DOI 10.1186/s12859-017-1665-1

BMC Bioinformatics

COMMENTARY Open Access

Solving the master equation for Indels

Ian H. Holmes

Abstract

Background: Despite the long-anticipated possibility of putting sequence alignment on the same footing as statistical phylogenetics, theorists have struggled to develop time-dependent evolutionary models for indels that are as tractable as the analogous models for substitution events.

Main text: This paper discusses progress in the area of insertion-deletion models, in view of recent work by Ezawa (BMC Bioinformatics 17:304, 2016); (BMC Bioinformatics 17:397, 2016); (BMC Bioinformatics 17:457, 2016) on the calculation of time-dependent gap length distributions in pairwise alignments, and current approaches for extending these approaches from ancestor-descendant pairs to phylogenetic trees.

Conclusions: While approximations that use finite-state machines (Pair HMMs and transducers) currently represent the most practical approach to problems such as sequence alignment and phylogeny, more rigorous approaches that work directly with the matrix exponential of the underlying continuous-time Markov chain also show promise, especially in view of recent advances.

Keywords: Phylogenetics, Alignment, Indels

Background

Models of sequence evolution, formulated as continuous-time discrete-state Markov chains, are central to statistical phylogenetics and bioinformatics. As descriptions of the process of nucleotide or amino acid substitution, their earliest uses were to estimate evolutionary distances [1], parameterize sequence alignment algorithms [2], and construct phylogenetic trees [3]. Variations on these models, including extra latent variables, have been used to estimate spatial variation in evolutionary rates [4, 5]; these patterns of spatial variation have been used to predict exon structures of protein-coding genes [6, 7], foldback structure of non-coding RNA genes [8, 9], regulatory elements [10], ultra-conserved elements [11], protein secondary structures [12], and transmembrane structures [13]. They are widely used to reconstruct ancestral sequences [14–22], a method that is finding increasing application in synthetic biology [16, 20–22]. Trees built using substitution models used to classify species [23], predict protein function [24], inform conservation efforts [25], or identify novel pathogens [26]. In the analysis of rapidly evolving pathogens, these methods are used to uncover population histories [27], analyze transmission dynamics [28], reconstruct key transmission events [29], and predict future evolutionary trends [30]. There are many other applications; the ones listed above were selected to give some idea of how influential these models have been.

Continuous-time Markov chains describe evolution in a state space Φ, for example the set of nucleotides $\Phi = \{A,C,G,T\}$. The stochastic process $\phi(t)$ at any given instant of time, t, takes one of the values in Φ. Let $\vec{p}(t)$ be a vector describing the marginal probability distribution of the process at a single point in time: $\vec{p}_i(t) = P(\phi(t) = i)$. The time-evolution of this vector is governed by a *master equation*

$$\frac{d}{dt}\vec{p}(t) = \vec{p}(t)\mathbf{R} \qquad (1)$$

where, for $i,j \in \Phi$ and $i \neq j$, $\mathbf{R}_{i,j}$ is the instantaneous rate of mutation from state i to state j. For probabilistic normalization of Eq. 1, it is then required that

$$\mathbf{R}_{i,i} = -\sum_{j\in\Phi, j\neq i} \mathbf{R}_{i,j}$$

The probability distribution of this process at equilibrium is given by the vector $\vec{\pi}$, which must satisfy the equation

$$\vec{\pi}\mathbf{R} = \vec{0}$$

Correspondence: ihh@berkeley.edu
Dept of Bioengineering, University of California, 94720 Berkeley, USA

The general solution to Eq. 1 can be written $\vec{p}(t) = \vec{p}(0)\mathbf{M}(t)$ where $\mathbf{M}(t)$ is the *matrix exponential*

$$\mathbf{M}(t) = \exp(\mathbf{R}t) \tag{2}$$

Entry $\mathbf{M}_{i,j}(t)$ of this matrix is the probability $P(\phi(t) = j|\phi(0) = i)$ that, conditional on starting in state i, the system will after time t be in state j. It follows, by definition, that this matrix satisfies the Chapman-Kolmogorov forward equation:

$$\mathbf{M}(t)\mathbf{M}(u) = \mathbf{M}(t+u) \tag{3}$$

That is, if $\mathbf{M}_{i,j}(t)$ is the probability that state i will, after a finite time interval t, have evolved into state j, and $\mathbf{M}_{j,k}(u)$ is the analogous probability that state j will after time u evolve into state k, then summing out j has the expected result:

$$\sum_{j \in \Phi} \mathbf{M}_{i,j}(t)\mathbf{M}_{j,k}(u) = \mathbf{M}_{i,k}(t+u) \quad \forall i,k \in \Phi$$

This is one way of stating the defining property of a Markov chain: its lack of historical memory previous to its current state. Equation 1 is just an instantaneous version of this equation, and Eq. 3 is the same equation in matrix form.

The conditional likelihood for an ancestor-descendant pair can be converted into a phylogenetic likelihood for a set of extant taxon states S related by a tree T, as follows. (I assume for convenience that T is a binary tree, though relaxing this constraint is straightforward.)

To compute the likelihood requires that one first computes, for every node n in the tree, the probability $\vec{F}_i(n)$ of all observed states at leaf nodes descended from node n, conditioned on node n being in state i. This is given by Felsenstein's *pruning recursion*:

$$\vec{F}(n) = \begin{cases} \left(\mathbf{M}(t_{nl})\vec{F}(l)\right) \circ \left(\mathbf{M}(t_{nr})\vec{F}(r)\right) & \text{if } n \text{ is an internal node with children } l, r \\ \vec{\Delta}(s_n) & \text{if } n \text{ is a leaf node in state } s_n \end{cases} \tag{4}$$

where t_{mn} denotes the length of the branch from tree node m to tree node n. I have used the notation $\vec{\Delta}(j)$ for the unit vector in dimension j, and the symbol $\circ$ to denote the *Hadamard product* (also known as the *pointwise product*), defined such that for any two vectors $\vec{u}, \vec{v}$ of the same size: $(\vec{u} \circ \vec{v})_i = \vec{u}_i\vec{v}_i$.

Supposing that node 1 is the root node of the tree, and that the distribution of states at this root node is given by $\vec{\rho}$, the likelihood can be written as

$$P(S|T, \mathbf{R}, \vec{\rho}) = \vec{\rho} \cdot \vec{F}(1) \tag{5}$$

where $\vec{u} \cdot \vec{v}$ denotes the scalar product of $\vec{u}$ and $\vec{v}$. It is common to assume that the root node is at equilibrium, so that $\vec{\rho} = \vec{\pi}$.

As mentioned above, this mathematical approach is fundamental to statistical phylogenetics and many applications in bioinformatics. For small state spaces Φ, such as (for example) the 20 amino acids or 61 sense codons, the matrix exponential $\mathbf{M}(t)$ in Eq. 2 can be solved exactly and practically by the technique of spectral decomposition (i.e. finding eigenvalues and eigenvectors). Such an approach informs the Dayhoff PAM matrix. It was also solved for certain specific parametric forms of the rate matrix $\mathbf{R}$ by Jukes and Cantor [1], Kimura [31], Felsenstein [3], and Hasegawa et al. [32], among others. This approach is used by all likelihood-based phylogenetics tools, such as RevBayes [33], BEAST [34], RAxML [35], HyPhy [36], PAML [37], PHYLIP [38], TREE-PUZZLE [39], and XRate [40]. Many more bioinformatics tools use the Dayhoff PAM matrix or other substitution matrix based on an underlying master equation of the form Eq. 1.

Homogeneity, stationarity, and reversibility

There exists a deep literature on Markov chains, to which this brief survey cannot remotely do justice, but several concepts must be mentioned in order to survey progress in this area.

A Markov chain is *time-homogeneous* if the elements of the rate matrix $\mathbf{R}$ in Eq. 1 are themselves independent of time. If a Markov chain is time-homogeneous and is known to be in equilibrium at a given time, for example $\vec{p}(0) = \vec{\pi}$, then (absent any other constraints) it will be in equilibrium at all times; such a chain is referred to as being *stationary*.

The time-scaling of these models is somewhat arbitrary: if the time parameter t is replaced by a scaled version t/κ, while the rate matrix $\mathbf{R}$ is replaced by $\mathbf{R}\kappa$, then the likelihood in Eq. 2 is unchanged. For some models, the rate is allowed to vary between sites [4, 5].

A Markov chain is *reversible* if it satisfies the instantaneous *detailed balance* condition $\vec{\pi}_i\mathbf{R}_{i,j} = \vec{\pi}_j\mathbf{R}_{j,i}$, or its finite-time equivalent $\vec{\pi}_i\mathbf{M}_{i,j} = \vec{\pi}_j\mathbf{M}_{j,i}$. This amounts to a symmetry constraint on the parameter space of the chain (specifically, the matrix $\mathbf{S}$ with elements $\mathbf{S}_{i,j} = \sqrt{\vec{\pi}_i/\vec{\pi}_j}\mathbf{R}_{i,j}$ is symmetric) which has several convenient advantages: it effectively halves the number of parameters that must be determined, it eases some of the matrix manipulations (symmetric matrices have real eigenvalues and the algorithms to find them are generally more stable), and it allows for some convenient manipulations, such as the so-called *pulley principle* allowing for arbitrary re-rooting of the tree [3]. From another angle, however, these supposed advantages may be viewed as drawbacks: reversibility is a simplification which ignores some unreversible aspects of real data, limits the expressiveness of the model, and makes the root node placement statistically unidentifiable.

Stationarity has similar advantages and drawbacks. If one assumes the process was started at equilibrium, that is one less set of parameters to worry about (since the equilibrium distribution is implied by the process itself), but it also renders the model less expressive and makes some kinds of inference impossible.

The early literature on substitution models involved generalizing from rate matrices $\mathbf{R}$ characterized only by a single rate parameter [1], to symmetry-breaking versions that allowed for different transition and transversion rates [31], non-uniform equilibrium distributions over nucleotides [3], and combinations of the above [32]. These models are all, however, reversible. A good deal of subsequent research has gone into the problem, in various guises, of generalizing results obtained for reversible, homogeneous and/or stationary models to the analogous irreversible, nonhomogeneous and nonstationary models. For examples, see [30, 41–43].

From individual residues to whole sequences

The question naturally arises: how to extend the model to describe the evolution of an entire sequence, not just individual sites? In such cases, when one talks about $\mathbf{M}(t)_{i,j}$ "the likelihood of an ancestor-descendant pair (i,j)" (or, more precisely, the probability that—given the system starts in ancestral state i—it will after time t have evolved into descendant state j) one must bear in mind that the states i and j now represent not just individual residues, but entire sequences.

As long as the allowed mutations are restricted to point substitutions and their mutation rates are independent of flanking context, then the extension to whole sequences is trivially easy: one can simply multiply together probabilities of independent sites.

However, many kinds of mutation violate this assumption of site independence; most notably context-dependent substitutions and indels, where the rates depend on neighboring sites. For these mutations the natural approach is to extend the state space Φ to be the set of all possible sequences over a given alphabet (for example, the set of all DNA or protein sequences). This state space is (countably) infinite; Eqs. 1-4 can still be used on an infinite state space, but solution by brute-force enumeration of eigenvalues and eigenvectors is no longer feasible, except in special cases where there is explicit structure to the rate matrix that allows identification of the eigensystem by algebraic approaches [44, 45].

It has turned out that whole-sequence evolutionary models have proved quite challenging for theorists. There is extensive evidence suggesting that indels, in particular, can be profoundly informative to phylogenetic studies, and to applications of phylogenetics in sequence analysis [46–51]. The field of efforts to unify alignment and phylogeny, and to build a theoretical framework for the evolutionary analysis of indels, has been dubbed *statistical alignment* by Hein, one of its pioneers [52]. Recent publications by Ezawa [53–55] and Rivas and Eddy [56] have highlighted this problem once again, directly leading to the present review.

Main text

In this paper I focus only on "local" mutations: mostly indel events (which may include local duplications), but also context-dependent substitutions. This is not because "nonlocal" events (such as rearrangements) are unimportant, but rather that they tend to defy phylogenetic reconstruction due to the rapid proliferation of possible histories after even a few such events [57].

The discussion here is separated into two parts. In the first part, I discuss the master equation (Eq. 1) and exact solutions thereof (Eq. 2), along with various approximations and their departure from the Chapman-Kolmogorov ideal (Eq. 3). This is an area in which recent progress has been reported in this journal. In the second part, I review the extension from pairwise probability distributions to phylogenetic likelihoods of multiple sequences, using analogs of Felsenstein's pruning recursion (Eq. 4).

Solving the master equation

This section begins with various approaches to finding the time-dependent probability distribution of gap lengths in a pairwise alignment, under several evolutionary models.

Exactly solved models on k-mer strings

As an approach to models on strings of unbounded length, one can consider short motifs of k residues. These can still be considered as finite state-space models; for example, a k-nucleotide model has 4^k possible states. Several such models have been analyzed, including models on codons where $k = 3$ [47, 58], dinucleotides involved in RNA base-pairs where $k = 2$ [59–61], and models over sequences of arbitrary length k [44, 62].

Mostly, these models handle short sequences (motifs) and do not allow the sequence length to change over time (so they model only substitutions and not indels). Some of the later models do allow the sequence length to change via insertions or deletions [62] though these models have not yet been analyzed in a way that would allow the computation of alignment likelihoods for sequences of realistic lengths.

Exactly solved models on strings of unbounded length

It is a remarkable reflection on the extremely challenging nature of this problem that, to date, the only exactly solved indel model on strings is the TKF91 model, named after the authors' initials and date of publication of this seminal paper [63]. While there has been progress in developing approximate models in the 25 years since the publication of this paper, and in extending it from pairwise to multiple sequence alignment, it remains the only model for which

1. the state space Φ is the set of all sequences (strings) over a finite alphabet,
2. the state space is ergodically explored by substitutions and indels (so there is a valid alignment and evolutionary trajectory between any two sequences $\phi(0)$ and $\phi(t)$),
3. Equation 2 can be calculated exactly (specifically, as a sum over alignments, where the individual alignment likelihoods can be written in closed form).

The TKF91 model allows single-residue context-independent events only. These include (i) single-residue substitutions, (ii) single-residue insertions (with the inserted residue drawn from the equilibrium distribution of the substitution process), and (iii) single-residue deletions (whose rates are independent of the residue being deleted). The rates of all these mutation events are independent of the flanking sequence.

This process is equivalent to a linear birth-death model with constant immigration [64]. Thorne et al. showed that an ancestral sequence can be split into independently evolving zones, one for each ancestral residue (or "links", as they call them). This leads to the very appealing result that the length distribution for observed gaps is geometric, which conveniently allows the joint probability $P(\phi(0), \phi(t))$ to be expressed as a paired-sequence Hidden Markov Model or "Pair HMM" [65]. The conditional probability $P(\phi(t)|\phi(0))$ can similarly be expressed as a weighted finite-state transducer [66–68]. Some interesting discussion of why the TKF91 model should be solvable at all can be found in [69] and in [42].

There are several variations on the TKF91 model. The case where there are no indels at all, only substitutions, can be viewed as a special case of TKF91, and can of course be solved exactly, as is well known. Another variation on the TKF91 model constrains the total indel rate to be independent of sequence length [70].

In the following section I cover some variants that use different state spaces.

Exactly solved models on state spaces other than strings

It is difficult to extend TKF91 to more realistic models wherein indels (or substitutions) can affect multiple residues at once. In such models, the fate of adjacent residues is no longer independent, since a single event can span multiple sites.

As a way around this difficulty, several researchers have developed evolutionary models where the state is not a pure DNA or protein sequence, but includes some extra "hidden" information, such as boundaries, markers or other latent structure. In some of these models the sequence of residues is replaced by a sequence of indivisible fragments, each of which can contain more than one residue [56, 69, 71]. These includes the TKF92 model [71] which is, essentially, TKF91 with residues replaced by fragments (so the alphabet itself is the countably infinite set of all sequences over some other, finite alphabet). Other models approximate indels as a kind of substitution that temporarily hides a residue, by augmenting the DNA or protein alphabet with an additional gap character [72–74].

These models can be used to calculate some form of likelihood for a pairwise alignment of two sequences, but since this likelihood is not derived from an underlying instantaneous model of indels, the equations do not, in general, satisfy the Chapman-Kolmogorov forward Eq. (3). That is, the probability of evolving from i to k comes out differently depending on whether or not one conditions on an intermediate sequence j. Clearly, something about this "seems wrong": the failure to obey Eq. 3 illustrates the ad hoc nature of these approaches. Ezawa [53] describes the Chapman-Kolmogorov property (Eq. 3) as *evolutionary consistency*; it can also be regarded as being the defining property of any correct solution to a continuous-time Markov chain. The above-mentioned approaches may be evolutionarily consistent if the state space is allowed to include the extra information that is introduced to make the model tractable, such as fragment boundaries. Lèbre and Michel have criticized other aspects of the Rivas-Eddy 2005 and 2008 models [73, 74]; in particular, incomplete separation of the indel and substitution processes [42].

Models which allow for heterogeneity of indel and substitution rates along the sequence also fall into this category of latent variable models. The usual way of allowing for such spatial variation in substitution models is to assume a latent rate-scaling parameter associated with each site [4, 5]. For indel models, this latent information must be extended to include hidden site boundaries [56].

Exactly solved models on graphs

Another variation on TKF91 is the TKF Structure Tree, which describes the evolutionary behavior of RNA structures with stem and loop regions which are subject to insertion and deletion [75]. Rather than describing the evolution of a sequence, this model essentially captures the time-evolution grammar of a tree-like graph whose individual edges are evolving according to the TKF91 model. Other evolutionary models have made use of graph grammars, for example to model pseudoknots [76] or context-dependent indels [77].

Finite-event trajectories

In tackling indel models where the indel events can insert or delete multiple residues at once, several authors have used the approximation that indels never overlap, so that any observed gap corresponds to a single indel event. This approximation, which is justified if one is considering evolutionary timespans $t \ll 1/(\delta\ell)$ where δ is the indel rate per site and ℓ is the gap length, considerably simplifies the task of calculating gap probabilities [67, 78–83].

At longer timescales, it is necessary to consider multiple-event trajectories, but (as a simplifying approximation) one can still truncate the trajectory at a finite number of events. A problem with this approach is that many different trajectories will generally be consistent with an observed mutation. Summing over all such trajectories, to compute the probability of observing a particular configuration after finite time (e.g. the observed gap length distribution), is a nontrivial problem.

In analyzing the *long indel* model, a generalization of TKF91 with arbitrary length distributions for instantaneous indel events, Miklós et al. [84] make the claim that the existence of a conserved residue implies the alignment probability is factorable at that point (since no indel has ever crossed the boundary). They use a numerical sum over indel trajectories to approximate the probability distribution of observed gap lengths. Although they used a reversible model, their approach generalizes readily to irreversible models. This work builds on an earlier model which allows long insertions, but only single-residue deletions [85]. Recent work by Ezawa has put this finite-event approximation on a more solid footing by developing a rigorous algebraic definition of equivalence classes for event trajectories [53–55].

Solutions obtained using finite-event approximations will not exactly satisfy Eq. 3. There will be some error in the probability, and in general the error will be greater on longer branches, as the main assumption behind the approximation (that there are no overlapping indels in the time interval, or that there is a finite limit to the number of overlapping indels) starts to break down. However, since these are principled approximations, it should be possible to form some conclusions as to the severity of the error, and its dependence on model parameters. Simulation studies have also been of some help in assessing the error of these approximations.

Taylor series approximations

For context-dependent substitution processes, such as models that include methylation-induced CpG-deamination, a clever approach was developed in [44]. Rather than considering a finite-event trajectory, they develop an explicit Taylor series for the matrix exponential (Eq. 2) and then truncate this Taylor series. Specifically, the rate matrix for a finite-length sequence is constructed as a sum of rate matrices operating locally on the sequence, using the Kronecker sum $\oplus$ and Kronecker product $\otimes$ to concatenate rate matrices. These operators may be understood as follows, for an alphabet of N symbols: suppose that $\mathcal{M}_m$ is the set of all matrices indexed by m-mers, so that if $\mathbf{A} \in \mathcal{M}_m$, then $\mathbf{A}$ is an $N^m \times N^m$ matrix. Let i, j be m-mers, k, l be n-mers, and ik, jl the concatenated $m + n$-mers. If $\mathbf{A} \in \mathcal{M}_m$ and $\mathbf{B} \in \mathcal{M}_n$ then $\mathbf{A} \oplus \mathbf{B}$ and $\mathbf{A} \otimes \mathbf{B}$ are both in $\mathcal{M}_{m+n}$ and are specified by

$$(\mathbf{A} \otimes \mathbf{B})_{ik,jl} = \mathbf{A}_{i,j}\mathbf{B}_{k,l} \tag{6}$$

$$(\mathbf{A} \oplus \mathbf{B})_{ik,jl} = \delta(k = l)\mathbf{A}_{i,j} + \delta(i = j)\mathbf{B}_{k,l} \tag{7}$$

where $\delta(i = j) = 1$ if $i = j$ and 0 if $i \neq j$. Furthermore, suppose $\mathbf{O}_n \in \mathcal{M}_n$ is the $N^k \times N^k$ null matrix containing only zeroes. Then $\mathbf{O}_m \oplus \mathbf{B}$ commutes with $\mathbf{A} \oplus \mathbf{O}_n$, and $\exp(\mathbf{A} \oplus \mathbf{B}) = \exp(\mathbf{A}) \otimes \exp(\mathbf{B})$. The rate matrix $\mathbf{R}$ for a length-L sequence operated on locally by a context-sensitive rate matrix $\mathbf{A} \in \mathcal{M}_m$ can be written as a sum of the form

$$\mathbf{R} = \sum_{n=0}^{L-m} \mathbf{O}_n \oplus \mathbf{A} \oplus \mathbf{O}_{L-m-n}$$

Commuting terms in the Taylor series for $\exp(\mathbf{R}t)$ can then be systematically rearranged into a quickly-converging dynamic programming recursion. This approach was first used by [44] and further developed including model-fitting algorithms [86] application to phylogenetic trees [87] and discussion of the associated eigensystem [45, 62]. It remains to be seen to what extent such an approach offers a practical solution for general indel models, where the instantaneous transitions are between sequences of differing lengths.

Simulation studies

Such is the difficulty of solving long indel models that several authors have performed simulations to investigate the empirical gap length distributions that are observed after finite time intervals for various given instantaneous indel-rate models. These observed gaps can arise from multiple overlapping indel events, in ways that have so far defied straightforward algebraic characterization.

In recent work, Rivas and Eddy [56] have shown that if an underlying model has a simple geometric length distribution over instantaneous indel events, the observed gap length distribution at finite times (accounting for the possibility of multiple, overlapping indels) cannot be geometric. Rivas and Eddy report simulation studies supporting this result, and go on to propose several models incorporating hidden information (such as fragment boundaries, *a la* TKF92) which have the advantage of being good fits to HMMs for their finite-time distributions.

It has long been known that the lengths of empirically-observed indels are more accurately described by a power-law distribution than a geometric distribution [46, 47, 88–91] and that alignment algorithms may benefit from using such length distributions, which lead to generalized convex gap penalties, rather than the computationally more convenient geometric distribution, which leads to an affine or linear gap penalty [92, 93]. For molecular evolution purposes in particular, it is known that overreliance on affine gap penalties leads to serious underestimates of the true lengths of natural indels [94]. For almost as long, it has been known that using a mixture of geometric distributions, or (considered in score space rather than probability space) a piecewise linear gap penalty, mitigates some of these problems in sequence alignment [94–96]. Taken together, these results suggest that simple HMM-like models, which are most efficient at modeling geometric length distributions, may be fundamentally limited in their ability to fully describe indels; that adding more states (yielding length distributions that arise from sums of geometric random variates, such as negative binomial distributions, or mixtures of geometric distributions) can lead to an improvement; and that generalized HMMs, which can model arbitrary length distributions at the cost of some computational efficiency [97], may be most appropriate. For example, the abovementioned "long indel" model of Miklós et al. uses a generalized Pair HMM [84], as does the HMM of [98]. It is even conceivable that some molecular evolution studies in the future will abandon HMMs altogether, although they remain very convenient for many applications.

The recent work of Ezawa has some parallels, but also differences, to the work of Rivas and Eddy [53–55]. Ezawa criticizes over-reliance on HMM-like models, and insists on a systematic derivation from simple instantaneous models. He puts the intuition of Miklós et al [84] on a more formal footing by introducing an explicit notation for indel trajectories and the concept of "local history set equivalence classes" for equivalent trajectories. Ezawa uses this concept to prove that alignment likelihoods for long-indel and related models are indeed factorable, and investigates, by numerical computation and analysis (with confirmation by simulation), the relative contribution of multiple-event trajectories to gap length distributions. Ezawa's results also show that the effects on the observed indel lengths due to overlapping indels become more significant as the indels get larger, making the problem particularly acute for genomic alignments where indels can be much larger than in proteins.

A number of excellent, realistic sequence simulators are available including DAWG [99], INDELible [100], and indel-Seq-Gen [101].

Extending from pairs to trees

Consider now the extension of these results from pairwise alignments, such as TKF91 and the "long indel" model, to multiple alignments (with associated phylogenies). Some of the approaches to this problem use Markov Chain Monte Carlo (MCMC); some of the approaches use finite-state automata; and there is also some overlap between these categories (i.e. MCMC approaches that use automata).

Approaches based on MCMC sampling

MCMC is the most principled approach to integrating phylogeny with multiple alignment. In principle an MCMC algorithm for phylogenetic alignment can yield the posterior distribution of alignments, trees, and parameters for any model whose pairwise distribution can be computed. This includes long indel models and also, in principle, other effects such as context-dependent substitutions.

Of the MCMC methods reported in the literature, some just focus on alignment and ancestral sequence reconstruction [65]; others on simultaneous alignment and phylogenetic reconstruction [79–81, 83, 102, 103]; some also include estimation of evolutionary parameters such as dN/dS [104]; and some (focused on RNA sequences) attempt prediction of secondary structure [105, 106].

In practise these mostly use HMMs, or dynamic programming of some form, in common with the methods of the following section. It is of course possible to use HMM-based or other MCMC approaches to propose candidate reconstructions, and then to accept or reject those proposals (in the manner of Metropolis-Hastings or importance sampling) using a more realistic formulation of the indel likelihood. Ezawa's methods, and others that build on them or are related to them, may be useful in this context. For example, Ezawa's formulation was used to calculate the indel component of the probability of a fixed multiple sequence alignment (MSA) resulting from sequence evolution along a fixed tree [53]. He also developed an algorithm to approximately calculate the indel component of the MSA probability using all MSA-compatible parsimonious indel histories [54], and applied it to some analyses of simulated MSAs [107]. Using such realistic likelihood calculations as a post-processing "filter" for coarser, more rapid MCMC approaches that sample the space of possible reconstructions could be a promising approach.

Approaches based on automata theory

The dynamic programming recursion for pairwise alignment reported for the TKF91 model [63] can be exactly extended to alignment of multiple sequences given a tree [108, 109]. This works essentially because the TKF91 joint distribution over ancestor and descendant sequences can be represented as a Pair HMM; the multiple-sequence version is a multi-sequence HMM [65].

This approach can be generalized, using *finite-state transducer* theory. Transducers were originally developed as modular representations of sequence-transforming operations for use in speech recognition [110]. In bioinformatics, they offer (among other things) a systematic way of extending HMM-like pairwise alignment likelihoods to trees [67, 68, 111, 112]. Other applications of transducer models in bioinformatics have included copy number variation in tumors [113], protein family classification [114], DNA-protein alignment [115] and error-correcting codes for DNA storage [116].

A finite-state transducer is a state machine that reads an input tape and writes to an output tape [117]. A probabilistically weighted finite-state transducer is the same, but its behavior is stochastic [110]. For the purposes of bioinformatics sequence analysis, a transducer can be thought of as being just like a Pair HMM; except where a Pair HMM's transition and emission probabilities have been normalized so as to describe joint probabilities, a transducer's probabilities are normalized so as to describe conditional probabilities like the entries of matrix $\mathbf{M}(t)$ (Eq. 2). More specifically, if i and j are sequences, then one can define the matrix entry $A_{i,j}$ to be the Forward score for those two sequences using transducer $\mathcal{A}$. Thus, the transducer is a compact encoding for a square matrix of countably infinite rank, indexed by sequence states (rather than nucleotide or amino acid states).

The utility of transducers arises since for many purposes they can be manipulated analogously to matrices, while being more compact than the corresponding matrix (as noted above, matrices describing evolution of arbitrary-length sequences are impractically—or even infinitely—large). If $\mathcal{A}$ and $\mathcal{B}$ are finite transducers encoding (potentially infinite) matrices $\mathbf{A}$ and $\mathbf{B}$, then there is a well-defined operation called *transducer composition* yielding a finite transducer $\mathcal{AB}$ that represents the matrix product $\mathbf{AB}$. There are other well-defined transducer operations corresponding to the various other linear algebra operations used in this paper: the Hadamard product ($\circ$) corresponds to *transducer intersection*, the Kronecker product ($\otimes$) corresponds to *transducer concatenation*, and the scalar product ($\cdot$) and the unit vector ($\vec{\Delta}$) can also readily be constructed using transducers. Consequently, Eq. 4 can be interpreted directly in terms of transducers [67, 68, 82].

This has several benefits. One is theoretical unification: Eq. 4, using the above linear algebra interpretation of transducer manipulations, turns out to be very similar to Sankoff's algorithm for phylogenetic multiple alignment [118]. Thus is a famous algorithm in bioinformatics unified with a famous algorithm in likelihood phylogenetics by using a tool from computational linguistics. (This excludes the RNA structure-prediction component of Sankoff's algorithm; that can, however, be included by extending the transducer framework to pushdown automata [119].) Practically, the phylogenetic transducer can be used for alignment [79, 81], parameter estimation [104], and ancestral reconstruction [67], with promising results for improved accuracy in multiple sequence alignment [112].

More broadly, one can think of the transducer as being in a family of methods that combine phylogenetic trees (modeling the temporal structure of evolution) with automata theory, grammars, and dynamic programming on sequences (modeling the spatial structure of evolution). The TKF Structure Tree, mentioned above, is in this family too: it can be viewed as a context-free grammar, or as a transducer with a pushdown stack [75].

The HMM-like nature of TKF91, and the ubiquity of HMMs and dynamic programming in sequence analysis, has motivated numerous approaches to indel analysis based on Pair HMMs [56, 71, 74, 78, 120], as well as many other applications of phylogenetic HMMs [6, 7, 12, 121, 122] and phylogenetic grammars [8, 10, 40, 60, 123, 124]. In most of these models, an alignment is assumed fixed and the HMM or grammar used to partition it; however, in principle, one can combine the ability of HMMs/grammars to model indels (and thus impute alignments) with the ability to partition sequences into differently evolving regions.

Conclusions

The promise of using continuous-time Markov chains to model indels has been partially realized by automata-theoretic approaches based on transducers and HMMs. Recent work by Rivas and Eddy [56] and by Ezawa [53–55] may be interpreted as both good and bad news for automata-theoretic approaches.

It appears that closed-form solutions for observed gap length distributions at finite times, and in particular the geometric distributions that simple automata are good at modeling, are still out of reach for realistic indel models, and indeed (for simple models) have been proven impossible [56]. Further, simulation results have demonstrated that geometric distributions are not a good fit to the observed gap length distributions when the underlying indel model has geometrically-distributed lengths for its instantaneous indel events [56]. If the lengths of the instantaneous indels follow biologically plausible power-law distributions, the evolutionary effects due to overlapping indels become larger as the gaps grow longer [54].

That is the bad news (at least for automata). The good news is that the simulation results also suggest that, for short branches and/or gaps (such that indels rarely overlap), the error may not be too bad to live with. Approximate-fit approaches that are common in Pair HMM modeling and pairwise sequence alignment—such as using a mixture of geometric distributions to approximate a gap length distribution (yielding a longer tail than can be modeled using a pure geometric distribution)—may help bridge the accuracy gap [96]. Given the power of automata-theoretic approaches, the best way forward (in the absence of a closed-form solution) may be to embrace such approximations and live with the ensuing error.

Interestingly, the authors of the two recent simulation studies that prompted this commentary come to different conclusions about the viability of automata-based dynamic programming approaches. Ezawa [53, 54], arguing that realism is paramount, advocates deeper study of the gap length distributions obtained from simple instantaneous models—while acknowledging that such gap length distributions may be more difficult to use in practice than the simple geometric distributions offered by HMM-like models. Rivas and Eddy [56], clearly targeting applications (particularly those such as profile HMMs), work backward from HMM-like models toward evolutionary models with embedded hidden information. These models may be somewhat mathematically contrived, but are easier to tailor so as to model effects such as position-specific conservation, thus trading (in a certain sense) purism for expressiveness.

Whichever approach is used, these results are unambiguously good news for the theoretical study of indel processes. The potential benefits of modeling alignment as an aspect of statistical phylogenetics are significant. One can reasonably hope that the advance of theoretical work in this area will continue to inform advances in both bioinformatics and statistical phylogenetics. After all, and in spite of the Cambrian explosion in bioinformatics subdisciplines, sequence alignment and phylogeny truly are closely related aspects of mathematical biology.

Acknowledgements
The author thanks Kiyoshi Ezawa, Elena Rivas, Sean Eddy, Jeff Thorne, Benjamin Redelings, Marc Suchard, and one anonymous referee for productive conversations that have informed this review.

Funding
This work was supported by NIH/NHGRI grant HG004483.

Availability of data and materials
Not applicable.

Authors' contributions
IH wrote the article.

Competing interests
The author declares that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Received: 9 January 2017 Accepted: 30 April 2017
Published online: 12 May 2017

References

1. Jukes TH, Cantor C. Evolution of protein molecules. In: Mammalian Protein Metabolism. New York: Academic Press; 1969. p. 21–132.
2. Dayhoff MO, Eck RV, Park CM. A model of evolutionary change in proteins Atlas of Protein Sequence and Structure In: Dayhoff MO, editor. Washington, DC: National Biomedical Research Foundation; 1972. p. 89–99.
3. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17:368–76.
4. Yang Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 1993;10:1396–401.
5. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 1994;39:306–14.
6. Pedersen JS, Hein J. Gene finding with a hidden Markov model of genome structure and evolution. Bioinformatics. 2003;19(2):219–27.
7. Siepel A, Haussler D. Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol. 2004;11(2-3):413–28.
8. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol. 2006;2(4):33.
9. Pollard KS, Salama SR, Lambert N, Lambot M, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares M, Vanderhaeghen P, Haussler D. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 2006;443(7108):167–72.
10. Pedersen JS, Meyer IM, Forsberg R, Simmonds P, Hein J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic Acids Res. 2004;32(16):4925–3.
11. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.
12. Goldman N, Thorne JL, Jones DT. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J Mol Biol. 1996;263(2):196–208.
13. Liò P, Goldman N. Using protein structural information in evolutionary inference: transmembrane proteins. Mol Biol Evol. 1999;16:1696–710.
14. Blanchette M, Green ED, Miller W, Haussler D. Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res. 2004;14(12):2412–23. Comparative Study.
15. Ugalde JA, Chang BS, Matz MV. Evolution of coral pigments recreated. Science. 2004;305(5689):1433.
16. Liberles DA. Ancestral sequence reconstruction. Oxford biosciences. Oxford, UK: OUP; 2007. https://books.google.com/books?id=1_uPZWm1nSYC.
17. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. Crystal structure of an ancient protein: evolution by conformational epistasis. Science. 2007;317(5844):1544–8.
18. Gaucher EA, Govindarajan S, Ganesh OK. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature. 2008;451(7179):704–7.
19. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. Fast M L: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 2012;40(Web Server issue):580–4.
20. Alcolombri U, Elias M, Tawfik DS. Directed evolution of sulfotransferases and paraoxonases by ancestral libraries. J Mol Biol. 2011;411(4):837–53.
21. Santiago-Ortiz J, Ojala DS, Westesson O, Weinstein JR, Wong SY, Steinsapir A, Kumar S, Holmes I, Schaffer DV. AAV ancestral reconstruction library enables selection of broadly infectious viral variants. Gene Ther. 2015;22(12):934–46.
22. Zakas PM, Brown HC, Knight K, Meeks SL, Spencer HT, Gaucher EA, Doering CB. Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. Nat Biotechnol. 2016;35(1):35–37.
23. Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD, McTavish EJ, Midford PE, Owen CL, Ree RH, Rees JA, Soltis DE, Williams T, Cranston KA. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci U S A. 2015;112(41):12764–9.

24. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol. 2005;1(5):e45.
25. Pollock LJ, Rosauer DF, Thornhill AH, Kujala H, Crisp MD, Miller JT, McCarthy MA. Phylogenetic diversity meets conservation policy: small areas are key to preserving eucalypt lineages. Philos Trans R Soc Lond B Biol Sci. 2015;370(1662):20140007.
26. Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA, Berger A, Burguiere AM, Cinatl J, Eickmann M, Escriou N, Grywna K, Kramme S, Manuguerra JC, Muller S, Rickerts V, Sturmer M, Vieth S, Klenk HD, Osterhaus AD, Schmitz H, Doerr HW. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med. 2003;348(20):1967–76.
27. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 2005;22(5):1185–92.
28. Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR, Arinaminpathy N, Stramer SL, Busch MP, Delwart EL. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proc Natl Acad Sci U S A. 2012;109(37):15066–71.
29. Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, Koblin BA, Heneine W, Lemey P, Jaffe HW. 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature. 2016;539(7627):98–101.
30. Bielejec F, Baele G, Rodrigo AG, Suchard MA, Lemey P. Identifying predictors of time-inhomogeneous viral evolutionary processes. Virus Evol. 2016;2(2):023.
31. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20.
32. Hasegawa M, Kishino H, Yano T. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22:160–74.
33. Hohna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol. 2016;65(4):726–36.
34. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007;7:214.
35. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22(21):2688–90.
36. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21(5):676–9.
37. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
38. Felsenstein J. PHYLIP - phylogeny inference package (version 3.2). Cladistics. 1989;5:164–6.
39. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 2002;18(3):502–4.
40. Westesson O, Holmes I. Developing and applying heterogeneous phylogenetic models with XRate. PLoS ONE. 2012;7(6):36898.
41. Gu X, Li WH. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc Natl Acad Sci U S A. 1998;95(11):5899–905.
42. Lèbre S, Michel CJ. An evolution model for sequence length based on residue insertion-deletion independent of substitution: an application to the GC content in bacterial genomes. Bull Math Biol. 2012;74(8):1764–88.
43. Bahi JM, Michel CJ. A stochastic gene evolution model with time dependent mutations. Bull Math Biol. 2004;66(4):763–78.
44. Lunter GA, Hein J. A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics. 2004;20 Suppl 1:216–23.
45. Benard E, Michel CJ. A generalization of substitution evolution models of nucleotides to genetic motifs. J Theor Biol. 2011;288:73–83.
46. Benner SA, Cohen MA, Gonnet GH. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Biol. 1993;229(4):1065–82.
47. Chang MS, Benner SA. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J Mol Biol. 2004;341(2):617–31.
48. Hsing M, Cherkasov A. Indel PDB: a database of structural insertions and deletions derived from sequence alignments of closely related proteins. BMC Bioinformatics. 2008;9:293.
49. Williams LE, Wernegreen JJ. Sequence context of indel mutations and their effect on protein evolution in a bacterial endosymbiont. Genome Biol Evol. 2013;5(3):599–605.
50. McCrow JP. Alignment of phylogenetically unambiguous indels in Shewanella. J Comput Biol. 2009;16(11):1517–28.
51. Caspi A, Pachter L. Identification of transposable elements using multiple alignments of related genomes. Genome Res. 2006;16(2): 260–70.
52. Hein J, Wiuf C, Knudsen B, Moller MB, Wibling G. Statistical alignment: computational properties, homology testing and goodness-of-fit. J Mol Biol. 2000;302:265–79.
53. Ezawa K. General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable? BMC Bioinformatics. 2016;17:304.
54. Ezawa K. General continuous-time Markov model of sequence evolution via insertions/deletions: local alignment probability computation. BMC Bioinformatics. 2016;17(1):397.
55. Ezawa K. Erratum to: General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable?. BMC Bioinformatics. 2016;17(1):457.
56. Rivas E, Eddy SR. Parameterizing sequence alignment with an explicit evolutionary model. BMC Bioinformatics. 2015;16:406.
57. Sankoff D, Blanchette M. Multiple genome rearrangement and breakpoint phylogeny. J Comput Biol. 1998;5(3):555–70.
58. Arquès DG, Michel CJ. Analytical expression of the purine/pyrimidine codon probability after and before random mutations. Bull Math Biol. 1993;55(6):1025–38.
59. Arquès DG, Michel CJ. Analytical solutions of the dinucleotide probability after and before random mutations. J Theor Biol. 1995;175(4): 533–44.
60. Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics. 1999;15(6):446–54.
61. Michel CJ. Evolution probabilities and phylogenetic distance of dinucleotides. J Theor Biol. 2007;249(2):271–7.
62. Benard E, Lèbre S, Michel CJ. Genome evolution by transformation, expansion and contraction (GETEC). BioSystems. 2015;135:15–34.
63. Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol. 1991;33:114–24.
64. Feller W. An introduction to probability theory and its applications, Vol II. New York: John Wiley and Sons; 1971.
65. Holmes I, Bruno WJ. Evolutionary HMMs: a Bayesian approach to multiple alignment. Bioinformatics. 2001;17(9):803–20.
66. Holmes I. Using guide trees to construct multiple-sequence evolutionary HMMs. Bioinformatics. 2003;19 Suppl. 1:147–57.
67. Westesson O, Lunter G, Paten B, Holmes I. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. PLoS ONE. 2012;7(4):34572.
68. Bouchard-Côté A. A note on probabilistic models over strings: the linear algebra approach. Bull Math Biol. 2013;75(12):2529–50.
69. Metzler D. Statistical alignment based on fragment insertion and deletion models. Bioinformatics. 2003;19(4):490–9.
70. Bouchard-Côté A, Jordan MI. Evolutionary inference via the poisson indel process. Proc Natl Acad Sci U S A. 2013;110(4):1160–6.
71. Thorne JL, Kishino H, Felsenstein J. Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol. 1992;34:3–16.
72. McGuire G, Denham MC, Balding DJ. Models of sequence evolution for DNA sequences containing gaps. Mol Biol Evol. 2001;18(4):481–90.
73. Rivas E. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics. 2005;6:63.
74. Rivas E, Eddy SR. Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput Biol. 2008;4:1000172.
75. Holmes I. A probabilistic model for the evolution of RNA structure. BMC Bioinformatics. 2004;5:166.
76. Matsui H, Sato K, Sakakibara Y. Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures. Bioinformatics. 2005;21:2611–7.

77. Hickey G, Blanchette M. A probabilistic model for sequence alignment with context-sensitive indels. Lect Notes Comput Sci. 2011;6577/2011: 85–103. doi:http://dx.doi.org/10.1007/978-3-642-20036-6_10.
78. Knudsen B, Miyamoto M. Sequence alignments and pair hidden Markov models using evolutionary history. J Mol Biol. 2003;333(2):453–60.
79. Redelings BD, Suchard MA. Joint Bayesian estimation of alignment and phylogeny. Syst Biol. 2005;54(3):401–18.
80. Suchard MA, Redelings BD. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics. 2006;22(16):2047–8.
81. Redelings BD, Suchard MA. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. BMC Evol Biol. 2007;7:40.
82. Westesson O, Lunter G, Paten B, Holmes I. Phylogenetic automata, pruning, and multiple alignment. 2011. arXiv:1103.4347.
83. Westesson O, Barquist L, Holmes I. Hand Align: Bayesian multiple sequence alignment, phylogeny, and ancestral reconstruction. Bioinformatics. 2012;28(8):1170–71.
84. Miklós I, Lunter G, Holmes I. A long indel model for evolutionary sequence alignment. Mol Biol Evol. 2004;21(3):529–40.
85. Miklós I, Toroczkai Z. An improved model for statistical alignment. In: First Workshop on Algorithms in Bioinformatics. Berlin, Heidelberg: Springer; 2001.
86. Hobolth A. A Markov Chain Monte Carlo Expectation Maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. J Comput Graph Stat. 2008;17(1):138–62.
87. Bérard J, Guéguen L. Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context. Syst Biol. 2012;61(3):510. doi:http://dx.doi.org/10.1093/sysbio/sys024.
88. Fan Y, Wang W, Ma G, Liang L, Shi Q, Tao S. Patterns of insertion and deletion in mammalian genomes. Curr Genomics. 2007;8(6):370–8.
89. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science. 1992;256(5062):1443–5.
90. Yamane K, Yano K, Kawahara T. Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. DNA Res. 2006;13(5):197–204.
91. Zhang Z, Gerstein M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. 2003;31(18):5338–48.
92. Gu X, Li WH. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J Mol Evol. 1995;40(4):464–73.
93. Cartwright RA. Problems and solutions for estimating indel rates and length distributions. Mol Biol Evol. 2009;26(2):473.
94. 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(2):298–309.
95. Miller W, Myers EW. Sequence comparison with concave weighting functions. 1988;50:97–120.
96. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 2005;15(2):330–40. Comparative Study.
97. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268(1):78–94.
98. Kim J, Sinha S. Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. Bioinformatics. 2007;23(3):289–97.
99. Cartwright RA. DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics. 2005;21 Suppl 3:31–8.
100. Fletcher W, Yang Z. INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol. 2009;26(8):1879–88.
101. Strope CL, Abel K, Scott SD, Moriyama EN. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. Mol Biol Evol. 2009;26(11):2581–93.
102. Novak A, Miklós I, Lyngsoe R, Hein J. StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics. 2008;24(20):2403–4.
103. Bouchard-Côté A, Klein D, Jordan MI. Advances in Neural Information Processing Systems 21 In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors. Vancouver, British Columbia, Canada: Curran Associates, Inc.; 2009. p. 177–84. http://papers.nips.cc/paper/3406-efficient-inference-in-phylogenetic-indel-trees.pdf.
104. Redelings B. Erasing errors due to alignment ambiguity when estimating positive selection. Mol Biol Evol. 2014;31(8):1979–93.
105. Arunapuram P, Edvardsson I, Golden M, Anderson JW, Novak A, Sukosd Z, Hein J. StatAlign 2.0: combining statistical alignment with RNA secondary structure prediction. Bioinformatics. 2013;29(5):654–5.
106. Meyer IM, Miklós I. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. PLoS Comput Biol. 2007;3(8):149.
107. Ezawa K. Characterization of multiple sequence alignment errors using complete-likelihood score and position-shift map. BMC Bioinformatics. 2016;17(1):133. doi:http://dx.doi.org/10.1186/s12859-016-0945-510.1186/s12859-016-0945-5.
108. Hein J. Pacific Symposium on Biocomputing In: Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE, editors. Singapore: World Scientific; 2001. p. 179–90.
109. Lunter GA, Miklós I, Song YS, Hein J. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. J Comput Biol. 2003;10(6):869–89.
110. Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition. Comput Speech Lang. 2002;16(1):69–88.
111. Searls DB, Murphy KP. Automata-theoretic models of mutation and alignment. Proc Int Conf Intell Syst Mol Biol. 1995;3:341–9.
112. Holmes IH. Historian: accurate reconstruction of ancestral sequences and evolutionary rates. Bioinformatics. 2017;33(8):1227–29.
113. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput Biol. 2014;10(4):1003535.
114. Eskin E, Noble WS, Singer Y. Protein family classification using sparse Markov transducers. J Comput Biol. 2003;10(2):187–213.
115. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14(5):988–95.
116. Holmes I. Modular non-repeating codes for DNA storage. 2016. arXiv:1606.01799.
117. Mealy GH. A method for synthesizing sequential circuits. Bell Syst Technical J. 1955;34:1045–79.
118. Sankoff D. Simultaneous solution of the RNA folding, alignment, and protosequence problems. SIAM J Appl Math. 1985;45:810–25.
119. Bradley RK, Holmes I. Evolutionary triplet models of structured RNA. PLoS Comput Biol. 2009;5(8):1000483.
120. Wang J, Keightley PD, Johnson T. MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. BMC Bioinformatics. 2006;7:292.
121. Felsenstein J, Churchill GA. A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol. 1996;13:93–104.
122. Siepel A, Haussler D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol. 2004;21(3): 468–88.
123. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 2003;31(13): 3423–428.
124. Klosterman PS, Uzilov AV, Bendana YR, Bradley RK, Chao S, Kosiol C, Goldman N, Holmes I. XRate: a fast prototyping, training and annotation tool for phylo-grammars. BMC Bioinformatics. 2006;7:428.