# Applying the Thorne-Kishino-Felsenstein model to sequence evolution on a star-shaped tree

August 21, 2000

M. STEEL[1] AND J. HEIN[2]
[1] Biomathematics Research Centre
University of Canterbury, Christchurch, New Zealand
[2] Department of Ecology and Genetics,
University of Aarhus, Ny Munkegade, Aarhus, Denmark

**Keywords**– DNA sequence evolution, insertion-deletion model, dynamic programming

**Abstract**

Stochastic models that allow site substitutions, insertions and deletions provide a useful framework for a statistical approach to DNA sequence evolution. Such a model, and recursions to calculate the probability of evolving two sequences, have been known for almost a decade. In this paper we show how the pairwise recursions can be generalised to a 3-sequence tree, and more generally to a $r$-sequence star-shaped tree.

## 1 INTRODUCTION

Proteins and DNA sequences evolve predominantly by substitutions, insertions and deletions of single amino acids/nucleotides or strings of these elements. During the last two decades, the analysis of the substitution process has improved considerably, and has increasingly been based on stochastic models. The process of insertions and deletions has not received the same attention and is presently being analysed by optimisation techniques (parsimony or optimising a similarity score). A pioneering paper by Thorne, Kishino and Felsenstein [1] proposed a well defined model for insertion and deletions that allowed a proper statistical analysis for two sequences. Such an analysis can be used to provide maximum likelihood (pairwise) sequence alignments, or to estimate the evolutionary distance between two sequences. A useful tool for applications is a recursion for calculating the joint probability of sequences, and such a recursion was described for pairs of sequences in [1]. However this approach has until now not been generalised to allow analysis of three or more sequences related by a tree (other authors who have tried alternative approaches include [2], [3], [4], [5], [6]). Here we present a recursion that leads to a polynomial-time algorithm for calculating the probability of three sequences that evolve on a tree (and more generally $r$ sequences that evolve on a star-shaped tree) according to the Thorne-Kishino-Felsenstein process.

[1]We thank the New Zealand Marsden Fund (UOC-MIS-003), the Danish Research Council (S.N.F.), and the University of Canterbury Erskine Fund for supporting this research, and two anonymous referees for helpful comments.

## 1.1 Preliminaries

**Definition** Let $\mathcal{A}$ be a fixed alphabet of letters. A *sequence* $A$ is a finite string of letters over $\mathcal{A}$ of length $l(A) \geq 0$. We let $\emptyset$ denote the empty sequence of length 0, and if $A = a_1, \ldots, a_{n-m}, \ldots, a_n$ where $m \geq 0, n \geq 1$, we let $A_j = a_j$, and let $A[m]$ denote the sequence $a_1, \ldots, a_{n-m}$, and $A(m)$ denote the sequence $a_{n-m}, \ldots, a_n$. Thus, $A(0)$ is the last term of A, while $A[1]$ is the sequence up to, but excluding, the last term. By convention, $A[l(A)] = \emptyset$.

Suppose $T$ is a (phylogenetic) tree, with root vertex $v$, and $r \geq 3$ leaf vertices. To each leaf vertex $i$ is associated a given sequence $A^i$. In this paper we will assume the sequences evolve according to the Thorne-Kishino-Felsenstein (1991) reversible Markov model of insertions, deletions and substitutions, [1], [7] denoted more briefly as the $\mathcal{TKF}$-model. Briefly, this model used three classes of variables. Firstly, to each nucleotide was associated a reversible substitution process (identical to the usual site substitution models that did not allow insertions or deletions). Secondly, to each nucleotide is associated a deletion stochastic variable, $D_i$ that is exponentially distributed with parameter $\mu$. If this $D_i$ fires, the $i$–th nucleotide is removed. Thirdly, to the right of every nucleotide an insertion stochastic variable, called a mortal link, $I_i$, was associated. It is exponentially distributed with parameter, $\lambda < \mu$. If $I_i$ fires a nucleotide is chosen from the stationary distribution of the substitution process and is placed (along with a new mortal link to its right) to the right of $I_i$. If the $i$–th nucleotide dies, $I_i$ dies with it. To the left of the complete sequence there is an immortal link that can give birth to nucleotides (with associated mortal links), at the same rate as a mortal link, but cannot die. This prevents the empty sequence from becoming a sink. This model has been generalised further to allow insertions and deletions of blocks (see [8]) but we do not consider this extension here.

As a simple illustration of this model, if we denote the immortal link by the symbol $\bullet$ and a mortal link by the symbol $\star$, the sequence AGTT is represented as $\bullet$A$\star$G$\star$T$\star$T$\star$. If the third mortal link fires, with the resulting selection of (say) nucleotide C from the equilibrium distribution, and then the second nucleotide (G) dies then we obtain the representation $\bullet$A$\star$T$\star$C$\star$T$\star$ which corresponds to the sequence ATTC.

Let $\mathbb{P}(A^1, \ldots, A^r)$ denote the joint probability of observing sequences $A^1, \ldots, A^r$ at the leaves $1, \ldots, r$ respectively, under this model. Here we will deal just with the case $r = 3$ but the results may be generalised as we indicate later. We will let $t_i$ denote the (scaled) time parameter that the Markov process operates for on the edge of the tree incident with leaf $i$. Henceforth we will, without loss of generality, regard the sequences as evolving from the ancestral sequence $X$ at the internal vertex of the three-sequence tree. Note that, from [1], $l(X)$ has a geometric distribution, with

$$\mathbb{P}(l(X) = l) = (1 - \frac{\lambda}{\mu})(\frac{\lambda}{\mu})^l \tag{1}$$

## 1.2 Notation

- For brevity, we will denote the triple $A^1, A^2, A^3$ by $\mathbf{A}$. Thus, we will write, for example, $\mathbb{P}(\mathbf{A})$ in place of $\mathbb{P}(A^1, A^2, A^3)$. For $\mathbf{n} = (n_1, n_2, n_3)$, where $n_i \in \{0, \ldots, l(A^i)\}$ we let $\mathbf{A}[\mathbf{n}]$ denote the triple $A^1[n_1], A^2[n_2], A^3[n_3]$.

- For events $A, B, C$, $\mathbb{P}(A|B)$ denotes the conditional probability of $A$ given $B$, and we will also write $\mathbb{P}(A, B|C)$ (resp. $\mathbb{P}(A|B, C)$) as shorthand for $\mathbb{P}(A \cap B|C)$ (resp. $\mathbb{P}(A|B \cap C)$).

- For a letter $a \in \mathcal{A}$ let $\pi(a)$ denote the associated equilibrium probability under the $\mathcal{TKF}$ model. Set $\pi(\emptyset) := 1$, and for a sequence $A$, let $\pi(A) := \prod_{i=1}^{l(A)} \pi(A_i)$.
- Let $\mathbf{N} = (N_1, N_2, N_3)$ where $N_i \in \{0, \ldots, l(A^i)\}$ be the random variable that denotes the total number of descendants of the rightmost link of $X$ in $A^i$, and let $U \subseteq \{1, 2, 3\}$ be the random variable that denotes those indices $i$ for which the rightmost link of $X$ survives in $A^i$. Note that, when $l(X) = 0$, the rightmost link of $X$ is the immortal link. Also, $N_i = 0$ implies $i \notin U$ (but not conversely, since a mortal link may die after having left behind descendants that survive in $A^i$).

We will further economize by writing $\mathbf{n}$ and $\mathbf{u}$ in place of the events $\mathbf{N} = \mathbf{n}$ and $\mathbf{U} = \mathbf{u}$, respectively. For example, for $l \geq 0$, and $\mathbf{n} = (n_1, n_2, n_3)$, where $n_i \in \{0, \ldots, l(A^i)\}$ and $u \subseteq \{1, 2, 3\}$, we will adopt the convention

$$\mathbb{P}(\mathbf{A}, \mathbf{n}, u|l) := \mathbb{P}(A^1, A^2, A^3, \mathbf{N} = \mathbf{n}, U = u|l(X) = l)$$

(the joint probability of observing sequences $A^1, A^2, A^3$ at the leaves $1, 2, 3$ respectively, and that $N_i = n_i$ for $i = 1, 2, 3$ and $U = u$, conditional upon the event that $l(X) = l$). Following [1], let $p_k(t)$ (resp. $p'_k(t)$) denote the probability that after duration $t$, a mortal link has exactly $k$ descendants, and one of these is (resp. is not) the original mortal link. Let $p''_k(t)$ denote the probability that after duration $t$ an immortal link has exactly $k$ descendants (including itself). From [1] and [7] we have:

$$p_k(t) = e^{-\mu t}[1 - \lambda\beta(t)][\lambda\beta(t)]^{k-1}; k > 0$$
$$p'_k(t) = [1 - e^{-\mu t} - \mu\beta(t)][1 - \lambda\beta(t)][\lambda\beta(t)]^{k-1}; k > 0$$
$$p'_0(t) = \mu\beta(t)$$

and

$$p''_k(t) = [1 - \lambda\beta(t)][\lambda\beta(t)]^{k-1}; k > 0$$

where

$$\beta(t) = \frac{1 - e^{(\lambda-\mu)t}}{\mu - \lambda e^{(\lambda-\mu)t}}, 0 < \lambda < \mu$$

First we calculate the probability of three empty sequences.

**Lemma 1**

$$\mathbb{P}(\emptyset, \emptyset, \emptyset) = (1 - \frac{\lambda}{\mu})\frac{\prod_{i=1}^{3}(1 - \lambda\beta(t_i))}{1 - \lambda\mu^2\beta(t_1)\beta(t_2)\beta(t_3)}$$

**Proof.** We have,

$$\mathbb{P}(\emptyset, \emptyset, \emptyset) = \sum_{l \geq 0} \mathbb{P}(\emptyset, \emptyset, \emptyset|l(X) = l)\mathbb{P}(l(X) = l),$$

and $\mathbb{P}(\emptyset, \emptyset, \emptyset|l(X) = l) = \prod_{i=1}^{3} p''_1(t_i)(p'_0(t_i))^l$. The result now follows from Equation (1), upon substituting the above formulae for $p''_k(t)$ and $p'_0(t)$ and then summing the resulting geometric series.

## 2 THE RECURSION

Our aim is to establish a recursion (Theorem 1) for $\mathbb{P}(\mathbf{A})$ in terms of the joint probabilities of initial segments of the sequences in $\mathbf{A}$. Letting $\mathbb{P}(\mathbf{A}|l) = \mathbb{P}(\mathbf{A}|l(X) = l)$, we have

$$\mathbb{P}(\mathbf{A}) = \sum_{l\geq 0} \mathbb{P}(\mathbf{A}|l)\mathbb{P}(l(X) = l)$$

and so, from Equation (1),

$$\mathbb{P}(\mathbf{A}) = (1 - \frac{\lambda}{\mu}) \sum_{l\geq 0} \mathbb{P}(\mathbf{A}|l)(\frac{\lambda}{\mu})^l \tag{2}$$

To evaluate $\mathbb{P}(\mathbf{A}|l)$ we deal with the two cases $l = 0$ and $l > 0$ separately. Firstly, for $l = 0$, under the $\mathcal{TKF}$-model,

$$\mathbb{P}(\mathbf{A}|l(X) = 0) = \prod_{i=1}^{3} \pi(A^i) p''_{l(A^i)}(t_i) \tag{3}$$

Now suppose $l \geq 1$. By elementary probability,

$$\mathbb{P}(\mathbf{A}|l) = \sum_{(\mathbf{n},u)\in\mathcal{I}(\mathbf{A})} \mathbb{P}(\mathbf{A}, \mathbf{n}, u|l) \tag{4}$$

where

$$\mathcal{I}(\mathbf{A}) := \{((n_1, n_2, n_3), u) : 0 \leq n_i \leq l(A^i), i = 1, 2, 3; u \subseteq \{1, 2, 3\}; n_i = 0 \Rightarrow i \notin u\}$$

denotes the subset of those values of $(\mathbf{n}, u)$ for which $\mathbb{P}(\mathbf{A}, \mathbf{n}, u|l)$ can take a positive value. Using the identity $\mathbb{P}(U, V|W) = \mathbb{P}(U|V, W)\mathbb{P}(V|W)$ (for any three events $U, V, W$) we obtain:

$$\mathbb{P}(\mathbf{A}, \mathbf{n}, u|l) = \mathbb{P}(\mathbf{A}|l, \mathbf{n}, u) \times \mathbb{P}(\mathbf{n}, u|l) \tag{5}$$

Now, conditional on the events that $l(X) = l$, $\mathbf{N} = \mathbf{n}$ and $U = u$ the sequences $\mathbf{A}$ evolve from $X$ precisely if the first $l - 1$ nucleotides of $X$ evolve into $\mathbf{A}[\mathbf{n}]$ while the remaining nucleotide (and the last mortal link) of $X$ evolves into the remainder of $\mathbf{A}$. Since we have conditioned on (i) exactly how many descendants the last link of $X$ possesses in each of $A^1, A^2, A^3$, and (ii) in which of these sequences the last link of $X$ has survived, we can express $\mathbb{P}(\mathbf{A}|l, \mathbf{n}, u)$ as follows.

$$\mathbb{P}(\mathbf{A}|l, \mathbf{n}, u) = \mathbb{P}(\mathbf{A}[\mathbf{n}]|l - 1) \times w_1(\mathbf{A}, \mathbf{n}, u) \tag{6}$$

where

$$w_1(\mathbf{A}, \mathbf{n}, u) = (\prod_{i\notin u: n_i\geq 1} \pi(x_i)) \times \mathbb{P}(\{(i, x_i) : i \in u\}) \times \prod_{1\leq i\leq 3: n_i\geq 2} \pi(A^i(n_i - 2))$$

and where $x_i = A^i_{l(A^i)-n_i+1}$ and for $S \subseteq \{1, 2, 3\}$, $\mathbb{P}(\{(i, x_i) : i \in S\})$ is the joint probability of observing the letter $x_i$ at leaf $i$ for each leaf $i \in S$.

This last joint probability term can easily be computed from the edge lengths $(t_1, t_2, t_3)$ and the rate matrix for the underlying site substitution process. Furthermore, since $l \geq 1$,

$$\mathbb{P}(\mathbf{n}, u|l) = \prod_{i=1}^{3} \mathbb{P}(N_i = n_i, U \cap \{i\} = u \cap \{i\}) = \prod_{i=1}^{3} p_{n_i}^{|u\cap\{i\}|}(t_i) \tag{7}$$

where $p_k^0(t) := p_k'(t)$ and $p_k^1(t) := p_k(t)$. Note that the right hand side of Equation (7) (with $l \geq 1$) is dependent only on $\mathbf{n}$ and $u$, and hence we can denote it as $w_2(\mathbf{n}, u)$. Let

$$w(\mathbf{A}, \mathbf{n}, u) := w_1(\mathbf{A}, \mathbf{n}, u) w_2(\mathbf{n}, u).$$

Then, by combining Equations (4), (5), (6) and (7) we have, for $l \geq 1$,

$$\mathbb{P}(\mathbf{A}|l) = \sum_{(\mathbf{n},u)\in\mathcal{I}(\mathbf{A})} w(\mathbf{A}, \mathbf{n}, u)\mathbb{P}(\mathbf{A}[\mathbf{n}]|l-1) \tag{8}$$

Combining Equations (2), (3) and (8) we have:

$$\mathbb{P}(\mathbf{A}) = (1-\frac{\lambda}{\mu})[\prod_{i=1}^{3} \pi(A^i) p''_{l(A^i)}(t_i) + \sum_{(\mathbf{n},u)\in\mathcal{I}(\mathbf{A})} w(\mathbf{A}, \mathbf{n}, u) \sum_{l\geq 1} \mathbb{P}(\mathbf{A}[\mathbf{n}]|l-1)(\frac{\lambda}{\mu})^l] \tag{9}$$

Rearranging this last equation we have:

$$\mathbb{P}(\mathbf{A}) = (1-\frac{\lambda}{\mu})[\prod_{i=1}^{3} \pi(A^i) p''_{l(A^i)}(t_i) + \frac{\lambda}{\mu} \sum_{(\mathbf{n},u)\in\mathcal{I}(\mathbf{A})} w(\mathbf{A}, \mathbf{n}, u) \sum_{s\geq 0} \mathbb{P}(\mathbf{A}[\mathbf{n}]|s)(\frac{\lambda}{\mu})^s]$$

and applying Equation (2) to the second term in this last equation gives:

**Theorem 1** *Under the $\mathcal{TKF}$-model, the probability $\mathbb{P}(\mathbf{A})$ of generating the three sequences $\mathbf{A} = A^1, A^2, A^3$ satisfies the following recurrence equation.*

$$\mathbb{P}(\mathbf{A}) = (1-\frac{\lambda}{\mu})[\prod_{i=1}^{3} \pi(A^i) p''_{l(A^i)}(t_i)] + \frac{\lambda}{\mu} \sum_{(\mathbf{n},u)\in\mathcal{I}(\mathbf{A})} w(\mathbf{A}, \mathbf{n}, u)\mathbb{P}(\mathbf{A}[\mathbf{n}])$$

### 2.1 Example

As a simple example, we can use Theorem 1 to compute $\mathbb{P}(\mathbf{A})$ for $\mathbf{A} = \emptyset, \emptyset, \emptyset$. In this case we have: $\mathcal{I}(\mathbf{A}) = \{(\mathbf{0}, \emptyset)\}$ (where $\mathbf{0} = (0,0,0)$) and so, by Theorem 1,

$$\mathbb{P}(\mathbf{A}) = (1-\frac{\lambda}{\mu}) \prod_{i=1}^{3} p''_0(t_i) + \frac{\lambda}{\mu} w(\mathbf{A}, \mathbf{0}, \emptyset)\mathbb{P}(\mathbf{A}[\mathbf{0}]).$$

Now, $\mathbf{A}[\mathbf{0}] = \mathbf{A}$, and for this example, $w(\mathbf{A}, \mathbf{0}, \emptyset) = w_2(\mathbf{0}, \emptyset) = \prod_{i=1}^{3} p'_0(t_i)$ which, upon substitution for $p'_0(t_i)$ and $p''_0(t_i)$, leads to the formula for $\mathbb{P}(\mathbf{A})$ described by Lemma 1.

### 2.2 A polynomial-time algorithm

Given three sequences $B^1, B^2, B^3$, one can recursively use Theorem 1 to compute $\mathbb{P}(\mathbf{A})$ for all initial segments $\mathbf{A} = A^1, A^2, A^3$ of these sequences. Note that we can rewrite Theorem 1 in the following form:

$$\mathbb{P}(\mathbf{A}) = (1-\frac{\lambda}{\mu} w(\mathbf{A}, \mathbf{0}, \emptyset))^{-1}((1-\frac{\lambda}{\mu})[\prod_{i=1}^{3} \pi(A^i) p''_{l(A^i)}(t_i)] + \frac{\lambda}{\mu} \sum_{(\mathbf{n},u)\in\mathcal{I}^*(\mathbf{A})} w(\mathbf{A}, \mathbf{n}, u)\mathbb{P}(\mathbf{A}[\mathbf{n}])) \tag{10}$$

where $\mathcal{I}^*(\mathbf{A}) = \mathcal{I}(\mathbf{A}) - (\mathbf{0}, \emptyset)$. The advantage of this representation is that each triple of sequences appearing on the right hand side of Equation (10) has a combined total length at most one less than that of the sequences in $\mathbf{A}$. In this way one can recursively compute $\mathbb{P}(\mathbf{A}[\mathbf{k}])$ for all triples $\mathbf{k} = k_1, k_2, k_3$ where $k_i \leq l(A^i)$, and thereby construct a polynomial-time algorithm for computing $\mathbb{P}(\mathbf{A})$ of complexity $O(l^6)$ where $l = \max\{l(A^i), i = 1, 2, 3\}$.

### 2.3 Extension to star-shaped trees

If $r \geq 4$ sequences were related by a *star-shaped tree*, that is, a tree with only one internal node, then an analogous recursion to (10) holds. For example, the probability of observing $r$ empty sequences at the leaves, obtained by a similar reasoning, is

$$\mathbb{P}(\emptyset, \emptyset, \emptyset \ldots, \emptyset) = (1 - \frac{\lambda}{\mu}) \frac{\prod_{i=1}^{r}(1 - \lambda\beta(t_i))}{1 - \lambda\mu^{r-1}\prod_{i=1}^{r}\beta(t_i)}$$

The other arguments given in the 3-sequence case generalise accordingly.

### 2.4 Summary

Two challenges are immediate: First, it seems plausible that an $O(l^r)$ algorithm is possible for the $r$-star-shaped tree (where $l = \max_i\{l(A^i)\}$). The algorithm in [1] can be formulated as a 2-sequence analogue to Equation (10) and this would lead to an $O(l^4)$ algorithm. However, a modified approach allows for an $O(l^2)$ algorithm, as described in [1]. This trick can most likely also be used in the cases considered in this paper. Secondly, the presentation of an algorithm for triplewise sequence calculations may lead to a more useful algorithm performing statistical alignment of many sequences related by a binary tree. Such an algorithm in itself is bound to be impractically slow. Nevertheless, the analogous generalisation of the pairwise sequence alignment algorithm to the $r$ sequence by Sankoff [9] was the subsequent inspiration of many useful approximate and heuristic methods.

## References


[1] J.L. Thorne, H. Kishino and J. Felsenstein, An evolutionary model for maximum likelihood alignment of DNA sequences, *J. Mol. Evol.* **33**, 114-124 (1991).

[2] L. Allison, C.S. Wallace and C.N. Yee, Minimum message length encoding, evolutionary trees and multiple alignment. In *Hawaii International Conference on System Sciences*, vol. 1, pp. 663-674 (1992).

[3] G. Mitchison, A probabilistic treatment of phylogeny and sequence alignment, *J. Mol. Evol.* **49(1)**, 11-22 (1999).

[4] G. Mitchison and R. Durbin, Tree based maximum likelihood substitution matrices and hidden Markov models, *J. Mol. Evol.* **41**, 1139-1151 (1995).

[5] M.J. Bishop and E.A. Thompson, Maximum likelihood alignment of DNA sequences, *J. Mol. Biol.* **190**, 159-165 (1986).

[6] J. Hein, C. Wiuf, B. Knudsen, M.B. Møller and G. Wibling, Statistical alignment: computational properties, homology testing and goodness-of-fit, *J. Mol. Biol.* in press (2000).

[7] J.L. Thorne, H. Kishino and J. Felsenstein, Erratum, An evolutionary model for maximum likelihood alignment of DNA sequences, *J. Mol. Evol.* **34**, 91-92 (1992).

[8] J.L. Thorne, H. Kishino, and J. Felsenstein, Inching toward reality: An improved likelihood model of sequence evolution, *J. Mol. Evol.* **34**, 3-16 (1992).

[9] D. Sankoff, Minimal mutation trees of sequences, *SIAM J. of Appl. Math.* **78**, 35-42 (1975).