Supplementary Materials for the Blueprint of
the "<u>A</u>lignment <u>N</u>eighborhood <u>Ex</u>plorer" (ANEX)
(tentatively named),
by Kiyoshi Ezawa

(Finished on June 4th, 2017; CC4 statement added on August 14th, 2020)

Table of Contents

* Figure S4 was added (on June 4 (Sun), 2017).

Supplementary appendix

SA-1. Practical factorability of indel configuration multiplication factor into contributions from isolated gap masses

For this purpose, it would be more convenient to use the ancestral-state-based equations derived in section SM-4 of (Ezawa, unpublished A 2016b), rather than to use the indel-history-based equations derived in subsection 4.2 of (Ezawa, Graur and Landan, Part I). Thus, we begin by recalling the key results of the former.

Provided that the indel model satisfies the conditions (i), (ii) and (iii) (see R6 and R7 of Ezawa, unpublished A 2016b), the probability, $P\left[\alpha[s_1, s_2,..., s_{N^X}] \mid T\right]$, that a given MSA ($\alpha[s_1, s_2,..., s_{N^X}]$) result from an indel process along a given tree (T) is factorized as in Eq.(SM-4.20) of (Ezawa, unpublished A 2016b):

$$P\left[\alpha[s_1, s_2,..., s_{N^X}] \mid T\right] = P_0\left[s_0^{Root} \mid T\right] \prod_{K=1}^{K_{max}} \widetilde{M}_P\left[\alpha[s_1, s_2,..., s_{N^X}]; s_0^{Root}; C_K \mid T\right].$$ --- Eq. (SA-1.1)

Here, s_0^{Root} denotes a "reference" root sequence state that is consistent with the MSA, and $P_0\left[s_0^{Root} \mid T\right]$ is the probability that the sequence state remained s_0^{Root} all across the tree (see Eq.(SM-4.21) of *ibid.* for details). And $\widetilde{M}_P\left[\alpha[s_1, s_2,..., s_{N^X}]; s_0^{Root}; C_K \mid T\right]$ is the multiplication factor contributed from a local region (C_K). It's specific expression is given by Eq.(SM-4.22) of *ibid.*:

$$\widetilde{M}_P\left[\alpha[s_1, s_2,..., s_{N^X}]; s_0^{Root}; C_K \mid T\right] = \sum_{\substack{\{s(n)-s_0^{Root}\}_{N^{IN}}[C_K] \\ \in \Delta_\Sigma\left[C_K; s_0^{Root}; \alpha[s_1,s_2,...,s_{N^X}]; \{n\in N^{IN}(T)\}; T\right]}} M_P\left[\alpha[s_1, s_2,..., s_{N^X}]; \{s(n)\}_{N^{IN}}; s_0^{Root}; C_K \mid T\right].$$ --- Eq.(SA-1.2)

Here, $\Delta_\Sigma\left[C_K; s_0^{Root}; \alpha[s_1, s_2,..., s_{N^X}]; \left\{n \in N^{IN}(T)\right\}; T\right]$ is the space of deviations of ancestral state spaces ($\left\{s(n)\right\}_{N^{IN}}$) from s_0^{Root} within C_K (denoted as $\left\{s(n)-s_0^{Root}\right\}_{N^{IN}}[C_K]$ in the range of summation). And the summand is given by Eq.(SM-4.18) of *ibid.*:

$$M_P\left[\alpha[s_1, s_2,..., s_{N^X}]; \left\{s(n)\right\}_{N^{IN}}; s_0^{Root}; C_K \mid T\right]$$
$$\equiv M_P\left[\alpha[s_1, s_2,..., s_{N^X}]; \left\{s(n)\right\}_{N^{IN}}; C_K \mid T\right] \mu_P\left[s\left(n^{Root}\right), s_0^{Root}, n^{Root}; C_K\right]$$ --- Eq.(SA-1.3)
$$\times \exp\left(-\sum_{b\in\{b\}_T} \int_{t\left(n^A(b)\right)}^{t\left(n^D(b)\right)} d\tau \, \delta R_X^{ID}(s^A(b), \, s_0^{Root}, \, \tau)[C_K]\right)$$

with Eq.(SM-4.13) of *ibid.*:

$$\mathbf{M}_P\left[\alpha[s_1, s_2, ..., s_{N^X}]; \{s(n)\}_{N^{IN}}; C_K \mid T\right]$$

$$\equiv \prod_{b \in \{b\}_T} \left\{ \prod_{\gamma_{\kappa_b}(b) \subseteq C_K} \tilde{\mu}_P\left[\left(\tilde{\Lambda}^{ID}\left[\gamma_{\kappa_b}(b); \alpha(s^A(b), s^D(b))\right], b\right) \mid (s^A(b), n^A(b))\right] \right\} \quad \text{--- Eq.(SA-1.4)}$$

Here, $\tilde{\mu}_P\left[\left(\tilde{\Lambda}^{ID}\left[\gamma_{\kappa_b}(b); \alpha(s^A(b), s^D(b))\right], b\right) \mid (s^A(b), n^A(b))\right]$ is the multiplication factor

contributed from the portion of the PWA ($\alpha(s^A(b), s^D(b))$) between the ancestral state (

$s^A(b)$) and the descendant state ($s^D(b)$) along the branch (b), confined in a region (

$\gamma_{\kappa_b}(b)$, which is within C_K). (See Eq.(SM-4.11), Eq.(R6.8). and Eq.(SM-2.14) of *ibid*.)

$\mu_P\left[s\left(n^{Root}\right), s_0^{Root}, n^{Root}; C_K\right]$ is the multiplicative difference of the probability of state

$s\left(n^{Root}\right)$ at the root (n^{Root}) from that of s_0^{Root} originated from C_K (see Eq.(SM-4.16) of

ibid.). And $\delta R_X^{ID}(s^A(b), s_0^{Root}, \tau)[C_K]$ is the increment of the exit rate of the ancestral state

($s^A(b)$), compared to that of the reference state (s_0^{Root}), coming from the state difference

confined in C_K (see Eq.(SM-4.14) of *ibid*.).

  For the current purpose, it would be more convenient to reorganize Eq.(SA-1.3)

accompanied by Eq.(SA-1.4) as follows:

$$\mathbf{M}_P\left[\alpha[s_1, s_2, ..., s_{N^X}]; \{s(n)\}_{N^{IN}}; s_0^{Root}; C_K \mid T\right]$$

$$= \mu_P\left[s\left(n^{Root}\right), s_0^{Root}, n^{Root}; C_K\right] \prod_{b \in \{b\}_T} \tilde{\mu}_P\left[\alpha(s^A(b), s^D(b)); C_K \mid (s^A(b), n^A(b))\right], \quad \text{--- Eq.}$$

(SA-1.3')

with

$$\tilde{\mu}_P\left[\alpha(s^A(b), s^D(b)); C_K \mid (s^A(b), n^A(b))\right]$$

$$\equiv \exp\left(-\int_{t\left(n^A(b)\right)}^{t\left(n^D(b)\right)} d\tau \, \delta R_X^{ID}(s^A(b), s_0^{Root}, \tau)[C_K]\right) \quad \text{--- Eq.(SA-1.4')}$$

$$\times \prod_{\gamma_{\kappa_b}(b) \subseteq C_K} \tilde{\mu}_P\left[\left(\tilde{\Lambda}^{ID}\left[\gamma_{\kappa_b}(b); \alpha(s^A(b), s^D(b))\right], b\right) \mid (s^A(b), n^A(b))\right]$$

  Now, we are ready to discuss the problem at hand, that is, calculating the

multiplication factor, Eq.(SA-1.2), when the local MSA contains at least two *isolated* gap-

masses, as in Figure 3C or Figure S3A. (We will use the latter for illustration.) In such a case,

some of the ancestral nodes ("R" and "a1" in the current example) have *virtually* fixed gap

states (Figure S3B). [NOTE: More precisely, as noted also in Appendix A2, the states at these

nodes in some non-parsimonious indel histories have extra sites that are destined to vanish

completely from the extant sequences (as in Figure S4). Each of such indel histories, however,

requires at least two additional indels that are coordinated exquisitely. Thus, in general, their

contributions are negligible.] Such ancestral nodes could be used to "partition" the set of all

internal nodes, $\mathbf{N}^{IN}$, into three sub-sets:

$$\mathbf{N}^{IN} = \mathbf{N}_0^{IN} \cup \mathbf{N}_1^{IN} \cup \mathbf{N}_2^{IN} . \quad \text{--- Eq.(SA-1.5)}$$

Here, $\mathbf{N}_0^{IN}$ is the subset consisting of the "partitioning" nodes with virtually fixed ancestral states (like the red-shaded nodes in Figure S3C); each of $\mathbf{N}_1^{IN}$ and $\mathbf{N}_2^{IN}$ is the subset consisting of internal nodes involved in one of the isolated gap-masses (like the nodes enclosed in dashed boxes in Figure S3C). Likewise, we can also decompose each set of all ancestral states, $\{s(n)\}_{\mathbf{N}^{IN}}$, as:

$$\{s(n)\}_{\mathbf{N}^{IN}} = \{s(n)\}_{\mathbf{N}_0^{IN}} \cup \{s(n)\}_{\mathbf{N}_1^{IN}} \cup \{s(n)\}_{\mathbf{N}_2^{IN}} . \quad \text{--- Eq.(SA-1.6)}$$

Here, $\{s(n)\}_{\mathbf{N}_0^{IN}}$ is virtually fixed. And $\{s(n)\}_{\mathbf{N}_1^{IN}}$ and $\{s(n)\}_{\mathbf{N}_2^{IN}}$ are virtually independent of each other. Hence, $\Delta_\Sigma\left[C_K ; s_0^{Root} ; \alpha[s_1, s_2, ..., s_{N^X}] ; \{n \in \mathbf{N}^{IN}(T)\} ; T\right]$ can be approximately expressed as a direct product:

$$\Delta_\Sigma\left[C_K ; s_0^{Root} ; \alpha[s_1, s_2, ..., s_{N^X}] ; \{n \in \mathbf{N}^{IN}(T)\} ; T\right] \approx \left(\Delta_\Sigma\right)_0 \times \left(\Delta_\Sigma\right)_1 \times \left(\Delta_\Sigma\right)_2 . \quad \text{--- Eq.(SA-1.7)}$$

Here, $\left(\Delta_\Sigma\right)_k$ ($k = 0, 1$ or 2) is a shorthand notation of <mark>a "component" of</mark> $\Delta_\Sigma\left[C_K ; s_0^{Root} ; \alpha[s_1, s_2, ..., s_{N^X}] ; \{n \in \mathbf{N}_k^{IN}(T)\} ; T\right]$, and consists of the sets of local-MSA-consistent ancestral states at nodes in $\mathbf{N}_k^{IN}$ (more precisely, their differences from s_0^{Root}). (Note that $\left(\Delta_\Sigma\right)_0 = \left\{\left\{s(n) - s_0^{Root}\right\}_{\mathbf{N}_0^{IN}}\right\}$.) Another essential element is the decomposition of the set of branches:

$$\{b\}_T = \{b\}_T^0 \cup \{b\}_T^1 \cup \{b\}_T^2 . \quad \text{--- Eq.(SA-1.8)}$$

Here, $\{b\}_T^k$ ($k = 1$ or 2) is the set of branches that are directly connected with at least one node in $\mathbf{N}_k^{IN}$ (like the branches in the dashed boxes in Figure S3C); $\{b\}_T^0$ is the set of remaining branches (*i.e.*, those connected *solely* with nodes in $\mathbf{N}_0^{IN}$).

Now, substituting Eqs.(SA-1.7,8) into Eq.(SA-1.2) accompanied by Eqs.(SA-1.3',4'), we get an approximate equation:

$$\widetilde{\mathbf{M}}_P\left[\alpha[s_1, s_2, ..., s_{N^X}] ; s_0^{Root} ; C_K \mid T\right] \approx \left(\widetilde{\mathbf{M}}_P\right)_0 \left(\widetilde{\mathbf{M}}_P\right)_1 \left(\widetilde{\mathbf{M}}_P\right)_2 . \quad \text{--- Eq.(SA-1.9)}$$

Here,

$$\left(\widetilde{\mathbf{M}}_P\right)_k = \sum_{\{s(n) - s_0^{Root}\}_{\mathbf{N}_k^{IN}}[C_K] \in (\Delta_\Sigma)_k} \left[\begin{array}{l} \Theta\left(\mu_P\left[s\left(n^{Root}\right), s_0^{Root}, n^{Root} ; C_K\right] ; n^{Root} ; \mathbf{N}_k^{IN}\right) \\ \times \prod_{b \in \{b\}_T^k} \tilde{\mu}_P\left[\alpha(s^A(b), s^D(b)) ; C_K \mid (s^A(b), n^A(b))\right] \end{array} \right]$$

4

--- Eq.(SA-1.10)

is the collection of contributions from the sub-histories of indels along the branches in $\{b\}_T^k$ ($k = 0, 1$ or 2). In this equation, it is tacitly agreed that the sequence states at nodes in $\{s(n)\}_{N_0^{IN}}$ are fixed as mentioned above. And we also defined the following function:

$$\theta(x; n; N') = \begin{cases} x & if \quad n \in N', \\ 1 & otherwise \end{cases} \quad \text{--- Eq.(SA-1.11)}$$

When there are $K (\geq 2)$ isolated gap-masses, Eq.(SA-1.9) can be generalized as:

$$\tilde{\mathbf{M}}_P \left[\alpha[s_1, s_2, ..., s_{N^X}]; s_0^{Root}; C_K \mid T \right] \approx \left(\tilde{\mathbf{M}}_P \right)_0 \prod_{k=1}^{K} \left(\tilde{\mathbf{M}}_P \right)_k, \quad \text{--- Eq.(SA-1.9')}$$

where $\left(\tilde{\mathbf{M}}_P \right)_k$ is the total contribution from the k-th gap mass (like Eq.(SA-1.10)), and

$\left(\tilde{\mathbf{M}}_P \right)_0$ is the contribution from the remaining, "fixed," part of the indel histories.

SA-2. Proof of necessary and sufficient condition, Eq.(SA-2.1), for independent effects of overlapping isolated shifts

As described in Appendix A5, the effects of the shifts of two isolated gap-blocks can be regarded as nearly independent of each other **if the approximate equation, Eq.(A5-1) holds**:

$$\log\left(P\left((c', c''', c'') \right) \right) + \log\left(P\left((-, c'', -) \right) \right) \overset{?}{\approx} \log\left(P\left((c', c''', -) \right) \right) + \log\left(P\left((-, c''', c'') \right) \right).$$

--- Eq.(SA-2.1)

Here, c' and c'' are the portions of MSA columns flanking the two gap-blocks in question, and c''' is the remaining portion of the column. (c', c''', c'') denotes the MSA column created by aligning these portions. $(c', c''', -)$ denotes the column made from (c', c''', c'') by replacing the residues in c'' with gaps. And other similar symbols can be interpreted accordingly. In this section, we will prove that Eq.(A5-1) (or Eq.(SA-2.1)) is indeed sufficient for the independent effects of the shifts of isolated gap-blocks. For this purpose, we here introduce a symbol, $\langle ... \rangle$, which is a shorthand notation of $\log\left(P\left((...) \right) \right)$. (For example, $\langle c', c''', c'' \rangle$ denotes $\log\left(P\left((c', c''', c'') \right) \right)$.)

We first consider the situation illustrated in Figure 12 B. We here use the following symbols for the portions of the columns:

c' denotes the shifted portion corresponding to the "CC" occupying sequences 1 and 2;

5

c'' denotes the shifted portion corresponding to the "C" occupying sequence 5;

c_1''' denotes the unmoved portion corresponding to the "TTT" occupying sequences 3, 4 and 5;

c_2''' denotes the unmoved portion corresponding to the "- - CC" occupying sequences 1, 2, 3 and 4; and

c_3''' denotes the unmoved portion corresponding to the "CT" occupying sequences 3 and 4.

Because we are considering a substitution model that allows each site to evolve independently of the other sites, the logarithm of the residue component of the MSA probability is given by a summation of column-wise log-likelihoods over all MSA columns. Therefore, when comparing the log-component of the MSA after the double-shift with that before the double-shift, it is sufficient to compare the contributions from the columns affected by the shifts, i.e., the columns shaded in blue and/or red. Then, the difference of the log-component after the move from that before the move is expressed as:

$$\left(\Delta \log P\right)\big|_{12B} \equiv \left\{\langle c', c_1''' \rangle + \langle c_2''', c'' \rangle + \langle -, c_3''', - \rangle\right\} - \left\{\langle -, c_1''' \rangle + \langle c_2''', - \rangle + \langle c', c_3''', c'' \rangle\right\}.$$

--- Eq.(SA-2.2)

On the other hand, when only each of the two shifted portions (i.e., c' and c'') is shifted from the MSA before the double-shift, we have two possible differences:

$$\left(\Delta \log P\right)\big|_{12B(c'\,alone)} \equiv \left\{\langle c', c_1''' \rangle + \langle -, c_3''', c'' \rangle\right\} - \left\{\langle -, c_1''' \rangle + \langle c', c_3''', c'' \rangle\right\}, \quad \text{--- Eq.(SA-2.3a)}$$

$$\left(\Delta \log P\right)\big|_{12B(c''\,alone)} \equiv \left\{\langle c_2''', c'' \rangle + \langle c', c_3''', - \rangle\right\} - \left\{\langle c_2''', - \rangle + \langle c', c_3''', c'' \rangle\right\}. \quad \text{--- Eq.(SA-2.3b)}$$

The independence of the effects of the shifts means that $\left(\Delta \log P\right)\big|_{12B}$ is equal to the summation of $\left(\Delta \log P\right)\big|_{12B(c'\,alone)}$ and $\left(\Delta \log P\right)\big|_{12B(c''\,alone)}$. Thus, using Eq.(SA-2.2) and Eqs.(SA-2.3a,b), we must have:

$$\begin{aligned} 0 &= \left(\Delta \log P\right)\big|_{12B} - \left\{\left(\Delta \log P\right)\big|_{12B(c'\,alone)} + \left(\Delta \log P\right)\big|_{12B(c''\,alone)}\right\} \\ &= \langle -, c_3''', - \rangle + \langle c', c_3''', c'' \rangle - \langle -, c_3''', c'' \rangle - \langle c', c_3''', - \rangle \end{aligned} . \quad \text{--- Eq.(SA-2.4)}$$

This is nothing other than Eq.(SA-2.1) (*i.e.*, Eq.(A5-1)).

Next we consider the situation illustrated in Figure 12 C. To represent the portions of the columns, we will use the same symbols as above, but associate them with slightly different meanings:

c' denotes the shifted portion corresponding to the "CC" occupying sequences 1 and 2;

c'' denotes the shifted portion corresponding to the "T" occupying sequence 5;

c_1''' denotes the unmoved portion corresponding to the "TT" occupying sequences 3 and 4;

c_2''' denotes the unmoved portion corresponding to the "CT" occupying sequences 3 and 4.

In this case, the difference of the log-component after the double-shift from that before the double-shift is expressed as:

$$\left(\Delta \log P \right)\big|_{12C} \equiv \left\{ \left\langle c', c_1''', - \right\rangle + \left\langle -, c_3''', c'' \right\rangle \right\} - \left\{ \left\langle -, c_1''', c'' \right\rangle + \left\langle c', c_3''', - \right\rangle \right\} . \quad \text{--- Eq.(SA-2.5)}$$

The effects of the single-shifts are calculated as:

$$\left(\Delta \log P \right)\big|_{12C\,(c'\,alone)} \equiv \left\{ \left\langle c', c_1''', c'' \right\rangle + \left\langle -, c_3''', - \right\rangle \right\} - \left\{ \left\langle -, c_1''', c'' \right\rangle + \left\langle c', c_3''', - \right\rangle \right\} , \quad \text{--- Eq.}$$

(SA-2.6a)

$$\left(\Delta \log P \right)\big|_{12C\,(c''\,alone)} \equiv \left\{ \left\langle -, c_1''', - \right\rangle + \left\langle c', c_3''', c'' \right\rangle \right\} - \left\{ \left\langle -, c_1''', c'' \right\rangle + \left\langle c', c_3''', - \right\rangle \right\} . \quad \text{--- Eq.}$$

(SA-2.6b)

Using Eq.(SA-2.5) and Eqs.(SA-2.6a,b), the independence of the effects of the shifts can be expressed as:

$$\begin{aligned}
0 \; &= \left(\Delta \log P \right)\big|_{12C} - \left\{ \left(\Delta \log P \right)\big|_{12C\,(c'\,alone)} + \left(\Delta \log P \right)\big|_{12C\,(c''\,alone)} \right\} \\
&= \left\{ \left\langle c', c_1''', - \right\rangle + \left\langle -, c_1''', c'' \right\rangle - \left\langle c', c_1''', c'' \right\rangle - \left\langle -, c_1''', - \right\rangle \right\} \\
&\quad + \left\{ \left\langle -, c_3''', c'' \right\rangle + \left\langle c', c_3''', - \right\rangle - \left\langle -, c_3''', - \right\rangle - \left\langle c', c_3''', c'' \right\rangle \right\}
\end{aligned} \quad \text{--- Eq.(SA-2.7)}$$

The expression in each pair of braces on the right hand side of Eq.(SA-2.7) vanishes if Eq.(A5-1) (*i.e.*, Eq.(SA-2.1)) holds. Thus, in conjunction with Eq.(SA-2.4), Eq.(SA-2.7) demonstrates that the effects of two overlapping isolated shifts are independent of each other if Eq.(A5-1) holds. (And the arguments in this section also indicates that Eq.(A5-1) is actually the necessary and sufficient condition for the independence of the effects of overlapping shifts.)

SA-3. Non-factorability of residue configuration probability change into contributions from shifts of *isolated* gap-blocks

As described in Appendix A5 (and proved in Supplementary appendix SA-2), we found that the effects of the shifts of two isolated gap-blocks can be regarded as nearly independent of each other **if the approximate equation, Eq.(A5-1) (or Eq.(SA-2.1)) holds**:

$$\log\left(P\left(\left(c', c''', c''\right)\right)\right) + \log\left(P\left(\left(-, c'', -\right)\right)\right) \stackrel{?}{\approx} \log\left(P\left(\left(c', c''', -\right)\right)\right) + \log\left(P\left(\left(-, c''', c''\right)\right)\right) \quad .$$

--- Eq.(SA-3.1)

Here, c' and c'' are the portions of MSA columns flanking the two gap-blocks in question, and c''' is the remaining portion of the column. $\left(c', c''', c''\right)$ denotes the MSA column created by aligning these portions. $\left(c', c''', -\right)$ denotes the column made from $\left(c', c''', c''\right)$ by replacing the residues in c'' with gaps. And other similar symbols can be interpreted accordingly.

In order to examine under what condition(s) Eq.(SA-3.1) (*i.e.*, Eq.(A5-1)) holds, we consider a situation illustrated in Figure 13. Let $P\left(\omega, \omega'; c'''\right)$ be the joint probability that we have ω and ω' at the upper-ends of branches delimiting c'' and c', respectively, and also have the residue configuration in c'''. Under the situation in Figure 13, it can be expressed as follows:

$$P\left(\omega, \omega'; c'''\right) = P\left(\omega; n^R\right) P\left(\omega \mapsto \omega'; b_1''\right) \tilde{P}_L\left(\omega; b_2''', c\right) \tilde{P}_L\left(\omega'; b_3''', c\right) \quad . \text{--- Eq.(SA-3.2)}$$

Using this probability, we can calculate the probability $P\left(\left(c', c''', c''\right)\right)$ as:

$$P\left(\left(c', c''', c''\right)\right) = \sum_{\omega \in \Omega} \sum_{\omega' \in \Omega} \left\{ P\left(\omega, \omega'; c'''\right) \tilde{P}_L\left(\omega; b'', c\right) \tilde{P}_L\left(\omega'; b', c\right) \right\} \quad . \text{--- Eq.(SA-}$$

3.3a)

If, for example, c'' is occupied solely with gaps, we have $\tilde{P}_L\left(\omega; b'', c\right) = 1$ for every ω. This and similar facts yield the following:

$$P\left(\left(-, c''', c''\right)\right) = \sum_{\omega \in \Omega} \left\{ P\left(\omega, \cdot; c'''\right) \tilde{P}_L\left(\omega; b'', c\right) \right\} \quad , \quad \text{--- Eq.(SA-3.3b)}$$

$$P\left(\left(c', c''', -\right)\right) = \sum_{\omega' \in \Omega} \left\{ P\left(\cdot, \omega'; c'''\right) \tilde{P}_L\left(\omega'; b', c\right) \right\} \quad , \quad \text{--- Eq.(SA-3.3c)}$$

$$P\left(\left(-, c''', -\right)\right) = P\left(\cdot, \cdot; c'''\right) \quad . \quad \text{---Eq.(SA-3.3d)}$$

Here, we introduced the notations:

$$P\left(\omega, \cdot; c'''\right) \equiv \sum_{\omega' \in \Omega} P\left(\omega, \omega'; c'''\right) \, , \, P\left(\cdot, \omega'; c'''\right) \equiv \sum_{\omega \in \Omega} P\left(\omega, \omega'; c'''\right) \, , \text{and}$$
$$P\left(\cdot, \cdot; c'''\right) \equiv \sum_{\omega \in \Omega} \sum_{\omega' \in \Omega} P\left(\omega, \omega'; c'''\right) \, . \quad \text{---Eq.(SA-3.4a,b,c)}$$

Thus, **for Eq.(SA-3.1) (*i.e.*, Eq.(A5-1)) to hold as an *exact* equation for any residue configurations of** c' **and** c''**,** (and thus for any probability vectors $\left\{ \tilde{P}_L\left(\omega'; b', c\right) \right\}_{\omega' \in \Omega}$ and $\left\{ \tilde{P}_L\left(\omega; b'', c\right) \right\}_{\omega \in \Omega}$,) **the probabilities** $\left\{ P\left(\omega, \omega'; c'''\right) \right\}_{\omega, \omega' \in \Omega}$ **must satisfy the following equation**:

$$P\left(\omega, \omega'; c'''\right) = P\left(\omega, \cdot; c'''\right) P\left(\cdot, \omega'; c'''\right) / P\left(\cdot, \cdot; c'''\right) \quad for \quad {}^\forall\left(\omega, \omega'\right) \in \Omega^2 \quad . \text{--- Eq.(SA-}$$

3.5)

In other words, their ω-dependence and ω'-dependence must decouple. In the situation at hand, where $\left\{ P(\omega, \omega'; c''') \right\}_{\omega, \omega' \in \Omega}$ is given by Eq.(SA-3.2), the condition approximately holds when branch b'''_1 is so long that $P(\omega \mapsto \omega'; b'''_1)$ nearly saturates. Or, unless Eq.(SA-3.5) holds even approximately, Eq.(SA-3.1) (*i.e.*, Eq.(A5-1)) could still be satisfied provided that either b' or b'' is quite long (and thus either $\tilde{P}_L(\omega'; b', c)$ or $\tilde{P}_L(\omega; b'', c)$ is nearly saturated).

      If, however, none of the branches b', b'' and b''' is quite long, it is unlikely that Eq.(SA-3.1) (*i.e.*, Eq.(A5-1)) should hold even approximately. **Thus, in general, it would be safe *not* to assume the approximate equation, Eq.(A5-1).** This restriction makes the calculation of the effects of two *"isolated"* shifts a bit cumbersome when they happen to affect the same column(s).

Supplementary Figures (with legends)

(A) Unlikely "vertical split" (of gap-blocks)
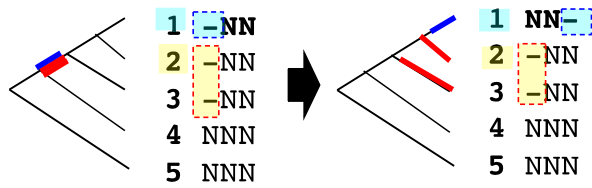
(B) Unlikely "vertical split" (of sequence-blocks)

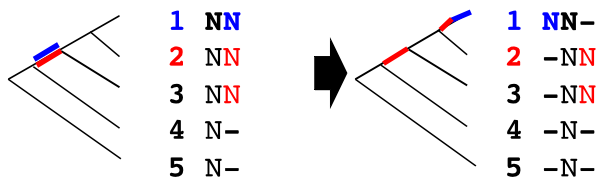**Figure S1. Unlikely instances of "vertical-split"s.**

A. An unlikely "vertical-split" (of gap-blocks). This is somewhat similar to a CII, though not exactly the one. **B.** An unlikely "vertical-split" (of sequence-blocks). This is actually an "ex-nihilo" aligned with a gap-block. These "vertical-splits" are unlikely, because they require one more indel than those in panels E&F of Figure 4.

(A) "Incomplete purge"

(B) "Merge + split" (same)

(C) "Incomplete merge" (complementary)
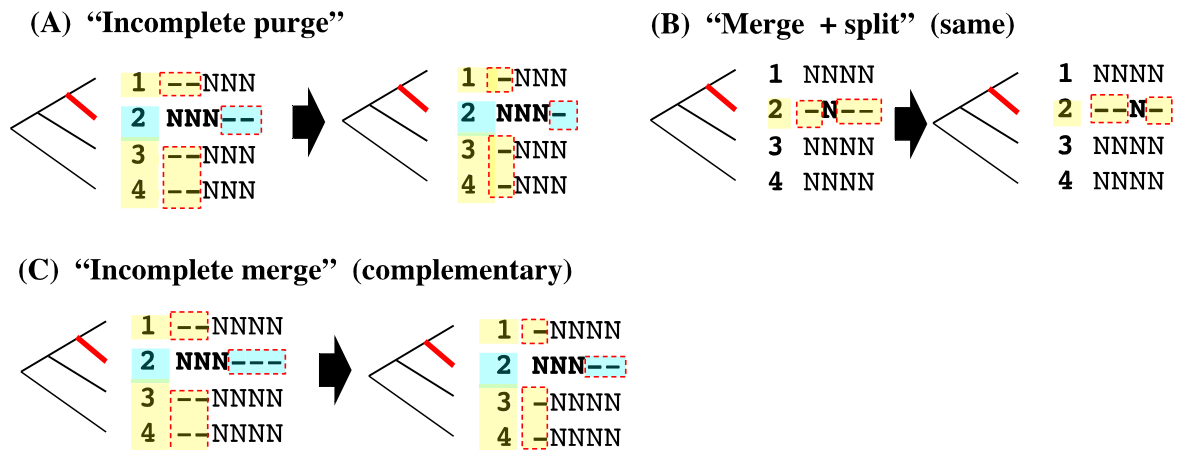
Figure S2. Secondary moves of gap-blocks.

A. An "incomplete purge." **B.** A "merge + split" of two gap-blocks affecting the same set of sequences. (It is likely to have actually resulted from the "shift" of the bridging gapless segment.) **C.** An "incomplete merge" of two gap-blocks affecting the complementary sets of sequences. (It could also be interpreted as a "merge + split" (complementary).)

It is relatively less likely that these moves (or their reverses) actually resulted in the reconstructed MSAs, because they do not change the number of indels and thus only slightly change the indel component of the likelihood.

(A) Tree **MSA**

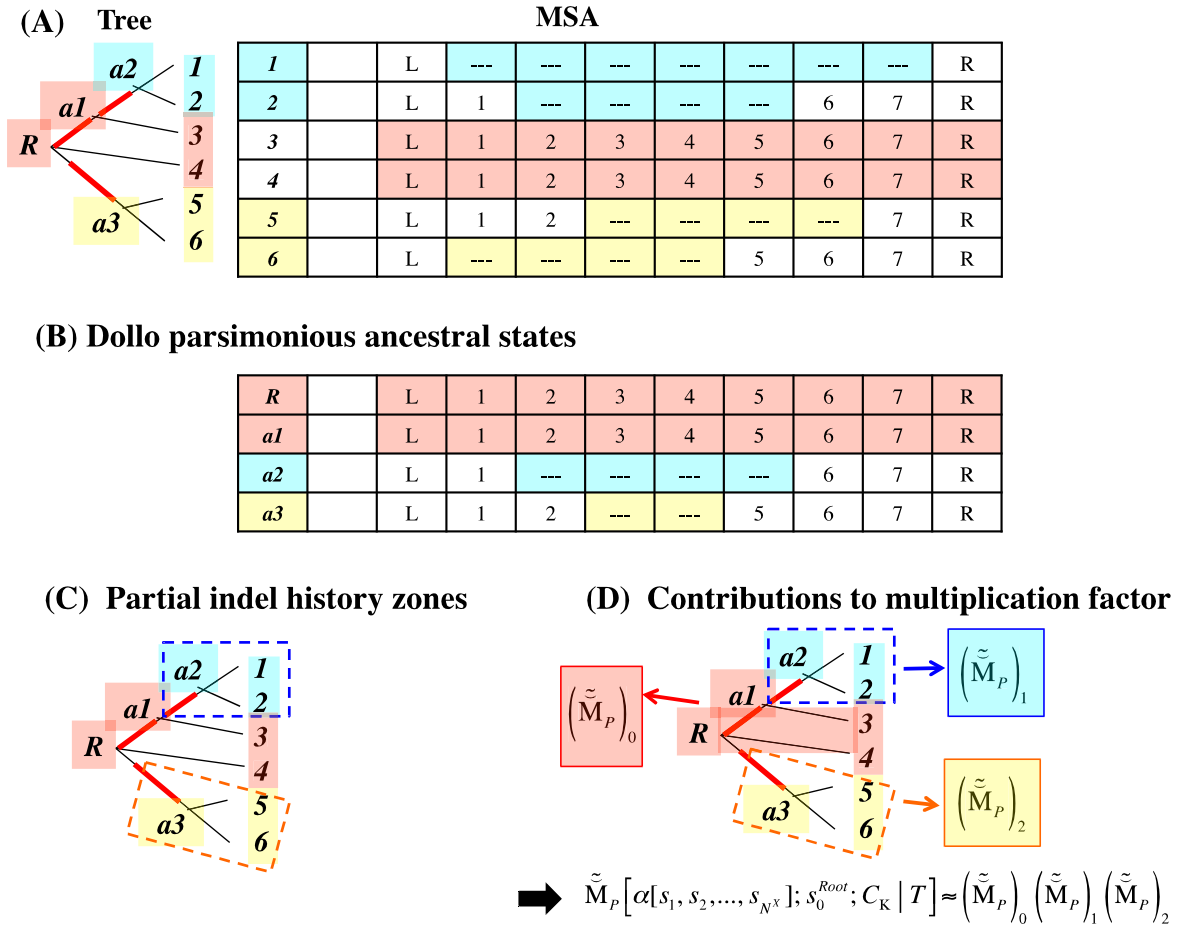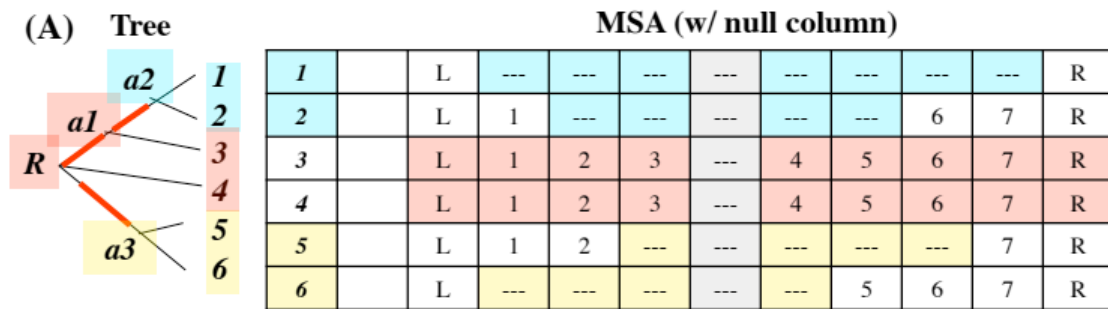|										
1		L	---	---	---	---	---	---	---	R
2		L	1	---	---	---	---	6	7	R
3		L	1	2	3	4	5	6	7	R
4		L	1	2	3	4	5	6	7	R
5		L	1	2	---	---	---	---	7	R
6		L	---	---	---	---	5	6	7	R

(B) Dollo parsimonious ancestral states

R		L	1	2	3	4	5	6	7	R
a1		L	1	2	3	4	5	6	7	R
a2		L	1	---	---	---	---	6	7	R
a3		L	1	2	---	---	5	6	7	R

(C) Partial indel history zones   **(D) Contributions to multiplication factor**

$$\tilde{\tilde{M}}_P\left[\alpha[s_1, s_2,..., s_{N^X}]; s_0^{Root}; C_K \mid T\right] \approx \left(\tilde{\tilde{M}}_P\right)_0 \left(\tilde{\tilde{M}}_P\right)_1 \left(\tilde{\tilde{M}}_P\right)_2$$

Figure S3. Indel histories that can yield horizontally overlapping gap-blocks.
A. An example local MSA. It is more complex than Figure 3A, though the two figures are the same in essence. Each cell in the MSA is assigned either an ancestry index (L, R, or arabic numeral) or a gap (a lump of triple-dashes). The italicized Arabic numeral on the left indicates the sequence on the external node labeled with the same numeral in the tree. The masses of gaps isolated from each other are shaded in cyan and yellow. **B.** The Dollo parsimonious states obtained from the MSA in **panel A**. Any ancestral states consistent with the MSA must keep the occupied sites in the Dollo parsimonious states. Thus, the states "*R*" and "*a1*" here (red-shaded) indicate that the two isolated gap masses must have been created independently from each other. **C.** As a result, the indel history (or the ancestral states) yielding the cyan gap mass (enclosed in blue dashed box) is virtually independent of the history (or the ancestral states) yielding the yellow gap mass (enclosed in orange dashed box), separated by the red-shaded sequence states. (See, *e.g.*, Eq.(SA-1.7) of Supplementary appendix SA-1.) **D.** The resulting multiplication factor, $\tilde{\tilde{M}}_P\left[\alpha[s_1, s_2,..., s_{N^X}]; s_0^{Root}; C_K \mid T\right]$, is approximately the

product of the contributions from different parts of the phylogenetic tree: one from the part with "fixed" ancestral states ($\left(\tilde{\mathbf{M}}_P \right)_0$), and the others from the isolated gap-masses ($\left(\tilde{\mathbf{M}}_P \right)_1$ and $\left(\tilde{\mathbf{M}}_P \right)_2$). (See Eq.(SA-1.9) of Supplementary appendix SA-1.)

NOTE: For more general gap patterns, this purely vertical partitioning will NOT work. See Section SSA-1 (and Figure SS1) of "suppl_addendum.xxx.pdf" for an extended method that can partition such more general cases.

(A) Tree

MSA (w/ null column)

1		L	---	---	---	---	---	---	---	---	R
2		L	1	---	---	---	---	---	6	7	R
3		L	1	2	3	---	4	5	6	7	R
4		L	1	2	3	---	4	5	6	7	R
5		L	1	2	---	---	---	---	---	7	R
6		L	---	---	---	---	---	5	6	7	R

(B) Ancestral states

R		L	1	2	3	x1	4	5	6	7	R
a1		L	1	2	3	x1	4	5	6	7	R
a2		L	1	---	---	---	---	---	6	7	R
a3		L	1	2	---	---	---	5	6	7	R

Figure S4. Example of non-parsimonious history that "connect" effectively independent indel blocks.

We will use the example MSA (and tree) given in Figure S3A.

In this case, an extra site (grey cell with ancestry "*x1*") in the ancestral states (*R* and *a1*) was deleted in all extant sequences (1-6), resulting in a null column (grey). (In this case, in addition to the four deletions needed for creating MSA in Figure S3A, two more deletions, along branches *a1*-3 and *R*-4, are necessary in order to completely delete the site with ancestry "*x1*".)

Such null columns will not usually be predicted via single-optimum-search aligners.