bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part IV: Incorporation of substitutions and other mutations

**Kiyoshi Ezawa [1, 2, *], Dan Graur [1], and Giddy Landan [1, 3]**

[1] Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001, USA
[2] Present address: Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka 820-8502, JAPAN
[3] Present address: Institute of Genomic Microbiology, Heinrich-Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, GERMANY

[*] Corresponding Author
(E-mail: kezawa@bio.kyutech.ac.jp, kezawa.ezawa3@gmail.com)

## Abstract

### Background

Insertions and deletions (indels) account for more nucleotide differences between two related DNA sequences than substitutions do, and thus it is imperative to develop a stochastic evolutionary model that enables us to reliably calculate the probability of the sequence evolution through indel processes. In a separate paper (Ezawa, Graur and Landan 2015a), we established the theoretical basis of our *ab initio* perturbative formulation of a continuous-time Markov model of the evolution of an *entire* sequence via insertions and deletions along time axis. In other separate papers (Ezawa, Graur and Landan 2015b,c), we also developed various analytical and computational methods to concretely calculate alignment probabilities via our formulation. In terms of frequencies, however, substitutions are usually more common than indels. Moreover, many experiments suggest that other mutations, such as genomic rearrangements and recombination, also play some important roles in sequence evolution.

### Results

Here, we extend our *ab initio* perturbative formulation of a *genuine* evolutionary model so that it can incorporate other mutations. We give a sufficient set of conditions that the probability of evolution via both indels and substitutions is factorable into the product of an overall factor and local contributions. We also show that, under a set of conditions, the probability can be factorized into two sub-probabilities, one via indels alone and the other via substitutions alone. Moreover, we show that our formulation can be extended so that it can also incorporate genomic rearrangements, such as inversions and duplications. We also discuss how to accommodate some other types of mutations within our formulation.

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**Conclusions**

Our *ab initio* perturbative formulation thus extended could in principle describe the stochastic evolution of an *entire* sequence along time axis via major types of mutations.

[This paper and three other papers (Ezawa, Graur and Landan 2015a,b,c) describe a series of our efforts to develop, apply, and extend the *ab initio* perturbative formulation of a general continuous-time Markov model of indels.]

## Keywords

Insertion/deletion (indel), substitution, genomic rearrangement, inversion, duplication, pairwise sequence alignment (PWA), multiple sequence alignment (MSA), probability, likelihood, continuous-time Markov model, perturbation theory

## List of abbreviations

HMM, hidden Markov model; indel, insertion/deletion; LHS, local history set; MSA, multiple sequence alignment; PAS, preserved ancestral site; PWA, pairwise alignment.

## Table of contents

**Introduction** **pp.3-4**
**Results & Discussion** **pp.5-11**
1. Incorporation of substitutions pp.5-7
2. Incorporation of inversions and duplications pp.7-10
3. Other mutational mechanisms pp.10-11
4. Boundary conditions and cut-off lengths p.11

**Conclusions** **p.12**
**Authors' contributions** **p.12**
**Acknowledgements** **pp.12-13**
**Appendix: Details on incorporation of substitutions** **pp.14-32**
A1. Formulating sequence evolution via both indels and substitutions pp.14-16
A2. Factorizing probability into regional contributions pp.16-20
A3. Factorizing probability into basic and residue components pp.21-30
A4. Pursuing further biological realism pp.30-32
**References** **pp.33-35**

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Introduction

The evolution of DNA, RNA, and protein sequences is driven by mutations such as base substitutions, insertions and deletions (indels), recombination, and other genomic rearrangements (*e.g.*, Graur and Li 2000; Gascuel 2005; Lynch 2007). Thus far, analyses on substitutions have predominated in the field of molecular evolutionary study, in particular using the probabilistic (or likelihood) theory of substitutions that is now widely accepted (e.g., Felsenstein 1981, 2004; Yang 2006). However, some recent comparative genomic analyses have revealed that indels account for more base differences between the genomes of closely related species than substitutions (*e.g.*, Britten 2002; Britten et al. 2003; Kent *et al*. 2003; The International Chimpanzee Chromosome 22 Consortium 2004; The Chimpanzee Sequencing and Analysis Consortium 2005). It is therefore imperative to develop a stochastic model that enables us to reliably calculate the probability of sequence evolution via mutations including insertions and deletions.

Traditionally, the computation of probabilities of indels has been based on hidden Markov models (HMMs) or transducer theories (see, *e.g.*, Rivas 2005; Bradley and Holmes 2007; Miklós et al. 2009). However, these methods have two fundamental problems, one regarding the theoretical grounds and the other regarding the biological realism. (See the "background" section in part I (Ezawa, Graur and Landan 2015a) for more details on these problems.)

As an unprecedented approach to these problems on the probabilistic models of indels, we proposed an *ab initio* perturbative formulation of a *genuine* stochastic evolutionary model, which describes the evolution of an *entire* sequence via indels along the time axis (Ezawa, Graur and Landan 2015a). Such a *genuine* evolutionary model is devoid of the aforementioned problems from the beginning. More specifically, our *genuine* evolutionary model is a general continuous-time Markov model of sequence evolution via indels. It allows any indel rate parameters including length distributions, but it does not impose any unnatural restrictions on indels. In part I of this series of study (Ezawa, Graur and Landan 2015a), we gave the theoretical basis of our *ab initio* formulation. Especially, we derived a sufficient and nearly necessary set of conditions under which the probability of an alignment is factorable, like a sort of HMM. In part II (Ezawa, Graur and Landan 2015b), we developed some analytical techniques for performing concrete perturbation analyses. In part III (Ezawa, Graur and Landan 2015c), we developed an algorithm that can calculate the first-approximate probability of each local MSA delimited by gapless columns, given an input MSA and under a given parameter setting including a phylogenetic tree.

Mainly in order to avoid unnecessary confusions of the readers, these studies (Ezawa, Graur and Landan 2015a,b,c) dealt with indels alone. In terms of frequencies, however, substitutions are usually more common than indels (*e.g.*, Lunter 2007; Cartwright 2009). Moreover, many experiments suggest that still other mutations, such as genomic rearrangements and recombination, also play some important roles in the evolution of protein and DNA sequences (*e.g.*, Graur and Li 2000; Lynch 2007; Gu et al. 2008). Thus, a natural question arises as to whether the methods and conclusions obtained in these papers are still valid, at least to some degree, even when we consider other types of mutations as well.

In this study, we extend our *ab initio* parturbative formulation of a *genuine* evolutionary model so that it can incorporate other mutations, such as substitutions and genomic rearrangements. In Section 1 of Results & Discussion, we consider the model of sequence evolution via both indels and substitutions. We give a sufficient set of conditions that the probability of an alignment under this model is factorable into

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the product of an overall factor and the contributions from local evolutionary processes. We also show that, under a set of conditions, the probability can be factorized into two components. One is the "basic" component given by the indel processes alone, and the other is the residue component that concerns the substitution processes alone. Our set of conditions is more general than the commonly used conditions that the indigenous and inserted residue frequencies are equal to the equilibrium frequencies of the substitution model (*e.g.*, Thorne et al. 1991, 1992; Miklós et al. 2004; Rivas and Eddy 2008). In Section 2, we show that our *ab initio* formulation of a genuine evolutionary model can be extended to incorporate genomic rearrangements, especially inversions and duplications. In Section 3, we discuss how we can accommodate some other types of mutations within our formulation. Appendix gives mathematical details on Section 1.

This paper is part IV of a series of our papers that documents our efforts to develop, apply, and extend the *ab initio* perturbative formulation of the general continuous-time Markov model of sequence evolution via indels. Part I (Ezawa, Graur and Landan 2015a) gives the theoretical basis of this entire study. Part II (Ezawa, Graur and Landan 2015b) describes concrete perturbation calculations and examines the applicable ranges of other probabilistic models of indels. Part III (Ezawa, Graur and Landan 2015c) describes our algorithm to calculate the first approximation of the probability of a given MSA and simulation analyses to validate the algorithm. Finally, part IV (this paper) discusses how our formulation can incorporate substitutions and other mutations, such as duplications and inversions.

This paper basically uses the same conventions as used in part I (Ezawa, Graur and Landan 2015a). Briefly, a sequence state $s\,(\in S)$ is represented as an array of sites, each of which is either blank or equipped with some specific attributes. And each indel event is represented as an operator acting on the bra-vector, $\langle s|$, representing a sequence state. More specifically, the operator $\hat{M}_I(x,l)$ denotes the insertion of $l$ sites between the $x$ th and $(x+1)$ th sites, and the operator $\hat{M}_D(x_B,x_E)$ denotes the deletion of a sub-array between (and including) the $x_B$ th and the $x_E$ th sites. See Section 2 of part I for more details. And, as in part I, the following terminology is used. The term "an indel process" means a series of successive indel events with both the order and the specific timings specified, and the term "an indel history" means a series of successive indel events with only the order specified. And, throughout this paper, the union symbol, such as in $A\bigcup B$ and $\bigcup_{i=1}^{I} A_i$, should be regarded as the union of *mutually disjoint* sets (*i.e.,* those satisfying $A\bigcap B=\varnothing$ and $A_i\bigcap A_j=\varnothing$ for $i\neq j\,(\in\{1,\ldots,I\})$, respectively, where $\varnothing$ is an empty set), unless otherwise stated.

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Results & Discussion

**1. Incorporation of substitutions**

For clarity, we focused on the description of indel processes and omitted substitutions in the bulk of the paper. Actually, it is not so difficult to incorporate substitutions into our framework. For this purpose, we first extend the sequence state space $S$ ($= S^I, S^{II}, or\ S^{III}$) so that each site of the sequence will accommodate a residue in the set $\Omega$ (see, *e.g.*, Subsection 2.1 of part I (Ezawa, Graur and Landan 2015a)). When $S = S^I$ ($\cong \mathrm{N}_0 \equiv \{0,1,2,...\}$), the extended space is $\breve{S} = \Omega^* \equiv \bigcup_{L=0}^{\infty} \Omega^L$, in which each site is assigned a residue $\omega \in \Omega$. When $S = S^{II}$ ($\subset \Upsilon^*$), the ancestry assigned to each site, $\upsilon \in \Upsilon$, is replaced with a pair, $(\upsilon, \omega) \in \Upsilon \times \Omega$. Thus, the extended space is $\breve{S}^{II} \subset \{\Upsilon \times \Omega\}^*$. Similarly, when $S = S^{III}$ ($\subset \{\mathrm{N}_0 \times \mathrm{N}_1\}^*$), the extended space is $\breve{S}^{III} \subset \{\mathrm{N}_0 \times \mathrm{N}_1 \times \Omega\}^*$, in which each site is assigned a trio, $(\sigma, \xi, \omega) \in \mathrm{N}_0 \times \mathrm{N}_1 \times \Omega$. (Here $\sigma$ is the source identifier of the inserted (or initial) subsequence, and $\xi$ is the relative coordinate within the inserted (or initial) subsequence.) An extended sequence state $\breve{s} \in \breve{S}$ will sometimes be represented as $\breve{s} = (s, \vec{\omega})$ to explicitly show that it is composed of a "basic" component $s \in S$ and a residue component $\vec{\omega} = [\omega_1, ..., \omega_{L(\breve{s})}] \in \Omega^{L(\breve{s})}$. (Here $L(\breve{s})$ is the length of the sequence $\breve{s}$.) Once we extended the sequence state space, we can now extend the indel rate operator $\hat{Q}^{ID}(t)$ (given in Eqs.(2.4.2a-d) of part I). This is done in two steps. (1) We add the substitution component of the rate operator, $\hat{Q}^S(t)$. The operator is a generalization of the one given in Eq.(1.2.1') of part I to a sequence with multiple sites and to a more general substitution model. And (2) we extend the indel component ($\hat{Q}^{ID}(t)$) to take explicit account of the residue dependence of indels, including the relative probabilities among residue states filling in each inserted array of sites. See Appendix A1 for details. Then, the total rate operator of the entire evolutionary model ($\hat{Q}^{SID}(t)$) is given by adding the indel and substitution rate operators:

$$\hat{Q}^{SID}(t) \equiv \hat{Q}^{ID}(t) + \hat{Q}^S(t). \qquad \text{--- Eq.(1.1)}$$

At least formally, we could apply the perturbation expansion technique to the stochastic evolution operator of the entire model, $\hat{P}^{SID}(t_I, t_F) \equiv T\left\{\exp\left(\int_{t_I}^{t_F} dt\, \hat{Q}^{SID}(t)\right)\right\}$. ($T\{...\}$ denotes the time-ordered product.) We can do this by decomposing $\hat{Q}^{SID}(t)$ into two parts: $\hat{Q}^{SID}(t) = \hat{Q}_0^{SID}(t) + \hat{V}(t)$, where $\hat{V}(t)$ is treated as a "perturbation" part. There are two major ways of doing this. (A) To decompose it into $\hat{Q}_0^{SID}(t) = \hat{Q}_X^{ID}(t) + \hat{Q}_X^S(t)$ and $\hat{V}(t) = \hat{Q}_M^I(t) + \hat{Q}_M^D(t) + \hat{Q}_M^S(t)$. And (B) to decompose it into $\hat{Q}_0^{SID}(t) = \hat{Q}_X^{ID}(t) + \hat{Q}^S(t)$ and $\hat{V}(t) = \hat{Q}_M^I(t) + \hat{Q}_M^D(t)$. Here $\hat{Q}_X^S(t)$ and $\hat{Q}_M^S(t)$ are the exit-rate part and the single-mutation part, respectively, of $\hat{Q}^S(t)$; $\hat{Q}_X^{ID}(t)$, $\hat{Q}_M^I(t)$ and $\hat{Q}_M^D(t)$ are the exit-rate part, the single-insertion part and the single-deletion part, respectively, of $\hat{Q}^{ID}(t)$.

By mainly using the decomposition (A) and by following the same line of arguments as in Sections 3 and 4 of part I, we can show that the probability of an

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

alignment (whether it is a PWA or a MSA) is factorable into the product of an overall factor and the contributions from regions that can potentially accommodate local evolutionary histories, this time including both indels and substitutions. We showed the factorability under the following sufficient set of conditions. (I) In each region, the rate of each of the substitutions and indels is independent of the portions of the extended sequence states in the other regions. (II) In each region, the increment of the total exit rate, $R_X^{SID}(\breve{s},t) = R_X^{S}(\breve{s},t) + R_X^{ID}(\breve{s},t)$, due to each of the substitutions and indels is independent of the portions of the extended states in the other regions. And, if considering a MSA, (III) the factorization of the root sequence state probability (*i.e.*, an extended version of Eq.(4.2.8) in part I) holds. See Appendix A2 for details on the derivation.

The factorization of the probability into the local contributions would definitely be helpful. However, it would be at least equally useful to factorize the probability under the entire evolutionary model into the "basic component" and the "residue" component." Here, the "basic component" is based only on possible indel histories, and the "residue component" is based only on possible substitution histories (and initial residue state distributions). Mainly using the aforementioned decomposition (B), we can show that such an "indel-substitution factorization" can indeed be carried out on the alignment probability if the following four conditions are satisfied. (i) The indel rates are independent of the residue states. (ii) Each finite-time evolution probability of the residue state via substitutions is factorable into the product of site-wise substitution probabilities. (iii) The residue state spectrum of each inserted sub-sequence is factorable into the product of site-wise residue probabilities over the inserted sites. And (iv) the site-wise inserted residue probabilities at each site, $\{p_I(\omega;\upsilon_j,t)\}_{\omega\in\Omega}$, where $\upsilon_j$ is the ancestry of the site, should satisfy the equation:

$$\sum_{\omega'\in\Omega} p_I(\omega';\upsilon_j,t)\langle\omega'|\hat{P}^S(t,t_F;\upsilon_j)|\omega^D(\upsilon_j)\rangle = p_I(\omega^D(\upsilon_j);\upsilon_j,t_F). \quad \text{--- Eq.(1.2)}$$

Here $\hat{P}^S(t,t_F;\upsilon_j)$ is the site-wise stochastic evolution operator via substitutions. When dealing with a MSA, the following equation also needs to be satisfied:

$$\sum_{\omega'\in\Omega} P\left[\left(\omega',n^{Root}\right)|\,\upsilon_i\right]\langle\omega'|\hat{P}^S(n^{Root},n^D(b);\upsilon_i)|\omega^D\rangle = p_I(\omega^D;\upsilon_i,n^D(b)). \quad \text{--- Eq.(1.3)}$$

Here, $P\left[\left(\omega',n^{Root}\right)|\,\upsilon_i\right]$ is the probability of the residue $\omega'$ at the site with ancestry $\upsilon_i$ in the sequence at the root node ($n^{Root}$). See Appendix A3 for details on the proof of the "indel-substitution factorization" under these conditions. Eq.(1.2) means that the inserted residue frequencies must evolve according to the site-wise stochastic evolution via substitutions. And Eq.(1.3) requires that, at every point along the tree, they should coincide with the residue frequencies that would have evolved through substitutions beginning with $\left\{P\left[\left(\omega,n^{Root}\right)|\,\upsilon_i\right]\right\}_{\omega\in\Omega}$ at the root. These equations are a generalized version of the following commonly assumed pair of conditions. (a) The residue frequencies are equilibrium frequencies, $\{\pi(\omega)\}_{\omega\in\Omega}$, that satisfy the detailed-balance conditions, $\sum_{\omega'\in\Omega}\pi(\omega')\langle\omega'|\hat{P}^S(t',t;\upsilon_i)|\omega\rangle = \pi(\omega)$. And (b) the inserted residue frequencies must coincide with these equilibrium frequencies (see, *e.g.*, Thorne et al. 1991, 1992; Miklós et al. 2004; Rivas and Eddy 2008).

In order to pursue the biological realism further, however, the aforementioned conditions (i)-(iv) would be too restrictive, even though the conditions were

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

somewhat relaxed compared to the common practice of imposing (a) and (b) (instead of (iv)). It may be relatively easy to relax the conditions (ii) and (iii) so that the substitution processes could depend on the residue states of neighboring sites, by, *e.g.*, introducing codon models (*e.g.*, Yang 2006) (but see also, *e.g.*, Lunter and Hein 2004; Arndt and Hwa 2005). The violation of the condition (i) might be tackled at least partially, if the indel rates depend on the residue states only locally, through some motifs that sparsely scatter along the sequence. Specifically, in such a case, we may first factorize the probability into the product of local contributions and then perform the "indel-substitution factorization" on the contributions from regions that are likely devoid of such motifs. The violation of the condition (iv) may be more prevalent and serious, especially for large-scale insertions (see, *e.g.*, Waterhouse and Russell 2006; Morgante et al. 2007; Chalopin et al. 2015). Recently, analytical methods where developed for examining the effects of deviation of the inserted residue composition from the substitution-inherent residue composition (Lèbre and Michel 2010, 2013). It may be worth trying to apply some of their methods to the situations where the condition (iv) is violated. Meanwhile, some recent data analyses showed that the substitution rate increases near the sites of indels (Tian et al. 2008; De and Babu 2010). If desired, such effects could be represented in our extended theoretical formulation (described in Appendix A4 in more details), and might be handled similarly to how the violation of the aforementioned condition (i) could be remedied. It remains to be seen whether the remedial methods suggested here actually work, or otherwise whether our formulation could be modified or further extended somehow to efficiently deal with the more biologically realistic features mentioned above.

## 2. Incorporation of inversions and duplications

At least formally, the theoretical formulation developed in this paper can be extended to incorporate other genomic rearrangements (*e.g.*, Gascuel 2005; Gu et al. 2008). Here, we discuss the incorporation of inversions (*e.g.*, Kelshner and Wendel 1996; Graham et al. 2000) and duplications (*e.g.*, Bailey and Eichler 2006; Lynch 2007; Ezawa et al. 2011), as two most important examples.

To incorporate inversion processes, it is convenient to extend the space state, especially $S^{II}$ or $S^{III}$, to accommodate the complement of each site. Specifically, we let a superscript "$C$" indicate that the state is of a site on the complementary strand of the site before inversion. For example, $\upsilon^C$ denotes the ancestry of the site complementary to a site with the ancestry $\upsilon$ (in the space $S^{II}$). And $(\sigma,\xi)^C$ denotes the attributes of the site complementary to a site with attributes $(\sigma,\xi)$ (in the space $S^{III}$). And we consider that the complement of the complement is the original: $\left(X^C\right)^C = X$. To incorporate duplication processes, we also introduce another indicator, $\chi$, telling that the site is on the $\chi$ th copy of the subsequence. For example, $\upsilon.\chi$ represents the ancestry of the $\chi$ th copy of the original site with the ancestry $\upsilon$ (in the space $S^{II}$), and $(\sigma,\xi,\chi)$ represents the attribute of the $\chi$ th copy of the original site with the attribute $(\sigma,\xi)$ (in the space $S^{III}$). The state spaces formed by extending the spaces $S^{II}$ and $S^{III}$ this way will be represented by $S^{IIe}$ and $S^{IIIe}$, respectively. [NOTE: As for the state space $S^{I}$, we do not need to extend it to accommodate inversions and duplications; we just invert or duplicate a sub-array of *unlabeled* sites of a sequence, as well as the *residue states* filling in the sub-array.]

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

An inversion event could be depicted by an inversion operator, $\hat{M}_V(x_B, x_E)$, which inverts a sub-array of sites between (and including) the $x_B$ th and $x_E$ th sites. For example, the action of $\hat{M}_V(2,4)$ on the basic state $s = [(1,1),(1,2),(2,1),(2,2,1),(2,2,2)]$ ($\in S^{IIIe}$) is expressed as:

$$\begin{aligned}&\langle[(1,1),(1,2),(2,1),(2,2,1),(2,2,2)]|\hat{M}_V(2,4)\\&= \langle[(1,1),(2,2,1)^C,(2,1)^C,(1,2)^C,(2,2,2)]|.\end{aligned} \quad \text{--- Eq.(2.1a)}$$

In principle, we could also define the inversion of a region sticking out of either sequence end. For example, the action of $\hat{M}_V(0,2)$ on the above state could be represented by something like the following:

$$\begin{aligned}&\langle[(1,1),(1,2),(2,1),(2,2,1),(2,2,2)]|\hat{M}_V(0,2)\\&= \langle[(1,1)^C,(3,1),(2,1),(2,2,1),(2,2,2)]|.\end{aligned} \quad \text{--- Eq.(2.1b)}$$

The point is that this operation replaces the 2nd site with the complement of the 0th site. The 0th site was outside of the region under consideration, and we formally assigned it a new attribute, $(3,1)^C$, before inversion. Thus, an inversion sticking out of either sequence end is effectively equivalent to a simultaneous operation of a deletion and an insertion, and also a smaller-scale inversion when the inverted region within the sequence is longer than the region sticking out. The inversion rate operator, $\hat{Q}^V(t)$, is also defined similarly to how the indel rate operator is defined via Eqs.(2.4.1a,b) of part I (Ezawa, Graur and Landan 2015a):

$$\langle s|\hat{Q}^V(t) = \left[\sum_{x_B=-\infty}^{L(s)}\sum_{x_E=\max\{x_B,1\}}^{+\infty} r_V(x_B,x_E;s,t)\langle s|\hat{M}_V(x_B,x_E)\right] - R_X^V(s,t)\langle s|, \quad \text{--- Eq.(2.2a)}$$

with the exit rate:

$$R_X^V(s,t) = \sum_{x_B=-\infty}^{L(s)}\sum_{x_E=\max\{x_B,1\}}^{+\infty} r_V(x_B,x_E;s,t)\ . \quad \text{--- Eq.(2.2b)}$$

If the inversion rates, $r_V(x_B,x_E;s,t)$, are space-homogeneous, the exit rate will be an affine function of the sequence length and the probability of a LHS equivalence class of inversion processes will be factorable. Unfortunately, inversions are known to occur preferentially on palindrome sequences or between inverted repeats (*e.g.*, Kelshner and Wendel 1996; Gu et al. 2008). Nevertheless, even if we take account of such structural dependence, the probability may be more or less factorable. This is because a simple inversion does not change the sequence length or much of the inverted repeat structure, and thus because $R_X^V(s,t)$ is expected to change little due to inversions.

A duplication event could be depicted via a duplication operator, $\hat{M}_C(+,x,[x_B,x_E])$ or $\hat{M}_C(-,x,[x_B,x_E])$, which copies the sub-array between (and including) the $x_B$ th and $x_E$ th sites and inserts the copy between the $x$ th and $(x+1)$ th sites. The "$+$" and "$-$" in the 1st argument represent the insertion on the original and complementary strands, respectively. For example, the actions of $\hat{M}_C(+,1,[3,5])$ and $\hat{M}_C(-,1,[3,5])$ on the above state could be expressed as:

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\begin{aligned}&\left\langle\left[(1,1),(1,2),(2,1),(2,2,1),(2,2,2)\right]\middle|\hat{M}_C(+,1,[3,5])\right.\\&=\left\langle\left[(1,1),(2,1,2),(2,2,3),(2,2,4),(1,2),(2,1,1),(2,2,1),(2,2,2)\right]\right|,\end{aligned} \quad \text{--- Eq.(2.3a)}$$

and

$$\begin{aligned}&\left\langle\left[(1,1),(1,2),(2,1),(2,2,1),(2,2,2)\right]\middle|\hat{M}_C(-,1,[3,5])\right.\\&=\left\langle\left[(1,1),(2,2,4)^C,(2,2,3)^C,(2,1,2)^C,(1,2),(2,1,1),(2,2,1),(2,2,2)\right]\right|,\end{aligned} \quad \text{--- Eq.(2.3b)}$$

respectively. Again in principle, we could also define the duplication of a region sticking out of either sequence end. For example, the action of $\hat{M}_C(+,3,[0,2])$ on the above state could be depicted as:

$$\begin{aligned}&\left\langle\left[(1,1),(1,2),(2,1),(2,2,1),(2,2,2)\right]\middle|\hat{M}_C(+,3,[0,2])\right.\\&=\left\langle\left[(1,1,1),(1,2,1),(2,1),(3,1),(1,1,2),(1,2,2),(2,2,1),(2,2,2)\right]\right|.\end{aligned} \quad \text{--- Eq.(2.3c)}$$

Again, the 0th site was out of consideration before the event, and thus was assigned a new attribute, $(3,1)$, when inserted. In general, the duplication of a region sticking out of either sequence end is effectively equivalent to the simultaneous operation of a duplication of the region within the sequence and an adjacent insertion of a new subsequence. Furthermore, we could define the duplication of a region totally out of the sequence under consideration, *i.e.*, $r_C(\varepsilon,x,[x_B,x_E])$ with $\varepsilon\in\{+,-\}$ and with either $x_B\le x_E<1$ or $L(s)<x_B\le x_E$. Under the current formulation, its effect is indistinguishable from that of the insertion, $r_I(x,x_E-x_B+1)$. Using these duplication operators, the duplication rate operator, $\hat{Q}^C(t)$, could be defined as:

$$\begin{aligned}\langle s|\hat{Q}^C(t)=&\sum_{\varepsilon\in\{+,-\}}\sum_{x=0}^{L(s)}\sum_{x_B=-\infty}^{+\infty}\sum_{x_E=x_B}^{+\infty}r_C(\varepsilon,x,[x_B,x_E])\langle s|\hat{M}_C(\varepsilon,x,[x_B,x_E])\\&-R_X^C(s,t)\langle s|\quad,\end{aligned} \quad \text{--- Eq.(2.4a)}$$

with the duplication exit rate:

$$R_X^C(s,t)=\sum_{\varepsilon\in\{+,-\}}\sum_{x=0}^{L(s)}\sum_{x_B=-\infty}^{+\infty}\sum_{x_E=x_B}^{+\infty}r_C(\varepsilon,x,[x_B,x_E])\ . \quad \text{--- Eq.(2.4b)}$$

If the duplication rates, $r_C(\varepsilon,x,[x_B,x_E])$, are space-homogeneous, $R_X^C(s,t)$ should become an affine function of the sequence length, and the probability of a LHS equivalence class of duplication processes could be factorable. But, again, it is unlikely that the duplication rates are space-homogeneous, because duplications preferentially occur between direct repeat motifs (*e.g.*, Bailey and Eichler 2006; Gu et al. 2008). It remains to be seen whether the probability is still factorable or not even after taking account of this factor. But some situations with interspersed repeat motifs may be modeled as in Eqs.(5.3.2a,b,c) in part I, at least to some degree, and thus the probabilities may be partially factorable.

Some transposon insertions (*e.g.*, Morgante et al. 2007; Chalopin et al. 2015) are essentially duplications (via copy-and-paste mechanisms). Thus, in some cases it might be beneficial to regard the transposons as duplicated using the formulation explained here, rather than handling them as simple insertions. On the other hand, the latter could be done in principle by associating particular sets of residue configurations with elevated insertion rates, via the theoretical formulation briefed in Section 1. Their description as duplications would be beneficial especially when two

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

or more transposons belonging to the same family were inserted into positions that are close to each other, because they could induce secondary genomic rearrangements.

In a traditional alignment (PWA or MSA), an inversion should manifest itself as a pair of equally long gapped regions, interpreted as a deletion and an adjacent insertion, and it will normally be penalized twice. Moreover, commonly used aligners will ignore the fact that the reverse complement of the inverted region can be well aligned with its original region. Meanwhile, a duplication event should cause a gap in a normal alignment, and will be interpreted as an insertion. However, unless we take account of the fact that the duplicated region is in fact homologous to, and thus can be aligned with, its original region, the resulting alignment could often be erroneous. Therefore, by taking explicit account of inversions and deletions when reconstructing an alignment, its accuracy might improve substantially. This kind of attempts has only a short history (see, *e.g*, Paten et al. 2008b, and references therein). The theoretical formulation briefed in this subsection may be conducive to the development of likelihood-based (or Bayesian-based) alignment programs that also take account of genomic rearrangements other than indels. To do so, the rate parameters, such as $r_V(x_B, x_E; s, t)$ and $r_C(\varepsilon, x, [x_B, x_E])$, will have to be estimated accurately. It would probably be difficult to estimate them directly from the sequence data to be analyzed, because the events are relatively rare. Thus, it would be helpful to estimate the parameter distribution, or relative rates as functions of the length of the duplicated/inverted region, the distance between the original region and the copy-insertion point, residue states of the flanking regions, etc., by analyzing genome-wide data from a large sample of organisms (or individuals from populations). Although the broad dependences of the duplication frequency on the original-copy distance and the orientation were examined in the past (*e.g.*, Ezawa et al. 2011), more extensive and thorough analyses will be necessary.

At the population genetic level, genome rearrangements are observed as genomic structural variations (SVs), including copy number variations (CNVs) (*e.g.*, Teshima and Innan 2012; Ezawa et al. 2013). In some aspects, population genetics could be regarded as molecular evolution on a very short time scale. Thus, the theoretical formulation unfolded in this paper may be applicable, possibly via some modifications, to the analyses of SVs as well.

## 3. Other mutational mechanisms

Mechanisms of genomic mutations are not limited to substitutions and genome rearrangements. Among the most important ones would be recombination (*e.g.*, Saitou and Kitano 2013) and gene conversion (*e.g.*, Chen et al. 2007; Ezawa et al. 2010; Fawcett and Innan 2013). Out of them, inter-locus gene conversion (*e.g.*, Ezawa et al. 2010; Fawcett and Innan 2013) may be depicted via something similar to the duplication operator proposed in Section 2, but with a modification. Specifically, instead of inserting a sub-sequence, a gene-conversion operator must make a sub-sequence replace another region that is usually paralogous. This description will work as long as the interacting regions are both within a single sequence to be analyzed. However, recombination between alleles (*e.g.*, Saitou and Kitano 2013), including inter-allelic gene conversion (*e.g.*, Chen et al. 2007), is a mechanism involving two orthologous sequences, and thus the above description will not be naïvely applicable to it. The description will not be applicable to inter-locus gene conversion, either, if one of the interacting regions is out of the subject sequence. A possible way to handle these cases would be to allow the phylogenetic tree to take different topologies and/or branch lengths in different regions of the sequence, as implemented in the molecular

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

evolution simulator, Dawg (Cartwright 2005). This measure should work especially when the alignment probability is factorable.

The copy number change of short tandem repeats, or microsatellites, is another important mutational mechanism (*e.g.*, Ellegren 2004; Sainudiin et al. 2004). However, the evolution of microsatellites would be quite refractory to a naïve analysis, because they have quite high mutation rates and show complex mutation patterns (*e.g.*, Ellegren 2004; Sainudiin et al. 2004). Thus, unless we are handling very short-term sequence evolution, we should avoid using the perturbation theory developed in this paper. Instead, we should try to solve "exactly," maybe via a numerical computation, an empirical evolutionary model of microsatellite evolution (*e.g.*, Sainudiin et al. 2004). Our current hunch is that, as far as an accurate alignment reconstruction is concerned, a best way would be to remove alleged microsatellites from the sequences before aligning them.

**4. Boundary conditions and cut-off lengths**

In this study, as in the previous studies (Ezawa, Graur and Landan 2015a,b,c), we only considered simple boundary conditions. Each sequence end was either freely mutable or flanked by a biologically essential region that allows no indels. Moreover, the constant cutoff lengths were introduced just for the sake of simplicity, to broadly take account of the effects of various factors that suppress very long rearrangements (such as selection, chromosome size, genome stability, etc.). In real sequence analyses, however, the situations are unlikely to be so simple. (See Discussion of part I (Ezawa, Graur and Landan 2015a) for more details.) In order to pursue further biological realism and to enable further accurate sequence analyses, it would be inevitable to address these issues seriously.

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## Conclusions

In a previous study (Ezawa, Graur, and Landan 2015a), we established the theoretical basis of an *ab initio* perturbative formulation of a general continuous-time Markov model, which is a *genuine* stochastic model describing the evolution of an *entire* sequence via indels along the time axis. Then, in two other previous studies (Ezawa, Graur, and Landan 2015b,c), we demonstrated how we can analytically and computationally calculate the probabilities of concrete alignments via our formulation. In these previous studies, we dealt only with insertions/deletions (indels), mainly for clarity of the arguments.

Here in this study, we attempted to incorporate other types of mutations into our *ab initio* perturbative formulation of a *genuine* evolutionary model. We first extended the model to accommodate substitutions. We showed that, under a set of conditions on the model parameters similar to that for the pure indel model, the probability of an alignment is factorable into local contributions also in this extended model. We also gave a set of conditions under which an alignment probability is factorable into the product of an "indel component" and a "substitution component." We next showed how our evolutionary model can be extended to accommodate two other types of genomic rearrangements, namely, inversions and duplications. We also discussed how to handle other types of mutations such as recombination within the framework of our formulation. Thus, at least in principle, our *ab initio* perturbative formulation could describe the evolution of an *entire* sequence via any major types of mutations along the time axis. It still remains to be seen whether all of these suggested extensions are *computationally* feasible and useful for *practical* sequence analyses, such as the reconstruction of generalized multiple sequence alignments (MSAs). If this proves to be the case, however, our formulation could open up a new venue for the theoretical study of sequence evolution.

## Authors' contributions

KE conceived of and mathematically formulated the theoretical framework in this paper, implemented the key algorithms, participated in designing the study, performed all the mathematical analyses, and drafted the manuscript. DG and GL participated in designing the study, helped with the interpretation of the data, and helped with the drafting of the manuscript.

## Acknowledgements

This study is dedicated to the late Dr. Keiji Kikkawa, who was a renowned theoretical physicist, one of the key pioneers of the string field theory of the elementary particle physics, and the best ever mentor of K.E. We are grateful to Dr. R. A. Cartwright at Arizona State University for his useful information and discussions that inspired this study. We appreciate the logistic support and the feedback of Dr. Tetsushi Yada at the Kyushu Institute of Technology. We would also like to thank the three anonymous referees of the predecessor manuscript entitled: "Framework that enables approximate lilelihood analysis of insertions/deletions on multiple sequence alignment." Their comments helped drastically improve the study itself, not to mention the manuscript. This work was a part of the project, "Error Correction in Multiple Sequence Alignments," which was funded by US National Library of Medicine [grant number LM010009-01 to Dan Graur and Giddy Landan at the University of Houston]. The

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

later stage of this work was also supported by Grants-in-Aid No. 221S0002, which was awarded to Tetsushi Yada by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Appendix

Here we give detailed arguments regarding the incorporation of substitutions into our theoretical formulation, as briefly described in Section 1 of Results. We first explain how to extend our formulation to incorporate substitutions in Section A1. Then, in Section A2, we will examine the conditions under which the probability of each alignment is factorable into the product of contributions from regions delimited by preserved ancestral sites (PASs). Next, in Section A3, we will examine the conditions under which the probability of each alignment can be factorized into the "basic" component concerning the indels and the residue component concerning the substitutions (and the initial residue state distribution at each site). Finally, in Section A4, we will discuss how to pursue the biological realism further.

**A1. Formulating sequence evolution via both indels and substitutions**

In order to incorporate substitutions into our theoretical formulation of sequence evolution via indels, we first extend the sequence state space $S$ $(= S^I, S^{II}, or\ S^{III})$ so that each site of the sequence will accommodate a residue in the set $\Omega$. (See Subsection 2.1 of part I (Ezawa, Graur and Landan 2015a) for details on the original state spaces $S^I, S^{II}$ *and* $S^{III}$.) When $S = S^I$ $(\cong \mathrm{N}_0 \equiv \{0,1,2,...\})$, the extended space is $\breve{S} = \Omega^* \equiv \bigcup_{L=0}^{\infty} \Omega^L$, in which each site is assigned a residue $\omega \in \Omega$. When $S = S^{II}$ $(\subset \Upsilon^*)$, the ancestry assigned to each site, $\upsilon \in \Upsilon$, is replaced with a pair, $(\upsilon, \omega) \in \Upsilon \times \Omega$. Thus, the extended space is $\breve{S}^{II} \subset \{\Upsilon \times \Omega\}^*$. Similarly, when $S = S^{III}$ $(\subset \{\mathrm{N}_0 \times \mathrm{N}_1\}^*)$, the extended space is $\breve{S}^{III} \subset \{\mathrm{N}_0 \times \mathrm{N}_1 \times \Omega\}^*$, in which each site is assigned a trio, $(\sigma, \xi, \omega) \in \mathrm{N}_0 \times \mathrm{N}_1 \times \Omega$. (Here $\sigma$ is the source identifier of the initial or inserted (sub)sequence, and $\xi$ is the relative coordinate within the initial or inserted (sub)sequence.) An extended sequence state $\breve{s} \in \breve{S}$ will sometimes be represented as $\breve{s} = (s, \vec{\omega})$ to explicitly show that it is composed of a "basic" component $s \in S$ and a residue component $\vec{\omega} = [\omega_1, ..., \omega_{L(\breve{s})}] \in \Omega^{L(\breve{s})}$. (Here $L(\breve{s})$ is the length of the sequence state $\breve{s}$.) Once we extended the sequence state space, we can now extend the indel rate operator (defined in Eqs.(2.4.2a-d) of part I). This is done in two steps: (1) adding the substitution component to the rate operator; and (2) extending the indel component to take explicit account of the residue dependence of indels.

First, we add the substitution component of the rate operator, $\hat{Q}^S(t)$, to the indel component, $\hat{Q}^{ID}(t)$ (defined in Eqs.(2.4.2a-d) of part I). This yields the total mutation rate operator:

$$\hat{Q}^{SID}(t) \equiv \hat{Q}^{ID}(t) + \hat{Q}^S(t). \quad \text{--- Eq.(A1.1)}$$

Similarly to the derivation of Eqs.(2.4.2a-d) of part I, the substitution component is written as $\hat{Q}^S(t) \equiv \sum_{\breve{s} \in \breve{S}} |\breve{s}\rangle\langle\breve{s}| \hat{Q}^{S(L(\breve{s}))}(t)$, where $\hat{Q}^{S(L)}(t)$ is the substitution rate operator defined on a sequence of length $L$. If we allow only for a single residue substitution at a time, $\hat{Q}^{S(L)}(t)$ is expanded as: $\hat{Q}^{S(L)}(t) = \sum_{x=1}^{L} \hat{Q}^S(x,t)$, where $\hat{Q}^S(x,t)$ is the operator representing the effect of a substitution at the $x$ th site at time $t$. Let $\hat{M}_S(x, \omega \mapsto \omega')$ be the operator representing the substitution from the residue

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$\omega$ to $\omega'(\neq\omega)$ at the $x$ th site. And let $\omega_x(\breve{s})$ be the residue state at the $x$ th site of a sequence with the extended state $\breve{s}$. Then, in general, $\hat{Q}^S(x,t)$ is defined by the following action on the extended sequence states $\breve{s}\in\breve{S}$ with $L(\breve{s})\geq x$:

$$\langle\breve{s}|\hat{Q}^S(x,t)=\left[\sum_{\substack{\omega'\in\Omega,\\ \omega'\neq\omega_x(\breve{s})}} r_S\left(x,\omega_x(\breve{s})\mapsto\omega';\breve{s},t\right)\langle\breve{s}|\hat{M}_S(x,\omega_x(\breve{s})\mapsto\omega')\right]-R_X^S(x;\breve{s},t)\langle\breve{s}|. \qquad \text{--- Eq.(A1.2)}$$

Here the position-specific exit rate, $R_X^S(x;\breve{s},t)\equiv\sum_{\omega'\in\Omega,\,\omega'\neq\omega_x(\breve{s})} r_S\left(x,\omega_x(\breve{s})\mapsto\omega';\breve{s},t\right)$, is made of the substitution rates at the $x$ th site, $r_S\left(x,\omega_x(\breve{s})\mapsto\omega';\breve{s},t\right)$, which could in general depend on the extended sequence state, site position, and time. In the special case where all indel rates are zero, the sequence states keep their length and the ancestries $\upsilon_x$ (or attributes $(\sigma_x,\xi_x)$) of the sites intact. Thus, the total stochastic evolution operator of the system ($\hat{P}^{SID}(t,t')$) is given solely by its substitution component, $\hat{P}^S(t,t')=T\left\{\exp\left(\int_t^{t'}d\tau\,\hat{Q}^S(\tau)\right)\right\}$. (Here $T\{...\}$ denotes the time-ordered product. See below Eq.(1.1.11) of part I for details.) If desired, we can decompose the substitution rate operator as $\hat{Q}^S(t)=\hat{Q}_M^S(t)+\hat{Q}_X^S(t)$. Here, $\hat{Q}_M^S(t)$ is a linear combination of substitution operators, and $\hat{Q}_X^S(t)$ keeps the residue state unchanged while letting the probability decay at the exit rate. Then, in the same line of reasoning as in Subsection 4.1 of part I, the conditional probability $P\left[(s',t')\,|\,(s,t)\right]=\langle s|\hat{P}^S(t,t')|s'\rangle$ can be factorized into the product of multiplication factors, each of which is contributed from a region that is independent of the others. Actually, in this case, the stochastic evolution operator ($\hat{P}^S(t,t')$) itself becomes factorable into a tensor product, as long as no mutation changes the interdependence structure among the substitution rates at different sites. Especially, when each of the substitution rates depends only on the state at the site of the substitution, *e.g.*, $r_S\left(x,\omega_x(\breve{s})\mapsto\omega';\breve{s},t\right)=g_S\left(\omega_x(\breve{s}),\omega';\upsilon_x(\breve{s}),t\right)$ (when $\breve{S}=\breve{S}^{II}$), we have:

$$\hat{P}^{S(L)}(t,t')=\bigotimes_{x=1}^{L}\hat{P}^S(t,t';x). \qquad \text{--- Eq.(A1.3)}$$

Here, $\hat{P}^S(t,t';x)\equiv T\left\{\exp\left(\int_t^{t'}d\tau\,\hat{Q}^S(x,\tau)\right)\right\}$, supplemented by Eq.(A1.2), is the stochastic operator describing the evolution of the $x$ th site via substitutions during the time interval $[t,t']$. At least conceptually, Eq.(A1.3) should be familiar to many researchers of molecular evolution (*e.g.*, Felsenstein 1981, 2004; Yang 2006).

Second, we extend the indel component of the rate operator, $\hat{Q}^{ID}(t)$ (given in Eqs.(2.4.2a-d) of part I), so that it will take *explicit* account of the possible residue state dependence of indels. First of all, we replace the summations, such as $\sum_{s\in S}(...)$ and $\sum_{s\in S}|s\rangle(...)\langle s|$, with the extended ones, *i.e.*, $\sum_{\breve{s}\in\breve{S}}(...)$ and $\sum_{\breve{s}\in\breve{S}}|\breve{s}\rangle(...)\langle\breve{s}|$. Then, consider the operations of indels on the extended state $\breve{s}\in\breve{S}$, as well as their rates. Regarding deletions, it is easy; we just need to extend the deletion rates $r_D(x_B,x_E;s,t)$ to $r_D(x_B,x_E;\breve{s},t)$, which explicitly depend also on the residue state component

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$\vec{\omega} \in \Omega^{L(\breve{s})}$ of the extended state $\breve{s}$. Regarding insertions, however, we need something more. We first introduce a new operator, $\hat{F}(x, \delta\vec{\omega}'[l])$, that fills in an array of newly inserted sites, from the $(x+1)$ th through the $(x+l)$ th sites in the post-insertion sequence, with an array of residues, $\delta\vec{\omega}'[l] \equiv [\omega'_{x+1}, ..., \omega'_{x+l}] \in \Omega^l$. In other words, $\hat{F}(x, \delta\vec{\omega}'[l])$ inserts $\delta\vec{\omega}'[l]$ between the $x$ th and $(x+1)$ th residues of $\vec{\omega} \in \Omega^{L(\breve{s})}$. Then, we replace the term $r_I(x,l;s,t)\langle s|\hat{M}_I(x,l)$ in Eq.(2.4.2b) of part I with $\sum_{\delta\vec{\omega}'[l]\in\Omega^l} r_I(x,l,\delta\vec{\omega}'[l];\breve{s},t)\langle\breve{s}|\hat{M}_I(x,l)\hat{F}(x,\delta\vec{\omega}'[l])$. And we also replace $r_I(x,l;s,t)$ in the exit rate (Eq.(2.4.1b) or Eq.(2.4.1b') of part I) with $r_I(x,l;\breve{s},t) \equiv \sum_{\delta\vec{\omega}'[l]\in\Omega^l} r_I(x,l,\delta\vec{\omega}'[l];\breve{s},t)$. If there is no correlation between the $\delta\vec{\omega}'[l] \in \Omega^l$ to be inserted and the state $\breve{s}$ to undergo the insertion, the dependence of the rates on them could be decoupled as $r_I(x,l,\delta\vec{\omega}'[l];\breve{s},t) = r_I(x,l;\breve{s},t)\, p_I(\delta\vec{\omega}'[l];t|l)$. Here $p_I(\delta\vec{\omega}'[l];t|l)$ is the probability (or the relative frequency) that $\delta\vec{\omega}'[l]$ is inserted at time $t$, conditioned on the insertion of $l$ sites. (The conditional probabilities satisfy $\sum_{\delta\vec{\omega}'[l]\in\Omega^l} p_I(\delta\vec{\omega}'[l];x,t|l) = 1$.) This implies that $r_I(x,l;s,t)$ used in the bulk of the paper were something like $r_I(x,l;\breve{s},t)$ above, considering that the former already *implicitly* depended on the residue states of the sequence before insertion. If the above decoupling holds, and if the dependence on the inserted residues can be factorized as $p_I(\delta\vec{\omega}'[l];t|l) = \prod_{i=1}^{l} \pi(\omega'_{x+i})$, then the inserted residues could be handled as were done in the past (*e.g.*, Thorne et al. 1991, 1992; Miklós et al. 2004; Rivas and Eddy 2008). Here $\pi(\omega)$ is the equilibrium frequency of a residue $\omega$ (with $\sum_{\omega\in\Omega} \pi(\omega) = 1$) under a time-reversible substitution model.

**A2. Factorizing probability into regional contributions**

The perturbation theory unfolded in this paper can also be applied, with some extensions, to the entire model defined by the rate operator $\hat{Q}^{SID}(t)$ in Eq.(1.1) of Results (*i.e.*, Eq.(A1.1)). Because indels are much less frequent than substitutions, a natural way would be to separate $\hat{Q}^{SID}(t)$ as:

$$\hat{Q}^{SID}(t) = \hat{Q}_0^{SID}(t) + \hat{Q}_M^{ID}(t). \qquad \text{--- Eq.(A2.1a)}$$

Here

$$\hat{Q}_0^{SID}(t) \equiv \hat{Q}_X^{ID}(t) + \hat{Q}^S(t) \qquad \text{--- Eq.(A2.1b)}$$

describes the sequence evolution via no indels. And $\hat{Q}_M^{ID}(t) = \hat{Q}_M^I(t) + \hat{Q}_M^D(t)$, which is the aforementioned extension of Eq.(3.1.1c) of part I (Ezawa, Graur and Landan 2015a), describes the sequence change via an insertion or a deletion. The point is that the *entire* substitution rate operator ($\hat{Q}^S(t)$) is included in the "unperturbed" part of the rate operator ($\hat{Q}_0^{SID}(t)$), and that *only* insertions/deletions are regarded as "perturbations." Then, the perturbation expansion of the *entire* stochastic evolution operator, $\hat{P}^{SID}(t_I,t_F) \equiv T\left\{\exp\left(\int_{t_I}^{t_F} dt\, \hat{Q}^{SID}(t)\right)\right\}$, is given by Eq.(3.1.3) of part I with the

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

extension of the state space and with

$$\hat{P}_0^{ID}(t,t') \equiv T\left\{\exp\left(\int_t^{t'} d\tau\, \hat{Q}_0^{ID}(\tau)\right)\right\} = \exp\left(\int_t^{t'} d\tau\, \hat{Q}_X^{ID}(\tau)\right) \text{ replaced by}$$

$$\hat{P}_0^{SID}(t,\ t') \equiv T\left\{\exp\left(\int_t^{t'} d\tau\, \hat{Q}_0^{SID}(\tau)\right)\right\} = T\left\{\exp\left(\int_t^{t'} d\tau\, \hat{Q}_X^{ID}(\tau) + \int_t^{t'} d\tau\, \hat{Q}^{S}(\tau)\right)\right\}.$$

--- Eq.(A2.2)

The Eq.(3.1.13) of part I for the probability of a PWA between an ancestral and a descendant sequence, $P\left[\left(\alpha(s^A,s^D),[t_I,t_F]\right)\middle|\,(s^A,t_I)\right]$, can also be extended as follows:

$$P\left[\left(\alpha(\breve{s}^A,\breve{s}^D),[t_I,t_F]\right)\middle|\,(\breve{s}^A,t_I)\right]$$

$$= \sum_{\substack{[\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N] \\ \in \tilde{\mathrm{H}}^{ID}[\alpha(s^A,s^D)]}} \sum_{\vec{\omega}_1\in\Omega^{L_1}} \cdots \sum_{\vec{\omega}_N\in\Omega^{L_N}} P\left[\left(\begin{matrix}[\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N], \\ [\vec{\omega}_1,\vec{\omega}_2,\ldots,\vec{\omega}_N],[t_I,t_F]\end{matrix}\right),(\vec{\omega}^D,t_F)\middle|\,(s^A,\vec{\omega}^A,t_I)\right].$$

--- Eq.(A2.3a)

Here we let $\breve{s}_\nu \in \breve{S}$ ($\nu = 1,2,\ldots,N$) be the extended sequence state "at the time of" the event $\hat{M}_\nu$, or, more precisely, the state immediately before and after the event if the event is a deletion and an insertion, respectively. We also let $L_\nu = L(\breve{s}_\nu)$ be the number of sites in $\breve{s}_\nu$. Let $\vec{\omega}_\nu \in \Omega^{L_\nu}$ and $s_\nu \in S$, respectively, be the residue state component and the basic component of $\breve{s}_\nu$. And let $\breve{s}(t_\nu^{(+)})$ and $\breve{s}(t_\nu^{(-)})$, respectively, denote the extended sequence states immediately after and before the time $t_\nu$ of the event $\hat{M}_\nu$ within an evolutionary process. Then, we have $\breve{s}_\nu = \breve{s}(t_\nu^{(-)})$ and $\breve{s}_\nu = \breve{s}(t_\nu^{(+)})$, respectively, when $\hat{M}_\nu$ is a deletion and insertion. This choice of $\breve{s}_\nu$ automatically takes account of the summations over $\delta\vec{\omega}_\nu[l] \in \Omega^l$ filled in by the accompanying operator $\hat{F}(x,\delta\vec{\omega}'[l])$ when $\hat{M}_\nu = \hat{M}_I(x,l)$. Moreover, we also used the representation $\breve{s}^A = (s^A,\vec{\omega}^A) \in \breve{S}$, with $s^A \in S$ and $\vec{\omega}^A \in \Omega^{L(\breve{s}^A)}$, and, similarly, $\breve{s}^D = (s^D,\vec{\omega}^D)$. Each summand on the right hand side of Eq.(A2.3a) represents the probability, conditioned on an ancestral state $\breve{s}^A \in \breve{S}$ at time $t_I$, that we have the following 3 outcomes at the same time: (1) an indel history $[\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N]$ consistent with $\alpha(s^A,s^D)$ occurred during the time interval $[t_I,t_F]$; (2) an array of residue states $\vec{\omega}_\nu$ was observed "at the time of" each event $\hat{M}_\nu$; and (3) a descendant array of residue states $\vec{\omega}^D$ resulted at time $t_F$. The summand is an extension of Eq.(3.1.8b) of part I and is specifically expressed as:

$$P\left[\left([\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N],[\vec{\omega}_1,\vec{\omega}_2,\ldots,\vec{\omega}_N],[t_I,t_F]\right),(\vec{\omega}^D,t_F)\middle|\,(s^A,\vec{\omega}^A,t_I)\right]$$

$$= \int_{t_I=t_0<t_1<\cdots<t_N<t_{N+1}=t_F}\!\!\cdots\!\int dt_1\cdots dt_N \left[\begin{matrix}\left(\prod_{\nu=1}^{N} r(\hat{M}_\nu;\breve{s}_\nu,t_\nu)\right) \\ \times\left(\prod_{\nu=0}^{N}\left\langle \breve{s}(t_\nu^{(+)})\middle|\hat{P}_0^{SID}(t_\nu,t_{\nu+1})\middle|\breve{s}(t_{\nu+1}^{(-)})\right\rangle\right)\end{matrix}\right]\Bigg|_{\substack{\left\{s(t_{\nu+1}^{(-)})=s(t_\nu^{(+)})\ \middle|\ \nu=0,\ldots,N\right\}, \\ \left\{\left\langle\breve{s}(t_\nu^{(+)})\right|=\left\langle\breve{s}(t_\nu^{(-)})\right|\hat{M}_\nu\hat{F}_\nu\ \middle|\ \nu=1,\ldots,N\right\}}}.$$

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

--- Eq.(A2.3b)

Here $r(\hat{M}_\nu;\breve{s}_\nu,t_\nu)$ denotes $r_I(x,l,\delta\vec{\omega}_\nu[l];\breve{s}(t_\nu^{(-)}),t_\nu)$ and $r_D(x_B,x_E;\breve{s}(t_\nu^{(-)}),t_\nu)$, respectively, if $\hat{M}_\nu$ is $\hat{M}_I(x,l)$ and $\hat{M}_D(x_B,x_E)$. (Here $\delta\vec{\omega}_\nu[l]\in\Omega^l$ denotes an array of residues filling in the sites inserted by $\hat{M}_\nu=\hat{M}_I(x,l)$. When $\hat{M}_\nu=\hat{M}_I(x,l)$, the combination of $\breve{s}(t_\nu^{(-)})$, the insertion position ($x$), and $\delta\vec{\omega}_\nu[l]$ holds information equivalent to that of $\breve{s}(t_\nu^{(+)})=\breve{s}_\nu$.) We also used the notation, $\hat{F}_\nu=\hat{F}(x,\delta\vec{\omega}_\nu[l])$ if $\hat{M}_\nu=\hat{M}_I(x,l)$, and $\hat{F}_\nu=\hat{I}$ (*i.e.*, does nothing) if $\hat{M}_\nu=\hat{M}_D(x_B,x_E)$. The factor $\left\langle\breve{s}(t_\nu^{(+)})\middle|\hat{P}_0^{SID}(t_\nu,t_{\nu+1})\middle|\breve{s}(t_{\nu+1}^{(-)})\right\rangle$ $\left(=\left\langle\breve{s}(t_\nu^{(+)})\middle|T\left\{\exp\left(\int_{t_\nu}^{t_{\nu+1}}dt\,\hat{Q}_X^{ID}(t)+\int_{t_\nu}^{t_{\nu+1}}dt\,\hat{Q}^S(t)\right)\right\}\middle|\breve{s}(t_{\nu+1}^{(-)})\right\rangle\right)$ is an extension of $\exp\left\{-\int_{t_\nu}^{t_{\nu+1}}dt\,R_X^{ID}(s(t_\nu^{(+)}),t)\right\}$ $\left(=\left\langle s(t_\nu^{(+)})\middle|T\left\{\exp\left(\int_{t_\nu}^{t_{\nu+1}}dt\,\hat{Q}_X^{ID}(t)\right)\right\}\middle|s(t_{\nu+1}^{(-)})\right\rangle\right)$ in Eq.(3.1.8b) of part I. (Here, the notation of the latter was changed from the original, to fit in context.) It should be noted that, in the evolutionary process under consideration, the basic state $s(t_\nu^{(+)})$ $(=s(t_{\nu+1}^{(-)}))$ remained unchanged during the open time interval $(t_\nu,t_{\nu+1})$. Thus, during this time interval, the evolutionary changes of the extended state $\breve{s}(t)=(s(t),\vec{\omega}(t))$ are limited to changes in the residue state $\vec{\omega}(t)$ purely via substitutions. In a special case where the indel exit rate, $R_X^{ID}(\breve{s}(t),t)\equiv\left\langle\breve{s}(t)\middle|\hat{Q}_X^{ID}(t)\middle|\breve{s}(t)\right\rangle$, is independent of $\vec{\omega}(t)$ (and thus depends only on $s(t)$ and $t$), the factor is calculated as:

$$\begin{aligned}&\left\langle\breve{s}(t_\nu^{(+)})\middle|\hat{P}_0^{SID}(t_\nu,t_{\nu+1})\middle|\breve{s}(t_{\nu+1}^{(-)})\right\rangle\\&=\exp\left\{-\int_{t_\nu}^{t_{\nu+1}}dt\,R_X^{ID}(s(t_\nu^{(+)}),t)\right\}\left\langle\breve{s}(t_\nu^{(+)})\middle|T\left\{\exp\left(\int_{t_\nu}^{t_{\nu+1}}dt\,\hat{Q}^S(t)\right)\right\}\middle|\breve{s}(t_{\nu+1}^{(-)})\right\rangle\\&=\exp\left\{-\int_{t_\nu}^{t_{\nu+1}}dt\,R_X^{ID}(s(t_\nu^{(+)}),t)\right\}\left\langle\vec{\omega}(t_\nu^{(+)})\middle|T\left\{\exp\left(\int_{t_\nu}^{t_{\nu+1}}dt\,\hat{Q}^S(t)\right)\right\}\middle|\vec{\omega}(t_{\nu+1}^{(-)})\right\rangle.\end{aligned}$$

--- Eq.(A2.4)

Its only difference from the factor in Eq.(3.1.8b) of part I is the multiplication by the transition probability from $\vec{\omega}(t_\nu^{(+)})$ to $\vec{\omega}(t_{\nu+1}^{(-)})$ via substitutions. In more general cases where the indel exit rate $R_X^{ID}(\breve{s}(t),t)=R_X^{ID}(s(t),\vec{\omega}(t),t)$ could depend on the residue state $\vec{\omega}(t)$, we will decompose the substitution rate operator as $\hat{Q}^S(t)=\hat{Q}_M^S(t)+\hat{Q}_X^S(t)$ as argued above on the pure substitution case, and follow a line of argument similar to that in Subsection 3.1 of part I. Let $\tilde{H}^S\left[\alpha_0\left(\vec{\omega}(t_\nu^{(+)}),\vec{\omega}(t_{\nu+1}^{(-)})\right)\right]$ be the set of all substitution histories consistent with the gap-free (*i.e.*, trivial) PWA, $\alpha_0\left(\vec{\omega}(t_\nu^{(+)}),\vec{\omega}(t_{\nu+1}^{(-)})\right)$, between the residue states at the time boundaries. Then, we get:

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\left\langle \breve{s}(t_\nu^{(+)}) \middle| \hat{P}_0^{SID}(t_\nu, t_{\nu+1}) \middle| \breve{s}(t_{\nu+1}^{(-)}) \right\rangle$$

$$= \sum_{\substack{[\hat{M}_{S;1},\ldots,\hat{M}_{S;N_\nu}] \\ \in \hat{H}^S\left[\alpha_0\left(\vec{\omega}(t_\nu^{(+)}),\vec{\omega}(t_{\nu+1}^{(-)})\right)\right]}} \left[ \begin{array}{l} \int\cdots\int_{\substack{t_\nu = t_{\nu;0} < t_{\nu;1} < \cdots \\ < t_{\nu;N_\nu} < t_{\nu;N_\nu+1} = t_{\nu+1}}} dt_{\nu;1}\cdots dt_{\nu;N_\nu} \left\{ \exp\left( -\sum_{\nu'=0}^{N_\nu} \int_{t_{\nu;\nu'}}^{t_{\nu;\nu'+1}} dt\, R_X^{ID}(s(t_\nu^{(+)}), \vec{\omega}_{\nu;\nu'}, t) \right) \right. \\ \left. \times \rho_P\left[ \left( \left[ (\hat{M}_{S;1}, t_{\nu;1}),\ldots,(\hat{M}_{S;N_\nu}, t_{\nu;N_\nu}) \right], [t_\nu, t_{\nu+1}] \right) \middle| \left( \vec{\omega}(t_\nu^{(+)}), t_\nu \right) \right] \right\} \end{array} \right].$$

--- Eq.(A2.5a)

Here, $\hat{M}_{S;\nu'}$ (with $\nu' = 1,\ldots,N_\nu$) is the $\nu'$ th substitution event, like $\hat{M}_S(\omega \mapsto \omega'; x)$ above, in a substitution history during the open time interval $(t_\nu, t_{\nu+1})$. $t_{\nu;\nu'}$ is the time at which the $\nu'$ th event occurred. And $\vec{\omega}_{\nu;\nu'}$ is the residue state of the sequence immediately after the $\nu'$ th event, with an exception $\vec{\omega}_{\nu;0} \equiv \vec{\omega}(t_\nu^{(+)})$. Also, $\rho_P[\ldots]$ in Eq.(A2.5a) is the probability density of the substitution process $\left[(\hat{M}_{S;1}, t_{\nu;1}),\ldots,(\hat{M}_{S;N_\nu}, t_{\nu;N_\nu})\right]$, conditioned on the initial residue state $\vec{\omega}(t_\nu^{(+)})$. It is formally similar to the integrand in Eq.(3.1.8b) of part I, and is given by:

$$\rho_P\left[ \left( \left[ (\hat{M}_{S;1}, t_{\nu;1}),\ldots,(\hat{M}_{S;N_\nu}, t_{\nu;N_\nu}) \right], [t_\nu, t_{\nu+1}] \right) \middle| \left( \vec{\omega}(t_\nu^{(+)}), t_\nu \right) \right]$$

$$= \left( \prod_{\nu'=1}^{N_\nu} r_S(\hat{M}_{S;\nu'}; s(t_\nu^{(+)}), \vec{\omega}_{\nu;\nu'}, t_{\nu;\nu'}) \right) \exp\left\{ -\sum_{\nu'=0}^{N_\nu} \int_{t_{\nu;\nu'}}^{t_{\nu;\nu'+1}} d\tau\, R_X^S(s(t_\nu^{(+)}), \vec{\omega}_{\nu;\nu'}, \tau) \right\} ,$$

--- Eq.(A2.5b)

where we used the same notation as in Eq.(A2.5a). The symbol $r_S(\hat{M}_{S;\nu'}; s, \vec{\omega}, t)$ denotes the rate parameter of the substitution $\hat{M}_{S;\nu'}$ on the extended state $\breve{s} = (s, \vec{\omega})$ at time $t$. And $R_X^S(\breve{s}, t) \equiv R_X^S(s, \vec{\omega}, t) \equiv \sum_{x=1}^{L(s)} R_X^S(x; s, \vec{\omega}, t) = \sum_{x=1}^{L(s)} \sum_{\omega' \in \Omega, \omega' \neq \omega_x} r_S\left(x, \omega_x \mapsto \omega'; s, \vec{\omega}, t\right)$ is the substitution exit rate of the extended state $\breve{s} = (s, \vec{\omega})$ at time $t$. When deriving Eq.(A2.5b), we used the equation: $\langle \breve{s} | \hat{Q}_X^S(t) = -R_X^S(\breve{s}, t) \langle \breve{s} |$. Eq.(A2.5a) suggests that the factor $\left\langle \breve{s}(t_\nu^{(+)}) \middle| \hat{P}_0^{SID}(t_\nu, t_{\nu+1}) \middle| \breve{s}(t_{\nu+1}^{(-)}) \right\rangle$ is a weighted summation of $\exp\left\{ -\int_{t_\nu}^{t_{\nu+1}} dt\, R_X^{ID}(s(t_\nu^{(+)}), \vec{\omega}(t), t) \right\}$ over all substitution processes (each represented by a trajectory $\left\{ \vec{\omega}(t) \middle| \; t \in (t_\nu, t_{\nu+1}), \vec{\omega}(t) \in \Omega^{L(\breve{s}(t_\nu^{(+)}))} \right\}$) consistent with the (trivial) PWA, $\alpha_0\left(\vec{\omega}(t_\nu^{(+)}), \vec{\omega}(t_{\nu+1}^{(-)})\right)$, with the weights given by the probability densities of the processes. Because Eq.(A2.5a) supplemented with Eq.(A2.5b) is similar in form to Eq.(3.1.13) of part I supplemented with Eq.(3.1.8b) of part I, a reasoning similar to that in Subsection 4.1 of part I is also applicable when examining the factorability of $\left\langle \breve{s}(t_\nu^{(+)}) \middle| \hat{P}_0^{SID}(t_\nu, t_{\nu+1}) \middle| \breve{s}(t_{\nu+1}^{(-)}) \right\rangle$. We see that it is factorable into the product of an overall factor and contributions from separate regions if the two conditions are met. (i) The rate parameter $r_S\left(x, \omega_x \mapsto \omega'; s, \vec{\omega}, t\right)$ of every substitution $\hat{M}_S(\omega_x \mapsto \omega'; x)$ in each region is independent of the portions of the residue states in the other regions. And (ii)

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the increment of the total exit rate $R_X^{SID}(\breve{s},t) = R_X^{S}(\breve{s},t) + R_X^{ID}(\breve{s},t)$, caused by every substitution $\hat{M}_S(\omega_x \mapsto \omega'; x)$ in each region, is independent of the portions of the residue states in the other regions.

Then, substituting the factorized Eq.(A2.5a) for $\left\langle \breve{s}(t_\nu^{(+)}) \middle| \hat{P}_0^{SID}(t_\nu, t_{\nu+1}) \middle| \breve{s}(t_{\nu+1}^{(-)}) \right\rangle$ into Eq.(A2.3b), and substituting the result into Eq.(A2.3a), we can examine the factorability of the total probability, $P\left[\left(\alpha(\breve{s}^A, \breve{s}^D), [t_I, t_F]\right) \middle| (\breve{s}^A, t_I)\right]$, of a PWA between the extended states of the ancestral and descendant sequences. Again, we can follow a line of reasoning similar to that in Subsection 4.1 of part I, with two differences: (a) here, $-\log\left[\left\langle \breve{s}(t_\nu^{(+)}) \middle| \hat{P}_0^{SID}(t_\nu, t_{\nu+1}) \middle| \breve{s}(t_{\nu+1}^{(-)}) \right\rangle\right]$ plays the role of $\int_{t_\nu}^{t_{\nu+1}} dt\, R_X^{ID}(s(t_\nu^{(+)}), \vec{\omega}(t), t)$; and (b) here, not only the basic states but also the residue states comes into question. Moreover, thanks to the preceding argument, we know that a change in $-\log\left[\left\langle \breve{s}(t_\nu^{(+)}) \middle| \hat{P}_0^{SID}(t_\nu, t_{\nu+1}) \middle| \breve{s}(t_{\nu+1}^{(-)}) \right\rangle\right]$ results from a collective effects of the changes in the substitution rates and the total exit rates in Eqs.(A2.5a,b). Thus, we find the following sufficient set of conditions under which $P\left[\left(\alpha(\breve{s}^A, \breve{s}^D), [t_I, t_F]\right) \middle| (\breve{s}^A, t_I)\right]$ is factorable into the product of an overall factor and contributions from separate regions.

**Condition (I):** In each region, the rate of each of the substitutions and indels is independent of the portions of the extended states in the other regions. And

**Condition (II):** In each region, the increment of the total exit rate, $R_X^{SID}(\breve{s},t) = R_X^{S}(\breve{s},t) + R_X^{ID}(\breve{s},t)$, due to each of the substitutions and indels, is independent of the portions of the extended states in the other regions.

As argued in Subsection 3.2 of part I, we could calculate the probability of a given MSA (under a given phylogenetic tree and a given evolutionary model) by assembling the PWAs between the ancestral and descendant sequence states along all branches and by summing over all possible sets of sequence states at internal nodes that are consistent with the MSA. Thus, once we know that the probabilities of PWAs of extended sequence states are factorable, we can also factorize the probability of a given MSA of extended sequence states under a given phylogenetic tree. We can do this by extending the line of arguments unfolded in Subsection 4.2 of part I so that it will incorporate the substitution processes and the resulting residue components of the sequence states. According to such an extended line of arguments, we find that the probability of a given MSA is factorable if the above conditions (I) and (II) holds and, in addition, if we have an extended version of the factorability of the root sequence probability (given in Eq.(4.2.8) of part I):

$$P\left[\left(\breve{s}^{Root}, n^{Root}\right)\right] = P\left[\left(\breve{s}_0^{Root}, n^{Root}\right)\right] \prod_{\breve{\mathrm{K}}=1}^{\breve{\mathrm{K}}_{\max}} \mu_P\left[\breve{s}^{Root}, \breve{s}_0^{Root}, n^{Root}; \breve{C}_{\breve{\mathrm{K}}}\right]. \text{ --- Eq.(A2.6)}$$

Here $\breve{s}_0^{Root} = (s_0^{Root}, \vec{\omega}_0^{Root})$ is the extended state of a "reference" root sequence. And the set of regions, $\left\{\breve{C}_{\breve{\mathrm{K}}}\right\}_{\breve{\mathrm{K}}=1,\ldots,\breve{\mathrm{K}}_{\max}}$, consists not only of the regions potentially accommodating local indel histories ($\left\{C_{\mathrm{K}}\right\}_{\mathrm{K}=1,\ldots,\mathrm{K}_{\max}}$) but also of the PASs (*i.e.*, gapless columns), which can never experience indels but can possibly experience substitutions.

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**A3. Factorizing probability into basic and residue components**

Thus far, we considered the probability of a given alignment as the summation of the probabilities with consistent indel histories, each of which, in turn, is the summation of the probabilities of substitution histories consistent with the alignment and the indel history (see, *e.g.*, Eqs.(A2.3a,b)). The factorization of the probability as discussed above in Subsection A2 could substantially reduce the computational burden of the calculation. However, we will be able to improve the computational efficiency further if we can factorize the entire probability into the product of the "basic" component, which concerns indel processes, and the residue component, which concerns substitution processes (and the initial states of residues). This is because the residue component will then need to be computed only once, instead of as many times as the indel LHSs to be considered. Here we will consider a sufficient set of conditions for such an "indel-substitution factorization." Because indels are usually at most 1/10 times as frequent as substitutions (*e.g.*, Lunter 2007; Cartwright 2009), we will consider, as in Eqs.(A2.3a,b) and Eqs.(A2.5a,b), that each indel history determines the "skeleton" of the alignment that are made up of the basic sequence states at the nodes. Hence we suppose that each substitution history determines the residue states that flesh out the "skeleton" to complete the alignment. First, to simplify the argument, we assume that the probability of the alignment skeleton itself does not depend on the residue states. This assumption is true if the indel rates are independent of the residue states of the sequence immediately before the indels:

$$\begin{aligned} r_D(x_B, x_E; \breve{s}, t) &= r'_D(x_B, x_E; s, t), \\ r_I(x, l, \delta\vec{\omega}'[l]; \breve{s}, t) &= r'_I(x, l; s, t)\, p_I(\delta\vec{\omega}'[l]; x, l; s, t). \end{aligned} \qquad \text{--- Eqs.(A3.1a,b)}$$

It should be noted that the insertion rates could still depend on the residue states of the *inserted* subsequence through $p_I(\delta\vec{\omega}'[l]; x, l; s, t)$'s, which satisfy $\sum_{\delta\vec{\omega}'[l]\in\Omega^l} p_I(\delta\vec{\omega}'[l]; x, l; s, t) = 1$. Under the condition Eqs.(A3.1a,b), the indel exit rate $R_X^{ID}(\breve{s}(t), t) \equiv \langle \breve{s}(t) | \hat{Q}_X^{ID}(t) | \breve{s}(t) \rangle$ is independent of the residue component of $\breve{s}(t)$, and thus the factor $\langle \breve{s}(t_v^{(+)}) | \hat{P}_0^{SID}(t_v, t_{v+1}) | \breve{s}(t_{v+1}^{(-)}) \rangle$ in Eq.(A2.3b) can be factorized as in Eq.(A2.4). Thus, the integrand in Eq.(A2.3b) is reduced to $\Pi_{ID} \times \Pi_S$, with

$$\begin{aligned} \Pi_{ID} &\equiv \left( \prod_{v=1}^{N} r'(\hat{M}_v; s_v, t_v) \right) \exp\left\{ -\sum_{v=0}^{N} \int_{t_v}^{t_{v+1}} dt\, R_X^{ID}(s(t_v^{(+)}), t) \right\}, \\ \Pi_S &\equiv \left( \prod_{v=1}^{N} p'(\hat{F}_v; s_v, t_v) \right) \left( \prod_{v=0}^{N} \langle \vec{\omega}(t_v^{(+)}) | \hat{P}^S(t_v, t_{v+1}) | \vec{\omega}(t_{v+1}^{(-)}) \rangle \right) \end{aligned} \qquad \text{--- Eqs.(A3.2a,b)}$$

under the same setting and notations as introduced around Eq.(A2.3b). Here $r'(\hat{M}_v; s_v, t_v)$ denotes $r'_I(x, l; s(t_v^{(-)}), t_v)$ and $r'_D(x_B, x_E; s(t_v^{(-)}), t_v)$, respectively, if $\hat{M}_v$ is $\hat{M}_I(x, l)$ and $\hat{M}_D(x_B, x_E)$. And $p'(\hat{F}_v; s_v, t_v)$ denotes $p_I(\delta\vec{\omega}_v[l]; x, l; s(t_v^{(-)}), t_v)$ and 1 (unity), respectively, if $\hat{F}_v = \hat{F}(x, \delta\vec{\omega}_v[l])$ and $\hat{F}_v = \hat{I}$. The product, $\Pi_{ID}$, is of the same form as the integrand in Eq.(3.1.8b) of part I (Ezawa, Graur and Landan 2015a), which represents the (residue-independent) probability distribution of an indel process, $\left[ (\hat{M}_1, t_1), (\hat{M}_2, t_2), \ldots, (\hat{M}_N, t_N) \right]$. The product, $\Pi_S$, is the joint probability of the residue states, $\{\vec{\omega}_v\}_{v=1,\ldots,N}$, "at the times of" the indel events and of the final residue state ($\vec{\omega}^D$), conditioned on the initial residue state ($\vec{\omega}^A$) and the indel process. One major difference between Eq.(3.1.13) of part I and its extension, Eq.(A2.3a), is that the latter performs the summation over all possible $\{\vec{\omega}_v\}_{v=1,\ldots,N}$'s. Therefore, if the summation of

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$\Pi_S$'s becomes independent of the indel process, then, it can be factored out of the multiple-time integration in Eq.(A2.3b) and further be factored out of the summation over the consistent indel histories in Eq.(A2.3a). This means that the total probability of a PWA can be factorized into the product of the probability of the "skeleton" of the PWA (as in Eq.(3.1.13) of part I) and the probability of the residue component of the PWA. For the latter, the calculation techniques have been developed well (*e.g.*, Felsenstein 1981, 2004; Yang 2006).

Thus the problem is reduced to the condition under which the aforementioned summation of $\Pi_S$'s over all $\{\vec{\omega}_\nu\}_{\nu=1,\ldots,N}$'s becomes independent of the indel process consistent with the PWA. It would be convenient to categorize the columns in the PWA according to their histories, *i.e.*, whether or not they are preserved throughout $[t_I, t_F]$, and, if not, the time of insertion ($t_{\nu(I)}$) and/or the time of deletion ($t_{\nu(D)}$). (If a column has both $t_{\nu(I)}$ and $t_{\nu(D)}$, $\nu(I) < \nu(D)$ should always hold.) And here, we will make a second simplifying assumption, that is, the substitution rates within each block of contiguous columns with a shared history are assumed as independent of the portions of the residue states in the rest of the sequence. Then, the factor $\langle \vec{\omega}(t_\nu^{(+)}) | \hat{P}^S(t_\nu, t_{\nu+1}) | \vec{\omega}(t_{\nu+1}^{(-)}) \rangle$ in Eq.(A3.2b) can be factorized into the product of partial probabilities, $\langle \vec{\omega}(t_\nu^{(+)})[B_i] | \hat{P}^S(t_\nu, t_{\nu+1})[B_i] | \vec{\omega}(t_{\nu+1}^{(-)})[B_i] \rangle$'s, of such blocks with shared histories (denoted as $B_i$ ($i = 1,\ldots,I_B$)). And we also assume that $p_I(\delta\vec{\omega}_\nu[l]; x, l; s(t_\nu^{(-)}), t_\nu)$ can be factorized into the product of components, denoted as $\prod_{B_i \subseteq [\upsilon_{x+1}(t_{\nu(I)}^{(+)}),\ldots,\upsilon_{x+l}(t_{\nu(I)}^{(+)})]} p_I(\vec{\omega}_\nu[B_i]; B_i; s(t_\nu^{(-)}), t_\nu)$. Here each component ($p_I(\vec{\omega}_\nu[B_i]; B_i; s(t_\nu^{(-)}), t_\nu)$) comes from a relevant block, $B_i \subseteq [\upsilon_{x+1}(t_{\nu(I)}^{(+)}),\ldots,\upsilon_{x+l}(t_{\nu(I)}^{(+)})]$. (Note that the blocks are positioned relative to the MSA (or the array of ancestries) but not relative to a particular sequence state.) These two assumptions make the summation of $\Pi_S$'s over the intermediate residue states also factorable into the contributions from such blocks. In the following, we consider the contributions from such blocks to the summation of $\Pi_S$'s. For this purpose, we broadly classify them into four classes: (1) when $B_i$ was preserved throughout $[t_I, t_F]$; (2) when it existed at $t_I$ but was deleted at $t_{\nu(D)}$; (3) when it was inserted at $t_{\nu(I)}$ and was deleted at $t_{\nu(D)}$; (4) when it was inserted at $t_{\nu(I)}$ and was preserved through $t_F$.

(1) When $B_i$ was preserved throughout $[t_I, t_F]$, the contribution from the block is:

$$\sum_{\vec{\omega}_1[B_i] \in \Omega^{L(B_i)}} \cdots \sum_{\vec{\omega}_N[B_i] \in \Omega^{L(B_i)}} \prod_{\nu=0}^{N} \langle \vec{\omega}_\nu[B_i] | \hat{P}^S(t_\nu, t_{\nu+1})[B_i] | \vec{\omega}_{\nu+1}[B_i] \rangle. \quad \text{--- Eq.(A3.3a)}$$

Because $\hat{P}^S(t,t')[B_i] = T\left\{\exp\left(\int_t^{t'} d\tau\, \hat{Q}^S(\tau)[B_i]\right)\right\}$ itself is a stochastic evolutionary operator made of the block-wise substitution rate operator $\hat{Q}^S(\tau)[B_i]$, it satisfies the Chapman-Kolmogorov (CK) equation:

$$\sum_{\vec{\omega}''[B_i] \in \Omega^{L(B_i)}} \langle \vec{\omega}[B_i] | \hat{P}^S(t,t'')[B_i] | \vec{\omega}''[B_i] \rangle \langle \vec{\omega}''[B_i] | \hat{P}^S(t'',t')[B_i] | \vec{\omega}'[B_i] \rangle \quad \text{--- Eq.(A3.4)}$$
$$= \langle \vec{\omega}[B_i] | \hat{P}^S(t,t')[B_i] | \vec{\omega}'[B_i] \rangle.$$

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Applying the CK equation successively regarding the intermediate residue states $\{\vec{\omega}_v[B_i]\}_{v=1,\ldots,N}$, Eq.(A3.3a) is reduced to $\langle \vec{\omega}^A[B_i] | \hat{P}^S(t_I, t_F)[B_i] | \vec{\omega}^D[B_i] \rangle$, which depends only on the initial and final time points, and the residue states at these time points.

(2) When $B_i$ existed at $t_I$ but was deleted at $t_{v(D)}$, the block's contribution is:

$$\sum_{\vec{\omega}_1[B_i]\in\Omega^{L(B_i)}} \cdots \sum_{\vec{\omega}_{v(D)}[B_i]\in\Omega^{L(B_i)}} \prod_{v=0}^{v(D)-1} \langle \vec{\omega}_v[B_i] | \hat{P}^S(t_v, t_{v+1})[B_i] | \vec{\omega}_{v+1}[B_i] \rangle. \quad \text{--- Eq.(A3.3b)}$$

The CK equations apply to the summations over $\{\vec{\omega}_v[B_i]\}_{v=1,\ldots,v(D)-1}$, and Eq.(A3.3b) is reduced to $\sum_{\vec{\omega}_{v(D)}[B_i]\in\Omega^{L(B_i)}} \langle \vec{\omega}^A[B_i] | \hat{P}^S(t_I, t_{v(D)})[B_i] | \vec{\omega}_{v(D)}[B_i] \rangle$. The summand is the probability of $\vec{\omega}_{v(D)}[B_i]$ at time $t_{v(D)}$ conditioned on the initial state $\vec{\omega}^A[B_i]$. Thus, in this case, the summation Eq.(A3.3b) gives $1$ (unity).

(3) When $B_i$ was inserted at $t_{v(I)}$ and was deleted at $t_{v(D)}$, the block's contribution is:

$$\sum_{\vec{\omega}_{v(I)}[B_i]\in\Omega^{L(B_i)}} \cdots \sum_{\vec{\omega}_{v(D)}[B_i]\in\Omega^{L(B_i)}} \left[ \begin{array}{l} p_I(\vec{\omega}_{v(I)}[B_i]; B_i; s(t_{v(I)}^{(-)}), t_{v(I)}) \\ \times \prod_{v=v(I)}^{v(D)-1} \langle \vec{\omega}_v[B_i] | \hat{P}^S(t_v, t_{v+1})[B_i] | \vec{\omega}_{v+1}[B_i] \rangle \end{array} \right]. \quad \text{--- Eq.(A3.3c)}$$

The CK equations apply to the summations over $\{\vec{\omega}_v[B_i]\}_{v=v(I)+1,\ldots,v(D)-1}$, and we have:

$$\sum_{\vec{\omega}_{v(I)}[B_i]\in\Omega^{L(B_i)}} \left[ \begin{array}{l} p_I(\vec{\omega}_{v(I)}[B_i]; B_i; s(t_{v(I)}^{(-)}), t_{v(I)}) \\ \times \left( \sum_{\vec{\omega}_{v(D)}[B_i]\in\Omega^{L(B_i)}} \langle \vec{\omega}_{v(I)}[B_i] | \hat{P}^S(t_{v(I)}, t_{v(D)})[B_i] | \vec{\omega}_{v(D)}[B_i] \rangle \right) \end{array} \right]. \quad \text{--- Eq.(A3.3c')}$$

As in case (2), the summation over $\vec{\omega}_{v(D)}[B_i]$'s in the parentheses in Eq.(A3.3c') gives unity, thus the equation is reduced to $\sum_{\vec{\omega}_{v(I)}[B_i]\in\Omega^{L(B_i)}} p_I(\vec{\omega}_{v(I)}[B_i]; B_i; s(t_{v(I)}^{(-)}), t_{v(I)})$. This summation is nothing other than $1$ (unity) thanks to the normalization condition of $p_I(\vec{\omega}_{v(I)}[B_i]; B_i; s(t_{v(I)}^{(-)}), t_{v(I)})$. Thus, in conjunction with (2), we see that the contribution from a deleted block to the *conditional* probability of a PWA is always unity, regardless of whether it already existed in the initial sequence or it was inserted.

(4) When $B_i$ was inserted at $t_{v(I)}$ and was preserved through $t_F$, the block's contribution is:

$$\sum_{\vec{\omega}_{v(I)}[B_i]\in\Omega^{L(B_i)}} \cdots \sum_{\vec{\omega}_N[B_i]\in\Omega^{L(B_i)}} \left[ \begin{array}{l} p_I(\vec{\omega}_{v(I)}[B_i]; B_i; s(t_{v(I)}^{(-)}), t_{v(I)}) \\ \times \prod_{v=v(I)}^{N} \langle \vec{\omega}_v[B_i] | \hat{P}^S(t_v, t_{v+1})[B_i] | \vec{\omega}_{v+1}[B_i] \rangle \end{array} \right]. \quad \text{--- Eq.(A3.3d)}$$

The CK equations apply to the summations over $\{\vec{\omega}_v[B_i]\}_{v=v(I)+1,\ldots,N}$, and we have:

$$\sum_{\vec{\omega}_{v(I)}[B_i]\in\Omega^{L(B_i)}} \left[ p_I(\vec{\omega}_{v(I)}[B_i]; B_i; s(t_{v(I)}^{(-)}), t_{v(I)}) \times \langle \vec{\omega}_{v(I)}[B_i] | \hat{P}^S(t_{v(I)}, t_F)[B_i] | \vec{\omega}^D[B_i] \rangle \right].$$

--- Eq.(A3.3d')

In order for the summation of $\Pi_S$'s to be independent of the indel process, Eq.(A3.3d') needs to be independent of $t_{v(I)}$, especially it needs to be equal to its

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

limit under $t_{v(I)} \to t_F$. In this limit, we have $\langle \vec{\omega}_{v(I)}[B_i] | \hat{P}^S(t_{v(I)}, t_F)[B_i] | \vec{\omega}^D[B_i] \rangle \to \delta(\vec{\omega}_{v(I)}[B_i], \vec{\omega}^D[B_i])$, thanks to the defining property of the stochastic evolutionary operator (Eq.(1.1.10b') of part I). Thus, the aforementioned condition can be expressed as:

$$\sum_{\vec{\omega}_{v(I)}[B_i] \in \Omega^{L(B_i)}} \left[ p_I(\vec{\omega}_{v(I)}[B_i]; B_i; s(t^{(-)}_{v(I)}), t_{v(I)}) \times \langle \vec{\omega}_{v(I)}[B_i] | \hat{P}^S(t_{v(I)}, t_F)[B_i] | \vec{\omega}^D[B_i] \rangle \right]$$
$$= \quad p_I(\vec{\omega}^D[B_i]; B_i; s^D, t_F) \quad .$$

--- Eq.(A3.5)

Here we used $\lim_{t_{v(I)} \to t_F} s(t^{(-)}_{v(I)}) = s^D$. The right hand side of Eq.(A3.5) depends only on the final time, a fixed block, and the portion of the extended sequence state in the block at the final time, and therefore it is totally independent of the indel process, as required.

Thus, let us assume that the four conditions we imposed above are satisfied. Namely, (i) the indel rates are independent of the residue states; (ii) each finite-time evolution probability of the residue state via substitutions is factorable into the product of the probabilities of blocks of particular histories; (iii) the residue state spectrum of each inserted sub-sequence is factorable into the product of block-wise contributions; and (iv) each block-wise spectrum of inserted residue states satisfies Eq.(A3.5). Under these conditions, the summation of $\Pi_S$'s (given in Eq.(A3.2b)) over all possible intermediate residue states is expressed as:

$$\left[ \prod_{B_i: class\,(1)} \langle \vec{\omega}^A[B_i] | \hat{P}^S(t_I, t_F)[B_i] | \vec{\omega}^D[B_i] \rangle \right] \times \left[ \prod_{B_j: class\,(4)} p_I(\vec{\omega}^D[B_j]; B_j; s^D, t_F) \right].$$

--- Eq.(A3.6)

Thus, those sites that didn't make it through $t_F$ do not contribute to Eq.(A3.6). (More precisely, each of them contributes a trivial multiplication factor of 1 (unity).) Eq.(A3.6) is independent of the indel process that provides the "skeleton," except the possible dependence on the particular way how the indel process partitions the PWA into blocks. If we consider all indel processes that are consistent with the PWA, there could be a variety of ways of partitioning it into blocks. A simplest way to assure the independence on the way of partitioning the PWA is to assume the following two properties. (1) The stochastic evolutionary operator of substitutions is factorable into the product of site-wise operators: $\hat{P}^{S(L)}(t,t') = \bigotimes_{x=1}^{L} \hat{P}^S(t,t'; \upsilon_x)$, as in Eq.(A1.3). And (2) the residue state spectrum of inserted sub-sequence, $p_I(\delta\vec{\omega}'[l]; x, l; s, t)$ in Eq.(A3.1b), is also factorable into the product of site-wise contributions: $p_I(\delta\vec{\omega}'[l]; x, l; s, t) = \prod_{i=1}^{l} p_I(\omega'_{x+i}; \upsilon_{x+i}(s'), t)$ with $\sum_{\omega \in \Omega} p_I(\omega; \upsilon, t) = 1$. Here, $(s', \vec{\omega}')$ is the extended sequence state immediately *after* the insertion, and we used the ancestries ($\upsilon_x$'s) instead of the site numbers ($x$'s) as arguments because the former is invariant through an indel process. Under this assumption, the condition Eq.(A3.5) is reduced to the following single-site condition:

$$\sum_{\omega \in \Omega} p_I(\omega; \upsilon_j, t) \langle \omega | \hat{P}^S(t, t_F; \upsilon_j) | \omega^D(\upsilon_j) \rangle = p_I(\omega^D(\upsilon_j); \upsilon_j, t_F) \qquad \text{--- Eq.(A3.5')}$$

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for $^{\forall}t \in [t_I, t_F]$. Then, after partitioning the PWA, $\alpha(\breve{s}^A, \breve{s}^D)$, into the columns with ancestries $\{\upsilon_1, \upsilon_1, ..., \upsilon_{L(\alpha)}\}$, where $L(\alpha)$ is the number of columns in the PWA, Eq.(A3.6) is further reduced to:

$$\left[\prod_{\upsilon_i: class\,(1)} \left\langle \omega^A(\upsilon_i) \middle| \hat{P}^S(t_I, t_F; \upsilon_i) \middle| \omega^D(\upsilon_i) \right\rangle\right] \times \left[\prod_{\upsilon_j: class\,(4)} p_I(\omega^D(\upsilon_j); \upsilon_j, t_F)\right]$$
$$= P\left[\left(\vec{\omega}^A = \vec{\rho}^A\left(\alpha(\breve{s}^A, \breve{s}^D)\right), \vec{\omega}^D = \vec{\rho}^D\left(\alpha(\breve{s}^A, \breve{s}^D)\right), [t_I, t_F]\right) \middle| \left(\breve{s}^A = (s^A, \vec{\omega}^A), t_I\right), \alpha(s^A, s^D)\right].$$

--- Eq.(A3.6')

Here $\omega^A(\upsilon_i)$ and $\omega^D(\upsilon_i)$, respectively, denote the ancestral and the descendant residues in the PWA column (*i.e.*, site) with ancestry $\upsilon_i$, and $\hat{P}^S(t_I, t_F; \upsilon_i)$ is the single-site evolutionary operator via substitutions in the site with ancestry $\upsilon_i$. On the right hand side, $\vec{\rho}^A\left(\alpha(\breve{s}^A, \breve{s}^D)\right)$ and $\vec{\rho}^D\left(\alpha(\breve{s}^A, \breve{s}^D)\right)$, respectively, symbolically represent the vector functions extracting the ancestral and descendant residue states from the PWA of extended sequence states ($\alpha(\breve{s}^A, \breve{s}^D)$). As desired, the left hand side of Eq.(A3.6') depends only on the ancestral and descendant states, as well as the homology structure of the PWA, and does not depend on any details of the indel processes. Hence, the right hand side follows. Thus, in this case, the summation of $\Pi_S$'s can indeed be factored out of the multiple-time integration in Eq.(A2.3b) and also of the summation over all indel histories consistent with the PWA in Eq.(A2.3a). This finally enables us to re-expresses Eq.(A2.3a), supplemented by Eq.(A2.3b), into the form we desire:

$$P\left[\left(\alpha\left(\breve{s}^A = (s^A, \vec{\omega}^A), \breve{s}^D = (s^D, \vec{\omega}^D)\right), [t_I, t_F]\right) \middle| \left(\breve{s}^A = (s^A, \vec{\omega}^A), t_I\right)\right]$$
$$= P\left[\left(\alpha(s^A, s^D), [t_I, t_F]\right) \middle| (s^A, t_I)\right]$$
$$\times P\left[\left(\vec{\omega}^A = \vec{\rho}^A\left(\alpha(\breve{s}^A, \breve{s}^D)\right), \vec{\omega}^D = \vec{\rho}^D\left(\alpha(\breve{s}^A, \breve{s}^D)\right), [t_I, t_F]\right) \middle| \left(\breve{s}^A = (s^A, \vec{\omega}^A), t_I\right), \alpha(s^A, s^D)\right].$$

--- Eq.(A3.7a)

On the right hand side, the second factor is given by Eq.(A3.6'), and the first factor is given by:

$$P\left[\left(\alpha(s^A, s^D), [t_I, t_F]\right) \middle| (s^A, t_I)\right]$$
$$= \sum_{\substack{[\hat{M}_1, \hat{M}_2, \cdots, \hat{M}_N] \\ \in \tilde{\mathrm{H}}^{ID}[\alpha(s^A, s^D)]}} \int_{t_I = t_0 < t_1 < \cdots < t_N < t_{N+1} = t_F} \cdots \int dt_1 \cdots dt_N \left[\left(\prod_{\nu=1}^{N} r'(\hat{M}_\nu; s_\nu, t_\nu)\right) \exp\left\{-\sum_{\nu=0}^{N} \int_{t_\nu}^{t_{\nu+1}} dt\, R_X^{ID}(s(t_\nu^{(+)}), t)\right\}\right].$$

--- Eq.(A3.7b)

Here we remind the equations, $s(t_0^{(+)}) = s_I$, $s(t_N^{(+)}) = s(t_{N+1}^{(-)}) = s_F$, and $\left\langle s(t_\nu^{(+)}) \right| = \left\langle s(t_\nu^{(-)}) \right| \hat{M}_\nu = \left\langle s(t_{\nu-1}^{(+)}) \right| \hat{M}_\nu$ for $\nu = 1, ..., N$. Eq.(A3.7b) corresponds exactly to Eq.(3.1.13) of part I supplemented by Eq.(3.1.8b) of part I, which gives the probability of the "skeleton" $\alpha(s^A, s^D)$ of the PWA $\alpha(\breve{s}^A, \breve{s}^D)$, conditioned on the basic ancestral state $s^A$ at the initial time, due to indel processes.

The above line of arguments implies that it would be very difficult, even if it is possible at all, to factorize a whole probability of a PWA into the product, Eq.(A3.7a), of the basic component (Eq.(A3.7b)) and the residue component (Eq.(A3.6')), *unless*

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the residue component of the stochastic evolution operator is factorable into the product of column-wise operators as in Eq.(A1.3), or *unless* the indel rate parameters can be expressed as in Eqs.(A3.1a,b), where $r'_D(x_B, x_E; s,t)$ and $r'_I(x,l; s,t)$ are independent of the residue states. Nevertheless, even if either of these conditions is violated, the whole probability of a PWA is still factorable into the product of the contributions from some separated regions, if the rate parameters of indels and substitutions in each region are independent of the portions of the extended sequence state outside of the region, as argued in Subsection A2. Therefore, even if the residue states at some sites have substantial impacts on the indel rates and/or the substitution rates, if such effects are localized in some narrow regions, we could first factorize the entire PWA probability into the product of regional contributions and then factorize most of such regional contributions into the basic and the substitution components. Then, we could separately handle the small portions whose "indel-substitution factorizations" are not possible. This way, the computational burden may still be mitigated considerably. Such a situation might apply to some mutagenic and/or functional motifs that are scattered along the sequence and which show quite rapid turnover. (If a motif is strongly conserved, in contrast, its effect will be well approximated by the *ancestry* dependence, instead of the residue dependence, of the rate parameters, and thus the "indel-substitution factorization" holds at least approximately.)

Now, let us assume the aforementioned conditions for the indel-substitution factorization of the conditional probability of PWAs, *i.e.*, that the indel rate parameters are of the forms in Eqs.(A3.1a,b) and that the substitution evolutionary operator is factorable into the product of the column-wise operators as in Eq.(A1.3). Under such conditions, we will examine whether the probability of a given MSA is also factorable into the indel (*i.e.*, "basic") and substitution (*i.e.*, residue) components. In this case, we generalize Eqs.(3.2.13a,b') of part I to the probability of an "extended MSA," *i.e.*, an alignment of multiple extended sequence states, $\alpha[\breve{s}_1, \breve{s}_2, ..., \breve{s}_{N^X}]$. The set of all sets of basic states at internal nodes consistent with the MSA, *i.e.*, $\Sigma\left[\alpha[s_1, s_2, ..., s_{N^X}]; \left\{n \in \mathrm{N}^{IN}(T)\right\}; T\right]$ in Eq.(3.2.13a) of part I, can be extended by filling in each of the sets of internal basic states with all possible residue states. Thus, the extended set is expressed as:

$$\breve{\Sigma}\left[\breve{\alpha}[\breve{s}_1, \breve{s}_2, ..., \breve{s}_{N^X}]; \left\{n \in \mathrm{N}^{IN}(T)\right\}; T\right]$$
$$\equiv \left\{ \left\{\breve{s}(n) = \left(s(n), \vec{\omega}(n)\right)\right\}_{n \in \mathrm{N}^{IN}(T)} \;\middle|\; \begin{array}{l} \left\{s(n)\right\}_{n \in \mathrm{N}^{IN}(T)} \in \Sigma\left[\alpha[s_1, s_2, ..., s_{N^X}]; \left\{n \in \mathrm{N}^{IN}(T)\right\}; T\right], \\ \vec{\omega}(n) \in \Omega^{L(s(n))} \quad for \quad {}^{\forall}n \in \mathrm{N}^{IN}(T) \end{array} \right\}.$$

--- Eq.(A3.8)

Hence, we have the extended version of Eq.(3.2.13a) of part I as follows:

$$P\left[\alpha[\breve{s}_1, \breve{s}_2, ..., \breve{s}_{N^X}] \,\middle|\, T\right] = \sum_{\substack{\{\breve{s}(n)\}_{\mathrm{N}^{IN}} \\ \in \breve{\Sigma}\left[\alpha[\breve{s}_1, \breve{s}_2, ..., \breve{s}_{N^X}]; \left\{n \in \mathrm{N}^{IN}(T)\right\}; T\right]}} P\left[\alpha[\breve{s}_1, \breve{s}_2, ..., \breve{s}_{N^X}]; \left\{\breve{s}(n)\right\}_{\mathrm{N}^{IN}} \,\middle|\, T\right]$$

$$= \sum_{\substack{\{s(n)\}_{\mathrm{N}^{IN}} \\ \in \Sigma\left[\alpha[s_1, s_2, ..., s_{N^X}]; \left\{n \in \mathrm{N}^{IN}(T)\right\}; T\right]}} \prod_{n \in \mathrm{N}^{IN}(T)} \left( \sum_{\vec{\omega}(n) \in \Omega^{L(s(n))}} \right) P\left[\alpha[\breve{s}_1, \breve{s}_2, ..., \breve{s}_{N^X}]; \left\{(s(n), \vec{\omega}(n))\right\}_{\mathrm{N}^{IN}} \,\middle|\, T\right].$$

--- Eq.(A3.9a)

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Here, $\prod_{n\in \mathrm{N}^{IN}(T)}\left(\sum_{\vec{\omega}(n)\in\Omega^{L(s(n))}}\right)$ represents the multiple summations over all possible sets of residue states at internal nodes. More precisely, in each possible set, $\{\vec{\omega}(n)\}_{n\in\mathrm{N}^{IN}(T)}$, each component state ($\vec{\omega}(n)\in\Omega^{L(s(n))}$) fills in a fixed basic state ($s(n)$) at each internal node ($n\in\mathrm{N}^{IN}(T)$). And the probability, $P\left[\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}];\{(s(n),\vec{\omega}(n))\}_{\mathrm{N}^{IN}}\,|\,T\right]$, is given by an extended version of Eq.(3.2.13b') of part I:

$$P\left[\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}];\{\breve{s}(n)\}_{\mathrm{N}^{IN}}\,|\,T\right]$$
$$=P\left[\left(\breve{s}^{Root},n^{Root}\right)\right]\prod_{b\in\{b\}_T}P\left[(\alpha(\breve{s}^A(b),\breve{s}^D(b)),b)\,\middle|\,(\breve{s}^A(b),n^A(b))\right]. \quad \text{--- Eq.(A3.9b)}$$

Here, as below Eq.(3.2.13b') of part I, $P\left[(\alpha(\breve{s}^A(b),\breve{s}^D(b)),b)\,\middle|\,(\breve{s}^A(b),n^A(b))\right]$ denotes the probability of a PWA between the *extended* states at the ancestral and descendant nodes of branch $b$. Under the present assumptions, each of such probabilities can be factorized into the basic and residue components, as in Eq.(A3.7a). In addition, we assume that the root state probability is also factorable as:

$$P\left[\left(\breve{s}^{Root}=(s^{Root},\vec{\omega}^{Root}),n^{Root}\right)\right]=P\left[\left(s^{Root},n^{Root}\right)\right]\times\prod_{x=1}^{L(s^{Root})}P\left[\left(\omega_x^{Root},n^{Root}\right)\,\middle|\,\upsilon_x^{Root}\right],$$

--- Eq.(A3.9c)

where $\omega_x^{Root}$ and $\upsilon_x^{Root}$ denote the residue and basic components, respectively, at the $x$ th site of the root sequence. By substituting Eqs.(A3.7a, 9c) into Eq.(A3.9b), we see that the summand in Eq.(A3.9b), *i.e.*, $P\left[\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}];\{(s(n),\vec{\omega}(n))\}_{\mathrm{N}^{IN}}\,|\,T\right]$, is factorized as:

$$P\left[\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}];\{(s(n),\vec{\omega}(n))\}_{\mathrm{N}^{IN}}\,|\,T\right]$$
$$=P\left[\alpha[s_1,s_2,...,s_{N^X}];\{s(n)\}_{\mathrm{N}^{IN}}\,|\,T\right]$$
$$\times P\left[\begin{matrix}\{\vec{\omega}_j=\vec{\rho}_j\left(\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]\right)\}_{j=1,2,...,N^X},\\ \{\vec{\omega}(n)=\vec{\rho}\left(\breve{s}(n)\right)\}_{n\in\mathrm{N}^{IN}}\end{matrix}\,\middle|\,T,\alpha[s_1,s_2,...,s_{N^X}];\{s(n)\}_{\mathrm{N}^{IN}}\right].$$

--- Eq.(A3.10a)

Here, $\vec{\rho}_j\left(\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]\right)$ (with $j=1,2,...,N^X$) symbolically represents the vector function extracting the residue state of $\breve{s}_j$ in the MSA of extended sequence states ($\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]$). And $\vec{\rho}\left(\breve{s}\right)$ represents the vector function extracting the residue state of an extended sequence state $\breve{s}$. On the right hand side of Eq.(A3.10a), the first factor, $P\left[\alpha[s_1,s_2,...,s_{N^X}];\{s(n)\}_{\mathrm{N}^{IN}}\,|\,T\right]$, is given by an equation that exactly corresponds to Eq.(3.2.13b') of part I. Meanwhile, the second factor is given by:

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$P\left[\begin{matrix}\left\{\vec{\omega}_j=\vec{\rho}_j\left(\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]\right)\right\}_{j=1,2,...,N^X}, \\ \left\{\vec{\omega}(n)=\vec{\rho}\left(\breve{s}(n)\right)\right\}_{n\in \mathrm{N}^{IN}}\end{matrix}\middle| T,\alpha[s_1,s_2,...,s_{N^X}];\left\{s(n)\right\}_{\mathrm{N}^{IN}}\right]$$

$$=\left(\prod_{x=1}^{L(s^{Root})}P\left[\left(\omega_x^{Root},n^{Root}\right)\middle|\,\upsilon_x^{Root}\right]\right)$$

$$\times\prod_{b\in\{b\}_T}P\left[\begin{matrix}\left(\vec{\omega}^A(b)=\vec{\rho}^A\left(\alpha(\breve{s}^A(b),\breve{s}^D(b))\right),\right. \\ \left.\vec{\omega}^D(b)=\vec{\rho}^D\left(\alpha(\breve{s}^A(b),\breve{s}^D(b))\right),\ [n^A(b),n^D(b)]\right)\end{matrix}\middle|\begin{matrix}\left(\breve{s}^A=(s^A,\vec{\omega}^A),n^A(b)\right), \\ \alpha(s^A,s^D)\end{matrix}\right].$$

--- Eq.(A3.10b)

Here each conditional probability in $\prod_{b\in\{b\}_T}(...)$ is given by the left hand side of Eq.(A3.6') slightly modified to fit the evolution along each branch $b$. Thus, the right hand side of Eq.(A3.10b) can be re-expressed as a product over contributions from single columns of $\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]$, where each single column's contribution is also a product of terms coming from different branches and nodes. If we can show that the summation of the right hand side of Eq.(A3.10b) over all possible internal residue states is independent of particular details of the basic states at internal nodes as long as they are consistent with the MSA, then, the summation can be factored out of the summation over $\Sigma\left[\alpha[s_1,s_2,...,s_{N^X}];\left\{n\in\mathrm{N}^{IN}(T)\right\};T\right]$ on the right hand side of Eq.(A3.9a). If so, the total probability of a given extended MSA, $P\left[\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]\,|\,T\right]$ given by Eq.(A3.9a), can be factorized into the basic and residue components:

$$P\left[\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]\,|\,T\right]=P\left[\alpha[s_1,s_2,...,s_{N^X}]\,|\,T\right]$$
$$\times P\left[\left\{\vec{\omega}_j=\vec{\rho}_j\left(\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]\right)\right\}_{j=1,2,...,N^X}\middle|\,T,\alpha[s_1,s_2,...,s_{N^X}]\right].$$

--- Eq.(A3.11)

Here $P\left[\alpha[s_1,s_2,...,s_{N^X}]\,|\,T\right]$ is given by an equation exactly corresponding to Eq.(3.2.13a) of part I. And $P\left[\left\{\vec{\omega}_j=\vec{\rho}_j\left(\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]\right)\right\}_{j=1,2,...,N^X}\middle|\,T,\alpha[s_1,s_2,...,s_{N^X}]\right]$ is the summation of Eq.(A3.10b) over all possible internal residue states filling in a fixed set of basic internal states ($\left\{s(n)\right\}_{\mathrm{N}^{IN}}$) consistent with the MSA. In the following, we will show that this summation is indeed independent of the details of $\left\{s(n)\right\}_{\mathrm{N}^{IN}}$.

First, because Eq.(A3.10b) with the reverse-substitution by Eq.(A3.6') is a product of single-site contributions, we can sort it into a product of column-wise probabilities ($P(\upsilon_i)$'s) over all MSA columns, which are assigned ancestries $\upsilon_1,\upsilon_2,...,\upsilon_{L(\alpha)}$ ($L(\alpha)$ is the number of columns in the MSA):

$$P\left[\begin{matrix}\left\{\vec{\omega}_j=\vec{\rho}_j\left(\alpha[\breve{s}_1,\breve{s}_2,...,\breve{s}_{N^X}]\right)\right\}_{j=1,2,...,N^X}, \\ \left\{\vec{\omega}(n)=\vec{\rho}\left(\breve{s}(n)\right)\right\}_{n\in\mathrm{N}^{IN}}\end{matrix}\middle| T,\alpha[s_1,s_2,...,s_{N^X}];\left\{s(n)\right\}_{\mathrm{N}^{IN}}\right]=\prod_{i=1}^{L(\alpha)}P(\upsilon_i).$$

--- Eq.(A3.12a)

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Because each MSA column, *i.e.*, each site with a given ancestry ($\upsilon_i$), can experience *at most* one insertion, each $P(\upsilon_i)$ can be broadly classified into two forms, as follows. (1) When the site did not experience an insertion, the site already existed at the root node. Thus, considering Eq.(A3.6') and Eq.(A3.9c), we have:

$$P(\upsilon_i) = P\left[\left(\omega(n^{Root};\upsilon_i), n^{Root}\right) \middle| \upsilon_i\right] \prod_{b \in \{b\}_T\left[\upsilon_i, \{s(n)\}_{\mathrm{N}^{IN}}\right]} \left\langle \omega(n^A(b);\upsilon_i) \middle| \hat{P}^S(n^A(b), n^D(b);\upsilon_i) \middle| \omega(n^D(b);\upsilon_i) \right\rangle .$$

--- Eq.(A3.12b)

Here, $\omega(n;\upsilon_i)$ denotes the residue state at the site with the ancestry $\upsilon_i$ in the extended sequence state at the node $n$. And $\{b\}_T\left[\upsilon_i;\{s(n)\}_{\mathrm{N}^{IN}}\right]$ is the set of branches through which the site $\upsilon_i$ continued to exist (*i.e.*, the site experienced no indel event along the branch), given a set of the internal node states ($\{s(n)\}_{\mathrm{N}^{IN}}$). (2) When the site experienced an insertion along a branch $b_I$, we have:

$$P(\upsilon_i) = p_I(\omega(n^D(b_I);\upsilon_i);\upsilon_i, n^D(b_I)) \times \prod_{b \in \{b\}_T\left[\upsilon_i, \{s(n)\}_{\mathrm{N}^{IN}}\right]} \left\langle \omega(n^A(b);\upsilon_i) \middle| \hat{P}^S(n^A(b), n^D(b);\upsilon_i) \middle| \omega(n^D(b);\upsilon_i) \right\rangle .$$

--- Eq.(A3.12c)

Here, inevitably, $\{b\}_T\left[\upsilon_i;\{s(n)\}_{\mathrm{N}^{IN}}\right]$ consists solely of some, but not necessarily all, of the descendant branches of $b_I$.

Now, in a MSA column (with ancestry $\upsilon_i$), consider the summation of $P(\upsilon_i)$'s over all possible residue states at relevant internal nodes. Let $\mathrm{N}^{IN}\left[\upsilon_i;\{s(n)\}_{\mathrm{N}^{IN}}\right]$ be the set of such internal nodes. According to the phylogenetic correctness condition (*e.g.*, Childelevitch et al. 2006) (as mentioned near the bottom of Subsection 3.2 of part I), the union of $\mathrm{N}^{IN}\left[\upsilon_i;\{s(n)\}_{\mathrm{N}^{IN}}\right]$, the set of external nodes with the site of ancestry $\upsilon_i$ ($\mathrm{N}^X[\upsilon_i]$), and $\{b\}_T\left[\upsilon_i;\{s(n)\}_{\mathrm{N}^{IN}}\right]$ always forms a single, connected web in the tree (see the lower part of Subsection 3.4 of part I, and Figures 4 and 7 of part I). The minimum web among such webs is given by the Dollo parsimonious indel history (Farris 1977), and is a union of $\mathrm{N}^X[\upsilon_i]$ and the shortest paths, each of which connects a pair of nodes in $\mathrm{N}^X[\upsilon_i]$. Other webs consistent with the MSA column are formed by continuously extending one or more paths from the minimum web. There are two types of web-extension: upward (toward the root) and downward (toward, but always short of, the external nodes not in $\mathrm{N}^X[\upsilon_i]$). Downward extensions could branch off, as long as the branches do not reach any external nodes. We first handle downward extensions and then handle upward extensions. At each lower-tip of a downward extension, we always encounter a summation of single conditional probabilities, such as $\sum_{\omega(n^D(b);\upsilon_i)\in\Omega} \left\langle \omega(n^A(b);\upsilon_i) \middle| \hat{P}^S(n^A(b), n^D(b);\upsilon_i) \middle| \omega(n^D(b);\upsilon_i) \right\rangle$, which always gives 1 (unity). After repeating this type of summations, the residue state probabilities at nodes on each downward extension (except its origin belonging to the minimum web)

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

leave no effects on the column-wise MSA probability via substitutions. Next, at each upper-tip of an upward extension, we encounter the following summation: either $\sum_{\omega(n^{Root};\upsilon_i)\in\Omega} P\left[\left(\omega(n^{Root};\upsilon_i), n^{Root}\right) \mid \upsilon_i\right]\left\langle\omega(n^{Root};\upsilon_i)\middle|\hat{P}^S(n^{Root}, n^D(b);\upsilon_i)\middle|\omega(n^D(b);\upsilon_i)\right\rangle$ if the tip is the root node (*i.e.*, $n^A(b) = n^{Root}$), or $\sum_{\omega(n^A(b);\upsilon_i)\in\Omega} p_I(\omega(n^A(b);\upsilon_i);\upsilon_i, n^A(b))\left\langle\omega(n^A(b);\upsilon_i)\middle|\hat{P}^S(n^A(b), n^D(b);\upsilon_i)\middle|\omega(n^D(b);\upsilon_i)\right\rangle$ otherwise (*i.e.*, if $n^A(b) = n^D(b_I)$). Each of them is a summation over the initial states, and each summand is the product of an "initial probability" and a single probability conditioned on the initial state. Thanks to Eq.(A3.5'), the latter type of summation can be performed, yielding $p_I(\omega(n^D(b);\upsilon_i);\upsilon_i, n^D(b))$. The former type of summation can also be performed *if we additionally assume* the following equation:

$$\sum_{\omega'\in\Omega} P\left[\left(\omega', n^{Root}\right) \mid \upsilon_i\right]\left\langle\omega'\middle|\hat{P}^S(n^{Root}, n^D(b);\upsilon_i)\middle|\omega(n^D(b);\upsilon_i)\right\rangle$$
$$= p_I(\omega(n^D(b);\upsilon_i);\upsilon_i, n^D(b)) \quad . \qquad \text{--- Eq.(A3.13)}$$

Thus, if Eq.(A3.13) holds, every upward extension can be receded down to its origin ($n^{Ori}$) belonging to the minimum web, providing $p_I(\omega(n^{Ori};\upsilon_i);\upsilon_i, n^{Ori})$. Finally, consider the contribution from a "null site" that is not kept at any external nodes. In this case, after successively performing the summations at the lower-tips of the downward extensions, we are always left with either $\sum_{\omega'\in\Omega} P\left[\left(\omega', n^{Root}\right) \mid \upsilon_i\right]$ or $\sum_{\omega'\in\Omega} p_I(\omega';\upsilon_i, n)$. Each of them always yields 1 (unity). Putting together all these arguments, we see that the probability of the residue states of a MSA column under any indel history can be reduced to that under the Dollo parsimonious indel history. This means that the residue component of the MSA probability, which is the summation of Eq.(A3.10b) over all possible residue states at internal nodes, is indeed independent of the basic sequence states at internal nodes, $\{s(n)\}_{N^{IN}}$. Thus, under the assumptions of Eqs.(A3.1a,b), the column-wise factorization of the substitution evolutionary operators, Eq.(A3.5') and Eq.(A3.13), the MSA probability can indeed be factorized into the basic and residue components, as in Eq.(A3.11).

**A4. Pursuing further biological realism**

Eq.(A3.5') and Eq.(A3.13) in Subsection A3 are *non-equilibrium* generalizations of the famous detailed-balance condition, $\sum_{\omega'\in\Omega}\pi(\omega')\langle\omega'|\hat{P}^S(t', t;\upsilon_i)|\omega\rangle = \pi(\omega)$, for a time-reversible substitution model with the equilibrium residue frequencies $\{\pi(\omega)\}_{\omega\in\Omega}$ and the assumption that the residue content of the inserted subsequences is also given by $\{\pi(\omega)\}_{\omega\in\Omega}$. These widely accepted assumptions played important roles to facilitate the calculations in the past studies with evolutionary models incorporating both substitutions and indels (*e.g.*, Thorne et al. 1991, 1992; Miklós et al. 2004; Rivas and Eddy 2008). However, even if generalized to Eq.(A3.5') and Eq.(A3.13), they may still be too restrictive to accommodate some biologically realistic features. For example, when transposable elements (*e.g.*, Morgante et al. 2007; Chalopin et al. 2015) or foreign DNA sequences (*e.g.*, Waterhouse and Russell 2006) are inserted, the residue content of such inserted sequences is likely to be substantially different from the residue content of the genome that underwent the

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

insertions. It remains to be seen whether we can further relax the conditions to accommodate such situations while keeping the "indel-substitution factorization" enabled. Recently, Lèbre and Michel (2010, 2013) developed some analytical models to examine the effects of the base composition of inserted sequences on the evolution of the base composition of an entire genome or of its subset. It might be interesting to see if their methods are applicable to the issue at hand.

Another potentially important biologically realistic feature is the observation by some studies that the substitution rate increases at sites surrounding insertions/deletions (*e.g.*, Tian et al. 2008; De and Babu 2010). If the incremental substitutions occurred simultaneously or almost simultaneously with the indel events, this feature could be formally incorporated into our theoretical framework by "dressing" each indel operator term in the action of the rate operator, $\langle \breve{s} | \hat{Q}^{SID}$, with substitution operators. The deletion operator, $\hat{M}_D(x_B, x_E)$, could be replaced with:

$$\hat{M}_D(x_B,x_E) \otimes \left[ \sum_{[\omega_{x_{B2}},\ldots,\omega_{x_{E2}}] \in \Omega^{L_{LF}^{CO}+L_{RF}^{CO}}} \sum_{[\omega'_{x_{B2}},\ldots,\omega'_{x_{E2}}] \in \Omega^{L_{LF}^{CO}+L_{RF}^{CO}}} \left\{ \begin{array}{l} p_{\Delta S}\left( \begin{array}{l} [\omega_{x_{B2}},\ldots,\omega_{x_{E2}}] \\ \mapsto [\omega'_{x_{B2}},\ldots,\omega'_{x_{E2}}]; \breve{s}', t \end{array} \right) \\ \times \left( \bigotimes_{x=x_{B2}}^{x_{E2}} \hat{M}_S(x, \omega_x \mapsto \omega'_x) \right) \end{array} \right\} \right] \Bigg|_{\substack{x_{B2} \equiv x_B - L_{LF}^{CO}, \\ x_{E2} \equiv x_B + L_{RF}^{CO} - 1, \\ \langle \breve{s}' | = \langle \breve{s} | \hat{M}_D(x_B, x_E)}} .$$

--- Eq.(A4.1)

Here $p_{\Delta S}\left( [\omega_{x_{B2}},\ldots,\omega_{x_{E2}}](\breve{s}') \mapsto [\omega'_{x_{B2}},\ldots,\omega'_{x_{E2}}]; \breve{s}', t \right)$ is the probability that the residue states of the intermediate state, $\langle \breve{s}' | \equiv \langle \breve{s} | \hat{M}_D(x_B, x_E)$, was replaced as indicated, and satisfies $\sum_{[\omega'_{x_{B2}},\ldots,\omega'_{x_{E2}}] \in \Omega^{L_{LF}^{CO}+L_{RF}^{CO}}} p_{\Delta S}\left( [\omega_{x_{B2}},\ldots,\omega_{x_{E2}}](\breve{s}') \mapsto [\omega'_{x_{B2}},\ldots,\omega'_{x_{E2}}]; \breve{s}', t \right) = 1$. For notational convenience, we have also set $\langle \omega'_x | \hat{M}_S(x, \omega_x \mapsto \omega_x) \equiv \delta(\omega'_x, \omega_x) \langle \omega_x |$, where the subscript $x$ in $\omega_x$ and $\omega'_x$ indicates that they are residue states at the $x$ th site. The $L_{LF}^{CO}$ and $L_{RF}^{CO}$ are the "cut-off" lengths of the left-flanking and right-flanking regions, respectively, that could accommodate the incremental substitutions. The cut-offs were introduced just for convenience. If we can assume that the incremental substitution at each site is independent of those at the other sites, Eq.(A4.1) is reduced to:

$$\hat{M}_D(x_B,x_E) \otimes \left[ \bigotimes_{x=x_{B2}}^{x_{E2}} \left\{ \sum_{\omega_x \in \Omega} \sum_{\omega'_x \in \Omega} \left( p_{\Delta S}(x, \omega_x \mapsto \omega'_x; \breve{s}', t)\, \hat{M}_S(x, \omega_x \mapsto \omega'_x) \right) \right\} \right] \Bigg|_{\substack{x_{B2} \equiv x_B - L_{LF}^{CO}, \\ x_{E2} \equiv x_B + L_{RF}^{CO} - 1, \\ \langle \breve{s}' | = \langle \breve{s} | \hat{M}_D(x_B, x_E)}} .$$

--- Eq.(A4.1')

Here the site-wise incremental substitution probability, $p_{\Delta S}(x, \omega_x \mapsto \omega'_x; \breve{s}', t)$, satisfies the equation, $\sum_{\omega'_x \in \Omega} p_{\Delta S}(x, \omega_x \mapsto \omega'_x; \breve{s}', t) = 1$. We can also "dress" the insertion operator, $\hat{M}_I(x,l)\hat{F}(x, \delta\vec{\omega}'[l])$, in a similar manner. The expression of a dressed insertion operator becomes bulkier than Eq.(A4.1), and thus is omitted here. Once the incremental substitutions are introduced as in Eq.(A4.1), we cannot easily perform the "indel-substitution factorization" of the alignment probabilities, because the expected number of substitutions increases with the number of indels in a local history.

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Therefore, if the incremental substitutions occur commonly along the genome, the above line of arguments is no longer applicable for the separation of basic and residue components of the *entire* alignment probability. Nevertheless, the alignment probability may still be factorable into the product of local contributions, possibly with some modifications in the arguments (in Section 4 of part I (Ezawa, Graur and Landan 2015a)) and the models (given in Section 5 of part I). It remains to be seen whether we can still factorize the alignment probability into the basic and residue components by substantially extending and/or modifying the arguments given in this subsection. Unless we can, one solution might be to develop an "effective substitution model" that takes beforehand account of the effects of such incremental substitutions in the vicinity of indels (including invisible ones).

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# References


Arndt PF, Hwa T. 2005. **Identification and measurement of neighbor-dependent nucleotide substitution processes.** *Bioinformatics* **21**:2322-2328.

Bailey JA, Eichler EE. 2006. **Primate segmental duplications: crucibles of evolution, diversity and disease.** *Nat Rev Genet* **7**:552-564.

Bradley RK, Holmes I. 2007. **Transducers: an emerging probabilistic framework for modeling indels on trees.** *Bioinformatics* **23**:3258-3262.

Britten RJ. 2002. **Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels.** *Proc. Natl. Acad. Sci. USA* **99**:13633-13635.

Britten RJ, Rowen L, Willians J, Cameron RA. 2003. **Majority of divergence between closely related DNA samples is due to indels.** *Proc. Natl. Acad. Sci. USA* **100**:4661-4665.

Cartwright RA. 2005. **DNA assembly with gap (Dawg): simulating sequence evolution.** *Bioinformatics* **21**:iii31-iii38.

Cartwright RA. 2009. **Problems and solutions for estimating indel rates and length distribution.** *Mol Biol Evol.* **26**:473-480.

Chalopin D, Naville M, Plard F, Galiana D, Volff JN. 2015. **Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates.** *Genome Biol Evol*. 7:567-580.

Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrionos GP. 2007. **Gene conversion: mechanisms, evolution and human disease.** *Nat Rev Genet*. **8**:762-775.

Chindelevitch L. Li Z, Blais E, Blanchette M. 2006. **On the inference of parsimonious evolutionary scenarios.** *J Bioinform Comput Biol.* **4**:721-744.

De S, Babu M. 2010. **A time-invariant principle of genome evolution.** *Proc Natl Acad Sci USA* **107**:13004-13009.

Ellegren H. 2004. **Microsatellites: simple sequences with complex evolution.** *Nat Rev Genet.* **5**:435-445.

Ezawa K, Ikeo K, Gojobori T, Saitou N. 2010. **Evolutionary pattern of gene homogenization between primate-specific paralogs after human and macaque speciation using the 4-2-4 method.** *Mol Biol Evol*. **27**:2152-2171.

Ezawa K, Ikeo K, Gojobori T, Saitou N. 2011. **Evolutionary patterns of recently emerged animal duplogs.** *Genome Biol Evol*. **3**:1119-1135.

Ezawa K, Landan G, Graur D. 2013. **Detecting negative selection on recurrent mutations using gene genealogy.** *BMC Genetics* **14**:37.

Ezawa K, Graur D, Landan G. 2015a. **Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part I: Theoretical basis.** (unpublished, to be posted on bioRxiv.)

Ezawa K, Graur D, Landan G. 2015b. **Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part II: Perturbation analyses.** (unpublished, to be posted on bioRxiv.)

Ezawa K, Graur D, Landan G. 2015c. **Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part III: Algorithm for first approximation.** (unpublished, to be posted on bioRxiv.)

Farris JS. 1977. **Phylogenetic analysis under Dollo's law.** *Syst Zool.* **26**:77-88.

Fawcett JA, Innan H. 2013. **The role of gene conversion in preserving rearrangement hotspots in the human genome.** *Trends in Genetics* **29**:561-568.

Felsenstein J. 1981. **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol.* **17**:368-376.

Felsenstein J. 2004. *Inferring Phylogenies.* Sunderland (MA), Sinauer Associates.

Gascuel O (editor). 2005. *Mathematics of Evolution and Phylogeny.* New York, Oxford University Press.

Graham SW, Reeves PA, Burns ACE, Olmstead RG. 2000. **Microstructural changes in noncoding chloroplast DNA: Interpretation, evolution, and utility of indels and**

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**inversions in basal angiosperm phylogenetic inference.** *Int J Plant Sci* **161**:S83-S96.

Graur D, Li WH. 2000. *Fundamentals of Molecular Evolution, 2nd ed.* Sunderland (MA), Sinauer Associates.

Gu W, Zhang F, Lupski JR. 2008. **Mechanisms for human genomic rearrangements.** *PathoGenetics* **1**:4.

Kelshner SA, Wendel JF. 1996. **Hairpins create minute inversions in non-coding regions of chloroplast DNA.** *Curr Genet* **30**:259-262.

Kent WJ, Baertsch R, Hinrichs A, Miller W, and Haussler D. 2003. **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.** *Proc Natl Acad Sci USA* **100**:11484-11489.

Lèbre S, Michel CJ. 2010. **A stochastic evolution model for residue insertion-deletion independent from substitution.** *Comput Biol Chem.* **34**:259-267.

Lèbre S, Michel CJ. 2013. **A new molecular evolution model for limited insertion independent of substitution.** *Math Biosci.* **245**:137-147.

Lunter G. 2007. **Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes.** *Bioinformatics* **23**:i289-i296.

Lunter G, Hein J. 2004. **A nucleotide substitution model with nearest-neighbour interactions.** *Bioinformatics* **20**:i216-i223.

Lynch M. 2007. *The Origins of Genome Architecture.* Sunderland (MA), Sinauer Associates.

Miklós I, Lunter GA, Holmes I. 2004. **A "long indel" model for evolutionary sequence alignment.** *Mol Biol Evol.* **21**:529-540.

Miklós I, Novák Á, Satija R, Lyngsø R, Hein J. 2009. **Stochastic models of sequence evolution including insertion-deletion events.** *Stat Methods Med Res.* **18**:453-485.

Morgante M, De Paoli E, Radovic S. 2007. **Transposable elements and the plant pan-genomics.** *Curr Opin Plant Biol.* **10**:149-155.

Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. 2008b. **Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs.** *Genome Res.* **18**:1814-1828.

Rivas E. 2005. **Evolutionary models for insertions and deletions in a probabilistic modeling framework.** *BMC Bioinformatics* **6**:63.

Rivas E, Eddy SR. 2008. **Probabilistic phylogenetic inference with insertions and deletions.** *PLoS Comput Biol.* **4**:e1000172.

Sainudiin R, Durrett RT, Aquadro CF, Nielsen R. 2004. **Microsatellite mutation models: insights from a comparison of humans and chimpanzees.** *Genetics* **168**:383-395.

Saitou N, Kitano T. 2013. **The PNarec method for detection of ancient recombinations through phylogenetic network analysis.** *Mol Phylogen Evol.* **66**:507-514.

Teshima KM, Innan H. 2012. **The coalescent with selection on copy number variants.** *Genetics* **190**:1077-1086.

The Chimpanzee Sequencing and Analysis Consortium. 2005. **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* **437**:69-87.

The International Chimpanzee Chromosome 22 Consotrium. 2004. **DNA sequence and comparative analysis of chimpanzee chromosome 22.** *Nature* **429**:382-388.

Thorne JL, Kishino H, Felsenstein J. 1991. **An evolutionary model for maximum likelihood alignment of DNA sequences.** *J Mol Evol.* **33**:114-124.

Thorne JL, Kishino H, Felsenstein J. 1992. **Inching toward reality: an improved likelihood model of sequence evolution.** *J Mol Evol.* **34**:3-16.

Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ. 2008. **Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes.** *Nature* **455**:105-108.

Waterhouse JC, Russell RR. 2006. **Dispensable genes and foreign DNA in Streptococcus mutants.** *Microbiology* **152**:1777-1788.

bioRxiv preprint doi: https://doi.org/10.1101/023622; this version posted July 31, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Yang Z. 2006. *Computational Molecular Evolution.* New York (NY), Oxford University Press.