SUPPLEMENTARY INFORMATION

This supplementary section describes how to calculate the sufficient statistics $\hat{S}$, $\hat{Z}$, $\hat{D}$ and $\hat{I}$. defined in section 3.3.

Usually $\hat{S}$ is apparent from the observed data (we typically observe the initial state of the system, or can treat one of our observations as the initial state using the "pulley principle" [26]). We thus need to calculate $\hat{Z}$, $\hat{D}$ and $\hat{I}$. These expectations contain both *observed* and *unobserved* contributions. The observed contributions are the indels that constitute the alignment; these correspond to the parsimonious edit distance, i.e. the minimum set of indels required for the observed alignment. The unobserved contributions are from mutually cancelling indel events, creating transient links that are not observed in the ancestor or descendant sequences.

We calculate the expectations by breaking the alignment (and the evolving sequence) into independent *zones*, one for each transition between states in the Pair HMM of Figure 1, with an observed link (or link-pair) at the start of each zone. This link is called the *marker link*; the first zone's marker link is the immortal link.

There are four types of zone; each transition in the pair HMM corresponds to a unique type of zone. The four zone types, illustrated in Figure 7, are (i) a zone whose marker link was deleted, with no surviving descendants to the right (transitions DM,DD,DE), symbolically $\circ$; (ii) a zone whose marker link was not deleted, with no surviving descendants to the right (transitions SM,SD,SE,MM,MD,ME,IM,ID,IE), symbolically $\bullet$; (iii) a zone whose marker link was deleted, with at least one surviving descendant to the right (transition DI), symbolically $\circ\bullet$; (iv) a zone whose marker link was not deleted, with at least one surviving descendant to the right (transitions SI,MI,II), symbolically $\bullet\bullet$. Examples of zone histories are shown in Figure 9.

In calculating the expectations $\hat{I}$, $\hat{D}$ and $\hat{Z}$ for each zone, we can assume without loss of generality that the marker link was an ancestral link. This may seem counter-intuitive: if the marker link was inserted, it will be younger than the ancestral links, and so will have accumulated fewer unobserved descendants. However, in such a case, the inserted marker link is itself descended from an ancestral link (directly or indirectly). Unobserved transient indels from this ancestral link (and the lineage connecting it to the marker link) can accumulate in the zone containing the marker link, exactly as if the marker link were itself ancestral. Thus the birth-time of a marker link is, in fact, independent of the expected number of transients in its zone.

5.1 Finding the expected number of insertions in a zone

To calculate the expected number of insertions for all types of zone in the TKF91 model, we need

- $\hat{I}_\circ(T)$, the expected number of insertions in the zone whose ancestral marker link died in the interval $[0,T]$ without any surviving child links to the right;
- $\hat{I}_\bullet(T)$, the expected number of insertions in the zone whose ancestral marker link survived the interval $[0,T]$ without any surviving child links to the right;
- $\hat{I}_{\circ\bullet}(T)$, the expected number of insertions in the zone whose ancestral marker link died in the interval $[0,T]$ with at least one surviving child link to the right; and
- $\hat{I}_{\bullet\bullet}(T)$, the expected number of insertions in the zone whose ancestral marker link survived the interval $[0,T]$ with at least one surviving child link to the right.

We first enumerate three possible fates for a link over a time interval: (i) it survives directly (i.e. the link is not deleted); (ii) it survives indirectly (i.e. the link is deleted, but has surviving descendants); (iii) it is extinct. We can also classify insertion events in this way: consider an insertion during the infinitesimal time interval $[t, t+dt]$. At the later time T, this inserted link may be directly surviving, indirectly surviving or extinct. The probability of each type of event during the infinitesimal time interval $[t, t+dt]$ is $\lambda_v(T-t)dt$ for directly surviving insertions, $\lambda_w(T-t)dt$ for indirectly surviving insertions and $\lambda_x(T-t)dt$ for extinct insertions, where

$$\begin{aligned}\lambda_v(t) &= \lambda\alpha(t)\\ \lambda_w(t) &= \lambda(1-\alpha(t))\gamma(t)\\ \lambda_x(t) &= \lambda(1-\alpha(t))(1-\gamma(t))\end{aligned}$$

Let $v(t|T)$ be the probability that, at time t, a link has not been deleted and has not had any insertions that will survive (indirectly or directly) to time T. This is given by the integral equation

$$\log v(t|T) = -\int_0^t (\mu + \lambda_v(T-t') + \lambda_w(T-t'))dt' \quad (8)$$

Let $a(t|T)$ be the posterior probability that an ancestral link exists at time t, given that by time T it has been deleted with no surviving descendants (defined in (13)). Given a single insertion at time t that is extinct by time T, the expected number of insertion events descended from this original insertion (including the original insertion itself) is $1+\hat{I}_\circ(T-t)$. This reasoning gives us integral equations for $\hat{I}_\circ(T)$ and $\hat{I}_\bullet(T)$

$$\hat{I}_\circ(T) = \int_{t=0}^{T} a(t|T)\lambda_x(T-t)(1+\hat{I}_\circ(T-t))dt \quad (9)$$

$$\hat{I}_\bullet(T) = \int_{t=0}^{T} \lambda_x(T-t)(1+\hat{I}_\circ(T-t))dt \quad (10)$$

For $\hat{I}_{\circ\bullet}(T)$ and $\hat{I}_{\bullet\bullet}(T)$, we need to condition on the last (directly or indirectly) surviving insertion from the ancestral link. We call this event the *propagation event*, written $\mathcal{P}$ for short. Examples of direct and indirect propagation events are shown in Figure 9.

Suppose that the propagation event happens at time t. The insertion events that we must count include: (i) $\mathcal{P}$ itself (totalling 1); (ii) any extinct insertions at the ancestral link that happened after $\mathcal{P}$ (totalling $\hat{I}_{\circ}(T-t)$ if the ancestral link is deleted, and $\hat{I}_{\bullet}(T-t)$ if it is not deleted); and (iii) any further extinct insertions to the left of the next observed link (totalling $\hat{I}_{\circ\bullet}(T-t)$ if $\mathcal{P}$ is an indirectly surviving insertion, and 0 if $\mathcal{P}$ is a directly surviving insertion).

We first consider $\hat{I}_{\circ\bullet}(T)$. In this scenario, the ancestral link is deleted, but has at least one surviving insertion. The probability that a link is not deleted during the interval $[0,t]$ and becomes extinct over the interval $[t,T]$ is $\alpha(t)q_0(T-t)$, while the probability that the link indirectly survives the interval $[0,T]$ is $(1-\alpha(T))\gamma(T)$; write the ratio of these two probabilities as $b(t|T)$. The posterior probability that $\mathcal{P}$ was a directly surviving insertion at time t is $b(t|T)\lambda_v(T-t)dt$, while the posterior probability that $\mathcal{P}$ was an indirectly surviving insertion at time t is $b(t|T)\lambda_w(T-t)dt$. This gives us the following integral equation for $\hat{I}_{\circ\bullet}(T)$

$$\begin{aligned}\hat{I}_{\circ\bullet}(T) &= \int_{t=0}^{T} b(t|T)\Big(\lambda_v(T-t)(1+\hat{I}_{\circ}(T-t)) \\ &\quad +\lambda_w(T-t)(1+\hat{I}_{\circ}(T-t)+\hat{I}_{\circ\bullet}(T-t))\Big)\,dt\end{aligned} \tag{11}$$

Now consider $\hat{I}_{\bullet\bullet}(T)$. In this scenario, the ancestral link is not deleted, and has at least one surviving insertion. The probability that a link has no surviving insertions over the interval $[t,T]$ is $1-\beta(T-t)$, while the probability that the link has surviving insertions over the interval $[0,T]$ is $\beta(T)$; write the ratio of these two probabilities as $c(t|T)$. The posterior probability that $\mathcal{P}$ was a directly surviving insertion at time t is $c(t|T)\lambda_v(T-t)dt$, while the posterior probability that $\mathcal{P}$ was an indirectly surviving insertion at time t is $c(t|T)\lambda_w(T-t)dt$. This gives us the following integral equation for $\hat{I}_{\bullet\bullet}(T)$

$$\begin{aligned}\hat{I}_{\bullet\bullet}(T) &= \int_{t=0}^{T} c(t|T)\Big(\lambda_v(T-t)(1+\hat{I}_{\bullet}(T-t)) \\ &\quad +\lambda_w(T-t)(1+\hat{I}_{\bullet}(T-t)+\hat{I}_{\circ\bullet}(T-t))\Big)\,dt\end{aligned} \tag{12}$$

The formulae for $a(t|T)$, $b(t|T)$ and $c(t|T)$ are

$$\begin{aligned}a(t|T) &= \frac{v(t|T)(1-\alpha(T-t))(1-\gamma(T-t))}{(1-\alpha(T))(1-\gamma(T))} \\ b(t|T) &= \frac{\alpha(t)(1-\alpha(T-t))(1-\gamma(T-t))}{(1-\alpha(T))\gamma(T)} \\ c(t|T) &= \frac{1-\beta(T-t)}{\beta(T)}\end{aligned} \tag{13}$$

It can be seen that all of the above calculations of zone fates require conditioning on the time of an insertion event (which, in the latter two zone types, can be a direct or indirect propagation event), then marginalising the time of this event by integration. This procedure is illustrated in Figure 8.

5.2 Finding the expected site time of a zone

To calculate the expected time accrued by all links in a zone, we note that this is given by $\hat{Z} = \langle \int_{t=0}^{T} l(t)dt \rangle$ where $l(t)$ is the sequence length at time t (a stochastic process). The integral $Z = \int_{t=0}^{T} l(t)dt$ is called the *site time*, since it is effectively the average number of sites multiplied by the total time.

To get the site time for all types of zone, we need

- $\hat{Z}_{\circ}(T)$, the expected site time for a zone whose ancestral marker link died in the interval $[0,T]$ without any surviving child links to the right;
- $\hat{Z}_{\bullet}(T)$, the expected site time for a zone whose ancestral marker link survived the interval $[0,T]$ without any surviving child links to the right;
- $\hat{Z}_{\circ\bullet}(T)$, the expected site time for a zone whose ancestral marker link died in the interval $[0,T]$ with at least one surviving child link to the right; and
- $\hat{Z}_{\bullet\bullet}(T)$, the expected site time for a zone whose ancestral marker link survived the interval $[0,T]$ with at least one surviving child link to the right.

We find $\hat{Z}_{\circ}, \hat{Z}_{\bullet}, \hat{Z}_{\circ\bullet}$ and $\hat{Z}_{\bullet\bullet}$ via the following integral equations

$$\hat{Z}_\circ(T) = \int_{t=0}^{T} a(t|T)(1+\lambda_x(T-t)\hat{Z}_\circ(T-t))dt \quad (14)$$

$$\hat{Z}_\bullet(T) = \int_{t=0}^{T} (1+\lambda_x(T-t)\hat{Z}_\circ(T-t))dt \quad (15)$$

$$\hat{Z}_{\circ\bullet}(T) = \int_{t=0}^{T} b(t|T)\Big(\lambda_v(T-t)\hat{Z}_\circ(T-t) + \lambda_w(T-t)(\hat{Z}_\circ(T-t)+\hat{Z}_{\circ\bullet}(T-t))\Big)\,dt \quad (16)$$

$$\hat{Z}_{\bullet\bullet}(T) = \int_{t=0}^{T} c(t|T)\Big(\lambda_v(T-t)(t+\hat{Z}_\bullet(T-t)) + \lambda_w(T-t)(t+\hat{Z}_\bullet(T-t)+\hat{Z}_{\circ\bullet}(T-t))\Big)\,dt \quad (17)$$

Again, these calculations are conditioned on the time of an insertion event that is then integrated out, as is illustrated in Figure 8.

5.3 Algebraic solution

This linear birth-death process with immigration is extensively studied, and a system of orthogonal polynomial eigenfunctions is known [27, 28, 17]. To make this work practical and tidy, a priority should be to find algebraic solutions to the integral equations (8)-(17).

5.4 Summing the zone expectations

Let $\hat{N}_{XY}$ be the expected number of transitions from state X to state Y in the Pair HMM, either conditioned on a trusted alignment or calculated for unaligned sequences using the Forward-Backward algorithm. Define the following sums

$$\begin{aligned}
\hat{N}_\circ &= \hat{N}_{DM}+\hat{N}_{DD}+\hat{N}_{DE}\\
\hat{N}_\bullet &= \hat{N}_{SM}+\hat{N}_{SD}+\hat{N}_{SE}+\hat{N}_{MM}+\hat{N}_{MD}+\hat{N}_{ME}\\
&\quad +\hat{N}_{IM}+\hat{N}_{ID}+\hat{N}_{IE}\\
\hat{N}_{\circ\bullet} &= \hat{N}_{DI}\\
\hat{N}_{\bullet\bullet} &= \hat{N}_{SI}+\hat{N}_{MI}+\hat{N}_{II}
\end{aligned}$$

The expectations $\hat{S}$, $\hat{I}$, $\hat{D}$ and $\hat{Z}$, defined in Section 3.3, are then given by

$$\begin{aligned}
\hat{S} &= \hat{N}_\circ+\hat{N}_\bullet-1\\
\hat{I} &= \hat{N}_\circ\hat{I}_\circ(T)+\hat{N}_\bullet\hat{I}_\bullet(T)+\hat{N}_{\circ\bullet}\hat{I}_{\circ\bullet}(T)+\hat{N}_{\bullet\bullet}\hat{I}_{\bullet\bullet}(T)\\
\hat{D} &= \hat{I}+\hat{N}_\circ-\hat{N}_{\bullet\bullet}\\
\hat{Z} &= \hat{N}_\circ\hat{Z}_\circ(T)+\hat{N}_\bullet\hat{Z}_\bullet(T)+\hat{N}_{\circ\bullet}\hat{Z}_{\circ\bullet}(T)+\hat{N}_{\bullet\bullet}\hat{Z}_{\bullet\bullet}(T)
\end{aligned}$$

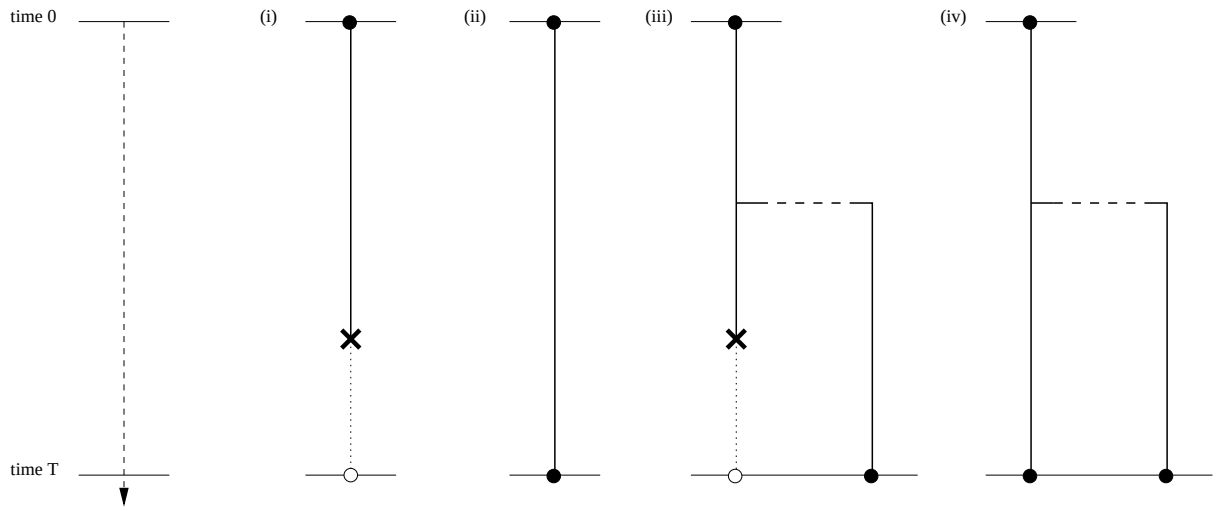

Fig. 7. The four different zone types for calculating indel and length expectations. (i) Link is deleted, with no surviving insertions (type ○). (ii) Link survives, with no surviving insertions (type ●). (iii) Link is deleted, with at least one surviving insertion (type ○●). (iv) Link survives, with at least one surviving insertion (type ●●).

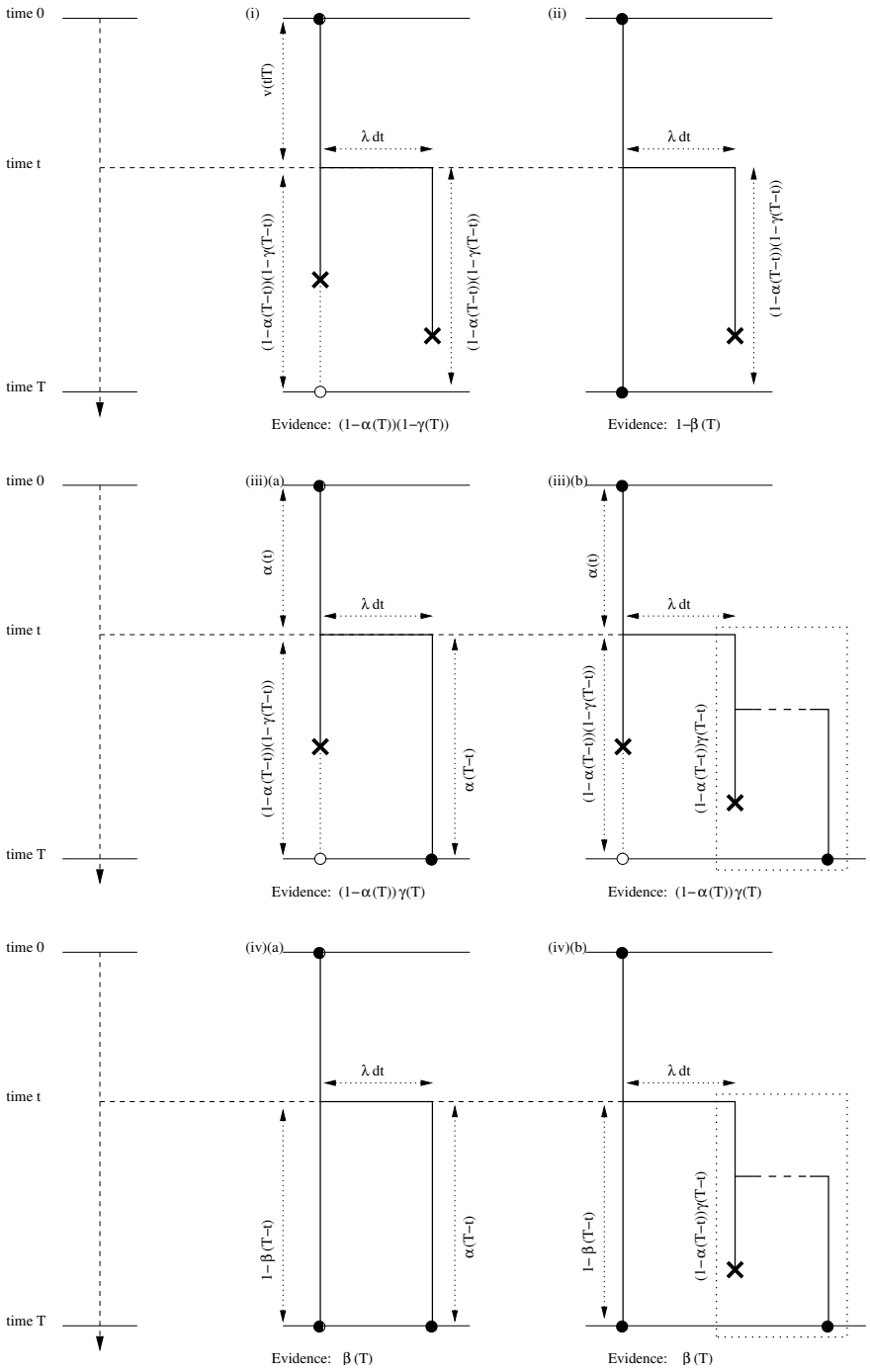

Fig. 8. Calculation of expectations for each type of zone requires conditioning on the time of an insertion event and then integrating out this time over the full time-range $[0, T]$. Case (i): Link is deleted, with no surviving insertions (Figure 7(i); type ○). An unobserved insertion at time t looks like Figure 7(i) for the remaining time $T - t$. Case (ii): Link survives, with no surviving insertions (Figure 7(ii); type ●). An unobserved insertion at time t looks like Figure 7(i) for the remaining time $T - t$. Cases (iii)(a) and (iii)(b): Link dies, with at least one surviving insertion (Figure 7(iii); type ○●). There are two possibilities: *direct propagation* or *indirect propagation*. Case (iii)(a): if a direct propagation event occurs at time t, then the ancestral link looks like Figure 7(i) for the remaining time $T - t$. Case (iii)(b): if an indirect propagation event occurs at time t, the ancestral link looks like Figure 7(i) for the remaining time $T - t$, while the (unobserved) descendant link looks like Figure 7(iii). Cases (iv)(a) and (iv)(b): Link survives, with at least one surviving insertion (Figure 7(iv); type ●●). Again, there are two possibilities: *direct propagation* or *indirect propagation*. Case (iv)(a): if a direct propagation event occurs at time t, then the ancestral link looks like Figure 7(ii) for the remaining time $T - t$. Case (iv)(b): if an indirect propagation event occurs at time t, the ancestral link looks like Figure 7(ii) for the remaining time $T - t$, while the (unobserved) descendant link looks like Figure 7(iii).

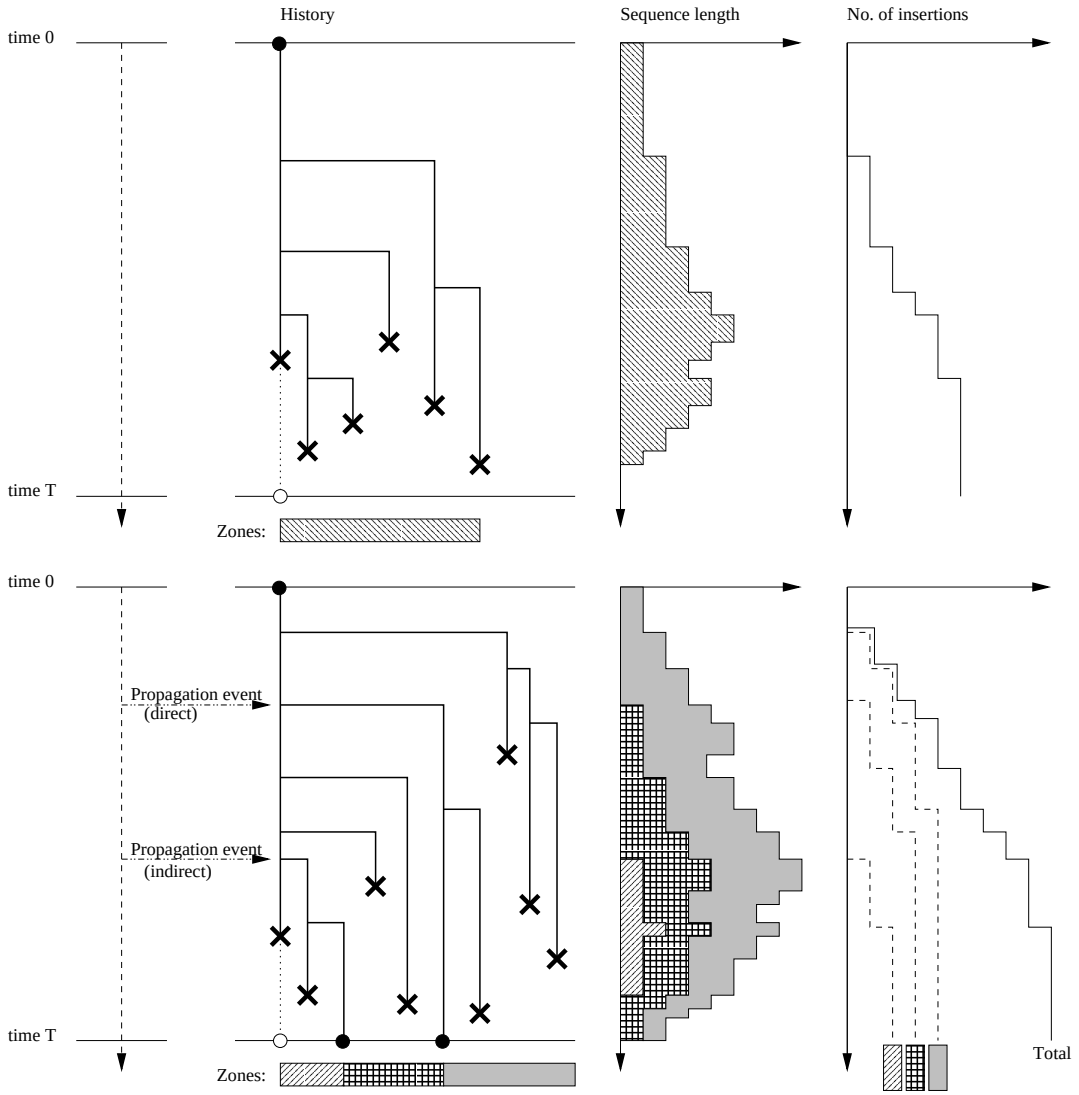

Fig. 9. Example zone histories. Top example: zone type ○ (Figure 7(i) and Figure 8(i)). Ancestor dies, with no orphaned survivors. Five insertions occur during the interval. Site time is shown as shaded area. Bottom example: one zone of type ○● (Figure 7(iii)) with indirect propagation (Figure 8(iii)(b)), followed by one zone of type ●● (Figure 7(iv)) with direct propagation (Figure 8(iv)(a)) and finally one zone of type ● (Figure 7(ii) and Figure 8(ii)). Ancestor survives, with two surviving insertions, creating three zones. Insertions and site times for each zone shown using different shadings.