z_v]{X_u} 0; \quad 0 \xrightarrow[Y_v]{Z_v} 0; \quad 0 \xrightarrow[X_v]{Y_v} 0.$$

The • character signals the sequence end for the machines:

$$(\tilde{\xi}_u = \ell) \equiv (X_{u,n+1} = \bullet) \qquad (\tilde{\eta}_v = s) \equiv (Y_{v,s+1} = \bullet) \qquad (\tilde{\eta}_v = s) \equiv (Z_{v,s+1} = \bullet).$$

The sequence ends also follow the phylogeny, because the automata stop after copying it to their output(s):

$$(\tilde{\xi}_u + 1) \xrightarrow[Z_v]{X_u} (\tilde{\eta}_v + 1); \quad (\tilde{\eta}_v + 1) \xrightarrow[Y_v]{Z_v} (\tilde{\eta}_v + 1); \quad (\tilde{\eta}_v + 1) \xrightarrow[X_v]{Y_v} (\tilde{\xi}_v + 1).$$

The relation $\to$ makes every output character belong to a single input character. Since the automata never step back, the relation is monotonic.

Lemma 2. *All three relations* $\xrightarrow[Z_w]{X_u}$, $\xrightarrow[Y_v]{Z_v}$ *and* $\xrightarrow[X_v]{Y_v}$ *are surjective and monotonic.*

Proof. The relation $\to$ is surjective, since it records the pair of reading and writing position when a character was output. Thus, at a difference machine

$$\forall s \in [0, \tilde{\eta}_v + 1]\, \exists \ell \colon \ell \xrightarrow[Z_v]{X_u} s,$$

where s takes values in the closed interval $[0, \tilde{\eta}_v + 1] = \{0, 1, \ldots, \tilde{\eta}_v + 1\}$. At an insert machine

$$\forall \ell \in [0, \tilde{\xi}_v + 1]\, \exists s \colon s \xrightarrow[X_v]{Y_v} \ell.$$

The relation is monotone because the automata never step back. At the insert machine for node v, for all $0 \le s < s' \le \tilde{\eta}_v + 1$ and for all $0 \le \ell, \ell' \le \tilde{\xi}_v + 1$, if $s \xrightarrow[X_v]{Y_v} \ell$ and $s' \xrightarrow[X_v]{Y_v} \ell'$ then $\ell < \ell'$. At the difference machine for edge $uv \in T$, for all $0 \le \ell < \ell' \le \tilde{\xi}_u + 1$ and for all $0 \le s, s' \le \tilde{\eta}_v + 1$, if $\ell \xrightarrow[Z_v]{X_u} s$ and $\ell' \xrightarrow[Z_v]{X_u}, s'$, then $s < s'$. The claim about $\xrightarrow[Y_v]{Z_v}$ is immediate, since the mutator machine reads and writes exactly one symbol on each transition: $s \xrightarrow[Z_v]{Y_v} s$ for all positions $0 \le s \le \tilde{\eta}_v + 1$. □

2.2 Pair-HMMs and Tree-HMMs as TKF91 ancestral reconstruction

The TKF91 model on every edge uv is reversible in the sense that the conditional distribution from child to parent $X_v \mid X_u$ is also a TKF91 process [1]. The corresponding alignment problem can be represented by a pair hidden Markov model [4], or *pair-HMM*, with three principal states for match, insert, and delete in X_v with respect to the parent's X_u. The resulting stochastic machine's delete and insert state transition probabilities issue directly from the process's p, q parameters, and the distribution of ξ_u determines the termination probabilities. The insert and match states in the usual transformation [4] are not based on positional homology, but rather group the replaced ancestral residue with the insert state and leave only the residue homologies to the match state.

Pair-HMMs are generalized from the simple match-delete-insert machine of parent-child alignment by making the gap extension parameters free for insert-insert and delete-delete transition probabilities. We claim that the such a pair-HMM is equivalent to aligning *sibling* sequences in the TKF91 model. Suppose that we want to align two *leaf* sequences X_v, X_w descending from the same parent X_u. The sibling homology $\bowtie$ is produced by the composition of the different processes: the residues $X_{u,i}$ and $X_{v,j}$ are homologous if they have a common ancestor:

$$j \bowtie k = \exists i\colon \underbrace{(i \bowtie_v j) \wedge (i \triangle_w k)}_{\textit{linked through } X_{u,i}}\}.$$

The resulting alignment is commonly shown as a sequence of columns enumerating the homologies in a monotone order: $\{(0,0),(2,1),(4,3)\}$, for example. The missing homologs show up as gaps in the alignment's representation:

$$\begin{bmatrix} 0 & 1 & 2 & & 3 & 4 \\ 0 & & 1 & 2 & & 3 \end{bmatrix}$$

The lack of homology, or *indel* is not a physical object: here $3 \bowtie \emptyset$ and $\emptyset \bowtie 2$, and there is no order between them. We can annotate the alignment columns, however, by the origin of the residues with respect to the ancestral sequence X_u in five ways: M for match (if $i \bowtie j$), L_w for loss in w (if $k \bowtie_v i$ but $k \triangledown_w \emptyset$), L_v for deletion in v (if $k \bowtie_w j$ but $k \triangledown_v \emptyset$), I_w for insertion in w (if $\emptyset \triangle_w j$), or I_v for insertion in v (if $\emptyset \triangle_v i$). A pair-HMM has just three principal states, fusing the different types of indel differences: M (match and start), $\mathsf{D} = \mathsf{L}_w + \mathsf{I}_v$ ("deletion" in X_w compared to X_v) and $\mathsf{I} = \mathsf{L}_v + \mathsf{I}_w$ ("insertion" into X_w compared to X_v).

The transition probabilities are regularized by incorporating a geometric prior for the parent's sequence length distribution $\mathbb{P}\{\tilde{\xi}_u = \ell\} = (1-q)q^\ell$ that is

used for transitions to the terminal state E:

$$\begin{aligned}
\tau(\mathsf{M},\mathsf{E}) &= (1-q)\frac{(1-q_v)(1-q_w)}{1-q_vq_w} \\
\tau(\mathsf{M},\mathsf{D}) &= \frac{1-q_w}{1-q_vq_w}\left(q_v + (1-q_v)q\frac{(1-p_v)p_w}{1-p_vp_w}\right) && = \delta_{\mathsf{D}} \\
\tau(\mathsf{M},\mathsf{I}) &= \frac{1-q_v}{1-q_vq_w}\left(q_w + (1-q_w)q\frac{p_v(1-p_w)}{1-p_vp_w}\right) && = \delta_{\mathsf{I}} \\
\tau(\mathsf{M},\mathsf{M}) &= \frac{(1-q_v)(1-q_w)}{1-q_vq_w}q\frac{(1-p_v)(1-p_w)}{1-p_vp_w} \\
\tau(\mathsf{D},\mathsf{E}) &= (1-q_v)(1-q) \\
\tau(\mathsf{D},\mathsf{D}) &= q_v + (1-q_v)q\frac{(1-p_v)p_w}{1-p_vp_w} && = \epsilon_{\mathsf{D}} \\
\tau(\mathsf{D},\mathsf{M}) &= (1-q_v)q\frac{(1-p_v)(1-p_w)}{1-p_vp_w} \\
\tau(\mathsf{D},\mathsf{I}) &= (1-q_v)q\frac{p_v(1-p_w)}{1-p_vp_w} \\
\tau(\mathsf{I},\mathsf{E}) &= (1-q_w)(1-q) \\
\tau(\mathsf{I},\mathsf{I}) &= q_w + (1-q_w)q\frac{p_v(1-p_w)}{1-p_vp_w} && = \epsilon_{\mathsf{I}} \\
\tau(\mathsf{I},\mathsf{M}) &= (1-q_w)q\frac{(1-p_v)(1-p_w)}{1-p_vp_w} \\
\tau(\mathsf{I},\mathsf{D}) &= (1-q_w)q\frac{(1-p_v)p_w}{1-p_vp_w}
\end{aligned}$$

In particular, the TKF91 model provides the parameters for a pair-HMM with gap opening (δ) and gap extension (ϵ) probabilities, separately for insertion and deletion. And vice versa: the formulas can be reversed to find equivalent q, q_u, q_v, p_u, p_v given five independent parameters in the HMM, which are $\delta_{\mathsf{D}}, \delta_{\mathsf{I}}$ (gap openings), $\epsilon_{\mathsf{D}}, \epsilon_{\mathsf{I}}$ (gap extensions), and either the termination probability $\tau(\mathsf{M},\mathsf{E})$ or the match extension probability $\tau(\mathsf{M},\mathsf{M})$.

Pair-HMMs are further generalized to multiple sequence alignment [5, 6] by placing a general pair-HMM on each edge. Given that a pair-HMM corresponds to a TKF91-alignment of sibling sequences, imagining a pair-HMM on each edge amounts to introducing a "mirror" node that reflects the sibling. Specifically, the alignment of X_u and X_v by pair-HMM is equivalent to sibling alignment between X_u and X_v, imagining a common ancestor w', which, by reversibility of $TKF91$, is in reality the same as the sibling w in T, with conserved sequence $X_{w'} = X_w$. The pair-HMM alignment on the other edge uw also imagines a common ancestor v'. Continuing towards the root with X_u will create a mirror of u for the parent, and so on. Figure 2 illustrates how tree-HMM alignment is looking for the TKF91 ancestral sequence alignment in reality. The equivalence between the two problems highlight why progressive multiple alignment is such a hard algorithmic problem. One has to insert conserved

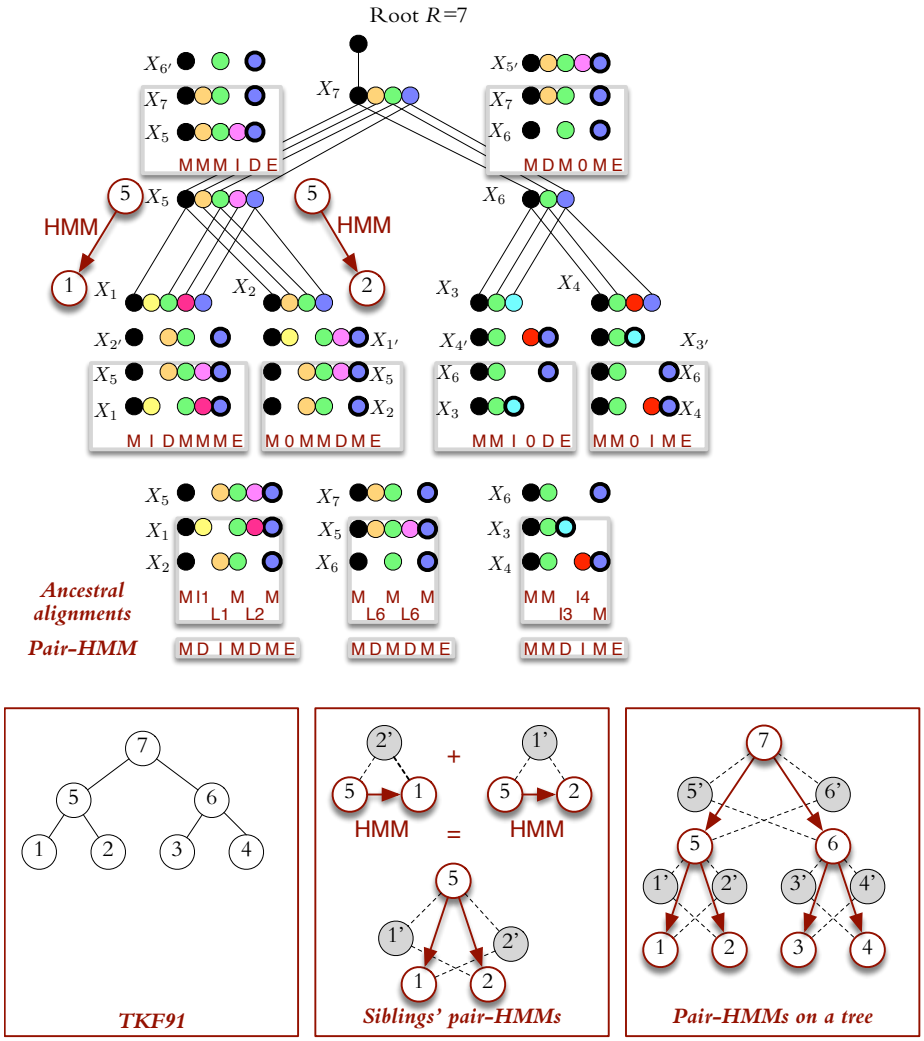

Figure 2: Pair-HMM alignments (columns $\mathsf{M}, \mathsf{D}, \mathsf{I}, \mathsf{E}$) and ancestral TKF91 alignments (columns $\mathsf{M}, \mathsf{L}_v, \mathsf{L}_w, \mathsf{I}_v, \mathsf{I}_w$). Placing a pair-HMM on an edge is equivalent to ancestral TKF91 alignment with the the sibling acting as the ancestor to the parent and the other child. Multiple alignment based on pair-HMMs on the phylogeny is the equivalent of looking for the original TKF91 reconstruction with interspersed mirror nodes (1', 2',..., 7', grey circles) reflecting the sibling sequences. The example illustrates the fundamental algorithmic difficulty in progressive alignment: one has to insert new columns when fusing two multiple alignments from the children (columns $\mathsf{0}$, occurring four times in this small example constructed without prejudice).

residues when fusing partial alignments from the children. Specifically, if one keeps track of the residue homologies only ($\Diamond$), then at an ancestral node u with children $uv, uw \in T$, residue homology may be inherited asymmetrically when position homologies are synchronized: $\{\ell \bowtie_u i\}$ and $\{\ell \bowtie_v j\}$ may be accompanied with $\{\ell \Diamond_v i\}$ and $\{\ell \Diamond_w \emptyset\}$. If some $\{\ell-1 \Diamond_w \emptyset\}$, then the a gap is extended, but if $\{\ell - 1 \Diamond_w k\}$, then we open a deletion gap after position k. The necessary tracking the indel state from the children in order to apply the appropriate gap penalties is *the* main challenge of statistical multiple alignment [7, 8]. We posit that more complicated pair-HMM models introducing multiple indel states [9] are likely to correspond to the TKF91 ancestral reconstruction, as well, with more complicated mirroring of cousin nodes and other friends and relations. Simply put, pair-HMMs model the homologies from the bottom (with gaps). The corresponding algorithms infer the homologies *progressively*, extending the gaps while moving up to the root. In contrast, ancestral sequence alignment infers the the homologies viewed from the top (conserved sequence without gaps), Conserved sequences are recovered *regressively*, eliminating residues and merging sibling sequences toward the root.

Fun: a snow machine

We can design an automaton that makes ancestor sequences, including the lost residues. Introduce a volatile character (snowflake ❄) for the non-conserved residues. The **snow machine** is between the insert machine and the copy machine. Transitions:

state	read	next	write	probability
○	x	○	x	$1-\epsilon_v$
○		○	❄	ϵ_v

The mutator machines lets them pass:

state	read	next	write	probability
○	❄	○	❄	1

The difference machine implements loss by $\tilde{p}_v, \tilde{p}_w$, independently. It has a silent transition that produces no output, for symmetric loss.

state	read	next	write1	write2	probability
○	❄	○	❄		$(1-\tilde{p}_v)\tilde{p}_w$
○	❄	○	❄	❄	$(1-\tilde{p}_v)(1-\tilde{p}_w)$
○	❄	○		❄	$\tilde{p}_v(1-\tilde{p}_w)$
○	❄	○			$\tilde{p}_v\tilde{p}_w = \epsilon_u$

And finally, the insert machine melts them

state	read	next	write	probability
○	❄	○		1

References

[1] J. L. Thorne, H. Kishino, J. Felsenstein, An evolutionary model for maximum likelihod alignment of DNA sequences, Journal of Molecular Evolution 33 (1991) 114–124.

[2] P. Liò, N. Goldman, Models of molecular evolution and phylogeny, Genome Research 8 (1998) 1233–1244.

[3] M. Csűrös, Gain-loss-duplication models on a phylogeny: exact algorithms for computing the likelihood and its gradient (2021).
URL `https://arxiv.org/abs/2107.11440`

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

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

[6] G. Mitchison, R. Durbin, Tree-based maximal likelihood substitution matrices and hidden Markov models, Journal of Molecular Evolution 4 (1995) 1139–1151.

[7] A. Löytynoja, N. Goldman, A model of evolution and structure of multiple sequence alignment, Philosophical Transactions of the Royal Society of London, Series B 363 (2008) 3913–3919.

[8] G. Lunter, A. J. Drummond, I. Miklós, J. Hein, Statistical alignment: Recent progress, new applications, and challenges, in: R. Nielsen (Ed.), Statistical Methods in Molecular Evolution, Springer-Verlag, Heidelberg, 2005, Ch. 14.

[9] R. K. Bradley, A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes, Fast statistical alignment, PLoS Computational Biology 5 (5) (2007) e1000392.