KEZW-BI-ME/00005

August 13, 2020

New Perturbation Method to Compute Probabilities of Mutually Adjoining Insertion-type and Deletion-type Gaps in Ancestor-Descendant Pairwise Sequence Alignment under *Genuine* Sequence Evolution Model with *Realistic* Insertions/Deletions: — the "Last Piece of the Puzzle" —

Kiyoshi Ezawa

Independent (ORCID: 0000-0003-4906-8578)

Current address: 3-1-33 Nakamura-machi, Chichibu 368-0051, JAPAN;

Phone & Fax: +81-494-22-5501; E-mail: kezawa.ezawa3@gmail.com

Abstract

Accurately estimating the probabilities of pairwise-alignments (PWAs) of ancestral and descendant sequences is essential when aiming for precise evolutionary analyses on homologous (*i.e.*, ancestor-sharing) sequences. Doing so in the presence of *realistic* insertions/deletions, however, has been hampered by many formidable technical challenges. Recently [1], based on the theoretical formulation dealing with stochastic models of sequence evolution with realistic insertions/deletions [2, 3], we invented an algorithm to compute practically exact probabilities of insertion-type gaps alone and those of deletion-type gaps alone. However, accurate computation of the final type of gapped segments (called case-(iv)), in each of which intersion-type gaps and deletion-type gaps coexist and adjoin each other, has been left unresolved as the *"last piece of the puzzle"* of accurately estimating the probabilities of ancestor-descendant PWAs.

Here, we construct a new perturbation method to provide this "last piece of the puzzle", that is, to accurately, and *systematically*, compute the probabilities of case-(iv) gapped segments. In short, this new method classifies (or "colors") the sites in the sub-sequence evolution in each case-(iv) segment into "ancestral" and "descendant" types; then it considers only those insertions/deletions which change the coloring-pattern of the sub-sequence as "perturbations"; and finally it computes the probabilities of the contributing evolutionary histories of the sub-sequence from the lowest-order perturbation terms upward. Our *in silico* "experiments" indicated that this new method computes probabilities quite accurately, even with only 2nd- and 3rd-order terms.

Combining the results of this study and of a previous work of ours [1], we should now be able to estimate the probabilities of ancestor-descendant PWAs quite accurately.

Thus, this study represents a significant step toward the ultimate goal of precise evolutionary analyses on homologous sequences. The method reported here has been implemented in an open-source package of prototype Perl scripts and modules, named "LAST-PIECE(_P)", which is available at the FTP repository of the ANEX project in Bioinformatics.org (https://www.bioinformatics.org/ftp/pub/anex/).

[**Keywords:** pairwise sequence alignment (PWA), probability, evolution, (stochastic) sequence evolution model, insertion/deletion (indel), accurate computation, perturbation method, DNA sequence, biological sequence]

[**Abbreviations:** hidden Markov model (HMM), multiple sequence alignment (MSA), preserved ancestral site (PAS), pairwise sequence alignment (PWA), Thorne-Kishino-Felsenstein (TKF)]

Contents

[1] The "perturbation method" unfolded here is somewhat similar to, yet significantly different from, the perturbation-theoretical approach in our previous studies [3, 1]; especially, the "base"-"perturbation" decomposition of the rate operator, $\hat{Q}^{ID}$, differs between the two approaches. Here, the probabilities computed via the latter approach (so that the evolutionary histories include as much indels as desired) are regarded as "exact"-solutions (under the "base" rate operator here), and thus the term, "perturbation", always refers to that in the method constructed here.

1 Introduction

1.1 Background

Aligning homologous, *i.e.*, ancestor-sharing, sequences is a cornerstone of homology-based analyses of bio-molecular sequences, such as DNA, RNA, and protein sequences (*e.g.*, [4, 5, 6, 7, 8, 9]). Traditionally, a sequence alignment has been represented in the form of a matrix, in which each row represents a sequence and each row, often called a "site", represents a set of residues descended from a common ancestral residue (*e.g.*, [4, 10]); when a sequence lacks a residue, a gap (usually denoted by a dash, "-") is placed in the corresponding cell. Depending on the number of sequences involved, there are two major categories of sequence alignments: a pairwise (sequence) alignment (PWA) that consists of two sequences (*e.g.*, [11, 12]), and a multiple sequence alignment (MSA) that consists of three or more sequences (*e.g.*, [13, 14, 4]). In principle, the alignment of some homologous sequences is a product of the evolution of the sequence, and should be determined *uniquely* by the evolutionary events

such as insertions and deletions (often referred collectively as "indels").[2] [3]

The inevitable serious problem is that the nature will *never* give us such a *true* sequence alignment *as it is*; all we are left with is a set of extant sequences consisting *only* of residues, *i.e.*, with all gaps removed. This means that we have to *infer*, or *reconstruct*, the alignment from the extant sequences, by placing gaps so that they will collectively represent some plausible (and hopefully close to true) evolutionary history of the sequences. At least up to now, the prevalent practice is to (attempt to) find an alignment that *optimizes* a score function prescribed by certain rules (*e.g.*, [11, 12, 24, 13, 14, 4, 25, 26, 27, 10, 28]).

Unfortunately, such reconstruction of sequence alignments has turned out to be a very tough, error-prone, process; some recent studies on the sequence alignment errors indicated that a considerable fraction, often even a majority or a near-majority, of gaps are erroneous (*e.g.*, [10, 29, 30]). Regarding the causes of such alignment errors, some studies revealed that the inherent stochasticity of the sequence evolution plays an important, or dominant, role (*e.g.*, [31, 32, 30]). In other words, the *true* sequence alignment very frequently is *not* the optimum, because the evolutionary processes are stochastic.[4] These studies suggest that *merely* finding a *single* optimum alignment is *not* enough, and that the most *truthful* way should be to present a *probability distribution* of alignments inferred from the set of extant sequences.

Actually, the idea of providing the probability distribution of alignments, often referred to

[2] An insertion of a residue creates a column with gaps in *all* sequences *except* the offsprings of the ancestral sequence that underwent the insertion. A deletion of a residue creates a column with gaps *in* the offsprings of the ancestral sequence that experienced the deletion.

[3] In this study, we deal with *collinear* sequence alignments; an alignment is called "collinear" (*e.g.*, [15]) if it is devoid of genomic rearrangements such as inversions, duplications, and translocations (*e.g.*, [16, 17, 18, 19]). (Possibly non-collinear) alignments of (usually very long) sequences that possibly underwent such genomic rearrangements are called "genome alignments", and they are *not* the subject of this study. Readers who are interested in genome alignments should refer, *e.g.*, to: [15, 20, 21, 22, 23].

[4] In fact, this is the case even with the ideal score of the log-probability under the stochastic evolution model that actually created the true alignment.

as "*statistical alignment*" [33], is not new. As far as we know, the first attempt to compute the probabilities of pairwise sequence alignments and use them as scores was made in the ground-breaking work (in 1986) by Bishop and Thompson [34], who used a simple probabilistic model to assign "probabilities" to gaps and gapless columns . Then, in another ground-breaking work (in 1991), Thorne, Kishino and Felsenstein proposed an evolution model of biological sequences allowing only single-residue insertions/deletions (indels) (TKF91, in [35]). Then, they further attempted to "inch toward" the realistic sequence evolution, by proposing an evolution model that permits multi-residue indels of some sorts by considering a biological sequence as a sequence of fragments, each of which consisting of one or more residues, and by allowing only single-fragment indels (TKF92, in [36]). The computation of alignment probabilities under these TKF models were later cast into standard Hidden Markov Models (HMMs) (*e.g.,* [33, 37]), which were easier to handle than the original computational procedures derived by TKF [35, 36]. Although standard HMMs have flaws similar to those of the TKF91 & TKF92 models, such as the inability to *faithfully* take account of nested and/or overlapping indels, there were still some attempts to incorporate some effects of overlapping indels into standard HMMs (*e.g.*, [37]). Meanwhile, as an attempt to handle more realistic sequence evolution models, Miklós and Toroczkai [38] proposed a method to compute alignment probabilities under an evolution model that allows any lengths of insertions (with geometrically distributed rates) but *only* single-residue deletions.

The evolution models handled up to then were dissatisfactory in the sense that their insertion/deletion processes are *not* realistic, in at least two ways: (1) most of them permit only single-residue indels or multi-residue indels with the geometrically distributed rates, whereas many empirical studies indicate that multi-residue indels occur *quite often*, and with the rates following *power-law* (*e.g.*, [39, 40, 41, 42, 43, 44, 45])[5]; and (2) most of them (or all except [38] and [37]) can*not* take account of nested and/or overlapping indels, which

[5]Some studies based on standard HMMs alleviated this flaw (*not fully* but to some extent) by using mixed geometric distributions (*e.g.*, [25, 31]).

are expected from *natural* evolution processes of biological sequences.

In our view, a satisfactory method was first proposed in the milestone work (in 2004) by Mikós, Lunter, and Holmes [2]. Their "long indel" model is a space-homogeneous[6] *genuine* sequence evolution model [7] of a biological sequence that can in principle incorporate any indel length distributions, including biologically realistic power-laws. This study [2] made a couple of new achievements: (i) they verbally proved that, *under the "long indel" model* [8], the probability of an ancestor-descendant PWA[9] can be factorized into the product of the probabilities of "chop zones", each of which is delimited by a gapless column (or the left-end of the alignment) and the next gapless column (or the right-end of the alignment), excluding the former column and including the latter one; (ii) after arguing that the probability of each chop zone *should* be approximated well by the summation of the probabilities of indel histories with less than an arbitrary number of indels which are shorter than an arbitrary

[6] A sequence evolution model is called "space-homogenous" if its rates of evolutionary events are uniform, *i.e.*, independent of the positions, along the sequence.

[7] A sequence evolution model is regarded as *genuine* only when it follows the evolutionary principle; The evolutionary principle requires that, when a certain time-interval is divided into two or more sub-time-intervals, the probability of each evolution process of a sequence during the time-interval must result from multiplying the probabilities of sub-processes of the sequence *as a whole* during all these sub-time-intervals. In other words, the evolutionary principle dictates that the probability of each evolution process can be factorized *vertically* (*i.e.*, along the time-interval), in contrast to horizontally (*i.e.*, along the sequence) as HMMs do. In the context of a continuous-time Markov model, the evolutionary principle, in conjunction with the "completeness" of the set of (intermediate) states, leads to the famous Chapman-Kolmogorov equation, which in turn is equivalent to the defining equation of the finite-time transition probability operator (*e.g.*, Eq. (R3.18) in [3]).

[8] This factorization of the PWA probability does not necessarily work in a *genuine* sequence evolution model *in general*. A previous study of ours [3] provided a set of conditions under which the alignment probability in a *genuine* sequence evolution model is factorable. It is exactly because the "long indel" model satisfies these conditions that the alignment probability in the model is factorable.

[9] Although they [2] discussed simply a PWA of two sequences in general, it is actually equivalent to an ancestor-descendant PWA, because the "long indel" model is time-reversible. Here we specifically put forth an ancestor-descendant PWA because it is the main focus of this study.

length (finite trajectory approximation), they provided an algorithm to recursively calculate the probability of each individual indel history (referred to as the "trajectory likelihood"); (iii) they provided a dynamic programming algorithm of $O(L^4)$ time-complexity, where L is the size of the sequence lengths, to compute the joint probability of two sequences from the probabilities of the chop zones; (iv) they also came up with an algorithm (a sort of Viterbi algorithm [46]) to search for an alignment with the largest probability under the "long indel" model; although they did not explicitly show it, they used it to validate their method using a set of structural sequence alginments (HOMSTRAD [47]).

Since the aforementioned milestone work by by Mikós, Lunter, and Holmes [2], more than a decade had passed without any particular advances regarding the application of the *genuine* sequence evolution models to the analytical or deductive studies, although some advances had been made on the study using the simulators of *genuine* sequence evolution (*e.g.*, [48, 49, 50]). It was our previous studies [3, 1, 30] that made some advances for the first time since [2] in the deductive study of statistical alignment under the *genuine* sequence evolution models. One of our major achievements was to answer the question on the factorability of alignment probabilities [3]. More precisely, after defining a "gapped segment" as delimited by a gapless column and the next gapless column (similarly to the definition of the aforementioned "chop zone") but *excluding both*, and we first reformulated the factorability problem so that we will ask whether the alignment probability (under *general*, *genuine* sequence evolution model) can be factorized into the product of an overall factor and the contributions from the gapped segments. Then, we provided the conditions on the indel rates and the exit rate under which the alignment probability is indeed factorable, for ancestor-descendant PWAs and for MSAs.

Another major achievement was to provide concrete methods to calculate the contribution from each gapped segment [1]. The "long indel" paper [2] *did* clearly provide methods to compute alignment probabilities and the joint probability of a sequence pair using the probabilities of chop zones as a building block, as well as an algorithm to compute the probability of each individual indel history that will contribute to the probability of a chop

zone. In that paper ([2]), however, it was unclear how the probability of each chop zone was computed. In a previous paper of ours [1], we addressed this problem. First, we classified the gapped segments (, each of which in an ancestor-descendant PWA is simply a chop-zone with the gapless column on its right removed,) into four major categories, which we referred to as case-(i), -(ii), -(iii), and -(iv) segments: a case-(i) "gapped" segment contains no ancestral or descendant residues; a case-(ii) segment contains some ancestral residues but no descendant residues; a case-(iii) segment contains some descendant residues but no ancestral residues; and a case-(iv) segment contains some ancestral residues and some descendant residues, and none of them are homologous to one another.[10] , [11] Then, applying either of the two defining integral equations of sequence evolution (Eqs.(R4.4) &(R4.5) in [3]) to each gapped segment, we derived the algorithms to *numerically* compute "practically exact" probabilities of (the gap configurations of) case-(i), (ii), and (iii) gapped segments [1]. *Because of technical difficulties,* however, we were not able to give an algorithm to compute practically, or almost, exact probabilities of case-(iv) gapped segments; so we settled for providing methods to compute *all* contributions from parsimonious and next-to-parsimonious indel histories that are responsible for each case-(iv) segment [1].

So, combining the results of Mikós, Lunter, and Holmes [2] and our previous results [3, 1], we are now *almost* in a position to compute "practically exact" probabilities of ancestor-descendant PWAs: we now know that ,under a *genuine* sequence evolution model satisfying

[10]Following the viewpoint of [2], we don't care about the relative order between the ancestral and descendant residues in each case-(iv) gapped segment,, because we consider an alignment as a homology structure [51, 52], in which the order makes sense only among residues in the same sequence(, and in which homologous residues are vertically aligned to each other, of course).

[11]In terms of gap-configurations, a case-(i) "gapped" segment contains no gaps, a case-(ii) segment contains only "deletion-type" gaps in the descendant sequence, a case-(iii) segment contains only "insertion-type" gaps in the ancestral sequence, and a case-(iv) segment contains both "insertion-type" gaps in the ancestral sequence and "deletion-type" gaps in the descendant sequence, which are adjoining each other, *i.e.*, *not* mediated by any gapless columns.

the factorability conditions, the alignment probability is factorized into the product of an overall factor and the contributions from gapped segments; and we already have algorithms to compute practically exact probabilities of case-(i), (ii) and (iii) gapped segments; we are now left with *only one* issue, which is computing the practically (or nearly) exact probabilities of case-(iv) gapped segments. The purpose of this study is to attempt to provide this ***last piece of the "puzzle"*** of computing the probability of ancestor-descendant PWAs. Once this "puzzle" is solved, it will be relatively easy to apply the solution to a wide range of problems, such as the construction of probabilistic ancestor-descendant PWAs, as well as the approximate construction of probabilistic MSAs. Therefore, this "last piece of the puzzle", if provided, will greatly enhance our ability to *truthfully* reconstruct the alignments of homologous sequences based on *genuine* sequence evolution models.

In this study, we will attempt to provide the "last piece" by constructing a *new* perturbation method, which divides the instantaneous rate operator of a *genuine* sequence evolution model into "base" and "perturbation" parts in a different manner than the *old* perturbation method previously provided [3, 1]. In the *old* method, the "base" part consisted only of the exit rate term and the "perturbation" parts contained *all* insertion and deletion operators and rates accompanying them. Therefore, the perturbation level in the old method is nothing other than the number of indel events constituting each indel history. In contrast, in the new method, even the zeroth-order probabilities may be contributed from some indel histories consisting of a large number of indels each. Therefore, there is a hope that, in this new perturbation method, even the summation of low order terms could give a good approximation of the exact probabilities (of case-(iv) gapped segments), and we will see that this is indeed the case.

1.2 Notes on recent developments

After having finished this study and while preparing this manuscript, we learned that there have been a couple of new developments regarding this subject of *accurate* PWA probability computation [53, 54, 55]. First, in an attempt to speed up Miklós et al.'s approach to compute PWA probabilities [2], a simulation-based approach has been devised [53]. A key strategy of theirs is to decouple the stage of computing the "chop-zone" probabilities and the stage of PWA inference, by re-using the "chop-zone" probabilities pre-computed in the first stage. [12] We view this development very favorably, and we hope that they will go on to extend their efforts to statistical MSA problems. (Although the current implementation seems to use only geometric length distributions, incorporating, *e.g.*, power-law distributions does not seem so hard.) In principle, this simulation-based approach could also solve exactly the same problem that we are about to address in this paper. *Nevertheless*, we strongly believe that this *deductive* study of ours is still *indispensable*, for *at least* two reasons. One reason is that simulation-based approaches and deductive approaches are two essential and *complementary* kinds of approaches, on *both of* which any scientific disciplines have been developed and advanced properly. It is *only if* these two kinds of approaches are advanced soundly that the study of a scientific discipline will develop in a wholesome manner. The other reason is that we, human-beings, are, after all, creatures of reasoning. Every time a new fact is revealed, there arise strong demands for the reason(s) behind it. Generally, deductive studies are much better at satisfying these demands than simulation-based studies. Therefore, we consider it *absolutely necessary* to disclose this *deductive* study of ours now, even after the advent of a simulation-based approach [53].

Second, there have been a couple of attempts to incorporate the effects of overlapping indels into the framework of (pair-)HMMs [54, 55], although with the instantaneous in-

[12] As you can see, this strategy is very similar to the strategy employed by ANEX, a program package we recently developed to address statistical MSA problem under *genuine* sequence evolution models [56]. See also footnote 30 below.

del length distributions being geometric. Especially in [55], an attempt has been made to approximate the master equation of continuous-time Markov model (with geometric indel length distributions) within (pair-)HMMs (or finite-state automata). We consider that these attempts are laudable. At the same time, however, we find it a pity that these studies confined themselves in biologically unrealistic geometric indel length distributions. We wish that they could have extended their efforts to more biologically realistic power-law indel length distributions, or *at least* to mixed geometric distributions. [13]

1.3 Structure of This Paper

This article is structured as follows. In Section 2, we explain the basic ideas and principles underlying the method proposed here. Then, in Section 3, we propose the basic framework based on the theory we previously provided [3, 1], which extends a space-homogeneous theory [2] (see also [48]) to more general situations. Based on the proposed basic framework, in Section 4, we construct a *new* "perturbation theory" that enables us to compute the probabilities of case-(iv) gap-patterns as accurately as desired. The next section (Section 5) unfolds the concrete computations at the 2nd- and 3rd-order levels in this new "perturbation

[13] Incidentally, many (or most) of the combinations of parameter values in the simulation study of [55] seem unrealistic for a study of homologous sequences. In our view, the product $t \times \mu$ (, where t is the time-lapse and μ is the deletion rate,) must be less than the upper-bound of about $1/4$ ($\approx \ 4 \times 1/16$, where the 4 (substitutions/site) is an (*extremely generous*) upper-limit of the evolutionary distance with detectable homology (via residue configurations), and the $1/16$ (deletions/substitution) is a half of the $1/8$ (indels/substitutions) estimated in a genome-wide analysis [57]); and the product should usually be *much less* than this upper-bound (*i.e.*, $1/4$). Otherwise, it should be extremely hard, or impossible, to detect homology between the ancestral and descendant sequences. For example, when $t = 8$ and $\mu = 0.5$, we expect that each site suffers (at least) 4 ($= 8 \times 0.5$) deletions on average!; (actually, though, each site can suffer *at most* one deletion, because it will never come back once it is deleted); how is it possible to detect homology in such a circumstance? We strongly urge the researchers (especially theoretical ones) in this field to have a decent sense of biologically realistic scales, which will help avoid futile disputes, for example.

theory." Then, in Section 6, we provide a prototype algorithm to compute the probabilities to the third-order level of "perturbation", and demonstrate how accurate its results are. Finally, in Section 7, we discuss some of possible further developments and outstanding issues.

For clarity, in this article, we focus only on the sequence evolution via insertions/deletions, by assuming that the probability of a sequence alignment (given a fixed phylogenetic tree, if necessary) can be decoupled into the product of the probability under the substitution model and the probability under the insertion/deletion model. This decoupling can be done if the sequence evolution model satisfies a set of conditions, for example, as we showed before [30] (via a generalization of the proof by Kim and Sinha [58]), the following three conditions will suffice: (i) the indel rates (excluding the multiplication factors assigned to the inserted residues) are independent of the residue state and the substitution process before the indel event; (ii) the substitution rates at each site are independent of the states and the evolutionary processes at other sites; and (iii) the probability of the residue state of each inserted sub-sequence (conditioned on the insertion) can be factorized into the product of residue probabilities (at the time) over the inserted sites. This focusing on insertions/deletions significantly simplifies the problem at hand.

Then, we also assume that the indel evolution model we deal with here satisfies the "sufficient and nearly necessary" set of conditions for factorable probabilities of ancestor-descendant PWAs [3], namely: (1) the rate of each (indel) event is independent of the region(s) outside of the site(s) it affect; and (2) the increment of the exit rate by each indel event is independent of the region(s) outside of the site(s) it affect.

2 Underlying ideas and principles

In a previous work of ours [1], we provided a pair of algorithms to compute the probabilities of isolated gaps, *i.e.*, case-(ii) gapped segments having only ancestral sites and case-(iii)

gapped segments having only descendant sites. As a matter of fact, these probabilities were relatively easy to compute, because all the sites under consideration (in between a pair of preserved ancestral sites (PASs) belong either only to the ancestor or only to the descendant; in the former case, those sites are destined to be deleted eventually, and in the latter case, NONE of those sites existed at the initial time (i.e., in the ancestral sequence). This means that, when dealing with each case-(ii) or -(iii) gapped segment, we do NOT need to keep track of the *origins* of the sites (in between the PASs) during the evolution, and this effectively enabled us to focus only on the evolution of the *number* of those sites (or, equivalently, the length of the region).

In contrast, each case-(iv) gapped segment consists of some ancestral sites and some descendant sites that are not homologous to any ancestral ones. Therefore, when considering the evolution of the region in question, we need to keep track of the origins of the sites, that is, whether each site existed from the beginning (of the time-interval) or was newly inserted at some point. This fact makes it much more difficult to compute the probabilities of of case-(iv) gap-patterns than to compute the probabilities of case-(ii) and -(iii) gap-patterns.

Nevertheless, by carefully considering the situations in question, as well as the nature of insertions and deletions, the problem could be surprisingly simplified, as we explain in the following.

Let us consider a general indel history that results in a case-(iv) gapped segment (*e.g.*, Figure 1). First, when focusing on the segment alone in a PWA, it consists of three kinds of ingredients: (i) two preserved ancestral sites (PASs), where an ancestral site is aligned with a descendant site, that flank the segment [14]; (ii) some "ancestral sites" that were deleted at some point in the time-interval; and (iii) some "descendant sites" that were inserted at some point in the time-interval. Next, when considering the indel history as well, there may also be (iv) "evanescent" (or "transient") sites, which were inserted at some point and deleted

[14]Strictly speaking, these PASs do *not* belong to the gapped segment. However, when computing the probability, we will consider that the segment includes the PASs.

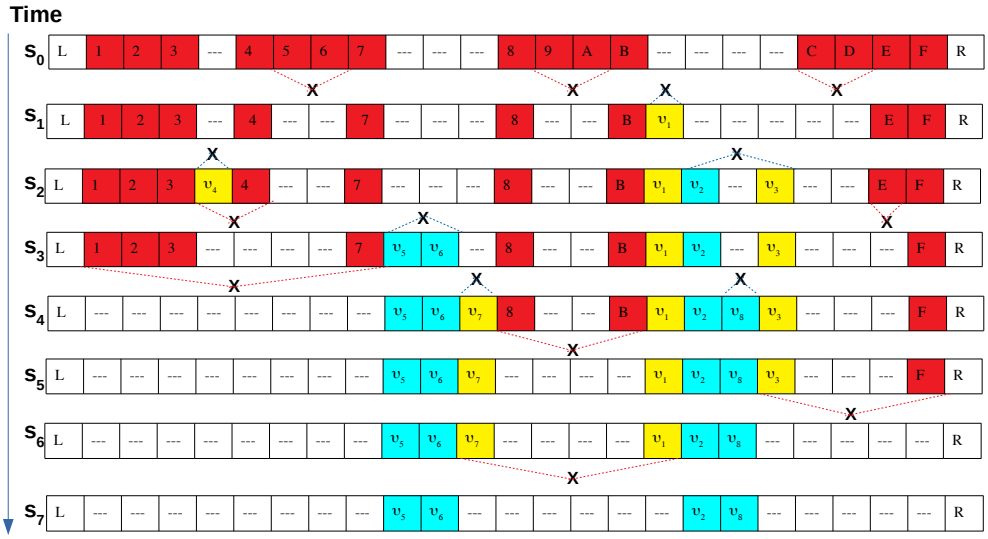

Figure 1: **Time-trajectories of ancestral (red), descendant (cyan), and evanescent (yellow) sites in example indel history that results in case-(iv) gapped segment.** Sequence states at different time-points, labelled $s_0, s_1, ..., s_7$, are aligned just as in a normal multiple sequence alignment (MSA). Sites sharing the same ancestry are aligned vertically. The '—' in a white cell indicates that the sequence lacked the site at that time. The pair of blue dashed lines diverging from an "X" represents an insertion. The pair of red dashed lines converging to an "X" represents a deletion. The "1, 2, ..., 9, A, ..., F" in the red cells are the ancestry indexes of the ancestral sites.The ancestry indexes of the inserted sites (cyan and yellow boxes), υ_j (with $j = 1, 2, ...8$), were assigned in the temporal order of creation.

at some later point.

For the moment from now, let us ignore the "evanescent" sites (as in Figure 2). When constructing a PWA, especially when the homology structure alone is important, we could completely segregate ancestral sites from descendant sites. When considering the time-trajectories of those sites, however, we observe that the ancestral sites and descendant sites were actually intermingled with one another, as shown in Figure 2. This is because the sites in a (sub-)sequence at each time-point must follow a strict spacial order; site A on the left of site B must NEVER be placed on the right of B. Taking another look at the time-trajectories of the sites (e.g., in Figure 2), we notice that, at each time-point, some ancestral sites are lumped together against the background of surrounding descendant sites, to form a "cluster"; similarly, some descendant sites form a "cluster" against the background of surrounding ancestral sites. By introducing symbols "A" and "D" to denote a "cluster" of ancestral sites and a "cluster" of descendant sites, respectively, we could represent the

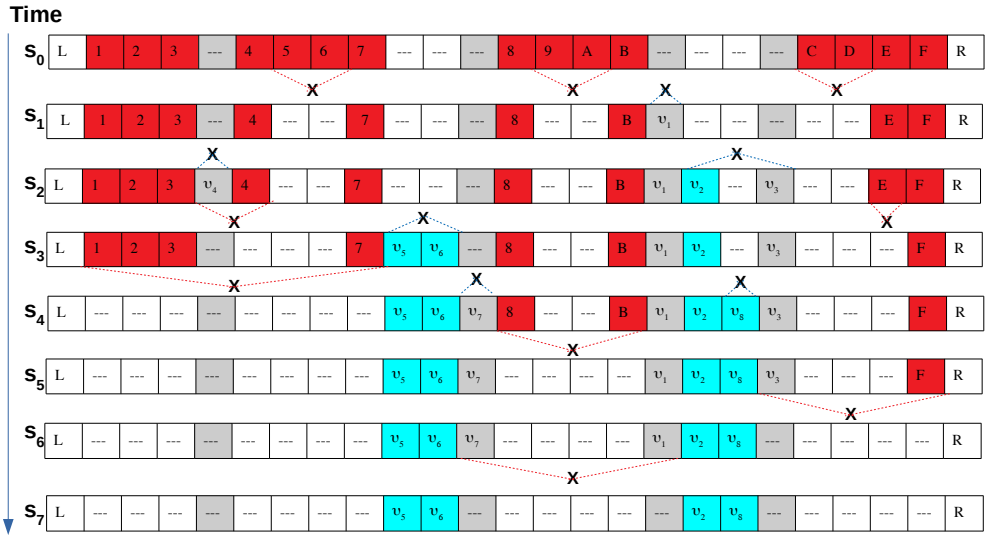

Figure 2: **Time-trajectories of ancestral (red) and descendant (cyan) sites, when we ignore evanescent sites (shaded in gray).** This figure differs from Figure 1 only in how the evanescent sites are colored/shaded. Now, the evolution of the "A"/"D"-coloring-pattern has become clearer. The pattern evolution here is: "A "(s_0 & s_1) → "ADA" (s_2) → "ADADA" (s_3) → "DADA" (s_4) → "DA" (s_5) → "D" (s_6 & s_7). The timing slightly differs from that when evanescent sites are also taken into account (Figure 3).

sub-sequence at a time-point as an alternating concatenation of "A"s and "D"s, such as "ADADA". This way, the indel history of the sub-sequence could be broadly represented by a "time-series", such as, "A → ADA → ADADA → DADA → DA → D" (Figure 2).

Now, let us incorporate the evanescent sites again, and pay attention to their surroundings (Figure 3). First, an evanescent site, or a lump of contiguous evanescent sites, is regarded as belonging to a "D"-region, if any of the following is satisfied: (i) it was inserted within, or at the end of, a (lump of) descendant site(s) (*e.g.*, v_7 in Figure 3); (ii) it was inserted in conjunction with one or more descendant site(s) (*e.g.*, v_3 in Figure 3); or (iii) it spawned one or more descendant site(s) within, or at the end of, it (*e.g.*, v_1 in Figure 3). Second, the evanescent site(s) will not be regarded as either "A" or "D", if they were created after the deletion of all ancestral sites and before the first creation of descendant sites. (These sites can still be incorporated into our framework without any problem. See below.) Then, each of the remaining evanescent sites is regarded as belonging to an "A"-region, which should be uniquely determined at the time of the site's creation (the surrounding of *e.g.*, v_4

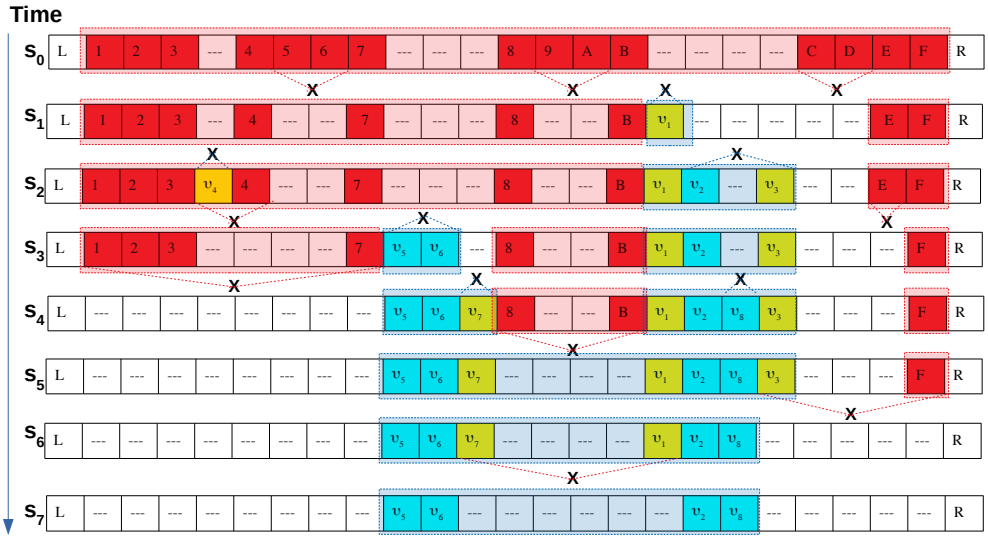

Figure 3: **Evolution of "A"/"D"-coloring pattern of region in between PASs ("L" and "R"), after taking account of evanescent sites again.** In each sequence state, the dashed-bordered transparent rectangles colored in red and blue represent the "A"-region and the "D"-region, respectively. The pattern evolution here is: "A "(s_0) $\rightarrow$ "ADA" (s_1 & s_2) $\rightarrow$ "ADADA" (s_3) $\rightarrow$ "DADA" (s_4) $\rightarrow$ "DA" (s_5) $\rightarrow$ "D" (s_6 & s_7). The timing slightly differs from that when evanescent sites are ignored (Figure 2). This figure was created by merely adding the "A"- and "D"-region indicators to Figure 1.

in Figure 3). An important corollary of these rules is that a (lump of) evanescent site(s) always belongs to a "D"-region if it was inserted at the boundary of an "A"-region and the "D"-region (*e.g.*, v_7 in Figure 3); this will play an important role below, when defining the "base" rate operator acting on each region.

Following the rules prescribed in the previous paragraph, the "A/D-coloring pattern evolution", or the "broad time-series", is nearly uniquely assigned to each indel history resulting in a case-(iv) gapped segment (Figure 3). This means that we may accurately compute the probability of each case-(iv) gapped segment as follows: (i) classify all the indel histories resulting in a given segment into the "broad time-series", (ii) compute the probability contributed by (all the histories belonging to) each "broad time-series", and (iii) sum all the probabilities computed as in (ii). In practice, however, the "A/D-coloring pattern" could be quite complex; for such complex cases, the probability computation might be too time-consuming to be practically feasible unless some smart time-saving algorithms are invented. Therefore, for the moment, it would be wise to deal with the problem via a

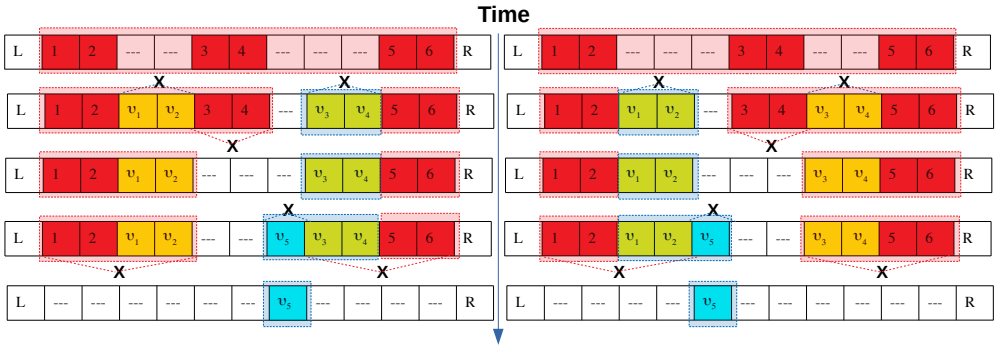

Figure 4: **Example indel history with "A/D"-coloring ambiguity.** The same notations as in Figure 3 are used here. The indel histories on the left and right are in fact identical to each other; the ancestral sites with indexes 3 and 4 and the descendant site with index v_5 cannot be uniquely ordered, because they are not homologous to each other, and because there are no sites determining their relative order. In this indel history, two lumps of evanescent sites, $[v_1, v_2]$ and $[v_3, v_4]$, were created. Depending on which of the lumps incorporates the descendant site (v_5), there are two ways of coloring the evanescent sites in "A" (transparent red) and "D" (transparent cyan), as shown here.

sort of "perturbation method",where the probabilities will be computed starting from the simplest patterns of "A/D-coloring pattern evolution" and gradually moving on to complex patterns. [15]

In the following couple of sections, we will concretely construct a new "perturbation-method" that embodies the computation strategy explained above.

[15]As a matter of fact, there could be an ambiguity on the "A/D"-coloring of evanescent sites, (probably only) in an exceptional type of indel histories, which satisfy the following conditions (see Figure 4 for an example): (1) two lumps of evanescent sites were independently created in an "A"-region; (2) then, the sites in between the two lumps were deleted; (3) then, some descendant sites were created exactly at the boundary between the two lumps; and (4) no other descendant sites were created within, or at the end of, the two lumps of evanescent sites. In such indel histories, either of the two lumps could be colored "A", and each such indel history could be counted twice in the perturbation method unfolded below. This means that some correction should be done in order to rectify such double-counting and obtain perfectly exact probabilities. Fortunately, each such indel history *inevitably* involves at least six insertions/deletions (*i.e.*, an insertion-deletion pair for each lump of evanescent sites, a deletion (of intervening "A"-sites), and an insertion (of the descendant sites)) that are in *extremely exquisite* spatial relationships to each other. Therefore, the total contributions from such indel histories are expected to be negligible in most practical analyses.

3 Basic Framework

We concretize the above computation strategy founded on the theoretical formulation we provided before [3, 1], which can handle stochastic sequence evolution models generalizing Miklós et al.'s "long-indel model" [2] (and also of the model used by a *genuine* sequence evolution simulator, Dawg [48]). For the problem at hand, i.e., regarding a case-(iv) gapped segment, it is sufficient to consider a sub-sequence (or a region) consisting of the two flanking preserved ancestral sites (PASs), labelled as "L" (for left) and "R" (for right) hereafter, and the sites in between them, labelled as $1, 2, ..., \Delta L(s)$, where the $\Delta L(s)$ is the number of sites in between the PASs and it varies while the sequence state, s, evolves. Aside from these labels, each of the sites, $x = L, 1, 2, ..., \Delta L(s), R$ is also assigned an "ancestry index", denoted $\upsilon(x)$ [3]. These indexes *as a function of* x also vary in the course of the evolution, because insertion(s)/deletion(s) could change the position (along the sequence) of the site with a particular ancestry. (Thus, if you want to make their time-dependence explicit, you could use the notation, $\upsilon(x, t)$.) [16] Since the sole role of these "ancestry index"es is to keep track of the time-trajectories of the sites involved, we could arbitrarily assign the indexes, $\{\ \upsilon(x)\ |\ x = L, 1, ..., \Delta L(s), R\ \}$, as long as the sites of different origins carry different indexes. Because the sites L and R are preserved throughout the evolution under consideration, the ancestry indexes at these sites are also fixed. Thus, we will henceforce assign $\upsilon(L) = L$ and $\upsilon(R) = R$. Then, the basic sequence state (concerning only insertions/deletions) is denoted as:

$$\langle s\ | = \ \langle\ [L, \upsilon(1), \upsilon(2), ..., \upsilon(\Delta L(s)), R]\ |\ . \tag{1}$$

(Or,

$$\langle s(t)\ | = \ \langle\ [L, \upsilon(1, t), \upsilon(2, t), ..., \upsilon(\Delta L(s), t), R]\ |$$

[16] It should be noted, however, that each site *must* keep its *unique* ancestry index during its evolutionary course, from the beginning to the end.

to explicitly represent its time-dependence.) Next, we define the time-frame. Let the time-interval in question begin at t_I and end at t_F. Thus, we will consider that the time t always be within the interval, $[t_I\ ,\ t_F]$.

As described in section R3 of a previous work of ours [3], the stochastic time-evolution of this sequence is controlled/prescribed by the (instantaneous) indel rate operator, $\hat{Q}^{ID}(t)$, which can be decomposed as:

$$\hat{Q}^{ID}(t) = \hat{Q}^{I}(t) + \hat{Q}^{D}(t)\ , \tag{2}$$

$$\hat{Q}^{m}(t) = \hat{Q}^{m}_{M}(t) + \hat{Q}^{m}_{X}(t) \quad (m = I, D)\ . \tag{3}$$

And the actions of the components on the sequence state, Eq. 1, are:

$$\langle s \mid \hat{Q}^{I}_{M}(t) = \sum_{x=-1}^{\Delta L(s)+1} \sum_{l=1}^{\infty} r_I(x, l; s, t) \langle s \mid \hat{M}_I(x, l)\ , \tag{4}$$

$$\langle s \mid \hat{Q}^{D}_{M}(t) = \sum_{x_B=-\infty}^{\Delta L(s)+1} \sum_{x_E=\max\{0, x_B\}}^{+\infty} r_D(x_B, x_E; s, t) \langle s \mid \hat{M}_D(x_B, x_E)\ , \tag{5}$$

$$\langle s \mid \hat{Q}^{m}_{X}(t) = -R^{m}_{X}(s, t) \langle s \mid \quad (m = I, D)\ . \tag{6}$$

Here, as in [3], we tacitly assumed that the sub-sequence in question is embedded in an infinitely long sequence. And, for computational convenience, we replaced $x = R$ and $x = L$ with $x = 0$ and $x = \Delta L(s) + 1$, respectively, for the site numbers of the flanking PASs. [17] The above equations use the same notations as in a previous work of ours [3]. Especially, $\hat{M}_I(x, l)$ is the operator that inserts a string of length l between the sites x and $x + 1$, and $r_I(x, l; s, t)$ is the instantaneous rate that such an insertion occurs on the sequence state $\langle s \mid$ at time t; $\hat{M}_D(x_B, x_E)$ is the operator that deletes the sites, $x = x_B$ through $x = x_E$ (both inclusive), and $r_D(x_B, x_E; s, t)$ is the instantaneous rate that such a deletion occurs on $\langle s \mid$ at time t. The $R^{I}_{X}(s, t)$ and $R^{D}_{X}(s, t)$ are the insertion and deletion components, respectively,

[17] The difference from the equations in [3] is that the site number, x, starts with $x = 0$ (and ends with $x = \Delta L(s) + 1$ $(= L(s) - 1$, where $L(s) \stackrel{\text{def}}{=} \Delta L(s) + 2$ is the number of sites in the sequence state, $s)$) here, whereas it starts with $x = 1$ (and ends with $x = L(s)$) in [3].

of the exit rate, $R_X^{ID}(s,t)$, of the state $\langle s \mid$ at time t via insertions/deletions. Specifically,

$$R_X^{ID}(s,t) = R_X^I(s,t) + R_X^D(s,t) \ , \tag{7}$$

$$R_X^I(s,t) = \sum_{x=-1}^{\Delta L(s)+1} \sum_{l=1}^{\infty} r_I(x,l;s,t) \ , \tag{8}$$

$$R_X^D(s,t) = \sum_{x_B=-\infty}^{\Delta L(s)+1} \sum_{x_E=\max\{0,x_B\}}^{+\infty} r_D(x_B,x_E;s,t) \ , \tag{9}$$

These equations, Eq.1 through Eq.9, provide the foundation for our strategy to compute the probability of each case-(iv) gapped segment. Using these equations, we could at least theoretically calculate the transition probability from a sequence state to another in between the PASs, L and R, just as described in a previous work of ours [3]. In general, however, such state transitions are not limited to those providing case-(iv) gap-patterns. To go further, we will resort to a sort of "perturbation theory", as unfolded in the next section.

4 New Perturbation Method [18] to Systematically Compute Probabilities of Case-(iv) Gap-Patterns up to Desired Level of Accuracy

4.1 Broad Account of the Procedures

As explained in Section 2, the time evolution of a sub-sequence yielding a case-(iv) gapped segment follows a particular time-series of "A(ncestral)/D(escendant)-coloring", such as "A

[18] The "perturbation method" unfolded here is somewhat similar to, yet significantly different from, the perturbation-theoretical approach in our previous studies [3, 1]; especially, the "base"-"perturbation" decomposition of the rate operator, $\hat{Q}^{ID}$, differs between the two approaches. Here, the probabilities computed via the latter approach (so that the evolutionary histories include as much indels as desired) are regarded as "exact"-solutions (under the "base" rate operator here), and thus the term, "perturbation", always refers to that in the method constructed here.

$\rightarrow$ ADA $\rightarrow$ ADADA $\rightarrow$ DADA $\rightarrow$ DA $\rightarrow$ D" (Figure 3); especially, it *must* always start with "A" (ancestral only) and end with "D" (descendant only). Now, we slice the broad evolutionary history at the times, t_k with $k = 1, 2, ..., N_{Sl} - 1$, when the "A/D"-coloring patterns of the sequence change (Figure 5). (The N_{Sl} above is the number of slices in the broad history.) Then, each slice, *e.g.*, in an open time-interval, (t_k, t_{k+1}), shows a fixed "A/D"-coloring pattern, *e.g.*, "DADA", in which "A" and "D" alternate. This suggests the strategy for computing the total probability that each particular time-series of "A/D"-coloring occurs: (i) within each slice, decompose the rate operator $\hat{Q}^{ID}(t)$ into the "base"- and "perturbation"-parts; (ii) within each slice, compute the probabilities of state transitions through each slice *under (or conditioned on) its fixed "A/D"-coloring pattern*, using the "base"-part of $\hat{Q}^{ID}(t)$; (iii) between two consecutive slices, compute the probabilities of the "coloring-pattern-transition" from each slice to the next, by multiplying the rates of the particular insertions/deletions that changes the "A/D"-coloring pattern(, which probably belong to the "perturbation" part); (iv) stack the probabilities of the slices (via (ii)) and the "coloring-pattern-transitions" (via (iii)), from the beginning to the end, and integrate over the times, $(t_I <) t_1 < t_2 < ... < t_{N_{Sl}-1} (< t_F)$, to get the probability contributed from the particular pattern of "A/D"-coloring evolution to the case-(iv) gapped segment.

4.2 "Base"-"Perturbation" Decomposition of Rate Operator

Within each slice of a fixed "A/D"-coloring pattern, we can formally decompose the rate operator $\hat{Q}^{ID}(t)$ into the "base"-and "perturbation"-parts. Hereafter in this subsection, we will consider a slice in the open time-interval, (t_k, t_{k+1}), and refer to it as "slice k". The k can range from 0 through $N_{Sl} - 1$, where $t_0 \stackrel{\text{def}}{=} t_I$ and $t_{N_{Sl}} \stackrel{\text{def}}{=} t_F$ are the initial and final times, respectively, of the whole time-interval considered here. Let $N_C(k)$ be the number of colored regions in slice k, and let C_i (with $i = 1, ..., N_C(k)$) be the i-th colored region, numbered from left to right. For example, in the above case of "DADA" (for slice k $(= 4)$),

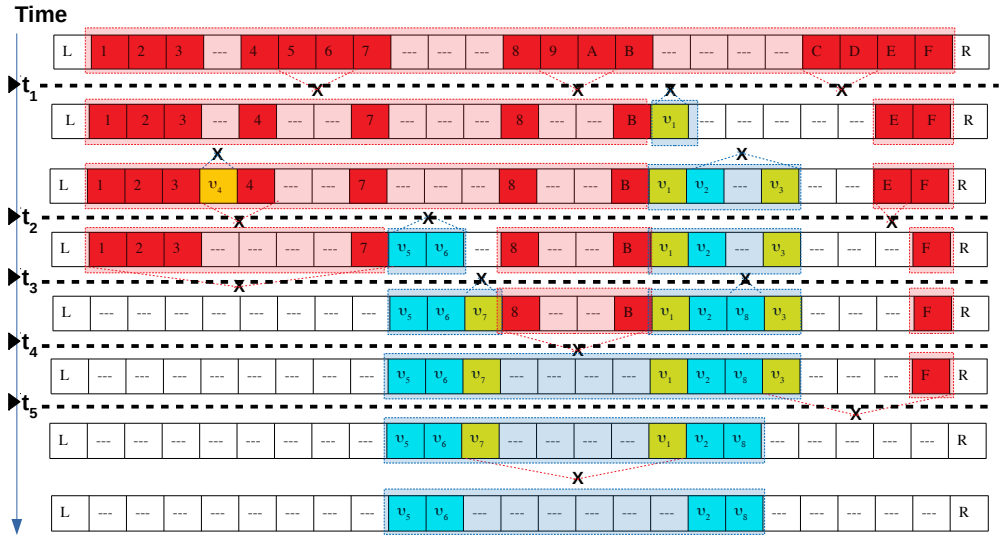

Figure 5: **Slicing "A"/"D"-coloring pattern evolution.** This figure was created from Figure 1 by adding the times, $t_1, ..., t_5$, at which the "A/D"-coloring-pattern changes, and by removing the state labels $(s_0, ..., s_7)$ for clarity. To clarify the slices, each of the time is accompanied by a triangle on the time-axis and a thick dashed horizontal line.

$N_C(k) = 4$, C_1 and C_3 represent the 1st and 2nd "D"-colored regions, respectively, and C_2 and C_4 represent the 1st and 2nd "A"-colored regions, respectively. Hereafter, the "D"- and "A"-colored regions will be abbreviated as the "A"- and "D"-regions, respectively. And let $x_B(i)$ and $x_E(i)$ be the site numbers at the beginning and end, respectively, of the region C_i. Especially, $x_B(i+1) = x_E(i) + 1$, $x_B(1) = 1$ and $x_E(N_C(k)) = \Delta L(t)$ always hold. Then, we *define* the "base" rate operator, $\hat{Q}_0^{ID}(i;t)$, for the region C_i, as follows:

$$\hat{Q}_0^{ID}(i;t) = \hat{Q}_{0\ M}^{I}(i;t) + \hat{Q}_{0\ M}^{D}(i;t) + \hat{Q}_{0\ X}^{ID}(i;t), \tag{10}$$

$$\langle s \mid \hat{Q}_{0\ M}^{I}(i;t) = \sum_{x=x_B(i)-\sigma_B(i)}^{x_E(i)-1+\sigma_E(i)} \sum_{l=1}^{\infty} r_I(x,l;s,t)\langle s \mid \hat{M}_I(x,l) \ , \tag{11}$$

$$\langle s \mid \hat{Q}_{0\ M}^{D}(i;t) = \sum_{\substack{x_B(i)\le x_B\le x_E\le x_E(i)\ , \\ (x_B,x_E)\neq(x_B(i),x_E(i))}} \sum_{x_B} \sum_{x_E} r_D(x_B,x_E;s,t)\langle s \mid \hat{M}_D(x_B,x_E) \ , \tag{12}$$

$$\langle s \mid \hat{Q}_{0\ X}^{ID}(i;t) = -\Delta R_X^{ID}(s,C_i,t)\langle s \mid \ . \tag{13}$$

First note that no deletions in the deletion component, Eq.12, stick out of the region C_i; thus, this "base" rate operator focuses on the deletions occurring totally within the region. Also note that Eq.12 does not include the deletion of an entire C_i (i.e., $\hat{M}_D(x_B(i), x_E(i))$; such a deletion will be included in the "perturbation" part (defined below). In the insertion

component, Eq.11, each of $\sigma_B(i)$ and $\sigma_E(i)$ takes only the values 0 and 1. These values are determined to realize the affiliation rules for the "evanescent" sites prescribed in Section 2. Specifically, $\sigma_B(i) = \sigma_E(i) = 1$ if C_i is a "D"-region; if C_i is an "A"-region, $\sigma_B(i) = 1$ still holds if $i = 1$, and $\sigma_E(i) = 1$ still holds if $i = N_C(t)$; in the remaining cases, each of them equals 0. In the exit-rate component, Eq.13, $\Delta R_X^{ID}(s, C_i, t) \stackrel{\text{def}}{=} \delta R_X^{ID}(s, \{s \setminus C_i\}, t)$ ($\stackrel{\text{def}}{=} R_X^{ID}(s,t) - R_X^{ID}(\{s \setminus C_i\}, t)$) is the increment of the exit rate caused by the existence of the region C_i in the sequence state s (at t). [19]

In fact, the insertion and deletion components of the "base" rate operator, Eqs.11 & 12, are defined so that $\hat{Q}_0^{ID}(i;t)$'s with different i's do not interfere with each other, *provided that* Condition (i) in section R6 of a previous study of ours [3] is satisfied between different C_i's. Besides, the exit rates, $\Delta R_X^{ID}(s, C_i, t)$'s, with different i's are also independent of each other, provided that the Condition (ii) for the factorability of the PWA probabilities (in section R6 of [3]) is satisfied. If these conditions hold, we actually have:

$$R_X^{ID}(s,t) = R_X^{ID}([L,R],t) + \sum_{i=1}^{N_C(t)} \Delta R_X^{ID}(s, C_i, t). \tag{14}$$

Because $R_X^{ID}([L,R],t)$ is like a "constant" independent of the specific state of the subsequence (as long as it has the PASs, L and R), it would be convenient to define an operator, $\hat{Q}_{0\ X}^{ID}(0;t)$, such that

$$\langle s \mid \hat{Q}_{0\ X}^{ID}(0;t) \stackrel{\text{def}}{=} -R_X^{ID}([L,R],t) \langle s \mid \ . \tag{15}$$

Then, for a positive $N_C(k)$, we *define* the "base" total rate operator as follows:

$$\hat{Q}_0^{ID}(t) \stackrel{\text{def}}{=} \hat{Q}_{0\ X}^{ID}(0;t) + \sum_{i=1}^{N_C(k)} \hat{Q}_0^{ID}(i;t). \tag{16}$$

From the above arguments, we see that the $\hat{Q}_0^{ID}(i;t)$'s with different i's do not interfere with one another. Thus, under the aforementioned conditions, they could be time-integrated independently from one another, to get the "base" finite-time transition operator, $\hat{P}_0^{ID}(t,t')$,

[19] Here, the $\{s \setminus C_i\}$ represents the sequence state formed by removing C_i from s.

as follows:

$$\hat{P}_0^{ID}(t,t') \stackrel{\text{def}}{=} T\left\{\exp\left(\int_t^{t'} d\tau\ \hat{Q}_0^{ID}(\tau)\right)\right\} \tag{17}$$

$$= \hat{P}_0^{ID}(0;t,t') \times \prod_{i=1}^{N_C(k)} \hat{P}_0^{ID}(i;t,t')\ , \quad \text{with}$$

$$\hat{P}_0^{ID}(0;t,t') = \exp\left(-\int_t^{t'} d\tau\ R_X^{ID}([L,R],\tau)\right) \times \hat{\mathbf{1}}\ , \tag{18}$$

$$\hat{P}_0^{ID}(i;t,t') \stackrel{\text{def}}{=} T\left\{\exp\left(\int_t^{t'} d\tau\ \hat{Q}_0^{ID}(i;\tau)\right)\right\} \quad (\text{with}\ \ i=1,...,N_C(k))\ . \tag{19}$$

Here, $T\{...\}$ represents the time-ordered product of operators in which an operator at time τ comes on the right of other operators acting earlier than τ and on the left of others acting later than τ. And the "$\hat{\mathbf{1}}$" above is the identity operator.

The "perturbation" part, denoted here as $\Delta\hat{Q}^{ID}(t)$, is simply defined as:

$$\Delta\hat{Q}^{ID}(t) \stackrel{\text{def}}{=} \hat{Q}^{ID}(t) - \hat{Q}_0^{ID}(t)\ . \tag{20}$$

Because the "base" operator, $\hat{Q}_0^{ID}(t)$ in Eq.16, includes the entire exit-rate component, $\hat{Q}_X^{ID}(t)(\stackrel{\text{def}}{=} \hat{Q}_X^I(t) + \hat{Q}_X^D(t))$, as well as the entire insertion-mutation component, $\hat{Q}_M^I(t)$, $\Delta\hat{Q}^{ID}(t)$ is purely a linear combination of the deletion operators representing deletions extending into two or more regions, as well as deletions of the whole regions of single "A"s and single "D"s. As it is, the explicit expression of $\Delta\hat{Q}^{ID}(t)$ is considerably complex.

Fortunately, the fact that we are now dealing only with a single case-(iv) gapped segment, as well as the natures of "A" and "D" regions, will considerably simplify the expression of the *"effective part"*, $\Delta\hat{Q}_{Eff}^{ID}(t)$, of the perturbation part, which collects all those deletions which could actually occur in the indel histories we want, as follows. First, we can ignore all deletions that affects L and/or R, because, *by definition*, the PASs, L and R, have never experienced deletions. Second, we can also ignore all deletions each of which deletes a whole "D"-region, because the "D"-regions, also *by definition*, must NOT be deleted entirely. (If so, they are just lumps of evanescent sites.) This also means that a deletion can always

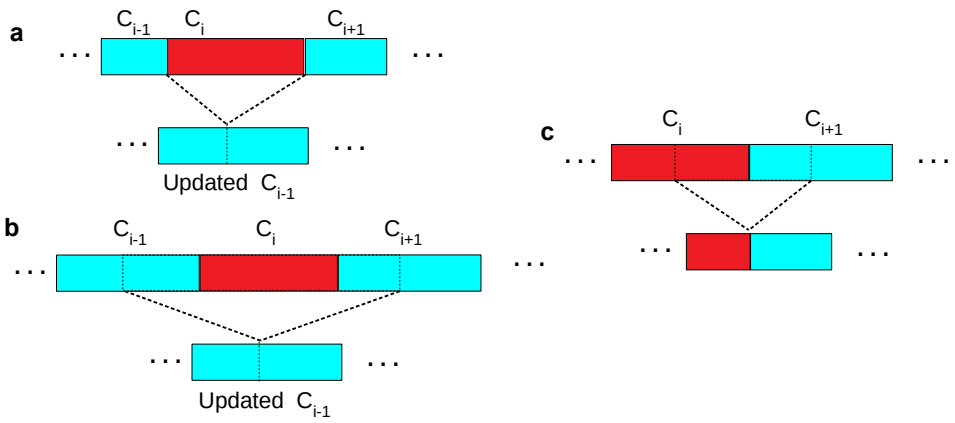

Figure 6: **Deletions included in the "effective" perturbation part ($\Delta\hat{Q}^{ID}_{Eff}(t)$) of the rate operator. a.** The deletion of an entire "A"-region *alone*. **b.** The simultaneous deletion of an entire "A"-region and parts of the flanking "D"-regions. **c.** A "boundary-eroding" deletion. In each panel, the red and cyan rectangles represent an "A"-region and a "D"-region, respectively.

be ignored if it deletes two or more contiguous regions, as "A" and "D" always alternate. Taking account of these two restrictions, we see that the constituent deletion operators of the "effective part", $\Delta\hat{Q}^{ID}_{Eff}(t)$, can be classified into two categories: (i) "A-deleting" deletions, each of which deletes an entire "A"-region (Figure 6 a), and may simultaneously delete (a) part(s), but never the whole, of the flanking "D"-region(s) (Figure 6 b) and (ii) "boundary-eroding" deletions, each of which straddles two contiguous regions, but does NOT delete any whole regions (Figure 6 c). These results can be expressed as follows:

$$\Delta\hat{Q}^{ID}_{Eff}(t) \;=\; \hat{Q}^{D}_{M:\text{A-del}}(t) \;+\; \hat{Q}^{D}_{M:\text{B-er}}(t)\;, \tag{21}$$

$$\hat{Q}^{D}_{M:\text{A-del}}(t) \;\stackrel{\text{def}}{=}\; \sum_{i=1}^{N_C(k)} \delta(\text{color}(i), A)\; \hat{Q}^{D}_{M:\text{A-del}}(i;t)\;, \tag{22}$$

$$\langle s \mid \hat{Q}^{D}_{M:\text{A-del}}(i;t) \;\stackrel{\text{def}}{=}\; \sum_{x_B=x_B(i-1)+1}^{x_B(i)} \sum_{x_E=x_E(i)}^{x_E(i+1)-1} r_D(x_B,x_E;s,t)\langle s \mid \hat{M}_D(x_B,x_E)\;, \tag{23}$$

$$\hat{Q}^{D}_{M:\text{B-er}}(t) \;=\; \sum_{i=1}^{N_C(k)-1} \hat{Q}^{D}_{M:\text{B-er}}(i;t)\;, \tag{24}$$

$$\langle s \mid \hat{Q}^{D}_{M:\text{B-er}}(i;t) \;\stackrel{\text{def}}{=}\; \sum_{x_B=x_B(i)+1}^{x_E(i)} \sum_{x_E=x_B(i+1)}^{x_E(i+1)-1} r_D(x_B,x_E;s,t)\langle s \mid \hat{M}_D(x_B,x_E)\;. \tag{25}$$

Here, the subscripts, "A-del" and "B-er", are short for "A-deleting" and "boundary-eroding", respectively. In Eq.22, the $\delta(\text{color}(i), A)$ is an analog of Cronecker's delta, and $\delta(\text{color}(i), A) = 1$ if C_i is an "A"-region, and $= 0$ otherwise. In Eq.23, we define $x_B(0) = 0$ and $x_E(N_C(k) + 1) = \Delta L(s) + 1$, to cover the cases where $\text{color}(C_1) = A$ and $\text{color}(C_{N_C(k)}) = A$, respectively. As indicated by their definitions, Eqs.23 and 25, the $\hat{Q}^D_{M:\text{A-del}}(i;t)$ is the collection of all deletions deleting the entire "A"-region, C_i, and the $\hat{Q}^D_{M:\text{B-er}}(i;t)$ is the collection of all "boundary-eroding" deletions affecting C_i and C_{i+1}.

For completeness, let us finally consider slices with $N_C(k) = 0$, which could occur if all ancestral sites are deleted before any "D"-regions are created. Such a slice consists only of evanescent sites, and it begins and ends with the sequence state, $[L, R]$. Therefore, the total probabilities for such a slice and a series of such slices (with various time-intervals) are exactly equal to those of case-(i) gapped segments, which can be computed using the whole rate-operator, $\hat{Q}^{ID}(t)$ in Eq.2, exactly as described in a previous study of ours [1] (, or, via a faster method identical in essence to that in subsection 4.3).

4.3 Computing Transition Probabilities within Each Slice

If we want to compute a transition probability within each slice, say, slice k in (t_k, t_{k+1}), we first need to specify the "initial" and "final" lengths of the regions, C_i with $i = 1, ..., N_C(k)$, at times $t_k^+ \overset{\text{def}}{=} t_k + \epsilon$ and $t_{k+1}^- \overset{\text{def}}{=} t_{k+1} - \epsilon$, respectively. (Here, ϵ is an infinitesimal value.) This is equivalent to specifying the "initial" and "final" values of $x_E(i)$, with $i = 1, ..., N_C(k)$; let $x_E(i;F)$ and $x_E(i;F)$ (both with $i = 1, ..., N_C(k)$) be such "initial" and "final" values, respectively. Similarly, let $x_B(i;I)$ and $x_B(i;F)$ be the "initial" and "final" values, respectively, of $x_B(i)$ (with $i = 1, ..., N_C(k)$). More generally, let $x_B(i;t)$ and $x_E(i;t)$ be the beginning and end coordinates, respectively, of the region C_i at time t in (t_k, t_{k+1}). (See Figure 7 for a schematic illustration of the situation considered here.)

As in the previous subsection, we consider that Conditions (i) and (ii) in section R6 of

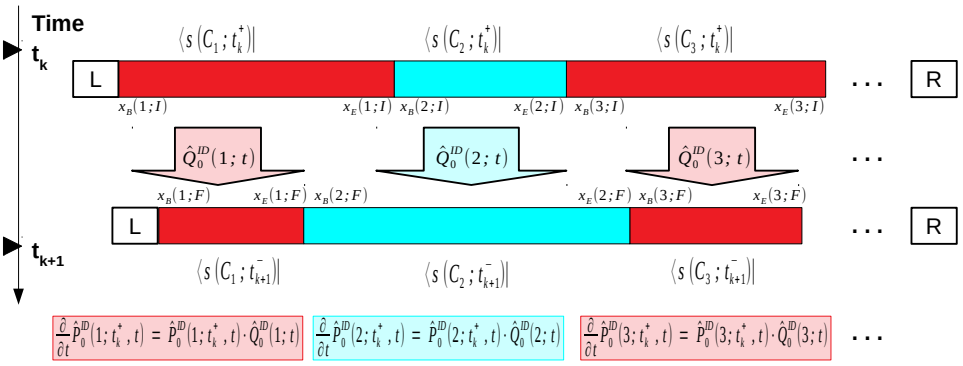

Figure 7: **Sequence evolution (via indels) under "base" rate operator, $\hat{Q}_0^{ID}(t)$, within slice k (in (t_k, t_{k+1})).** As before, the red and cyan rectangles represent an "A"-region and a "D"-region, respectively. Each thick, light-colored downward arrow suggests that the stochastic evolution of the corresponding region (C_i) is dictated by the relevant "base" rate operator ($\hat{Q}_0^{ID}(i;t)$).

a previous study of ours [3] are satisfied between the "base" rate operators, $\hat{Q}_0^{ID}(i;t)$'s, on different C_i's. Then, *in the current perturbation framework*, the "zero-th approximation" of the transition probabilities within slice k can be obtained by sandwiching the factorized form of the "base" transition operator, Eq.(17), with the initial and final sequence states. As far as this "zero-th approximation" is concerned, we could express the sequence state, $\langle s \mid$ in Eq.1, as a "tensor product":

$$\langle s(t) \mid \quad = \quad \langle s_0 \mid \otimes \left[\bigotimes_{i=1}^{N_C(k)} \langle s(C_i;t) \mid \right] \quad , \quad \text{with} \tag{26}$$

$$\langle s_0 \mid \quad \stackrel{\text{def}}{=} \quad \langle [L, R] \mid \, , \tag{27}$$

$$\langle s(C_i;t) \mid \quad \stackrel{\text{def}}{=} \quad \langle [\upsilon(x_B(i;t),t), ..., \upsilon(x_E(i;t),t)] \mid \quad . \tag{28}$$

Here, $\upsilon(x,t)$ is the ancestry index of the x th site at time t. Hence, using this "tensor product" expression, the "zero-th approximation" of the transition probability can be calculated formally as:

$$P_0 \left[\left(s(t_{k+1}^-), t_{k+1}^- \right) \mid \left(s(t_k^+), t_k^+ \right) \right] \stackrel{\text{def}}{=} \langle s(t_k^+) \mid \hat{P}_0^{ID}(t_k^+, t_{k+1}^-) \mid s(t_{k+1}^-) \rangle$$

$$= \exp\left(-\int_{t_k^+}^{t_{k+1}^-} d\tau \; R_X^{ID}([L,R],\tau)\right) \times \prod_{i=1}^{N_C(k)} \langle s(C_i;t_k^+) \mid \hat{P}_0^{ID}(i;t_k^+,t_{k+1}^-) \mid s(C_i;t_{k+1}^-)\rangle . \quad (29)$$

To attain the second equation, we used the fact, $\langle s_0 \mid \hat{P}_0^{ID}(0;t,t') \mid s_0\rangle = \exp\left(-\int_t^{t'} d\tau \; R_X^{ID}([L,R],\tau)\right)$. Therefore, the problem of computing the "zero-th approximation" of the transition probabilities within each slice can be reduced to that of computing the multiplication factors, or transition probabilities within each colored region, denoted here as:

$$\mu_{P_0}\left[(s(C_i;t_{k+1}^-),t_{k+1}^-) \mid (s(C_i;t_k^+),t_k^+)\right] \stackrel{\text{def}}{=} \langle s(C_i;t_k^+) \mid \hat{P}_0^{ID}(i;t_k^+,t_{k+1}^-) \mid s(C_i;t_{k+1}^-)\rangle \; , \quad (30)$$

for each C_i (with $i = 1, ..., N_C(k)$). [20] (To remember the situation under consideration, see Figure 7.)

For a locally space-homogeneous model, whose insertion/deletion rates are homogeneous within each of most gapped segments, we derived a system of iterative equations to give effectively exact multiplication factors for isolated gaps. (See section SM-3 in additional file 1 of a previous work of ours [1].) It should be noted that, within each C_i and under the "base" rate operator, *we can forget about the origins (or ancestries) of the sites in the region*, because all of them will eventually be deleted anyway if C_i is an "A"-region, and because none of them belonged to the ancestral (or initial) sequence state, anyway, if C_i is a "D"-region. It follows that, under a locally space-homogenous model, the state, $\langle s(C_i;t) \mid$, depends only on the number of sites in C_i, *i.e.*, $\Delta L(i;t) \stackrel{\text{def}}{=} x_E(i;t) - x_B(i;t) + 1$; we can thus express this fact explicitly as: $\langle s(C_i;t) \mid = \langle \Delta L(i;t) \mid$, just as in the derivation for an isolated gap. Moreover, $\Delta R_X^{ID}(s,C_i,t)$ also depends only on $\Delta L(i;t)$ (and possibly t); let us remember this fact by using an "alias", $\Delta R_X^{ID}(\Delta L(i;t),t)$, of $\Delta R_X^{ID}(s,C_i,t)$. Therefore, following almost exactly the same procedure (as in SM-3 of a previous study of ours [1]), we

[20] In the current "perturbation" method, the "zero-th" approximation does *not* mean that there are no insertions/deletions during the time-interval, (t_k^+, t_{k+1}^-); actually, there could be any numbers of insertions/deletions, as long as they occur totally within each of the regions (C_i's) and as long as they do *not* delete any of the entire C_i's.

could derive a system of iterative equations also for effectively exact multiplication factors for each C_i in slice (t_k, t_{k+1}), under a locally space-homogenous model.

The only differences from the previous method [1] are: (i) the exit rate used here is the increment, $\Delta R_X^{ID}(s, C_i, t)$ (see the description below Eq.13), instead of the whole rate, $R_X^{ID}(s,t)$ (in Eq.7); (ii) both initial and final states have nonzero sites in between L and R, whereas, for an isolated gap, either of them has no sites in between; and (iii) the "effective" rate operator differs slightly, specifically, the one used here lacks the deletion of the entire C_i, and it may also lack insertions at both or either end(s) if C_i is an "A"-region. (See appendix A for the specific computation along this line.)

Here, however, we provide a *much faster* approach; it is based on a *direct* approximation of the following definition:

$$\begin{aligned} \hat{P}_0^{ID}(i; t_F, t_I) &\stackrel{\text{def}}{=} T\left\{\exp\left[\int_{t_I}^{t_F} d\tau\ \hat{Q}_0^{ID}(i;\tau)\right]\right\} \\ &\stackrel{\text{def}}{=} \lim_{N_P\to\infty}\left(\hat{\mathbf{1}} + \Delta_{N_P}t\cdot\hat{Q}_0^{ID}(i;\bar{t}_1)\right)\cdots\left(\hat{\mathbf{1}} + \Delta_{N_P}t\cdot\hat{Q}_0^{ID}(i;\bar{t}_{N_P})\right)\ , \quad (31) \end{aligned}$$

where $\bar{t}_j = t_I + (j-1/2)\Delta_{N_P}t \quad (j = 1, 2, ..., N_P)$, with $\Delta_{N_P}t \stackrel{\text{def}}{=} (t_F - t_I)/N_P$. This should be realized with an algorithm of space-complexity $O(N_P L^{CO})$ and time-complexity $O(N_P\{L^{CO}\}^2)$.

Now, for a specific region C_i in a given slice in (t_k, t_{k+1}) (or in $[t_k^+, t_{k+1}^-]$), let us compute the transition probabilities (or the multiplication factors), $\mu_{P_0}\left[(\Delta L(i;F), t_{k+1}^-) \mid (\Delta L(i;I), t_k^+)\right](\sigma_B E(i))$, with ranging $\Delta L(i;I)$, $\Delta L(i;F)$, and $[t_k^+, t_{k+1}^-] \in [t_I, t_F]$, by *directly* approximating the definition, Eq.31, of the finite-time transition operator. The first thing we need to do is discretize the time-interval, $[t_I, t_F]$, with a number of (usually equal-spaced) points in between t_I and t_F; Here, we choose, $t_j = t_I + j\cdot\Delta_{N_P}t$ $(j = 1, 2, ..., N_P - 1)$, with $\Delta_{N_P}t \stackrel{\text{def}}{=} (t_F - t_I)/N_P$. And we *set* $t_0 \stackrel{\text{def}}{=} t_I$ and $t_{N_P} \stackrel{\text{def}}{=} t_F$. [21] Then, we define a series of "discretized" finite-time transition operators:

$$\hat{P}_0^{ID\ [N_P]}(i; t_I, t_j) \stackrel{\text{def}}{=} \left(\hat{\mathbf{1}} + \Delta_{N_P}t\cdot\hat{Q}_0^{ID}(i;\bar{t}_1)\right)\cdots\left(\hat{\mathbf{1}} + \Delta_{N_P}t\cdot\hat{Q}_0^{ID}(i;\bar{t}_j)\right) \quad (32)$$

[21] The t_j's defined here are related to the $\bar{t}_j$'s defined above via the relations, $\bar{t}_j = (t_{j-1} + t_j)/2$.

for $j = 1, 2, ..., N_P$, and $\hat{P}_0^{ID\ [N_P]}(i; t_I, t_0) = \hat{\mathbf{1}}$. To express this definition more precisely, we can use the recursion relation:

$$\hat{P}_0^{ID\ [N_P]}(i; t_I, t_j) = \hat{P}_0^{ID\ [N_P]}(i; t_I, t_{j-1}) \cdot \Big(\hat{\mathbf{1}} + \Delta_{N_P} t \cdot \hat{Q}_0^{ID}(i; \bar{t}_j) \Big) \tag{33}$$

for $j = 1, 2, ..., N_P$, and, again, $\hat{P}_0^{ID[N_P]}(i; t_I, t_0) = \hat{\mathbf{1}}$.

Next, we define the multiplication factors:

$$\mu_{P_0}^{[N_P]} \left[(s(C_i; t_j), t_j) \mid (s(C_i; t_I), t_I) \right] \stackrel{\text{def}}{=} \langle s(C_i; t_I) \mid \hat{P}_0^{ID\ [N_P]}(i; t_I, t_j) \mid s(C_i; t_j) \rangle \ , \tag{34}$$

for $j = 0, 1, ..., N_P$. Then, by substituting Eq.33 (for $j = 1, ..., N_P$) into the right-hand side of the above equation, and using $\sum_{s(C_i, t_{j-1}) \in S(C_i)} \mid s(C_i; t_{j-1}) \rangle \langle s(C_i; t_{j-1}) \mid = \hat{\mathbf{1}}$(, where $S(C_i)$ denotes the space of all possible states in the region C_i), we get:

$$\begin{aligned} & \mu_{P_0}^{[N_P]} \left[(s(C_i; t_j), t_j) \mid (s(C_i; t_I), t_I) \right] \\ = & \sum_{s' \in S(C_i)} \Bigg[\mu_{P_0}^{[N_P]} \left[(s', t_{j-1}) \mid (s(C_i; t_I), t_I) \right] \times \\ & \times \Big(\delta(s', s(C_i; t_j)) + \Delta_{N_P} t \cdot \langle s' \mid \hat{Q}_0^{ID}(i; \bar{t}_j) \mid s(C_i; t_j) \rangle \Big) \Bigg] \ . \end{aligned} \tag{35}$$

Here, we set $s' = s(C_i; t_{j-1})$ for simplicity. This gives the recursion relations for the "directly approximated" multiplication factors at time-points, $t_1, t_2, ..., t_{N_P} (= t_F)$, and the relations can be computed by iteration. As argued above, in the locally space-homogeneous model we deal with, the state *within each region* (C_i) is determined by the number of sites, $\Delta L(i; t)$, in between the PASs. Thus, under the same setting as for Eq.56, the recursion relations, Eq.35, become:

$$\begin{aligned} & \mu_{P_0}^{[N_P]} \left[(\Delta L_j, t_j) \mid (\Delta L_I, t_I) \right] (\sigma_{BE}(i)) \\ = & \Big(1 - \Delta_{N_P} t \cdot \Delta R_X^{ID}(\Delta L_j, \bar{t}_j) \Big) \cdot \mu_{P_0}^{[N_P]} \left[(\Delta L_j, t_{j-1}) \mid (\Delta L_I, t_I) \right] (\sigma_{BE}(i)) \\ & + \Delta_{N_P} t \cdot \Bigg[\sum_{l=1}^{\min(L_I^{CO},\ \Delta L_j - 1)} \Big\{ (\Delta L_j - l - 1 + \sigma_{BE}(i)) \times g_I(l, \bar{t}_j) \times \\ & \qquad \times \mu_{P_0}^{[N_P]} \left[(\Delta L_j - l, t_{j-1}) \mid (\Delta L_I, t_I) \right] (\sigma_{BE}(i)) \Big\} \end{aligned}$$

$$+(\Delta L_j+1)\sum_{l=1}^{\min(L_D^{CO},\ L^{CO}-\Delta L_j)}\Big\{\ g_D(l,\bar{t}_j)\times$$
$$\times\mu_{P_0}^{[N_P]}\left[(\Delta L_j+l,t_{j-1})\mid(\Delta L_I,t_I)\right](\sigma_{BE}(i))\ \Big\}\Big]\ , \tag{36}$$

with the initial condition:

$$\mu_{P_0}^{[N_P]}\left[(\Delta L_0,t_0)\mid(\Delta L_I,t_I)\right](\sigma_{BE}(i))=\delta(\Delta L_0,\Delta L_I)\ .$$

(It should be noted that $t_I = t_0$.) Here, for clarity, we used the short-hand notations, $\Delta L_I = \Delta L(i;I)$ and $\Delta L_j = \Delta L(i;j)$. Here, we also used the fact that the dependence of the factors on the region C_i is only through $\sigma_{BE}(i)$ ($\stackrel{\text{def}}{=} \sigma_B(i)+\sigma_E(i)$). [22] When t_I is fixed, the recursion relations, Eq.36, with ranging ΔL_j, ΔL_I, and t_j has the space- and time-complexities of $O(N_P\{L^{CO}\}^2)$ and $O(N_P\{L^{CO}\}^3)$, respectively. If the model is time-homogeneous as well, this will be sufficient, because $\mu_{P_0}^{[N_P]}[\cdots]$'s depend on t_I and t_j only through $t_j - t_I$. If the model is not time-homogeneous, however, we also need to compute the Eq.36 with ranging t_I. If such computations are performed *serially*, the total time-complexity becomes $O(\{N_P\}^2\{L^{CO}\}^3)$.

Alternatively, we could extend the time-interval backward, from $[t_F, t_F]$. See appendix B if you are interested in the specific expression of the resulting recursion relation.

To further raise the level of approximation, we need to switch on $\hat{Q}^D_{M:\text{B-er}}(t)$ in Eq.21, because this term also preserves the "A/D"-coloring pattern of the sub-sequence.

Although we could, *at least theoretically*, directly approximate the definition of the finite-time transition operator:

$$\hat{P}^{ID}_{0+\text{B-er}}(t,t')\stackrel{\text{def}}{=}T\left\{\exp\left(\int_t^{t'}d\tau\ \left[\hat{Q}^{ID}_0(\tau)+\hat{Q}^D_{M:\text{B-er}}(\tau)\right]\right)\right\}\ , \tag{37}$$

such computations will get practically impossible as the number of colored regions increases

[22] The above recursion relation is valid for $\Delta L_j = 1,...,L^{CO}$, and $\Delta L_I = 1,...,L^{CO}$; $\Delta L_j = 0$ needs to be excluded because $\hat{Q}^{ID}_0(i;t)$ does NOT include the deletion of an entire region; and $\Delta L_I = 0$ is excluded because we are here considering the evolution of a colored-region *after* its creation.

(to, *e.g.*, 3 or 4), because we need to keep at least $O(\{L^{CO}\}^{2N_C})$ multiplication factors in the memory during the computation for the time-slice with N_C colored regions.

In this regard, a more promising approach would be to iteratively solve the integral equation in which the "base"-operator is $\hat{Q}_0^{ID}(\tau)$ and the "perturbation" operator is $\hat{Q}^D_{M:\text{B-er}}(t)$, because the factors, $\hat{Q}^D_{M:\text{B-er}}(i;t)$'s (with $i = 1, ..., N_C - 1$), could be dealt with separately. See appendix C for more details.

[NOTE:] In this study, however, we *regard* the "boundary-eroding" deletions in Eq.25 also as "coloring-pattern-changing events", and deal with them as "perturbations" just as the "A-deleting" deletions in Eq.22 and the "D"-creating insertions (in the "base" rate operator). This means that, *hereafter*, the evolution of each colored-region in each slice is described by the "base" rate operator, $\hat{Q}_0^{ID}(i;t)$ (defined in Eq.10), and the resulting "base" multiplication factor, $\mu_{P_0}\left[(\Delta L(i;F), t^-_{k+1}) \mid (\Delta L(i;I), t^+_k)\right]$.

4.4 Computing Probability of Change in "A/D"-Coloring-Pattern

Broadly speaking, changes in the "A/D"-coloring patterns can be classified into two categories: (a) deletion of an "A"-region, possibly accompanying the size reduction of either or both of the flanking "D"-regions, as well as their merger; and (b) creation of a new "D"-region, possibly accompanying the split of the "parent" "A"-region into two "A"-regions. Changes in category (a) are caused by the "A-deleting" rate operator, $\hat{Q}^D_{M:\text{A-del}}(t)$ in Eq.22. In contrast, changes in category (b) are caused by the insertion operators on the "A"-regions in $\hat{Q}_0^{ID}(t)$ (in Eq.16). (In addition, in this study, we include (c) "boundary-eroding" deletions, caused by $\hat{Q}^D_{M:\text{B-er}}(t)$ in Eq.24, into the "perturbation".)

Deferring the summation of the contributions from different indel operators to the next subsection, here we focus on the computation of the action of each *single* indel operator. Each rate operator that accommodates the indel operator always takes the form of a distribution, which originates from the indel rates that the indel operators are multiplied by. Therefore,

the action of a rate operator, e.g., $\hat{Q}^D_{M:\text{A-del}}(t)$ at a "moment", t, always occurs in the form, $dt\ \hat{Q}^D_{M:\text{A-del}}(t)$, where dt is an infinitesimal time-element. In numerical computation, the minimum time-width, denoted here as Δt (e.g., $= (t_F - t_I)/N_P$), or its multiple in some cases, will play the role of dt. When a pattern-changing indel occurs at (actually, closely around) time t_k, we pick the minimum time-interval, $[t_a, t_a + \Delta t]$, that encompasses t_k. We assume that the indel occurs somewhere within the aforementioned time-interval, but that we do not know exactly when. Naïvely, the action of the relevant term in $dt\ \hat{Q}^D_{M:\text{A-del}}(t)$ (or $dt\ \hat{Q}^{ID}_0(t)$) should be the combination of the following:

(1) deleting a specified set of sites, or inserting a specified number of sites at a specified position, in the sequence state before the indel; and

(2) multiplying the probability (computed up to there) by Δt times the relevant indel rate.

When Δt is *sufficiently* small, *i.e.*, $\Delta t \ll 1/|\Delta R^{ID}(\Delta L, t)|$, where ΔL is the length change caused by the relevant indel, this naïve prescription works well. If, however, Δt is relatively large, *e.g.*, $\Delta t \times |\Delta R^{ID}(\Delta L, t)| \geq O(1)$, this may not be the case, because other indel events can get involved with a non-negligible probability, especially when a long subsequence is inserted/deleted. In order to keep the high accuracy of the computation even when Δt is relatively large, one way would be to use a sufficiently small time-interval , say, $\Delta' t (\ll \Delta t)$, *only* to compute the probabilities of "A/D"-coloring changes, and to graft them with the transition probabilities (for the time-lapse $\Delta t - \Delta' t$) under an unchanged "A/D"-coloring pattern. This should work as long as the "A/D"-coloring changes are relatively rare.

4.5 Computing Total Contributions from Each Pattern of "A/D"-Coloring Evolution

In the previous subsection, we focused on the contribution from each single coloring-pattern-changing insertion/deletion event. Here, we consider the collective effects of each type of

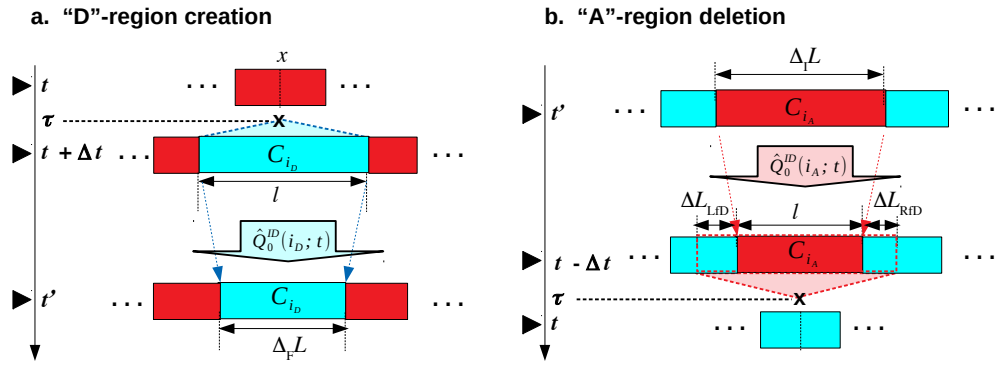

Figure 8: **Partial coloring-pattern evolution underlying multiplication factors on "D"-region creation and "A"-region deletion.** **a.** Evolution underlying the "D"-creation factor, $\mu_{\text{D-cr}}[(x,[t,t+\Delta t])\ ;\ (\Delta_F L, t')]$, in Eq.38. **b.** Evolution underlying the "A"-deletion factor, $\mu_{\text{A-del}}[(\Delta_I L, t')\ ;\ (0,-\delta\Delta L_{fD},[t-\Delta t,t])](\sigma_{BE}(=0))$, in Eq.41. As in previous figures, the red and cyan rectangles represent an "A"-region and a "D"-region, respectively. Note that, in panel **b**, $\delta\Delta L_{fD} \stackrel{\text{def}}{=} \delta\Delta L_{LfD} + \delta\Delta L_{RfD}$. See the text for more details on each setting.

such insertions/deletions . (As in the previous subsections, a locally space-homogeneous model is assumed also here. Thus, $g_I(l,t)$ is the rate of an insertion of length l at time t, and $g_D(l,t)$ is the rate of the deletion of a length l sub-sequence at time t; these rates are independent of the positions of the events as long as they occur in between the PASs, L and R.)

With the current definition of the "A/D"-coloring (given in Section 2), the creation of a "D"-region is relatively easy to deal with, because each "D"-region is created by a single insertion event, and because all sites inserted by the event belong to the "D"-region at its beginning (Figure 8 a). It should also be noted that the insertion event creating the "D"-region belongs to the "base" rate-operator, $\hat{Q}_0(i_A;\tau)$, where τ is the time of the creation, and C_{i_A} at τ is the "A"-region from which the "D"-region was created. Hence, the probability (more precisely, the multiplication factor) that a "D"-region(, referred to as C_{i_D},) was created at a specific site, say x in C_{i_A}, at some time in a (small) time-interval, $[t,t+\Delta t]$, and that

the region evolved, without vanishing, to have $\Delta_F L$ sites at a later time, t', is calculated as:

$$\begin{aligned}
&\mu_{\text{D-cr}}\left[(x,[t,t+\Delta t])\ ;\ (\Delta_F L, t')\right] \\
&= \sum_{l=1}^{\infty}\left[\int_{t}^{t+\Delta t} d\tau\ g_I(l,\tau)\cdot\mu_{P_0}\left[(\Delta_F L,t')\mid(l,\tau)\right](\sigma_{BE}=2)\ \right].
\end{aligned} \tag{38}$$

On the left-hand side, we omitted the trivial dependence of the $\mu_{\text{D-cr}}[\cdots]$ on C_{i_A} and C_{i_D}. And the $\mu_{P_0}[\cdots]$ on the right-hand side can be computed as described in subsection 4.3, with $\sigma_{BE}\ (=\sigma_B(i_D)+\sigma_E(i_D))=2$, as explicitly shown as an argument of $\mu_{P_0}[\cdots]$.

When numerically computing Eq.38, we apply the prescription given in subsection 4.4, and also set the upper-bound L_I^{CO} of the insertion length. When Δt is *sufficiently* small, the right-hand side of Eq.38 can be approximated as:

$$\Delta t\times\sum_{l=1}^{L_I^{CO}}\Big[g_I(l,t)\cdot\mu_{P_0}\left[(\Delta_F L,t')\mid(l,t+\Delta t)\right](\sigma_{BE}=2)\Big]. \tag{39}$$

And, when Δt is relatively large, the measure explained in subsection 4.4 is specifically expressed as follows (using $\Delta' t(\ll\Delta t)$):

$$\begin{aligned}
&\mu_{\text{D-cr}}\left[(x,[t,t+\Delta' t])\ ;\ (\Delta_F L,t')\right] \\
\approx\ &\sum_{l'=1}^{\infty}\Big[\Delta' t\times\sum_{l=1}^{L_I^{CO}}\Big[g_I(l,t)\cdot\mu_{P_0}\left[(l',t+\Delta t)\mid(l,t+\Delta' t)\right](\sigma_{BE}=2)\Big] \\
&\times\mu_{P_0}\left[(\Delta_F L,t')\mid(l',t+\Delta t)\right](\sigma_{BE}=2)\ \Big]. \\
=\ &\Delta' t\times\sum_{l=1}^{L_I^{CO}}\Big[g_I(l,t)\cdot\mu_{P_0}\left[(\Delta_F L,t')\mid(l,t+\Delta' t)\right](\sigma_{BE}=2)\Big].
\end{aligned} \tag{40}$$

The space-complexity required to store all these factors is $O(N_P\ L^{CO})$ (except for $\mu_{P_0}[\cdots]$'s that require $O(N_P\{L^{CO}\}^2)$ memory space), and the time-complexity required to compute these equations is $O(N_P\ \{L^{CO}\}^2)$, making the computation feasible. Because the relevant "D"-region did not exist before t, the accompanying multiplication factor is 1 (unity) in this case.

Next, we will deal with a relatively difficult case, which is the deletion of an "A"-region (Figure 8 b). As indicated by Eq.23, the deletion of an "A"-region could, in general, ad-

ditionally delete some sites in the flanking "D"-regions. Therefore, when computing their contributions to the probabilities of case-(iv) gapped segments, the number of such additionally deleted sites must also be recorded. Let $\delta\Delta L_{LfD}$ and $\delta\Delta L_{RfD}$, respectively, be the numbers of such deleted sites in the left-flanking and right-flanking "D"-regions. In a general insertion/deletion model, both $\delta\Delta L_{LfD}$ and $\delta\Delta L_{RfD}$ must be recorded, which means that we must store $O(N_P\ \{L^{CO}\}^3)$ factors, which may be quite hard on the computer(s). (Note that one $O(L^{CO})$ comes from the "initial" length of the "A"-region to be deleted.) Fortunately, in a locally space-homogeneous model considered here, *after factoring out the transition probabilities for the flanking "D"-regions*, the remaining factor depends on the number of deleted "D"-sites *only through* the summation, $\delta\Delta L_{fD} \stackrel{\text{def}}{=} \delta\Delta L_{LfD} + \delta\Delta L_{RfD}$, because the deletion rate, $g_D(l,t)$, depends *only* on the deletion length, *i.e.*, the total number of sites deleted.

Thus, the factor we need to store should be the probability (more precisely, the multiplication factor) that an "A"-region(, referred to as C_{i_A},), which had $\Delta_I L$ sites at time t', was later deleted at some time in a (small) time-interval, $[t-\Delta t, t]$, simultaneously with the deletion of $\delta\Delta L_{fD}$ sites in the flanking "D"-region(s). It is expressed as:

$$\begin{aligned} & \mu_{\text{A-del}}\left[(\Delta_I L, t')\ ;\ (0, -\delta\Delta L_{fD}, [t-\Delta t, t])\right](\sigma_{BE}) \\ = & \sum_{l=1}^{\infty}\left[\int_{t-\Delta t}^{t} d\tau\ g_D(l+\delta\Delta L_{fD}, \tau)\cdot\mu_{P_0}\left[(l,\tau)\mid(\Delta_I L, t')\right](\sigma_{BE})\right]. \end{aligned} \tag{41}$$

Here, again, on both hand sides, we explicitly recorded the dependence of the multiplication factors on $\sigma_{BE} \stackrel{\text{def}}{=} \sigma_B + \sigma_E$, which could significantly change the results. Especially, it should be noted that only $\delta\Delta L_{fD} = 0$ is meaningful when $\sigma_{BE} = 2$, because there are no flanking "D"-regions in such a case.

In the numerical computation, we do as in subsection 4.4 and set the upper-bound, L_D^{CO}, of the deletion length. When Δt is *sufficiently* small, the right-hand side of Eq.41 can be approximated as:

$$\Delta t \times \sum_{l=1}^{L_D^{CO}-\delta\Delta L_{fD}}\Big[g_D(l+\delta\Delta L_{fD}, t)\cdot\mu_{P_0}\left[(l, t-\Delta t)\mid(\Delta_I L, t')\right](\sigma_{BE})\Big]. \tag{42}$$

When, Δt is relatively large, the measure explained in subsection 4.4 is specifically expressed (using $\Delta' t (\ll \Delta t)$), just as in the "creation" case above:

$$\begin{aligned} & \mu_{\text{A-del}}\left[(\Delta_I L, t')\ ;\ (0, -\delta\Delta L_{fD}, [t-\Delta' t, t])\right](\sigma_{BE}) \\ \approx\ & \Delta' t \times \sum_{l=1}^{L_D^{CO}-\delta\Delta L_{fD}} \Big[g_D(l+\delta\Delta L_{fD}, t) \cdot \mu_{P_0}\left[(l, t-\Delta' t)\mid(\Delta_I L, t')\right](\sigma_{BE}) \Big]. \end{aligned} \tag{43}$$

If nothing else is done to *effectively* adjust the exponential factor, any of these factors, Eqs.41, 42, 43, *could* be accompanied by the multiplication factor:

$$\exp\left\{ -\int_{t-\Delta t}^{t} d\tau \Delta R^{ID}(\Delta L_{afD}(t-\Delta t), \tau) \right\}\ , \tag{44}$$

where $\Delta L_{afD}(t-\Delta t)$ is the total number of sites in the affected flanking "D"-region(s) at time $t-\Delta t$. [23]

Now, *at last*, we can compute the probabilities (or their total) of the insertion/deletion histories having a given pattern of "A/D"-coloring evolution. As described in the beginning of this section (section 4), we first identify the times at which the coloring-pattern changes, [24] and slice the coloring pattern history at these times. Then, for each colored region, say C_i in each slicegiven by an (open) time-interval, *e.g.*, (t_k, t_{k+1}), assign

$$\mu_{P_0}\left[(\Delta L(i; t_{k+1}^-), t_{k+1}^-)\mid(\Delta L(i; t_k^+), t_k^+)\right](\sigma_{BE}(i))$$

if it remained existing at both t_k and t_{k+1},

$$\mu_{\text{D-cr}}\left[(x, [t_k, t_k+\Delta t])\ ;\ (\Delta L(i; t_{k+1}^-), t_{k+1}^-)\right]$$

[23] It should be noted, however, that the introduction of multiplication factors like this is equivalent to assuming that *no* indel events hit the relevant regions during the time-interval in question ($[t-\Delta t, t]$ in the above case). Such an assumption should hold well if $\Delta t \cdot |\Delta R^{ID}(\Delta L, \tau)| \ll 1$, where ΔL is the total length of the relevant regions. Otherwise, it would be better to "pad" the time-interval with the transition probabilities under the unchanged "A/D"-coloring pattern, though it might sometimes be time-consuming.

[24] Remember that we also regard the "boundary-eroding" deletions in Eq.25 as belonging to "pattern-changing" events.

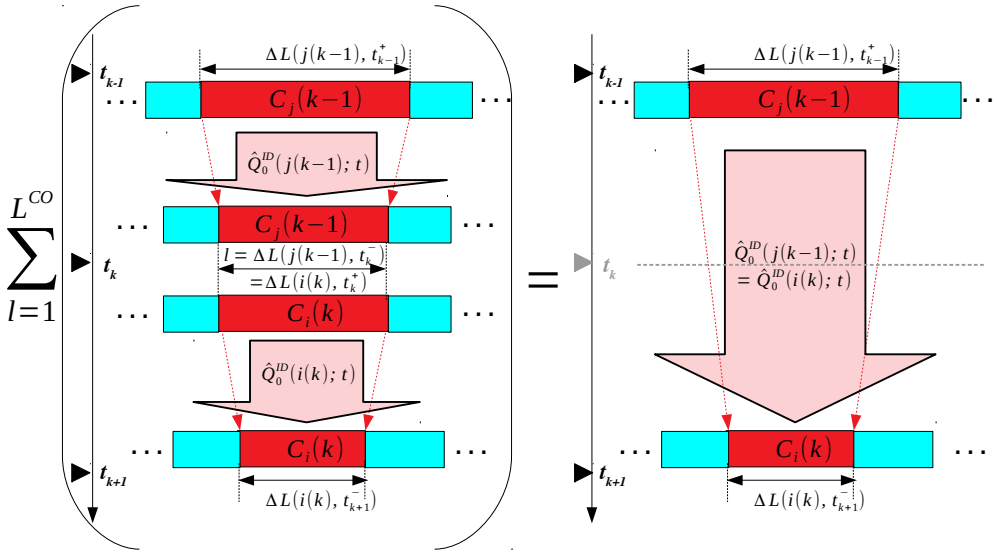

Figure 9: **Colored region extending across slice boundary.** When a colored region (in this case, $C_i(k-1)$ or $C_j(k)$) extends across the boundary (t_k) between two slices ($k-1$ over (t_{k-1}, t_k) and slice k over (t_k, t_{k+1})), the multiplication factor for this region can span these slices. All you have to do is match the region lengths at the boundary, and sum over the lengths, as shown in this schematic equation. In this figure, the notations $j(k-1)$ and $i(k)$ represent the j th region in slice $k-1$ and the i th region in slice k, respectively. Although an "A"-region (red) extends across the boundary in this figure, a "D"-region (cyan) can also extend similarly. (Remember that, for an "A"-region to extend, its σ_{BE} needs to remain unchanged.)

if it was created immediately after t_k, and

$$\mu_{\text{A-del}}\left[(\Delta L(i; t_k^+), t_k^+)\ ;\ (0, -\delta\Delta L_{fD}, [t_{k+1} - \Delta t, t_{k+1}])\right](\sigma_{BE})$$

if it was deleted immediately before t_{k+1}. Then, at each time t_k, sum over all possible lengths, $\Delta L(i; t_k^+)$'s and $\Delta L(j, t_k^-)$'s, of the regions involved ($C_i(k)$'s in the slice after t_k, and $C_j(k-1)$'s in the slice before t_k). If a colored region extends across the time t_k without being affected by any perturbations (Figure 9)(, which means that even σ_{BE} remains unchanged), the relevant summation (and the accompanying "matching" of the pair of lengths across t_k) "join"s the multiplication factors before and after t_k, giving a multiplication factor defined in the *joint* interval, say, (t_{k-1}, t_{k+1}).

This way, we are left only with the summations at the "perturbation" events, each regarding the regions affected by the event. After computing these summations, we can, *at least in principle*, obtain the total contribution for a given pattern evolution by "integrating" over the times of the events. *In practice*, however, especially when multiple events are in-

volved, the required multiple-time integration could be extremely time-consuming, *if naïvely performed*. It should therefore be desirable to, *whenever possible,* deal with the "perturbations" one after another, each time computing the summation and the time-integration required for each event, to store the "interim" results, and then to re-use them for the "next step" of the computation. This strategy is somewhat similar to that of the "pruning" (aka, "peeling") algorithm for the computation of the likelihood of a phylogenetic tree under a given substitution model, given a sequence data set [59] [60] [61]. In the next section, we will apply these procedures to the concrete computations of the probabilities when the number of perturbations are relatively small.

5 Concretely Computing Contributions to Case-(iv) Probabilities at Given Perturbation Levels

In Section 4, we mathematically derived the "ingredients" to compute the probabilities of case-(iv) gap-patterns. Now, by putting these "ingredients" together, we will concretely compute the probabilities at relatively low perturbation levels, to illustrate the computation procedures explained at the bottom of subsection 4.5. Throughout this section, again, it is assumed that the computations are performed under a locally space-homogeneous model. Moreover, we assume that we consider that the ancestral sequence existed at the "initial time", t_I, and the descendant sequence was sampled (or examined) at the "final time", t_F. This means that the time-interval, $[t_I, t_F]$, gives the entire time-frame for the evolution of the sequence we are interested in. And let $\Delta_I L_A$ be the number of ancestral sites at time t_I, and let $\Delta_F L_D$ be the number of descendant sites at time t_F, both excluding the PASs, L and R. Then, what we want is the probability of having descendant sites at t_F *conditioned on* totally non-homologous ancestral sites at t_I:

$$P_{\text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right]$$

$$\stackrel{\text{def}}{=} P\left[([L, \upsilon'(1), \cdots, \upsilon'(\Delta_F L_D), R],\ t_F) \mid ([L, \upsilon(1), \cdots, \upsilon(\Delta_I L_A), R],\ t_I)\right] \Big|_{\substack{\upsilon(i) \neq \upsilon'(j) \\ \text{for } {}^\forall i = 1, ..., \Delta_I L_A; \\ {}^\forall j = 1, ..., \Delta_F L_D}} \quad (45)$$

When we compute the probabilities of ancestor-descendant PWAs, however, what we actually need is the multiplication factor:

$$\mu_{P\ \text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] \stackrel{\text{def}}{=} \frac{P_{\text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right]}{P\left[([\ ], [t_I, t_F]) \mid (\Delta_I L_A, t_I)\right]}\ , \quad (46)$$

where $P\left[([\ ], [t_I, t_F]) \mid (\Delta_I L_A, t_I)\right]$ is the probability that no indel events hit the region during $[t_I, t_F]$, conditioned on $\Delta_I L_A$ ancestral sites in between the PASs, L and R, at t_I. Then, in our perturbation method, the multiplication factor (Eq.46) is expanded as:

$$\begin{aligned} & \mu_{P\ \text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] \\ = \ & \mu^{\text{2nd}}_{P\ \text{case-(iv)}}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \ +\ \mu^{\text{3rd}}_{P\ \text{case-(iv)}}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \ +\ \cdots \end{aligned} \quad (47)$$

Here, $\mu^{\text{2nd}}_{P\ \text{case-(iv)}}[\cdots]$ and $\mu^{\text{3rd}}_{P\ \text{case-(iv)}}[\cdots]$ are the total summations of all 2nd-order contributions and all 3rd-order contributions, respectively, to the multiplication factor, Eq.46.

It should be noted that, under the locally space-homogeneous model we are considering, the identity:

$$\begin{aligned} & P_{\text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] \times \exp\left(+\int_{t_I}^{t_F} d\tau R^{ID}_X([L, R], \tau)\right) \\ \equiv \ & \mu_{P\ \text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] \times \exp\left(-\int_{t_I}^{t_F} d\tau \Delta R^{ID}_X(\Delta_I L_A, \tau)\right) \end{aligned} \quad (48)$$

holds, thanks to the equation:

$$P\left[([\ ], [t_I, t_F]) \mid (\Delta_I L_A, t_I)\right] = \exp\left(-\int_{t_I}^{t_F} d\tau \Big[R^{ID}_X([L, R], \tau) + \Delta R^{ID}_X(\Delta_I L_A, \tau)\Big]\right)\ .$$

Actually, the identities similar to Eq.48 hold also for the partial contributions to the probabilities of case-(iv) gapped segments, and they will be used frequently in the following sub-sections.

Once we obtain the analytical form of each contribution, which includes some time-integrations, the numerical computation of each time-integration could be most simply done

with the trapezoidal formula (see, *e.g.*, Press et al. [62]). If you would, however, Sympson's formula could also be applied (again, see, *e.g.*, Press et al. [62]), with some extra care in the vicinity of the boundaries when the time-interval is an odd number times the time-element (Δt). In the following, we will only give the analytical forms of the contributions.

5.1 Second-order Contributions

To have a case-(iv) gapped segment, at least two pattern-changing events are necessary: a "deletion" of the ancestral sites, and an "insertion" of the descendant sites. This means that the minimum perturbation level for case-(iv) is the second order. Varying in the time-order of the insertion and the deletion, as well as in their spatial-order, the second-order patterns of "A/D"-coloring evolution are broadly classified into three (Figure 10): (i) A $\rightarrow$ "$\emptyset$" $\rightarrow$ D [25] ; (ii) A $\rightarrow$ AD $\rightarrow$ D; and (iii) A $\rightarrow$ DA $\rightarrow$ D. Let us derive analytical expressions of the contributions from these patterns one by one.

5.1.1 (i) A $\rightarrow$ "$\emptyset$" $\rightarrow$ D

Let us assume that the "A"-region was completely deleted at some time in $[t_1 - dt_1, t_1]$, and that the "D"-region was created at some time in $[t_2, t_2 + dt_2]$ (with $t_1 \leq t_2$). Then the analytical expression for the contribution of pattern (i) to the probability (and multiplication factor) we want is:

$$
\begin{aligned}
& P^{\text{2nd (i)}}_{\text{case-(iv)}}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \times \exp\left(+\int_{t_I}^{t_F} d\tau R_X^{ID}([L,R],\tau)\right) \\
\equiv\ & \mu^{\text{2nd (i)}}_{P\ \text{case-(iv)}}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \times \exp\left(-\int_{t_I}^{t_F} d\tau \Delta R_X^{ID}(\Delta_I L_A,\tau)\right) \\
=\ & \int_{t_I+dt_1}^{t_F-dt_2} dt_1 \int_{t_1}^{t_F-dt_2} dt_2 \Big[\left(\mu_{\text{A-del}}\left[(\Delta_I L_A, t_I)\ ;\ (0,0,[t_1-dt_1,t_1])\right](\sigma_{BE}=2)/dt_1\right) \times \\
& \times \mu_{P\ \text{case-(i)}}\left[(0,t_2) \mid (0,t_1)\right] \times \left(\mu_{\text{D-cr}}\left[(x,[t_2,t_2+dt_2])\ ;\ (\Delta_F L_D, t_F)\right]/dt_2\right)\Big] \ . \qquad (49)
\end{aligned}
$$

[25] Here, the $\emptyset$ (for "empty-set") is double-quoted, to remind us that some evanescent sites may actually have existed between the deletion of the "A"-region and the creation of the "D"-region.

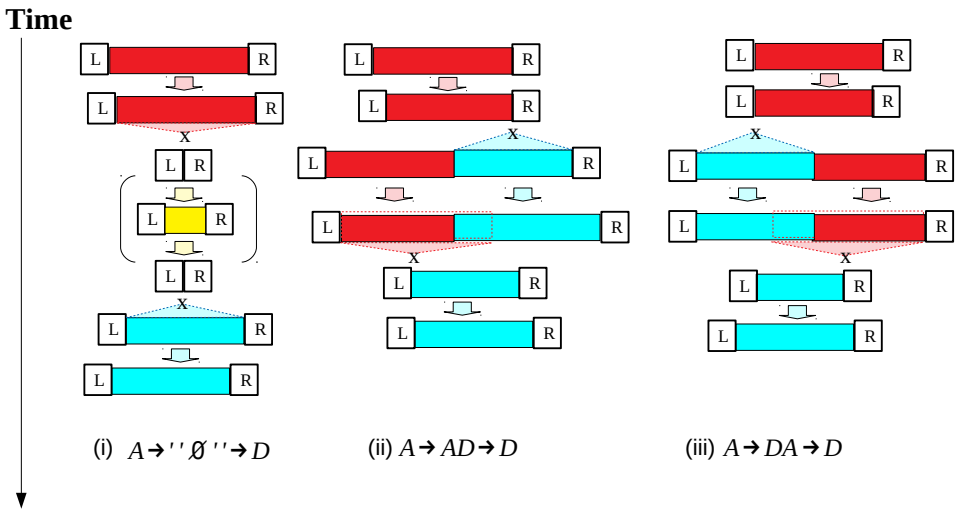

Figure 10: **Topologies of 2nd-order A/D-coloring pattern histories that result in case-(iv) gapped segments. (Left)** pattern (i) (A → "∅" → D); **(Center)** pattern (ii) (A → AD → D); **(Right)** pattern (iii) (A → DA → D). As in the previous figures, the red and cyan rectangles represent an "A"-region and a "D"-region, respectively. And the yellow rectangle (in panel a) represents a lump of evanescent sites that are not colored either "A" or "D". The transparent red triangle converging to an "X" represents the complete deletion of an "A"-region; the transparent cyan triangle diverging from an "X" represents the creation of a "D"-region. The thin-colored downward arrow represents evolution via the "base" rate operator.

Here, $\mu_{P\ \text{case-(i)}}\left[(0,t_2)\mid(0,t_1)\right]$ is the multiplication factor for the "case-(i) gapped segment", which does not have any ancestral or descendant sites in between L and R, at both times t_1 and t_2. (We do not care whether or not any evanescent sites existed during the open interval, (t_1,t_2).) This factor could be computed practically exactly as a by-product of the algorithm in SM-3 of [1], or via a faster algorithm based on the recursion relation *nearly* identical to Eq.36 with $\sigma_{BE}(i)=2$; it requires *only two* modifications: (1) inclusion of $\Delta L_I=0$ (actually, the factors with this value are all we need to compute here) and $\Delta L_j=0$, and (2) replacement of the upper-bound of the 1st summation with $\min(L_I^{CO},\ \Delta L_j)$. [26] It should also be noted that the $\mu_{\text{A-del}}[\cdots]$ and $\mu_{\text{D-cr}}[\cdots]$ above are infinitesimal probabilities of $O(dt_1)$ and $O(dt_2)$, respectively; thus their division by dt_1 and dt_2, respectively, give

[26] Alternatively, we could devise another faster algorithm based on the recursion relation *nearly* identical to Eq.62 with $\sigma_{BE}(i)=2$; again, we need *only two* modifications: (1) inclusion of $\Delta L_F=0$ (actually, the factors with this value are all we need to compute here) and $\Delta L_j=0$, and (2) replacement of the upper-bound of the 2nd summation with $\min(L_D^{CO},\ \Delta L_j)$.

probability densities. Another point to note is the existence of the dt_1, dt_2, etc. in the upper- and lower-bounds of the time-integrations; we *deliberately* employed this *extremely unconventional* notation, in order to facilitate the transition from the analytical expression to the numerical computation; (we will keep using this notation hereafter, too); if, instead, you intend to calculate these equations *completely analytically*, just ignore these infinitesimal time-elements in the integration boundaries. [27]

As argued at the bottom of subsection 4.5, it would save time to perform the calculation associated with individual events one after another. See appendix D for details on such a series of calculations. The calculations can be done with the maximum time-complexity of $O(\{N_P\}^2\{L^{CO}\}^2)$, and the space-complexity of *at most* $O(N_P\{L^{CO}\}^2)$, where N_P is the number of sub-time-intervals and L^{CO} is the upper-bounds of the number of sites included in each colored region.

5.1.2 (ii) A $\rightarrow$ AD $\rightarrow$ D

As opposed to the pattern (i), let us assume here that the "D"-region is created at some time in $[t_1, t_1 + dt_1]$, and that the "A"-region is completely deleted at some time in $[t_2 - dt_2, t_2]$ (with $t_1 + dt_1 \leq t_2 - dt_2$). In this pattern, the "base" rate operator, $\hat{Q}_0^{ID}(i = 1; t)$, for the "A"-region (C_1) has $\sigma_{BE} = 2$ for $t < t_1$ and $\sigma_{BE} = 1$ for $t_1 < t < t_2 - dt_2$. Meanwhile, the operator for the "D"-region always has $\sigma_{BE} = 2$. Let ΔL_1 be the length of the "A"-region at t_1, $\delta\Delta L_2$ be the number of sites in the "D"-region deleted in conjunction with the entire "A"-region (in $[t_2 - dt_2, t_2]$), and ΔL_2 be the length of the "D"-region iat t_2. (We also assume here that, during $[t_2 - dt_2, t_2]$, the "D"-region only suffered the deletion that erased the entire "A"-region.) Figure 11 illustrates the situation.

Under this setting, we obtain the following expression for the contribution from pattern

[27] At the top of the above equation, we explicitly wrote down the expression including the portion of $P_{\text{case-(iv)}}[\cdots]$, because this is the first concrete expression given in this section. Hereafter, however, we will omit such expressions.

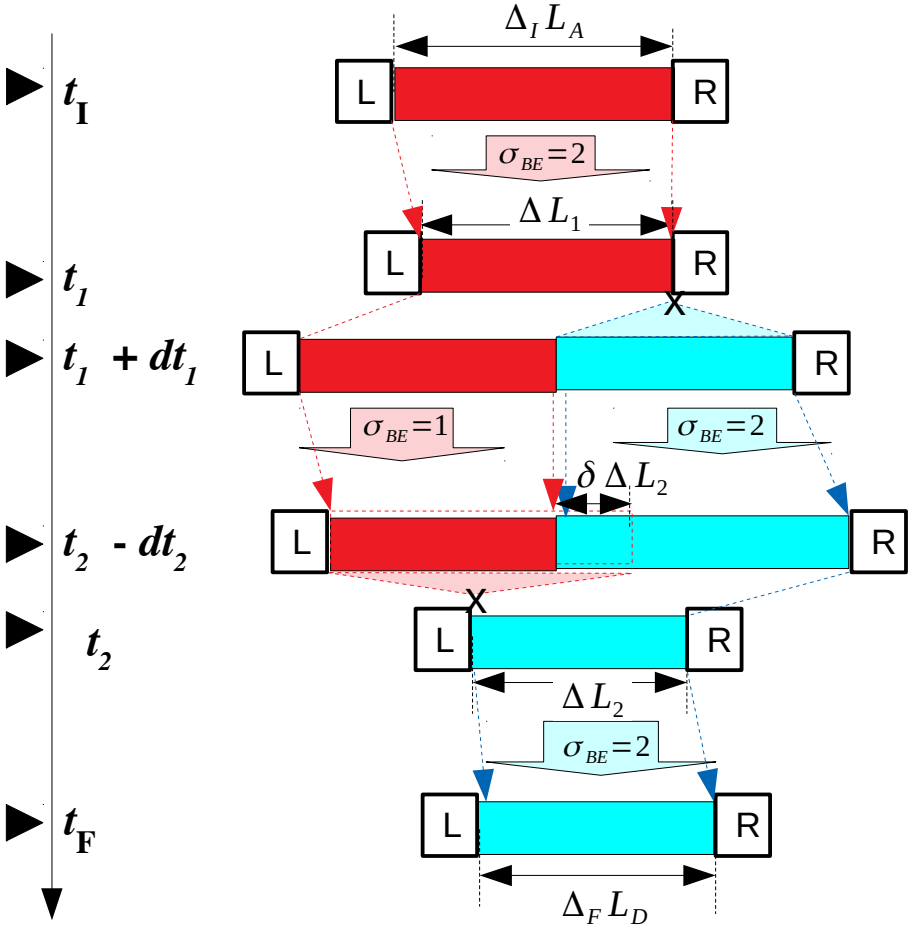

Figure 11: **Setting for computing probabilities of 2nd-order pattern (ii) (A → AD → D), in Eq.50.** This figure was created from the center panel of Figure 10, by adding further annotations. In each transparent-colored downward arrow, which indicates evolution via the "base" rate operator, only the value of σ_{BE} is shown. Note that, in this figure (and in previous figures), dt_1, etc. are made disproportionately large, to clearly show changes via insertions/deletions.

(ii):

$$
\begin{aligned}
& \mu^{\text{2nd (ii)}}_{P\ \text{case-(iv)}}\left[(\Delta_F L_D,\ t_F)\mid(\Delta_I L_A,\ t_I)\right]\times\exp\left(-\int_{t_I}^{t_F} d\tau \Delta R^{ID}_X(\Delta_I L_A,\tau)\right) \\
= & \sum_{\Delta L_1=1}^{\infty}\sum_{\Delta L_2=1}^{\infty}\sum_{\delta\Delta L_2=0}^{\infty}\int_{t_I}^{t_F-dt_1-dt_2} dt_1 \int_{t_1+dt_1+dt_2}^{t_F} dt_2 \Big[\mu_{P_0}\left[(\Delta L_1,t_1)\mid(\Delta_I L_A,t_I)\right](\sigma_{BE}=2)\times \\
& \times\left(\mu_{\text{D-cr}}\left[(x=\Delta L_1,[t_1,t_1+dt_1])\ ;\ (\Delta L_2+\delta\Delta L_2,t_2-dt_2)\right]/dt_1\right)\times \\
& \times\exp\left(-\int_{t_2-dt_2}^{t_2} d\tau_2\ \Delta R^{ID}_X(\Delta L_2+\delta\Delta L_2,\tau_2)\right)\times \\
& \times\left(\mu_{\text{A-del}}\left[(\Delta L_1,t_1)\ ;\ (0,-\delta\Delta L_2,[t_2-dt_2,t_2])\right](\sigma_{BE}=1)/dt_2\right)\times
\end{aligned}
$$

$$\times \mu_{P_0}\left[(\Delta_F L_D, t_F) \mid (\Delta L_2, t_2)\right](\sigma_{BE} = 2)\Big] \ . \tag{50}$$

It should be noted here that, *normally*, the argument, $t_2 - dt_2$ of the $\mu_{\text{D-cr}}[\cdots]$ should be expressed as t_2, because the result remain unchanged for the integration of a well-behaved function of time, and the exponential factor in the 3rd last line should be omitted because it usually gives 1 (unity). Here, however, we *deliberately* chose these expressions, to facilitate the *translation* to the numerical computation. [28]

In numerical computation, the upper-limits of the above summations must be finite numbers, instead of infinity. For example, if we consider each region-length to be L^{CO} or less, the above triple-summations will be replaced by:

$$\sum_{\Delta L_1=1}^{L^{CO}} \sum_{\Delta L_2=1}^{L^{CO}} \sum_{\delta\Delta L_2=0}^{L^{CO}-\Delta L_2} \ .$$

As usual, it should be time-efficient to perform the computations associated with individual events separately, one after another. Such "pruning-like" computation in this case is detailed in appendix E. The series of computations can be performed with the maximum time-complexity of $O(\{N_P\}^2\{L^{CO}\}^3)$ and the maximum space-complexity of $O(N_P\{L^{CO}\}^2)$.

5.1.3 (iii) A $\rightarrow$ DA $\rightarrow$ D

Actually, contributions from this pattern is identical to those from pattern (ii), as long as the indel rates are symmetric regarding the reversal of the spacial order of sites, which is indeed the case with the locally space-homogeneous model we are now dealing with. Thus, we can just "borrow" the results of the pattern (ii), to compute the contributions from the pattern (iii) here. This is most easily implemented by *doubling* the results in the previous sub-subsection. (Even if the model is not spatially symmetric, we can easily modify the equations for the pattern (ii), to obtain those for the pattern (iii).)

[28] As already argued in footnote 23, it would be better to replace the exponential factor with appropriate transition probabilities under the unchanged "A/D"-coloring pattern, if the numerical counterpart of dt is *not* sufficiently small.

5.1.4 Summary

By combining all these results obtained in the previous sub-subsections, we can obtain the total contributions from all the second-order terms, as follows:

$$\begin{aligned} & \mu_{P\ \text{case-(iv)}}^{2\text{nd}}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \\ = & \sum_{\alpha=\ \text{i},\ \text{ii},\ \text{iii}} \mu_{P\ \text{case-(iv)}}^{2\text{nd }(\alpha)}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \\ = & \ \mu_{P\ \text{case-(iv)}}^{2\text{nd (i)}}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] + 2 \times \mu_{P\ \text{case-(iv)}}^{2\text{nd (ii)}}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \ . \quad (51) \end{aligned}$$

5.2 Third-order Contributions

The third-order terms are contributed by "A/D"-coloring evolution patterns with three, *i.e.*, one plus the minimum (=2), pattern-changing events. In addition to the creation of a "D"-region and the complete deletion of an "A"-region, the additional event could be an insertion or a deletion. Besides, the additional event could be a "boundary-eroding" deletion included in the perturbation, Eq.21. Varying in (1) the kind of the additional event, (2) the time-order of the events, and (3) their spatial relationships, the third-order "A/D"-coloring evolution patterns can be broadly classified into the following six (Figure 12): (a) A $\rightarrow$ ADA $\rightarrow$ AD $\rightarrow$ D; (b) A $\rightarrow$ ADA $\rightarrow$ DA $\rightarrow$ D; (c) A $\rightarrow$ DA $\rightarrow$ DAD $\rightarrow$ D; (d) A $\rightarrow$ AD $\rightarrow$ DAD $\rightarrow$ D; (e) A $\rightarrow$ DA $\stackrel{\text{B-er}}{\rightarrow}$ DA $\rightarrow$ D; and (f) A $\rightarrow$ AD $\stackrel{\text{B-er}}{\rightarrow}$ AD $\rightarrow$ D. In (e) and (f), the "B-er" indicates that a "boundary-eroding" deletion occurred then. As you can see, the patterns (b), (d) and (f) are the space-reversal of the patterns (a), (c) and (e), respectively. If the indel evolution model we consider is symmetric under the space-reversal(, which is indeed the case here), (b), (d) and (f) give contributions identical to those of (a), (c) and (e), respectively. (Even if otherwise, the analytical expressions of the former's contributions can be easily derived by properly modifying those of the latter's.) Hence, in the following, we will only describe contributions from the patterns, (a), (c), and (e).

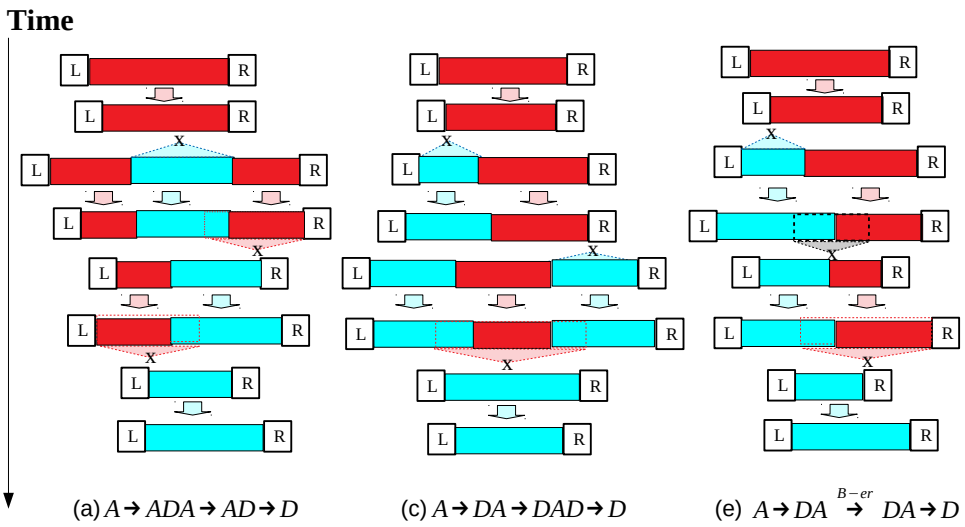

Figure 12: **Topologies of 3rd-order "A/D"-coloring pattern histories that result in case-(iv) gapped segments. (Left)** pattern (a) (A $\rightarrow$ ADA $\rightarrow$ AD $\rightarrow$ D); **(Center)** pattern (c) (A $\rightarrow$ DA $\rightarrow$ DAD $\rightarrow$ D); **(Right)** pattern (e) (A $\rightarrow$ DA $\overset{\text{B-er}}{\rightarrow}$ DA $\rightarrow$ D). The notations are basically the same as in the 2nd-order "A/D"-coloring-pattern histories (Figure 10). The transparent black triangle converging to an "X" represents a "boundary-eroding" deletion. Patterns (b), (d) and (f) were omitted here; they can be obtained via the space-reversal of patterns (a), (c) and (e), respectively.

5.2.1 (a) A $\rightarrow$ ADA $\rightarrow$ AD $\rightarrow$ D

Let us assume that the "D"-region was created at some time in $[t_1, t_1 + dt_1]$, that the "A"-region on the right was completely deleted at some time in $[t_2 - dt_2, t_2]$, and that the remaining "A"-region was completely deleted at some time in $[t_3 - dt_3, t_3]$, with $t_1 + dt_1 \leq t_2 - dt_2$ and $t_2 \leq t_3 - dt_3$. In this pattern, the "base" rate operators, $\hat{Q}_0^{ID}(i;t)$'s, for the "A"-regions have $\sigma_{BE} = 2$ before the "D"-creation (i.e., $t < t_1$), and $\sigma_{BE} = 1$ after the "D"-creation (i.e., $t_1 < t (< t_3)$). As usual, the "D"-region always has $\sigma_{BE} = 2$. Let $\Delta_L L_1$ and $\Delta_R L_1$ be the sizes of the left- and right-fragments, respectively, of the "A"-region at t_1. [29] Let $\delta\Delta L_2$ be the number of sites in the "D"-region deleted in conjunction with the entire "A"-region on the right (in $[t_2 - dt_2, t_2]$), and ΔL_2 be the size of the "D"-region immediately after this deletion (and also assume that the "D"-region suffered no other indels during $[t_2 - dt_2, t_2]$).

[29] The "left- and right-fragments" here mean the fragments of the "A"-region that are on the left and right, respectively, of the position where the "D"-region was inserted. (We assume that indels involving the position did not occur after t_1 and before the insertion of the "D"-region.)

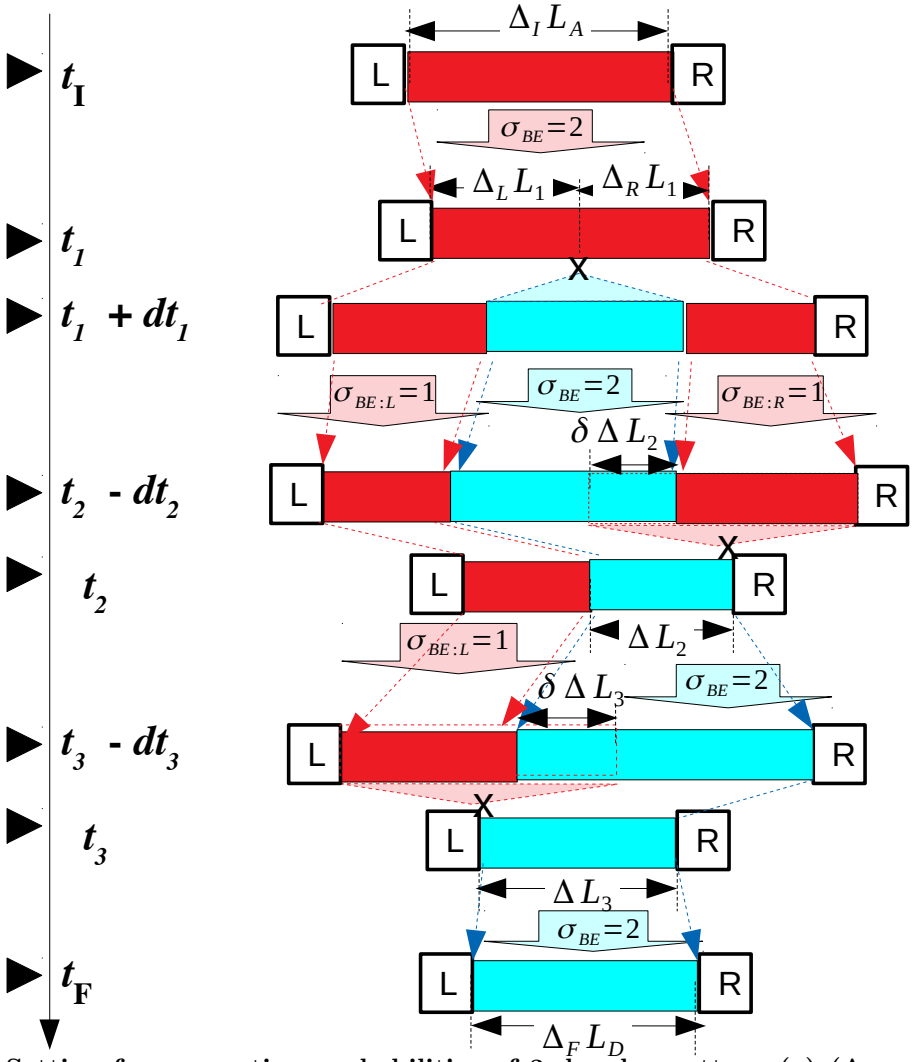

Figure 13: **Setting for computing probabilities of 3rd-order pattern (a) (A → ADA → AD → D), in Eq.52.** This figure was created from the left panel of Figure 12, by adding further annotations. Notes similar to those on Figure 11 apply also here.

And let $\delta\Delta L_3$ be the number of sites in the "D"-region deleted in conjunction with the entire remaining "A"-region (in $[t_3 - dt_3, t_3]$), and ΔL_3 be the site of the "D"-region immediately after this deletion (and also assume that the "D" region suffered no other indels during $[t_3 - dt_3, t_3]$. Figure 13 illustrates the setting.

Under this setting, the contributions from the pattern (a) can be *analytically* expressed as:

$$\mu^{\text{3rd (a)}}_{P\ \text{case-(iv)}}\left[(\Delta_F L_D,\ t_F)\mid(\Delta_I L_A,\ t_I)\right]\times\exp\left(-\int_{t_I}^{t_F} d\tau \Delta R_X^{ID}(\Delta_I L_A,\tau)\right)$$

$$
\begin{aligned}
= \sum_{\Delta_L L_1=1}^{\infty} \sum_{\Delta_R L_1=1}^{\infty} \sum_{\Delta L_2=1}^{\infty} \sum_{\delta\Delta L_2=0}^{\infty} \sum_{\Delta L_3=1}^{\infty} \sum_{\delta\Delta L_3=0}^{\infty} \int_{t_I}^{t_F-dt_1-dt_2-dt_3} dt_1 \int_{t_1+dt_1+dt_2}^{t_F-dt_3} dt_2 \int_{t_2+dt_3}^{t_F} dt_3 \Big[\\
& \mu_{P_0}\left[(\Delta L_1 = \Delta_L L_1 + \Delta_R L_1, t_1) \mid (\Delta_I L_A, t_I)\right](\sigma_{BE}=2) \times \\
& \times \left(\mu_{\text{D-cr}}\left[(x=\Delta_L L_1, [t_1, t_1+dt_1])\ ;\ (\Delta L_2 + \delta\Delta L_2, t_2 - dt_2)\right]/dt_1\right) \times \\
& \times \exp\left(-\int_{t_2-dt_2}^{t_2} d\tau_2\ \Delta R_X^{ID}(\Delta L_2 + \delta\Delta L_2, \tau_2)\right) \times \\
& \times \left(\mu_{\text{A-del}}\left[(\Delta_R L_1, t_1)\ ;\ (0, -\delta\Delta L_2, [t_2 - dt_2, t_2])\right](\sigma_{BE:R}=1)/dt_2\right) \times \\
& \times \mu_{P_0}\left[(\Delta L_3 + \delta\Delta L_3, t_3 - dt_3) \mid (\Delta L_2, t_2)\right](\sigma_{BE}=2) \\
& \times \exp\left(-\int_{t_3-dt_3}^{t_3} d\tau_3\ \Delta R_X^{ID}(\Delta L_3 + \delta\Delta L_3, \tau_3)\right) \times \\
& \times \left(\mu_{\text{A-del}}\left[(\Delta_L L_1, t_1)\ ;\ (0, -\delta\Delta L_3, [t_3 - dt_3, t_3])\right](\sigma_{BE:L}=1)/dt_3\right) \times \\
& \times \mu_{P_0}\left[(\Delta_F L_D, t_F) \mid (\Delta L_3, t_3)\right](\sigma_{BE}=2)\Big] \ . \qquad (52)
\end{aligned}
$$

(Here, the same notes as in the previous subsection apply to the notations and factors with analytically no effects, which were introduced merely to facilitate the *translation* to the numerical computation.)

This expression involves summations over six lengths and integrations over three time-intervals. Thus, if naïvely performed, it could take too long to finish within a reasonable amount of time. Therefore, following the general strategy already proposed, we will perform computations associated with the individual events, one after another. See appendix F for details. The series of computations can be performed with the maximum time-complexity of max [$O(\{N_P\}^2\{L^{CO}\}^3),\ O(N_P\{L^{CO}\}^4)$] and the maximum space-complexity of $O(N_P\{L^{CO}\}^2)$.

5.2.2 (c) A → DA → DAD → D

Let us assume that the "D"-regions on the left and on the right were created at some times in $[t_1, t_1+dt_1]$ and $[t_2, t_2+dt_2]$, respectively, and that the "A"-region was completely deleted at some time in $[t_3-dt_3, t_3]$, with $t_1+dt_1 \le t_2$ and $t_2+dt_2 \le t_3-dt_3$. Here, the "base" rate operators, $\hat{Q}_0^{ID}(i;t)$'s, for the "A"-region has $\sigma_{BE}=2$ for $t<t_1$, $\sigma_{BE}=1$ for $t_1<t<t_2$,

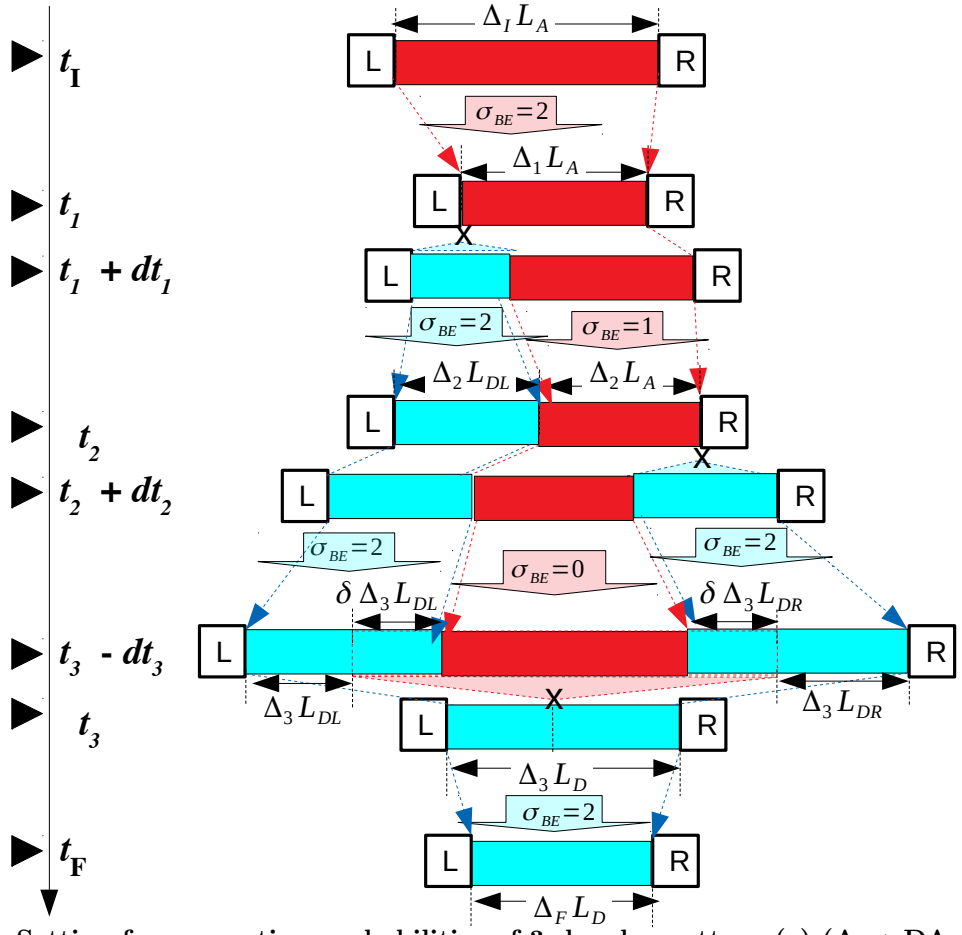

Figure 14: **Setting for computing probabilities of 3rd-order pattern (c) (A → DA → DAD → D), in Eq.53.** This figure was created from the center panel of Figure 12, by adding further annotations. Notes similar to those on Figure 11 apply also here.

and $\sigma_{BE} = 0$ for $t_2 < t < t_3$. Let $\Delta_1 L_A$ and $\Delta_2 L_A$ be the sizes of the "A"-region at t_1, t_2, respectively; let $\Delta_3 L_{DL}$ be the size of the "D"-region on the left at t_3; let $\Delta_3 L_{DR}$ be the size of the "D"-region on the right at t_3. And let $\delta\Delta_3 L_{DL}$ and $\delta\Delta_3 L_{DR}$ be the numbers of sites in the "D"-regions on the left and right, respectively, that were deleted in conjunction with the complete deletion of the "A"-region (in $[t_3 - dt_3]$). (And we also assume that the "D"-regions suffered no other indels during $[t_3 - dt_3, t_3]$.) We also define $\Delta_3 L_D \stackrel{\text{def}}{=} \Delta_3 L_{DL} + \Delta_3 L_{DR}$ and $\delta\Delta_3 L_D \stackrel{\text{def}}{=} \delta\Delta_3 L_{DL} + \delta\Delta_3 L_{DR}$. Figure 14 illustrates the setting.

Under this setting, the contributions from the pattern (c) can be *analytically* expressed

as:

$$
\begin{aligned}
&\mu^{\text{3rd (c)}}_{P\ \text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] \times \exp\left(-\int_{t_I}^{t_F} d\tau \Delta R_X^{ID}(\Delta_I L_A, \tau)\right) \\
=&\sum_{\Delta_1 L_A=1}^{\infty}\sum_{\Delta_2 L_A=1}^{\infty}\sum_{\Delta_3 L_{DL}=1}^{\infty}\sum_{\Delta_3 L_{DR}=1}^{\infty}\sum_{\delta\Delta_3 L_{DL}=0}^{\infty}\sum_{\delta\Delta_3 L_{DR}=0}^{\infty}\int_{t_I}^{t_F-dt_1-dt_2-dt_3} dt_1 \int_{t_1+dt_1}^{t_F-dt_2-dt_3} dt_2 \int_{t_2+dt_2+dt_3}^{t_F} dt_3 \Big[\\
&\mu_{P_0}\left[(\Delta_1 L_A, t_1) \mid (\Delta_I L_A, t_I)\right](\sigma_{BE}=2) \times \\
&\times \mu_{P_0}\left[(\Delta_2 L_A, t_2) \mid (\Delta_1 L_A, t_1)\right](\sigma_{BE}=1) \times \\
&\times \left(\mu_{\text{D-cr}}\left[(x_1 = 1, [t_1, t_1+dt_1])\ ;\ (\Delta_3 L_{DL} + \delta\Delta_3 L_{DL}, t_3 - dt_3)\right]/dt_1\right) \times \\
&\times \left(\mu_{\text{D-cr}}\left[(x_2 = \Delta_2 L_{DL} + \Delta_2 L_A, [t_2, t_2+dt_2])\ ;\ (\Delta_3 L_{DR} + \delta\Delta_3 L_{DR}, t_3 - dt_3)\right]/dt_2\right) \times \\
&\times \exp\left(-\int_{t_3-dt_3}^{t_3} d\tau_3 [\Delta R_X^{ID}(\Delta_3 L_{DL} + \delta\Delta_3 L_{DL}, \tau_3) + \Delta R_X^{ID}(\Delta_3 L_{DR} + \delta\Delta_3 L_{DR}, \tau_3)]\right) \times \\
&\times \left(\mu_{\text{A-del}}\left[(\Delta_2 L_A, t_2)\ ;\ (0, -\delta\Delta_3 L_D = -(\delta\Delta_3 L_{DL} + \delta\Delta_3 L_{DR}), [t_3 - dt_3, t_3])\right](\sigma_{BE}=0)/dt_3\right) \times \\
&\times \mu_{P_0}\left[(\Delta_F L_D, t_F) \mid (\Delta_3 L_D = \Delta_3 L_{DL} + \Delta_3 L_{DR}, t_3)\right](\sigma_{BE}=2)\Big] \ . \qquad (53)
\end{aligned}
$$

As in the previous cases, we attempt to save time and memory by performing computations associated with the individual events, one after another. See appendix G for details. The series of computations can be performed with the maximum time-complexity of $O(\{N_P\}^2\{L^{CO}\}^4)$ and the maximum space-complexity of $O(N_P\{L^{CO}\}^3)$, which could be too large for a single personal computer. Fortunately, the computations can be easily cast into parallel or distributed computing, which reduces the maximum time- and space-complexities to $O(\{N_P\}^2\{L^{CO}\}^3)$ and $O(N_P\{L^{CO}\}^2)$, respectively.

5.2.3 (e) $\mathbf{A \to DA \overset{B\text{-er}}{\to} DA \to D}$

Now, we assume that the "D"-region was created at some time in $[t_1, t_1 + dt_1]$, that the "boundary-eroding" deletion occurred at some time in $[t_2 - dt_2, t_2]$, and that the "A"-region was completely deleted at some time in $[t_3 - dt_3, t_3]$, with $t_1 + dt_1 \leq t_2 - dt_2$ and $t_2 \leq t_3 - dt_3$. Here, the "base" rate operators, $\hat{Q}_0^{ID}$'s, for the "A"-region has $\sigma_{BE} = 2$ for $t < t_1$, and $\sigma_{BE} = 1$ for $t > t_1$. Let $\Delta_1 L_A$ and $\Delta_2 L_A$ be the sizes of the "A"-region at t_1 and t_2, respectively, and let $\delta\Delta_2 L_A$ be the number of sites in the "A"-region deleted by the

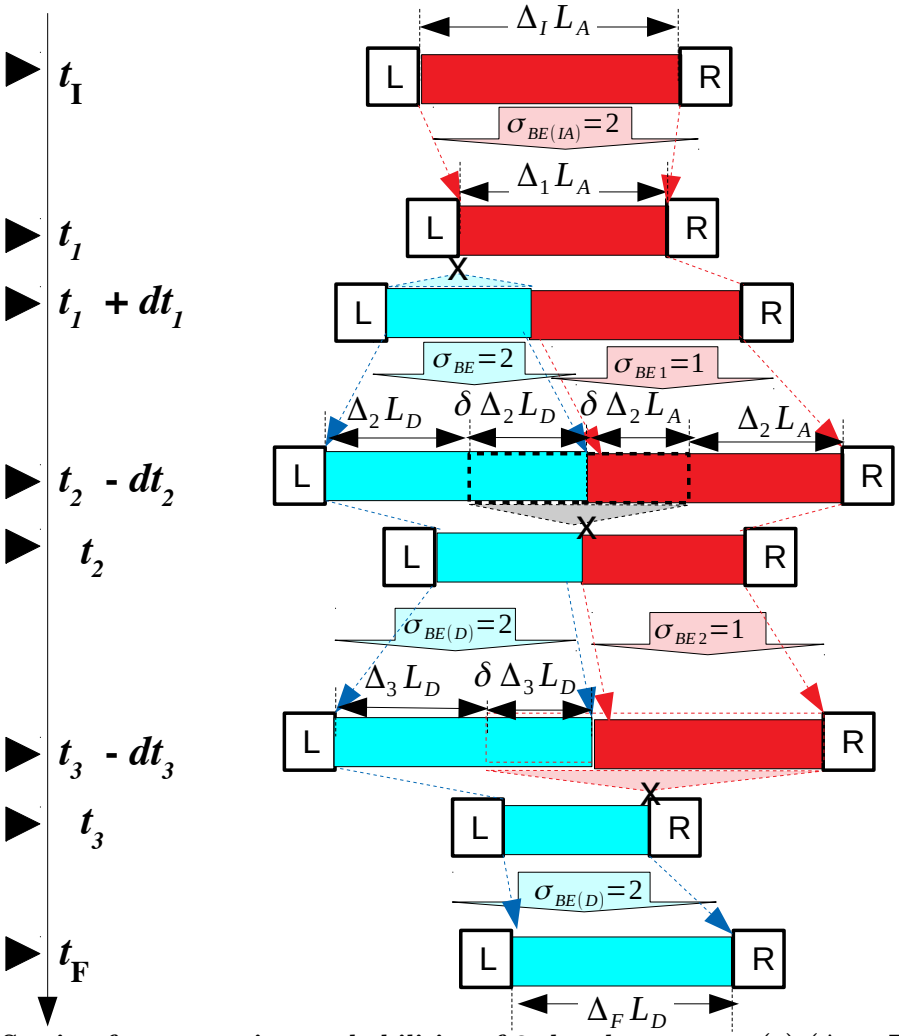

Figure 15: **Setting for computing probabilities of 3rd-order pattern (e) (A $\rightarrow$ DA $\overset{\text{B-er}}{\rightarrow}$ DA $\rightarrow$ D), in Eq.54.** This figure was created from the right panel of Figure 12, by adding further annotations. Notes similar to those on Figure 11 apply also here.

"boundary-eroding" deletion. (And we also assume that the "A"-region suffered no other indels during $[t_2 - dt_2, t_2]$.) Let $\Delta_2 L_D$ and $\Delta_3 L_D$ be the sizes of the "D"-region at t_2 and t_3, respectively, and let $\delta\Delta_2 L_D$ and $\delta\Delta_3 L_D$ be the numbers of sites in the "D"-region deleted by the "boundary-eroding" deletion (in $[t_2 - dt_2, t_2]$) and sites deleted in conjunction with the complete deletion of the "A"-region (in $[t_3 - dt_3, t_3]$), respectively. (And we also assume that the "D"-region suffered no other indels during $[t_2 - dt_2, t_2]$ and $[t_3 - dt_3, t_3]$.) Figure 15 illustrates the setting.

Under this setting, the contributions from the pattern (e) can be *analytically* expressed

as:

$$
\begin{aligned}
&\mu_{P\ \text{case-(iv)}}^{\text{3rd (e)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] \times \exp\left(-\int_{t_I}^{t_F} d\tau \Delta R_X^{ID}(\Delta_I L_A, \tau)\right) \\
=&\sum_{\Delta_1 L_A=1}^{\infty}\sum_{\Delta_2 L_A=1}^{\infty}\sum_{\Delta_2 L_D=1}^{\infty}\sum_{\delta\Delta_2 L_A=1}^{\infty}\sum_{\delta\Delta_2 L_D=1}^{\infty}\sum_{\Delta_3 L_D=1}^{\infty}\sum_{\delta\Delta_3 L_D=0}^{\infty}\int_{t_I}^{t_F-dt_1-dt_2-dt_3} dt_1 \int_{t_1+dt_1+dt_2}^{t_F-dt_3} dt_2 \int_{t_2+dt_3}^{t_F} dt_3 \Big[\\
&\mu_{P_0}\left[(\Delta_1 L_A, t_1) \mid (\Delta_I L_A, t_I)\right](\sigma_{BE(IA)}=2) \times \\
&\times \mu_{P_0}\left[(\Delta_2 L_A + \delta\Delta_2 L_A, t_2 - dt_2) \mid (\Delta_1 L_A, t_1)\right](\sigma_{BE1}=1) \times \\
&\times \left(\mu_{\text{D-cr}}\left[(x=1, [t_1, t_1+dt_1])\ ;\ (\Delta_2 L_D + \delta\Delta_2 L_D, t_2 - dt_2)\right]/dt_1\right) \times \\
&\times \exp\left\{-\int_{t_2-dt_2}^{t_2} d\tau_2 \Big[\Delta R_X^{ID}(\Delta_2 L_A + \delta\Delta_2 L_A, \tau_2) + \Delta R_X^{ID}(\Delta_2 L_D + \delta\Delta_2 L_D, \tau_2)\Big]\right\} \times \\
&\times g_D(\delta\Delta_2 L_A + \delta\Delta_2 L_D, t_2) \times \\
&\times \left(\mu_{\text{A-del}}\left[(\Delta_2 L_A, t_2)\ ;\ (0, -\delta\Delta_3 L_D, [t_3 - dt_3, t_3])\right](\sigma_{BE2}=1)/dt_3\right) \times \\
&\times \mu_{P_0}\left[(\Delta_3 L_D + \delta\Delta_3 L_D, t_3 - dt_3) \mid (\Delta_2 L_D, t_2)\right](\sigma_{BE(D)}=2) \times \\
&\times \exp\left\{-\int_{t_3-dt_3}^{t_3} d\tau_3\ \Delta R_X^{ID}(\Delta_3 L_D + \delta\Delta_3 L_D, \tau_3)\right\} \times \\
&\times \mu_{P_0}\left[(\Delta_F L_D, t_F) \mid (\Delta_3 L_D, t_3)\right](\sigma_{BE(D)}=2)\ \Big]\ . \qquad (54)
\end{aligned}
$$

As we can see, this computation could be harder than the computations in the patterns (a) and (c), because it involves summations over seven lengths, more than the six lengths for the patterns (a) and (c)!! As in the previous cases, we will follow the general strategy, and perform the computations serially. See appendix H for details. The largest computation step in this series of computations can be performed with the time-complexity of $O(\{N_P\}^2\{L^{CO}\}^4)$ and the space-complexity of $O(N_P\{L^{CO}\}^3)$, which could be too large for a single personal computer. Fortunately, the computation can be easily cast into parallel or distributed computing, and be "decomposed" into $O(\{L^{CO}\}^2)$ smaller-sized computations, each of which has the time- and space-complexities of $O(\{N_P\}^2\{L^{CO}\}^2)$ and $O(N_P\{L^{CO}\}^2)$, respectively.

5.2.4 Summary

By combining all these results obtained in the previous sub-subsections, we can obtain the total contributions from all the third-order terms, as follows:

$$\begin{aligned} & \mu_{P\ \text{case-(iv)}}^{\text{3rd}}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \\ = & \sum_{\alpha=\ \text{a,}\ \text{b,}\ \text{c,}\ \text{d,}\ \text{e,}\ \text{f}} \mu_{P\ \text{case-(iv)}}^{\text{3rd}\ (\alpha)}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \\ = & \ 2 \times \left\{ \sum_{\alpha=\ \text{a,}\ \text{c,}\ \text{e}} \mu_{P\ \text{case-(iv)}}^{\text{3rd}\ (\alpha)}\left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I)\right] \right\}. \end{aligned} \tag{55}$$

5.3 Notes on probabilities of gapped segments on boundaries

Provided that the evolution model at hand dictates how to handle indel rates on the boundary, it is not so hard to derive the formulas for the probabilities of gapped segments on the sequence boundary: in each of the formulas provided in this section, just replace one of the indel rates in the bulk with the indel rates on the boundary, when the entity in question is on the boundary. One thing that must be kept in mind is that the probabilities of horizontally symmetrical patterns (*e.g.*, the 2nd-order patterns (ii) and (iii)) are *not* equal any longer. Therefore, the summations given in sub-subsections 5.1.4 and 5.2.4 can*not* be reduced to their final forms, and you must compute the contributions from every single pattern diligently. This can double the computational time compared to that for the bulk probabilities.

In the actual sequence study, boundary indel rates can *vary greatly* depending on a number of factors, including how the aligned sequences were prepared, and whether any biological important sites or regions are on (or near) the boundaries. Therefore, *without* any clear ideas on these factors, it may be a good idea to exclude gapped segments on the boundaries from your data analyses.

6 Implementing and Validating the New "Perturbation" Method

6.1 Implementation

The new "perturbation" method proposed here was implemented into a package of prototype Perl scripts and Perl packages, named "LASTPIECE(_P)," which abbreviates "**L**ocal **A**lignment-**ST**ate **P**robability that **I**nsertion-type and d**E**letion-type gaps **C**o-**E**xist(, **P**erl-version)." The package can do the following:

1. it computes the multiplication factors of the probabilities of case-(i), (ii), (iii) and (iv) gapped segments of ancestor-descendant PWAs under a (locally) space-homogeneous *genuine* stochastic sequence evolution model;

2. currently, it explicitly incorporates the evolution model used by Dawg [48] with either power-law or geometric indel length distributions, although it can easily be modified to accept, any *genuine* evolution model (as long as it is (locally) space-homogeneous) and/or any indel length distributions;

3. Its main master script, 'lastpiece.alpha.pl', *collectively* computes the multiplication factors of the gap-configurations with 1, 2, ..., N_{UB} ancestral sites and/or with 1, 2, ..., N_{UB} descendant sites, where N_{UB} is a user-specified upper-bound, for a range of time-lapses, from near zero to a user-specified maximum value;

4. by default, it computes the multiplication factors of the case-(iv) gapped segments up to (and including) the 3rd-order perturbation level;

5. it does not only output the final results, i.e., the "practically exact" multiplication factors of case-(i), (ii) and (iii) gapped segments, as well as the factors of case-(iv) gapped segments up to 3rd-order, but it does also output the total contributions from the 2nd-order patterns (i) & (ii), and those from the 3rd-order patterns (a), (c) & (e);

6. the package also has some supplementary scripts that enable simple analyses on the computed multiplication factors, such as collectively calculating theoretically predicted frequencies of gap-configurations, collectively calculating the ratios, each of a case-(iv) multiplication factor to the corresponding product of case-(ii) and case-(iii) factors, tallying the frequencies of gap-configurations in a set of simulated ancestor-descendant PWAs, *etc.*

7. *unfortunately*, the current version computes *only* the probabilities *in the bulk*; however, it *is* possible to implement the computation of probabilities *on the boundaries* (see subsection 5.3); this is left as a future task.

It is worth noting that, once the main outputs of the main master script ("lastpiece.alpha.pl") are obtained, they can be fed *again and again* into algorithms to compute probabilities of ancestor-descendant PWAs or MSAs(, as long as the computational parameters match); such algorithms are integral parts of other program packages of ours, namely, LOLIPOG [1], ComplLiMment [30], and ANEX [56]. [30]

This package (LASEPIECE(_P)) is available as an open-source package at the FTP repository of the ANEX project in Bioinformatics.org (https://www.bioinformatics.org/ftp/pub/anex/). (Currently, the package runs on the Terminal of Mac OS X; it should run also on some other UNIX platforms, including Linux, although we have not yet confirmed that it does.)

In the current version (ver. 0.3), the main master script ("lastpiece.alpha.pl") performs all the computational steps *serially*; it thus uses only a single CPU (or core) and is considerably slow. As explained in appendixes E through H, however, most of the computational steps

[30] This strategy of pre-computing and re-using the multiplication factors has already been suggested in Additional File 1 of [1], and has actually been implemented since version 0.6 of LOLIPOG(_P) [1] and version 0.6 of ComplLiMment(_P) [30], both of which were released in 2015. We learned that a recent "simulation-based" approach to statistical PWA [53] also employs this strategy of pre-computing and re-using the probabilities of gapped segments (more precisely, "chop-zones" in their study). We do not know whether they have just borrowed our idea or *independently* come up with this strategy by themselves.

can be easily cast into parallel or distributed computing; we expect that, once this is done, and if the main scripts are also translated into C, for example, the computation could be more than 1000 times faster (if distributed to about 100 CPUs).

6.2 Validation *via in silico* experiments

In order to validate the method presented here, we simulated the evolution of 100,000 DNA sequences, each of which was 10,000 bases long initially, down each of the time-intervals of 0.2, 0.5, and 1.0 (in the unit of the expected number of substitutions per site); then, we counted how frequently each configuration of the gapped segments occur in the bulk (*i.e.*, not on either end) of the resulting ancestor-descendant PWAs; and finally, compared such "observed" frequencies [31] of the gap-configurations with their theoretically expected frequencies computed from the multiplication factors obtained by running "lastpiece.alpha.pl" described in the previous subsection. (Appendix I explains how we computed the theoretically expected frequencies from the multiplication factors.)

The simulation parameters used are: insertion rate = deletion rate = 0.1 (events/site/unit-time); upper-bound of the insertion length = upper-bound of the deletion length = 100; insertion- and deletion-length distributions are both power-law with the exponent of -1.6 (*i.e.*, $\{\text{frequency}\} \propto L^{-1.6}$, where L is the indel length).[32]

We matched these parameters with the corresponding parameters for "lastpiece.alpha.pl". The remaining important parameters for "lastpiece.alpha.pl" are set as follows: upper-bound of the number of initial or final sites for the multiplication factors to be output = 100; upper-bound of the number of sites as the arguments of multiplication factors (such as

[31] It should be kept in mind that the "observed" frequency of each configuration is a stochastic variable, which in general approximately follows a Poisson distribution. Thus, when dealing with an "observed" frequency, don't forget that it is actually fluctuating, with its standard error well-approximated by its square root. (For example, when an "observed" frequency is 100, its standard error is about 10 ($= \sqrt{100}$).)

[32] Parameters for substitutions are irrelevant, because we are interested *exclusively* in insertions/deletions here.

$\mu_{P_0}\left[(\Delta L_2, t_2) \mid (\Delta_I L_1, t_1)\right](\sigma_{BE})$) *to be computed* = 150 [33]; size of the sub-time-interval (corresponding to dt) = 0.01. On our Mac Pro (Late 2013) desktop computer (with OS version 10.11.6, with one 3.5 GHz 6-core Intel Xeon E5 Processor and 16 GB physical memory), using only a single core, it took 307 hours and 51 minutes, or 12.83 days, to finish this computation.

The full results of the numerical computation are freely available as an archive file that accompanies LASTPIECE(_P) at the Bioinformatics.org FTP repository (https://www.bioinformatics.org/ftp
Here, we only show the results at some sample points, which we believe are enough to demonstrate the accuracy of our new perturbation method.

In a previous study of ours [1], we compared the multiplication factors of case (ii) and case (iii) gapped segments at various perturbation levels to the "practically exact" solutions; in that study, however, the "practically exact" solutions themselves were *not* validated. Here, we first conduct this validation, by comparing the theoretically expected frequencies of case (ii) and case (iii) gap-configurations computed from the "practically exact" multiplication factors to the frequencies "observed" in the simulated PWAs. As shown in Table 1, the "practically exact" theoretical frequencies *do indeed* approximate the actually "observed" frequencies *extremely well* ! (Although the table here shows the results for case (ii) gapped segments only, the approximations are actually extremely good also for case (iii).)

Then, we go on to validate our main subject, *i.e.*, the multiplication factors of the case-(iv) gapped segments. Tables 2, 3, and 4 show the results with the time-lapses of 0.2, 0.5, and 1.0, respectively [34]; each table compares some observed frequencies with the corresponding theoretical expectations by only parsimonious indel histories, by parsimonious plus next-to-parsimonious indel histories, by our new perturbation method (2nd-order only), and by our

[33] This value was chosen in order to take some account of the effects of intermediate states with more than 100 sites, while keeping the computation-time from growing tremendously.

[34] These translate to 0.04, 0.1, and 0.2 indels/site, respectively, under the current setting of {total insertion rate} = {total deletion rate} = 0.1 events/unit-time.

Table 1: "Practically exact" theoretical frequencies compared to "observed" frequences: for case (ii) gap-configurations

	Time-lapse = 0.2		Time-lapse = 0.5		Time-lapse = 1.0	
n_A *	Theor.†	Obs.‡ (Ratio♭)	Theor.†	Obs.‡ (Ratio♭)	Theor.†	Obs.‡ (Ratio♭)
1	7647035	7639772 (1.001)	14545190	14520543 (1.002)	18462846	18425261 (1.002)
2	2555279	2556655 (0.999)	4954761	4956088 (1.000)	6476811	6471697 (1.001)
5	592030	592456 (0.999)	1154814	1154915 (1.000)	1524810	1523018 (1.001)
10	195150	195096 (1.000)	380309	379167 (1.003)	501694	500482 (1.002)
20	64347	63781 (1.009)	125332	124964 (1.003)	165252	164916 (1.002)
30	33672	33714 (0.999)	65690	65579 (1.002)	86816	86379 (1.005)
40	21285	20923 (1.017)	41621	41318 (1.007)	55183	54745 (1.008)
50	14914	14798 (1.008)	29218	28968 (1.009)	38844	38681 (1.004)
60	11141	11046 (1.009)	21830	21688 (1.007)	29054	28959 (1.003)
70	8679	8631 (1.006)	16950	16559 (1.024)	22528	22171 (1.016)
80	6934	6867 (1.010)	13411	13438 (0.998)	17755	17477 (1.016)
90	5548	5442 (1.019)	10491	10163 (1.032)	13866	13563 (1.022)
100	3511	3375 (1.040)	6867	6724 (1.021)	9763	9673 (1.009)

(The total indel rate is 0.2 indels/site/unit-time.)

* The number of ancestral sites.

† The theoretically expected frequency computed from the "practically exact" multiplication factor (rounded to the nearest whole number).

‡ The frequency observed in simulated PWAs.

♭ The ratio of the theoretically expected frequency to the observed frequency (rounded to the nearest thousandth).

new perturbation method (2nd-order plus 3rd-order). [35]

These tables clearly indicate that the new perturbation method provides a dramatic improvement in the accuracy, compared to our previous method based only on parsimonious (and possibly next-to-parsimonious) contributions; the accuracy improvement gets more remarkable as the numbers of sites increase. In fact, the theoretical prediction by the new perturbation method is amazingly accurate: even with only the 2nd-order terms, the prediction is more than a half of the observation up to (and including) 75 ancestral (=descendant) sites (when time-lapse = 0.5); if we incorporate the 3rd-order terms, this is the case even up to (and including) 95 sites and even with time-lapse = 1.0!! [36]

Still, accounting for just a little over $\frac{1}{2}$ of the observation may not be satisfactory for some people or some sorts of applications. As far as we can think of, there are at least three potential causes of these underestimations: (1) insufficiently fine partition of the time-interval, (2) insufficient incorporation of the effects of gapped segments containing more sites than the upper-bound for the output, and (3) lack of the 4th- or higher order terms. They are detailed in the following.

(1) In our *in silico* experiment, we used the sub-time interval of 0.01 (= Δt). This may

[35]The frequencies compared here are "double-cumulative" frequencies. The double-cumulative frequency at (x, y) is defined by the summation of the frequencies over $n_A = x, ..., n_A^{UB}$ and $n_D = y, ..., n_D^{UB}$. Here, n_A is the number of ancestral sites in the gapped segment, n_D is the number of descendant sites in the gapped segment; n_A^{UB} and n_D^{UB} are the upper-bounds of n_A and n_D, respectively. In this validation study, $n_A^{UB} = n_D^{UB} = 100$. These double-cumulative frequencies were used here because each particular case-(iv) configuration occurred only less than once in our simulated PWAs if both ancestral sites and descendant sites are many ($\sim$100).

[36]These tables (*i.e.*, tables 2, 3, and 4) may give you an impression that the theoretically expected frequencies *substantially* underestimate the observed frequencies at small values of n_{AD}. Such *ostensible* underestimates actually resulted mostly from the underestimated frequencies of larger gap-configurations. (Remember that the tables show double-cumulative frequencies.) Indeed, when comparing the expected *raw* frequencies of *individual* gap-configurations with the observed ones, the expected frequencies (*via* 2nd + 3rd order) are nearly 100% of the observed ones at $n_{AD} = 1$, and about 95% at $n_{AD} = 10$.

not have been sufficiently fine, especially when considering a large number of sites. For example, when $\Delta L = 100$ and $g_I(\tau) = g_D(\tau) = 0.1$, the increment of the exact rate is $\Delta R_X^{ID}(\Delta L,\ \tau) = (g_I(\tau)+g_D(\tau))\times\Delta L = 0.2\times 100 = 20$. Thus, $\Delta t\times\Delta R_X^{ID}(\Delta L,\ \tau) = 0.2$. This means that the region having 100 sites could suffer an indel in nearly 20% of the cases even during the smallest time-interval, Δt; this in turn means that a considerable fraction of events (or histories) could have been ignored in our numerical computation. This effect of "coarse-graining" is also indicated by the somewhat unexpectedly poor accuracy for the cases with the time-interval = 0.2, which is approximated by only 20 sub-intervals. At present, the only remedy for this problem is to use a smaller sub-time interval. Since the computational steps are at most $O(\{N_P\}^2)$, where N_P is the number of partitions (of the largest time-interval, which is 1.0 in this study), the computational time is expected to become 4 times as long if Δt is halved, and 16 times as long if Δt is quartered.

(2) In numerical computations, we always need to set an upper-bound (say, 100) of the number of sites that the subject region can contain. On the other hand, the actual evolution of sequences should *not* care about such an *artificial* upper-bound; in other words, there could be evolutionary histories in which the subject region contained more sites than the upper-bound *once(, or twice, ...)* in the middle of the evolutionary course, and in which the number of its sites finally got within the upper-bound. The contributions from such evolutionary histories are destined to be ignored, and the effects of such ignored histories are expected to be bigger as the number of (initial or final) sites approaches the upper-bound. One way to alleviate the effect of these ignored histories is to set *two kinds of* upper-bounds, one for the purpose of computation and the other for the purpose of the output. (The former is usually larger than the latter.) Actually, we took this measure: setting the upper-bound of 150 for computation and 100 for the output (shown in the tables in this paper). We also perform the computation using the upper-bound of 100 for both computation and the output (*data not shown*); the comparison of both results clearly indicated that this "enlarged upper-bound for computation" —em does work, sometimes enhancing the prediction up to

2- or 3-fold (especially when the time-lapse is large). However, it remains to be seen whether this upper-bound of 150 for computation was enough or not. For example, increasing the upper-bound to 200 may dramatically improve the accuracy further. The problem is that the site-number dependence of the computational time could be $O(\{L^{CO}\}^4)$, where L^{CO} is the upper-bound of the site number for computation. Thus, using 200 instead of 150 could more than triple the computational time.

(3) The current version of LASTPIECE(_P) incorporates up to (and including) the 3rd-order contributions. It is quite natural to expect that the accuracy will improve further if the 4th or higher-order contributions are also incorporated. We expect that the 4th-order terms will substantially increase the accuracy, because the 4th-order terms includes the evolutionary histories in which both insertions and deletions occur in the middle of the region. Insertions/deletions (indels) occurring in the middle of the region can occur at multiple alternative positions, whereas indels occurring at either end of the have only two options, at most. Thus, the number of possible histories could substantially increase if the indels occur in the middle. This is the reason why the 4th-order terms are expected to improve the accuracy substantially. We are not sure, however, how much the 5th or higher order terms will improve the accuracy.

Of the above three potential causes, (1) and (2) should be rectified relatively easily: as long as you have enough time, computer memory and storage, just changing the parameters and runring the current version of LASTPIECE_P would solve (or ease) the problem. Still, this consumes a tremendous amount of computational time, and thus translating into C and/or introducing parallel (or distributed) computing will greatly enhance the utility of the method. We recommend addressing problem (3) *only after* problems (1) and (2) are taken care of. It is possible that, as long as L^{CO} is around 100, solving problems (1) and (2) should provide enough accuracy, and that solving (3) should gain only a limited improvement. It is certain, however, incorporating higher-order terms should improve the accuracy if L^{CO} increases further.

Table 2: Comparing "observed" frequences * with various theoretical expectations: for case (iv) gap-configurations with time-lapse = 0.2 (or 0.04 indels/site)

n_{AD} *	Observed†	Theoretically expected ‡			
	—	parsimonious	parsimonious + next-to-parsimonious	new-perturbation♭, 2nd order	new-perturbation♭, 2nd + 3rd order
1	839485	584428 (0.70)♣	735068 (0.88)	759127 (0.90)	782203 (0.93)
2	389944	206152 (0.53)	304891 (0.78)	338223 (0.87)	357031 (0.92)
5	164745	56895 (0.35)	104936 (0.64)	135579 (0.82)	147156 (0.89)
10	79684	18007 (0.23)	40373 (0.51)	63333 (0.79)	70104 (0.88)
25	21078	2024 (0.096)	6213 (0.29)	15684 (0.74)	17759 (0.84)
50	3219	119 (0.037)	475 (0.15)	2328 (0.72)	2656 (0.83)
75	285	5.8 (0.020)	25 (0.088)	206 (0.72)	234 (0.82)
90	21	0.40 (0.019)	1.5 (0.071)	15 (0.71)	17 (0.81)
95	6	0.088 (0.015)	0.29 (0.048)	2.8 (0.47)	3.2 (0.53)

* The frequencies compared here are "diagonal", "double-cumulative" frequencies, which at n_{AD} is defined here as the summation over #{ancestral sites} = $n_{AD}, ..., 100$ and #{descendant sites} = $n_{AD}, ..., 100$.

† The frequency observed in simulated PWAs.

‡ The theoretically expected frequency computed from the "practically exact" multiplication factor. (Theoretically expected frequencies are written to two significant figures if they are less than 10; otherwise, they are rounded to the nearest whole number.)

♭ The new perturbation method proposed in this paper.

♣ The parenthesized number in each cell of the 3rd, .., or 6th column is the ratio of the theoretically expected frequency to the observed frequency (written to two significant figures).

Table 3: Comparing "observed" frequences * with various theoretical expectations: for case (iv) gap-configurations with time-lapse = 0.5 (or 0.1 indels/site)

n_{AD} *	Observed†	Theoretically expected ‡			
	—	parsimonious	parsimonious + next-to-parsimonious	new-perturbation♭, 2nd order	new-perturbation♭, 2nd + 3rd order
1	4115710	2114384 (0.51)♣	2932548 (0.71)	3667248 (0.89)	3940229 (0.96)
2	1983127	594118 (0.30)	1053356 (0.53)	1633121 (0.82)	1854776 (0.94)
5	862274	106693 (0.12)	259875 (0.30)	641992 (0.74)	776779 (0.90)
10	430149	20742 (0.048)	66275 (0.15)	296020 (0.69)	374083 (0.87)
25	118736	893 (0.0075)	4109 (0.035)	72998 (0.61)	96877 (0.82)
50	19752	29 (0.0015)	136 (0.0069)	11109 (0.56)	15063 (0.76)
75	2041	2.1 (0.0010)	7.7 (0.0038)	1068 (0.52)	1456 (0.71)
90	214	0.22 (0.0010)	0.74 (0.0035)	93 (0.43)	130 (0.61)
95	53	0.057 (0.0010)	0.18 (0.0034)	21 (0.40)	30 (0.57)

*, †, ‡, ♭, ♣ the same notes as in Table 2 apply.

Table 4: Comparing "observed" frequences* with various theoretical expectations: for case (iv) gap-configurations with time-lapse = 1.0 (or 0.2 indels/site)

n_{AD} *	Observed†	Theoretically expected ‡			
	—	parsimonious	parsimonious + next-to-parsimonious	new-perturbation♭, 2nd order	new-perturbation♭, 2nd + 3rd order
1	10964544	4031861 (0.37)♣	6088386 (0.56)	9172625 (0.84)	10448969 (0.95)
2	5596227	879561 (0.16)	1850172 (0.33)	4085212 (0.73)	5114311 (0.91)
5	2548574	95576 (0.038)	294546 (0.12)	1557902 (0.61)	2170348 (0.85)
10	1322899	10819 (0.0082)	43926 (0.033)	703358 (0.53)	1051858 (0.80)
25	393051	288 (0.00073)	1249 (0.0032)	172182 (0.44)	278141 (0.71)
50	72832	16 (0.00022)	58 (0.00080)	27321 (0.38)	45839 (0.63)
75	9131	1.3 (0.00014)	4.8 (0.00053)	2969 (0.33)	5077 (0.56)
90	1045	0.15 (0.00014)	0.50 (0.00048)	311 (0.30)	543 (0.52)
95	262	0.037 (0.00014)	0.12 (0.00046)	77 (0.29)	136 (0.52)

*, †, ‡, ♭, ♣ the same notes as in Table 2 apply.

6.3 Significance of *genuine* sequence evolution model

Since we confirmed that the new perturbation method enables us to compute the multiplication factors of case-(iv) gapped segments *quite accurately* already at the 3rd-order level, we can use now these multiplication factors to compare the case-(iv) multiplication factors to the product of the corresponding case-(ii) and case-(iii) multiplication factors at the same *finite* time-lapse. Previously [1], this comparison was made only in the limit where the time-lapse approaches 0 (zero).

Table 5 shows the results, in terms of the ratio of a *raw* case-(iv) factor, which was computed up to (and including) the 3rd order, to the product of practically exact case-(ii) and -(iii) factors. (The table shows only the ratios of some "diagonal" configurations, with #ancestral residues = #descendant residues = n_{AD}.) It is noteworthy that the ratio vary greatly, from around 1.5 at $n_{AD} = 1$ to over 15 at $n_{AD} = 50$; the ratio does not seem to depend so much on the time-lapse at least up to $n_{AD} = 50$. We consider that the results for $n_{AD} > 50$ are more or less due to artifacts caused by the finite upper-bound of the indel lengths (, which is 100 here); especially with a small time-lapse, the ratio is mostly determined by the overlapping indels in the 2nd-order patterns (ii) and (iii), and the number of such indels decreases as n_{AD} increases.

It should be noted that, in the simple generalized HMM of Kim and Sinha [58], this ratio should be 1 (unity), independently of the numbers of ancestral and descendant residues. Standard HMMs (*e.g.*, [37]) are also expected to give more or less similar results. Therefore, this Table 5 demonstrates how important it is to use *genuine* sequence evolution models when *accurately* computing (and comparing) the probabilities of ancestor-descendant PWAs, which in turn will provide building blocks in the computation of MSA probabilities (*e.g.*, [3, 30], which borrowed the concept of phylogenetic MSA construction from [63, 64]).

Some might argue that generalized HMMs (*e.g.*, [65]) should provide enough flexibility to incorporate such variations in the relative probabilities of gapped segments (see, *e.g.*, [66];

but also see [67] to prevent your view from being biased). Although their claim may be true *in a sense*, you *must* remember that generalized HMMs should suffer from **the "agony of the rich"**; that is, generalized HMMs should in general have a *compromised* predictive power because it can accommodate *much more* degrees of freedom *than necessary*. It should be noted that the variation in the ratio shown in Table 5 is nothing other than a consequence of the evolutionary principle, and thus it is a *theoretical prediction* that emerged *naturally* from the nearly accurate computation under the *genuine* sequence evolution model; to obtain it with a *genuine* evolution model, you don't need to *artificially* adjust any parameters. The above consideration definitely argues for the necessity of using *genuine* sequence evolution models for an accurate computation of alignment probabilities.

Of course, you *could* use some generalized HMMs with the parameters adjusted so that they will reproduce the evolutionary features as shown in Table 5. In this case, however, you *already* depended on the *genuine* sequence evolution model, and thus its indispensability will *never* be compromised at all. Besides, there may be other yet undiscovered features of the *genuine* sequence evolution model that cannot be reproduced by the aforementioned parameter-adjusted generalized HMMs. [37]

7 Discussions

As indicated by previous studies (*e.g.*, [10, 29, 30]), reconstructing a sequence alignment is a crucial yet highly error-prone step, and one of the major causes of errors is the inherent stochasticity of sequence evolution processes (*e.g.*, [31, 32, 30]). This fact necessitates the probability distribution of alignments under the *genuine* sequence evolution model, and makes it all the more crucial to compute the probabilities of sequence alignments as ac-

[37] The exceptions to these arguments are generalized HMMs that have been *directly derived* from *genuine* sequence evolution models. For example, the "long indel" model [2] and the models proposed (and used) in our previous works [3, 1, 30] fall into this category.

Table 5: Comparing case-(iv) multiplication factors * to product of case-(ii) and case-(iii) multiplication factors

n_{AD} *	Time-lapse = 0.2	Time-lapse = 0.5	Time-lapse = 1.0
1	1.58 †	1.62	1.62
2	1.94	2.03	2.09
5	3.18	3.36	3.51
10	5.26	5.62	5.88
25	10.91	11.76	12.23
50	16.22	17.34	17.87
75	13.52	14.79	16.51
90	7.58	10.13	13.96
95	5.15	8.65	13.44
100	3.39	8.63	14.39

(The total indel rate is 0.2 indels/site/unit-time.)

* The case-(iv) multiplication factors compared here are raw, "diagonal", factors, with #{ancestral sites} = #{descendant sites} = n_{AD}.

† The number in each cell is the ratio of the case-(iv) multiplication factor (2nd + 3rd order) to the product of corresponding case-(ii) and case-(iii) multiplication factors. More precisely:

$$\mu^{\text{2nd + 3rd}}_{P\ \text{case-(iv)}}\left[(n_{AD},\ t_F) \mid (n_{AD},\ t_I)\right] \Big/ \left\{\mu_{P\ \text{case-(ii)}}\left[(-,\ t_F) \mid (n_{AD},\ t_I)\right] \times \mu_{P\ \text{case-(iii)}}\left[(n_{AD},\ t_F) \mid (-,\ t_I)\right]\right\},$$

with $t_F - t_I = 0.2$, 0.5, or 1.0. Each ratio is written to the nearest hundredth.

curately as possible. Fortunately, the probability of alignments under a *genuine* sequence evolution model is factorable into the product of an overall factor and contributions from gapped segments (*e.g.*, [2]), *provided that* the model satisfies a couple of conditions [3]. Regarding ancestor-descendant PWAs, the gapped segments were classified into the four categories, namely, case-(i), -(ii), -(iii) and -(iv) segments [1]. In a previous work of ours [1], we provided a pair of methods to numerically compute *practically exact* probabilities of case-(i), -(ii) and -(iii) segments, but left the computation of practically or nearly exact probabilities of case-(iv) segments unresolved as the ***"last piece of the puzzle"***.

In this study, aiming to resolve this *"last piece of the puzzle"*, we constructed a new "perturbation" method, by reformulating the previous perturbation framework of *genuine* sequence evolution model [3, 1]. The validation analyses indicated that new "perturbation" method works considerably well, computing the probabilities of case-(iv) segments fairly accurately, already at the 3rd-order level (when the 2nd-order is the lowest), even when both the numbers of ancestral and descendant residues are near the upper-bound of 100. (See Tables 2, 3, and 4.) Now that we can compute fairly accurate probabilities of case-(iv) segments, we should be able to compute the probabilities of ancestor-descendant PWAs fairly accurately. Moreover, using the technique of stacking up ancestor-descendant PWAs along the phylogenetic tree [63, 64, 3, 30], we could also compute the probabilities of MSAs much more accurately than in the previous efforts (*e.g.*, [30]). In fact, in view of the result that the approximation by parsimonious and next-to-parsimonious indel histories are pretty poor when gaps are considerably long (Tables 2, 3, and 4), it may be better to revisit the previous analyses on MSA errors [30], using the results of this new perturbation method.

Meanwhile, the results of our analysis (Table 5) revealed an important feature of *genuine* sequence evolution that can *never* be reproduced by other, *non-evolutionary*, probabilistic models of sequence alignments, including standard HMMs, reconfirming our previous claim [1] that using *genuine* sequence evolution models is *indispensable* to the computation of *accurate* probability distribution of sequence alignments, hence to the *truthful* "reconstruction"

of sequence alignments.

Each term of the case-(iv) probability in our new perturbation method (in section 5) involves multiple time-integrations and multiple summations over intermediate states (in terms of the number of residues), and its naïve numerical computation could take practically forever even with current state-of-the art desktop computers [38] Fortunately, such multiple time-integrations combined with multiple summations could be unraveled into a series of much lighter (*i.e.*, less complex) computations, just as in the pruning algorithm for computing the likelihood of a tree given an MSA (*e.g.*, [59, 60, 61]).

We implemented the computational method up to (and including) the 3rd-order into a package of Perl scripts and modules, named "LASTPIECE(_P)," which is short for "**L**ocal **A**lignment-**ST**ate **P**robability that **I**nsertion-type and d**E**letion-type gaps **C**o-**E**xist(, **P**erl-version)." Although the current version (ver. 0.3) is fairly slow, it could potentially be speeded up 1,000-fold or more, if it is translated into a more time-efficient language, such as C, and if a few parts of particularly heavy computations are cast into parallel (or distributed) computing. And you must not forget that, once the program finishes running, the obtained results can be reused over and over again, and fed into other programs to compute *fairly accurate* probabilities of PWAs or MSAs, including our new program package, "ANEX_P" ("Alignment Neighborhood EXplorer(, Perl-version)"), that approximately computes the probability distribution of alternative MSAs under some *genuine* sequence evolution model with realistic indels [56]. [39]

We expect that extending the computation to the 4th- or 5th-order will further improve the accuracy, at some expense of computational time. Although you may be able to implement the 4th-order computations in an *ad-hoc* manner, as we did in the 2nd- and 3rd-orders,

[38] The situation may change once quantum computation technologies become feasible and widely available in the future.

[39] This strategy of pre-computing and re-using the multiplication factors is very similar to the strategy employed by a recent "simulation-based" approach to statistical PWA [53]. We are sure, however, that we have devised the strategy by ourselves. See footnote 30 for more details.

it would be a greatly painstaking job. And we expect that an *ad-hoc* implementation of the 5th-order computations would be almost impossible, unless a large number of excellent experts are involved. Thus, the key to implement such higher-order computations should be to devise a way to unravel the higher-order computation *automatically*, like the pruning algorithm for computing the tree likelihood under a given residue configuration of an MSA (*e.g.*, [59, 60, 61]). It should be remembered that the computation of 2nd- and 3rd-order terms were performed by *merging* the building blocks one after another in a bottom-up manner, and there seem to be some patterns regarding how the blocks are merged, although we are not sure whether the patterns have already been exhausted or not. In any case, implementing individual merger processes in these patterns and unraveling each term into a series of these merger patterns should be the key to the *automated* computation of higher-order terms.

The success of our new perturbation method implies that a similar method may work also for the probabilities of MSAs. We believe that the key to the success of this new perturbation method was its smart classification of indel events into "base" and "perturbation" indels. Specifically, "base" indels are indels not changing the coloring pattern, which can occur relatively frequently because of their rich positional degrees of freedom; "perturbation" indels, on the other hand, are indels changing the coloring pattern, which occur relatively rarely because of their positional constraints. When dealing with MSAs, (the topology of) the *ancestral states*, *i.e.*, sequence states at internal nodes, may be the analog of the coloring pattern for an ancestor-descendant PWA. Indeed, indel histories that keep the ancestral-state topology includes all the indels histories that keep the ancestor-descendant PWAs along individual branches. And, using the results of this study and a previous study of ours [1], we are now capable of computing nearly exact probabilities of ancestor-descendant PWAs. Therefore, by reformulating the probability of an MSA, from the summation over all indel histories that creates the gap-configuration of the MSA to the summation over all sets of ancestral states that are consistent with the MSA (exactly as the reformulation from Eqs.(R7.4&2) to

Eqs.(R7.5&6) of [3]), and by sorting the sets of ancestral states in order of increasing complexity, we may construct a "perturbation method" to compute MSA probabilities. And, as the results of this paper suggest, *even* the contributions from parsimonious (and maybe plus next-to-parsimonious) sets of ancestral states [40] *alone* may provide fairly accurate probabilities of MSAs (more precisely, probabilities of gapped segments in MSAs). For example, we previously estimated the frequencies of gapped segments in the MSAs of three sequences using *only* parsiminous indel histories, and found a small but non-negligible subset of (probably long) segments whose frequencies were substantially underestimated (Figure 6 b of [1]). Using the method suggested in this paragraph could *dramatically* reduce such underestimates of (long-)segment frequencies. (Of course, it should be examined whether these expectations are indeed the case or not ,*e.g.*, via *in silico* experiments.)

It should be noted here that the method presented in this paper is applicable to any *natural* continuous-time Markov models of insertions/deletions, with any (yet biologically meaningful) total rates and any length distributions of insertions and deletions *as long as* the rates and distributions are uniform *at least* within each gapped segment(, and *as long as* the entire sequence evolution model satisfies the factorability conditions [3]). This means that, *in principle*, [41] the method may be applied also to indel length distributions *other than* the power-law distributions (*e.g.*, [39, 40, 41, 42, 43, 44, 45]), which in a sense represent "average" behaviors, to incorporate more "genome-specific" or "region-specific" behaviors. For example, insertions of transposable elements *e.g.*, [68, 69]) may be incorporated; the simplest way to do this would be to merely add a delta-function-like spike in the insertion length distribution; a more refined way may be to allow the insertions at inter-site positions whose flanking sites show some specific "motif"s, or their approximations with runs of an-

[40] These should *not* be confused with parsimonious (and maybe plus next-to-parsimonious) indel histories, because the former can result in indel histories with *practically unlimited* numbers of indels, *albeit* following strict constraints at all internal nodes.

[41] The current version of LASTPIECE(_P) has *not* implemented this feature *yet*.

cestry indexes (assuming that the "motif"s are nearly conserved during the time-interval in question). Another example would be the incorporation of the evolution of tandem repeat arrays (like micro-satellites and mini-satellites) (*e.g.*, [70, 71, 72]). It may be harder to incorporate it, because tandem arrays show complex evolutionary patterns (*e.g.*, [70]). However, *provided that* their evolution can be well-approximated with continuous-time Markov models, and *provided that* the model allows the decomposition into the "base" and "perturbation" parts (as in subsection 4.2 of this study), the method in this paper should be applicable (maybe with some modifications).

Incidentally, the creation and deletion of "D(escendant)" and "A(ncestral)" segments in this new "perturbation" method (as seen in section 2) may be reminiscent of the "birth-death" processes of single sites in the TKF91 model [35]. In this sense, this new "perturbation" method may be considered as an "extension" of TKF91's "birth-death" approach [35] to *realistic* models of insertion/deletion, although we are not sure yet as to how rigorously this "extension" makes sense. In any case, delving further into the theoretical side of this new "perturbation" method may lead to some (analytical or numerical) methods that are much faster and/or much more accurate than the numerical computational method presented in this paper.

To summarize, the new perturbation method constructed in this study enable us to compute case-(iv) probabilities *fairly accurately*, and thus providing the *"last piece of the puzzle"* of accurately computing the probabilities of ancestor-descendant PWAs under *genuine* sequence evolution models (with any natural length distributions of insertions/deletions). Combined with our new program package, "ANEX_P," to approximately compute the probability distribution of alternative MSAs [56], this study (and its main product, "LAST-PIECE_P") will open up the possibility of the *truthful* reconstruction of sequence alignments, and thus for more accurate evolutionary analyses of homologous biological sequences.

7.1 Final Note

Some of our comments in this paper or other papers may sound like harsh criticisms on other researchers or their works. We strongly urge the readers to understand that such comments are our candid expressions of our sincere and pure hope for the advance of the science *in the right direction*, and that we have no intension to attack, harm, or hurt anybody or anybody's works. It should be kept in mind that we, all *hard-working* researchers in the world, are *not* enemies to each other but actually *comrades* to each other, who are fighting against the *common enemies*, *i.e.*, insufficient understanding of the Mother Nature and the lack of tools potent enough to uncover the essence of natural phenomena, as well as being complacent of the status quo like that. We truly hope for the future where we, all researchers, go hand-in-hand with each other to improve our understanding of the Mother Nature, by bringing together ones' own strengths under the common cause *instead of* competing against each other or even sabotaging each other's studies , and by sharing all information with each other*instead of* keeping crucial information to oneself. Then, our understanding of the Mother Nature should surely improve much faster than we've ever experienced. (If, however, there are, by any chance, *corrupt* researchers who are indulging in the complacency and/or who attempt to deform the scientific truths to their own interests, we *wlll* resolutely fight against them.)

8 Acknowledgments

The author (K.E.) greatly thanks Prof. Tetsushi Yada at Kyushu Institute of Techonology, Japan for the logistic support and encouragements during the middle third of this project, which includes this study and some others [3, 1, 30, 73, 56, 74, 75], and which was conducted first in the author's home in Yokosuka, Kanagawa, Japan, second in Kyushu Institute of

Technology, Japan, and last in the author's home in Chichibu, Saitama, Japan. He is also grateful to Prof. Dan Graur at University of Houston, TX, US and Dr. Giddy Landan at Christian-Albrechts-University of Kiel, Germany for letting me participate in their project, "Error Correction in Multiple Sequence Alignments", which was funded by US National Library of Medicine (grand number: LM010009-01 to Dan Graur and Giddy Landan, then at the University of Houston), from September 2009 till June 2011; partly inspired by their project, the author came up with this project. The author appreciates the inspiring discussions with Dr. Reed A Cartwright at Arizona State University, US and with Dr. Ian Holmes at University of California, Berkeley, US. He is also grateful to Prof. Naruya Saitou at National Institute of Genetics (NIG), Japan, and Dr. Kirill Kryukov at Tokai University, Japan for helping his interest in sequence alignment methods originate and grow, while he was studying with them at NIG. Last but not least, the author appreciates all of his family members, relatives, mentors, (ex-)friends, (ex-)supervisors, and (ex-)colleagues, for their support since his infancy, which enabled him to tread (or wander?) the scientific path and to manage to "finish" this project through all those difficulties and tough times.

The project including this study was in part supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan (grant numbers: KAKENHI Grant numbers 221S0002, 15H01358, both to Tetsushi Yada).

Appendixes

A Iteratively Solving Integral Equations for "Basic" Multiplication Factors, Eq.30

In subsection 4.2, we provided a very fast method to compute the "basic" multiplication factors (Eq.30) in a practically exact manner, by *directly* approximating the definition of the finite-time evolution operator (Eq.31). Here, we describe another approach to the computation, based on the iterative solution of the integral equations, like the practically exact computation of case-(i), (ii), and (iii) multiplication factors that we derived previously [1]. Although this approach is somewhat slower than the definition-based approach, it has an advantage that the upper-bound of the number of insertions/deletions (indels) incorporated can be controlled *totally independently* of the accuracy of numerical computation of the time integrations. Therefore, the approach provided here should be more suitable than the definition-based approach when examining, *e.g.*, how the accuracy of the computed multiplication factors depends on the maximum number of indels considered, as we did previously regarding case-(ii) and (iii) multiplication factors [1]. [42]

First, by sandwiching the integral equation,Eq.(R4.5) in a previous paper of ours [3] (rewritten for $\hat{P}_0^{ID}(i;t_k^+,t_{k+1}^-)$), with $\langle \Delta L(i;I) \mid$ and $\mid \Delta L(i;F)\rangle$, and by explicitly recording the dependence on $\sigma_{BE}(i)$ ($\stackrel{\rm def}{=} \sigma_B(i)+\sigma_E(i)$), we get:

$$
\begin{aligned}
& \mu_{P_0}\left[(\Delta L(i;F),t_{k+1}^-) \mid (\Delta L(i;I),t_k^+)\right](\sigma_{BE}(i)) \\
= \;& \delta(\Delta L(i;I),\Delta L(i;F))\exp\left\{-\int_{t_k^+}^{t_{k+1}^-} dt \Delta R_X^{ID}(\Delta L(i;I),t)\right\}
\end{aligned}
$$

[42] Another possible merit would be that the method provided here involves time integrations, which could be done via a more accurate algorithm, whereas the definition-based method involves multiple-multiplications to approximate an infinite-multiplications; (I do not know any methods that substantially improve the accuracy of the infinite-multiplications over multiple-multiplications.)

$$
\begin{aligned}
&+ \quad (\Delta L(i;I) - 1 + \sigma_{BE}(i)) \sum_{l=1}^{L_I^{CO}} \int_{t_k^+}^{t_{k+1}^-} dt \Big[\exp\left\{ - \int_{t_k^+}^{t} d\tau \Delta R_X^{ID}(\Delta L(i;I), \tau) \right\} \times \\
&\quad \times g_I(l,t) \times \mu_{P_0} \left[(\Delta L(i;F), t_{k+1}^-) \mid (\Delta L(i;I) + l, t) \right] (\sigma_{BE}(i)) \Big] \\
&+ \quad \sum_{l=1}^{\min(\Delta L(i;I)-1, L_D^{CO})} (\Delta L(i;I) - l + 1) \int_{t_k^+}^{t_{k+1}^-} dt \Big[\exp\left\{ - \int_{t_k^+}^{t} d\tau \Delta R_X^{ID}(\Delta L(i;I), \tau) \right\} \times \\
&\quad \times g_D(l,t) \times \mu_{P_0} \left[(\Delta L(i;F), t_{k+1}^-) \mid (\Delta L(i;I) - l, t) \right] (\sigma_{BE}(i)) \Big] \ . \qquad (56)
\end{aligned}
$$

Here, $\delta(\Delta L, \Delta L'^{|})$ is Kronecker's delta, which equals 1 if $\Delta L = \Delta L'^{|}$, and 0 otherwise; $g_I(l,t)$ is the rate of an insertion of length l between a given pair of contiguous sites at time t, and $g_D(l,t)$ is the rate of the deletion of a specific sub-sequence of length l at time t. Here, we also introduced the cut-off lengths, L_I^{CO} and L_D^{CO}, for insertions and deletions, respectively.

As in section SM-3 of a previous study of ours [1], the above integral equation system can be solved by iteration. The starting point is the "zero-event approximation" :

$$
\begin{aligned}
& \mu_{P_0}^{\langle 0 \rangle} \left[(\Delta L_F, t_{k+1}^-) \mid (\Delta L_t, t) \right] (\sigma_{BE}) \\
= \quad & \delta(\Delta L_t, \Delta L_F) \exp\left\{ - \int_t^{t_{k+1}^-} d\tau \Delta R_X^{ID}(\Delta L_t, \tau) \right\} \ . \qquad (57)
\end{aligned}
$$

Here and below, we use the "short-hand" notations, ΔL_F for $\Delta L(i;F)$ ΔL_t for $\Delta L(i;t)$, and σ_{BE} for $\sigma_{BE}(i)$, purely for clarity. And let $\mu_{P_0}^{\langle n_S \rangle} \left[(\Delta L_F, t_k^+) \mid (\Delta L_t, t) \right]$ be the approximation at the n_S th step, which includes all the responsible indel histories (in the "base" rate operator) with n_S or less indels. Then, by solving the following recursion relation, we can improve the approximation one by one, to the accuracy level we desire (provided that we have enough time to do so):

$$
\begin{aligned}
& \mu_{P_0}^{\langle n_S \rangle} \left[(\Delta L_F, t_{k+1}^-) \mid (\Delta L_t, t) \right] (\sigma_{BE}) \\
= \quad & \delta(\Delta L_t, \Delta L_F) \exp\left\{ - \int_t^{t_{k+1}^-} d\tau \Delta R_X^{ID}(\Delta L_t, \tau) \right\} \\
+ \quad & (\Delta L_t - 1 + \sigma_{BE}) \sum_{l=1}^{L_I^{CO}} \int_t^{t_{k+1}^-} dt' \Big[\exp\left\{ - \int_t^{t'} d\tau \Delta R_X^{ID}(\Delta L_t, \tau) \right\} \times \\
& \times g_I(l,t') \times \mu_{P_0}^{\langle n_S - 1 \rangle} \left[(\Delta L_F, t_{k+1}^-) \mid (\Delta L_t + l, t') \right] (\sigma_{BE}) \Big]
\end{aligned}
$$

$$+ \sum_{l=1}^{\min(\Delta L_t-1, L_D^{CO})} (\Delta L_t - l + 1) \int_t^{t_{k+1}^-} dt' \Big[\exp\left\{ - \int_t^{t'} d\tau \Delta R_X^{ID}(\Delta L_t, \tau) \right\} \times$$
$$\times g_D(l, t') \times \mu_{P_0}^{\langle n_S - 1 \rangle} \left[(\Delta L_F, t_{k+1}^-) \mid (\Delta L_t - l, t') \right] (\sigma_{BE}) \Big] \ . \quad (58)$$

As in section SM-3 of a previous work of ours [1], the above recursion relation could be rewritten as:

$$\begin{aligned} & \mu_{P_0}^{\langle n_S \rangle} \left[(\Delta L_F, t_{k+1}^-) \mid (\Delta L_t, t) \right] (\sigma_{BE}) \\ = \ & \delta(\Delta L_t, \Delta L_F) \exp\left\{ - \int_t^{t_{k+1}^-} d\tau \Delta R_X^{ID}(\Delta L_t, \tau) \right\} \\ + \ & \int_t^{t_{k+1}^-} dt' \Big[\exp\left\{ - \int_t^{t'} d\tau \Delta R_X^{ID}(\Delta L_t, \tau) \right\} \times \\ & \times \Phi_{\mu\ 0}^{\langle n_S \rangle} \left[(\Delta L_F, t_{k+1}^-) \ ; \ (\Delta L_t, t') \right] (\sigma_{BE}) \Big] \ , \end{aligned} \quad (59)$$

with the "auxiliary function",

$$\begin{aligned} & \Phi_{\mu\ 0}^{\langle n_S \rangle} \left[(\Delta L_F, t_{k+1}^-) \ ; \ (\Delta L_t, t') \right] (\sigma_{BE}) \\ \stackrel{\text{def}}{=} \ & (\Delta L_t - 1 + \sigma_{BE}) \sum_{l=1}^{L_I^{CO}} \Big[g_I(l, t') \times \mu_{P_0}^{\langle n_S - 1 \rangle} \left[(\Delta L_F, t_{k+1}^-) \mid (\Delta L_t + l, t') \right] (\sigma_{BE}) \Big] \\ + \ & \sum_{l=1}^{\min(\Delta L_t-1, L_D^{CO})} \Big[(\Delta L_t - l + 1) g_D(l, t') \times \mu_{P_0}^{\langle n_S - 1 \rangle} \left[(\Delta L_F, t_{k+1}^-) \mid (\Delta L_t - l, t') \right] (\sigma_{BE}) \Big] \ . \end{aligned} \quad (60)$$

As an additional note, although we have formally dealt with $\Delta R_X^{ID}(\Delta L, t')$ thus far, it is actually a linear function:

$$\begin{aligned} & \Delta R_X^{ID}(\Delta L, t') = G^{ID}[L_I^{CO}, L_D^{CO}, t'] \times \Delta L \\ & \Big(\text{ with } \qquad G^{ID}[L_I^{CO}, L_D^{CO}, t'] \stackrel{\text{def}}{=} \sum_{l=1}^{L_I^{CO}} g_I(l, t') + \sum_{l=1}^{L_D^{CO}} g_D(l, t') \ \Big) \ , \end{aligned} \quad (61)$$

under the locally space-homogeneous model. Knowing this fact will considerably speed up the computation.

For a fixed ΔL_F, the above system of recursion equations, Eq.59 and Eq.60 with ranging ΔL_t and t (and t'), could be numerically solved via an algorithm of time-complexity

$O(N_{ID}L^{CO}(L^{CO}+N_P)N_P)$ and space-complexity $O(L^{CO}N_P)$, just as in SM-3 of [1]. Here, L^{CO} is the upper-bound of ΔL_t (and ΔL_F), N_P is the number of sub-intervals into which the time-interval (t_k^+, t_{k+1}^-) (or rather $[t_I, t_F]$) is partitioned, and N_{ID} is the maximum number of indels that we take account of. Because the recursion equation systems with different ΔL_F's can be solved independently of each other, the computation could easily be parallelized (or rather, cast into distributed computing). And each computation should take as much time as in the case of isolated gaps (i.e., Eqs.(SM-3.4a,b) in [1]). Therefore, when L^{CO} is considerably large, say, 1000, although it may take too much time to numerically solve this recursion equation system with a single CPU (or core), the computation could finish typically in a few hours if you can access, e.g., hundreds of CPUs simultaneously. Moreover, once computed, the results could be stored and re-used for various analyses under the same evolution model. In this sense, this computation itself could be feasible.

By substituting the solutions of the recursion equation system, Eq.59 and Eq.60, into Eq.29, we can obtain the *practically exact* "zero-th approximation" of the transition probabilities within each slice.

B Backward-Extension to Directly Approximate Definition of Finite-Time Evolution Operator, Eq.31

In subsection 4.2, we *directly* approximated the definition of the finite-time evolution operator (Eq.31) *forward*, i.e., from the initial time (t_I) to the final time (t_F).

Alternatively, we could extend the time-interval backward, from $[t_F, t_F]$. In this case, we obtain the following recursion relation:

$$\begin{aligned} & \mu_{P_0}^{[N_P]}\left[(\Delta L_F, t_F) \mid (\Delta L_j, t_j)\right](\sigma_{BE}(i)) \\ = \; & \left(1 - \Delta_{N_P} t \cdot \Delta R_X^{ID}(\Delta L_j, \bar{t}_{j+1})\right) \cdot \mu_{P_0}^{[N_P]}\left[(\Delta L_F, t_F) \mid (\Delta L_j, t_{j+1})\right](\sigma_{BE}(i)) \end{aligned}$$

$$+\Delta_{N_P} t \cdot \Bigg[\; (\Delta L_j - 1 + \sigma_{BE}(i)) \times \sum_{l=1}^{\min(L_I^{CO},\; L^{CO}-\Delta L_j)} \Big\{ \; g_I(l, \bar{t}_{j+1}) \times$$

$$\times \mu_{P_0}^{[N_P]} \left[(\Delta L_F, t_F) \mid (\Delta L_j + l, t_{j+1}) \right] (\sigma_{BE}(i)) \; \Big\}$$

$$+ \sum_{l=1}^{\min(L_D^{CO},\; \Delta L_j - 1)} \Big\{ \; (\Delta L_j - l + 1) \times g_D(l, \bar{t}_{j+1}) \times$$

$$\times \mu_{P_0}^{[N_P]} \left[(\Delta L_F, t_F) \mid (\Delta L_j - l, t_{j+1}) \right] (\sigma_{BE}(i)) \; \Big\} \; \Bigg] \; , \tag{62}$$

with the "initial" condition:

$$\mu_{P_0}^{[N_P]} \left[(\Delta L_F, t_F) \mid (\Delta L_{N_P}, t_{N_P}) \right] (\sigma_{BE}(i)) = \delta(\Delta L_F, \Delta L_{N_P}) \; .$$

This recursion relation is valid for $\Delta L_j \; = 1, ..., L^{CO}$, and $\Delta L_F \; = 1, ..., L^{CO}$, for reasons similar to those given below Eq. 36 (in subsection 4.2).

C Iteratively Solving Integral Equation for Finite-Time Transition Probabilities when "Boundary-Eroding" Deletions are Switched on

Here, let us *formally* attempt to compute the finite-time transition probabilities when the "boundary-eroding" rate operator, $\hat{Q}^D_{M:\text{B-er}}(t)$ (in Eq.21), is switched on.

In Eq.24, $\hat{Q}^D_{M:\text{B-er}}(t)$ was expanded into the summation of operators, $\hat{Q}^D_{M:\text{B-er}}(i;t)$'s, acting only on individual region-boundaries. Thus, we will *define* a series of "perturbed" rate operators:

$$\begin{aligned} \hat{Q}^{ID(0)}_{\text{B-er}}(t) &\stackrel{\text{def}}{=} \hat{Q}^{ID}_0(t) \; , \\ \hat{Q}^{ID(i)}_{\text{B-er}}(t) &\stackrel{\text{def}}{=} \hat{Q}^{ID(i-1)}_{\text{B-er}}(t) + \hat{Q}^D_{M:\text{B-er}}(i;t) \quad (i = 1, ..., N_C(k) - 1). \end{aligned} \tag{63}$$

Then, the "perturbed" rate operator, $\hat{Q}^{ID(N_C(k)-1)}_{\text{B-er}}(t) = \hat{Q}^{ID}_0(t) + \hat{Q}^D_{M:\text{B-er}}(t)$, incorporates all "boundary-eroding" deletions. Using the above series of "perturbed" rate operators, and fol-

lowing the same procedure as when obtaining the fundamental integral equation, Eq.(R4.5) in [3], we can derive a series of integral equations for $\hat{P}^{ID(i)}_{\text{B-er}}(t,t') \stackrel{\text{def}}{=} T\left\{\exp\left[\int_t^{t'} d\tau\ \hat{Q}^{ID(i)}_{\text{B-er}}(\tau)\right]\right\}$:

$$\hat{P}^{ID(i)}_{\text{B-er}}(t,t') = \hat{P}^{ID(i-1)}_{\text{B-er}}(t,t') + \int_t^{t'} d\tau\ \hat{P}^{ID(i-1)}_{\text{B-er}}(t,\tau)\ \hat{Q}^{D}_{M:\text{B-er}}(i;\tau)\ \hat{P}^{ID(i)}_{\text{B-er}}(\tau,t')\ , \tag{64}$$

with $i = 1, ..., N_C(k) - 1$. Each of these integral equations could be solved by iteration, as we solved the integral equation, Eq.56, above. Because the perturbation term, $\hat{Q}^{D}_{M:\text{B-er}}(i;\tau)$, in each of the above integral equations acts only on two contiguous colored regions, we need to track only changes in the lengths of the two regions; this may significantly reduce the working space-complexity. At least theoretically, if you will, you could directly solve the "one-shot" integral equation:

$$\hat{P}^{ID}_{\text{B-er}}(t,t') = \hat{P}^{ID}_{0}(t,t') + \int_t^{t'} d\tau\ \hat{P}^{ID}_{0}(t,\tau)\ \hat{Q}^{D}_{M:\text{B-er}}(\tau)\ \hat{P}^{ID}_{\text{B-er}}(\tau,t')\ , \tag{65}$$

where we set $\hat{P}^{ID}_{\text{B-er}}(t,t') \stackrel{\text{def}}{=} \hat{P}^{ID(N_C(k)-1)}_{\text{B-er}}(\tau,t')$ $(= T\left\{\exp\left[\int_t^{t'} d\tau\ \left(\hat{Q}^{ID}_{0}(\tau) + \hat{Q}^{D}_{M:\text{B-er}}(\tau)\right)\right]\right\})$. In practice, a possibly big problem is that solving this equation could require a tremendous amount of memory, because you need to keep track of changes in the lengths of all colored regions. Provided that you have adequate resources to do that, of course, solving Eq.65 alone may be preferable to solving Eq.64's serially. In any case, Eq.64 and Eq.65 are identical if the slice in question consists of only two colored regions, and they do not even exist if the slice has only one region.

D Time-Efficient Computation of Contributions from 2nd-order Pattern (i): A $\rightarrow$ "$\emptyset$" $\rightarrow$ D

The contributions from the 2nd-order pattern (i), Eq. 49, were given in subsection 5.1.1. As argued at the bottom of subsection 4.5, it would save time to perform the calculation associated with individual events one after another.

In the case at hand, this could be done in two ways. The first method performs the integration over $(t_I + dt_1 <) t_1 (< t_2)$ first, and then integrates over $(t_I + dt_1 <) t_2 (< t_F - dt_2)$. It can be written as:

$$\begin{aligned} & \mu_{P \text{ case-(iv)}}^{\text{2nd (i)}} \left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I) \right] \times \exp\left(- \int_{t_I}^{t_F} d\tau \Delta R_X^{ID}(\Delta_I L_A, \tau) \right) \\ = & \int_{t_I + dt_1}^{t_F - dt_2} dt_2 \Big[\ \mu_{\text{A} \rightarrow \text{"}\emptyset\text{"}} \left[(\Delta_I L_A, t_I)\ ;\ (0,0)\ ;\ t_2 \right] (\sigma_{BE} = 2) \times \\ & \times \left(\mu_{\text{D-cr}} \left[(x = 1, [t_2, t_2 + dt_2])\ ;\ (\Delta_F L_D, t_F) \right] / dt_2 \right) \Big] \ , \end{aligned} \tag{66}$$

with the "extended" multiplication factor, $\mu_{\text{A} \rightarrow \text{"}\emptyset\text{"}}[\cdots]$, defined as:

$$\begin{aligned} & \mu_{\text{A} \rightarrow \text{"}\emptyset\text{"}} \left[(\Delta_I L_A, t_I)\ ;\ (0,0)\ ;\ t_2 \right] (\sigma_{BE} = 2) \\ \overset{\text{def}}{=} & \int_{t_I + dt_1}^{t_2} dt_1 \Big[\ (\mu_{\text{A-del}} \left[(\Delta_I L_A, t_I)\ ;\ (0,0,[t_1 - dt_1, t_1]) \right] (\sigma_{BE} = 2) / dt_1) \times \\ & \times \mu_{P \text{ case-(i)}} \left[(t_2, 0) \mid (t_1, 0) \right] \Big] \ . \end{aligned} \tag{67}$$

By (numerically) pre-computing Eq.67 with $t_2 \in [t_I + dt_1, t_F - dt_2]$ and with $\Delta_I L_A = 1, 2, ..., L^{CO}$, and storing the results, we can avoid the repeated computation, which could occur if we directly compute Eq.49.

The second method performs the integration over $(t_1 <) t_2 (< t_F - dt_2)$ first, and then integrates over $(t_I + dt_1 <) t_1 (< t_F - dt_2)$. We have:

$$\begin{aligned} & \mu_{P \text{ case-(iv)}}^{\text{2nd (i)}} \left[(\Delta_F L_D,\ t_F) \mid (\Delta_I L_A,\ t_I) \right] \times \exp\left(- \int_{t_I}^{t_F} d\tau \Delta R_X^{ID}(\Delta_I L_A, \tau) \right) \\ = & \int_{t_I + dt_1}^{t_F - dt_2} dt_1 \Big[\ (\mu_{\text{A-del}} \left[(\Delta_I L_A, t_I)\ ;\ (0,0,[t_1 - dt_1, t_1]) \right] (\sigma_{BE} = 2) / dt_1) \times \\ & \times \mu_{\text{"}\emptyset\text{"} \rightarrow \text{D}} \left[t_1\ ;\ x = 1\ ;\ (\Delta_F L_D, t_F) \right] \Big] \ , \end{aligned} \tag{68}$$

with the definition:

$$\begin{aligned} & \mu_{\text{"}\emptyset\text{"} \rightarrow \text{D}} \left[t_1\ ;\ x = 1\ ;\ (\Delta_F L_D, t_F) \right] \\ \overset{\text{def}}{=} & \int_{t_1}^{t_F - dt_2} dt_2 \Big[\ \mu_{P \text{ case-(i)}} \left[(t_2, 0) \mid (t_1, 0) \right] \times \\ & \times \left(\mu_{\text{D-cr}} \left[(x, [t_2, t_2 + dt_2])\ ;\ (\Delta_F L_D, t_F) \right] / dt_2 \right) \Big] \ . \end{aligned} \tag{69}$$

Similarly to the 1st method, the pre-computation of Eq.69 will avoid the repeated computation encountered in the naïve (numerical) computation of Eq.49.

E Time-Efficient Computation of Contributions from 2nd-order Pattern (ii): A $\rightarrow$ AD $\rightarrow$ D

The contributions from the 2nd-order pattern (iI), Eq. 50, were given in subsection 5.1.2. As usual, it should be time-efficient to perform the computations associated with individual events separately, one after another.

In this case, such "computational factorization" is done as follows. First, we (analytically) *define* the following "complex" multiplication factor (of $O(dt_1 \cdot dt_2)$):

$$
\begin{aligned}
& \mu_{\text{D-cr} \rightarrow \text{A-del}} \left[(\Delta L_1, x, [t_1, t_1 + dt_1]) \ ; \ (\Delta L_2, [t_2 - dt_2, t_2]) \right] (\sigma_{BE} = 1) \\
\stackrel{\text{def}}{=} & \sum_{\delta \Delta L_2 = 0}^{\infty} \Big[\mu_{\text{D-cr}} \left[(x = \Delta L_1, [t_1, t_1 + dt_1]) \ ; \ (\Delta L_2 + \delta \Delta L_2, t_2 - dt_2) \right] \times \\
& \times \exp \left(- \int_{t_2 - dt_2}^{t_2} d\tau_2 \ \Delta R_X^{ID} (\Delta L_2 + \delta \Delta L_2, \tau_2) \right) \times \\
& \times \mu_{\text{A-del}} \left[(\Delta L_1, t_1) \ ; \ (0, -\delta \Delta L_2, [t_2 - dt_2, t_2]) \right] (\sigma_{BE} = 1) \Big] \ , \qquad (70)
\end{aligned}
$$

which, after being divided by $dt_1 \cdot dt_2$, gives the (finite) probability density that a "D"-region was created between x and $x + 1$(, which is on either end, in any case,) of an "A"-region of size ΔL_1, immediately after t_1, and that the "A"-region (evolved with $\sigma_{BE} = 1$) was completely deleted (immediately before t_2), leaving the "D"-region of size ΔL_2 at time t_2. When dealing with a time-homogeneous model, the collection of these multiplication factors with ranging ΔL_1, ΔL_2 and $t_2 - t_1$ can be numerically computed with the time-complexity of $O(N_P \{L^{CO}\}^3)$, because the summation, $\sum_{\delta \Delta L_2 = 0}^{\infty}$ should be of $O(L^{CO})$ time-complexity (in numerical computation). Regarding the space-complexity, storing the inputs, $\mu_{\text{D-cr}}[\cdots]$ and $\mu_{\text{A-del}}[\cdots]$, requires the memories of $O(N_P L^{CO})$ and $O(N_P \{L^{CO}\}^2)$, respectively, and storing the output, $\mu_{\text{D-cr} \rightarrow \text{A-del}}[\cdots]$, also requires the memory of $O(N_P \{L^{CO}\}^2)$. Thus, the

space-complexity for this definition is $O(N_P\{L^{CO}\}^2)$.

As Figure 11 indicates, the aforementioned $\mu_{\text{D-cr}\,\rightarrow\,\text{A-del}}[\cdots]$ covers the "amputated" portion of the coloring-pattern evolution, *after* the initial period of a single "A" and *before* the final period of a single "D". We will express this fact with the equation:

$$\begin{aligned}
&\left(\mu_{P\ \text{case-(iv)}}^{\text{2nd (ii)}}\right)_{\text{Amp.}}\left[(\Delta L_2,\ t_2)\mid(\Delta L_1,\ t_1)\right]\times(dt_1\cdot dt_2)\\
\overset{\text{def}}{=}\ &\mu_{\text{D-cr}\,\rightarrow\,\text{A-del}}\left[(\Delta L_1,x,[t_1,t_1+dt_1])\ ;\ (\Delta L_2,[t_2-dt_2,t_2])\right](\sigma_{BE}=1).
\end{aligned}\tag{71}$$

Here, the subscript, "Amp.", on the left-hand side stands for "amputated". Now, using this "amputated" multiplication factor, Eq.50 is rewritten as:

$$\begin{aligned}
&\mu_{P\ \text{case-(iv)}}^{\text{2nd (ii)}}\left[(\Delta_F L_D,\ t_F)\mid(\Delta_I L_A,\ t_I)\right]\times\exp\left(-\int_{t_I}^{t_F}d\tau\Delta R_X^{ID}(\Delta_I L_A,\tau)\right)\\
=\ &\sum_{\Delta L_1=1}^{\infty}\sum_{\Delta L_2=1}^{\infty}\int_{t_I}^{t_F-dt_1-dt_2}dt_1\int_{t_1+dt_1+dt_2}^{t_F}dt_2\Big[\mu_{P_0}\left[(\Delta L_1,t_1)\mid(\Delta_I L_A,t_I)\right](\sigma_{BE}=2)\times\\
&\times\left(\mu_{P\ \text{case-(iv)}}^{\text{2nd (ii)}}\right)_{\text{Amp.}}\left[(\Delta L_2,\ t_2)\mid(\Delta L_1,\ t_1)\right]\times\\
&\times\mu_{P_0}\left[(\Delta_F L_D,t_F)\mid(\Delta L_2,t_2)\right](\sigma_{BE}=2)\Big]\ .
\end{aligned}\tag{72}$$

This equation tells us that the full multiplication factors for the 2nd-order evolution pattern (ii) can be obtained by "extending" their "amputated" version, *both* to the initial time and to the final time, each via "convolution" with $\mu_{P_0}[\cdots](\sigma_{BE}=2)$.

Eq.72 can be made further time- and space-efficient by separately performing the computations associated with $(\Delta L_1,t_1)$ and $(\Delta L_2,t_2)$. First, we (analytically) *define* the following "extended" version of the "amputated" factor:

$$\begin{aligned}
&\left(\mu_{P\ \text{case-(iv)}}^{\text{2nd (ii)}}\right)_{\text{Amp.\&Ext.-I}}\left[(\Delta L_2,\ t_2)\mid(\Delta_I L_A,\ t_I)\right]\\
\overset{\text{def}}{=}\ &\sum_{\Delta L_1=1}^{\infty}\int_{t_I}^{t_2-dt_1-dt_2}dt_1\Big[\mu_{P_0}\left[(\Delta L_1,t_1)\mid(\Delta_I L_A,t_I)\right](\sigma_{BE}=2)\times\\
&\times\left(\mu_{P\ \text{case-(iv)}}^{\text{2nd (ii)}}\right)_{\text{Amp.}}\left[(\Delta L_2,\ t_2)\mid(\Delta L_1,\ t_1)\right]\Big]\ .
\end{aligned}\tag{73}$$

Here, the subscript, "Amp.&Ext.-I", stands for "amputated and extended to the initial time

(*i.e.*, t_I)". Using this "extended" version, Eq.72 can be rewritten as:

$$\begin{aligned}
&\mu^{\text{2nd (ii)}}_{P\ \text{case-(iv)}}\left[(\Delta_F L_D,\ t_F)\mid(\Delta_I L_A,\ t_I)\right]\times\exp\left(-\int_{t_I}^{t_F}d\tau\Delta R^{ID}_X(\Delta_I L_A,\tau)\right)\\
=&\sum_{\Delta L_2=1}^{\infty}\int_{t_I+dt_1+dt_2}^{t_F}dt_2\Big[\left(\mu^{\text{2nd (ii)}}_{P\ \text{case-(iv)}}\right)_{\text{Amp.\&Ext.-I}}\left[(\Delta L_2,\ t_2)\mid(\Delta_I L_A,\ t_I)\right]\times\\
&\times\mu_{P_0}\left[(\Delta_F L_D,t_F)\mid(\Delta L_2,t_2)\right](\sigma_{BE}=2)\Big]\ .
\end{aligned}\tag{74}$$

The numerical computation of the definition, Eq.73, with ranging $\Delta_I L_A$, ΔL_2, and t_2-t_I, requires the space-complexity of $O(N_P\{L^{CO}\}^2)$ and the time-complexity of $O(\{N_P\}^2\{L^{CO}\}^3)$. And so does the numerical computation of Eq.74 with ranging $\Delta_I L_A$, $\Delta_F L_D$ and t_F-t_I. The $O(\{N_P\}^2\{L^{CO}\}^3)$ time-complexity may be relatively time-consuming. Once the input factors become available, however, the computation for different combinations of the region-lengths, $(\Delta_I L_A,\ \Delta_F L_D)$, can be done independently, and thus this computation can easily be parallelized (or cast into distributed computing). (For example, if these computations with different $\Delta_I L_A$'s are distributed, time-complexity of each computation reduces to $O(\{N_P\}^2\{L^{CO}\}^2)$, while its space-complexity remaining the same.) Thus, you may manage to numerically compute these contributions from the pattern (ii).

As in appendix D, the order of the time-integrations may be reversed, integrating over $(t_1+dt_1+dt_2<)t_2(<t_F)$ first and then integrating over $(t_I<)t_1(<t_F-dt_1-dt_2)$. In this case, we (analytically) *define* the "amputated-and-extended" factor:

$$\begin{aligned}
&\left(\mu^{\text{2nd (ii)}}_{P\ \text{case-(iv)}}\right)_{\text{Amp.\&Ext.-F}}\left[(\Delta_F L_D,t_F)\mid(\Delta L_1,\ t_1)\right]\\
\stackrel{\text{def}}{=}&\sum_{\Delta L_2=1}^{\infty}\int_{t_1+dt_1+dt_2}^{t_F}dt_2\Big[\left(\mu^{\text{2nd (ii)}}_{P\ \text{case-(iv)}}\right)_{\text{Amp.}}\left[(\Delta L_2,\ t_2)\mid(\Delta L_1,\ t_1)\right]\times\\
&\times\mu_{P_0}\left[(\Delta_F L_D,t_F)\mid(\Delta L_2,t_2)\right](\sigma_{BE}=2)\Big]\ .
\end{aligned}\tag{75}$$

Here, the subscript, "Amp.&Ext.-F", stands for "amputated and extended to the final time (*i.e.*, t_F)". Using this, Eq.72 can be rewritten as:

$$\mu^{\text{2nd (ii)}}_{P\ \text{case-(iv)}}\left[(\Delta_F L_D,\ t_F)\mid(\Delta_I L_A,\ t_I)\right]\times\exp\left(-\int_{t_I}^{t_F}d\tau\Delta R^{ID}_X(\Delta_I L_A,\tau)\right)$$

$$= \sum_{\Delta L_1=1}^{\infty} \int_{t_I}^{t_F - dt_1 - dt_2} dt_1 \Big[\mu_{P_0} \left[(\Delta L_1, t_1) \mid (\Delta_I L_A, t_I) \right] (\sigma_{BE} = 2) \times$$
$$\times \left(\mu_{P \text{ case-(iv)}}^{\text{2nd (ii)}} \right)_{\text{Amp.\&Ext.-F}} \left[(\Delta_F L_D, t_F) \mid (\Delta L_1,\ t_1) \right] \Big] \ . \tag{76}$$

The numerical computations resulting from this reverse-order procedure also have the time-complexity of $O(\{N_P\}^2\{L^{CO}\}^3)$ and the space-complexity of $O(N_P\{L^{CO}\}^2)$, and can easily be "parallelized", *e.g.*, by distributing the computations with different $\Delta_F L_D$'s.

Here, for later use, we give an "alias" to $\left(\mu_{P \text{ case-(iv)}}^{\text{2nd (ii)}} \right)_{\text{Amp.\&Ext.-F}} [\cdots]$, as follows:

$$\mu_{\text{D-cr} \to \text{AD} \to \text{D}} \left[(\Delta L_1, x, [t_1, t_1 + dt_1]) \ ; \ () \ ; \ (\Delta_F L_D, t_F) \right] (\sigma_{BE} = 1)$$
$$\stackrel{\text{def}}{=} \left(\mu_{P \text{ case-(iv)}}^{\text{2nd (ii)}} \right)_{\text{Amp.\&Ext.-F}} \left[(\Delta_F L_D, t_F) \mid (\Delta L_1,\ t_1) \right] \times dt_1 \ . \tag{77}$$

This will be used as an ingredient for the 3rd-order pattern (a).

F Time-Efficient Computation of Contributions from 3rd-order Pattern (a): A $\to$ ADA $\to$ AD $\to$ D

The contributions from the 3rd-order pattern (a), Eq. 52, were given in subsection 5.2.1. These contributions can be computed in a time-efficient manner, as in appendix E.

First, we observe that, during the time interval, $[t_1, t_3 - dt_3]$, if we ignore the left-fragment of the "A"-region, the remaining portion (consisting of the "D"-region and the right-fragment of the "A"-region) is identical to those defining the factor, $\mu_{\text{D-cr} \to \text{AD} \to \text{D}}[\cdots]$ defined in Eq.77 (, which is $O(dt_1)$ here,) with some modifications on the arguments. Thus, we can formally perform the time-integration over $(t_1 + dt_1 + dt_2 <) t_2 (< t_3 - dt_3)$ and the summations over $\delta\Delta L_2$ and ΔL_2, to get:

$$\mu_{P \text{ case-(iv)}}^{\text{3rd (a)}} \left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I) \right] \times \exp\left(- \int_{t_I}^{t_F} d\tau \Delta R_X^{ID}(\Delta_I L_A, \tau) \right)$$
$$= \sum_{\Delta_L L_1=1}^{\infty} \sum_{\Delta_R L_1=1}^{\infty} \sum_{\Delta L_3=1}^{\infty} \sum_{\delta\Delta L_3=0}^{\infty} \int_{t_I}^{t_F - dt_1 - dt_2 - dt_3} dt_1 \int_{t_1 + dt_1 + dt_2 + dt_3}^{t_F} dt_3 \Big[$$

$$
\begin{aligned}
&\mu_{P_0}\left[(\Delta L_1 = \Delta_L L_1 + \Delta_R L_1, t_1) \mid (\Delta_I L_A, t_I)\right](\sigma_{BE} = 2) \times \\
&\times \left(\mu_{\text{D-cr}\,\to\,\text{AD}\,\to\,\text{D}}\left[(\Delta_R L_1, x = \Delta_L L_1, [t_1, t_1 + dt_1])\ ;\ ()\ ;\ (\Delta L_3 + \delta\Delta L_3, t_3 - dt_3)\right](\sigma_{BE:R} = 1)/dt_1\right) \times \\
&\quad \times \exp\left(-\int_{t_3 - dt_3}^{t_3} d\tau_3\ \Delta R_X^{ID}(\Delta L_3 + \delta\Delta L_3, \tau_3)\right) \times \\
&\quad \times \left(\mu_{\text{A-del}}\left[(\Delta_L L_1, t_1)\ ;\ (0, -\delta\Delta L_3, [t_3 - dt_3, t_3])\right](\sigma_{BE:L} = 1)/dt_3\right) \times \\
&\quad \times \mu_{P_0}\left[(\Delta_F L_D, t_F) \mid (\Delta L_3, t_3)\right](\sigma_{BE} = 2)\Big] \ . \qquad (78)
\end{aligned}
$$

Second, we lump together the factors, $\mu_{\text{D-cr}\,\to\,\text{AD}\,\to\,\text{D}}[\cdots]$ and $\mu_{\text{A-del}}[\cdots]$, as well as the exponential factor in between, and perform the summation over $\delta\Delta L_3$. This gives a set of multiplication factors (of $O(dt_1 \cdot dt_3)$) that depend on the lengths $\Delta_L L_1$, $\Delta_R L_1$ and ΔL_3, and the times, t_1 and t_3, thus requiring the space-complexity of $O(\{N_P\}^2\{L^{CO}\}^3)$ if you want to store all of them. Fortunately, when the evolution model is time-homogeneous, the factors depend on the times only through $t_3 - t_1$. Besides, in the computation of Eq.52, all we need regarding the dependence on $(\Delta_L L_1, \Delta_R L_1)$ is through $\Delta L_1 = \Delta_L L_1 + \Delta_R L_1$. Thus, we could reduce the space-complexity to $O(N_P\{L^{CO}\}^2)$ by defining the following "complex" factors(, which are of $O(dt_1 \cdot dt_3)$):

$$
\begin{aligned}
&\mu_{\text{D-cr}\,\to\,\text{A-del}\,\to\,\text{A-del}}\left[(\Delta L_1, \cdot, [t_1, t_1 + dt_1])\ ;\ ()\ ;\ (\Delta L_3, [t_3 - dt_3, t_3])\right](\sigma_{BE:R} = \sigma_{BE:L} = 1) \\
&\stackrel{\text{def}}{=} \sum_{\Delta_R L_1 = 1}^{\Delta L_1 - 1} \sum_{\delta\Delta L_3 = 0}^{\infty} \Big[\\
&\mu_{\text{D-cr}\,\to\,\text{AD}\,\to\,\text{D}}\left[(\Delta_R L_1, x = \Delta_L L_1, [t_1, t_1 + dt_1])\ ;\ ()\ ;\ (\Delta L_3 + \delta\Delta L_3, t_3 - dt_3)\right](\sigma_{BE:R} = 1) \times \\
&\quad \times \exp\left(-\int_{t_3 - dt_3}^{t_3} d\tau_3\ \Delta R_X^{ID}(\Delta L_3 + \delta\Delta L_3, \tau_3)\right) \times \\
&\quad \times \mu_{\text{A-del}}\left[(\Delta_L L_1 = \Delta L_1 - \Delta_R L_1, t_1)\ ;\ (0, -\delta\Delta L_3, [t_3 - dt_3, t_3])\right](\sigma_{BE:L} = 1)\Big] \ . \qquad (79)
\end{aligned}
$$

These factors for ranging ΔL_1, ΔL_3, and $t_3 - t_1$ can be numerically computed with the space-complexity of $O(N_P\{L^{CO}\}^2)$ and the time-complexity of $O(N_P\{L^{CO}\}^4)$, and the computation is easily cast into parallel or distributed computing. (For example, if the computations with different ΔL_1's (or ΔL_3's) are distributed, each computation has the time-complexity of $O(N_P\{L^{CO}\}^3)$, albeit with an unchanged space-complexity.)

Now, by referring to Figure 13, we notice that the aforementioned $\mu_{\text{D-cr}\,\to\,\text{A-del}\,\to\,\text{A-del}}[\cdots]$ covers the "amputated" portion of the coloring-pattern evolution for the 3rd-order class (a). Thus, we can *define* the "amputated" version of $\mu^{\text{3rd (a)}}_{P\ \text{case-(iv)}}[\cdots]$ as:

$$\begin{aligned}
&\left(\mu^{\text{3rd (a)}}_{P\ \text{case-(iv)}}\right)_{\text{Amp.}}[(\Delta L_3,\ t_3)\mid(\Delta L_1,\ t_1)]\times(dt_1\cdot dt_3)\\
\overset{\text{def}}{=}\ &\mu_{\text{D-cr}\,\to\,\text{A-del}\,\to\,\text{A-del}}\left[(\Delta L_1,\cdot,[t_1,t_1+dt_1])\ ;\ ()\ ;\ (\Delta L_3,[t_3-dt_3,t_3])\right](\sigma_{BE:R}=\sigma_{BE:L}=1).
\end{aligned}\tag{80}$$

Then, the rest of the computation procedure for this pattern (a) is prescribed with the equations almost identical to Eq.72 and Eqs.73 & 74 (or Eqs.75 & 76). The only differences are: (1) the superscript, "2nd (ii)" should be replaced with "3rd (a)"; (2) $(\Delta L_2, t_2)$ should be replaced with $(\Delta L_3, t_3)$; and (3) $\int_{t_I}^{t_F-dt_1-dt_2} dt_1 \int_{t_1+dt_1+dt_2}^{t_F} dt_2$ should be replaced with $\int_{t_I}^{t_F-dt_1-dt_2-dt_3} dt_1 \int_{t_1+dt_1+dt_2+dt_3}^{t_F} dt_3$. And, of course, arguments on the space- and time-complexities also hold in the same way.

G Time-Efficient Computation of Contributions from 3rd-order Pattern (c): A $\to$ DA $\to$ DAD $\to$ D

The contributions from the 3rd-order pattern (c), Eq. 53, were given in subsection 5.2.2. As in appendix G, the equation can be cast into a series of equations to time-efficiently calculate these contributions.

First, we deal with the creation of the "D"-region on the right (in $[t_2, t_2+dt_2]$) (and a part of the complete deletion of the "A"-region (in $[t_3-dt_3, t_3]$)), *in two steps*. In the first step, we define the following complex factor(, which is $O(dt_2\cdot dt_3)$):

$$\begin{aligned}
&\mu_{\text{D-cr(R)}\,\to\,\cdot\text{A-del}}\left[(\Delta_2 L_A, x_2, [t_2, t_2+dt_2])\ ;\ (-\delta\Delta_3 L_{DL}, \Delta_3 L_{DR}, [t_3-dt_3, t_3])\right](\sigma_{BE3}=0)\\
\overset{\text{def}}{=}\ &\sum_{\delta\Delta_3 L_{DR}=0}^{\infty}\Big[\ \mu_{\text{D-cr}}\left[(x_2,[t_2,t_2+dt_2])\ ;\ (\Delta_3 L_{DR}+\delta\Delta_3 L_{DR}, t_3-dt_3)\right]\times
\end{aligned}$$

$$\times \exp\left(-\int_{t_3-dt_3}^{t_3} d\tau_3\ \Delta R_X^{ID}(\Delta_3 L_{DR}+\delta\Delta_3 L_{DR},\tau_3)\right)\times$$
$$\times\mu_{\text{A-del}}\left[(\Delta_2 L_A,t_2)\ ;\ (0,-\delta\Delta_3 L_D=-(\delta\Delta_3 L_{DL}+\delta\Delta_3 L_{DR}),[t_3-dt_3,t_3])\right](\sigma_{BE}=0)\ \Big]\ , \tag{81}$$

where x_2 $(=\Delta_2 L_{DL}+\Delta_2 L_A)$ only serves as a reminder of the insertion position of the "D"-region. With ranging $\Delta_2 L_A$, $\delta\Delta_3 L_{DL}$, $\Delta_3 L_{DR}$ and t_3-t_1, this computation has the time-complexity of $O(N_P\{L^{CO}\}^4)$ and the space-complexity of $O(N_P\{L^{CO}\}^3)$. If, however, we only memorize the output, *e.g.*, for each fixed $\Delta_3 L_{DR}$, and output the results each time when all the factors for the fixed $\Delta_3 L_{DR}$ are computed, the space-complexity reduces to $O(N_P\{L^{CO}\}^2)$. (The same holds true for the parallel (or distributed) computing, in which case the time-complexity also reduces to $O(N_P\{L^{CO}\}^3)$.)

In the second step, we define its "extended" version(, which is $O(dt_3)$):

$$\mu_{\cdot\text{A}\to\cdot\text{AD}\to\cdot\text{A-del}}\left[(\Delta_1 L_A,t_1)\ ;\ x_2\ ;\ (-\delta\Delta_3 L_{DL},\Delta_3 L_{DR},[t_3-dt_3,t_3])\right](\sigma_{BE2}=1,\sigma_{BE3}=0)$$
$$\stackrel{\text{def}}{=}\sum_{\Delta_2 L_A=1}^{\infty}\int_{t_1+dt_1}^{t_3-dt_2-dt_3}dt_2\Big[\ \mu_{P_0}\left[(\Delta_2 L_A,t_2)\mid(\Delta_1 L_A,t_1)\right](\sigma_{BE2}=1)\times$$
$$\times\left(\mu_{\text{D-cr(R)}\to\cdot\text{A-del}}\left[(\Delta_2 L_A,x_2,[t_2,t_2+dt_2])\ ;\ (-\delta\Delta_3 L_{DL},\Delta_3 L_{DR},[t_3-dt_3,t_3])\right](\sigma_{BE3}=0)/dt_2\right)\ \Big]\ . \tag{82}$$

With ranging $\Delta_1 L_A$, $\delta\Delta_3 L_{DL}$, $\Delta_3 L_{DR}$, and t_3-t_1, this definition can be computed with the space-complexity of $O(N_P\{L^{CO}\}^3)$ and the time-complexity of $O(\{N_P\}^2\{L^{CO}\}^4)$. If, for example, the computations with different $\Delta_3 L_{DR}$'s are distributed (or parallelized), each computation has the space-complexity of $O(N_P\{L^{CO}\}^2)$ and the time-complexity of $O(\{N_P\}^2\{L^{CO}\}^3)$.

Using the above definition of $\mu_{\cdot\text{A}\to\cdot\text{AD}\to\cdot\text{A-del}}[\cdots]$, the expression we desire, *i.e.*, Eq.53, reduces to:

$$\mu_{P\ \text{case-(iv)}}^{\text{3rd (c)}}\left[(\Delta_F L_D,t_F)\mid(\Delta_I L_A,t_I)\right]\times\exp\left(-\int_{t_I}^{t_F}d\tau\Delta R_X^{ID}(\Delta_I L_A,\tau)\right)$$
$$=\sum_{\Delta_1 L_A=1}^{\infty}\sum_{\Delta_3 L_{DL}=1}^{\infty}\sum_{\Delta_3 L_{DR}=1}^{\infty}\sum_{\delta\Delta_3 L_{DL}=0}^{\infty}\int_{t_I}^{t_F-dt_1-dt_2-dt_3}dt_1\int_{t_1+dt_1+dt_2+dt_3}^{t_F}dt_3\Big[$$
$$\mu_{P_0}\left[(\Delta_1 L_A,t_1)\mid(\Delta_I L_A,t_I)\right](\sigma_{BE}=2)\times$$
$$\times\left(\mu_{\text{D-cr}}\left[(x_1=1,[t_1,t_1+dt_1])\ ;\ (\Delta_3 L_{DL}+\delta\Delta_3 L_{DL},t_3-dt_3)\right]/dt_1\right)\times$$

$$\times \exp\left(-\int_{t_3-dt_3}^{t_3} d\tau_3\ \Delta R_X^{ID}(\Delta_3 L_{DL}+\delta\Delta_3 L_{DL},\tau_3)\right)\times$$
$$\times\Big(\mu_{\cdot\mathrm{A}\,\to\,\cdot\mathrm{AD}\,\to\,\cdot\mathrm{A\text{-}del}}\left[(\Delta_1 L_A,t_1)\ ;\ x_2\ ;\ (-\delta\Delta_3 L_{DL},\Delta_3 L_{DR},[t_3-dt_3,t_3])\right](\sigma_{BE2}=1,\sigma_{BE3}=0)/dt_3\Big)\times$$
$$\times\mu_{P_0}\left[(\Delta_F L_D,t_F)\mid(\Delta_3 L_D=\Delta_3 L_{DL}+\Delta_3 L_{DR},t_3)\right](\sigma_{BE}=2)\Big]\ . \tag{83}$$

Next, we deal *fully* with the complete deletion of the "A"-region (in $[t_3-dt_3,t_3]$). Although, in the current setting, the results of the deletion depend on $\Delta_3 L_{DL}$ and $\Delta_3 L_{DR}$, what we finally want is the function of the *total length*, $\Delta_3 L_D \stackrel{\text{def}}{=} \Delta_3 L_{DL}+\Delta_3 L_{DR}$, of the "D"-region after the deletion. Hence, we define the following complex factor(, which is $O(dt_1\cdot dt_3)$):

$$\mu_{\text{D-cr(L)}\,\to\,\text{D-cr(R)}\,\to\,\text{A-del}}\left[(\Delta_1 L_A,x_1=1,[t_1,t_1+dt_1])\ ;\ x_2\ ;\ (\Delta_3 L_D,[t_3-dt_3,t_3])\right](\sigma_{BE2}=1,\sigma_{BE3}=0)$$
$$\stackrel{\text{def}}{=}\ \sum_{\Delta_3 L_{DL}=1}^{\Delta_3 L_D-1}\ \sum_{\delta\Delta_3 L_{DL}=0}^{\infty}\Big[\exp\left(-\int_{t_3-dt_3}^{t_3} d\tau_3\ \Delta R_X^{ID}(\Delta_3 L_{DL}+\delta\Delta_3 L_{DL},\tau_3)\right)\times$$
$$\times\mu_{\text{D-cr}}\left[(x_1=1,[t_1,t_1+dt_1])\ ;\ (\Delta_3 L_{DL}+\delta\Delta_3 L_{DL},t_3-dt_3)\right]\times$$
$$\times\mu_{\cdot\mathrm{A}\,\to\,\cdot\mathrm{AD}\,\to\,\cdot\mathrm{A\text{-}del}}\left[(\Delta_1 L_A,t_1)\ ;\ x_2\right.$$
$$\left.\ ;\ (-\delta\Delta_3 L_{DL},\Delta_3 L_{DR}=\Delta_3 L_D-\Delta_3 L_{DL},[t_3-dt_3,t_3])\right](\sigma_{BE2}=1,\sigma_{BE3}=0)\Big]\ . \tag{84}$$

With ranging $\Delta_1 L_A$, $\Delta_3 L_D$ and t_3-t_1, this computation has the space-complexity of $O(N_P\{L^{CO}\}^3)$ and the time-complexity of $O(N_P\{L^{CO}\}^4)$, which may be quite hard on a single computer. However, if computations with, *e.g.*, different $\Delta_1 L_A$'s are distributed (or parallelized), the space- and time-complexities of each computation reduce to $O(N_P\{L^{CO}\}^2)$ and $O(N_P\{L^{CO}\}^3)$, respectively. [43]

Now, Figure 14 suggests that the aforementioned $\mu_{\text{D-cr(L)}\,\to\,\text{D-cr(R)}\,\to\,\text{A-del}}[\cdots]$ covers the "amputated" portion of the coloring-pattern evolution for the 3rd-order class (c). Thus, we have the "amputated" multiplication factor:

$$\left(\mu_{P\ \text{case-(iv)}}^{\text{3rd (c)}}\right)_{\text{Amp.}}\left[(\Delta_3 L_D,\ t_3)\mid(\Delta_1 L_A,\ t_1)\right]\times(dt_1\cdot dt_3)$$

[43]The entire result of this definition is of size $O(N_P\{L^{CO}\}^2)$, thus may be stored in the memory of a common contemporary computer.

$$\stackrel{\text{def}}{=} \mu_{\text{D-cr(L)} \to \text{D-cr(R)} \to \text{A-del}} \left[(\Delta_1 L_A, x_1 = 1, [t_1, t_1 + dt_1]) \ ; \right.$$
$$\left. x_2 \ ; \ (\Delta_3 L_D, [t_3 - dt_3, t_3]) \right] (\sigma_{BE2} = 1, \sigma_{BE3} = 0) \ . \tag{85}$$

As in appendix F the rest of the computation procedure for this pattern (c) is prescribed with the equations almost identical to Eq.72 and Eqs.73 & 74 (or Eqs.75 & 76). The only differences are: (1) the superscript, "2nd (ii)" should be replaced with "3rd (c)"; (2) $(\Delta L_1, t_1)$ and $(\Delta L_2, t_2)$ should be replaced with $(\Delta_1 L_A, \ t_1)$ and $(\Delta_3 L_D, \ t_3)$, respectively; and (3) $\int_{t_I}^{t_F - dt_1 - dt_2} dt_1 \int_{t_1 + dt_1 + dt_2}^{t_F} dt_2$ should be replaced with $\int_{t_I}^{t_F - dt_1 - dt_2 - dt_3} dt_1 \int_{t_1 + dt_1 + dt_2 + dt_3}^{t_F} dt_3$. Again, arguments on the space- and time-complexities remain the same.

H Time-Efficient Computation of Contributions from 3rd-order Pattern (e): A $\to$ DA $\stackrel{\text{B-er}}{\to}$ DA $\to$ D

The contributions from the 3rd-order pattern (e), Eq. 54, were given in subsection 5.2.3. Here is a series of measures to quickly compute these contributions.

First, paying attention to the "boundary-eroding" deletion, we define the following complex factor(, which is $O(dt_1 \cdot dt_2)$):

$$\mu_{\text{D-cr} \to \text{B-er.}} \left[(x = 1, [t_1, t_1 + dt_1]) \ ; \ (\Delta_2 L_D, -\delta\Delta_2 L_A, [t_2 - dt_2, t_2]) \right]$$
$$\stackrel{\text{def}}{=} \sum_{\delta\Delta_2 L_D = 1}^{\infty} \Big[\ \mu_{\text{D-cr}} \left[(x, [t_1, t_1 + dt_1]) \ ; \ (\Delta_2 L_D + \delta\Delta_2 L_D, t_2 - dt_2) \right] \times$$
$$\times \exp \left\{ - \int_{t_2 - dt_2}^{t_2} d\tau_2 \ \Delta R_X^{ID} (\Delta_2 L_D + \delta\Delta_2 L_D, \tau_2) \right\} \times dt_2 \ \tilde{g}_D (\delta\Delta_2 L_A + \delta\Delta_2 L_D, t_2) \ \Big] \ . \tag{86}$$

With ranging $\Delta_2 L_D$, $\delta\Delta_2 L_A$ and $t_2 - t_1$, this computation has the space- and time-complexities of $O(N_P \{L^{CO}\}^2)$ and $O(N_P \{L^{CO}\}^3)$, respectively.

Second, paying attention to the "boundary-eroding" deletion, again, we define the following(, which is also $O(dt_1 \cdot dt_2)$):

$$\mu_{(\text{A} \to) \text{ D-cr(L)} \to \text{B-er}} \left[(\Delta_1 L_A, x = 1, [t_1, t_1 + dt_1]) \ ; \ (\Delta_2 L_D, \Delta_2 L_A, [t_2 - dt_2, t_2]) \right] (\sigma_{BE1} = 1)$$

$$
\begin{aligned}
&\stackrel{\text{def}}{=} \sum_{\delta\Delta_2 L_A=1}^{\infty} \Big[\, \mu_{\text{D-cr}\,\rightarrow\,\text{B-er.}}\left[(x=1,[t_1,t_1+dt_1])\ ;\ (\Delta_2 L_D, -\delta\Delta_2 L_A,[t_2-dt_2,t_2])\right]\times \\
&\quad \times\mu_{P_0}\left[(\Delta_2 L_A+\delta\Delta_2 L_A, t_2-dt_2)\mid(\Delta_1 L_A,t_1)\right](\sigma_{BE1}=1)\times \\
&\quad \times\exp\left\{-\int_{t_2-dt_2}^{t_2} d\tau_2\Big[\Delta R_X^{ID}(\Delta_2 L_A+\delta\Delta_2 L_A,\tau_2)\Big]\right\}\Big]\ . \qquad (87)
\end{aligned}
$$

With ranging $\Delta_1 L_A$, $\Delta_2 L_D$ and $\Delta_2 L_A$, and t_2-t_1, this computation has the space- and time-complexities of $O(N_P\{L^{CO}\}^3)$ and $O(N_P\{L^{CO}\}^4)$, respectively. If, however, we distribute the computations, *e.g.*, with different $\Delta_1 L_A$'s, the space- and time-complexities of each computation become $O(N_P\{L^{CO}\}^2)$ and $O(N_P\{L^{CO}\}^3)$, respectively.

Third, paying attention to the complete deletion of the "A"-region (in $[t_3-dt_3,t_3]$), we define the following complex factor(, which is $O(dt_3)$):

$$
\begin{aligned}
&\mu_{\text{DA}\,\rightarrow\,\text{A-del}}\left[(\Delta_2 L_D,\Delta_2 L_A,t_2)\ ;\ (\Delta_3 L_D,[t_3-dt_3,t_3])\right](\sigma_{BE2}=1) \\
&\stackrel{\text{def}}{=} \sum_{\delta\Delta_3 L_D=0}^{\infty}\Big[\,\mu_{\text{A-del}}\left[(\Delta_2 L_A,t_2)\ ;\ (0,-\delta\Delta_3 L_D,[t_3-dt_3,t_3])\right](\sigma_{BE2}=1)\times \\
&\quad\times\mu_{P_0}\left[(\Delta_3 L_D+\delta\Delta_3 L_D,t_3-dt_3)\mid(\Delta_2 L_D,t_2)\right](\sigma_{BE(D)}=2)\times \\
&\quad\times\exp\left\{-\int_{t_3-dt_3}^{t_3} d\tau_3\ \Delta R_X^{ID}(\Delta_3 L_D+\delta\Delta_3 L_D,\tau_3)\right\}\Big]\ . \qquad (88)
\end{aligned}
$$

With ranging $\Delta_2 L_D$, $\Delta_2 L_A$, $\Delta_3 L_D$ and t_3-t_2, this computation also has the space- and time-complexities of $O(N_P\{L^{CO}\}^3)$ and $O(N_P\{L^{CO}\}^4)$, respectively. If, however, we distribute the computations, *e.g.*, with different $\Delta_3 L_D$'s, the space- and time-complexities of each computation become $O(N_P\{L^{CO}\}^2)$ and $O(N_P\{L^{CO}\}^3)$, respectively, as in the previous computation.

Next, we further combine the complex factors, Eq.87 and Eq.88, and define the following higher-level complex factor(, which is $O(dt_1\cdot dt_3)$):

$$
\begin{aligned}
&\mu_{(\text{A}\,\rightarrow)\ \text{D-cr(L)}\,\rightarrow\,\text{B-er}\,\rightarrow\,\text{A-del}}\left[(\Delta_1 L_A, x=1,[t_1,t_1+dt_1])\ ;\ ()\ ;\ (\Delta_3 L_D,[t_3-dt_3,t_3])\right](\sigma_{BE1}=\sigma_{BE2}=1) \\
&\stackrel{\text{def}}{=} \sum_{\Delta_2 L_D=1}^{\infty}\sum_{\Delta_2 L_A=1}^{\infty}\int_{t_1+dt_1+dt_2}^{t_3-dt_3} dt_2\Big[\\
&\quad\Big(\mu_{(\text{A}\,\rightarrow)\ \text{D-cr(L)}\,\rightarrow\,\text{B-er}}\left[(\Delta_1 L_A,x=1,[t_1,t_1+dt_1])\ ;\ (\Delta_2 L_D,\Delta_2 L_A,[t_2-dt_2,t_2])\right](\sigma_{BE1}=1)/dt_2\Big)\times \\
&\quad\times\mu_{\text{DA}\,\rightarrow\,\text{A-del}}\left[(\Delta_2 L_D,\Delta_2 L_A,t_2)\ ;\ (\Delta_3 L_D,[t_3-dt_3,t_3])\right](\sigma_{BE2}=1)\ \Big]\ . \qquad (89)
\end{aligned}
$$

With ranging $\Delta_1 L_A$, $\Delta_3 L_D$ and $t_3 - t_1$, this computation has the space- and time-complexities of $O(N_P\{L^{CO}\}^3)$ and $O(\{N_P\}^2\{L^{CO}\}^4)$, respectively; this is considerably hard on a single computer. One big problem is the large size of the input data, $\mu_{(\mathrm{A}\,\rightarrow)\ \mathrm{D\text{-}cr(L)}\ \rightarrow\ \mathrm{B\text{-}er}}[\cdots]$ and $\mu_{\mathrm{DA}\ \rightarrow\ \mathrm{A\text{-}del}}[\cdots]$, each of which takes $O(N_P\{L^{CO}\}^3)$ memory space if you want to store all of its elements. (In contrast, the output data, $\mu_{(\mathrm{A}\,\rightarrow)\ \mathrm{D\text{-}cr(L)}\ \rightarrow\ \mathrm{B\text{-}er}\ \rightarrow\ \mathrm{A\text{-}del}}[\cdots]$, take up only $O(N_P\{L^{CO}\}^2)$ memory space.) The only way to avoid this problem of large space-complexity (using a single computer) is to read in only $\mu_{(\mathrm{A}\,\rightarrow)\ \mathrm{D\text{-}cr(L)}\ \rightarrow\ \mathrm{B\text{-}er}}[\cdots]$ with a particular $\Delta_1 L_A$ and only $\mu_{\mathrm{DA}\ \rightarrow\ \mathrm{A\text{-}del}}[\cdots]$ with a particular $\Delta_3 L_D$, compute $\mu_{(\mathrm{A}\,\rightarrow)\ \mathrm{D\text{-}cr(L)}\ \rightarrow\ \mathrm{B\text{-}er}\ \rightarrow\ \mathrm{A\text{-}del}}[\cdots]$ with this particular combination, $(\Delta_1 L_A, \Delta_3 L_D)$, and free the memory space for the input data after the end of this particular computation. Then, the required memory space reduces to $O(N_P\{L^{CO}\}^2)$, and repeating this procedure for all combinations of $(\Delta_1 L_A, \Delta_3 L_D)$ will give the complete definition, Eq.89. When performing parallel (or distributed) computing, as well, it should be wise to distribute the computations with different $(\Delta_1 L_A, \Delta_3 L_D)$'s, not just those differing in ,*e.g.*, $\Delta_1 L_A$'s *alone*, to reduce the space-complexity of each computation (also to $O(N_P\{L^{CO}\}^2)$); in this case, the time-complexity of each computation also reduces to $O(\{N_P\}^2\{L^{CO}\}^2)$.

Now, as Figure 15 suggests, the higher-level complex factor defined in Eq.89 covers the "amputated" portion of the coloring-pattern evolution for the 3rd-order pattern (e). Thus, we *define* the following "amputated" multiplication factor:

$$\begin{aligned} & \left(\mu_{P\ \text{case-(iv)}}^{\text{3rd (e)}}\right)_{\text{Amp.}}[(\Delta_3 L_D,\ t_3)\mid(\Delta_1 L_A,\ t_1)]\times(dt_1\cdot dt_3) \\ \overset{\text{def}}{=}\ & \mu_{(\mathrm{A}\,\rightarrow)\ \mathrm{D\text{-}cr(L)}\ \rightarrow\ \mathrm{B\text{-}er}\ \rightarrow\ \mathrm{A\text{-}del}}\ [(\Delta_1 L_A, x=1, [t_1, t_1+dt_1])\ ;\ () \\ & \qquad ;\ (\Delta_3 L_D, [t_3-dt_3, t_3])](\sigma_{BE1}=\sigma_{BE2}=1)\ . \end{aligned} \tag{90}$$

Then, again, the remaining procedure is prescribed with the equations almost identical to Eq.72 and Eqs.73 & 74 (or Eqs.75 & 76). The only differences are: (1) the superscript, "2nd (ii)" should be replaced with "3rd (e)"; (2) $(\Delta L_1, t_1)$ and $(\Delta L_2, t_2)$ should be replaced with $(\Delta_1 L_A,\ t_1)$ and $(\Delta_3 L_D,\ t_3)$, respectively; and (3) $\int_{t_I}^{t_F - dt_1 - dt_2} dt_1 \int_{t_1+dt_1+dt_2}^{t_F} dt_2$ should be

replaced with $\int_{t_I}^{t_F-dt_1-dt_2-dt_3} dt_1 \int_{t_1+dt_1+dt_2+dt_3}^{t_F} dt_3$. Again, the space- and time-complexities of the computations are the same as in "2nd (ii)".

I Computing Theoretically Expected Frequencies of Gap-Configurations

To validate the new perturbation method presented in this study, we computed theoretically expected frequencies of the configurations of gapped segments in the bulk (*i.e.*, *not* on either end) of ancestor-descendant PWAs, and compared them with their frequencies actually "observed" in the bulk of simulated PWAs. Here we explain how the expected frequencies are computed.

First, the multiplication factor, *e.g.*, Eq. 47 for case-(iv) segments, was transformed into the probability that the segment, *including both flanking preserved ancestral sites (PASs)*, occurs in the PWAs. This can be done by modifying the identity, Eq. 48, as follows:

$$\begin{aligned} & P_{\text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] \\ \equiv\ & \mu_{P\ \text{case-(iv)}}\left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] \times \exp\left\{ - \int_{t_I}^{t_F} d\tau \left[R_X^{ID}([L,R],\tau) + \Delta R_X^{ID}(\Delta_I L_A, \tau) \right] \right\} . \end{aligned} \tag{91}$$

We know that, in (locally) space-homogeneous model, $\Delta R_X^{ID}(\Delta_I L_A, \tau) = (g_I(\tau) + g_D(\tau)) \cdot \Delta_I L_A$ holds [48, 3], where $g_I(\tau)$ ($\stackrel{\text{def}}{=} \sum_{l=1}^{L_I^{CO}} g_I(l,\tau)$) is the total insertion rate per site at time τ, and $g_D(\tau)$ ($\stackrel{\text{def}}{=} \sum_{l=1}^{L_D^{CO}} g_D(l,\tau)$) is the total deletion rate per site at time τ.

The question is: what on earth is the $R_X^{ID}([L,R],\tau)$? Actually, the answer to this question can vary, depending on how we treat the "surrounding regions" flanking the PASs, L and R, especially, insertions occurring there (discussed to some extent in section R8 of [3]). For example, the evolution model used by Dawg [48] made a quite natural choice, considering that insertions occur also in these regions at the same rate as in the subject sequence, and *incorporating* such insertions into the evolution of the sequence. In this model,

$R_X^{ID\ (Dawg)}([L,R],\tau)\ |\ =\ \ g_I(\tau) + \sum_{l=1}^{L_D^{CO}} \{(l-1)\ g_D(l,\tau)\}\ +\ 2\ (g_I(\tau) + g_D(\tau))$, where the summation in the 2nd term comes from deletions sticking out of the subject sequence [48]. Meanwhile, the "long indel" model of Miklós et al. [2] made another choice, giving such insertions the rates that are "mirror-images" of the rates of deletions reaching either end of the sequence, in order to keep the model time-reversible. (The expression of $R_X^{ID}([L,R],\tau)$ for the "long indel" model is omitted here because it is somewhat complex.)

In this study, we make yet another choice, that is, we *don't care at all* whether any insertions occur or not in the "surrounding regions". This choice should be appropriate for the problem at hand, because insertions in the "surrounding regions", if at all, will *never* erode the gapped segment (including the PASs, L and R). Thus, the $R_X^{ID}([L,R],\tau)$ we want is:

$$\begin{aligned} R_X^{ID\ (Our)}([L,R],\tau) &= R_X^{ID\ (Dawg)}([L,R],\tau) - 2\ g_I(\tau) \\ &= g_I(\tau) + \sum_{l=1}^{L_D^{CO}} \{(l-1)\ g_D(l,\tau)\}\ +\ 2\ g_D(\tau)\ . \end{aligned} \tag{92}$$

(Incidentally, if we time-integrate this equation across a phylogenetic tree, the result becomes equivalent to the total summation of the exponentiated factors in Eq.(SM-7.1) of [1], as it should be.)

Substituting these results into Eq. 91, we get:

$$\begin{aligned} & P_{\text{case-(iv)}}\left[(\Delta_F L_D, t_F)\ |\ (\Delta_I L_A, t_I)\right] \\ = \ & \mu_{P\ \text{case-(iv)}}\left[(\Delta_F L_D, t_F)\ |\ (\Delta_I L_A, t_I)\right] \times \\ \times \ & \exp\left\{ - \int_{t_I}^{t_F} d\tau \Big[g_I(\tau) + \sum_{l=1}^{L_D^{CO}} \{(l-1)\ g_D(l,\tau)\}\ +\ 2\ g_D(\tau)\ + (g_I(\tau) + g_D(\tau)) \cdot \Delta_I L_A \Big] \right\} . \end{aligned} \tag{93}$$

To compute the expected frequency of each configuration of the case (iv) gapped-segment, which is *uniquely* specified by $(\Delta_I L_A, \Delta_F L_D)$, we simply multiply the probability, Eq. 93, by the total number, N_T, of regions that could potentially become the gapped segment in

question. Thus, we have:

$$\{\text{Theoretically expected frequency of } (\Delta_I L_A, \Delta_F L_D)\} = N_T \times P_{\text{case-(iv)}} \left[(\Delta_F L_D, t_F) \mid (\Delta_I L_A, t_I)\right] . \tag{94}$$

Although we discussed case-(iv) gapped segments as an example, we could also obtain the theoretically expected frequencies of the configurations of case-(ii) and -(iii) gapped segments in totally similar manners.

In this study, we *naïvely* substituted the total number of ancestral sites, $N_{S0} \times L_{S0}$, for N_T. Here N_{S0} and L_{S0} are the number of ancestral sequences and the length of the ancestral sequences, respectively. In this study, $N_{S0} = 10^{+5}$ and $L_{S0} = 10^{+4}$. Therefore, $N_T \approx 10^{+5} \times 10^{+4} = 10^{+9}$. There are at least two factors to consider in order to determine whether this *naïve* measure and Eq. 94 work well. First, the actual number of available regions *must* be the number of sub-sequences embedded in the ancestral sequences. This should be $N_{S0} \times (L_{S0} - \Delta_I L_A - 1)$ for the configuration, $(\Delta_I L_A, \Delta_F L_D)$. In this study, $\Delta_I L_A \leq 100 \ll 10^{+4} = L_{S0}$. Therefore, the aforementioned actual number should be well approximated (albeit slightly overestimated) by $N_{S0} \times L_{S0}$. Second, Eq. 94 exactly holds *only when* the available N_T regions are independent of each other, whereas some nearby regions actually overlap each other; this poses the problem of auto-correlations. As long as the gapped segments are distributed sparsely enough within the sequences, such auto-correlations should be negligible. This must be indeed the case when the time-interval is 0.2 or 0.5, although it may not necessarily the case when the time-interval is 1.0. Anyway, in this study, we will use the *naïve* equation, Eq. 94, and solving the problem of auto-correlations will be left for future studies. [44]

[44] A brief consideration suggests that the auto-correlation here should cause an "exclusion effect"; two (or more) different gapped segment should *never* overlap each other, thus the frequency of actual occurrences should be somewhat lower than predicted by Eq. 94, provided that the probability is practically exact.

References

[1] K Ezawa. General continuous-time Markov model of sequence evolution via insertions/deletions: local alignment probability computation. *BMC Bioinformatics.*, 17:397, 2016.

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

[3] K Ezawa. General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable? *BMC Bioinformatics.*, 17:304, 2016. Erratum in: BMC Bioinformatics (2016) 17:457.

[4] D Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press, Cambridge, UK, 1997.

[5] S Kumar and A Filipski. Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res.*, 17:127–135, 2007.

[6] MR Aniba, O Poch, and JD Thompson. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Res.*, 38:7353–7363, 2010.

[7] A Löytynoja. Alignment methods: strategies, challenges, benchmarking, and comparative overview. In M Anisimova, editor, *Evolutionary Genomics. Methods in Molecular Biology (Methods and Protocols), vol. 855*, pages 203–235. Humana Press, Totowa, NJ, 2012.

[8] S Iantorno, K Gori, N Goldman, M Gil, and C Dessimoz. Who watches the watchmen? an appraisal of benchmarks for multiple sequence alignment. In D Russell, editor, *Multiple Sequence Alignment Methods. Methods in Molecular Biology (Methods and Protocols), vol. 1079*, pages 59–73. Humana Press, Totowa, NJ, 2014.

[9] T Warnow. *Computational Phylogenetics: An introduction to designing methods for phylogeny estimation*, chapter 9. Cambridge University Press, 2017.

[10] A Löytynoja and N Goldman. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science.*, 320:1632–1635, 2008.

[11] SB Needleman and CD Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J Mol Biol.*, 48:443–453, 1970.

[12] TF Smith and MS Waterman. Identification of common molecular subsequences. *J Mol Biol.*, 147:195–197, 1981.

[13] DF Feng and RF Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol.*, 25:351–360, 1987.

[14] JD Thompson, DG Higgins, and TJ Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[15] E Margulies and E Birney. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat Rev Genet.*, 9:303–313, 2008.

[16] JR Lupski. Genomic rearrangements and sporadic disease. *Nat Genet.*, 39(7 Suppl):S43–S47, 2007.

[17] M Lynch. *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA, 2007.

[18] W Gu, F Zhang, and JR Lupski. Mechanisms for human genomic rearrangements. *Pathogenetics.*, 1:4, 2008.

[19] TH Lee, J Kim, JS Robertson, and AH Paterson. Plant genome duplication database. In A. van Dijk, editor, *Plant Genomic Databases. Methods in Molecular Biology, vol. 1533*, pages 267–277. Humana Press, New York, NY, 2017.

[20] CN Dewey. Whole-genome alignment. In M Anisimova, editor, *Evolutionary Genomics. Methods in Molecular Biology (Methods and Protocols), vol. 855*, pages 237–257. Humana Press, Totowa, NJ, 2012.

[21] D Earl, N Nguyen, G Hickey, RS Harris, S Fitzgerald, and K et. al. Beal. Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res.*, 24:2077–2089, 2014.

[22] JD Thompson. *Statistics for Bioinformatics: Methods for Multiple Sequence Alignment, 1st Edition*, chapter 6. ISTE Press - Elsevier, 2016.

[23] T Yada. Genome alignment. In S Ranganathan, K Nakai, and C Schonbach, editors, *The Encyclopedia of Bioinformatics and Computational Biology*, pages 268–283. Elsevier, Amsterdam, NL, 2019.

[24] O Gotoh. An improved algorithm for matching biological sequences. *J Mol Biol.*, 162:705–708, 1982.

[25] CB Do, MS Mahabhashyam, M Brudno, and S Batzoglou. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, 15:330–340, 2005.

[26] C Notredame. Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol.*, 3:e123, 2007.

[27] A Wilm, DG Higgins, and C Notredame. R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, 36:e52, 2008.

[28] K Kryukov and N Saitou. MISHIMA–a new method for high speed multiple alignment of nucleotide sequences of bacterial genome scale data. *BMC Bioinformatics.*, 11:142, 2010.

[29] G Landan and D Graur. Characterization of pairwise and multiple sequence alignment errors. *Gene.*, 441:141–147, 2009.

[30] K Ezawa. Characterization of multiple sequence alignment errors using complete-likelihood score and position-shift map. *BMC Bioinformatics.*, 17:133, 2016.

[31] G Lunter, A Rocco, N Mimouni, A Heger, A Caldeira, and J Hein. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, 18:298–309, 2008.

[32] RA Cartwright. Problems and solutions for estimating indel rates and length distributions. *Mol Biol Evol.*, 26:473–480, 2009.

[33] J Hein, C Wiuf, B Knudsen, MB Møller, and G Wibling. Statistical alignment: computational properties, homology testing and goodness-of-fit. *J Mol Biol.*, 302:265–279, 2000.

[34] MJ Bishop and EA Thompson. Maximum likelihood alignment of DNA sequences. *J Mol Biol.*, 190:159–165, 1986.

[35] JL Thorne, H Kishino, and J Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Biol.*, 33:114–124, 1991.

[36] JL Thorne, H Kishino, and J Felsenstein. Inching toward reality: An improved likelihood model of sequence evolution. *J Mol Biol.*, 34:3–16, 1992.

[37] B Knudsen and MM Miyamoto. Sequence alignments and pair hidden Markov models using evolutionary history. *J Mol Biol.*, 333:453–460, 2003.

[38] I Miklós and Z Toroczkai. An improved model for statistical alignment. *WABI 2001.*, LNCS 2149:1–10, 2001.

[39] GH Gonnet, MA Cohen, and SA Benner. Exhaustive matching of the entire protein sequence database. *Science.*, 256:1443–1445, 1992.

[40] SA Benner, MA Cohen, and GH Gonnet. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J Mol Biol.*, 229:1065–1082, 1993.

[41] X Gu and WH Li. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J Mol Evol.*, 40:464–473, 1995.

[42] Z Zhang and M Gerstein. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.*, 31:5338–5348, 2003.

[43] MS Chang and SA Benner. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. *J Mol Biol.*, 341:617–631, 2004.

[44] K Yamane, K Yano, and T Kawahara. Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. *DNA Res.*, 13:197–204, 2006.

[45] Y Fan, W Wang, G Ma, L Liang, Q Shi, and S Tao. Patterns of insertion and deletion in mammalian genomes. *Curr Genomics.*, 8:370–378, 2007.

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

[47] K Mizuguchi, CM Dean, TL Blundell, and JP Overington. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, 7:2469–2471, 1998.

[48] RA Cartwright. DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics.*, 21 (Suppl. 3):iii31–iii38, 2005.

[49] W Fletcher and Z Yang. INDELible: A flexible simulator of biological sequence evolution. *Mol Biol Evol.*, 26:1879–1888, 2009.

[50] CL Strope, K Abel, SD Scott, and EN Moriyama. Biological sequence simulation for testing complex evolutionary hypothesis: indel-Seq-Gen version 2.0. *Mol Biol Evol.*, 26:2581–2593, 2009.

[51] G Lunter, I Miklós, A Drummond, J Jensen, and J Hein. Baysian phylogenetic inference under a statistical indel model. *Lecture Notes in Bioinformatics.*, 2812:228–244, 2003.

[52] G Lunter, I Miklós, A Drummond, JL Jensen, and J Hein. Baysian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics.*, 6:83, 2005.

[53] E Levy Karin, H Ashkenazy, J Hein, and T Pupko. A simulation-based approach to statistical alignment. *Syst Biol.*, 68:252–266, 2019.

[54] N De Maio. The cumulative indel model: fast and accurate statistical evolutionary alignment. *Syst Biol.*, 2020. (available as E-pub).

[55] I Holmes. Application of indel evolution by differential calculus of finite state automata. available in bioRxiv with doi: 10.1101/2020.06.29.178764., 2020.

[56] K Ezawa. Alingment Neighborhood EXplorer (ANEX): First attempt to apply *genuine* sequence evolution model with realistic insertions/deletions to Multiple Sequence Alignment reconstruction problem. preprint (KEZW_BI_ME00006.anex.pdf) available at: https://www.bioinformatics.org/ftp/pub/anex/Documents/Preprints/., 2020.

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

[58] J Kim and S Sinha. Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. *Bioinformatics.*, 23:289–297, 2007.

[59] J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.*, 17:368–376, 1981.

[60] J. Felsenstein. *Inferring Phylogenetics*. Sinauer, Sunderland, Massachusetts, 2004.

[61] Z. Yang. *Computational Molecular Evolution*. Oxford Univ. Press, Oxford, UK, 2006.

[62] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numecial Recipes in C: The Art of Scientific Computing (2nd Ed.)*. Cambridge Univ. Press, Cambridge, UK, 1992.

[63] I Holmes and WJ Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics.*, 17:803–820, 2001.

[64] I Holmes. Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics.*, 19(Suppl 1):i147–i157, 2003.

[65] C Burge and S Karlin. Prediction of complete gene structures in human genomic dna. *J Mol Biol.*, 268:78–94, 1997.

[66] IH Holmes. Solving the master equation for indels. *BMC Bioinformatics.*, 18:255, 2017.

[67] K Ezawa. Review of the Commentary: "Solving the master equation for indels". Posted on PubPeer (https://pubpeer.com)., 2017.

[68] M Morgante, E De Paoli, and S Radovic. Transposable elements and the plant pangenomics. *Curr Opin Plant Biol.*, 10:1039–1049, 2007.

[69] D Chalopin, M Navile, F Plard, D Galiana, and JN Vollff. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol Evol.*, 7:567–580, 2015.

[70] J Ellegren. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.*, 5:435–445, 2004.

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

[72] F Famoud, M Schwartz, and J Bruck. Estimation of duplication history under a stochastic model for tandem repeats. *BMC Bioinformatics.*, 20:64, 2019.

[73] K Ezawa, D Graur, and G Landan. Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part IV: incorporation of substitutions and other mutations. available in bioRxiv with doi: 10.1101/023622., 2015.

[74] K Ezawa. Substitutional Residue-Difference Map (SRD Map) to help locate mis-alignments in Multiple Sequence Alignment (MSA): toward Artificial-Intelilgence-assisted probability distribution of alternative MSAs. preprint (KEZW_BI_ME00007.srdmap.pdf) available at: https://www.bioinformatics.org/ftp/pub/anex/Documents/Preprints/., 2020.

[75] K Ezawa. (Approximate) Solutions to some technical issues on alignment probability calculation under *genuine* sequence evolution model with realistic insertions/deletions. (in preparation.).