Ezawa *BMC Bioinformatics* (2016) 17:304
DOI 10.1186/s12859-016-1105-7

BMC Bioinformatics

RESEARCH ARTICLE Open Access

General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable?

Kiyoshi Ezawa[1,2]

Abstract

Background: Insertions and deletions (indels) account for more nucleotide differences between two related DNA sequences than substitutions do, and thus it is imperative to develop a stochastic evolutionary model that enables us to reliably calculate the probability of the sequence evolution through indel processes. Recently, indel probabilistic models are mostly based on either hidden Markov models (HMMs) or transducer theories, both of which give the indel component of the probability of a given sequence alignment as a product of either probabilities of column-to-column transitions or block-wise contributions along the alignment. However, it is not *a priori* clear how these models are related with any genuine stochastic evolutionary model, which describes the stochastic evolution of an *entire* sequence along the time-axis. Moreover, currently none of these models can fully accommodate biologically realistic features, such as overlapping indels, power-law indel-length distributions, and indel rate variation across regions.

Results: Here, we theoretically dissect the *ab initio* calculation of the probability of a given sequence alignment under a genuine stochastic evolutionary model, more specifically, a general continuous-time Markov model of the evolution of an entire sequence via insertions and deletions. Our model is a simple extension of the general "substitution/insertion/deletion (SID) model". Using the operator representation of indels and the technique of time-dependent perturbation theory, we express the *ab initio* probability as a summation over all alignment-consistent indel histories. Exploiting the equivalence relations between different indel histories, we find a "sufficient and nearly necessary" set of conditions under which the probability can be factorized into the product of an overall factor and the contributions from regions separated by gapless columns of the alignment, thus providing a sort of generalized HMM. The conditions distinguish evolutionary models with factorable alignment probabilities from those without ones. The former category includes the "long indel" model (a space-homogeneous SID model) and the model used by Dawg, a genuine sequence evolution simulator.

Conclusions: With intuitive clarity and mathematical preciseness, our theoretical formulation will help further advance the *ab initio* calculation of alignment probabilities under biologically realistic models of sequence evolution via indels.

Keywords: Stochastic evolutionary model, Insertion/deletion (indel), Sequence alignment probability, Factorability, Biological realism, Power-law distribution, Rate variation, Non-equilibrium evolution

Correspondence: kezawa.ezawa3@gmail.com; kezawa@bio.kyutech.ac.jp
[1]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka 820-8502, Japan
[2]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001, USA

Background

The evolution of DNA, RNA and protein sequences is driven by mutations such as base substitutions, insertions and deletions (indels), recombination and other genomic rearrangements (*e.g.*, [1–3]). Some recent comparative genomic analyses revealed that indels account for more base differences between closely related genomes than base substitutions do (*e.g.*, [4–7]). It is therefore imperative to develop a method to reliably calculate the probability of sequence evolution via mutations including indels. Since the groundbreaking works by Bishop and Thompson [8] and by Thorne, Kishino and Felsenstein [9], many studies have been made to calculate the probabilities of pairwise alignments (PWAs) and multiple sequence alignments (MSAs) under probabilistic models aiming to incorporate the effects of indels. And the methods have greatly improved in terms of the computational efficiency and the scope of application (reviewed, *e.g.*, in [10–12]). Most of these studies are based on hidden Markov models (HMMs) (*e.g.*, [13]) or transducer theories (*e.g.*, [14]). Even today, the studies on these methods are steadily advancing (*e.g.*, [15, 16]), and it seems that their mathematical and algorithmic bases are about to be established.

However, it is important to remember that these desirable properties *alone* are not sufficient for a model in natural science to be satisfactory. In addition to having the mathematical (or algorithmic) soundness, a satisfactory model must also approximate well, or at least decently, the real phenomena it is intended to describe. In the case of an indel probabilistic model, there are two key elements for this requisite: one is the evolutionary consistency of the model, and the other is the model's flexibility to accommodate various biologically realistic features, *i.e.*, the biological realism, of indels.

Let us first explain the evolutionary consistency. In natural sequence evolution, the probability (density) of an indel process must be given *vertically*, as a multiplicative accumulation of the probabilities of transitions between states of an *entire* sequence, each from one time point to the next one, along the time axis (or along a phylogenetic tree when dealing with MSAs) (Fig. 1a). If a probabilistic model gives the probability density of each evolutionary process according to this natural design, we call it a "genuine stochastic evolutionary model", or simply an "evolutionary model" for short. And we will consider an indel probabilistic model as "evolutionarily consistent", if its alignment probabilities can be derived directly from an evolutionary model, even if it does not appear to follow the aforementioned natural design. By definition, each of HMMs and transducers calculates (the indel component of) an alignment probability *horizontally*, as a product of either inter-column transition probabilities or block-wise contributions (Fig. 1c). Therefore, it is *a priori* unclear whether each HMM or transducer is evolutionarily consistent or not, or, if it is, how (Fig. 1). It would be worth mentioning that some models were indeed derived explicitly from some sorts of evolutionary models (*e.g.*, [9, 17, 18]). Unfortunately, all such studies in the past imposed some unnatural assumptions, such as the prohibition of overlapping indels and the restriction of deletions to single-base ones. Such assumptions were necessary for an alignment probability to be trivially factorable, at least as a product of block-wise contributions. Thus, they are unsatisfactory from the viewpoint of the second key element, *i.e.*, the biological realism.

Regarding the biological realism, past empirical studies revealed some properties of real indels, in addition to their possibilities to affect multiple contiguous residues at a time and to overlap others. Among the most important would be the studies that showed power-law distributions of indel lengths (see, *e.g.*, [19] and references therein). On the contrary, standard HMMs and transducers can usually implement geometric distributions of indel lengths, or at best mixed geometric distributions (*e.g.*, [20]), but cannot implement the power-law distributions themselves. But some generalized HMMs (or transducers) (*e.g.*, [21, 22]) can incorporate power-law indel length distributions. For example, the HMM of Kim and Sinha [22] is quite flexible, and it can incorporate the power-law distributions and also do away with the commonly imposed time-reversibility. As discussed *e.g.*, in [21] and [23], there is no biological reason for imposing the time reversibility, and they were usually imposed to reduce the computational time. In this sense, the HMM of Kim and Sinha is two steps closer to the biological reality than the standard HMMs (and transducers). Unfortunately, similarly to the standard HMMs and transducers, their HMM is not evolutionarily consistent and thus cannot correctly handle overlapping indels along the same branch, though they can handle overlapping indels along different branches. Another possibly important biologically realistic feature is the indel rate variation across sites (or regions) (*e.g.*, [24]), due to selection and the mutational predispositions (caused, *e.g.*, by the sequence or epigenetic contexts). Thus far, attempts to incorporate this feature have been rare (*e.g.*, [25]), and most studies have handled space-homogeneous models, whose indel rates are homogeneous along the sequence.

As far as we know, except the models implemented in some genuine sequence evolution simulators (*e.g.*, [26–28]), there is only one class of genuine stochastic evolutionary models discussed thus far that is also considerably biologically realistic, which are the "substitution/insertion/deletion (SID) models" proposed by Miklós et al. [21]. The SID models in general do not impose the aforementioned unnatural restrictions on indels. Moreover, the general SID

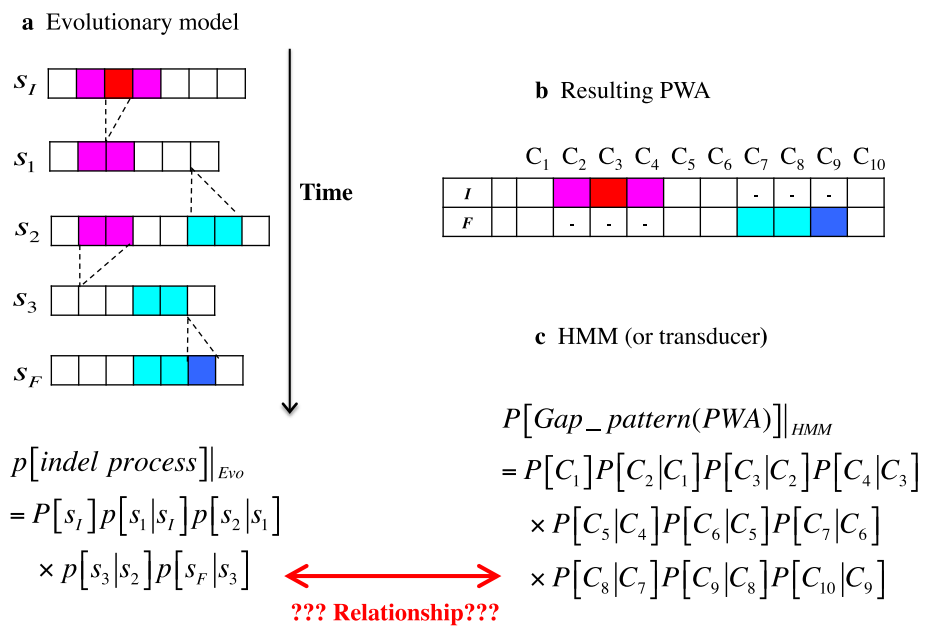

Fig. 1 Genuine stochastic evolutionary model vs. HMM (or transducer). **a** Probability density calculation via a genuine stochastic evolutionary model. Each sequence state is represented as an array of sites (boxes). Sites to be deleted are shaded in red or magenta. Inserted sites are shaded in blue or cyan. The s_I and s_F, respectively, denote the initial and final states. The s_v ($v = 1, 2, 3$) is an intermediate state. (The "P[…]" denotes a probability, the "p[…]" denotes a probability density.) **b** A pairwise alignment (PWA) between the initial (*I*) and final (*F*) states resulted from the indel process in panel **a**. The C_i ($i = 1, \ldots, 10$) labels the alignment column below. **c** Probability calculation via a HMM (or a transducer). It is a priori unclear whether or how the methods in panels **a** and **c** are related with each other. For clarity, residue states and substitutions were omitted. (Note that the equation in panel **a** is merely a rough expression to give a broad idea on the issue. Rigorous expressions will be given in *Results and discussion*.) Panels **a** and **b** of this figure were adapted from panels B and F of Fig. 1 of [32]

model can accommodate any indel length distributions, and also some indel rate variations across sites (albeit through the residue state context alone). Unfortunately, however, we have not seen any further theoretical development of the general SID model since it was proposed. Instead, Miklós et al. developed the "long indel" model [21], which is a space-homogeneous, time-reversible SID model. (More precisely, the insertion rate depends on the inserted sequence only through the product of the frequencies of its constituent residues. As mentioned above, the time-reversibility was introduced just for computational convenience, and it could be dispensed with if desired.) And they gave a verbal justification that the probability of a PWA under the "long indel" model can be calculated via a generalized HMM, as a product of contributions from "chop-zones" delimited by gapless columns. In the present viewpoint, this is the most satisfactory HMM that we know was used for actual sequence analyses, because it satisfies both the evolutionary consistency and the biological realism to some degree. Nevertheless, their justification, although plausible, has two problems. First, it is unclear exactly how their HMM is related to the *ab initio* probabilities of evolutionary (especially indel) processes under their evolutionary model. And second, their justification takes advantage of the space-homogeneity of the model, which makes it unclear how their HMM can be extended to be space-heterogeneous while keeping the evolutionary consistency. To solve these two problems, we need at least to get back to their origin, *i.e.*, the general SID model. It seems, however, that this model has never been theoretically dissected thus far, possibly due to the lack of mathematical or conceptual tools to handle it easily.

In this study, we examine a general continuous-time Markov model of the evolution of an *entire* sequence via insertions, deletions and substitutions. This model could be regarded as an extension of the general SID model, in the sense that it allows explicit (or inherent) rate variation across sites, not only due to residue state contexts. Such rate variation could be regarded as the effect of, *e.g.*, the epigenetic context and/or the context within the 3D structure of the protein product (*e.g.*, [25]).[1] To theoretically dissect the *ab initio* calculation of alignment probabilities under this model, we introduce some useful tools. Among the most important would be the operator representation of mutations, namely, insertions, deletions and substitutions. This enabled us to shift our focus from the trajectory of sequence states, which played a central role in [21], to the history of mutations (especially indels). Moreover, the operator representation enabled to algebraically define the equivalence relationships between two

different series of indels. They, in cooperation with the focus shift, enabled to define the local history set (LHS) equivalence classes. These equivalence classes play an essential role when deriving the "sufficient and nearly necessary" set of conditions under which alignment probabilities are indeed factorable, thus providing a sort of generalized HMM. We also adapt techniques from the time-dependent perturbation expansion in quantum mechanics [29, 30], expanding each alignment probability into a summation over contributing mutational histories with different numbers of indels. It should be noted, however, that we formally deal with all terms in the expansion.[2] Thus, at least formally, the probabilities we deal with are exact solutions of the model's defining equation. For clarity, we will focus on insertions/deletions in the bulk of the manuscript. However, we can also incorporate substitutions; see, *e.g.*, [31] for more details.

This paper describes the backbone of our study (more extensively recorded in an unpublished paper [32]) to give the theoretical basis of our *ab initio* probability calculation under the general continuous-time Markov model of indels. Peripheral topics surrounding the study can be found in [32].[3] Throughout the paper, we suppose that each probability is calculated under a given evolutionary model setting, including the phylogenetic tree of the sequences. In section R1 of Results and discussion, we briefly review the most general form of the SID model [21]. Then, in section R2, we introduce two important tools, namely, the ancestry index and the operator representation of mutations including indels. Using the results of sections R1 and R2, we define our general continuous-time Markov model in section R3, and formally give the general solution to its defining equation in terms of the operator representation. In section R4, we formally express the *ab initio* probability of a given PWA in a perturbation expansion. Then, using the concept of the LHS equivalence classes defined in section R5, we derive in section R6 the conditions under which the PWA probability is factorable. In section R7, the derivation is extended to the probability of a given MSA. In section S8, some examples are given to illustrate models with factorable and non-factorable alignment probabilities. The former category includes the indel evolutionary model of Dawg [26] and the "long indel" model [21], among others. In section R9, we discuss the merits, possible uses and extensions, as well as some outstanding issues, of the results in this study. In Table 1, we summarize the key concepts and results of this paper, mainly for those who want its gist quickly. Likewise, Table S1 (in Additional file 1) summarizes mathematical symbols used commonly in this paper, to facilitate the readers' cruise through the equations. Supplementary methods (in Additional file 1) and Supplementary appendix (in Additional file 2) give detailed derivations of some important results. The former is more essential and accessible to a wider audience; the latter is for those who are interested in further mathematical details.

We end this section with two notes. First, in this paper, the term "an evolutionary (or indel) process" means a series of successive mutation (or indel) events with both the order and the specific timing specified, and the term "an evolutionary (or indel) history" means a series of successive events with only the order specified. This usage should conform to the common practice in this field. Second, we will describe the results in the bra-ket notation, similar to that in quantum mechanics [29, 33]. However, those who are unfamiliar with the notation need not worry about it. Our formulation via the bra-ket notation can be proven to be equivalent to the standard formulation of the continuous-time Markov model via the vector-matrix notation. (We refer the interested readers to Supplementary appendix SA-1 in Additional file 2.) Therefore, if desired, the symbols of a bra ($\langle x|$), a ket ($|y\rangle$), and an operator ($\hat{O}$) could be regarded simply as convenient reminders of a row vector, a column vector, and a matrix, respectively.[4]

Results and discussion

The key concepts and results proposed/obtained in this paper are summarized in Table 1. Readers can use the table to quickly grasp an overview of this paper, as well as to easily locate what they look for. Also, most mathematical symbols are briefly explained in Table S1 in Additional file 1.

R1. Brief review of general SID model

Miklós et al. [21] proposed a class of evolutionary models, which they called the "substitution/insertion/deletion (SID) models". They are continuous-time Markov models defined on the space of strings (*i.e.*, sequences) of any lengths, each of which consists of letters (*i.e.*, residues, such as bases or amino acids) from a given alphabet (denoted as Ω here). Following [21], their state space will be denoted as: $\Omega^* \equiv \cup_{L=0}^{\infty} \Omega^L$, whose component, Ω^L, is the space of all sequences of length L. If desired, a sequence state, $s \in \Omega^L$, could be represented as: $s = [\omega_1, \omega_2, \ldots, \omega_L]$ (with $\omega_x \in \Omega$ for $x = 1, 2, \ldots, L$) (see Fig. 2a). In this model, mutations are defined as transitions from a sequence state to another, and their instantaneous rates can be given via the following "rate grammar" they proposed:

$$\text{Substitution: } s = s_L \omega s_R \xrightarrow{\rho_S(s_L, \omega, \omega', s_R)} s' = s_L \omega' s_R; \quad \text{(R1.1)}$$

$$\text{Insertion: } s = s_L s_R \xrightarrow{\rho_I(s_L, s_I, s_R)} s' = s_L s_I s_R; \quad \text{(R1.2)}$$

$$\text{Deletion: } s = s_L s_D s_R \xrightarrow{\rho_D(s_L, s_D, s_R)} s' = s_L s_R. \quad \text{(R1.3)}$$

Table 1 Key concepts and results in this paper

Concept/result	Description	Main location
Ancestry index	An ancestry index is assigned to each site. Sharing of an ancestry index among sites indicates the sites' mutual homology. As a fringe benefit, the indices enable the mutation rates to vary across regions (or sites) beyond the mere dependence on the residue state of the sequence.	Section R2 (1st and 2nd paragraphs), Fig. 2
Operator representation of mutations	This enables the intuitively clear and yet mathematically precise description of mutations, especially insertions/deletions, on sequence states. This is a core tool in our *ab initio* theoretical formulation of the genuine stochastic evolutionary model.	Section R2 (3rd paragraph), **Fig. 3**
Rate operator	An operator version of the rate matrix, which specifies the rates of the instantaneous transitions between the states in our evolutionary model. In other words, the rate operator describes the instantaneous stochastic effects of single mutations on a given sequence state.	Section R3, Eqs. (R3.1-R3.9) (full mutational model), **Eqs. (R3.2**,**R3.6**,**R3.11**-**R3.15)** (indel model)
Finite-time transition operator	An operator version of the finite-time transition matrix, each element of which gives the probability of transition from a state to another after a finite time-lapse. This results from the cumulative effects of the rate operator during a finite time-interval.	Section R3, **Eq. (R3.17)**, Eq. (R3.18)
Defining equations (differential)	1st-order time differential equations (forward and backward) that define our indel evolutionary model. They are operator versions of the standard defining equations of a continuous-time Markov model.	Section R3, Eqs. (R3.19,R3.21) (forward), Eqs. (R3.20,R3.21) (backward)
Defining equations (integral)	Two integral equations (forward and backward) that are equivalent to the aforementioned differential equations defining our indel evolutionary model. They play an essential role when deriving the perturbation expansion of the finite-time transition operator.	Section R4, **Eq. (R4.4)** (forward), **Eq. (R4.5)** (backward)
Perturbation expansion (transition operator)	The perturbation expansion of the finite-time transition operator. It was derived in an intuitively clear yet mathematically precise manner, by using the aforementioned defining integral equations.	Section R4, **Eqs. (R4.6,R4.7)**
Perturbation expansion (*ab initio* PWA probability)	The perturbation expansion of the *ab initio* probability of a given PWA, conditioned on the ancestral sequence state, under a given model setting.	Section R4, Eq. (R4.8) or Eq. (R4.9)
Binary equivalence relation	An equivalence relation between the products of two indel operators each. The relations play key roles when defining LHS equivalence classes.	Section R5, **Eqs. (R5.2a-R5.2d)**
Local-history-set (LHS) equivalence class	An equivalence class consisting of global indel histories that share all local history components. The classes play an essential role when proving the factorability of a given PWA probability.	Section R5, **below Eq. (R5.4)**, (*e.g.*, Fig. 5)
Factorability (*ab initio* PWA probability)	We proved that, under **conditions (i) and (ii)** (below Eq. (R6.4)), the *ab initio* probability of a given PWA is factorable into the product of an overall factor and contributions from local PWAs.	Section R6, **Eqs. (R6.7,R6.8)**, (see also Eqs. (R6.2,R6.3,R6.4))
Perturbation expansion (*ab initio* MSA probability)	The "perturbation expansion" of the *ab initio* probability of a given MSA, under a given model setting including a given phylogenetic tree.	Section R7, Eqs. (R7.2,R7.3,R7.4)
Factorability (*ab initio* MSA probability)	We proved that, under **conditions (i), (ii)** (below Eq. (R6.4) and **(iii)** (**Eq. (R7.8)**), the *ab initio* probability of a given MSA is factorable into the product of an overall factor and contributions from local MSAs.	Section R7, **Eq. (R7.9)**
Totally space-homogeneous model	Such a model gives factorable PWA probabilities, because the exit rate is an affine function of the sequence length (regardless of whether indel rates are time-dependent or not). The indel model of Dawg [26] and the "long indel" model [21] belong to this class.	Subsection R8-1, Eqs. (R8-1.1,R8-1.2), **Eqs. (R8-1.3,R8-1.4)**

Table 1 Key concepts and results in this paper *(Continued)*

Equivalence (with caveat) of the "chop-zone" method and our *ab initio* method	We showed that the "chop-zone" method in [21], adapted to calculate the probability of a given LHS equivalence class, is equivalent to our *ab initio* method, at least if the indel model is spatiotemporally homogeneous.	Subsection R8-1, **Supplementary appendix SA-3**
Model with simple insertion rate variation	If the deletion rates are space-homogeneous and the insertion rates depend only on the insertions' flanking sites, the PWA probabilities are still factorable.	Subsection R8-1, Eq. (R8-1.5)
Space-homogenous model flanked by essential sites	This kind of model is a simplest example of the indel model whose *ab initio* PWA probabilities are **non-factorable**.	Subsection R8-2, **Eqs. (R8-2.1,R8-2.3)**
Degree of non-factorability	The "difference of exit-rate differences" (Eq. (R8-2.4)) could measure the "degree of non-factorability."	Subsection R8-2, Eq. (R8-2.4)
Space-heterogeneous model with factorable PWA probability	We found that a class of indel models with rate-heterogeneity across regions (Eqs. (R8-3.1,R8-3.2)) have partially factorable PWA probabilities.	Subsection R8-3, **Eqs. (R8-3.1,R8-3.2)**, Eqs. (R8-3.3,R8-3.4,R8-3.5), Figure S3

NOTE: Especially important things are in boldface

Here, $\rho_m(...)$ with $m = S, I\,and\,D$ denote the rates of the substitution, the insertion, and the deletion, respectively, possibly depending on the arguments in the parentheses. In each of the above rules, s and s', respectively, denote the sequence states before and after the mutation. The symbols s_L and s_R denote the subsequences flanking the mutated portion from the left and from the right, respectively.[5] These SID models equipped with this "rate grammar" are genuine stochastic evolutionary models, and thus do not usually impose unnatural restrictions on the mutations (except possibly through restrictions on mutation rates). And the most general SID model can accommodate quite general mutation rates, including indel length distributions, by allowing their dependence on the sequence states before and after the mutation.[6] As far as we know, however, this most general SID model was not theoretically examined further (at least thus far), maybe because adequate mathematical or conceptual tools were not devised and because of some other reasons (mentioned below). In the following sections, we will provide such tools, which in turn will help theoretically dissect an extended version of the most general SID model.

R2. Ancestry indices and operator representation of mutations

First, we slightly extend the framework of the SID models, by assigning an ancestry to each site, which is a unit position in the sequence that accommodates a single residue. Hereafter, we will consider that it is the sites, instead of the residues, that are inserted/deleted. For example, the above example sequence state, $s = [\omega_1, \omega_2, ..., \omega_L](\in\Omega^L)$, can be extended as: $\check{s} = [(\upsilon_1, \omega_1), (\upsilon_2, \omega_2), ..., (\upsilon_L, \omega_L)]\left(\in(\Upsilon \times \Omega)^L\right)$ (Fig. 2b). Here, $\upsilon_x(\in\Upsilon)$ is the ancestry index assigned to the x-th site of the sequence (with $x = 1, 2, ..., L$). Alternatively, the extended sequence state could also be represented as: $\check{s} = \left(\vec{\upsilon}, \vec{\omega}\right)$,[7] where $\vec{\upsilon} = [\upsilon_1, \upsilon_2..., \upsilon_L]\left(\in\Upsilon^L\right)$ is an array of ancestry indices assigned to the sites, and $\vec{\omega} = [\omega_1, \omega_2..., \omega_L]\left(\in\Omega^L\right)$ is an array of residue states that fill in the sites. (Note that $\vec{\omega}$ corresponds to the sequence state (s) in the SID models (in section R1)). The ancestry indices follow a number of rules: (i) different sites in the same sequence always have different ancestry indices; (ii) the ancestry index of a site remain unchanged as long as the site exists; and (iii)

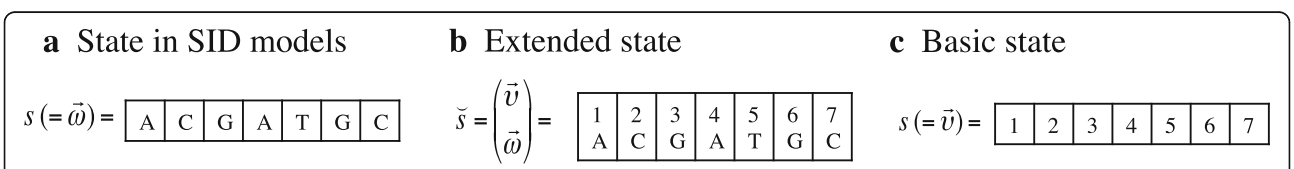

Fig. 2 Sequence states. **a** A sequence state in the SID models [21]. Each site (cell) is assigned a residue state (A, T, G or C). **b** The corresponding extended sequence state in our evolutionary model. Each site is assigned an ancestry index (number) in addition to a residue state. **c** The corresponding basic sequence state. Each site is assigned an ancestry index alone. The $\vec{\omega}$ represents the set of residue states assigned to all sites. The $\vec{\upsilon}$ represents the set of ancestry indices assigned to all sites. (Note that the identical symbols (*s*'s) used in panels **a** and **c** represent different types of states (of the same sequence))

every time when an insertion takes place, new ancestry indices are assigned to the newly inserted sites. Other than these rules, the assignment of the indices is arbitrary. Especially, their values themselves are not so important. The most essential thing is whether two sites of different sequences share the same ancestry index or not; if so, the sites are mutually homologous (actually orthologous unless duplications are considered). Another important thing is the spatial relationship of the site having each ancestry with other sites, especially preserved ancestral sites (PASs, explained shortly). For the space of ancestry indices, Υ, we will *tentatively* use the set of all positive integers ($N_1 \equiv \{1, 2, 3, \ldots\}$), although there should be a more appropriate mathematical entity.[8] Because of the rules imposed above, the space of the extended sequence states (denoted as $\check{S}^{II}$ in [31]) is included in but never equal to $\{\Upsilon \times \Omega\}^* = \cup_{L=0}^{\infty} \{\Upsilon \times \Omega\}^L$.

We devised the ancestry indices to facilitate the description of indel histories by keeping track of the evolutionary course of each site. For example, consider that a sequence whose initial state had the ancestry indices $\vec{v}_I = [1, 2, 3, 4, 5, 6, 7]$ evolved into the final state with $\vec{v}_F = [1, 5, 6, 8, 9, A, 7]$. (Here the "A" is an abbreviation of 10.) Then, we can immediately infer what happened during its evolution by aligning the two sequences (represented by the arrays of ancestry indices):

$$\begin{array}{ll} \vec{v}_I & 123456{-}{-}{-}7. \\ \vec{v}_F & 1{-}{-}{-}5689A7 \end{array} \qquad \text{(R2.1)}$$

This alignment tells that the sites with ancestries 2, 3 and 4 were deleted and that the sites with ancestries 8, 9 and A were inserted. We can also see that the sites with 1, 5, 6 and 7 were preserved during the evolution. We will henceforth refer to such sites as "preserved ancestral sites (PASs)". The PASs indicate that no indels occurred at or through the sites during the evolution under consideration. Thus, they can be used to narrow down the possible indel histories that might have resulted in the pairwise alignment (PWA) (as argued, *e.g.*, in [21]). A fringe benefit of the ancestry indices is that they enable the mutation rates to vary beyond the dependence on the residue states (section R3). Hereafter, we refer to an array of ancestry indices (like $\vec{v}$) as a "basic sequence state" (abbreviated as a "sequence state" or a "basic state"), which is a backbone to be fleshed out by the residue states (like $\vec{\omega}$) to give the extended sequence state (like $\check{s}$). Hereafter, the basic sequence state will often be denoted, *e.g.*, as s (Fig. 2c). (Henceforth, symbols like s will never denote a residue state (like $\vec{\omega}$), which was a sequence state in section R1). And $S^{II} (\subset \Upsilon^* = \cup_{L=0}^{\infty} \Upsilon^L)$ denotes the space of the basic states.

Another, probably more important, tool we introduce here is the operator representation of mutations. When considering the evolutionary processes (or histories), we symbolically represent each (extended) sequence state as a bra-vector, like $\langle \check{s}_I |$, which could be regarded as an abstract extension of a row vector in the normal representation of the continuous-time Markov model. Then, each mutation of a sequence can be represented as a linear operator (regarded as an abstract extension of a matrix) acting on a bra-vector (Fig. 3). Operator $\hat{M}_S(x, \omega \mapsto \omega')$ denotes the substitution of the residue to $\omega' (\neq \omega)$ if it was ω at the x th site (or the null action otherwise) (Fig. 3a). Operator $\hat{M}_I(x, l)$ denotes the insertion of l sites between the x-th and $(x+1)$-th sites (Fig. 3b). The insertion operator *alone* does not determine the residue states of the inserted sites. It is the job of the "fill-in" operator, $\hat{F}\left(x, \delta\vec{\omega}'[l]\right)$, which fills in the l inserted sites with a new array of residues, $\delta\vec{\omega}'[l] \left(\in \Omega^l\right)$. (This "division of labor" facilitates the decoupling of the substitution component and the indel component of an alignment probability. See Appendix A1 of [31] for more details). Finally, operator $\hat{M}_D(x_B, x_E)$ denotes the deletion of the subsequence between (and including) the x_B-th and x_E-th sites in the sequence immediately before the event (Fig. 3c). The action of multiple successive mutation events can be expressed as a product of mutation operators on an initial (extended) sequence state. For example, the indel history illustrated in panels a and b of Fig. 4 can be represented as a series of indel events, $\left[\hat{M}_D(3,3), \hat{M}_I(5,2), \hat{M}_D(2,3), \hat{M}_I(5,1)\right]$, on the initial basic state s_I (given above). Then, the final result of this indel history is expressed as: $\langle s_I | \hat{M}_D(3,3) \hat{M}_I(5,2) \hat{M}_D(2,3) \hat{M}_I(5,1)$. Figure 4c shows the MSA among the initial, intermediate and final sequence states. Figure 4d shows the resulting PWA between the initial and final sequence states.

These new tools, the ancestry indices and the operator representation of mutations, will play essential roles in our theoretical development described below. In the SID models [21], each evolutionary process was expressed as a (time-recorded) trajectory of sequence states, each of which was represented by an array of residues (without ancestry assignments). In consequence, an instantaneous transition from a state to the next state was often expressed as a summation of multiple possible mutations. (For example, the transition from $\vec{\omega} = [A, A]$ to $\vec{\omega}' = [A]$ could result from either $\hat{M}_D(1,1)$ or $\hat{M}_D(2,2)$). Ancestry indices help avoid such ambiguous channels by

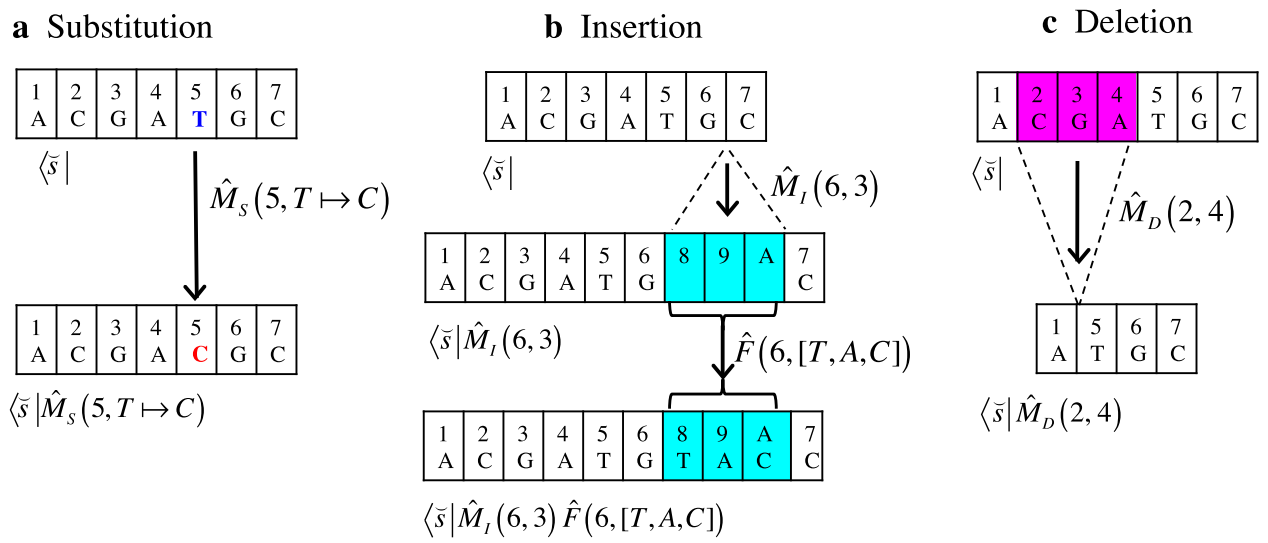

Fig. 3 Operator representation of mutations. **a** A substitution operator, $\hat{M}_S(5, T \mapsto C)$. The residues before and after the substitution are in boldface in blue and red, respectively. **b** An insertion operator, $\hat{M}_I(6, 3)$, and a fill-in operator, $\hat{F}(6, [T, A, C])$. The inserted sites are shaded in cyan. (Note that "A" at the top of the rightmost inserted column means the ancestry index of 10, not the residue state of A). **c** A deletion operator, $\hat{M}_D(2, 4)$. The sites to be deleted are shaded in magenta. In this figure, the extended sequence states were used for illustration. The bra-vector below each array denotes the state. The extended state, $\breve{s}$, is identical to that in Fig. 2b. Each vertical arrow indicates the action of the mutation operator beside it. Note that the first arguments of all operators and the second argument of the deletion operator specify positions along the sequence, and not ancestries (specified at the top of the sites)

uniquely defining each instantaneous state-to-state transition as an action of a single mutation. (In the above example, the former causes the transition from $\vec{v} = [1, 2]$ to $\vec{v}_{(1)}{}' = [2]$, and the latter yields the transition from $\vec{v}$ to $\vec{v}_{(2)}{}' = [1]$). This, in conjunction with the operator representation of mutations, enables us to shift the focus from the trajectory of sequence states to the history of mutations, especially indels. This shift of focus, as well

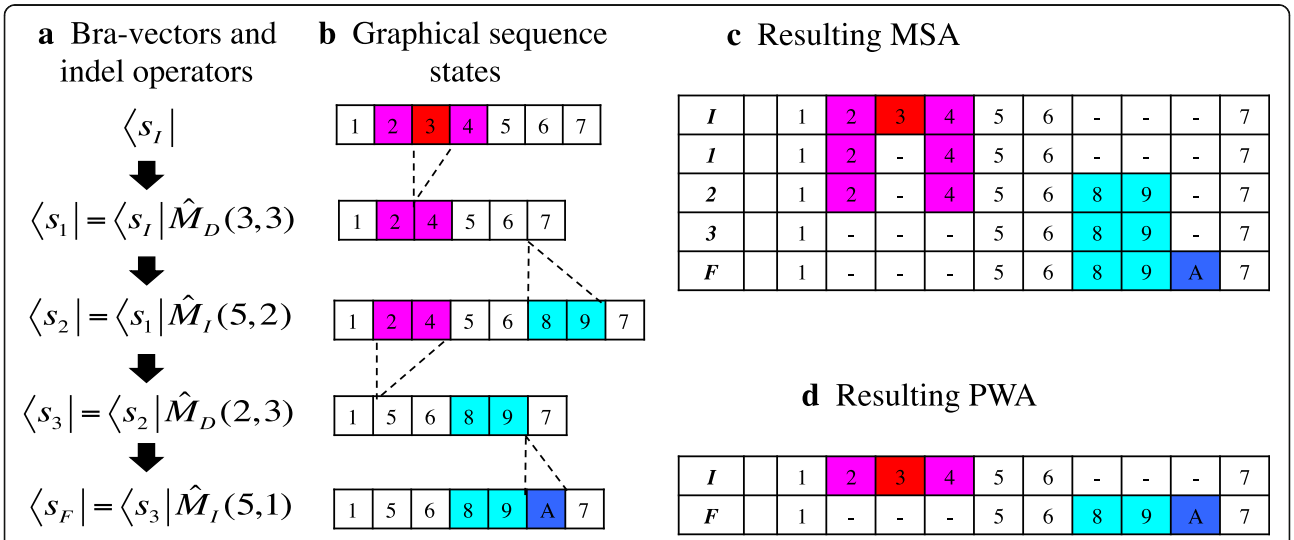

Fig. 4 Example indel history and resulting alignments. **a** An example indel history in terms of the bra-vectors of sequence states and indel operators. **b** The graphical illustration of the history using basic sequence states. Each sequence state in panel **a** is horizontally aligned with its graphical representation in panel **b**. **c** The resulting MSA among the sequence states that the indel history went through. **d** The resulting PWA between the initial and final sequences. In both **c** and **d**, the bold italicized characters in the leftmost column are the suffixes indicating the sequence states in panel **a**. In panels **b**, **c**, and **d**, the number in each site (cell) represents its ancestry, but not necessarily its position along the sequence. The 'A' in the final sequence abbreviates 10. The same shading scheme as in Fig. 1 is used. The figure was adapted from Fig. 1 of [32]

as the "equivalence relations" between the products of operators (section R5), facilitates the examination of the factorability of alignment probabilities, as we will see in sections R4-R6.

R3. Instantaneous transition (or rate) operator and finite-time transition operator

Now we are ready to define our evolutionary model, *i.e.*, the general continuous-time Markov model to describe the evolution of an entire sequence along a time axis. If desired, this could be done by extending the rate grammar, Eqs. (R1.1,2,3), so that each mutation rate depend on the entire extended sequence states (*i.e.*, not only on the residue states) immediately before and after the mutation. (We will also introduce the explicit time dependence. This may allow us to incorporate the effects of changes in the physiology, genomic contexts, external environments, etc. (See also section 2.4 of [32].)) However, to define the model more neatly, we will parametrize the mutation rates in coordination with the mutation operators defined in section R2. Although the parametrization is different from (the extended version of) Eqs. (R1.1,2,3), the two sets of mutation rates are equivalent. (To remember these differences in the parametrization, we will use the symbol $r_m(\ldots)$ ($m = S, I$ *and* D) instead of $\rho_m(\ldots)$.) Let $\check{s}$ be the extended sequence state immediately before the mutation. And let t be the time at which the mutation occurred. Then, $r_S(x, \omega \mapsto \omega'; \check{s}, t)$ is the rate of the substitution, $\hat{M}_S(x, \omega \mapsto \omega')$. It must be zero unless ω is at the x-th site of $\check{s}$. $r_I\left(x, l, \delta\vec{\omega}\,'[l]; \check{s}, t\right)$ is the rate of the insertion, $\hat{M}_I(x, l)$ (accompanied by $\hat{F}\left(x, \delta\vec{\omega}\,'[l]\right)$). And $r_D\left(x_B, x_E; \check{s}, t\right)$ is the rate of the deletion, $\hat{M}_D(x_B, x_E)$. Using these mutation rates that accompany the mutation operators, we can define our evolutionary model in a manner closer to the standard, by defining the instantaneous transition rate operator (or the "rate operator" for short), which is an analog of the instantaneous transition rate matrix in a continuous-time Markov model. Because the state space we are working in, $\check{S}^{II}$, is essentially infinite, we cannot give the explicit matrix expression of the rate operator on the entire state space. Nevertheless, the rate operator can be defined if we give its action on every state in $\check{S}^{II}$. Let $\hat{Q}^{SID}(t)$ denote our rate operator (at time t). It is convenient to decompose it as follows:

$$\hat{Q}^{SID}(t) = \hat{Q}^{S}(t) + \hat{Q}^{I}(t) + \hat{Q}^{D}(t), \tag{R3.1}$$

where $\hat{Q}^{m}(t)$ with $m = S, I$ *and* D, respectively, are the substitution, insertion, and deletion components of the rate operator. Each of the three components can be further decomposed as:

$$\hat{Q}^{m}(t) = \hat{Q}^{m}_{M}(t) + \hat{Q}^{m}_{X}(t) \quad (m = S, I \text{ and } D). \tag{R3.2}$$

Here, the "mutation part", $\hat{Q}^{m}_{M}(t)$, describes the transitions to different states via mutations of type m (=S, I *or* D). And the "exit rate part", $\hat{Q}^{m}_{X}(t)$, attenuates the state retention probability at the exit rate, $R^{m}_{X}(\check{s}, t)$, which is determined by the type m mutations. It guarantees that the state probabilities sum up to unity at any time. Specifically, the mutation parts are defined by the following actions on every state:

$$\langle \check{s}|\hat{Q}^{S}_{M}(t) \equiv \sum_{x=1}^{L(\check{s})} \sum_{\substack{\omega' \in \Omega, \\ \omega' \neq \omega_x(\check{s})}} \Big[r_S\left(x, \omega_x(\check{s}) \mapsto \omega'; \check{s}, t\right) \times \langle \check{s}|\hat{M}_S\left(x, \omega_x(\check{s}) \mapsto \omega'\right) \Big], \tag{R3.3}$$

$$\langle \check{s}|\hat{Q}^{I}_{M}(t) \equiv \sum_{x=0}^{L(\check{s})} \sum_{l=1}^{\infty} \sum_{\delta\vec{\omega}'[l] \in \Omega^{l}} \Big[r_I\left(x, l, \delta\vec{\omega}\,'[l]; \check{s}, t\right) \times \langle \check{s}|\hat{M}_I(x, l)\hat{F}\left(x, \delta\vec{\omega}\,'[l]\right) \Big], \tag{R3.4}$$

$$\langle \check{s}|\hat{Q}^{D}_{M}(t) \equiv \sum_{x_B=-\infty}^{L(\check{s})} \sum_{x_E=max\{1, x_B\}}^{+\infty} \Big[r_D\left(x_B, x_E; \check{s}, t\right) \times \langle \check{s}|\hat{M}_D\left(x_B, x_E\right) \Big]. \tag{R3.5}$$

Here, $L(\check{s})$ denotes the length (*i.e.*, the number of sites) of $\check{s}$, and $\omega_x(\check{s})$ denotes the residue at the x-th site of $\check{s}$. Figure 3 exemplifies the states on the right hand sides of Eqs. (R3.3-3.5). The terms with $x = 0$ and with $x = L(\check{s})$ in Eq. (R3.4) represent insertions at the left and right ends, respectively, of the sequence. How to deal with such insertions varies depending on various factors including the model setting, and can be implemented by adjusting the insertion rates accordingly (see, *e.g.*, [21, 26]). The terms with $x_B < 1$ or $x_E > L(\check{s})$ in Eq. (R3.5) represent the deletions of the subsequences sticking out of the subject sequence. These terms were included because the subject sequence is regarded as embedded in a "chromosome" with a virtually infinite length (*e.g.*, [21, 26]).[9] The exit rate parts are defined in nearly the same form:

$$\langle \check{s}|\hat{Q}^{m}_{X}(t) \equiv -R^{m}_{X}(\check{s}, t)\langle \check{s}| \quad (m = S, I \text{ and } D). \tag{R3.6}$$

They differ only in the exit rates:

$$R^{S}_{X}(\check{s}, t) = \sum_{x=1}^{L(\check{s})} \sum_{\substack{\omega' \in \Omega, \\ \omega' \neq \omega_x(\check{s})}} r_S(x, \omega_x(\check{s}) \mapsto \omega'; \check{s}, t), \tag{R3.7}$$

$$R_X^I(\check{s},t)=\sum_{x=0}^{L(\check{s})}\sum_{l=1}^{\infty}\sum_{\delta\vec{\omega}'[l]\in\Omega'} r_I\left(x,l,\delta\vec{\omega}\,'[l];\check{s},t\right), \tag{R3.8}$$

$$R_X^D(\check{s},t)=\sum_{x_B=-\infty}^{L(\check{s})}\sum_{x_E=max\{1,x_B\}}^{+\infty} r_D\left(x_B,x_E,\check{s},t\right). \tag{R3.9}$$

These equations, Eqs. (R3.1-R3.9), cooperatively define the instantaneous transition rate operator, $\hat{Q}^{SID}(t)$, and thus define our evolutionary model. If desired, $\hat{Q}^{SID}(t)$ could be decomposed as:

$$\hat{Q}^{SID}(t)=\hat{Q}_M^{SID}(t)+\hat{Q}_X^{SID}(t). \tag{R3.10}$$

Here $\hat{Q}_M^{SID}(t)\equiv\hat{Q}_M^S(t)+\hat{Q}_M^I(t)+\hat{Q}_M^D(t)$ is the collection of all mutational transition terms, and $\hat{Q}_X^{SID}(t)\equiv\hat{Q}_X^S(t)+\hat{Q}_X^I(t)+\hat{Q}_X^D(t)$ is the entire exit rate part, which attenuates the state retention probability at the total exit rate, $R_X^{SID}(\check{s},t)\equiv R_X^S(\check{s},t)+R_X^I(\check{s},t)+R_X^D(\check{s},t)$. Actually, the total mutational part, $\hat{Q}_M^{SID}(t)$, is equivalent to (the extended version of) the "instantaneous rate matrix" of the general SID model (defined by Eq. (1) in [21]), although the two expressions appear quite different from each other. The difference mainly stems from the parametrization and the state representation. We use our parametrization because we believe it to clarify the meaning of each term in the rate operator. And, in this section, the sequence state after the mutation (say, $\langle\check{s}'|$) was represented as the result of the mutation operator (say, $\hat{M}_m(\ldots)$) acting on the state before the mutation (say, $\langle\check{s}|$), like $\langle\check{s}'|=\langle\check{s}|\hat{M}_m(\ldots)$. This is legitimate because a mutation transfers a subject state uniquely to another state. This state representation will facilitate the unfolding of our theory below.

Thus far, we included substitutions (and residue states) mainly in order to discuss the relationship between our evolutionary model and the general SID model. However, our main interest here is in calculating the probability of the *skeleton* of a sequence alignment, composed only of the sites and gaps, which will then be filled in with residues to form a full alignment. Because such a skeleton can be created only through an evolutionary process of indels, we will hereafter omit the description of substitutions and the resulting changes in the residue state ($\vec{\omega}$). (This includes omitting the "fill-in" operators, $\hat{F}\left(x,\delta\vec{\omega}\,'[l]\right)$'s.) (Henceforth, the alignment skeleton will be called the "alignment" for short.) But we will retain the basic sequence state consisting of ancestry indices ($s=\vec{v}$). The rate heterogeneity will be realized only through the ancestry dependence. This would be almost sufficient for representing the dependence of the rates on, *e.g.*, the epigenetic or 3D structural context (*e.g.*, [25]). And this may also be able to approximate the dependence on some residue state contexts, especially in highly conserved regions. Now, the rate operator defined by Eqs. (R3.1-R3.9) is reduced as follows. (The reduced total rate operator will be denoted as $\hat{Q}^{ID}(t)$).

$$\hat{Q}^{ID}(t)=\hat{Q}^I(t)+\hat{Q}^D(t), \tag{R3.11}$$

$$\langle s|\hat{Q}_M^I(t)\equiv\sum_{x=0}^{L(s)}\sum_{l=1}^{\infty} r_I(x,l;s,t)\langle s|\hat{M}_I(x,l), \tag{R3.12}$$

$$\langle s|\hat{Q}_M^D(t)\equiv\sum_{x_B=-\infty}^{L(s)}\sum_{x_E=max\{1,x_B\}}^{+\infty} r_D(x_B,x_E;s,t)\langle s|\hat{M}_D(x_B,x_E), \tag{R3.13}$$

$$R_X^I(s,t)=\sum_{x=0}^{L(s)}\sum_{l=1}^{\infty} r_I(x,l;s,t), \tag{R3.14}$$

$$R_X^D(s,t)=\sum_{x_B=-\infty}^{L(s)}\sum_{x_E=max\{1,x_B\}}^{+\infty} r_D(x_B,x_E;s,t). \tag{R3.15}$$

Equations (R3.2,R3.6) remain unchanged except the exclusion of $m=S$ and the replacement of $\check{s}$ by s. In Eqs. (R3.12,R3.14), the insertion rates could be related to those in Eqs. (R3.4,R3.8) by the equation:

$$r_I(x,l;s,t)\equiv\sum_{\delta\vec{\omega}'[l]\in\Omega'} r_I\left(x,l,\delta\vec{\omega}'[l];s,t\right), \tag{R3.16}$$

where the $\vec{\omega}$-dependence of the right hand side was omitted.

Now we can calculate the operator that gives the finite-time transition probabilities between states. We will call it the "finite-time transition operator". Let $\hat{P}^{ID}\left(t,t'\right)$ be such an operator describing the state transition via indels alone from an initial time, t, to a final time, t' ($>t$). Operator $\hat{P}^{ID}\left(t,t'\right)$ is defined to give the following equations:

$$\langle s|\hat{P}^{ID}\left(t,t'\right)|s'\rangle=P\left[\left(s',t'\right)|(s,t)\right]\ for\quad {}^{\forall}(s,s')\in\left(S^{II}\right)^2. \tag{R3.17}$$

On the right hand side, $P[(s',t')|(s,t)]$ is the probability that the sequence is in state s' at time t' conditioned on that it was in state s at time t. On the left hand side, $|s'\rangle$ denotes a ket-vector (an abstract extension of a column vector), whose exclusive role here is to give "inner-products" with bra-vectors: $\langle s|\,s'\rangle=1$ (if $s=s'$), $=0$ (otherwise). From the evolutionary principle that our model must satisfy, or equivalently, from the fundamental properties (such as the Chapman-Kolmogorov equation) of the continuous-time Markov model, the finite-time transition operator must be given by the

multiplicative accumulation of the effects of the rate operators along the time axis. Thus, $\hat{P}^{ID}(t,t')$ is formally calculated as:

$$\begin{aligned}\hat{P}^{ID}(t,t') = &\lim_{N_P\to\infty}\left(\hat{I}+\tfrac{t'-t}{N_P}\hat{Q}^{ID}\left(t_1^{(N_P)}\right)\right)\\ &\times\left(\hat{I}+\tfrac{t'-t}{N_P}\hat{Q}^{ID}\left(t_2^{(N_P)}\right)\right)\cdots\left(\hat{I}+\tfrac{t'-t}{N_P}\hat{Q}^{ID}\left(t_{N_P}^{(N_P)}\right)\right)\\ &\equiv T\left\{exp\left(\int_t^{t'}d\tau\hat{Q}^{ID}(\tau)\right)\right\}.\end{aligned} \tag{R3.18}$$

In the middle of the equation, $\hat{I}$ is the identity operator (i.e., $\langle s|\hat{I}=\langle s|$ for every $s\in S^{II}$), and $t_k^{(N_P)}\equiv t+(k-\frac{1}{2})\frac{t'-t}{N_P}$. On the rightmost side, $T\{...\}$ denotes the meta-operator that rearranges the operators in each operator product term in the temporal order so that the earliest operator comes leftmost. Another way to give $\hat{P}^{ID}(t,t')$ is through the first-order time-differential equation. Again, from the fundamental properties of the continuous-time Markov model, or equivalently, from Eq. (R3.18), we can show that the operator satisfies the following differential equations:

$$\frac{\partial}{\partial t'}\hat{P}^{ID}(t,t') = \hat{P}^{ID}(t,t')\hat{Q}^{ID}(t'), \tag{R3.19}$$

$$\frac{\partial}{\partial t}\hat{P}^{ID}(t,t') = -\hat{Q}^{ID}(t)\hat{P}^{ID}(t,t'). \tag{R3.20}$$

Equation (R3.19) is the "forward equation", and Eq. (R3.20) is the "backward equation". And the evolutionary principle naturally includes the following equation:

$$\hat{P}^{ID}(t,t) = \hat{I}\quad for\quad {}^{\forall}t\in[t_I,t_F], \tag{R3.21}$$

where $[t_I, t_F]$ is the time interval in which the model is defined. This equation could be used as the initial condition for each of Eqs. (R3.19,R3.20). In the next section, we will obtain the solutions for Eqs. (R3.19, R3.20, R3.21) in a more tractable form than the defining solution, Eq. (R3.18).

R4. Perturbation expansion of finite-time transition operator and pairwise alignment probability: brief description

In time-dependent perturbation theory of quantum mechanics (*e.g.*, [29, 30]), the instantaneous time evolution operator (Hamiltonian $\hat{H}(t)$) is considered as a sum of two operators, $\hat{H}(t)=\hat{H}_0(t)+\hat{V}(t)$, and the time evolution of the system is described as if the system mostly evolves according to the well-solvable instantaneous time-evolution operator ($\hat{H}_0(t)$) and is occasionally perturbed by the "interaction" operator ($\hat{V}(t)$). We adapt such a technique of time-dependent perturbation expansion to our evolutionary model. Here, we briefly describe the results. For their detailed derivations, see Supplementary methods SM-1 in Additional file 1. We first re-express our rate operator as:

$$\hat{Q}^{ID}(t) = \hat{Q}_0^{ID}(t) + \hat{Q}_M^{ID}(t). \tag{R4.1}$$

Here $\hat{Q}_0^{ID}(t)\equiv\hat{Q}_X^I(t)+\hat{Q}_X^D(t)$ describes the mutation-free evolution, and $\hat{Q}_M^{ID}(t)\equiv\hat{Q}_M^I(t)+\hat{Q}_M^D(t)$ describes the single-mutation transition between states. From the reduced form of Eq. (R3.6), we get:

$$\langle s|\hat{Q}_0^{ID}(t) = -R_X^{ID}(s,t)\langle s|, \tag{R4.2}$$

$$\text{with}\quad R_X^{ID}(s,t)\equiv R_X^I(s,t)+R_X^D(s,t). \tag{R4.3}$$

Using Eq. (R4.1), the forward equation (Eq. (R3.19)) accompanied by the initial condition (Eq. (R3.21)) can be shown to be equivalent to a crucial integral equation:

$$\hat{P}^{ID}(t,t') = \hat{P}_0^{ID}(t,t') + \int_t^{t'}d\tau\hat{P}^{ID}(t,\tau)\hat{Q}_M^{ID}(\tau)\hat{P}_0^{ID}(\tau,t'). \tag{R4.4}$$

Here, $\hat{P}_0^{ID}(t',t'')\equiv T\left\{exp\left(\int_{t'}^{t''}d\tau\hat{Q}_0^{ID}(\tau)\right)\right\}$. Similarly, the backward equation (Eq. (R3.20)) accompanied by Eq. (R3.21) is equivalent to another crucial integral equation:

$$\hat{P}^{ID}(t,t') = \hat{P}_0^{ID}(t,t') + \int_t^{t'}d\tau\hat{P}_0^{ID}(t,\tau)\hat{Q}_M^{ID}(\tau)\hat{P}^{ID}(\tau,t'). \tag{R4.5}$$

Now, to formally solve Eq. (R4.4), we assume that the solution can be expanded as: $\hat{P}^{ID}(t,t')=\sum_{N=0}^{\infty}\hat{P}_{(N)}^{ID}(t,t')$, where $\hat{P}_{(N)}^{ID}(t,t')$ is the collection of terms containing N indel operators each. Substituting this expansion into Eq. (R4.4) and performing some formal calculations, we get the final form of the *ab initio* solution we desire:

$$\begin{aligned}\langle s_0|\hat{P}^{ID}(t_I,t_F) = &\sum_{N=0}^{\infty}\sum_{\left[\hat{M}_1,\,\hat{M}2,\cdots,\,\hat{M}_N\right]\in H^{ID}(N;s_0)}\\ &P\left[\left(\left[\hat{M}_1,\hat{M}_2,...,\hat{M}_N\right],[t_I,t_F]\right)|(s_0,t_I)\right]\langle s_0|\hat{M}_1\hat{M}_2\cdots\hat{M}_N.\end{aligned} \tag{R4.6}$$

Here, $H^{ID}(N;s_0)$ denotes the space of all possible histories of N indels each beginning with the sequence state, s_0. And

$$P\left[\left(\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N\right],[t_I,t_F]\right)|(s_0,t_I))]$$
$$=\int_{t_I=\tau_0<\tau_1<\cdots<\tau_N<\tau_{N+1}=t_F}\cdots\int d\tau_1\cdots d\tau_N\left(\prod_{\nu=1}^{N}r(\hat{M}_\nu;s_{\nu-1},\tau_\nu)\right)$$
$$\exp\left\{-\sum_{\nu=0}^{N}\int_{\tau_\nu}^{\tau_{\nu+1}}d\tau R_X^{ID}(s_\nu,\tau)\right\}\Big|_{\left\{\langle s_\nu|=\langle s_{\nu-1}|\hat{M}_\nu|\nu=1,\ldots,N\right\}} \tag{R4.7}$$

is the probability that an N -event indel history, $[\hat{M}_1, \hat{M}_2, \cdots, \hat{M}_N]$ (with $\hat{M}_\nu$ $(\nu = 1, 2, ..., N)$ being the ν-th event), occurred during time interval $[t_I, t_F]$, given an initial sequence state (s_0) at time t_I. The rate, $r(\hat{M}_\nu; s_{\nu-1}, \tau_\nu)$, is $r_I(x, l; s_{\nu-1}, \tau_\nu)$ if $\hat{M}_\nu = \hat{M}_I(x, l)$, and it is $r_D(x_B, x_E; s_{\nu-1}, \tau_\nu)$ if $\hat{M}_\nu = \hat{M}_D(x_B, x_E)$. It should be noted that $H^{ID}(N = 0; s_0) \equiv \{(s_0, [])\}$ consists only of the history with zero indel, [ ], whose conditional probability is: $P\left[([], [t_I, t_F])|(s_0, t_I)\right] = \exp\left\{-\int_{t_I}^{t_F} d\tau R_X^{ID}(s_0, \tau)\right\}$. Eq. (R4.6) supplemented with Eq. (R4.7) is also the solution of Eq. (R4.5). (Mathematically, Eq. (R4.7) is a multiple-time integral over all possible timing, whose integrand is the probability density of an evolutionary process of N indels with particular timing, $(\tau_1, \tau_2, ..., \tau_N)$).

Equation (R4.6) states that the finite-time transition operator (acting on $\langle s_0|$) is the collection of the effects of all possible indel histories starting with s_0, each weighted by its probability (Eq. (R4.7)). Thus, it mathematically underpins Gillespie's [34] famous stochastic simulation algorithm, which provides the basis of genuine molecular evolution simulators (*e.g.,* [26–28]). Our derivation of Eq. (R4.6) and Eq. (R4.7) through the integral equation (Eq. (R4.4) or Eq. (R4.5)) bridges Gillespie's own intuitive reasoning and Feller's [35] mathematically rigorous proof of the solution.

Now, substitute an "ancestral" sequence state, $s^A(\in S^{II})$, for s_0 in Eq. (R4.6), and take the inner product between it and the ket-vector, $|s^D\rangle$, of a "descendant" sequence state, $s^D(\in S^{II})$. Comparing the two sequence states in S^{II} naturally gives a PWA, $\alpha(s^A, s^D)$ (*e.g.,* Eq. (R2.1)).[10] Hence, the summation of $\langle s^A|\hat{P}^{ID}(t_I, t_F)|s^D\rangle$'s over all "equivalent" s^D's providing the same $\alpha(s^A, s^D)$ must be $P[(\alpha(s^A, s^D), [t_I, t_F])|(s^A, t_I)]$, which is the probability that $\alpha(s^A, s^D)$ results from sequence evolution during $[t_I, t_F]$, given s^A at t_I. Similarly to the derivation of Eq. (R4.6), we obtain its formal expression as:

$$P\left[\left(\alpha(s^A,s^D),[t_I,t_F]\right)|\left(s^A,t_I\right)\right]=\sum_{N=N_{min}[\alpha(s^A,s^D)]}^{\infty}\ \sum_{[\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N]\in H^{ID}[N;\alpha(s^A,s^D)]}$$
$$P\left[\left([\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N],[t_I,t_F]\right)\Big|\left(s^A,t_I\right)\right]. \tag{R4.8}$$

Here, $H^{ID}[N; \alpha(s^A, s^D)]$ denotes the set of all indel histories with N indels each that can result in $\alpha(s^A, s^D)$, and $N_{min}[\alpha(s^A, s^D)]$ is the minimum number of indels required for creating the PWA.

Using the set of all PWA-consistent histories, $\tilde{H}^{ID}[\alpha(s^A, s^D)] \equiv \bigcup_{N=N_{min}[\alpha(s^A,s^D)]}^{\infty} H^{ID}\left[N; \alpha(s^A, s^D)\right]$, Eq. (R4.8) can be further simplified as:

$$P\left[\left(\alpha(s^A,s^D),[t_I,t_F]\right)|\left(s^A,t_I\right)\right]=\sum_{[\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N]\in\tilde{H}^{ID}[\alpha(s^A,s^D)]}$$
$$P\left[\left([\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N],[t_I,t_F]\right)\Big|\left(s^A,t_I\right)\right]. \tag{R4.9}$$

Equation (R4.8) and Eq. (R4.9) are the formal expressions of the occurrence probability of $\alpha(s^A, s^D)$ derived in effect from the defining equations, Eqs. (R3.19, R3.20, R3.21), of our evolutionary model. Thus, they are the "*ab initio* probability" of the PWA. In the following, we will examine its factorability.

R5. Local history-set equivalence class of indel histories

Before advancing to the factorability of general PWA probabilities, we will introduce an essential concept here. For this purpose, we first consider the very simple PWA, Eq. (R2.1), as an example. (Here we make the substitutions, $s^A = \vec{v}_I$ and $s^D = \vec{v}_F$). In this case, $N_{min}[\alpha(s^A, s^D)] = 2$, and there are two 2-indel histories that can yield this PWA: one is $\left[\hat{M}_D(2,4), \hat{M}_I(3,3)\right]$ (Fig. 5a) and the other is $\left[\hat{M}_I(6,3), \hat{M}_D(2,4)\right]$ (Fig. 5b). Thus, these two indel histories result in the same final state: $\langle s_F| = \langle s_I|\hat{M}_D(2,4)\hat{M}_I(3,3) = \langle s_I|\hat{M}_I(6,3)\hat{M}_D(2,4)$ (Fig. 5, panels a and b). In other words, the two different successive actions of two indel operators have the same effect on the sequence states (in space S^{II}). This fact will be phrased as "the two products of the operators are *equivalent*", and represented by the relationship:

$$\hat{M}_I(6,3)\hat{M}_D(2,4) \sim \hat{M}_D(2,4)\hat{M}_I(3,3). \tag{R5.1}$$

This "binary equivalence" can be generalized to the following relationships between two indel events separated at least by a PAS:

$$\hat{M}_I(x_1,l_1)\hat{M}_I(x_2,l_2) \sim \hat{M}_I(x_2,l_2)\hat{M}_I(x_1+l_2,l_1)\ for\ x_1 > x_2, \tag{R5.2a}$$

$$\hat{M}_D(x_B,x_E)\hat{M}_I(x,l) \sim \hat{M}_I(x,l)\hat{M}_D(x_B+l,x_E+l)\ for\ x_B > x+1, \tag{R5.2b}$$

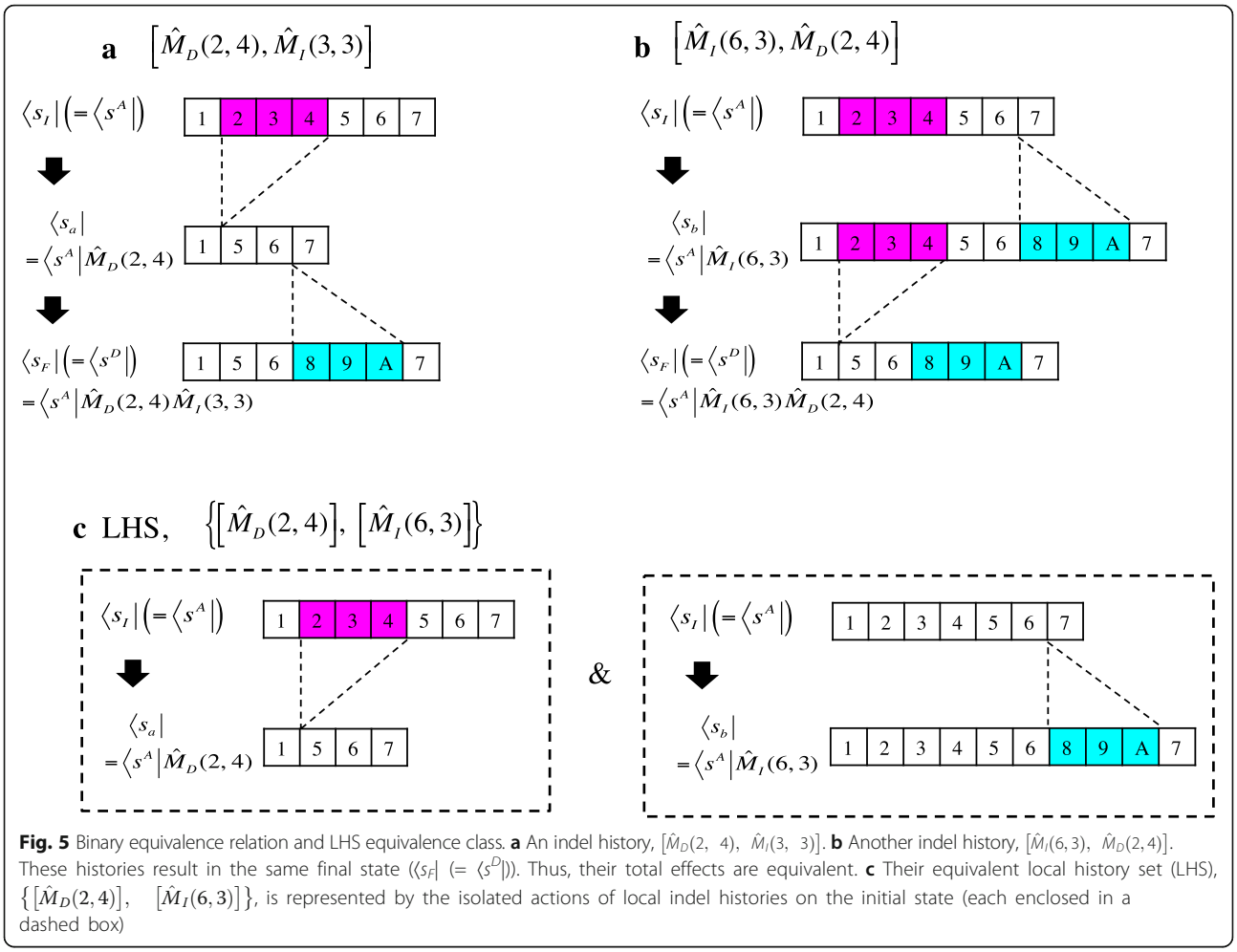

Fig. 5 Binary equivalence relation and LHS equivalence class. **a** An indel history, $[\hat{M}_D(2,\ 4),\ \hat{M}_I(3,\ 3)]$. **b** Another indel history, $[\hat{M}_I(6,3),\ \hat{M}_D(2,4)]$. These histories result in the same final state ($\langle s_F| = \langle s^D|$). Thus, their total effects are equivalent. **c** Their equivalent local history set (LHS), $\{[\hat{M}_D(2,4)],\ [\hat{M}_I(6,3)]\}$, is represented by the isolated actions of local indel histories on the initial state (each enclosed in a dashed box)

$$\hat{M}_I(x,l)\hat{M}_D(x_B,x_E) \sim \hat{M}_D(x_B,x_E)\hat{M}_I\left(x-l',l\right) \ for\ x > x_E, \tag{R5.2c}$$

$$\hat{M}_D(x_{B1},x_{E1})\hat{M}_D(x_{B2},x_{E2}) \sim \hat{M}_D(x_{B2},x_{E2})\hat{M}_D\left(x_{B1}-l'_2,x_{E1}-l'_2\right) \ for\ x_{B1} > x_{E2}+1. \tag{R5.2d}$$

Here, $l' \equiv min\{x_E - x_B + 1, x_E\}$ in Eq. (R5.2c), and $l'_2 \equiv min\{x_{E2} - x_{B2} + 1, x_{E2}\}$ in Eq. (R5.2d). If you will, these equivalence relations could be phrased as follows. "The operator representing the event on the left along the sequence will not change whether it comes first or second. The operator representing the event on the right will shift its operational position to the left/right by the number of sites deleted/inserted before its operation, when it comes second".[11]

Now, we will extend the binary equivalence relations, Eqs. (R5.2a-R5.2d), to the equivalence relations among more general complex indel histories, each consisting of more than two indel events. Let us consider a *global* history of N indel events, $[\hat{M}_1, \hat{M}_2, ..., \hat{M}_N]$, which begins with an "ancestral state", $s^A(\in S^{II})$, and ends with a "descendant state", $s^D(\in S^{II})$. Obviously, the two states must satisfy the equation:

$$\langle s^D| = \langle s^A|\hat{M}_1\hat{M}_2\cdots\hat{M}_N. \tag{R5.3}$$

Given an indel history, we can identify PASs unambiguously. Suppose that such PASs separate the indel events, $\hat{M}_\nu$ ($\nu = 1, 2, ..., N$) in $[\hat{M}_1, \hat{M}_2, ..., \hat{M}_N]$, into K *local* subsets of indels, each of which is confined either between a pair of PASs or between a PAS and an end of the resulting PWA. Number the K local subsets as $k = 1, 2, ..., K$ from left to right, and let N_k be the number of indel events in the k th local subset. Naturally, $\Sigma_{k=1}^{K} N_k = N$. And let $\hat{M}'[k, i_k]$ be the element of $\{\hat{M}_\nu\}_{\nu=1,2,...,N}$ representing the i_k th event (in the temporal order) in the k th local subset ($i_k = 1, 2, ..., N_k$; $k = 1, 2, ..., K$). (The prime here indicates that the operator is equivalent to the prime-less version). Then, repeatedly applying the binary equivalence relations, Eqs. (R5.2a-R5.2d), between the indel operators belonging to *different* local

subsets, we can move the operators around in the product in Eq. (R5.3) and get the following equation:

$$\langle s^D| = \langle s^A|\left[\hat{M}[K,1]\cdots\hat{M}[K,N_K]\right]\cdots\left[\hat{M}[1,1]\cdots\hat{M}[K,N_1]\right]. \tag{R5.4}$$

Here $\hat{M}[k,i_k]$ is an operator that was obtained from $\hat{M}'[k,i_k]$ through the series of equivalence relations Eqs. (R5.2a-R5.2d) that brought Eq. (R5.3) into Eq. (R5.4). As in Eq. (R5.3), the operators in each pair of large square parentheses in Eq. (R5.4) are arranged in temporal order, so that the earliest event in each local subset will come leftmost. But it should be noted that the order among the pairs of large square parentheses is the opposite of the actual spatial order among the local subsets, so that the rightmost one along the sequence (the K th one here) will come leftmost. In this way, the operators in each local subset, *e.g.*, $\left\{\hat{M}[k,1],\ldots,\hat{M}[k,N_k]\right\}$, are exactly the same as those when the events in the subset alone struck $\langle s^A|$. Thus the series of operators, $\left[\hat{M}[k,1],\ldots,\hat{M}[k,N_k]\right]$, for the k th local subset defines the k th *local indel history* that was isolated from the global indel history, $[\hat{M}_1,\hat{M}_2,\ldots,\hat{M}_N]$, on $s_I \in S$.

Now, let us consider a history of N indel events other than $[\hat{M}_1,\hat{M}_2,\ldots,\hat{M}_N]$. If the temporal operator product of the history is shown to be equivalent to Eq. (R5.4) through a series of Eqs. (R5.2a-R5.2d), then it should also be connected to Eq. (R5.3) though another series of Eqs. (R5.2a-R5.2d). Therefore, it should be equivalent to $[\hat{M}_1,\hat{M}_2,\ldots,\hat{M}_N]$ in this sense. Hence, we can define a particular equivalence class, which is the set of all global indel histories that can be "decomposed" into the identical set of local indel histories, such as Eq. (R5.4), only through a series of Eqs. (R5.2a-R5.2d), between indel operators separated by at least a PAS. We will call it the "**local-history-set (LHS) equivalence class**". In the equivalence class defined by a local history set (LHS), $\left\{\left[\hat{M}[k,1],\ldots,\hat{M}[k,N_k]\right]\right\}_{k=1,\ldots,K}$ (with $\sum_{k=1}^{K} N_k = N$), on an initial sequence state $s^A \in S^{II}$, there are $\frac{N!}{\prod_{k=1}^{K} N_k}$ LHS-equivalent global indel histories beginning with s^A. Each of the global histories corresponds to a way of reordering N indel events while retaining the relative temporal order among N_k events within the k th local indel history (for every $k = 1,\ldots,K$).

In the simplest example at the beginning of this section (Eq. (R5.1) and above), the corresponding LHS is: $\left\{\left[\hat{M}_D(2,4)\right],\left[\hat{M}_I(6,3)\right]\right\}$. The LHS consists of two local histories, each of which is a single-indel history (Fig. 5c). As a slightly more complex example, consider the history, $\left[\hat{M}_D(3,3),\hat{M}_I(5,2),\hat{M}_D(2,3),\hat{M}_I(5,1)\right]$, illustrated in Fig. 4. This history belongs to the LHS equivalence class represented by the LHS: $\left\{\left[\hat{M}_D(3,3),\hat{M}_D(2,3)\right],\left[\hat{M}_I(6,2),\hat{M}_I(8,1)\right]\right\}$, which consists of two 2-indel local histories. If this LHS is recast into the form in Eq. (R5.4), we have: $\hat{M}[1,1] = \hat{M}_D(3,3)$, $\hat{M}[1,2] = \hat{M}_D(2,3)$, $\hat{M}[2,1] = \hat{M}_I(6,2)$, and $\hat{M}[2,2] = \hat{M}_I(8,1)$.

R6. Factorability of pairwise alignment probability: brief description

Now we are ready to examine the factorability of the *ab initio* probability of PWA $\alpha(s^A,s^D)$, $P[(\alpha(s^A,s^D),[t_I,t_F])|(s^A,t_I)]$ in Eq. (R4.9), given the ancestral state (s^A) at the initial time (t_I). Here the "factorability" means that the PWA probability can be re-expressed as the product of an overall factor and contributions from local regions. Natural candidates for the local regions would be those in between the PASs, because we know that indels never hit or pierced PASs (if the alignment is correct). We are not interested in trivial factorability. Thus, we only consider PWAs (or global histories) each of which requires at least two local indel histories. In the following, we will only briefly describe our demonstration of the PWA probability factorization. Its more detailed yet rather intuitive description is given in Supplementary methods SM-2 in Additional file 1. (It is complemented by mathematically rigorous arguments in Supplementary appendix SA-2 in Additional file 2).

We first notice that each component probability, $P[([\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N],[t_I,t_F])|(s^A,t_I)]$ given by Eq. (R4.7), will not be factorable. This is because its domain of multiple-time integration is not a direct product. So, lets consider the total probability of a LHS equivalence class, $\left[\overrightarrow{\overrightarrow{\hat{M}}}\right]_{LHS}$ (with $\overrightarrow{\overrightarrow{\hat{M}}}$ abbreviating $\left\{\left[\hat{M}[k,1],\ldots,\hat{M}[k,N_k]\right]\right\}_{k=1,\ldots,K}$):

$$P\left[\left(\left[\overrightarrow{\overrightarrow{\hat{M}}}\right]_{LHS},[t_I,t_F]\right)|(s^A,t_I)\right] \equiv \sum_{[\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N]\in\left[\overrightarrow{\overrightarrow{\hat{M}}}\right]_{LHS}} P\left[([\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_N],[t_I,t_F])|(s^A,t_I)\right]. \tag{R6.1}$$

We can show that this probability can be factorized as:

$$\mu_P\left[\left(\left[\overrightarrow{\overrightarrow{\hat{M}}}\right]_{LHS},[t_I,t_F]\right)|(s^A,t_I)\right] = \prod_{k=1}^{K}\mu_P\left[\left(\left[\hat{M}[k,1],\ldots,\hat{M}[k,N_k]\right],[t_I,t_F]\right)|(s^A,t_I)\right], \tag{R6.2}$$

where

$$\mu_P\Big[\big(\big[\hat{M}[k,1],...,\hat{M}\big[k,N_k\big]\big],[t_I,t_F]\big)\Big|\big(s^A,t_I\big)\Big] \equiv P[\big(\big[\hat{M}[k,1],...,\hat{M}\big[k,N_k\big]\big],[t_I,t_F]\big)\big|\big(s^A,t_I\big)\big]/P\big[([],[t_I,t_F])\big|\big(s^A,t_I\big)\big], \tag{R6.3}$$

$$\mu_P\Big[\Big(\Big[\overrightarrow{\hat{M}}\Big]_{LHS},[t_I,t_F]\Big)\Big|\big(s^A,t_I\big)\Big] \equiv P\Big[\Big(\Big[\overrightarrow{\hat{M}}\Big]_{LHS},[t_I,t_F]\Big)\Big|\big(s^A,t_I\big)\Big]/P[([],[t_I,t_F])|\big(s^A,t_I\big)], \tag{R6.4}$$

if the following two conditions are satisfied.

Condition (i): The rate of an indel event ($r(\hat{M}_v;s_{v-1},\tau_v)$) is independent of the portion of the sequence state (s_{v-1}) outside of the region of the local history the event ($\hat{M}_v$) belongs to.

Condition (ii): The increment of the exit rate due to an indel event ($\delta R_X^{ID}(s_v,s_{v-1},\tau) \equiv R_X^{ID}(s_v,\tau) - R_X^{ID}(s_{v-1},\tau)$, with $\langle s_v| = \langle s_{v-1}|\hat{M}_v\rangle$ is independent of the portion of the sequence state (s_{v-1}) outside of the region of the local history the event ($\hat{M}_v$) belongs to.

See Supplementary appendix SA-2 in Additional file 2 for the derivation of the mathematically rigorous version of this set of conditions. (For illustration, in Supplementary methods SM-3 in Additional file 1, the factorability of the probability was examined for the simplest concrete LHS equivalence class (in Fig. 5)). Condition (i) is somewhat similar to the "context-independence" condition imposed on the "long indel" model [21], though our condition is slightly less restrictive. Condition (ii) has never been found or even discussed thus far. In fact, the "long indel" model trivially satisfies this condition (see subsection R8-1), thus [21] did not need to pay attention to it. However, this condition is not always satisfied. Indeed, as exemplified in subsection R8-2, some models have non-factorable alignment probabilities due to the violation of this condition even though condition (i) is satisfied.

Each global indel history (in the set of all PWA-consistent indel histories, $\tilde{\mathrm{H}}^{ID}\big[\alpha(s^A,s^D)\big]$) belongs to a single LHS equivalence class (say, $\Big[\overrightarrow{\hat{M}}\Big]_{LHS}$). And all elements of $\Big[\overrightarrow{\hat{M}}\Big]_{LHS}$ can result in the same PWA. Therefore, we get the direct sum structure:

$$\tilde{\mathrm{H}}^{ID}\big[\alpha(s^A,s^D)\big] = \bigcup_{\overrightarrow{\hat{M}}\in\tilde{\Lambda}^{ID}[\alpha(s^A,\ s^D)]}\Big[\overrightarrow{\hat{M}}\Big]_{LHS}, \tag{R6.5}$$

where $\tilde{\Lambda}^{ID}\big[\alpha(s^A,s^D)\big]$ is the set of all LHSs consistent with $\alpha(s^A,s^D)$. Hence, the PWA probability, Eq. (R4.9), can be rewritten as:

$$P\big[\big(\alpha(s^A,s^D),[t_I,t_F]\big)|\big(s^A,t_I\big)\big] = \sum_{\overrightarrow{\hat{M}}\in\tilde{\Lambda}^{ID}[\alpha(s^A,s^D)]} P\Big[\Big(\Big[\overrightarrow{\hat{M}}\Big]_{LHS},[t_I,t_F]\Big)|\big(s^A,t_I\big)\Big]. \tag{R6.6}$$

We are considering *all* indel histories, including non-parsimonious ones, that can yield $\alpha(s^A,s^D)$. Thus, the LHSs belonging to $\tilde{\Lambda}^{ID}\big[\alpha(s^A,s^D)\big]$ may consist of different numbers of local histories. We will choose the maximum possible set of PASs in the given PWA, which separates the PWA into the finest potentially local-history-accommodating regions, denoted as $\gamma_1,\gamma_2,...,\gamma_{\kappa_{max}}$. ($\kappa_{max}$ is uniquely determined by the PWA and the evolutionary model).[12] Then, we can represent any $\overrightarrow{\hat{M}} = \big\{\big[\hat{M}[k,1],...,\hat{M}[k,N_k]\big]\big\}_{k=1,...,K}\in\tilde{\Lambda}^{ID}\big[\alpha(s^A,s^D)\big]$ as a vector with κ_{max} components: $\overrightarrow{\hat{M}} = \Big(\overrightarrow{\hat{M}}[\gamma_1],\ \overrightarrow{\hat{M}}[\gamma_2],...,\overrightarrow{\hat{M}}\Big[\gamma_{\kappa_{max}}\Big]\Big)$. (See Figure S1 in Additional file 1). Substituting Eqs. (R6.2,R6.4) into Eq. (R6.6), and exploiting the vector representation of the LHSs, we can reach the desired expression:

$$P\Big[\big(\alpha(s^A,s^D),[t_I,t_F]\big)|\big(s^A,t_I\big)\Big] = P\Big[([],[t_I,t_F])|\big(s^A,t_I\big)\Big] \times\prod_{\kappa=1}^{\kappa_{\max}}\tilde{\mu}_P\Big[\big(\tilde{\Lambda}^{ID}\big[\gamma_\kappa;\alpha(s^A,s^D)\big],[t_I,t_F]\big)|\big(s^A,t_I\big)\Big]. \tag{R6.7}$$

Here, $\tilde{\Lambda}^{ID}\big[\gamma_\kappa;\alpha(s^A,s^D)\big]$ denotes the set of local indel histories that can give rise to the sub-PWA of $\alpha(s^A,s^D)$ confined in γ_κ, and the multiplication factor,

$$\tilde{\mu}_P\Big[\big(\tilde{\Lambda}^{ID}\big[\gamma_\kappa;\alpha(s^A,s^D)\big],[t_I,t_F]\big)|\big(s^A,t_I\big)\Big] \equiv \sum_{\overrightarrow{\hat{M}}\big[\gamma_\kappa\big]\in\tilde{\Lambda}^{ID}\left[\gamma_\kappa;\alpha(s^A,s^D)\right]}\mu_P\Big[\Big(\overrightarrow{\hat{M}}\big[\gamma_\kappa\big],[t_I,t_F]\Big)\Big|\big(s^A,t_I\big)\Big], \tag{R6.8}$$

represents the total contribution from $\tilde{\Lambda}^{ID}\big[\gamma_\kappa;\alpha(s^A,s^D)\big]$ to the PWA probability. (Here $\mu_P[([],[t_I,t_F])|(s^A,t_I)] = 1$ should be remembered).

Equation (R6.7) states that the PWA probability is factorized into the product of an overall factor ($P[([],[t_I,t_F])|(s^A,t_I)]$) and contributions from regions accommodating local indel histories ($\tilde{\mu}_P\big[\big(\tilde{\Lambda}^{ID}\big[\gamma_\kappa;\alpha(s^A,s^D)\big],[t_I,t_F]\big)\big|\big(s^A,t_I\big)\big]$'s). Therefore, the set of conditions, (i) and (ii), is sufficient for the factorability of the PWA

probability. At present, we are not sure whether the set of conditions is also necessary or not. This may not be the case in the *rigorous* sense, and there may be some instances with factorable PWA probabilities despite the violation of condition (i) or (ii). Nevertheless, even if there are, we suspect that such cases should be isolated, requiring intricate cancellations of the terms. Thus, we will refer to the conditions (i) and (ii) as the "sufficient and nearly necessary set of conditions" for factorable PWA probabilities.

R7. Factorability of multiple sequence alignment probability: brief description

Thus far, we only examined the probability of a given PWA, conditioned on an ancestral state at initial time. Actually, once we know how to calculate such conditional PWA probabilities, we can build them up along the phylogenetic tree to calculate the probability of a given MSA, as described in the introductions of [13] and [14]. (See also [36] for an essentially equivalent method that appears different.) Here, we basically follow their procedures. However, it should be stressed that the MSA probability here will be calculated *ab initio* under a genuine evolutionary stochastic model and not under a HMM or a transducer, which is not necessarily evolutionarily consistent. This section briefly explains the derivation of the factorization of an *ab initio* MSA probability. For details on the derivation, see Supplementary methods SM-4 in Additional file 1.

In this section, we formally calculate the *ab initio* probability of a MSA given a rooted phylogenetic tree, $T = (\{n\}_T, \{b\}_T)$, where $\{n\}_T$ is the set of all nodes of the tree, and $\{b\}_T$ is the set of all branches of the tree. We decompose the set of all nodes as: $\{n\}_T = N^{IN}(T) + N^X(T)$, where $N^{IN}(T)$ is the set of all internal nodes and $N^X(T) = \{n_1, \ldots, , n_{N^X}\}$ is the set of all external nodes. (The $N^X \equiv |N^X(T)|$ is the number of external nodes.) The root node plays an important role and will be denoted as n^{Root}. Because the tree is rooted, each branch b is directed. Thus, let $n^A(b)$ denote the "ancestral node" on the upstream end of b, and let $n^D(b)$ denote the "descendant node" on the downstream end of b. Let $s(n) \in S^{II}$ be a sequence state at the node $n \in \{n\}_T$. Especially, we use abbreviations: $s^A(b) \equiv s(n^A(b)) \in S^{II}$ and $s^D(b) \equiv s(n^D(b)) \in S^{II}$. Finally, as mentioned in Background, we suppose that the branch lengths, $\{|b| | b \in \{b\}_T\}$, and the indel model parameters, $\{\Theta_{ID}(b)\}_T \equiv \{\Theta_{ID}(b) | b \in \{b\}_T\}$, are all given. Note that the model parameters $\Theta_{ID}(b)$ could depend on the branch, at least theoretically.

First, we extend the ideas proposed in [13, 14, 36] to each indel history along a tree, by regarding the indel history along a branch as a map (or a transformation) from the ancestral state to the descendant state, as follows. An indel history along a tree consists of indel histories along all branches of the tree that are interdependent, in the sense that the indel process of a branch b determines a sequence state $s^D(b)$ at its descendant node $n^D(b)$, on which the indel processes along its downstream branches depend. Thus, an indel history on a given root sequence state $s^{Root} = s(n^{Root}) \in S^{II}$ automatically determines the sequence states at all nodes, $\{s(n) \in S^{II} \text{ for } {}^\forall n \in \{n\}_T\}$. Let $\tilde{\mathrm{H}}^{ID}(s_0) \equiv \cup_{N=0}^{\infty} \mathrm{H}^{ID}(N; s_0)$ (with $\mathrm{H}^{ID}(N; s_0)$ defined below Eq. (R4.6)) be the set of all indel histories along a time axis (or a branch) starting with state s_0. Then, each indel history, $\left\{\overrightarrow{\hat{M}}(b)\right\}_T$, along tree T and starting with s^{Root} can be specifically expressed as:

$$\left\{ \begin{array}{l} \overrightarrow{\hat{M}}(b) = [\hat{M}_1(b), \ldots, \hat{M}_{N(b)}(b)] \in \tilde{\mathrm{H}}^{ID}\left(s^A(b)\right) \text{ and} \\ \langle s^D(b)| = \langle s^A(b)|\hat{M}_1(b) \cdots \hat{M}_{N(b)}(b) \quad \text{for } {}^\forall b \in \{b\}_T \end{array} \middle| s\left(n^{Root}\right) = s^{Root} \right\}. \tag{R7.1}$$

Here, the symbol, $\hat{M}_\nu(b)$, denotes the ν th event in the indel history along branch $b \in \{b\}_T$. The probability of the indel history, Eq. (R7.1), can be calculated as:

$$P\left[\{\overrightarrow{\hat{M}}(b)\}_T | \left(s^{Root}, n^{Root}\right)\right] = \left(\prod_{b \in \{b\}_T} P\left[\left(\overrightarrow{\hat{M}}(b), b\right) | \left(s^A(b), n^A(b)\right)\right] \right) \Bigg|_{\substack{s(n^{Root}) = s^{Root}, \\ \langle s^D(b)| = \langle s^A(b)|\hat{M}_1(b) \cdots \hat{M}_{N(b)}(b) \\ \text{for } {}^\forall b \in \{b\}_T}} \tag{R7.2}$$

Here, the probability of an indel history, $\overrightarrow{\hat{M}}(b) = \left[\hat{M}_1(b), \ldots, \hat{M}_{N(b)}(b)\right] \in \tilde{\mathrm{H}}^{ID}\left(s^A(b)\right)$, along branch $b \in \{b\}_T$ is given by the probability during the corresponding time interval, $[t(n^A(b)), t(n^D(b))]$:

$$\begin{aligned} & P[(\overrightarrow{\hat{M}}(b), b) | \left(s^A(b), n^A(b)\right)] \\ & \equiv P[([\hat{M}_1(b), \cdots, \hat{M}_{N(b)}(b)], [t\left(n^A(b)\right), t(n^D(b))]) | (s^A(b), \\ & \quad t\left(n^A(b)\right))] \big|_{\Theta_{ID}(b).} \end{aligned} \tag{R7.3}$$

Here we explicitly showed the branch-dependence of the model parameters.

Now, consider a MSA, $\alpha[s_1, s_2, \ldots, s_{N^X}]$, among the sequences at the external nodes, $s_i = s(n_i) \in S^{II}$ $(n_i \in N^X(T))$. (Remember that the term "MSA" here means its homology structure, as noted in endnote (10)). Let $\left(s^{Root}, \left\{\overrightarrow{\hat{M}}(b)\right\}_T\right)$ be a pair of a root state and an indel history along T starting with the state. And let $\tilde{\Psi}^{ID}[\alpha[s_1, s_2, \ldots, s_{N^X}]; T]$ be the set of all such pairs defined on T consistent with $\alpha[s_1, s_2, \ldots, s_{N^X}]$. Then, analogously to

Eq. (R4.9) supplemented with Eq. (R4.7) for a PWA, the probability of a given MSA under a given model setting (including T) should be expressed as:

$$P\left[\alpha[s_1, s_2, ..., s_{N^X}]|T\right] = \sum_{\substack{\left(s^{Root},\ \left\{\vec{\hat{M}}(b)\right\}_T\right) \\ \in\tilde{\Psi}^{ID}\left[\alpha[s_1, s_2, ..., s_{N^X}];\ T\right]}} P\left[(s^{Root}, n^{Root})\right] P\left[\left\{\vec{\hat{M}}(b)\right\}_T \middle| (s^{Root}, n^{Root})\right], \tag{R7.4}$$

which is supplemented with Eq. (R7.2). Here, $P[(s^{Root}, n^{Root})]$ is the probability of state s^{Root} at the root node. (It may be interpreted as the prior in a Bayesian formalism.) If you will, Eqs. (R7.4) and (R7.2) could be considered as the "perturbation expansion" of an *ab initio* MSA probability. To make this formal expansion more tractable, let $\{s(n)\}_{N^{IN}} \equiv \left\{s(n) \in S \middle| n \in N^{IN}(T)\right\}$ denote a set of ancestral states at all internal nodes (, or, more precisely, its equivalence class in the sense of endnote (8)). To be consistent with a given MSA, the ancestral states must satisfy the "phylogenetic correctness" condition in each MSA column (*e.g.*, [37, 38]).[13] Let $\Sigma[\alpha[s_1, s_2, ..., s_{N^X}];\ \left\{n \in N^{IN}(T)\right\};\ T]$ be the set of all $\{s(n)\}_{N^{IN}}$'s consistent with $\alpha[s_1, s_2, ..., s_{N^X}]$ (and tree T). Then, Eq. (R7.4) supplemented with Eq. (R7.2) can be rewritten as:

$$P\left[\alpha[s_1, s_2, ..., s_{N^X}]|T\right] = \sum_{\substack{\{s(n)\}_{N^{IN}} \\ \in\Sigma[\alpha[s_1, s_2, ..., s_{N^X}];\ \left\{n \in N^{IN}(T)\right\};\ T]}} P\left[\alpha[s_1, s_2, ..., s_{N^X}];\ \{s(n)\}_{N^{IN}}|T\right]. \tag{R7.5}$$

Here, $P[\alpha[s_1, s_2, ..., s_{N^X}];\ \{s(n)\}_{N^{IN}}|T]$ is the probability of simultaneously getting $\alpha[s_1, s_2, ..., s_{N^X}]$ and $\{s(n)\}_{N^{IN}}$. This probability is the sum of contributions from all indel histories sharing the same homology structure among sequence states at all nodes. Especially, the sequence states at internal nodes have homology structures (with the states at other nodes) fixed for respective nodes. Thus, through some reasoning (explained in SM-4), we get:

$$P\left[\alpha[s_1, s_2, ..., s_{N^X}];\ \{s(n)\}_{N^{IN}}|T\right] = P\left[(s^{Root}, n^{Root})\right] \prod_{b \in \{b\}_T} P\left[\left(\alpha\left(s^A(b), s^D(b)\right), b\right) \middle| \left(s^A(b), n^A(b)\right)\right]. \tag{R7.6}$$

Here,

$$P\left[\left(\alpha\left(s^A(b), s^D(b)\right), b\right) \middle| \left(s^A(b), n^A(b)\right)\right] \equiv P\left[\left(\alpha\left(s^A(b), s^D(b)\right), \left[t\left(n^A(b)\right), t(n^D(b))\right]\right) \middle| \left(s^A(b), t\left(n^A(b)\right)\right)\right]\Big|_{\Theta_{ID}(b)} \tag{R7.7}$$

is the probability of the ancestor-descendant PWA along branch b. This Eq. (R7.6) is basically the expression proposed in [13, 14], and we demonstrated in effect that their proposal also holds even with a genuine stochastic evolutionary model. Usually, Eq. (R7.5) supplemented with Eq. (R7.6) is much more tractable than Eq. (R7.4) supplemented with Eq. (R7.2), because of the two reasons. (1) Usually, it is not the indel history along the tree but (the homology structure of) the set of ancestral sequence states that is inferred from a given MSA. (2) The probability of each indel history along the tree (Eq. (R7.2)) is not factorable in general, whereas Eq. (R7.6) is a product of PWA probabilities, each of which should be factorable if the conditions (i) and (ii) in section R6 are satisfied.

Now, we can show that, if the "condition (iii)" given below in addition to conditions (i) and (ii) is satisfied, we can factorize the MSA probability into a form somewhat similar to Eq. (R6.7) for the PWA probability. In subsection 4.2 of [32], we demonstrated it using the history-based expansion of the MSA probability (*i.e.*, Eq. (R7.4) supplemented with Eq. (R7.2)). In Supplementary methods SM-4, we will use the ancestral-state-based expansion (*i.e.*, Eq. (R7.5) supplemented with Eq. (R7.6)), as was only briefly sketched at the bottom of subsection 4.2 of *(ibid.)*. In a MSA, gapless columns play almost the same role as PASs in a PWA. Because of the aforementioned "phylogenetic correctness" condition, a gapless column indicates that no indel event hit or pierced the site. Therefore, gapless columns will partition a MSA into regions each of which accommodates a local subset of every global history. Analogously to the argument above Eq. (R6.7), let $C_1, C_2, ..., C_{K_{max}}$ be the maximum possible set of such regions determined by a given MSA and a model setting (including the tree) (Figure S2 in Additional file 1). (As argued in subsection R8-3, all gapless columns are not necessarily needed to delimit the regions.) Because the summation in Eq. (R7.5) involves the summation over all MSA-consistent root states, it would be convenient to specify a "reference" root state, s_0^{Root}. It can be anything, as long as it is the state at the root consistent with $\alpha[s_1, s_2, ..., s_{N^X}]$. Technically, one good candidate for s_0^{Root} would be a root state obtained by applying the Dollo parsimony principle [39] to each column of the MSA, because it is arguably the most readily available state that satisfies the phylogenetic

correctness condition along the entire MSA. Then, we will impose the following condition.

Condition (iii):

$$P\left[\left(s^{Root}, n^{Root}\right)\right] = P\left[\left(s_0^{Root}, n^{Root}\right)\right]\prod_{K=1}^{K_{max}} \mu_P[s^{Root}, s_0^{Root}, n^{Root}; C_K]. \tag{R7.8}$$

Here the multiplication factor, $\mu_P[s^{Root}, s_0^{Root}, n^{Root}; C_K]$, represents the change in the state probability at the root due to the difference between s^{Root} and s_0^{Root} within C_K. This equation holds, *e.g.*, when $P[(s^{Root}, n^{Root})]$ is a geometric distribution or a uniform distribution of the root sequence length.[14]

Under the conditions (i), (ii) and (iii), through a series of formal calculations and reasoning, Eq. (R7.5) supplemented with Eq. (R7.6) can be re-expressed into the final factorized form:

$$P[\alpha[s_1, s_2, ..., s_{N^X}]|T] = P_0[s_0^{Root}|T]\prod_{K=1}^{K_{max}} \widetilde{M}_P[\alpha[s_1, s_2, ..., s_{N^X}]; s_0^{Root}; C_K|T]. \tag{R7.9}$$

Here,

$$P_0\left[s_0^{Root}|T\right] \equiv P\left[\left(s_0^{Root}, n^{Root}\right)\right] P\left[\{[]\}_T \middle| \left(s_0^{Root}, n^{Root}\right)\right] \tag{R7.10}$$

is the probability of having state s_0^{Root} that has been intact all across tree T, and $\widetilde{M}_P[\alpha[s_1, s_2, ..., s_{N^X}]; s_0^{Root}; C_K|T]$ is the multiplication factor contributed from all local indel histories (along T) confined in C_K.[15] Briefly, the multiplication factor is the summation of terms over all possible sets of the MSA-consistent ancestral states in C_K. And each of the terms is the product of the local PWA multiplication factors (Eq. R6.8) confined in C_K (Figure S2 in Additional file 1), the exponential of minus the summation over all b's of the time-integrated exit rate differences between $s^A(b)$ and s_0^{Root} coming from C_K, and $\mu_P[s^{Root}, s_0^{Root}, n^{Root}; C_K]$ for the root state probability. (For the factor's exact expression and the detailed derivation of Eq. (R7.9), see Supplementary methods SM-4 in Additional file 1).

R8. Examples: Models with factorable/non-factorable alignment probabilities

A merit of conditions (i) and (ii) given in section R6 is that they can draw the line between evolutionary models with factorable PWA probabilities and those with non-factorable ones. To illustrate the use of these conditions, we here give three examples: (1) a simple model with factorable probabilities, (2) a simple model with non-factorable probabilities, and (3) a non-trivial model with factorable probabilities. (For more examples, see section 5 of [32]).

R8-1. Totally space-homogeneous model

The simplest conceivable indel models would be those whose indel rate parameters are space-homogeneous, *i.e.*, independent of the positions where the indels occur:

$$r_I(x, l_1; s, t) = g_I(l_1, t), \tag{R8 − 1.1}$$

$$r_D(x_B, x_B + l_2 - 1; s, t) = g_D(l_2, t). \tag{R8 − 1.2}$$

In fully space-homogeneous models, these equations hold for $1 \le x \le L(s) - 1$, $1 \le l_1 \le L_I^{CO}$, $1 \le l_2 \le L_D^{CO}$, and $2 - l_2 \le x_B \le L(s)$, where L_I^{CO} and L_D^{CO} are the "cut-off lengths" for insertions and deletions, respectively. (Depending on the model, $r_I(0, l; s, t) = g_{I;L}(l, t)$ and $r_I(L(s), l; s, t) = g_{I;R}(l, t)$ could differ from $g_I(l, t)$ in Eq. (R8-1.1)). In fact, these conditions were imposed in nearly all continuous-time Markov models of indels that were studied in the past. Note that the rate parameters in Eqs. (R8-1.1,R8-1.2) could depend on time, although most studies thus far assumed that the rates are time-independent as well. Eqs. (R8-1.1,R8-1.2) automatically guarantees condition (i). Thus, all we have to do is check whether or not condition (ii) is also satisfied. Indeed, we can show it is. The exit rate of this model is calculated by substituting Eqs. (R8-1.1,R8-1.2) into Eq. (R4.3) (supplemented with Eqs. (R3.14,R3.15)), and we find that it is an affine function of the sequence length ($L(s)$):

$$R_X^{ID}(s, t) = A(t)L(s) + B(t), \tag{R8 − 1.3}$$

with $A(t) = \sum_{l=1}^{L_I^{CO}} g_I(l, t) + \sum_{l=1}^{L_D^{CO}} g_D(l, t)$ and $B(t) = \sum_{l=1}^{L_D^{CO}} (l-1) g_D(l, t) - \sum_{l=1}^{L_I^{CO}} g_I(l, t) + \sum_{l=1}^{L_I^{CO}} \left(g_{I;L}(l, t) + g_{I;R}(l, t)\right)$. If the exit rate is affine, we have, for $\langle s(\nu)| = \langle s(\nu-1)|\hat{M}_\nu$:

$$\delta R_X^{ID}(s(\nu), s(\nu-1), t) \equiv R_X^{ID}(s(\nu), t) - R_X^{ID}(s(\nu-1), t) = A(t)[L(s(\nu)) - L(s(\nu-1))] = A(t)\delta L\left(\hat{M}_\nu\right). \tag{R8 − 1.4}$$

Here $\delta L\left(\hat{M}_\nu\right)$ is the length change caused by the event, $\hat{M}_\nu$. The rightmost hand side of this equation depends only on $\hat{M}_\nu$ and the time it occurred, but not on the sequence state ($s(\nu - 1)$). Thus, condition (ii) is always satisfied under fully space-homogenous models, which means that alignment probabilities calculated *ab initio* (as in section R4) under such models are factorable, as shown in section R6.

An important special case of the space-homogeneous model is the model used by Dawg [26], whose indel rate

parameters are given as: $g_I(l,t) = g_{I;L}(l,t) = g_{I;R}(l,t) = \lambda_I f_I(l)$ and $g_D(l,t) = \lambda_D f_D(l)$. Because this is a special case of Eqs. (R8-1.1,R8-1.2), it naturally provides factorable alignment probabilities. This model is probably among the most flexible indel evolutionary models used thus far. The model accommodates any distributions of indel lengths ($f_I(l)$ and $f_D(l)$) that are independent of each other, and independent total rates for insertions and deletions (λ_I and λ_D). Some of our studies [40, 41] are mostly based on this model.

Another important special case is the "long indel" model [21], whose (time-independent) rate parameters are given by: $g_I(l,t) = \lambda_l$, $g_{I;L}(l,t) = g_{I;R}(l,t) = \tilde{\lambda}_l^{(end)}$ (if $L(s) > 0$), $g_{I;L}(l,t) = \tilde{\tilde{\lambda}}_l^{(whole)}$ (if $L(s) = 0$), and $g_D(l,t) = \mu_l$. This model is less flexible than Dawg's model, because its indel rates are subject to the detailed-balance conditions: $\lambda_l = (\lambda_1/\mu_1)^l \mu_l$, $\tilde{\lambda}_l^{(end)} = (\lambda_1/\mu_1)^l \sum_{l'=l}^{L_D^{CO}} \mu_{l'}$ and $\tilde{\tilde{\lambda}}_l^{(whole)} = (\lambda_1/\mu_1)^l \sum_{l'=l}^{L_D^{CO}} \left(l' - 1 + 1\right)\mu_{l'}$. Like Dawg's model, this model is a special case of the model defined by Eqs. (R8-1.1,R8-1.2). Thus, the alignment probabilities calculated under it are indeed factorable, as verbally justified in [21]. Indeed, we can explicitly show that, as far as each LHS equivalence class is concerned, the indel component of its probability calculated according to the recipe of [21] equals the product of $P[([], [t_I, t_F])|(s^A, t_I)]$ and Eq. (R6.2), *i.e.*, the "total probability" of the LHS equivalence class via our *ab initio* formulation, calculated with the aforementioned indel rate parameters. The proof is given in Supplementary appendix SA-3 in Additional file 2. It should be stressed that, although [21] ignored condition (ii), this caused no problem thanks to Eq. (R8-1.4) satisfied by any fully space-homogeneous models. Actually, it is this condition (ii) that guarantees the equivalence of the probabilities calculated via the two methods, because it equates each increment of the exit rate of a chop-zone with that of an entire sequence. The equivalence can be extended to between PWA probabilities, *provided that the contributing local indel histories are correctly enumerated.* (We are uncertain about whether this extended equivalence indeed holds, because [21] did not explicitly describe how the local indel histories were enumerated).

Regarding the insertion rates, we could somewhat relax the space-homogeneity without compromising the factorability of alignment probabilities. For example, the insertion rates could depend on the ancestries, $v(s,x)$ and $v(s,x+1)$, of sites flanking the event:

$$r_I(x,l;s,t) = g_I(v(s,x), v(s,x+1), l, t). \qquad (\mathrm{R8}-1.5)$$

These rates satisfy condition (i). Eq. (R8-1.5) combined with the space-homogeneous deletion rates, Eq. (R8-1.2), still gives an exit rate whose increment due to an indel event depends only on the inserted/deleted subsequence (and flanking sites) but not on the regions separated from it by at least a PAS. Hence the model also satisfies condition (ii), thus providing factorable alignment probabilities. Relaxing the space-homogeneity of deletion rates, however, is somewhat difficult, particularly because of condition (ii). In subsection R8-3, we will attempt it.

R8-2. Space-homogeneous model flanked by biologically essential sites/regions

The space-homogeneous models discussed above may decently approximate the evolution of a sequence region under no selective pressure. A real genome, however, is scattered with regions and sites under strong or weak purifying selection. Here, we consider one of the simplest such cases, in which biologically essential sites or regions flank a neutrally evolving region from both sides.[16] The insertion rates of this model are given by Eq. (R8-1.1) with the same domain, and the deletion rates are:

$$r_D(x_B, x_E; s, t) = \begin{cases} g_D(x_E - x_B + 1,\ t) & for\quad 1 \le x_B \le x_E \le L(s) \quad and \quad 1 \le x_E - x_B + 1 \le L_D^{CO}, \\ 0 & for\quad x_B \le 0, \quad x_E > L(s) \quad or \quad x_E - x_B + 1 > L_D^{CO}. \end{cases} \qquad (\mathrm{R8}-2.1)$$

The exit rate for this model is calculated as:

$$R_X^{ID}(s,t) = (L(s)-1)\sum_{l=1}^{L_I^{CO}} g_I(l,t) + \sum_{l=1}^{L_I^{CO}} \left(g_{I;L}(l,t) + g_{I;R}(l,t)\right) + \sum_{l=1}^{min\left\{L(s),\ L_D^{CO}\right\}} (L(s) - l + 1) g_D(l,t). \qquad (\mathrm{R8}-2.2)$$

For $L(s) \ge L_D^{CO}$, this is affine, and given by Eq. (R8-1.3) with exactly the same $A(t)$ as before and $B(t) = -\sum_{l=1}^{L_D^{CO}} (l-1) g_D(l,t) - \sum_{l=1}^{L_I^{CO}} g_I(l,t) + \sum_{l=1}^{L_I^{CO}} \left(g_{I;L}(l,t) + g_{I;R}(l,t)\right)$. Therefore, with such a sequence length, the alignment probability is still factorable even under this model. For $L(s) < L_D^{CO}$, in contrast, it exhibits a *non-affine* form:

$$R_X^{ID}(s,t) = (L(s)-1)\sum_{l=1}^{L_I^{CO}} g_I(l,t) + \sum_{l=1}^{L_I^{CO}} \left(g_{I;L}(l,t) + g_{I;R}(l,t)\right) + \sum_{l=1}^{L(s)} (L(s) - l + 1) g_D(l,t). \qquad (\mathrm{R8}-2.3)$$

Thus, in this case, condition (ii) will not be satisfied in general, whereas condition (i) is satisfied. This case gives the simplest example of a model with non-factorable PWA probabilities despite space-homogeneous rates of indels (as long as they are allowed). In a model with

non-factorable alignment probabilities, the "difference of exit rate differences":

$$\delta\delta R_X^{ID}\left(s''', s'', s', s, t\right) \equiv \delta R_X^{ID}\left(s''', s'', t\right) - \delta R_X^{ID}\left(s', s, t\right), \quad (\mathrm{R8}-2.4)$$

where $\langle s'| = \langle s|\hat{M}_{\nu_1}$, $\langle s''| = \langle s|\hat{M}_{\nu_2}$ and $\langle s'''| = \langle s|\hat{M}_{\nu_2}\hat{M}_{\nu_1}$, indicates the "degree of non-factorability" due to the pair of events, $\hat{M}_{\nu_1}$ and $\hat{M}_{\nu_2}$, that belong to different local histories. (See the argument around Eq. (5.2.6) of [32] for more details.)

R8-3. Model with rate-heterogeneity across regions

It is not only space-homogenous models but also some space-heterogeneous models that satisfy both conditions (i) and (ii), albeit partially. Here we give an example. We first define a set of non-overlapping regions, $\mathrm{E}_y(s_I) \equiv [x^0_{B;y}, x^0_{E;y}]$ (with $y = 1, \ldots, Y$), that existed in the initial state, $s_I \in S^{II}$. We define the "descendant region", $\mathrm{E}_y(s)$, of $\mathrm{E}_y(s_I)$ in a descendant state (s) by the closed interval, $\mathrm{E}_y(s) \equiv [x_{B;y}(s), x_{E;y}(s)]$, where $x_{B;y}(s)$ and $x_{E;y}(s)$ are the leftmost and the rightmost sites, respectively, among those descended from the sites in $\mathrm{E}_y(s_I)$. Then, based on them, we define an indel model whose rate parameters are given by:

$$r_I(x, l; s, t) = r_{I;Base}(x, l; s, t) + \sum\nolimits_{y=1}^{Y} \Delta r_I\left[E_y\right](x, l; s, t). \quad (\mathrm{R8}-3.1)$$

$$r_D(x_B, x_E; s, t) = r_{D;Base}(x_B, x_E; s, t) + \sum\nolimits_{y=1}^{Y} \Delta r_D\left[E_y\right](x_B, x_E; s, t). \quad (\mathrm{R8}-3.2)$$

Here, the "baseline" indel rates, $\{r_{I;Base}(x, l; s, t)\}_{x,l}$ and $\left\{r_{D;Base}(x_B, x_E; s, t)\right\}_{x_B, x_E}$, are given by Eq. (R8-1.5) and Eq. (R8-1.2), respectively. The region-specific increments, $\{\Delta r_I[\mathrm{E}_y](x, l; s, t)\}_{x,l}$ and $\left\{\Delta r_D\left[\mathrm{E}_y\right](x_B, x_E; s, t)\right\}_{x_B, x_E}$, can be non-zero *only within* $\mathrm{E}_y(s) \equiv [x_{B;y}(s), x_{E;y}(s)]$ defined above (panel a of Figure S3 in Additional file 1). Moreover, the increments can depend only on the portion of the sequence state within $\mathrm{E}_y(s)$. The increments can be negative, as long as the entire rates, Eqs. (R8-3.1,R8-3.2), are non-negative. From Eqs. (R8-3.1,R8-3.2), the exit rates can be decomposed as:

$$R_X^{ID}(s, t) = R_{X;Base}^{ID}(s, t) + \sum_{y=1}^{Y} \Delta R_X^{ID}\left[E_y\right](s, t). \quad (\mathrm{R8}-3.3)$$

Here,

$$R_{X;Base}^{ID}(s, t) = \sum\nolimits_{x=0}^{L(s)} \sum\nolimits_{l=1}^{L_I^{CO}} r_{I;Base}(x, l; s, t) + \sum\nolimits_{x_B=-L_D^{CO}+2}^{L(s)} \sum\nolimits_{x_E=\max\{x_B,1\}}^{x_B+L_D^{CO}-1} r_{D;Base}(x_B, x_E; s, t) \quad (\mathrm{R8}-3.4)$$

is the baseline exit rate. And

$$\Delta R_X^{ID}\left[\mathrm{E}_y\right](s, t) \equiv \sum\nolimits_{x=x_{B;y}(s)}^{x_{E;y}(s)-1} \sum\nolimits_{l=1}^{L_I^{CO}} \Delta r_I\left[\mathrm{E}_y\right](x, l; s, t) + \sum\nolimits_{x_B=x_{B;y}(s)}^{x_{E;y}(s)} \sum\nolimits_{x_E=x_B}^{x_{E;y}(s)} \Delta r_D\left[\mathrm{E}_y\right](x_B, x_E; s, t) \quad (\mathrm{R8}-3.5)$$

is the increment of the exit rate confined in, and dependent only on, $\mathrm{E}_y(s)$ ($y = 1, \ldots, Y$). As explained at the bottom of subsection R8-1, $R_{X;Base}^{ID}(s, t)$ alone gives factorable alignment probabilities. And the increments, $\{\Delta R_X^{ID}[\mathrm{E}_y](s, t)\}_{y=1,\ldots,Y}$, behave independently of the portions of sequence states outside $\mathrm{E}_y(s)$. Thus, if each indel event is *completely* confined in any of the $\mathrm{E}_y(s)$'s or in any spacer regions between neighboring $\mathrm{E}_y(s)$'s (Figure S3, panel a), the alignment probability can be expressed as the product of the overall factor and the contributions from $\mathrm{E}_y(s)$'s and within spacer regions. Even if some events within a $\mathrm{E}_y(s)$ are separated from the others by at least a PAS, they must be put together into a single local indel history (panel a). A complexity arises because deletions may stick out of a $\mathrm{E}_y(s)$, or even bridge two or more regions (panels b and c). The rates of such deletions and indels that are completely outside of the regions are given by the baseline rates. When a deletion sticks out of a region, the region will be extended to encompass the deletion, and all events within the extended region are lumped into a single local indel history (panel b). When a deletion bridges two or more regions, a "meta-region" encompassing all bridged regions is defined, and all events within the meta-region will form a single local indel history (panel c). In contrast, the indels completely outside of the regions should be independent of each other as long as they are separated by at least a PAS. Hence, under this model, the PWA probabilities are "factorable" in this somewhat non-trivial sense.

In Supplementary appendix SA-3 in Additional file 2, we explicitly showed that the probability of a LHS equivalence class via the recipe of [21] is identical to that calculated via our *ab initio* formulation. Although we assumed the space-homogeneity there, the proof can probably be extended to the model in this subsection as well, by slightly modifying the definition of the "chop-zone".

R9. Merits, possible extensions & applications, and outstanding issues

In this paper, we presented a theoretical formulation built up by tools that help mathematically precise

dissection of the *ab initio* calculation of alignment probabilities under genuine stochastic evolutionary models. Another merit of this formulation is that it gives intuitively clear pictures. For example, the insertion and deletion operators simply mathematically represent the intuitive pictures of the indels naturally acting on sequences (Fig. 3). Thus, the action of the rate operator, given by Eqs. (R3.6-R3.10) (or Eqs. (R3.11-R3.15)), is intuitively understandable as merely the collection of all possible single-mutational channels from a given sequence state (and some compensating terms). Then, the expansion formula for the action of the finite-time transition operator, Eqs. (R4.6,R4.7), can also be intuitively interpreted as the collection of contributions from all possible mutational processes starting with an initial sequence state. Importantly, this expansion was not posed via a hand-waving argument but rigorously derived as the solution of the defining equations of the model (Eqs. (R3.19-R3.21)), which justifies its *ab initio* status. And the integral equations, Eqs. (R4.4,R4.5), bridged the expansion's mathematically rigorous and intuitive aspects. Finally, the binary equivalence relations, Eqs. (R5.2a-R5.2d) (*e.g.*, Fig. 5), and their resulting LHS equivalence classes also allow intuitive interpretations as the invariance of the local effects of indels under their relative orders with spatially separate events. Therefore, the conditions for the factorability of PWA probabilities are also intuitively understandable. Although their mathematically rigorous derivations (in Supplementary appendix SA-2 in Additional file 2) might appear somewhat formidable, they are actually not so difficult once the aforementioned intuitive pictures are understood. Hence, by coupling the mathematical preciseness with the intuitive clarity, our theoretical formulation is expected to facilitate further advances of the study of *ab initio* alignment probabilities under genuine stochastic evolutionary models with some biological realism.

For clarity, this study focused only on indels among various mutational types, because indels are essential for creating a sequence alignment. If desired, however, our theoretical formulation could also incorporate substitutions ([31]; see also [42]). Moreover, the formulation could also be extended to incorporate other genome rearrangements, such as duplications and inversions. (See [31] for an initial, rudimentary attempt.) Such an extended formulation will provide tools to enable concrete analyses of "rate grammars" extended to incorporate genome rearrangements (briefly mentioned in [21]).

The practical use of our formulation depends on how efficiently it can calculate quite accurate alignment probabilities. Although the factorability of alignment probabilities will help greatly speed up the computation, the contribution from each local region (*e.g.*, Eq. (R6.8) or Eq. (SM-4.22) in Additional file 1) is still composed of infinitely many terms. Good news is that the first approximation of each local contribution, which is the summation of the terms from parsimonious local indel histories alone, is quite accurate, as long as the gap lengths and the branch lengths are at most moderate (Ezawa, unpublished; draft manuscripts: [40, 41]). Thus, considerably efficient computation of considerably accurate alignment probabilities may be possible based on our formulation. Especially, at least when the model is spatiotemporally homogeneous, our *ab initio* calculation was shown to be equivalent to the calculation of [21], with a caveat (see subsection R8-1). Thus, their dynamic programming (DP) may be applicable, possibly with some modifications, to a wider class of models with factorable probabilities.

Despite these favorable aspects, our theoretical formulation still has some limitations and outstanding issues. First, we did not examine whether our "sufficient and nearly necessary" set of conditions for the alignment probability factorization is exactly necessary and sufficient or not. Nor did we provide any counterexamples that violate our set of conditions and still have factorable alignment probabilities. Solving these problems may be interesting at least mathematically.

Second, although in this study we *tentatively* used the set of positive integers to represent the space of ancestry indices (Υ), it is obviously not the best choice. Finding, or establishing, a mathematical entity (either a set or a space) that is more suitable for representing Υ should be another mathematically interesting issue.

Third, in section R8 (and in section 5 of [32]), we only considered simple boundary conditions. Each sequence end was either freely mutable or flanked by a biologically essential region that allows no indels. These boundary conditions may remain good approximations if the subject sequences were extracted from well-characterized genomic regions. In real sequence analyses, however, the ends of the aligned sequences are often determined by artificial factors, such as the methods to sequence the genome, detect sequence homology, and annotate the sequences. Moreover, the constant cutoff lengths (L_I^{CO} and L_D^{CO}) were introduced just for the sake of simplicity, to broadly take account of the effects of various factors that suppress very long indels (such as selection, chromosome size, genome stability, *etc.*). In reality, it is much more likely that the cutoff lengths will vary across regions. Then, the alignment probabilities would be only approximately factorable, as in subsection R8-2. In order to pursue further biological realism and to enable more accurate sequence analyses, it should be inevitable to address these issues seriously. Eq. (R8-2.4) may be useful for such studies.

Fourth, we strongly caution the readers that, at this point, a naïve application of our formulation or its algorithmic implementation [41] to a *reconstructed* MSA is

fraught with high risks of incorrect predictions of indel histories, *etc*. This is because reconstructed MSAs, *even if they were reconstructed via state-of-the-art aligners* (reviewed, *e.g.*, in [43]), are known to be considerably erroneous (*e.g.*, [42, 44, 45]). Thus, it should be preferable to first develop a method or a program that accurately assesses and rectifies alignment errors, preferably by estimating the distribution of fairly likely alternative MSAs (as, *e.g.*, in [16, 46, 47]), before using our formulation to make some evolutionary or biological predictions.

Fifth, in this study, the phylogenetic tree was regarded as given. In many cases, however, the phylogenetic trees must also be inferred from the input sequence data. A theoretically ideal way would be to infer the joint distribution of MSAs and phylogenetic trees, as it is expected to minimize possible prediction biases (*e.g.*, [13, 48–50]). A major problem is that such an analysis would be tremendously time-consuming in general. At present, it is a totally open question whether our formulation can be adapted to infer a quite accurate joint distribution efficiently enough.

Conclusions

To the best of our knowledge, this is the first study to theoretically dissect the *ab initio* calculation of alignment probabilities under a *genuine* stochastic evolutionary model, which describes the evolution of an *entire* sequence via insertions and deletions (indels) along the time axis. The model handled here extends the previously most general evolutionary model, *i.e.*, the general form of the "substitution/inserton/deletion models" [21]. It should be noted that we did not impose any unnatural restrictions such as the prohibition of overlapping indels. Nor did we make the pre-proof assumption that the probability is factorable into the product of column-to-column transition probabilities or block-wise contributions. The essential tool introduced in this study was the operator representation of indels. This enabled us to shift the focus from the trajectory of sequence states (as in [21]) to the series of indel events, and to define local-history-set (LHS) equivalence classes of indel histories. Moreover, the operator representation also facilitated the adaptation of the time-dependent perturbation expansion (*e.g.*, [29, 30, 33]), which enabled us to express the probability of an alignment as a summation of probabilities over all alignment-consistent indel histories. Then, under a most general set of indel rate parameters, we exploited the LHS equivalence classes and found a "sufficient and nearly necessary" set of conditions on the indel rate parameters and exit rates under which the *ab initio* alignment probabilities can be factorized to provide a sort of generalized HMMs. We also showed that quite a wide variety of indel models could satisfy this set of conditions. Such models include not only the "long indel" model [21] and the indel model of a genuine molecular evolution simulator, Dawg [26], but also some sorts of models with rate variation across regions. Moreover, we explicitly showed (in Supplementary Appendix SA-3 in Additional file 2) that, as far as each LHS equivalence class is concerned, the probability calculated via the method of [21] is equivalent to that calculated via our *ab initio* formulation, at least under their spatiotemporally homogeneous indel model.

To summarize, by depending purely on the first principle and by providing intuitively clear pictures, this study established firm theoretical grounds that will help further advance the *ab initio* calculation of alignment probabilities under genuine stochastic evolutionary models with some biological realism. And our theoretical formulation will also provide other indel probabilistic models with a sound reference point, *provided that* there exist approximate methods that can quite accurately estimate the *ab initio* alignment probabilities fairly efficiently. Such approximate methods will be the subject of a related study (Ezawa, unpublished; draft manuscripts available [40, 41]).

Methods

Methodological details in this study are described in Supplementary methods in Additional file 1, or in Supplementary appendix in Additional file 2.

Endnotes

[1]In a sense, the models implemented in the genuine sequence evolution simulators (*e.g.*, [26–28]) could also be considered as special cases of this evolutionary model.

[2]This poses no problem here because the summation is always bounded from above (by a number less than or equal to unity) and all terms in the summation are non-negative, which guarantees the convergence of the expansion.

[3]However, the two manuscripts differ in some aspects, *e.g.*, their starting points. In [32], we described our model from scratch, as a continuous-time Markov model, and only briefly discussed its relationship with the general SID model [21]. In contrast, this manuscript explicitly presents our model as an extension of the general SID model, because that would put our model into the historical context of studies on indel evolutionary models.

[4]It may be worth mentioning, though, that the operator notation allows flexible representations of general state changes (by mutations). In the vector-matrix notation, this is possible only via abstract mutation matrices (and state vectors); concrete matrices (and vectors) can at most describe some fixed specific indel histories.

[5]In Eq.(R1.1), $\omega(\in\Omega)$ and $\omega'(\in\Omega)$, respectively, are the residues before and after the substitution. s_I in Eq.(R1.2) denotes the inserted subsequence, and s_D in Eq.(R1.3) denotes the deleted subsequence. And the equation, $s = s_L\omega s_R$, for example, states that the sequence s is formed by concatenating the three factors, s_L, ω and s_R.

[6]Besides, as claimed in [21], the rate grammar could be extended to formally describe a wide variety of evolutionary processes, including duplication, inversion and translocation.

[7]This should be understood as a brief representation. The precise representation is: $\check{s} = (\vec{v}^{\,t}, \vec{\omega}^{\,t})^t$, where A^t denotes the transposition (*i.e.*, swapping the roles of rows and columns) of matrix A. (See Figure 2b.)

[8]Inspired by profile HMMs (*e.g.*, [51]) and the idea of position-specific evolutionary rates [25], we originally devised ancestry indices in order to distinguish columns of a given alignment from one another. Their values themselves are not considered so important. Thus, it would be convenient to *re-assign* the ancestry indices as follows, after an indel process (or history) created an alignment, α, whether it is among only extant sequences or among sequence states at all nodes of the phylogenetic tree. (1) Re-assign 1, 2, ..., $|\alpha|$ (= the number of columns in α) to the sites corresponding to the columns of α, from left to right. (2) Re-assign $|\alpha| + 1$, $|\alpha| + 2$, ... to the "evanescent" sites with no corresponding alignment columns, again from left to right. This re-assignment can be considered as replacing a set of aligned sequence states created by an indel process (or history) with a "representative" set "equivalent" to the former set in the sense that both give the same homology structure [48]. The re-assignment may facilitate the understanding of the arguments in sections R5 through R8. (The figures in this paper do not undergo this re-assignment, though.)

[9]As far as the sequence (with state $\check{s}$) is concerned, $\hat{M}_D(x_B, x_E)$ with $x_B < 1$, is indistinguishable from $\hat{M}_D(1, x_E)$, and $\hat{M}_D(x_B, x_E)$ with $x_E > L(\check{s})$ is indistinguishable from $\hat{M}_D(x_B, L(\check{s}))$.

[10]In this paper, $\alpha(\ldots)$ (or $\alpha[\ldots]$) is intended to denote the *homology structure* [48] of an alignment, and thus the symbol doesn't care about details on the ancestry indices or residue states filling in the alignment other than the distinction of different columns in the alignment. And, hereafter, the terms "alignment", "PWA", and "MSA" mean their homology structures.

[11]Actually, we could also define the equivalence relationships between products of non-separated indel operators (non-exhaustively listed in Appendix A1 of [32]), and they may assist further theoretical developments. However, discussing these extra equivalences is beyond the scope of this manuscript.

[12]Such a maximum set does not necessarily consist of *all* PASs in the PWA. An example is given in subsection R8-3.

[13]The "phylogenetic correctness" condition guarantees that the sites aligned in a MSA column should share an ancestry. The condition could be rephrased as: "if a site corresponding to the column is present at two points in the phylogenetic tree, the site must also be present all along the shortest path connecting the two points".

[14]HMMs commonly use geometric distributions of sequence lengths. The uniform distribution may be a good approximation if we can assume that the ancestral sequence was sampled randomly from a chromosome of length L_C. In this case, the distribution of the sequence length, $L(s)(<< L_C)$, would be proportional to $(1 - (L(s) - 1)/L_C) \approx 1$.

[15]$\widetilde{\widetilde{\mathrm{M}}}_P[\alpha[s_1, s_2, \ldots, s_{N^X}]; s_0^{Root}; C_K|T]$ should be equivalent to $\widetilde{\widetilde{\mathrm{M}}}_P\left[\tilde{\Lambda}_\Psi^{ID}[C_\mathrm{K}; \alpha[s_1, s_2, \ldots, s_{N^X}]; T]|T\right]$ given in Eq.(4.2.9c) of [32], although the two expressions may appear quite different at first glance.

[16]Of course, we can also consider a model where only a single essential site/region flanks a rate-homogeneous neutral region. Alignment probabilities of this model can be shown to be factorable because the exit rate is an affine function of the sequence length (subsection 5.2 of [32]).

Additional files

Additional file 1: A PDF file that consists of Supplementary methods (sections SM-1 through SM-4), Figures S1, S2, S3, and Table S1. The sections of Supplementary methods provide detailed derivations of some essential results. Figures S1-S3 graphically illustrate via MSAs some important concepts and ideas introduced in the main text. Table S1 summarizes mathematical symbols commonly used in this paper. (PDF 11751 kb)

Additional file 2: A PDF file that consists of Supplementary appendix (sections SA-1, SA-2 and SA-3). The sections of Supplementary appendix provide mathematical derivations of some important results. (PDF 9581 kb)

Abbreviations

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

Acknowledgements

This study is dedicated to the late Dr. Keiji Kikkawa, a key pioneer of string theory of elementary particle physics and the author's best mentor. The author would like to greatly thank Dr. Dan Graur at the University of Houston and Dr. Giddy Landan at Heinrich-Heine University Düsseldorf for their indispensable assistance with this study. The author is grateful to Dr. Reed A. Cartwright at Arizona State University for his useful information and discussions that inspired this study. The author appreciates the logistic support and the feedback of Dr. Tetsushi Yada at Kyushu Institute of Technology, as well as the encouragement and the feedback of Dr. Toshiaki Takitani, a professor emeritus at Nagoya University. The author would also like to thank Dr. Ian Holmes at the University of California, Berkeley, for his critical comments on the previous version of the manuscript and his information on some relevant previous studies, both of which helped substantially improve the manuscript. The author is also grateful to all of the reviewers and the handling

editors of this manuscript and its predecessors, as their feedback definitely helped improve the manuscript.

Funding
This work was a part of the project, "Error Correction in Multiple Sequence Alignments", which was funded by US National Library of Medicine (grant number: LM010009-01 to Dan Graur and Giddy Landan, then at the University of Houston). The later stage of this work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan (grant numbers: MEXT KAKENHI Grant numbers 221S0002, 15H01358, both to Tetsushi Yada).

Availability of data and material
The data sets supporting the conclusions of this article are included within the article and its additional files.

Author's contributions
Not applicable because the paper was written by a single author (KE).

Author's information
The author (KE) used to be a mathematical physicist, who studied theoretical elementary particle physics and quantum gravitational theories from 1991 till 1999. Then, since 2002, after having studied theoretical biophysics from 1999 till 2002, he has studied molecular evolution (including population genetics), mainly focusing on homology-based computational analyses of DNA and protein sequences. For more detailed information, including the list of his publications, refer to his ORCID record [52].

Competing interests
The author declares that he has no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Received: 2 March 2016 Accepted: 26 May 2016
Published online: 11 August 2016

References

1. Graur D, Li WH. Fundamentals of Molecular Evolution. 2nd ed. Sunderland: Sinauer Associates; 2000.
2. Gascuel O, editor. Mathematics of Evolution and Phylogeny. New York: Oxford University Press; 2005.
3. Lynch M. The Origins of Genome Architecture. Sunderland: Sinauer Associates; 2007.
4. Britten RJ. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. P Natl Acad Sci USA. 2002;99:13633–5.
5. Britten RJ, Rowen L, Willians J, Cameron RA. Majority of divergence between closely related DNA samples is due to indels. P Natl Acad Sci USA. 2003;100:4661–5.
6. The International Chimpanzee Chromosome 22 Consotrium. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature. 2004;429:382–8.
7. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437:69–87.
8. Bishop MJ, Thompson EA. Maximum likelihood alignment of DNA sequences. J Mol Biol. 1986;190:159–65.
9. Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol. 1991;33:114–24.
10. Rivas E. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics. 2005;6:63.
11. Bradley RK, Holmes I. Transducers: an emerging probabilistic framework for modeling indels on trees. Bioinformatics. 2007;23:3258–62.
12. Miklós I, Novák Á, Satija R, Lyngsø R, Hein J. Stochastic models of sequence evolution including insertion-deletion events. Stat Methods Med Res. 2009;18:453–85.
13. Holmes I, Bruno WJ. Evolutionary HMMs: a Bayesian approach to multiple sequence alignment. Bioinformatics. 2001;17:803–20.
14. Holmes I. Using guide trees to construct multiple-sequence evolutionary HMMs. Bioinformatics. 2003;19:i147–57.
15. Bouchard-Côté A. A note on probabilistic models over strings: The linear algebra approach. Bull Math Biol. 2013;75:2529–50.
16. Herman JL, Novák Á, Lyngsø R, Szabó A, Miklós I, Hein J. Efficient representation of uncertainty in multiple sequence alignments using directed acyclic graphs. BMC Bioinformatics. 2015;16:108.
17. Thorne JL, Kishino H, Felsenstein J. Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol. 1992;34:3–16.
18. Miklós I, Toroczkai Z. An improved model for statistical alignment. In: Gascuel O, Moret BME, editors. WABI 2001, LNCS 2249. Heidelberg: Splinger-Verlag; 2001.
19. Cartwright RA. Problems and solutions for estimating indel rates and length distribution. Mol Biol Evol. 2009;26:473–80.
20. Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. Genome Res. 2008;18:298–309.
21. Miklós I, Lunter GA, Holmes I. A "long indel" model for evolutionary sequence alignment. Mol Biol Evol. 2004;21:529–40.
22. Kim J, Sinha S. Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. Bioinformatics. 2007;23:289–97.
23. Rivas E, Eddy SR. Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput Biol. 2008;4:e1000172.
24. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. PathoGenetics. 2008;1:4.
25. Rivas E, Eddy SR. Parameterizing sequence alignment with an explicit evolutionary model. BMC Bioinformatics. 2015;16:406.
26. Cartwright RA. DNA assembly with gap (Dawg): simulating sequence evolution. Bioinformatics. 2005;21:iii31–8.
27. Fletcher W, Yang Z. INDELible: A flexible simulator of biological sequence evolution. Mol Biol Evol. 2009;26:1879–88.
28. Strope CL, Abel K, Scott SD, Moriyama EN. Biological sequence simulation for testing complex evolutionary hypothesis: indel-Seq-Gen version 2.0. Mol Biol Evol. 2009;26:2581–93.
29. Dirac PAM. The Principles of Quantum Mechanics. 4th ed. London: Oxford University Press; 1958.
30. Messiah A. Quantum Mechanics, Volume II. (Translated from French to English by Potter J). Amsterdam: North-Holland; 1961.
31. Ezawa K, Graur D, Landan G. Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part IV: Incorporation of substitutions and other mutations. bioRxiv. 2015. doi:10.1101/023622. Accessed 4 Aug 2015.
32. Ezawa K, Graur D, Landan G. Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part I: Theoretical basis. bioRxiv. 2015. doi:10.1101/023598. Accessed 4 Feb 2016.
33. Messiah A. Quantum Mechanics, Volume 1. (Translated from French to English by Temmer GM). Amsterdam: North-Holland; 1961.
34. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81:2340–61.
35. Feller W. On the integro-differential equations of purely discontinuous markov processes. T Am Math Soc. 1940;48:488–515.
36. Redelings BD, Suchard MA. Joint Bayesian estimation of alignment and phylogeny. Syst Biol. 2005;54:401–18.
37. Chindelevitch L, Li Z, Blais E, Blanchette M. On the inference of parsimonious evolutionary scenarios. J Bioinform Comput Biol. 2006;4:721–44.
38. Diallo AB, Makarenkov V, Blanchette M. Exact and heuristic algorithms for the indel maximum likelihood problem. J Comput Biol. 2007;14:446–61.
39. Farris JS. Phylogenetic analysis under Dollo's law. Syst Zool. 1977;26:77–88.
40. Ezawa K, Graur D, Landan G. Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part II: Perturbation analyses. bioRxiv. 2015. doi:10.1101/023606. Accessed 4 Aug 2015.
41. Ezawa K, Graur D, Landan G. Perturbative formulation of general continuous-time Markov model of sequence evolution via insertions/deletions, Part III: Algorithm for first approximation. bioRxiv. 2015. doi:10.1101/023614. Accessed 4 Aug 2015.
42. Ezawa K. Characterization of multiple sequence alignment errors using complete-likelihood score and position-shift map. BMC Bioinformatics. 2016;17:133.
43. Notredame C. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol. 2007;3:e123.

44. Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008;320:1632–5.
45. Landan G, Graur D. Characterization of pairwise and multiple sequence alignment errors. Gene. 2009;441:141–7.
46. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 2008;18:1829–43.
47. Westesson O, Lunter G, Paten B, Holmes I. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. PLoS One. 2012;7:e34572.
48. Lunter GA, Miklós I, Drummond A, Jensen JL, Hein J. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics. 2005;6:83.
49. Suchard MA, Redelings BD. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics. 2006;22:2047–8.
50. Novák Á, Miklós I, Lyngsø R, Hein J. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics. 2008;24:2403–4.
51. Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press; 1998.
52. The ORCID register of Kiyoshi Ezawa. http://orcid.org/0000-0003-4906-8578. Accessed May 19, 2016.