Stochastic Insertion-Deletion Processes and Statistical Sequence Alignment

D. Metzler[1], R. Fleißner[2], A. Wakolbinger[3], and A. von Haeseler[4,5]

[1] FB Biologie und Informatik, Goethe-Universität, 60054 Frankfurt am Main, Germany, metzler@informatik.uni-frankfurt.de
[2] Department of Mathematics, University of Idaho, P.O. Box 441103. Moscow, ID 83844-1103, USA, fleissne@uidaho.edu
[3] FB Mathematik, Goethe-Universität, 60054 Frankfurt am Main, Germany, wakolbin@math.uni-frankfurt.de
[4] Bioinformatik, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany, haeseler@cs.uni-duesseldorf.de
[5] Forschungszentrum Jülich, John von Neumann-Institut für Computing (NIC), Forschungsgruppe Bioinformatik, 52425 Jülich, Germany

Summary. The reconstruction of the history of a set of sequences is a central problem in molecular evolutionary biology. Typically this history is summarized in a phylogenetic tree. In current practice the estimation of a phylogenetic tree is a two-step procedure: first a multiple alignment is computed and subsequently a phylogenetic tree is reconstructed, based on the alignment. However, it is well known that the alignment and the tree reconstruction problem are intertwined. Thus, it is of great interest to estimate alignment and tree simultaneously. We present here a stochastic framework for this joint estimation. We discuss a variant of the Thorne-Kishino-Felsenstein model, having equal rates of insertions and of deletions of sequence fragments, for $\ell \geq 2$ sequences related by a phylogenetic tree. Finally, we review novel approaches to tree reconstruction based on insertion-deletion models.

1 Introduction

Sequence Evolution and Alignments

Biological (DNA or amino acid) sequences change over time due to the random process of evolution. Two ingredients play a major role:

(i) a process of *substitutions*, which in biological parlance replace a nucleotide or an amino acid by another one, or, more formally, change the *labels* of the *positions* in the sequence,

(ii) a process of *insertions* and *deletions* (briefly called *indel process*) of single positions or (more generally) sequence fragments.

The evolutionary process leading to observed sequences is modelled by a stochastic process indexed by a binary tree $\mathcal{T}$. The states of the process are the *labelled*

sequences $(a_1, ..., a_n)$, $n \in \mathbb{N}_0, a_i \in \mathcal{A}$, where $\mathcal{A}$ is a finite *alphabet*. (In case of DNA, $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.) We will refer to a_i as the label of position no. i in the sequence. The sequences are thought to evolve from an ancestral sequence placed in an inner node which figures as the *root* of $\mathcal{T}$. What is observed are the sequences appearing at the *leaves* of $\mathcal{T}$. The edge lengths of $\mathcal{T}$ are measured in units of "evolutionary distance" (or *time* for simplicity). $\mathcal{T}$ is called *phylogenetic tree*, and the graph structure (or "tree topology") of $\mathcal{T}$ is called the *phylogeny* of the observed sequences. Positions in the observed sequences which are connected by a path of the evolution process without insertion or deletion are called *homologous*. An *alignment* of the observed sequences is an array each of whose lines consist of one of the sequences, possibly augmented by "gaps", such that positions assumed to be homologous appear in one and the same column.

It is a widespread practice in biology to construct in a first step an "optimal" alignment of the observed sequences through some scoring rule which penalizes mismatches (i.e. aligned positions carrying different labels) and gaps ([5]), and in a second step to infer, on the basis of the aligned positions without gaps, the underlying phylogenetic tree ([17]). With the advent of powerful computer technology it became more popular to incorporate some model for the substitution process, and to estimate a maximum likelihood tree that relates the observed sequences. To model the substitution process one usually considers a Markovian jump dynamics on $\mathcal{A}$ which acts independently from position to position ([19]).

However, the reliability of an alignment which optimizes some score is hard to judge. Also, basing the analysis on a single alignment may result in underestimating the variability of the parameter estimates of the substitution process and of the reconstructed phylogeny. Finally, ignoring the gaps and their clues about insertion and deletion events may lead to a considerable loss of valuable information ([15]). All these problems call for an explicit stochastic modelling of the insertion-deletion process.

Outline of the Paper

In section 2 we will review various models of insertion-deletion processes and discuss a number of their basic properties. We will start from the nowadays classical models of Thorne, Kishino and Felsenstein for the insertion and deletion of single positions ("TKF", [21]) and of indivisible fragments ("TKF2", [22]). As will become clear, both models can be extended to the case of equal rates of insertion and deletion. We will show that although these dynamics have no equilibrium distribution on the sequences of finite length, reversibility is guaranteed, which appeases a caveat in the recent monograph [2] of J. Felsenstein. These models are building blocks for describing the evolution of sequences along a phylogenetic tree, and for obtaining a "multiple statistical alignment" of ℓ observed sequences.

In section 3 we turn to tree-indexed insertion-deletion processes, which form the mathematical basis for multiple statistical alignment. Indeed, a remarkable progress in multiple statistical alignment (based on the TKF model) was recently made by the group around J. Hein in Oxford. The basic idea is to think of the random multiple alignment as being generated by $\mathcal{T}$-indexed "events" which are brought into a well-defined order. Stimulated by recent communication with G. Lunter and I. Miklós in Oxford, we describe how this can be understood through a reading of the (now $\mathcal{T}$-indexed) indel genealogy "from left to right" and we show how to extend this

approach from the TKF model to the fragment insertion-deletion model defined in [11].

Finally, in section 4 we will review some novel approaches to tree reconstruction based on insertion-deletion models.

2 Sequence Evolution Models with Insertion and Deletion

2.1 Stochastic Indel Dynamics

In the prototype insertion-deletion ("indel") model suggested by Thorne, Kishino and Felsenstein [21] in 1991 (see subsection 2.2), single positons are inserted with rate λ and deleted with rate μ, independently of each other. The assumption $\lambda < \mu$ is a tribute to the existence of a reversible equilibrium distribution. We will refer to this as the TKF model for short. It is assumed that the rate of insertions and deletions is not influenced by the labels of the positions, which makes it possible to construct the substitution process "on top" of the indel process.

Consider first the case of $\ell = 2$ observed sequences. It turns out that one can read a nice "genealogical" structure into the TKF dynamics, which enriches the insertion-deletion path of the evolutionary process and renders a forest of Galton-Watson trees with immigration, see subsection 2.3. Reading the forest from left to right allows to generate the alignment by a Markov chain with three states $\binom{p}{p}, \binom{-}{p}, \binom{p}{-}$, where $\binom{p}{p}$ stands for a homologous pair of positions, and $\binom{-}{p}$ and $\binom{p}{-}$ denote positions appearing in only one of the two observed sequences. This chain is "hidden" in the sense that its path cannot be observed from the data. What *is* observed is the result of all the states' *emissions*. E.g. for the sequence of states $\binom{p}{p}, \binom{p}{-}, \binom{p}{p}, \binom{-}{p}, \binom{-}{p}$ the result of emissions could be $\binom{AGT}{ACGG}$. This hidden Markov structure allows to apply dynamic programming, which is a powerful tool not only for computing likelihoods but also for the sampling of alignments from the conditional distribution, given the observed sequences. Alternating the sampling of alignments with a Metropolis-Hastings algorithm for sampling the evolution parameters gives a Markov chain Monte Carlo method which makes it possible to assess the variability of the joint estimation of parameters and alignments [12].

In [11], a variant of the TKF model was introduced which assumes *equal* insertion and deletion rates $\lambda = \mu$. In the present paper, this will be referred to as the cTKF model (c for critical). The cTKF model is a special case of the fragment insertion deletion ("FID") model ([11]), which in turn is a "critical" variant of an indel model introduced by Thorne et al. in [22]. This "TKF2" model assumes insertion (at rate λ) and deletion (at rate $\mu > \lambda$) of (indivisible) *fragments* whose lengths are independent, geometrically distributed with expectation $(1-\rho)^{-1}$. In his recent monograph [2], J. Felsenstein quotes the FID model (with reference to [11]) with the caveat: " It is not clear that equality [of λ and μ] is tenable, as the resulting model of sequence-length variation then has no equilibrium distribution, and reversibility of the process is not guaranteed." However, as we will see in Proposition 1, these doubts can be remedied. In fact, the FID process which takes a sequence of length n_1 into a sequence of length n_2 can be thought of as "cut out" from an indel process on *infinite sequences*, with the condition that the development between the two sequences is flanked by an

(invisible) homologous pair to the left, and another one to the right of the observed sequences.

The assumption made in FID (and TKF2) that inserted pieces are unbreakable entities is unrealistic. Abandoning this assumption, one arrives at a more general insertion-deletion model (introduced in [11] as "GID"), which however is computationally much less tractable. In [11], the FID and GID models were compared with regard to robustness of estimates. Computer simulations showed that estimation procedures for the parameters which are based on the FID assumptions also work well when applied to data generated without the fragmentation restrictions.

The GID model, in turn, is related to the class of *long indel models* recently studied by Miklós et al [13] (see also [14]).

2.2 TKF Bridges

In the pioneering paper [21], Thorne, Kishino and Felsenstein introduced the following evolution model for finite sequences (where a *sequence* consists of *positions*, arranged in linear order).

Definition 1. *(TKF(λ, μ) dynamics for finite sequences)*
Each position is deleted at rate μ and between any two neighbouring positions (as well as to the left and to the right of the current sequence), a new position is inserted at rate λ.

The *insertion-deletion path* records the ordering of all the positions alive at any time s. In particular, the indel path keeps track of the times of insertions and deletions. For fixed parameters $\lambda, \mu > 0$ we will write $\mathrm{TKF}_{n,[0,t]}$ for the distribution of an indel path starting from a sequence of length n and evolving over time t (or "evolutionary distance") according to the TKF(λ, μ)-dynamics.

Remark 1. For $\mu > \lambda$, the TKF(λ, μ) dynamics has a reversible equilibrium, the random sequence length L in equilibrium being geometrically distributed with weights $\gamma^n(1-\gamma)$, $n \in \mathbb{N}_0$, where $\gamma = \lambda/\mu$. Consequently, L has expectation $\lambda/(\mu - \lambda)$.

Thorne et al. [21] consider only the case $\mu > \lambda$. In this case let us write

$$\mathrm{TKF}_{[0,t]} = \sum_{n \geq 0} \gamma^n (1-\gamma) \mathrm{TKF}_{n,[0,t]}$$

for the equilibrium distribution of an indel path evolving in the time interval $[0, t]$ under the TKF(λ, μ)-dynamics. The following definition makes sense also for general $\lambda, \mu > 0$.

Definition 2. *Let $\lambda, \mu, t > 0$. For given natural numbers n_1, n_2, we define the* TKF bridge $\mathrm{TKF}_{n_1,n_2;[0,t]}$ *as the distribution of the TKF(λ, μ) process, conditioned to take a sequence of length n_1 into a sequence of length n_2 over time t.*

The TKF bridge for $\lambda = \mu$ is a special case of the *fragment insertion deletion model* introduced in [11], see subsection 2.5 below. Although for $\lambda \geq \mu$ the TKF process has no equilibrium on the finite sequences, the distribution of the TKF bridge does not depend on the chosen direction of time, as the corollary to the following proposition tells.

Proposition 1. *For $\lambda, \mu > 0$ and $\gamma = \lambda/\mu$, the measure*

$$M := \sum_{n \geq 0} \gamma^n \, TKF_{n,[0,t]}$$

is invariant under time reversal (where the insertions and deletions in an indel path interchange roles).

Proof. Since, given the birth and death times, the birth and death positions within the current sequence are exchangeable, it suffices to consider the process $L = (L_s)_{0 \leq s \leq t}$ of sequence lengths. Under the $\mathrm{TKF}(\lambda, \mu)$ dynamics, this is a birth and death process on $\mathbb{N}_0$ with birth rates $b(n) := (n+1)\lambda$, $n \geq 0$, and death rates $d(n) := n\mu$, $n \geq 1$, hence

$$b(n) = \gamma d(n+1), \quad n \geq 0. \tag{1}$$

Write

$$P_s(m,n) := \mathbb{P}[L_s = n | L_0 = m]$$

for the transition probability of L, and define the measure g on $\mathbb{N}_0$ by

$$g(n) = \gamma^n, n \geq 0 \tag{2}$$

The detailed balance condition (1) guarantees reversiblity of the measure g:

$$g(m)P_s(m,n) = g(n)P_s(n,m), \; m,n \in \mathbb{N}_0, s \geq 0. \tag{3}$$

This extends to reversibility of the measure

$$M(L \in (.)) = \sum_{m \geq 0} g(m)\mathbb{P}[L \in (.) | L_0 = m] :$$

$$M(L_0 = m, L \in B, L_t = n) = M(L_0 = n, L \in R(B), L_t = m). \tag{4}$$

Here, B is a measurable set of $\mathbb{N}_0$-valued (right continuous) paths with jump size 1, and R is the time reversal operator mapping a path $(x_s)_{0 \leq s \leq t}$ into the right continuous modification of its time reversal $(x_{t-s})_{0 \leq s \leq t}$.

Corollary 1. *For all $\lambda, \mu > 0$ the bridge $TKF_{n_1,n_2;[0,t]}$ with parameters λ, μ equals the time reversal of the bridge $TKF_{n_2,n_1;[0,t]}$ with parameters λ, μ.*

Proof. For any (measurable) set A of indel paths,

$$\gamma^{n_1}\mathbb{P}[L_t = n_2 | L_0 = n_1]\mathrm{TKF}_{n_1,n_2;[0,t]}(A) = M(A; L_0 = n_1, L_t = n_2),$$

where M is the measure defined in Proposition 1. Hence the claim follows from the reversibility of M.

Let us in particular single out the "critical" case $\lambda = \mu$. In this case we will speak of the $\mathrm{cTKF}(\lambda)$ (instead of the $\mathrm{TKF}(\lambda, \lambda)$) dynamics, and denote the measure M in Proposition 1 by

$$\mathrm{cTKF}_{[0,t]} := \sum_{n \geq 0} \mathrm{TKF}_{n;[0,t]}.$$

Proposition 1 then specializes to

Corollary 2. *For each $\lambda > 0$, the (infinite) measure $cTKF_{[0,t]}$ with parameter λ is invariant under time reversal.*

2.3 A Genealogy of Positions

Following Thorne et al [21], we upgrade an indel path with a genealogy of positions by decreeing that each inserted position *is born by its left neigbour*. A position which inserted to the very left of the sequence is declared to be born by an invisible, immortal position to its left. Equivalently, one can interprete an insertion of this type as an *immigration* at the left end of the current sequence. Thus, the TKF(λ, μ) process turns into a (binary, continuous time) Galton-Watson process with birth rate and immigration rate equal to λ, and death rate equal to μ.

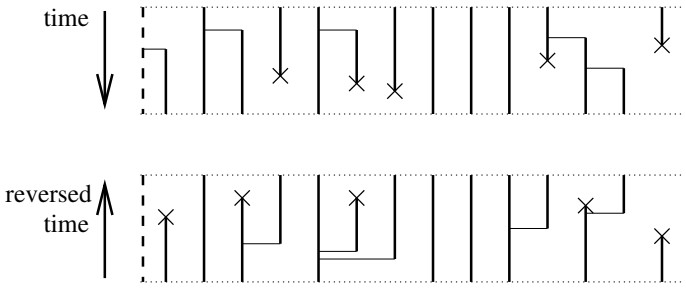

Fig. 1. Galton-Watson forest for a TKF-process. Applying the TKF-convention to the time-reversed realization changes the alignment (as seen at the right end) but not the homology structure. The immigrant (leftmost line in the forward-time figure) is considered to be born by an immortal position to its left (dashed line).

This can be illustrated with a graphical construction (see Figure 1, upper part): Let time run downwards. If a newly inserted position has a left neigbour in the sequence (that is, does not appear at the left end of the sequence), then draw a horizontal line at the time of insertion between the newly inserted position and its left neighbour. If a new insertion happens at the left end of the sequence, then draw a horizonal line to the invisible, immortal position thought to sit left of the sequence.

In this way the $\mathrm{TKF}_{n_1,n_2;[0,t]}$ distribution gives rise to a Galton-Watson forest with immigration, starting with n_1 mother positions at time 0, and conditioned to a total number of n_2 positions living at t. The picture can be reversed: at each deletion, draw a horizontal line to the current left neighbour, or to the invisible, immortal position if the deletion happens at the left end. Corollary 1 then turns into a statement about conditional Galton-Watson forests: a forest generated by $\mathrm{TKF}_{n_1,n_2;[0,t]}$ (with parameters $\lambda, \mu > 0$), when reversed in the described manner, equals in distribution a forest generated by $\mathrm{TKF}_{n_2,n_1;[0,t]}$, with the same parameters λ, μ (see Figure 1).

2.4 Reading an indel forest from left to right

Let us read the branches of an indel forest over the time interval $[0, t]$ from left to right, and write $\binom{p}{p}$ for a position which is conserved over the time interval $[0, t]$, $\binom{p}{-}$ for a position present at time 0 and deleted before time t, and $\binom{-}{p}$ for a position inserted after time 0 and surviving till time t. The indel forest generated by $\mathrm{cTKF}_{[0,t]}$

(or by $\mathrm{TKF}_{[0,t]}$ if $\lambda < \mu$) leads to a Markov chain with state space $\binom{p}{p}, \binom{p}{-}, \binom{-}{p}$. The transition probabilities of this chain can be phrased (and easily computed, see [11]) in terms of a binary Galton-Watson process.

For further use, let us think of time 0 as corresponding to the *root*, and time t to the *daughter node* of a "tree" consisting of a single edge.

Following a clever convention introduced in [9] we write a B for every position in the root, and the following symbols in the daughter node: H for $\binom{p}{p}$, N for $\binom{p-}{-p}$, E for $\binom{p}{-}$ not followed by $\binom{-}{p}$, and B for $\binom{-}{p}$ not preceded by $\binom{p}{-}$. Whereas H means that the position survives, N means that the position dies but leaves a (leftmost) child. In both cases H and N we will address the successor position as the *heir* of the ancestor position. The symbol B in the daughter node stands for either an immigrant position or a descendant position which is not an heir (these will be called *β-children* for short). For example, the *indel history* $\binom{--p-p\,p--}{pp\,pp\,--pp}$ translates into $F = \binom{B\;\;BB}{BBHBENB}$, or $(BBHBENB)$ for short (see Figure 2). Every F starts

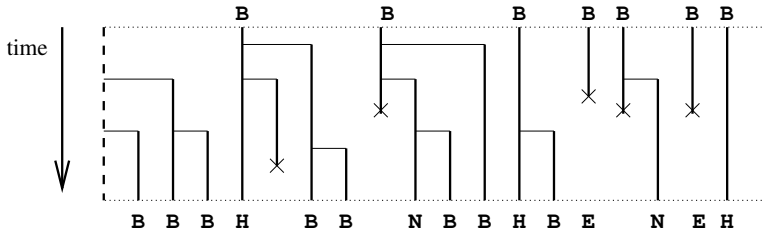

Fig. 2. $\{B, H, N, E\}$-notation of an alignment for a pair of sequences.

with an initial block $F^{(0)}$ (of immigrants) of the form $\binom{}{B\ldots B}$, followed by n blocks $F^{(1)}, ..., F^{(n)}$, where n is the length of the sequence in the root. Each of these blocks is of type H, N or E, that is, of the form $\binom{B}{HB\ldots B}$, $\binom{B}{NB\ldots B}$ or $\binom{B}{E}$. A block of the latter type means that the ancestral position is extinct. In the type H and type N blocks, the number k of B's in the second line is geometrically distributed with probability weight $(1-\pi_t)\pi_t^k$, where the parameter π_t depends also on λ and μ. In other words, right to any H, N or B in the second line, the current block is continued (with a B in the second line) with probability π_t; otherwise (as long as the current number of blocks is $\leq n$, a new block starts. Independently of what is to its left, a new block is of type H, N or E with probability π_t^H, π_t^N and π_t^E, where these three probabilites depend on λ and μ, and sum to 1. Under the infinite measure $\mathrm{cTKF}_{[0,t]}$ every n is assigned mass 1, whereas under the probability measure $\mathrm{TKF}_{[0,t]}$, n gets mass $\gamma^n(1-\gamma)$ (where $\gamma = \lambda/\mu$). The latter can of course also be realized by an independent stopping, which, after each n, continues with a new block with probability γ and jumps to an `End` state with probability $1-\gamma$.

Note that an alignment under the cTKF dynamics can be produced by an $\{B, H, N, E\}$-valued Markov chain with transition probabilities

$$\pi_t^{HB} = \pi_t^{NB} = \pi_t^{BB} = \pi_t; \quad \pi_t^{HH} = \pi_t^{NH} = \pi_t^{BH} = (1-\pi_t)\pi_t^H$$

$$\pi_t^{HN} = \pi_t^{NN} = \pi_t^{BN} = (1-\pi_t)\pi_t^N; \quad \pi_t^{HE} = \pi_t^{NE} = \pi_t^{BE} = (1-\pi_t)\pi_t^E;$$

$$\pi_t^{EH} = \pi_t^H, \quad \pi_t^{EN} = \pi_t^N, \quad \pi_t^{EE} = \pi_t^E, \quad \pi_t^{EB} = 0.$$

This chain starts in the initial distribution giving weight π_t to B and weights $(1-\pi_t)\pi_t^H, (1-\pi_t)\pi_t^N, (1-\pi_t)\pi_t^E$ to H, N and E, respectively.

2.5 A Fragment Insertion-Deletion Model

Let us consider the analogue of the cTKF dynamics (with parameter λ) acting on indivisible *fragments* whose lengths are independent and geometrically distributed with expectation $(1-\rho)^{-1}$, $\rho \in [0,1)$. The resulting indel dynamics on the finite sequences will be called *fragment insertion deletion dynamics* with parameters λ and ρ , abbreviated as FID(λ, ρ). The indel histories will be coded in a similar way as described in the previous subsection. For example, imagine a fragment of length 4 which survives and leaves one daughter fragment with length 3, say. Then the "second line" of the indel history reads as $(HHHHBBB)$. In case the mother fragment dies and leaves one daughter fragment with length 3, we would have $(EEENBB)$. Thanks to the properties of the geometric distribution, under the FID measure the alignment builds up as a Markov chain with transition probabilites π_t^{IJ}, $I, J \in \{H, N, E, B\}$, which are only slightly more complicated than those in the cTKF case, see [11] for details. Notably, for $\rho > 0$ one has $\pi_t^{NN} = \pi_t^{HN} = \pi_t^{BN}$ but $\pi_t^{HH} > \pi_t^{BH} = \pi_t^{NH}$, and $\pi_t^{BB} = \pi_t^{NB} > \pi_t^{HB}$, because compared to the cTKF model there is some additional probability for a position to be in the same fragment as its left neighbour.

3 Tree Indexed Indel Processes

3.1 Multiple TKF Bridges

Let $\mathbb{T}$ be a finite binary tree with ℓ leaves. Its sets of edges, nodes and leaves will be denoted by $\mathcal{E}$, $\mathcal{N}$, and $\mathcal{L}$. The edges ϵ are labelled by positive numbers (*lengths*) t_ϵ. The *branch* b_ϵ is represented as $\{\epsilon\} \times (0, t_\epsilon)$, and the *(labelled) tree* $\mathcal{T}$ is represented as

$$\mathcal{T} = \mathcal{N} \cup \bigcup_{\epsilon \in \mathcal{E}} b_\epsilon,$$

equipped with the obvious tree distance along the branches.

For the moment, one of the nodes of $\mathbb{T}$ is distinguished as the *root* of $\mathbb{T}$ and denoted by r. This choice assigns to each edge ϵ a direction (from the root to the leaves), and to each node $\nu \neq r$ its *mother node* and its *mother edge*. We will write $t(\nu)$ for the length of the mother edge of ν, and $\mathcal{T}_r$ for the pair $(\mathcal{T}, r)$.

Fix $0 < \lambda \leq \mu$, and consider the TKF(λ, μ) indel dynamics on the finite sequences.

Definition 3. *The* indel process indexed by $\mathcal{T}_r$ *starts with an ancestral sequence at the root and lets a copy of this sequence evolve independently according to the TKF(λ, μ) dynamics along each branch leading away from the root. In every inner node (different from the root), the process continues with two identical copies which evolve independently along the two branches descending from this node.*

The $\mathcal{T}_r$-indexed indel process induces a distribution $\mathrm{TKF}_{n;\mathcal{T},r}$ on the $\mathcal{T}$-indexed indel paths starting with length n in the root r. As in section 2 we put

$$\mathrm{cTKF}_{\mathcal{T}} = \sum_{n\geq 0} \mathrm{TKF}_{n;\mathcal{T},r} \quad \text{if} \quad \lambda = \mu,$$

$$\mathrm{TKF}_{\mathcal{T}} = \sum_{n\geq 0} \gamma^n(1-\gamma)\mathrm{TKF}_{n;\mathcal{T},r} \quad \text{if} \quad \lambda < \mu.$$

Note that due to Proposition 1 these measures indeed do not depend on the choice of the root. The same is true for the *multiple TKF bridge* $\mathrm{TKF}_{n_1,\ldots,n_\ell;\mathcal{T}}$ which arises by conditioning the tree-indexed indel process to produce given sequence lengths $n_1, ..., n_\ell$ in the leaves.

3.2 Decomposing a Tree Indexed Indel Path Into Heirs Lines

Let us consider a $\mathcal{T}_r$-indexed indel path w, and the $\mathbb{T}_r$-indexed indel history $\tilde{w}$ induced by w along the nodes of $\mathbb{T}$. As described in subsection 2.4, each position p at a node $\nu \in \mathcal{N}, \nu \neq r$, is either an immigrant or an heir or a β-child, depending on its genealogy along the mother edge of ν. Immigrant positions and positions at the root are called *founder positions*. Founder positions are marked with B, whereas heirs are marked by H or N, depending whether they are survivors or not (see subsection 2.4).

For a fixed node ν, $\mathbb{T}(\nu)$ denotes the subtree of $\mathbb{T}$ which is rooted in ν. Let p be a position at node ν which is marked by B. Following the heirs (and heirs of heirs...) descending from p along the nodes of $\mathbb{T}(\nu)$ we obtain a mapping from the nodes of $\mathbb{T}(\nu)$ which assigns a B to ν and an H, N, E or "-" to the other nodes of $\mathbb{T}(\nu)$ (where the convention is such that any node descending from a node which carries an E is assigned a "-"). We call this mapping the *tree indexed heirs line* ("`tihl`") e initiated by p, and say that the node ν is the *origin* r_e of the `tihl` e.

Let $\mathbb{T}_e$ be the subtree of $\mathbb{T}(\nu)$ consisting of all the nodes to which e assigns a B, H or N. Denote by $\mathrm{supp}(e)$ the *support* of e, i.e. the set of all those nodes which are connected with p by an heirs line, or in other words, the set of all those nodes of $\mathbb{T}(\nu)$ to which e assigns H or N. The set $\mathrm{supp}(e) \cup \{r_e\}$ will be called the *rooted support* of e, and by $\overline{\mathrm{supp}}(e)$ we will denote the *extended support* of e, i.e. the set of all those nodes of $\mathbb{T}(\nu)$ to which e assigns H, N or E.

A `tihl` initiated by a founder position will be called a *founder* `tihl`. If p is a β-child, then it has a left neighbour p', and we call the `tihl` e' to which p' belongs the *mother* of the `tihl` e initiated by p, and e a *daughter* of e'. Every `tihl` in the tree-indexed indel history $\tilde{w}$ has an ancestral line tracing back to a founder `tihl`, and in this sense we obtain a "family decomposition" of all the `tihl`s in $\tilde{w}$.

The idea is now to build up the tree-indexed indel history $\tilde{w}$ successively through the `tihl`s it consists of. To this end we introduce an order on the nodes of $\mathbb{T}$ as follows. First we put $\nu_2 \preceq \nu_1$ if the node ν_1 is on the path from ν_2 to the root. Following [9] we fix a total order $\leq$ on the nodes of the tree which extends the partial order $\preceq$. We say that ν_1 *has a smaller rank than* ν_2 if $\nu_1 < \nu_2$ in the total order, see Figure 3. The first `tihl` to be filled in is the one initiated by the leftmost founder at the node with the smallest possible rank. Now proceed inductively: as long as there remain `tihl`'s which are daughters of the ones already filled in, the next `tihl` to be filled in is the one among all these which is initiated by the leftmost β-child at the node with the smallest possible rank. After completion of a family of `tihl`'s, proceed with the `tihl` initiated by the leftmost founder at the node with the next smallest possible rank.

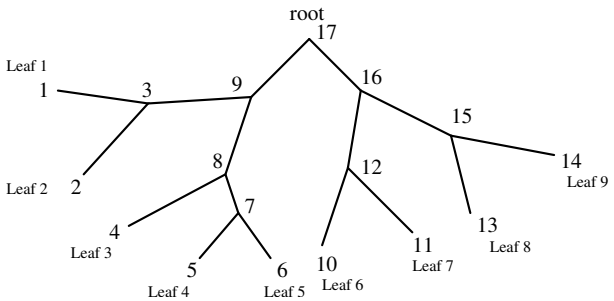

Fig. 3. Example for a total order "$\leq$" on the nodes of a tree: For each node ν the nodes in the left subtree stemming from ν have lower ranks than those in the right subtree .

3.3 Building an Indel History by a Markov Chain of Tree Indexed Heirs Lines and Sets of Active Nodes

Consider the measure $\mathrm{cTKF}_{\mathcal{T}}$ (with parameter λ) or the measure $\mathrm{TKF}_{\mathcal{T}}$ (with parameters $\lambda < \mu$) on the $\mathcal{T}_r$-indexed indel paths w. By the mapping described in subsection 3.2, this measure is transported into a measure on the finite sequences of `tihl`s. Any such sequence of `tihl`s starts with a block of `tihl` families founded by immigrants, and continues with n `tihl` families founded by positions at the root. Again (cf. subsection 2.4), n gets mass $(\lambda/\mu)^n(1-\lambda/\mu)$ under the probability measure $\mathrm{TKF}_{\mathcal{T}}$, whereas under the infinite measure $\mathrm{cTKF}_{\mathcal{T}}$ every n is assigned mass 1. This can be expressed in a unified way as follows: Independently after the completion of each `tihl` family, a new `tihl` family starts with probability $\gamma = \lambda/\mu$.

The process of `tihl`s is not Markov. However, one may keep track in parallel of the *set of active nodes* ("`soan`s"), and in fact the process of pairs of `soan`s and `tihl`s *is* Markov. We will call a node ν *active* if it is a candidate for the origin of the next `tihl` to be inserted. Initially, all nodes of $\mathbb{T}$ are active. Each newly inserted `tihl` re-activates all nodes in its support, and de-activates all those nodes which have a smaller rank than its origin and do not belong to its support. Formally, for a set $\mathcal{S} \subseteq \mathcal{N}$, and a `tihl` e with $r_e \in \mathcal{S}$, we put

$$[\mathcal{S}, e] := (\mathcal{S} \setminus \{\sigma \in \mathcal{N} | \sigma < r_e\}) \cup \operatorname{supp} e. \tag{5}$$

For a `tihl` e, we write

$$p(e) = \pi_{t(r_e)} \prod_{\sigma \in \overline{\mathrm{supp}}(e)} \pi_{t(\sigma)}^{e(\sigma)},$$

where we put

$$\pi_{t(r)} := \gamma.$$

Given the current set of active nodes is $\mathcal{R}$, the probability that the next step leads to the `tihl` e is

$$P(\mathcal{R}, e) = p(e) \prod_{\sigma \in \mathcal{R}: \sigma < r_e} (1 - \pi_{t(\sigma)}), \tag{6}$$

whereas the probability that the process is stopped (i.e. jumps to an extra state `End`) is $P(\mathcal{R}, \mathtt{End}) = \prod_{\sigma \in \mathcal{R}} (1 - \pi_{t(\sigma)})$. (Since the root r in any case belongs to $\mathcal{R}$, this latter transition probability is zero in the cTKF model.)

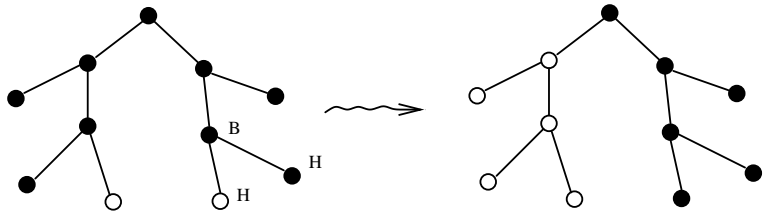

Fig. 4. Transition from one `soan` to the next via a `tihl`. Active nodes are drawn black. We assume the same type of order as shown in Figure 3.

Let us consider the `soan-tihl-` *process* $Y = (\mathcal{S}_0, e_1, \mathcal{S}_1, e_2, \mathcal{S}_2, ...)$ which starts from $\mathcal{S}_0 = \mathcal{N}$ and whose dynamics is given by the transition probability (6) and the rule $\mathcal{S}_{i+1} = [\mathcal{S}_i, e_{i+1}]$, see (5). Note that the transition probability on the `soan`s is

$$P(\mathcal{R}, \mathcal{S}) = \sum_{e: S=[R,e]} P(\mathcal{R}, e).$$

For a `tihl` e, and $j = 1, ..., \ell$, let us write $v_e(j) = 1$ if the leaf l_j belongs to supp e, and $v_e(j) = 0$ otherwise. We define

$$\mathbf{v}_e = (v_e(1), ..., v_e(\ell)).$$

For a realization $e_1, e_2, ...$ of the `tihl` process, let $\bar{m}$ be such that $e_{\bar{m}+1} = \texttt{End}$ if the process is stopped, and $\bar{m} = \infty$ otherwise. In order to count how many positions are emitted into the leaves by the first m `tihl`s in the process Y, we put

$$K_j(m) = \sum_{i=1}^{m \wedge \bar{m}} v_{e_i}(j),\ j = 1, ..., \ell; \quad \mathbf{K}(m) = (K_1(m), ...K_\ell(m)).$$

The multiple counting process $\mathbf{K}$ will be important in the next subsections; note that it is adapted to the process Y in the sense that $\mathbf{K}(m)$ can be read off from $(Y_1, ..., Y_m)$.

3.4 Generating Labelled Sequences in the Leaves

As a model for the *substitution process*, consider a time homogeneous Markov process X on a finite set $\mathcal{A}$ of *letters* in a reversible equilibrium α. For any `tihl` e, this induces a $\mathbb{T}_e$-indexed $\mathcal{A}$-valued process X^e in the following way: Start in r_e in the equilibrium α. Along an edge ϵ of $\mathbb{T}_e$ leading to a node which carries an H, the process develops (for the time t_ϵ) according to the substitution dynamics, whereas in a node carrying an N the process starts independently in distribution α.

By the process X^e, a `tihl` e assigns a random letter to each element of $\mathcal{L}_e := \mathcal{L} \cap \operatorname{supp} e$. The joint distribution of these letters will be denoted by α^e. Given a sequence $(e_1, e_2, ...)$ we assume that the processes $X^{e_1}, X^{e_2}, ...$ are independent.

Let $\mathbf{A}^j(m) = (A_1^j, ..., A_{K_j(m)}^j)$ be the sequence of letters assigned to leaf l_j by the first m `tihl`s $e_1, ..., e_m$ in the process Y, and put $\mathbf{A}(m) = (\mathbf{A}^1(m), ..., \mathbf{A}^\ell(m))$. The process $\mathbf{A}(m), m = 1, 2, ..$ is a variant of a multiple hidden Markov model in the

sense of [7]: the role of the hidden Markov chain is played by the `soan-tihl` process, which in each step emits letters into some subset of $\mathcal{L}$.

In the sequel, we denote by h the `tihl` which originates in the root r and assigns H to all $\nu \in \mathcal{N} \setminus \{r\}$.

Definition 4. *Fix* $\mathbf{k} = (k_1, ..., k_\ell) \in \mathbb{N}_0^\ell$ *and* $\mathbf{a_k} = (\mathbf{a}^1_{k_1}, ..., \mathbf{a}^\ell_{k_\ell})$, *where* $\mathbf{a}^j_{k_j} = (a^j_1, ..., a^j_{k_j})$ *is an* $\mathcal{A}$*-valued sequence of length* k_j *for* $j = 1, ..., \ell$.

a) (case TKF, i.e. $\gamma < 1$*) Let*

$$P(\mathbf{k}) = \mathbb{P}[\mathbf{A}(m) = \mathbf{a_k}; e_{m+1} = \texttt{End} \quad \text{for some } m \in \mathbb{N}_0].$$

For $\mathcal{S} \subseteq \mathcal{N}$*, let*

$$P_{\mathcal{S}}(\mathbf{k}) = \mathbb{P}[\mathbf{A}(m) = \mathbf{a_k};\ \mathcal{S}_m = \mathcal{S};\ e_{m+1} = \texttt{End} \quad \text{for some } m \in \mathbb{N}_0].$$

b) (case cTKF, i.e. $\gamma = 1$*) Let*

$$P(\mathbf{k}) = \mathbb{P}[\mathbf{A}(m) = \mathbf{a_k}; e_{m+1} = h \quad \text{for some } m \in \mathbb{N}_0].$$

For $\mathcal{S} \subseteq \mathcal{N}$*, let*

$$P_{\mathcal{S}}(\mathbf{k}) = \mathbb{P}[\mathbf{A}(m) = \mathbf{a_k};\ \mathcal{S}_m = \mathcal{S};\ e_{m+1} = h \quad \text{for some } m \in \mathbb{N}_0].$$

For a `tihl` e, and $\mathbf{k}, \mathbf{a_k}$ as in Definition 4, put

$$q(e) := \prod_{\sigma \in \operatorname{supp} e} (1 - \pi_{t(\sigma)}); \quad \vartheta(e, \mathbf{k}) = \alpha^e((a^j_{k_j})_{j \in \mathcal{L}_e}).$$

In words: $\vartheta(e, \mathbf{k})$ is the probability that the `tihl` e emits the letters $a^j_{k_j}$ into all leaves l_j which it reaches. The next lemma gives a recursion as well as the initial condition ($\mathbf{k} = \mathbf{0} = (0, ..., 0)$) for the $P_{\mathcal{S}}(\mathbf{k})$. In the TKF case this is in [8, 9]; we include a proof for convenience.

Lemma 1. *i) For* $\mathbf{k}, \mathbf{a_k}$ *as in Definition 4, and all* $\mathcal{S} \subseteq \mathcal{N}$*,*

$$P_{\mathcal{S}}(\mathbf{k}) = \sum_{(\mathcal{R}, e):\, \mathcal{S} = [\mathcal{R}, e]} p(e) q(e) P_{\mathcal{R}}(\mathbf{k} - \mathbf{v}_e) \vartheta(e, \mathbf{k}). \tag{7}$$

ii) In the TKF case (see Def. 4 a))

$$P_{\mathcal{N}}(\mathbf{0}) = P(\mathcal{N}, \texttt{End}) = \prod_{\sigma \in \mathcal{N}} (1 - \pi_{t(\sigma)}),$$

and in the cTKF case (see Def. 4 b))

$$P_{\mathcal{N}}(\mathbf{0}) = P(\mathcal{N}, h) = \prod_{\sigma \in \mathcal{N} \setminus \{r\}} (1 - \pi_{t(\sigma)}).$$

Proof. i) Because of (6) we have

$$\mathbb{P}[\mathbf{A}(m) = \mathbf{a_k},\ \mathcal{S}_m = \mathcal{S}] = \sum_{(\mathcal{R}, e):\, \mathcal{S} = [\mathcal{R}, e]} \mathbb{P}[\mathbf{A}(m-1) = \mathbf{a}_{\mathbf{k} - \mathbf{v}_e};\ \mathcal{S}_{m-1} = \mathcal{R}] \prod_{\sigma \in \mathcal{R}: \sigma < r_e} (1 - \pi_{t(\sigma)}) p(e) \vartheta(e, \mathbf{k}). \tag{8}$$

For all $\mathcal{R}$ and e such that $[\mathcal{R}, e] = \mathcal{S}$, one checks easily that

$$P(\mathcal{S}, \texttt{End}) \prod_{\sigma \in \mathcal{R}:\, \sigma < r_e} (1 - \pi_{t(\sigma)}) = P(\mathcal{R}, \texttt{End})\, q(e) \tag{9}$$

in the TKF case, and

$$P(\mathcal{S}, h) \prod_{\sigma \in \mathcal{R}: \sigma < r_e} (1 - \pi_{t(\sigma)}) = P(\mathcal{R}, h)\, q(e) \tag{10}$$

in the cTKF case. Multiplying both sides of (8) with $P(\mathcal{S}, \texttt{End})$ (or $P(\mathcal{S}, h)$), summing over m and using (9) (or (10)) we obtain the assertion.

ii) Because of the rule (5) it is impossible to have $\mathcal{S}_n = \mathcal{N}$ for some $n \geq 1$ and `tihl`s $e_1, ..., e_n$ with $\mathbf{v}_{e_1} = ... \mathbf{v}_{e_n} = 0$ (since any such `tihl` de-activates some leaf which can be re-activated only with some leaf e with $\mathbf{v}_e \neq 0$). On the other hand, any `tihl` e with $\mathbf{v}_e \neq 0$ leads to an increase of some coordinates of $\mathbf{k}$. Thus, the only contribution to $P_{\mathcal{N}}(\mathbf{0})$ comes from jumping from $\mathcal{N}$ to `End` (in the TKF case) or h (in the cTKF case) without any other `tihl` in between.

3.5 Computing Multiple Alignment Likelihoods

We proceed by explaining the method of Lunter et al. [9], which, as we saw in the previous subsection, generalizes easily to the cTKF case.

Noting that

$$P_{\mathcal{S}}(\mathbf{0}) = \sum_{(\mathcal{R}, e):\, \mathcal{S} = [\mathcal{R}, e]} p(e) q(e) P_{\mathcal{R}}(\mathbf{0}); \quad \mathcal{S} \neq \mathcal{N}, \tag{11}$$

we obtain from (7) and (11) by summing over all $\mathcal{S}$:

$$P(\mathbf{k}) = \sum_e p(e) q(e) P(\mathbf{k} - \mathbf{v}_e) \vartheta(e, \mathbf{k}). \tag{12}$$

$$P(\mathbf{0}) = P_{\mathcal{N}}(\mathbf{0}) + \sum_{e: \mathbf{v}_e = 0} p(e) q(e) P(\mathbf{0}) \tag{13}$$

Note that $P(\mathbf{k})$ and $P(\mathbf{0})$ appear on both sides of the equations. It is, however, straightforward to turn them into a recursion:

$$P(\mathbf{k}) = \frac{\sum_{e: \mathbf{v}_e \neq \mathbf{0}} p(e) q(e) \vartheta(e, \mathbf{k}) P(\mathbf{k} - \mathbf{v}_e)}{1 - \sum_{e: \mathbf{v}_e = \mathbf{0}} p(e) q(e)} \tag{14}$$

$$P(\mathbf{0}) = \frac{P_{\mathcal{N}}(\mathbf{0})}{1 - \sum_{e: \mathbf{v}_e = \mathbf{0}} p(e) q(e)} \tag{15}$$

Lunter et al. [9] make the computation still more efficient by using the accelerated chain $Y_{\tau(1)}, Y_{\tau(2)}, Y_{\tau(3)}, \ldots$ where $\tau(1) = 1$ and $\tau(m)$, $m > 1$ is the first time when either `End` occurs, or a `tihl` e which *overlaps* with one of the `tihl`s $e_{\tau(m-1)}$, ..., $e_{\tau(m)-1}$ in the sense that their rooted supports intersect, see Figure 5. Thus the set of `tihl`s occurring between $\tau(m-1)$ to $\tau(m) - 1$ are non-overlapping. Each *set of non-overlapping* `tihl`s (called *a set of nested events* in [9]) can be ordered, say $e_1 < e_2 < \ldots < e_j$, corresponding to the ordering of the $r(e_i)$ in the total order

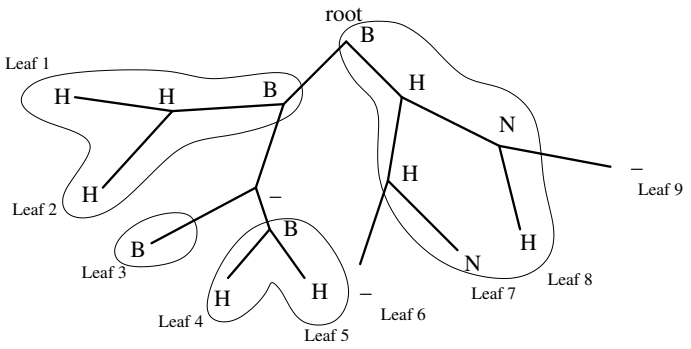

Fig. 5. A set of non-overlapping `tihl`s.

on $\mathcal{N}$. (This is precisely the order in which the non-overlapping `tihl`s occur in the process Y.) Call a set $\mathcal{E}$ of non-overlapping `tihl`s *silent* if $\mathbf{v}_e = \mathbf{0}$ for all $e \in \mathcal{E}$.

Let $\mathfrak{E}_0$ ($\mathfrak{E}_1$) denote the family of all silent (non-silent) sets of non-overlapping `tihl`s, and put $\mathfrak{E} := \mathfrak{E}_0 \cup \mathfrak{E}_1$. Similar as before but with an additional inclusion-exclusion argument, one can sum over the probabilities of all sets of non-overlapping `tihl`s that occur between $\tau(m-1)$ and $\tau(m)-1$:

$$P(\mathbf{k}) = \sum_j (-1)^{j-1} \sum_{\{e_1,\dots,e_j\}\in\mathfrak{E}} P(\mathbf{k} - \sum_i \mathbf{v}_{e_i}) \cdot \prod_i p(e_i) \cdot q(e_i) \cdot \vartheta(e_i, \mathbf{k}) \quad (16)$$

This leads to the recursion

$$P(\mathbf{k}) = \frac{\sum_j (-1)^{j-1} \sum_{\{e_1,\dots,e_j\}\in\mathfrak{E}_1} P(\mathbf{k} - \sum_i \mathbf{v}_{e_i}) \cdot \prod_i p(e_i) \cdot q(e_i) \cdot \vartheta(e_i, \mathbf{k})}{1 - \sum_j (-1)^{j-1} \sum_{\{e_1,\dots,e_j\}\in\mathfrak{E}_0} \prod_i p(e_i) \cdot q(e_i)} \quad (17)$$

Each $\mathcal{E} = \{e_1, ..., e_j\} \in \mathfrak{E}$ induces a mapping f from $\mathcal{N}$ to $\mathcal{Z} = \{-, B, H, N, E\}$ as follows: Put $f(\sigma) = e_i(\sigma)$ if $\sigma \preceq r(e_i)$ for some $i = 1, ..., j$, and $f(\sigma) = -$ otherwise. Lunter et al. [9] realized that the r.h.s. of (17) can be computed efficiently by grouping the elements $\mathcal{E} = \{e_1, ..., e_j\} \in \mathfrak{E}$ with respect to $\mathbf{V} = \mathbf{V}(\mathcal{E}) = \sum \mathbf{v}_{e_i}$. For $\mathbf{V} \in \{0,1\}^{\mathcal{L}}$ put

$$\mathcal{L}_{\mathbf{V}} =: \{l_\iota \in \mathcal{L} | \mathbf{V}_\iota = 1\}; \quad \mathfrak{E}_{\mathbf{V}} := \{\mathcal{E} = \{e_1, ..., e_j\} | \sum \mathbf{v}_{e_i} = \mathbf{V}\}.$$

Following [9], one can then compute for each node $\sigma \in \mathcal{N}$ and each $Z \in \mathcal{Z}$ the contribution to $P(\mathbf{k})$ from sets $\mathcal{E}$ of non-overlapping `tihl`s with $\mathbf{V}(\mathcal{E}) = \mathbf{V}$ that assign a Z to σ:

$$F(\mathbf{V}, \sigma, Z) := \sum_j (-1)^{j-1} \sum_{\{e_1,\dots,e_j\}} P(\mathbf{k} - \mathbf{V}) \cdot \prod_i p(e_i) \cdot q(e_i) \cdot \vartheta_\sigma(e_i, \mathbf{k}),$$

where the sum is taken over all $\mathcal{E} = \{e_1, ..., e_j\} \in \mathfrak{E}_{\mathbf{V}}$ that include an e_i with $e_i(\sigma) = Z$, and $\vartheta_\sigma(e_i, \mathbf{k})$ is the probability that, given the `tihl` e_i, the substitution process along e_i produces the labels $a^\iota_{k_\iota}$ in all the leaves l_ι which stem from σ and belong to $\mathcal{L}_{\mathbf{V}}$. The value $F(\mathbf{V}, \sigma, Z)$ can be computed directly from the corresponding values of the daughters of σ. Thus, using the dynamic programming idea

of Felsenstein's pruning algorithm [2], one can efficiently compute the values of all nodes, starting in the leaves and ending in the root. The numerator in (17) then results as $\sum_{\mathbf{V}\neq\mathbf{0}} F(\mathbf{V},r,B)P(\mathbf{k}-\mathbf{V})$, whereas the sum in the denominator of (17) equals $F(\mathbf{0},r,B)$. However, one still has to do this for all $\mathbf{k} \leq (n_1,....,n_\ell)$, so the time complexity is essentially the product of all sequence lengths n_ι and thus is exponential in the number ℓ of input sequences.

3.6 Extension to fragment insertions and deletions

When extending the TKF2 model (see subsection 2.1) or the FID model along a tree, one can either impose indivisibility of the fragments along the whole tree (as suggested in [3] for the TKF2 model), or one can allow that the fragmentation changes from edge to edge, which we assume in the sequel.

In the TKF and cTKF model, as soon as a `tihl` is born in some node of the tree, it grows independently of the indel history "to its left". This is different in the FID model: if the left neighbour position at some node σ carries an H, this enhances the probability of an H in the `tihl` at σ. Consequently, the `soan-tihl` process defined in subsection 3.3 is not Markov any more. To obtain a Markovian sequence, we have to keep track not only if a node is active, but also to remember if the previous (when looking back to the left) position carried a B, H or N. Our new state space now consists of all mappings φ from $\mathcal{N}$ to $\{B,H,N,E,B^-,H^-,N^-\}$. Let us put $S_\varphi := \{\sigma \in \mathcal{N} \,|\, \varphi(\sigma) \in \{B,H,N\}\}$. We say that a node σ is φ-*active* if $\sigma \in S_\varphi$, and φ-*inactive* otherwise. For a pair (ψ,φ) of such mappings we say that φ *activates* a ψ-inactive node σ if $\psi(\sigma) = \varphi(\sigma)^-$, and we say that φ *de–activates* a $\sigma \in S_\psi$ if $\varphi(\sigma) = \psi(\sigma)^-$.

The initial state is $\varphi_0 = h$, which maps r to B and all other nodes to H. The update rule replacing (5) is now

$$[\psi, e] = \varphi$$

where φ de-activates all nodes in $(S_\psi \cap \{\sigma | \sigma < r_e\}) \setminus \mathrm{supp}(e)$, re-activates all ψ-inactive nodes in $\mathrm{supp}(e)$, and sets $\varphi(r_e) = B$ and $\varphi(\sigma) = \psi(\sigma)$ for all $\sigma > r_e$. For $U \in \{B,H,N\}$ and $W \in \{H,N\}$ we put $\pi_t^{U^-B} = 0$ and $\pi_t^{U^-W} = \pi_t^{UW}$. Given the current state is ψ, the probability that the next inserted `tihl` is e equals

$$P(\psi,e) = p_\psi(e) \prod_{\min S_\psi \leq \sigma < r_e} (1 - \pi_{t(\sigma)}^{\psi(\sigma)B}),$$

where

$$p_\psi(e) = \pi_{t(r_e)}^{\psi(r_e)B} \prod_{\sigma \in \overline{\mathrm{supp}}\,(e)} \pi_{t(\sigma)}^{\psi(\sigma)e(\sigma)}.$$

Note that $\pi_t^{BU} = \pi_t^{NU}$ and $\pi_t^{B^-U} = \pi_t^{N^-U}$ for all $U \in \{B,H,N,E\}$. As a consequence, one can in fact restrict the domain of the functions φ to $\{B,H,E,B^-,H^-\}$. This decreases the number of functions φ, which is favourable for computational purposes.

Defining $P_\varphi(\mathbf{k})$ in analogy to Definition 4, we obtain in a similar way as in Lemma 1 for $(\mathbf{k},\varphi) \neq (\mathbf{0},\varphi_0)$ the "forward equation"

$$P_\varphi(\mathbf{k}) = \sum_{(\psi,e):\, \varphi=[\psi,e]} P_\psi(\mathbf{k}-\mathbf{v}_e)p_\psi(e)\vartheta(e,\mathbf{k}) \cdot \cdot \frac{p_\varphi(h)}{p_\psi(h)} \cdot \frac{1-\pi_{t(r_e)}^{\varphi(r_e)B}}{1-\pi_{t(r_e)}^{\psi(r_e)B}} \prod_{\sigma\in\mathrm{supp}(e)} (1-\pi_{t(\sigma)}^{\varphi(\sigma)B}) \tag{18}$$

and the start condition

$$P_{\varphi_0}(\mathbf{0}) = p_{\varphi_0}(h) \prod_{\sigma\neq r} (1-\pi_{t(\sigma)}^{HB}).$$

In order to turn (18) into a recursion one has to solve it for the vector $(P_\varphi(\mathbf{k}))_\varphi$, since $\mathbf{v}_e$ can be equal to $\mathbf{0}$. This seems only tractable as long as the number ℓ of input sequences is small. An alternative is to neglect all `tihl`s e with $\mathbf{v}_e = 0$, which seems legitimate if the indel rates are low. Another approach, which is briefly discussed in the next section, is to apply algorithms like simulated annealing or Markov chain Monte Carlo methods that assign sequences to the internal nodes (cf. [4], [7]).

4 Indel Models and Tree Reconstruction

Based on a stochastic insertion-deletion model, one can ask for the joint estimation of a multiple alignment and a phylogenetic tree by maximizing the probability of the alignment and the likelihood of the tree. *Exact* computations seem hopeless for large data sets. There are however promising heuristic approaches to this problem.

Thorne and Kishino [20] used the TKF model to get maximum-likelihood estimates of the pairwise evolutionary distances between the sequences in a data set from which they then reconstructed a neighbor-joining tree [16]. Although this approach might not be the best way to estimate the sequences' phylogeny, such a tree can be used as a guide tree for progressive statistical alignment and thus is a convenient starting point for further analysis. *Progressive statistical alignment* was introduced in [7] for the case of the TKF model: Given a tree $\mathcal{T}$, distinguish one of its nodes as the root and proceed from the leaves of $\mathcal{T}$ towards the root in the following way. Infer the most probable alignment for every pair of neighbouring nodes. Then, estimate a sequence at their parent node together with the indel history of this parent and its two children, which is compatible to the alignment of the offspring sequences. The sequence at the parent node is then aligned to its sibling, and so on. At the end of this procedure one has sequences for every node of the tree together with their indel history. This indel history and these inferred sequences depend, however, on the order in which the tree is traversed. In order to decrease this effect it is convenient to visit every node and every branch in random order, thereby emitting a new sequence at the respective node and optimizing the indel history along the respective branch. The sampling procedures for sequences at internal nodes use mutiple HMMs, where the observable emissions are the offspring sequences and the sequences at the neighbouring nodes, respectively.

In the heuristic outlined in [4] we use simulated annealing to approach a phylogenetic tree and an indel history whose posterior probability given a set of sequences $\mathbf{A}$ is maximal. We start with the progressive alignment procedure described above

and then modify the trees and indel histories in the following way: Inspired by the tree sampling procedure proposed in [10], we change the edge lengths in the current tree $\mathcal{T}^i$ independently by a small random amount, which results in a proposed tree $\mathcal{T}^*$. For short internal edges, the proposed length may come out negative, which we interprete as proposed change in the tree topology. Whenever this situation occurs one has to sample new sequences and a new indel history for the newly introduced internal edge, conditioned on the neighbouring sequences and their current alignment. Again, this is done with a multiple HMM, where the observable emissions are the sequences at the neighbouring positions. The indel history is further refined by producing a new sequence for each of the tree's internal nodes and by optimizing the pairwise alignments along the tree's branches. Thus, altering the tree also results in proposing a new indel history $\tilde{w}^*$. The proposal is accepted with probability $\min\left\{1, \left(\frac{p_{\mathcal{T}^*}(\mathbf{A},\tilde{w}^*)}{p_{\mathcal{T}^i}(\mathbf{A},\tilde{w}^i)}\right)^{c(i)}\right\}$ where $p_{\mathcal{T}}(\mathbf{A},\tilde{w})$ denotes the likelihood of $\mathcal{T}$, $\tilde{w}^i$ is the indel history in step i, and $c(i)$ increases with i.

In [4] also some alternative proposal chains are suggested which move faster through tree space, one of them based on nearest-neighbour interchanges (cf. [17]).

The estimation procedure becomes rather time consuming when applied to a data set consisting of a larger number of sequences. A way out might be to resort to a heuristics for tree reconstruction based on trees with four leaves only, known as quartet puzzling ([18]). This can be done in the following way: Find for every quartet of sequences the maximum-likelihood tree topology with the help of the recursion in section 3.5. Then, use the quartet puzzling algorithm to repeatedly combine all these quartet trees into a tree for the entire data set. Compute a consensus tree of all these intermediate trees and use this consensus tree as a guide tree for progressive statistical alignment Finally, optimize the branch lengths and the indel history by simulated annealing.

It is a challenging task and object of ongoing research to further improve the optimization heuristics for multiple alignment and phylogenetic trees. We have no doubt that this is most adequately done in the framework of stochastic insertion-deletion models and statistical alignment.

Acknowledgement

Our project was embedded into the DFG program "Interacting stochastic systems of high complexity". Financial support by the DFG and stimulating scientific exchange with other groups in the program, in particular those in Berlin, Erlangen and Frankfurt, is gratefully acknowledged. We would especially like to thank Matthias Birkner, Gerton Lunter and István Miklos for fruitful discussions.

References

1. J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368-376, 1981.
2. J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, 2004.
3. R. Fleißner, Sequence alignment and phylogenetic inference, PhD Thesis, Universität Düsseldorf, 2003.

4. R. Fleißner, D. Metzler, A. von Haeseler, Simultaneous statistical multiple alignment and phylogeny reconstruction, Preprint.
5. O. Gotoh, Multiple sequence alignments: algorithms and applications. *Adv. Biophys.* 36:159-206, 1999.
6. J. Hein, C. Wiuf, B. Knudsen, M.B. Møller, G. Wibling, Statistical alignment: Computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.*, 302:265-179, 2000.
7. I. Holmes, W. J. Bruno, Evolutionary HMMs: a Bayesian approach to multiple alignment, *Bioinformatics* 17:803-820, 2001.
8. G.A. Lunter, I. Miklós, J.L. Jensen, A. Drummond, J. Hein, Bayesian phylogenetic inference under a statistical insertion-deletion model, Proceedings of the Third International Workshop on Algorithms in Bioinformatics, Budapest, G. Benson, R.D.M. Page (eds), *Lecture Notes in Comp. Sc.* 2812, pp. 228-244, Springer 2003.
9. G.A. Lunter, I. Miklós, Y.S. Song, J. Hein, An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.* 10(6):869-889, 2003.
10. B. Mau, M. A. Newton, Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo, *J. Computational and Graphical Statistics* 6:122-131. 1997.
11. D. Metzler, Statistical alignment based on fragment insertion and deletion models, *Bioinformatics* 19:490-499, 2003.
12. D. Metzler, R. Fleißner, A. Wakolbinger, A. von Haeseler, Assessing variability by joint sampling of alignments and mutation rates, *J. Mol Evol.* 53:660-669, 2001.
13. I. Miklós, G.A. Lunter, I. Holmes, A "long indel" model for evolutionary sequence alignment. *Mol. Biol. Evol.* (accepted)
14. I. Miklós, Z. Toroczkai, Z. An improved model for statistical alignment. In: *Algorithms in bioinformatics* (O. Gascuel and B. M. E.Moret, eds) pp. 1-10, Springer.
15. G. J. Mitchison, A probabilistic treatment of phylogeny and sequence alignment, *J. Mol. Evol.* 49:11-22, 1999.
16. N. Saitou, M. Nei, The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425, 1987.
17. D. L. Swofford, G. L. Olsen, P. J. Waddell, D. M. Hillis, Phylogenetic Inference, in: *Molecular Systematics* (D. M. Hillis, C. Moritz, B. K. Mable, eds) pp. 407-514, Sinauer Accociates, Sunderland.
18. K. Strimmer, A. von Haeseler, Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13:964-969, 1996.
19. S. Tavaré, Some probabilistic and statistical problems on the analysis of DNA sequences, *Lec. Math. Life Sci.* 17, 57-86, 1986.
20. J. L. Thorne, H. Kishino, Freeing phylogenies from artifacts of alignment. *Mol. Biol. Evol.* 9:1148-1162, 1992.
21. J. L. Thorne, H. Kishino, J. Felsenstein, An evolutionary model for maximum likelihood alignment of DNA sequences, *J. Mol. Evol.* 33:114-124, 1991.
22. J. L. Thorne, H. Kishino, J. Felsenstein, Inching towards reality: an improved likelihood model for sequence evolution. *J. Mol. Evol.* 34:3-16, 1992.