Addendum to Supplementary Materials for the Blueprint of the "Alignment Neighborhood Explorer" (ANEX) (tentatively named),

by Kiyoshi Ezawa

(Finished on June 7th, 2017; added CC4 statement on August 14th, 2020)

Table of Contents

Supplementary-Supplementary Appendix

SSA-1. Revisiting the practical factorability of indel configuration multiplication factor (discussed in SA-1).

The block-wise (contrasted to segment-wise) factorization formulas, Eqs.(SA-1.9,10) in "suppl_blueprint1.xxx.doc", are actually NOT so practically useful for the following three reasons:

(1) They are not applicable to the gapped segments containing insertion-type gaps, as in Figure SS1 A.

(2) They are not so useful when actually calculating the contributions to a *given* MSA (of extant sequences), instead of to a given set of states at all nodes (including both extant and ancestral sequences).

(3) They are not so useful, either, when attempting to calculate the *increment(s) of* the contributions (or the entire MSA probability) caused by a move in the gap-configuration(s) of an MSA.

Thus, we hereby attempt to give practically more useful formulas for the block-wise factorization of the contributions.

First, we will specify a reference ancestral sequence state, $s_0(n)$, at each internal node ($n \in \mathrm{N}^{IN}$). (Thus far, only one reference sequence state (s_0^{Root}) was specified in each gapped segment, only at the root node (n^{Root}).) A simplest candidate of a set of such reference states would be the ancestral gap states uniquely (and fairly quickly) given by the Dollo parsimony principle (Farris 1977). (See Figure SS1 B for an example. Indeed, our program implementation of this calculation will use the Dollo parsimonious gap states as the reference states, $\left\{ s_0(n) \right\}_{n \in \mathrm{N}^{IN}}$.) Then, using the identity:

$$\delta R_X^{ID}(s^A(b),\ s_0^{Root},\ \tau)[C_{\mathrm{K}}] \ = \ \delta R_X^{ID}(s_0^A(b),\ s_0^{Root},\ \tau)[C_{\mathrm{K}}] \ + \ \delta R_X^{ID}(s^A(b),\ s_0^A(b),\ \tau)[C_{\mathrm{K}}],$$

where $s_0^A(b) \equiv s_0\left(n^A(b) \right)$, we can rewrite Eq.(SA-1.3) supplemented with Eq.(SA-1.4) as follows:

$$\begin{aligned}
&\mathbf{M}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{ s(n) \right\}_{\mathrm{N}^{IN}}; s_0^{Root}; C_{\mathrm{K}} \mid T \right] \\
&\equiv \mathbf{M}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{ s(n) \right\}_{\mathrm{N}^{IN}}; C_{\mathrm{K}} \mid T \right] \ \mu_P\left[s\left(n^{Root} \right), s_0^{Root}, n^{Root}; C_{\mathrm{K}} \right] \\
&\times \exp\left(-\Phi_0 [T, \left\{ s_0(n) \right\}_{\mathrm{N}^{IN}}, C_{\mathrm{K}}; \left\{ \Theta^{ID}(b) \right\}_T] \right) \times \exp\left(- \sum_{b \in \{b\}_T} \int_{t\left(n^A(b) \right)}^{t\left(n^D(b) \right)} d\tau \, \delta R_X^{ID}(s^A(b),\ s_0^A(b),\ \tau)[C_{\mathrm{K}}] \right)
\end{aligned}$$

.

--- Eq.(SSA-1.1)

Here,

$$\Phi_0[T, \left\{ s_0(n) \right\}_{\mathrm{N}^{IN}}, C_{\mathrm{K}}; \left\{ \Theta^{ID}(b) \right\}_T] \quad \equiv \quad \sum_{b \in \{b\}_T} \int_{t(n^A(b))}^{t(n^D(b))} d\tau \, \delta R_X^{ID}(s^A(b),\ s_0^{Root},\ \tau)[C_{\mathrm{K}}]$$

$$\Phi_0[T, \left\{ s_0(n) \right\}_{\mathrm{N}^{IN}}, C_{\mathrm{K}}; \left\{ \Theta^{ID}(b) \right\}_T] \quad \equiv \quad \sum_{b \in \{b\}_T} \int_{t(n^A(b))}^{t(n^D(b))} d\tau \, \delta R_X^{ID}(s_0^A(b),\ s_0^{Root},\ \tau)[C_{\mathrm{K}}]$$

--- Eq.(SSA-1.2)

is the “reference” phase factor determined uniquely by the tree (T), the gapped segment (C_{K}) and the reference sequence states ($\left\{ s_0(n) \right\}_{\mathrm{N}^{IN}}$) (and the indel model ($\left\{ \Theta^{ID}(b) \right\}_T$)); the factor is independent of any other specific local indel histories that are compatible with the MSA within C_{K} , hence it can be computed easily (and fairly quickly).

Second, we will re-define **effectively independent indel blocks** (, which were referred to as “partial indel history zones” in Figure S3 C (in “suppl_blueprint1_ANEX.xxx.pdf”,) so that they can also cover more complex cases including insertions. For this purpose, let us recall that we are now considering gap-state configurations within the direct product space, $C_{\mathrm{K}} \times T$ (as represented by the array of trees in Figure SS1 B), or more simply, $C_{\mathrm{K}} \times \mathbf{N}^{IN}(T)$. The latter is usually adequate for the current purpose, because we are considering all possible sets of ancestral gap-states compatible with the local MSA, instead of all possible indel histories compatible with the local MSA. (Please remember that gap-states at external nodes ($n \in \mathbf{N}^X$) are already fixed, given a local MSA.) Now, remember that the Dollo parsimonious ancestral state contains the smallest number of ancestral sites occupied by residues, because the Dollo parsimonious history consists of the shortest paths (along the tree) connecting residue-occupied sites at external nodes (see, e.g., Figure 4 of Ezawa, Graur and Landan Part I). In other words, all other MSA-compatible ancestral gap-states can be made from the Dollo parsimonious state by filling some empty sites (i.e., gaps) at ancestral nodes with residues in a phylogenetically consistent manner (e.g., Chindelevitch et al. 2006; Diallo et al. 2007), which is enabled by extending some “paths” of residue-occupied sites from the network of such sites representing the Dollo parsimonious states (see, e.g., Figure 4 of Ezawa, Graur and Landan Part I). This “minimal” nature of the Dollo parsimonious indel history enables it to partition each gapped segment ($C_{\mathrm{K}} \times T$) into some blocks, $\mathbf{\Gamma}_{\mathrm{K};\,i}$ ($i = 1, 2, \ldots, I_{\mathrm{K}}$), each of which can accommodate some indels (in parsimonious indel histories), and a “partitioning” network of residue-occupied sites of the Dollo parsimonious states extended across the tree, denoted as $\mathbf{\Gamma}_{\mathrm{K};\,0}$. (Figure SS1 C illustrates such a partitioning.) This partitioning can be represented in an abstract equation using the symbol, “ $\cup$ ”, for a union of sets:

$$C_{\mathrm{K}} \times T \quad = \quad \mathbf{\Gamma}_{\mathrm{K};0} \cup \left\{ \bigcup_{i=1}^{I_{\mathrm{K}}} \mathbf{\Gamma}_{\mathrm{K};\,i} \right\} .$$ --- Eq.(SSA-1.3)

(Here, it should be noted that the sets involved in the right-hand side are mutually disjoint from one another.) By definition, every pair of a node and a site (referred henceforth as a "node-site pair") belonging to $\boldsymbol{\Gamma}_{\mathrm{K};\,0}$ must always be occupied by a residue, regardless of the indel histories (as long as they are compatible with the MSA). In contrast, each node-site pair belonging to $\boldsymbol{\Gamma}_{\mathrm{K};\,i}$ ($i = 1, 2, \ldots, I_{\mathrm{K}}$) can be either empty or occupied with a residue depending on the MSA-compatible history, although it should always be empty in the Dollo parsimonious indel history. And it should also be noted that every indel event in every MSA-compatible parsimonious (or next-parsimonious) indel histories should be completely confined in one of $\left\{\boldsymbol{\Gamma}_{\mathrm{K};\,i}\right\}_{i=1,2,\ldots,I_{\mathrm{K}}}$; otherwise, the "partitioning" network, $\boldsymbol{\Gamma}_{\mathrm{K};\,0}$, is not working as it should, thus it must be re-defined. **Therefore, each of $\left\{\boldsymbol{\Gamma}_{\mathrm{K};\,i}\right\}_{i=1,2,\ldots,I_{\mathrm{K}}}$ defines an "effectively independent indel block" (previously referred to as "partial indel history zone"), or an "indel block" for short.** (We consider that two node-site pairs belong to the same indel block if they are connected via a path of node-site pairs that are empty in the Dollo parsimonious history. Otherwise, that is, if they are *clearly* separated by at least a node-site pair that is occupied by a residue in the Dollo parsimonious history, we consider them as belonging to different indel blocks.)

Now, let us extend Eq.(SA-1.7) (in "suppl_blueprint1_ANEX.xxx.pdf") for the purely vertical partitioning of a gapped segment, in order to create a formula suitable for the more general partitioning Eq.(SSA-1.3). For this purpose, let $P\left\{\boldsymbol{\Gamma}_{\mathrm{K};i}\right\}|_n$ be the projection of an indel block $\boldsymbol{\Gamma}_{\mathrm{K};\,i}$ onto a node n (, which could be either internal or external), which is nothing other than the set of all sites (both empty and residue-occupied) in the ancestral sequence at n belonging to $\boldsymbol{\Gamma}_{\mathrm{K};\,i}$. Similarly, let $P\left\{\boldsymbol{\Gamma}_{\mathrm{K};0}\right\}|_n$ be the projection of the partitioning network $\boldsymbol{\Gamma}_{\mathrm{K};\,0}$ onto n. (Panels D & E of Figure SS1 illustrate these projections.) Then, thanks to Eq.(SSA-1.3), the following decomposition always holds at node n:

$$C_{\mathrm{K}} \;=\; P\left\{\Gamma_{\mathrm{K};0}\right\}|_n \cup \left\{\bigcup_{i=1}^{I_{\mathrm{K}}} P\left\{\Gamma_{\mathrm{K};i}\right\}|_n\right\} . \quad \text{--- Eq.(SSA-1.4)}$$

(Again, the sets involved in the right-hand side are mutually disjoint. It should also be noted that $P\left\{\boldsymbol{\Gamma}_{\mathrm{K};i}\right\}|_n$'s may be empty sets for some i's.) Now, each element of $\Delta_{\Sigma}\left[C_{\mathrm{K}};\, s_0^{Root};\, \alpha[s_1, s_2, \ldots, s_{N^X}];\left\{n \in \mathrm{N}^{IN}(T)\right\};\, T\right]$, i.e., a set of differences of internal gap states (within the region C_{K}) from the reference root state (s_0^{Root}), $\left\{s(n) - s_0^{Root}\right\}_{\mathrm{N}^{IN}}[C_{\mathrm{K}}]$, could be considered as an N^{IN}-tuple, and that each of its components is the difference of an internal gap state ($s(n)$) at an internal node (n) from s_0^{Root}, which will be denoted as $\left(s(n)-s_0^{Root}\right)[C_{\mathrm{K}}]$ hereafter. (It should be noted that the components of the N^{IN}-tuples are NOT mutually independent of each other, because they have to satisfy the phylogenetic consistency condition.) Because differences in the indel blocks are effectively independent of one another, we almost always have:

$$\begin{aligned}&\left(s(n)-s_0^{Root}\right)[C_{\mathrm{K}}]\\ =\ &\left(s_0(n)-s_0^{Root}\right)[C_{\mathrm{K}}]+\left(s(n)-s_0(n)\right)\left[P\left\{\mathbf{\Gamma}_{\mathrm{K};0}\right\}|_n\right]+\sum_{i=1}^{I_{\mathrm{K}}}\left(\left(s(n)-s_0(n)\right)\left[P\left\{\mathbf{\Gamma}_{\mathrm{K};i}\right\}|_n\right]\right)\end{aligned}.$$

--- Eq.(SSA-1.5)

Here, $\left(s(n)-s_0^{Root}\right)\left[P\left\{\mathbf{\Gamma}_{\mathrm{K};i}\right\}|_n\right]$ is the gap-state difference within the indel block, $\mathbf{\Gamma}_{\mathrm{K};\,i}$.

Because the gap-state within the partitioning network ($\mathbf{\Gamma}_{\mathrm{K};\,0}$) is almost always identical to that for $s_0(n)$, the second term on the right-hand side almost always vanishes, hence we have:

$$\left(s(n)-s_0^{Root}\right)[C_{\mathrm{K}}]\ =\ \left(s_0(n)-s_0^{Root}\right)[C_{\mathrm{K}}]+\sum_{i=1}^{I_{\mathrm{K}}}\left(\left(s(n)-s_0(n)\right)\left[P\left\{\mathbf{\Gamma}_{\mathrm{K};i}\right\}|_n\right]\right).$$

--- Eq.(SSA-1.6)

Now, consider the N^{IN}-tuple again. As already noted, its different components are NOT independent of one another. Nevertheless, *as long as the partitioning network remains intact*, it is sufficient to consider the phylogenetic consistency conditions among the components, $\left\{\left(s(n)-s_0(n)\right)\left[P\left\{\mathbf{\Gamma}_{\mathrm{K};i}\right\}|_n\right]\middle|\ n\in \mathrm{N}^{IN}(T)\right\}$, *within each indel block* ($\mathbf{\Gamma}_{\mathrm{K};\,i}$). In such cases, components within different indel blocks can be treated independently of one another. Thus, the space of ancestral state differences, $\Delta_\Sigma\left[C_{\mathrm{K}};s_0^{Root};\alpha[s_1,s_2,...,s_{N^X}];\left\{n\in\mathrm{N}^{IN}(T)\right\};T\right]$' can be approximately decomposed as follows. First, separate the constant differences between the reference ancestral states and the reference root state from the remaining variable parts:

$$\begin{aligned}&\Delta_\Sigma\left[C_{\mathrm{K}};s_0^{Root};\alpha[s_1,s_2,...,s_{N^X}];\left\{n\in\mathrm{N}^{IN}(T)\right\};T\right]\\ =\ &\left\{s_0(n)-s_0^{Root}\right\}_{\mathrm{N}^{IN}(T)}\ +\ \Delta_\Sigma\left[C_{\mathrm{K}};\left\{s_0(n)\right\}_{\mathrm{N}^{IN}(T)};\alpha[s_1,s_2,...,s_{N^X}];\left\{n\in\mathrm{N}^{IN}(T)\right\};T\right]\end{aligned}.$$

--- Eq.(SSA-1.7)

On the right-hand side, the first term represents the set of constant differences, and the second term represents the remaining variable parts. Then, the second term can be approximately decomposed as follows:

$$\begin{aligned}&\Delta_{\Sigma}\left[C_{\mathrm{K}};\left\{s_0(n)\right\}_{\mathrm{N}^{IN}(T)};\alpha[s_1,s_2,\ldots,s_{N^X}];\left\{n\in\mathrm{N}^{IN}(T)\right\};T\right]\\ \approx\ &\mathop{\times}_{i=1}^{I_{\mathrm{K}}}\Delta_{\Sigma}\left[\Gamma_{\mathrm{K};i};\left\{s_0(n)\right\}_{\mathrm{N}^{IN}(T)};\alpha[s_1,s_2,\ldots,s_{N^X}];\left\{n\in\mathrm{N}^{IN}(T)\right\};T\right]\end{aligned}\ . \quad \text{--- Eq.(SSA-1.8)}$$

Here, each element of the component space, $\Delta_{\Sigma}\left[\Gamma_{\mathrm{K};i};\left\{s_0(n)\right\}_{\mathrm{N}^{IN}(T)};\alpha[s_1,s_2,\ldots,s_{N^X}];\left\{n\in\mathrm{N}^{IN}(T)\right\};T\right]$, is a phylogenetically consistent set, $\left\{s(n)-s_0(n)\right\}_{\mathrm{N}^{IN}}\left[\Gamma_{\mathrm{K};i}\right]\equiv\left\{\left(s(n)-s_0(n)\right)\left[P\left\{\Gamma_{\mathrm{K};i}\right\}\big|_n\right]\Big|\ n\in\mathrm{N}^{IN}(T)\right\}$, of ancestral gap state differences within the given indel block, $\boldsymbol{\Gamma}_{\mathrm{K};\,i}$.

Now, in order to extend the purely vertical factorization of the multiplication factor, Eq.(SA-1.9) (or Eq.(SA-1.9')) in "suppl_blueprint1_ANEX.xxx.pdf", we need two assumptions, in addition to the approximate space decomposition given by Eqs.(SSA-1.7,8). One is the assumption that the region-wise increment of the exit rate can also be further decomposed (at least approximately) just as in Eq.(SSA-1.6):

$$\delta R_X^{ID}(s(n),\ s_0(n),\ \tau)[C_{\mathrm{K}}]\ \approx\ \sum_{i=1}^{I_{\mathrm{K}}}\left(\delta R_X^{ID}(s(n),\ s_0(n),\ \tau)\left[\Gamma_{\mathrm{K};i}\right]\right). \quad \text{--- Eq.(SSA-1.9)}$$

(Here we omitted the increment confined in $\boldsymbol{\Gamma}_{\mathrm{K};\,0}$, because the ancestral gap states within the partitioning network are almost always unchanged.) The other is the assumption that the (multiplicative) increment of the prior probability of the root state confined in C_{K} can also be further factorized (at least approximately) as:

$$\begin{aligned}&\mu_P\left[s\left(n^{Root}\right),\,s_0^{Root},\,n^{Root};\,C_{\mathrm{K}}\right]\\ \approx\ &\mu_P\left[s\left(n^{Root}\right),\,s_0^{Root},\,n^{Root};\,P\left\{\Gamma_{\mathrm{K};0}\right\}\big|_{n^{Root}}\right]\times\prod_{i=1}^{I_{\mathrm{K}}}\mu_P\left[s\left(n^{Root}\right),\,s_0^{Root},\,n^{Root};\,P\left\{\Gamma_{\mathrm{K};i}\right\}\big|_{n^{Root}}\right]\end{aligned}\ .$$

--- Eq.(SSA-1.10)

(Whether Eqs.(SSA-1.9.10) indeed hold or not is non-trivial in general, especially when the considered model has indel variation across sites (or regions). Here, however, we *assume* that they hold at least approximately.)

Using Eqs.(SSA-1.9,10), the fact that different indel blocks ($\boldsymbol{\Gamma}_{\mathrm{K};\,i}$'s)are almost always independent of each other, and the fact that the partitioning network ($\boldsymbol{\Gamma}_{\mathrm{K};\,0}$) almost always remains intact, the multiplication factor, Eq.(SA-1.3) supplemented with Eq.(SA-1.4) (both in "suppl_blueprint1_ANEX.xxx.pdf"), which has been rewritten here as Eq.(SSA-1.1) supplemented with Eq.(SSA-1.2), can be further factorized (at least approximately) as follows:

$$
\begin{aligned}
&\mathrm{M}_P\left[\underset{\smile}{\alpha}[s_1, s_2, \ldots, s_{N^X}]; \left\{s(n)\right\}_{\mathrm{N}^{IN}}; s_0^{Root}; C_{\mathrm{K}} \mid T\right] \\
\approx\ &\mathrm{M}_P\left[\alpha[\underset{\smile}{s_1}, s_2, \ldots, s_{N^X}]; \left\{s_0(n)\right\}_{\mathrm{N}^{IN}}; s_0^{Root}; \Gamma_{\mathrm{K};0} \mid T\right] \\
&\times \prod_{k=1}^{I_{\mathrm{K}}} \mathrm{M}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{s(n)\right\}_{\mathrm{N}^{IN}}; \left\{s_0(n)\right\}_{\mathrm{N}^{IN}}; \Gamma_{\mathrm{K};i} \mid T\right]
\end{aligned}
\quad . \text{--- Eq.(SSA-1.11)}
$$

Here,

$$
\begin{aligned}
&\mathrm{M}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{s_0(n)\right\}_{\mathrm{N}^{IN}}; s_0^{Root}; \Gamma_{\mathrm{K};0} \mid T\right] \\
\equiv\ &\mu_P\left[s\left(n^{Root}\right), s_0^{Root}, n^{Root}; P\left\{\Gamma_{\mathrm{K};0}\right\}\Big|_{n^{Root}}\right] \times \exp\left(-\Phi_0\left[T, \left\{s_0(n)\right\}_{\mathrm{N}^{IN}}, C_{\mathrm{K}}; \left\{\Theta^{ID}(b)\right\}_T\right]\right)
\end{aligned}
$$

--- Eq.(SSA-1.12)

is the portion of the multiplication factor associated with the ancestral gap states, $\left\{s(n)\right\}_{\mathrm{N}^{IN}}$, within the gapped segment, C_{K}, contributed from the partitioning network, $\mathit{\Gamma}_{K;0}$. It should be noted that this portion remains unchanged regardless of the ancestral gap states, $\left\{s(n)\right\}_{N^{IN}}$.

And

$$
\begin{aligned}
&\breve{M}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{s(n)\right\}_{N^{IN}}; \left\{s_0(n)\right\}_{N^{IN}}; \mathit{\Gamma}_{K;i} \mid T\right] \\
\text{¿}\ &M_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{s(n)\right\}_{N^{IN}}; \mathit{\Gamma}_{K;i} \mid T\right]\ \mu_P\left[s\left(n^{Root}\right), s_0^{Root}, n^{Root}; P\left\{\mathit{\Gamma}_{K;i}\right\}\Big|_{n^{Root}}\right] \\
&\times \exp\left(-\sum_{b \in \{b\}_T} \int_{t\left(n^A(b)\right)}^{t\left(n^D(b)\right)} d\tau\ \delta R_X^{ID}\left(s^A(b), s_0^A(b), \tau\right)\left[\mathit{\Gamma}_{K;i}\right]\right)
\end{aligned}
$$

--- Eq.(SSA-1.13)

, with

$$
\begin{aligned}
&M_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{s(n)\right\}_{N^{IN}}; \mathit{\Gamma}_{K;i} \mid T\right] \\
\text{¿}\ &\prod_{b \in \{b\}_T} \left\{ \prod_{\gamma_{\kappa_b}(b) \subseteq \mathit{\Gamma}_{K;i}} \tilde{\mu}_P\left[\left(\tilde{\mathit{\Lambda}}^{ID}\left[\gamma_{\kappa_b}(b); \alpha\left(s^A(b), s^D(b)\right)\right], b\right) \mid \left(s^A(b), n^A(b)\right)\right]\right\}
\end{aligned}
$$

--- Eq.(SSA-1.14)

is the portion of the multiplication factor contributed from the indel block, $\mathit{\Gamma}_{K;i}$.

Then, substituting Eq.(SSA-1.11) into Eq.(SA-1.2) in “suppl_blueprint1_ANEX.xxx.pdf”, and using Eqs.(SSA-1.7,8), we get the approximate factorization of the entire multiplication factor for the local MSA within C_K:

$$
\begin{aligned}
&\breve{\tilde{M}}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; s_0^{Root}; C_K \mid T\right] \\
\text{¿}\ &\breve{M}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{s_0(n)\right\}_{N^{IN}}; s_0^{Root}; \mathit{\Gamma}_{K;0} \mid T\right] \\
&\times \prod_{i=1}^{I_K} \left(\breve{\tilde{M}}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{s_0(n)\right\}_{N^{IN}}; \mathit{\Gamma}_{K;i} \mid T\right]\right)
\end{aligned}
\quad \text{--- Eq.(SSA-1.15)}
$$

. Here, $\breve{M}_P\left[\alpha[s_1, s_2, \ldots, s_{N^X}]; \left\{s_0(n)\right\}_{N^{IN}}; s_0^{Root}; \mathit{\Gamma}_{K;0} \mid T\right]$ has already been defined in Eq.(SSA-1.12), and

$$\sum_{\text{¿}\ \Delta_{\Sigma}\left[\Gamma_{K;i};\ \{s_0(n)\}_{N^{IN}};\ \alpha[s_1,s_2,\ldots,s_{N^X}];\ \{n\in N^{IN}(T)\};\ T\right]\text{¿}} \text{¿}\ \{s(n)-s_0(n)\}_{N^{IN}}[\Gamma_{K;i}]\text{¿¿¿¿}$$

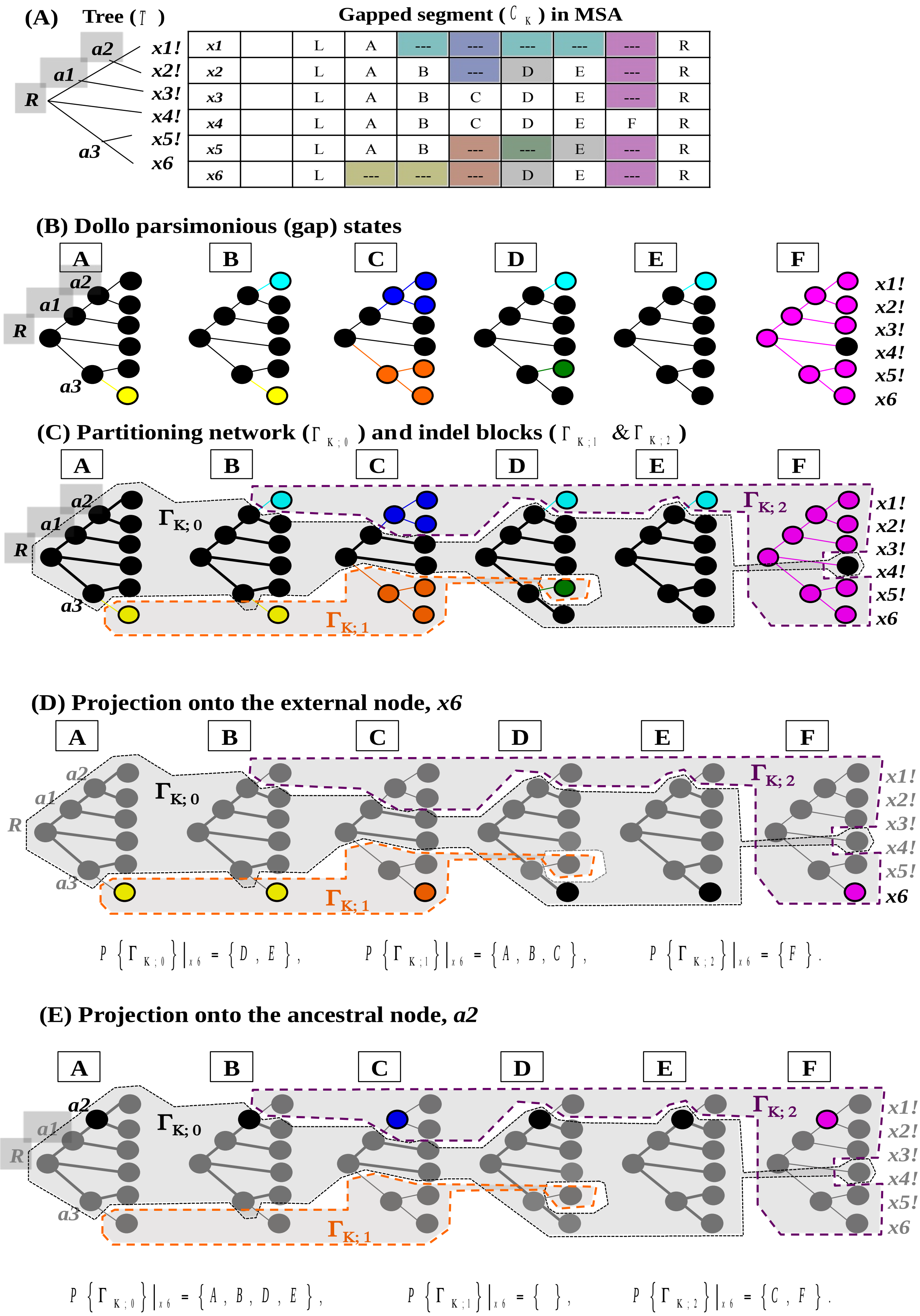
$$\sum_{\text{¿}\ \Delta_{\Sigma}\left[\Gamma_{K;i};\ \{s_0(n)\}_{N^{IN}};\ \alpha[s_1,s_2,\ldots,s_{N^X}];\ \{n\in N^{IN}(T)\};\ T\right]\text{¿}} \text{¿}\ \{s(n)-s_0(n)\}_{N^{IN}}[\Gamma_{K;i}]\text{¿¿¿¿}$$

--- Eq.(SSA-1.16)

(with the summands defined in Eq.(SSA-1.13)) is the total multiplicative contribution from the indel block $\Gamma_{K;i}$.

Because each of the multiplicative factors in the approximate factorization formula, Eq.(SSA-1.15), can in principle be calculated independently of one another, this should facilitate the computation of the entire multiplication factor for each gapped segment, and also the computation of its change in response to a move of the gap-pattern in the gapped segment and its neighbors. Thus, this extension of the vertical partitioning ((SA-1.9,10) in "suppl_blueprint1.xxx.doc") will resolve the three drawbacks of its predecessor mentioned at the top of this section.

Supplementary-Supplementary Figures (with legends)



(A) Tree (T) Gapped segment (C K) in MSA

x1		L	A	---	---	---	---	---	R
x2		L	A	B	---	D	E	---	R
x3		L	A	B	C	D	E	---	R
x4		L	A	B	C	D	E	F	R
x5		L	A	B	---	---	E	---	R
x6		L	---	---	---	D	E	---	R

(B) Dollo parsimonious (gap) states

(C) Partitioning network ($\Gamma_{K;0}$) and indel blocks ($\Gamma_{K;1}$ & $\Gamma_{K;2}$)

(D) Projection onto the external node, *x6*

$P\{\Gamma_{K;0}\}|_{x6} = \{D, E\}$, $P\{\Gamma_{K;1}\}|_{x6} = \{A, B, C\}$, $P\{\Gamma_{K;2}\}|_{x6} = \{F\}$.

(E) Projection onto the ancestral node, *a2*

$P\{\Gamma_{K;0}\}|_{x6} = \{A, B, D, E\}$, $P\{\Gamma_{K;1}\}|_{x6} = \{\ \}$, $P\{\Gamma_{K;2}\}|_{x6} = \{C, F\}$.

Figure SS1. Extending notion of purely vertical partitioning of gapped segment.

(A) An example of an input data set, consisting of a tree (T) and a gapped segment (C_K) of a multiple sequence alignment (MSA). In this example, purely vertical partitioning (i.e., partitioning purely in terms of nodes (or the tree), as in Figure S3 or section SA-1 of "suppl_blueprint1_ANEX.xxx.pdf") does NOT work, because of the insertion of site F. (In this figure, capital roman alphabets label the sites in the MSA.)

(B) Minimal gap states reconstructed under the Dollo parsimony principle. The black-filled circles represent the sites filled with residues at the nodes, and color-shaded open circles represent the empty sites at the nodes. The colors correspond to the colors in panel A. And a black branch (edge) indicates that the site remains filled with a residue along the branch, whereas a colored branch indicates that the site could become (or remain) empty along the branch.

(C) According to the Dollo parsimonious gap states in panel B, we can define a partitioning network ($\Gamma_{K;0}$) and indel blocks ($\Gamma_{K;i}$, with $i=1,\ldots,I_K$). (Here, $I_K = 2$.) The partitioning network is constructed by connecting all black-filled nodes and black branches. An indel block is constructed by connecting *contiguous* open nodes and colored branches.

(D) & (E) Projections of the partitioning network and indel blocks onto the external node, $x6$, and their projections onto the ancestral node, $a2$, respectively. In these panels, the relevant nodes are highlighted by coloring the other nodes and branches in grey. In each panel, the results of the projections are shown below the array of trees (, which represents $C_K \times T$).