Likelihood computation under TKF91 model for a 3 taxa tree

Problem

Given a 3 taxa phylogenetic tree (Figure 1), where Ω denotes the root node (always placed on a taxa), the model parameters θ (substitution rates, insertion rate per link λ and deletion rate per link μ) and an MSA α, the proposed algorithm calculates the likelihood under the TKF91 model.

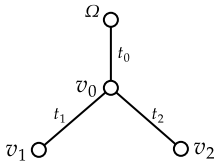

Figure 1: 3 taxa tree

Solution

The likelihood under TKF91 can be factorized in this way

$$P(\alpha \mid \theta) = P(\alpha, \alpha' \mid \theta) = P(\alpha \mid \alpha', \theta) P(\alpha' \mid \theta)$$

where $P(\alpha' | \theta)$ calculates the geometric probability $\gamma_N(\lambda, \mu)$ that the ancestral sequence has N links and considers only the structure of the MSA α', that is insertions and deletions, converting the original MSA into character/gap states (#/-) while $P(\alpha | \alpha', \theta)$ accounts for the substitution process given the MSA structure and the model parameters. Next we describe how to calculate the likelihood $P(\alpha' | \theta)$.

The likelihood of the MSA structure $P(\alpha' | \theta)$ is composed by the product of the likelihood of each block where lk_i is the likelihood of block i

$$P(\alpha' | \theta) = \gamma_N(\lambda, \mu) \prod_i \mathrm{lk}_i$$

The probabilities for the different possible fates of a single link (character) evolving along a tree branch of length t are equal to

$$\begin{aligned}
P_n''(t) &= \left(1 - \lambda\beta(t)\right) \left(\lambda\beta(t)\right)^{n-1} && n > 0 \\
P_0'(t) &= \mu\beta(t) \\
P_n'(t) &= \left(1 - e^{-\mu t}\mu\beta(t)\right) \left(1 - \lambda\beta\left(t\right)\right) \left(\lambda\beta\left(t\right)\right)^{n-1} && n > 0 \\
P_n(t) &= e^{-\mu t} \left(1 - \lambda\beta\left(t\right)\right) \left(\lambda\beta(t)\right)^{n-1} && n > 0
\end{aligned}$$

(for their derivation refer to the original paper).

Marginalization of (not observable) empty column

All children born from the ancestor link who die along t_1 and t_2 are not visible in the sequences in the nodes v_1 and v_2. The number of these links is not observable and must be marginalized. Below is the analytical marginalization for these empty columns.

case $P_n(t)$

$$\begin{aligned}
&P_n(t_0) + P_{n+1}(t_0)P_0{}'(t_1)P_0{}'(t_2) + P_{n+2}(t_0)P_0{}'(t_1)^2P_0{}'(t_2)^2 + \ldots = \\
&\quad = P_n(t_0)\left(\lambda\beta(t_0)\,\mu\beta(t_1)\,\mu\beta(t_2) + \left(\lambda\beta(t_0)\,\mu\beta(t_1)\,\mu\beta(t_2)\right)^2 + \ldots\right) = \\
&\quad = P_n(t_0)\sum_{k=0}^{\infty}\left(\lambda\beta(t_0)\,\mu\beta(t_1)\,\mu\beta(t_2)\right)^k = \\
&\quad = P_n(t_0)P^\star(t_0,t_1,t_2)
\end{aligned}$$

where

$$P^\star(t_0,t_1,t_2) = \frac{1}{1-\lambda\mu^2\beta(t_0)\,\beta(t_1)\,\beta(t_2)} = \frac{1}{1-\frac{\lambda}{\mu}\prod_{k=0}^{2}P_0{}'(t_k)}$$

case $P_n{}'(t)$

$$\begin{aligned}
&P_n{}'(t_0) + P'_{n+1}(t_0)P_0{}'(t_1)P_0{}'(t_2) + P'_{n+2}(t_0)P_0{}'(t_1)^2P_0{}'(t_2)^2 + \ldots = \\
&\quad = P_n{}'(t_0)\sum_{k=0}^{\infty}\left(\lambda\beta(t_0)\,\mu\beta(t_1)\,\mu\beta(t_2)\right)^k = \\
&\quad = P_n{}'(t_0)P^\star(t_0,t_1,t_2)
\end{aligned}$$

case $P_n{}''(t)$

$$\begin{aligned}
&P_n{}''(t_0) + P''_{n+1}(t_0)P_0{}'(t_1)P_0{}'(t_2) + P''_{n+2}(t_0)P_0{}'(t_1)^2P_0{}'(t_2)^2 + \ldots = \\
&\quad = P_n{}''(t_0)P^\star(t_0,t_1,t_2)
\end{aligned}$$

Therefore $P^\star(t_0,t_1,t_2)$ is the marginal probability for all links in v_0 that die in both v_1 and v_2, regardless of whether they extend $P_n(t)$, $P'_n(t)$ or $P''_n(t)$.

Split the MSA into blocks

The first block (if visible) is treated differently from all of the following because it is generated by the immortal link (see Section First block).

Each block represents the history of a single ancestor character. The character can generate children and possibly be deleted, therefore in each block there is only one ancestor character and all the others were generated by itself or its descendants.

Ω	-	-	#	-	-	-	#	#	-	#	-	#	#	-	-
v_1	#	#	#	#	-	-	#	-	#	-	#	#	-	#	-
v_2	-	#	-	-	#	#	-	#	-	-	-	#	-	#	#

Figure 2: MSA splitted into blocks (gray scale)

From the N blocks - and therefore from the total number of N ancestral characters - we can calculate the geometric probability $\gamma_N(\lambda, \mu) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^N$.

The expected number of characters at the root Ω is $\mathbb{E}(N) = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}}$.

Column categories

In this section we describe the procedure for an internal block (not generated by the immortal link). The block is divided into 6 sub-blocks (note that some sub-blocks may be empty) with the following characteristics:

	A	B			C		D			E		F	
Ω	#	-	-	-	-	-	-	-	-	-	-	-	-
v_0	?	?	?	?	?	?	?	?	?	#	#	#	#
v_1	#/-	#	#	#	#	#	-	-	-	-	-	-	-
v_2	#/-	#	#	#	-	-	#	#	#	-	-	-	-

Figure 3: 5 categories of sub-blocks

A: the first column (which contains always the ancestor at the root Ω)

B: all columns having 2 characters, the number of these columns is denoted with k

C: all columns with only one character in v_1 but not in v_2. The number of these columns is denoted with L

D: all columns with only one character in v_2 but not in v_1. The number of these columns is denoted with R

E: extra columns (not in the original block, they extend the canonical block). These columns contain an ancestor in v_0 that dies in both v_1 and v_2 leaving descendants in the canonical block (in C and/or D). The number of these columns is denoted with W.

F: the ancestors in v_0 of these columns die in both v_1 and v_2 without leaving any survivors. The likelihood for these columns is marginalized for an infinite and unobservable number of such columns in $P^{\star}(t_0, t_1, t_2)$.

Category A

The first column of the block can be found in 4 combinations, 3 make type I and 1 type II.

	l_1	l_2	l_3	II
Ω	#	#	#	#
v_0	#	#	#	#/-
v_1	#	#	-	-
v_2	#	-	#	-

Figure 4: 4 possible configurations of the 1st column and their type (I/II)

In type I (l_1,l_2,l_3) the status of the ancestor in node v_0 can only be a link since a gap cannot generate characters (links). In type II we have 2 possible states, both link and gap, so we have to marginalize over these 2 possibilities.

Category B

Columns in the sub-block of category B have only one possible state in v_0 and that is a link, a gap could not yield homologous 2 characters in v_1 and v_2.

Category C and D

Ancestors must be assigned to the columns of these 2 categories. The characters in these columns can be either children of the first column (if it contains a character) or of one of the link present in the columns B, or they can have a character themselves

(in their column) or come from other columns that are not part of the canonical block, E (extra columns), that cannot be observed in the MSA.

If the first column is of type I then $K = 1$, the case the first column is of type II will be described later. Then the k columns of type B are summed up together obtaining $K = 1 + k$. Note that even with $k = 0$, K is always greater than 0.

Parent assignment

The first thing to note is that all columns of type A I and B have a clear ancestor while the parents of all columns C and D remain to be assigned. Here are 2 examples (Figure 5 and 6).

Ω	#	-	-	-	-	-
v_0	#	-	#	-	-	#
v_1	-	#	-	-	#	-
v_2	#	-	#	#	-	-

Figure 5: The extra column (right) leaves no offsprings in the 2 sequences, this case is calculated in the factor $P^{\star}(t_0, t_1, t_2)$. This column is of type F.

Ω	#	-	-	-	-	-
v_0	#	-	-	-	-	#
v_1	-	#	-	-	#	-
v_2	#	-	#	#	-	-

Figure 6: The extra column, even if not observable in the block, leaves some offspring observable in either of the 2 taxa. This column is of type E.

Case I_1

To assign a parent to all characters in C and D we have $K = 1 + k + c$ with $c = 0,1,\ldots, L + R$ possible parents in the ancestral sequence in v_0. With $c = 0$ we assign all the characters in C and D to the already known characters present (note that if $k = 0$, $K = 1$ and therefore we always have at least one candidate parent). With $c = L + R$ each character has only 1 descendant (all characters survive without giving birth to children). The case $c > L + R$ means that all characters beyond $c = L + R$ die in both sequences. This case is marginalized in the factor

$P^{\star}(t_0, t_1, t_2)$. With $K = 1 + k$ we must assign the $L + R$ characters to the $1 + k$ possible parents belonging to the 5 categories A, B, C, D, E but since $c = 0$ we cannot add new fathers therefore C, D, E get 0 and therefore there are only 2 categories to assign the children to, *i.e.* A and B. With the first column of type I_1 then categories A and B can be merged into one single category AB. With $k = 0$ then there is no sub-block of type B.

Let's make an example. Sub-blocks C and D have together 9 links, $L = 4$ and $R = 5$. The links are generated by 3 candidate links $k = 2$ links in sub-block AB. Now we show that it is not necessary to partition the 9 links in these 3 fathers because the likelihoods are equivalent.

Let's first compute the partition of 9 in 1 part (the complete table of partition 9 can be found below):

$\mathbb{Q}(9,1) =$

AB
9

Here we use the symbol $\mathbb{Q}$ for the partition of candidate fathers into sub-blocks while the symbol Q refers to the partition of children between links.

Let's try partitioning 9 into 3 candidate fathers, $Q(9,3)$, and for this example we use 2 different partitions, *i.e.* 3|3|3 and 2|2|5. Of these 2 partitions we show one of the possible split into v_1 and v_2 as follows:

A)

link 1		link 2		link 3	
3		3		3	
v_1	v_2	v_1	v_2	v_1	v_2
2	1	2	1	0	3

B)

link 1		link 2		link 3	
2		2		5	
v_1	v_2	v_1	v_2	v_1	v_2
1	1	0	2	3	2

case A) the likelihood gets

$$\mathrm{lk_A} = \left(P_{1+2}(t_1)P_{1+1}(t_2)\right)\left(P_{1+2}(t_1)P_{1+1}(t_2)\right)\left(P_{1+0}(t_1)P_{1+3}(t_2)\right)$$

case B)

$$\mathrm{lk_B} = \left(P_{1+1}(t_1)P_{1+1}(t_2)\right)\left(P_{1+0}(t_1)P_{1+2}(t_2)\right)\left(P_{1+3}(t_1)P_{1+2}(t_2)\right)$$

which can we rewritten as follow:

$$\mathrm{lk_A} = P_1(t_1)\left(\lambda\beta(t_1)\right)^2 P_1(t_2)\left(\lambda\beta(t_2)\right) P_1(t_1)\left(\lambda\beta(t_1)\right)^2 P_1(t_2)\left(\lambda\beta(t_2)\right) P_1(t_1)P_1(t_2)\left(\lambda\beta(t_2)\right)^3$$

$$= P_1(t_1)P_1(t_2)P_1(t_1)P_1(t_2)P_1(t_1)P_1(t_2)\left(\lambda\beta(t_1)\right)^4\left(\lambda\beta(t_2)\right)^5$$

and

$$\mathrm{lk_B} = P_1(t_1)\left(\lambda\beta(t_1)\right) P_1(t_2)\left(\lambda\beta(t_2)\right) P_1(t_1)P_1(t_2)\left(\lambda\beta(t_2)\right)^2 P_1(t_1)\left(\lambda\beta(t_1)\right)^3 P_1(t_2)\left(\lambda\beta(t_2)\right)^2$$

$$= P_1(t_1)P_1(t_2)P_1(t_1)P_1(t_2)P_1(t_1)P_1(t_2)\left(\lambda\beta(t_1)\right)^4 \left(\lambda\beta(t_2)\right)^5$$

so $\mathrm{lk_A} = \mathrm{lk_B}$.

Loop

$c = 1$

At this point the loop increases the number of new parents (variable c) in the internal state. Increasing the number of fathers also changes $P_K(t_0)$ to $P_{K+c}(t_0) = P_K(t_0)\left(\lambda\beta(t_0)\right)^c$, the probability that the ancestor character in Ω generates one more character along the branch of length t_0. The new ancestor in v_0 must be assigned to a sub-block, denoted $\mathbb{Q}$. We have 3 possible sub-blocks, C, D and also sub-block E. We calculate the possible partitions of 1 new father in 3 parts (C,D,E) and we get

$\mathbb{Q}(1,3) =$

1	0	0

the number of unique permutations, denoted with $\tilde{\mathbb{Q}}$ is:

$\tilde{\mathbb{Q}}(1,3) =$

C	D	E
1	0	0
0	1	0
0	0	1

(note that the two '0' are not distinguishable from each other). The table shows 3 possible solutions to assign the new father character to sub-blocks. We have to marginalize the 3 possibilities by adapting the L and R values. In fact when we choose the case 1|0|0 we have one less character in the C block to assign to a father candidate, when we calculate the 0|1|0 case we have one less character in the D block to assign and when we finally calculate the 0|0|1 case, L and R remain unchanged because we added an extra column to E, that is $W = 1$.

$c = 7$

We now show the case with $c = 7$, which means that there are in total 10 candidate parent links ($K = 1 + k + c = 1 + 2 + 7 = 10$) to be splitted into 3 sub-block types, *i.e.* C, D, E. Let's calculate first

$\mathbb{Q}(7,3) =$

7	0	0
6	1	0
5	1	1
4	2	1
3	3	1
5	2	0
3	2	2
4	3	0

and then

$\tilde{\mathbb{Q}}(7,3) =$

C	D	E
7	0	0
0	7	0
0	0	7
6	1	0
1	6	0
6	0	1
0	6	1
0	1	6
1	0	6
5	1	1
1	1	5
1	5	1

C	D	E
4	2	1
2	4	1
2	1	4
1	2	4
1	4	2
4	1	2
3	3	1
3	1	3
1	3	3
5	2	0
5	0	2
2	5	0

C	D	E
0	5	2
2	0	5
0	2	5
3	2	2
2	3	2
2	2	3
4	3	0
0	4	3
0	3	4
3	4	0
4	0	3
3	0	4

This table contains configurations that are not possible (constrained by $L = 4$ or $R = 5$) and must be deleted (highlighted in red), all the others are plausible and must be marginalized.

Let's see some examples:

<u>case 0|0|7</u>

In this case L and R remain unchanged, $W = 7$ extra columns are added to the block E, this means that there are 7 new fathers in sub-block E. They have all to become at least a child because, the case that one or more of them die without giving birth is already considered in the marginal $P^{\star}(t_0, t_1, t_2)$ (sub-block F). We have now to assign the 9 links to the sub-block AB, E while ensuring at least 7 links to sub-block E. In sub-blocks C and D there are currently no candidate parents and therefore they cannot receive children. Here the integer partion of 9:

$Q(9) =$

9	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0
7	1	1	0	0	0	0	0	0
6	1	1	1	0	0	0	0	0
5	1	1	1	1	0	0	0	0
4	1	1	1	1	1	0	0	0
3	1	1	1	1	1	1	0	0
2	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1
2	2	1	1	1	1	1	0	0
3	2	1	1	1	1	0	0	0
4	2	1	1	1	0	0	0	0
2	2	2	1	1	1	0	0	0
3	3	1	1	1	0	0	0	0
5	2	1	1	0	0	0	0	0

3	2	2	1	1	0	0	0	0
4	3	1	1	0	0	0	0	0
6	2	1	0	0	0	0	0	0
4	2	2	1	0	0	0	0	0
2	2	2	2	1	0	0	0	0
3	3	2	1	0	0	0	0	0
5	3	1	0	0	0	0	0	0
4	4	1	0	0	0	0	0	0
7	2	0	0	0	0	0	0	0
5	2	2	0	0	0	0	0	0
3	2	2	2	0	0	0	0	0
4	3	2	0	0	0	0	0	0
6	3	0	0	0	0	0	0	0
3	3	3	0	0	0	0	0	0
5	4	0	0	0	0	0	0	0

and the integer partition of 9 into 2 sub-blocks, with permutations, is

$\tilde{\mathbb{Q}}(9,2) =$

AB	E
9	0
8	1
7	2
6	3
5	4

AB	E
0	9
1	8
2	7
3	6
4	5

In red are highlighted all partitions that do not meet the requirements of W, in fact, if there are 7 columns, sub-block E must receive at least 7 children.

The only 3 possible solutions are therefore

AB	E
0	9
1	8
2	7

As we have seen before, for the case 2|7 there is no need to partition the 2 children among the 3 links in AB, but both for 1|8 and 2|7 we have to select if AB takes from L or from R. We have to compute the possible permutation of the partition of 0, 1 and 2 respectively. So here the possible solution:

	AB		E	
	L	R	L	R
1)	0	0	4	5
2)	1	0	3	5
3)	0	1	4	4
4)	2	0	2	5
5)	0	2	4	3
6)	1	1	3	4

case 1) We have 2 ways to assign 9 children among 7 links, that is $Q(9,7)$:

$Q(9,7) =$

	l_1	l_2	l_3	l_4	l_5	l_6	l_7
a)	3	1	1	1	1	1	1
b)	2	2	1	1	1	1	1

case a) One link in the extra column of sub-block E dies both on v_1 and v_2 giving birth to 3 links in the canonical block. 3 can be partitioned, with permutation, into 2, denoted $\tilde{Q}(3,2)$, in this way

	l_1	
	v_1	v_2
i)	3	0
ii)	2	1
iii)	1	2
iv)	0	3

the corresponding likelihoods are

$$\mathrm{lk}_{1,a,i} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_3'(t_1)P_0'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)\left(P_0'(t_1)P_1'(t_2)\right)^5$$

$$\mathrm{lk}_{1,a,ii} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_2'(t_1)P_1'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^2\left(P_0'(t_1)P_1'(t_2)\right)^4$$

$$\mathrm{lk}_{1,a,iii} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_1'(t_1)P_2'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^3\left(P_0'(t_1)P_1'(t_2)\right)^3$$

$$\mathrm{lk}_{1,a,iv} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_0'(t_1)P_3'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^4\left(P_0'(t_1)P_1'(t_2)\right)^2$$

case b) Two links in the extra column of sub-block E die both on v_1 and v_2 giving birth to 2 links each. 2 can be partitioned, with permutations, into 2, $\tilde{Q}(2,2)$ in this way

	l_1		l_2	
	v_1	v_2	v_1	v_2
i)	2	0	2	0
ii)	2	0	1	1
iii)	2	0	0	2

	l_1		l_2	
	v_1	v_2	v_1	v_2
iv)	1	1	2	0
v)	1	1	1	1
vi)	1	1	0	2

	l_1		l_2	
	v_1	v_2	v_1	v_2
vii)	0	2	2	0
viii)	0	2	1	1
ix)	0	2	0	2

note that violet, pink and green coloured rows are identical to each other (permutation of links).

The corresponding likelihoods are

$$\mathrm{lk}_{1,b,i} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_2'(t_1)P_0'(t_2)\right)^2\left(P_0'(t_1)P_1'(t_2)\right)^5$$

$$\mathrm{lk}_{1,b,ii} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_2'(t_1)P_0'(t_2)\right)\left(P_1'(t_1)P_1'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)\left(P_0'(t_1)P_1'(t_2)\right)^4$$

$$\mathrm{lk}_{1,b,iii} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_2'(t_1)P_0'(t_2)\right)\left(P_0'(t_1)P_2'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^2\left(P_0'(t_1)P_1'(t_2)\right)^3$$

$$\mathrm{lk}_{1,b,iv} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_1'(t_1)P_1'(t_2)\right)\left(P_2'(t_1)P_0'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)\left(P_0'(t_1)P_1'(t_2)\right)^4$$

$$\mathrm{lk}_{1,b,v} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_1'(t_1)P_1'(t_2)\right)\left(P_1'(t_1)P_1'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^2\left(P_0'(t_1)P_1'(t_2)\right)^3$$

$$\mathrm{lk}_{1,b,vi} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_1'(t_1)P_1'(t_2)\right)\left(P_0'(t_1)P_2'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^3\left(P_0'(t_1)P_1'(t_2)\right)^2$$

$$\mathrm{lk}_{1,b,vii} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_0'(t_1)P_2'(t_2)\right)\left(P_2'(t_1)P_0'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^2\left(P_0'(t_1)P_1'(t_2)\right)^3$$

$$\mathrm{lk}_{1,b,viii} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_0'(t_1)P_2'(t_2)\right)\left(P_1'(t_1)P_1'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^3\left(P_0'(t_1)P_1'(t_2)\right)^2$$

$$\mathrm{lk}_{1,b,ix} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^3\left(P_0'(t_1)P_2'(t_2)\right)\left(P_0'(t_1)P_2'(t_2)\right)\left(P_1'(t_1)P_0'(t_2)\right)^4\left(P_0'(t_1)P_1'(t_2)\right)$$

from these formulas (and the colours in the table) it 's clear that

$$\mathrm{lk}_{1,b,ii} = \mathrm{lk}_{1,b,iv}$$

$\mathrm{lk}_{1,\mathrm{b,vi}} = \mathrm{lk}_{1,\mathrm{b,viii}}$

$\mathrm{lk}_{1,\mathrm{b,iii}} = \mathrm{lk}_{1,\mathrm{b,vii}}$

moreover, $\mathrm{lk}_{1,a,ii}$ can be rewritten as

$$\mathrm{lk}_{1,a,ii} = P_{10}(t_0)\big(P_1(t_1)P_1(t_2)\big)^3 P_1'(t_1)^2 P_0'(t_1)^5 P_1'(t_2)^5 P_0'(t_2)^2 \big(\lambda\beta(t_1)\big)^2$$

and

$$\mathrm{lk}_{1,b,ii} = \mathrm{lk}_{1,b,iv} = = P_{10}(t_0)\big(P_1(t_1)P_1(t_2)\big)^3 P_1'(t_1)^2 P_0'(t_1)^5 P_1'(t_2)^5 P_0'(t_2)^2 \big(\lambda\beta(t_1)\big)^2$$

and they are also equivalent.

case 2) To assign 8 children among 7 links we have to compute $Q(8,7)$:

$Q(8,7) =$

	l_1	l_2	l_3	l_4	l_5	l_6	l_7
a)	2	1	1	1	1	1	1

One link in the extra column of sub-block E dies both on v_1 and v_2 giving birth to 2 links. The number 2 can be partitioned , with permutations, in 3 ways:

$\tilde{Q}(2,2) =$

	l_1	
	v_1	v_2
i)	2	0
ii)	1	1
iii)	0	2

then the likelihoods are

$$\mathrm{lk}_{2,\mathrm{a,i}} = P_{10}(t_0)\big(P_1(t_1)P_1(t_2)\big)^2 \big(P_2(t_1)P_1(t_2)\big)\big(P_2'(t_1)P_0'(t_2)\big)\big(P_1'(t_1)P_0'(t_2)\big)\big(P_0'(t_1)P_1'(t_2)\big)^5$$

$$\mathrm{lk}_{2,\mathrm{a,ii}} = P_{10}(t_0)\big(P_1(t_1)P_1(t_2)\big)^2 \big(P_2(t_1)P_1(t_2)\big)\big(P_1'(t_1)P_1'(t_2)\big)\big(P_1'(t_1)P_0'(t_2)\big)^2\big(P_0'(t_1)P_1'(t_2)\big)^4$$

$$\mathrm{lk}_{2,\mathrm{a,iii}} = P_{10}(t_0)\big(P_1(t_1)P_1(t_2)\big)^2 \big(P_2(t_1)P_1(t_2)\big)\big(P_0'(t_1)P_2'(t_2)\big)\big(P_1'(t_1)P_0'(t_2)\big)^3\big(P_0'(t_1)P_1'(t_2)\big)^3$$

case 3) Using the previous result we get the likelihood

$$\mathrm{lk}_{3,\mathrm{a,i}} = P_{10}(t_0)\big(P_1(t_1)P_1(t_2)\big)^2 \big(P_1(t_1)P_2(t_2)\big)\big(P_2'(t_1)P_0'(t_2)\big)\big(P_1'(t_1)P_0'(t_2)\big)^2\big(P_0'(t_1)P_1'(t_2)\big)^4$$

$$\mathrm{lk}_{3,\mathrm{a,ii}} = P_{10}(t_0)\big(P_1(t_1)P_1(t_2)\big)^2 \big(P_1(t_1)P_2(t_2)\big)\big(P_1'(t_1)P_1'(t_2)\big)\big(P_1'(t_1)P_0'(t_2)\big)^3\big(P_0'(t_1)P_1'(t_2)\big)^3$$

$$\mathrm{lk}_{3,\mathrm{a,iii}} = P_{10}(t_0)\big(P_1(t_1)P_1(t_2)\big)^2 \big(P_1(t_1)P_2(t_2)\big)\big(P_0'(t_1)P_2'(t_2)\big)\big(P_1'(t_1)P_0'(t_2)\big)^4\big(P_0'(t_1)P_1'(t_2)\big)^2$$

case 4) There is only one way to assign 7 children among 7 links

$Q(7,7) =$ a)

l_1	l_2	l_3	l_4	l_5	l_6	l_7
1	1	1	1	1	1	1

2 links are associated to block AB, computing $Q(2)$ we have 2 ways of assigning the children:

$Q(2) =$ i)

2	0

ii)

1	1

with likelihoods:

$$\mathrm{lk}_{4,a,i} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^2\left(P_3(t_1)P_1(t_2)\right)\left(P'_1(t_1)P'_0(t_2)\right)^2\left(P'_0(t_1)P'_1(t_2)\right)^5$$

$$\mathrm{lk}_{4,a,ii} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)\left(P_2(t_1)P_1(t_2)\right)^2\left(P'_1(t_1)P'_0(t_2)\right)^2\left(P'_0(t_1)P'_1(t_2)\right)^5$$

case 5) In a similar way to the previous one we get

$$\mathrm{lk}_{5,a,i} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)^2\left(P_1(t_1)P_3(t_2)\right)\left(P'_1(t_1)P'_0(t_2)\right)^4\left(P'_0(t_1)P'_1(t_2)\right)^3$$

$$\mathrm{lk}_{5,a,ii} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)\left(P_1(t_1)P_3(t_2)\right)^2\left(P'_1(t_1)P'_0(t_2)\right)^4\left(P'_0(t_1)P'_1(t_2)\right)^3$$

case 6)

$$\mathrm{lk}_{6,a} = P_{10}(t_0)\left(P_1(t_1)P_1(t_2)\right)\left(P_2(t_1)P_1(t_2)\right)\left(P_1(t_1)P_2(t_2)\right)\left(P'_1(t_1)P'_0(t_2)\right)^3\left(P'_0(t_1)P'_1(t_2)\right)^4$$

<u>case 0|1|6</u>

L remains unchanged, while R gets $R^* = R - 1$ and 6 extra columns are added to the block E for the new 6 candidate fathers. We have to assign the 9 links to the sub-block AB, D, E while ensuring at least 6 links to sub-block E. We compute

$\mathbb{Q}(9,3) =$

AB	D	E
9	0	0
8	1	0
7	1	1
6	2	1
5	3	1
4	4	1
7	2	0
5	2	2
4	3	2
6	3	0
3	3	3
5	4	0

and then

$\tilde{\mathbb{Q}}(9,3) =$

AB	D	E
9	0	0
0	9	0
0	0	9
8	1	0
1	8	0
1	0	8
0	8	1
8	0	1
0	1	8
7	1	1
1	7	1

AB	D	E
1	1	7
3	6	0
2	1	6
1	6	2
2	6	1
1	2	6
6	1	2
5	3	1
3	1	5
1	5	3
5	1	3

AB	D	E
3	5	1
1	3	5
4	4	1
4	1	4
1	4	4
7	2	0
2	0	7
0	7	2
0	2	7
2	7	0
7	0	2

AB	D	E
5	2	2
2	2	5
2	5	2
4	3	2
3	4	2
2	3	4
4	2	3
2	4	3
3	2	4
6	3	0
0	6	3

AB	D	E
3	6	0
6	0	3
3	0	6
0	3	6
3	3	3
5	4	0
0	5	4
4	5	0
0	4	5
5	0	4
4	0	5

Highlighted in red are all partitions that do not meet the requirements either of R or W.

The possible solutions are

AB	D	E
0	1	8
1	1	7
2	1	6
1	2	6
0	2	7
0	3	6

By partitioning the possible solutions in v_1 and v_2 we obtain the following table

	AB		D		E	
	L	R	-	R	L	R
1)	0	0	-	1	4	4
2)	1	0	-	1	3	4
3)	0	1	-	1	4	3
4)	2	0	-	1	2	4
5)	0	2	-	1	4	2
6)	1	1	-	1	3	3

	AB		D		E	
	L	R	-	R	L	R
7)	1	0	-	2	3	3
8)	0	1	-	2	4	2
9)	0	0	-	2	4	3
10)	0	0	-	3	4	2

The tables highlight cases that require attention. In purple there are cases where there is at least one link in sub-block E that generates more than 1 link, so first we have to partition in 6 father links and then the partition in v_1 and v_2 should be calculated.

In orange are the cases where 2 and 3 links, respectively, are assigned to sub-block D. Here again we must first calculate the partition in 5 candidate father links and then in v_1 and v_2.

In green are highlighted the cases where more links are assigned to block AB. Also in these cases it must be calculated whether only one father generates 2 children or 2 fathers generate 1 child each.

case 2|2|3

Let's analyse case 2|2|3.

2 random columns in sub-block C now have an ancestor in the their column, 2 columns in sub-block D also have their father in their respective column. 3 new columns have been added in sub-block E. It means that now $L^* = L - 2 = 2$, $R^* = R - 2 = 3$ and $W = 3$. The remaining $L^* + R^* = 5$ characters must be assigned to the sub-blocks AB, C, D, E. We have to compute $\tilde{\mathbb{Q}}(9,4)$. From the table $Q(9)$ we first extract $Q(9,4)$ (cells highlighted in orange)

9	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0
7	1	1	0	0	0	0	0	0
6	1	1	1	0	0	0	0	0
5	1	1	1	1	0	0	0	0
4	1	1	1	1	1	0	0	0
3	1	1	1	1	1	1	0	0
2	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1
2	2	1	1	1	1	1	0	0
3	2	1	1	1	1	0	0	0
4	2	1	1	1	0	0	0	0
2	2	2	1	1	1	0	0	0
3	3	1	1	1	0	0	0	0
5	2	1	1	0	0	0	0	0

3	2	2	1	1	0	0	0	0
4	3	1	1	0	0	0	0	0
6	2	1	0	0	0	0	0	0
4	2	2	1	0	0	0	0	0
2	2	2	2	1	0	0	0	0
3	3	2	1	0	0	0	0	0
5	3	1	0	0	0	0	0	0
4	4	1	0	0	0	0	0	0
7	2	0	0	0	0	0	0	0
5	2	2	0	0	0	0	0	0
3	2	2	2	0	0	0	0	0
4	3	2	0	0	0	0	0	0
6	3	0	0	0	0	0	0	0
3	3	3	0	0	0	0	0	0
5	4	0	0	0	0	0	0	0

from $Q(9,4)$ we get $\tilde{\mathbb{Q}}(9,4)$ which is is a table of size 220×4. Filtering out the rows that do not meet the requirement of L, R, W we obtain

$\tilde{\mathbb{Q}}(9,4) =$

AB	C	D	E
1	2	2	4
1	3	2	3
1	2	3	3
0	2	2	5
2	2	2	3

AB	C	D	E
0	3	2	4
0	2	4	3
0	4	2	3
0	2	3	4
0	3	3	3

and then considering the partition in v_1 and v_2:

	AB		C		D		E	
	L	R	L	-	-	R	L	R
1)	1	0	2	-	-	2	1	3
2)	0	1	2	-	-	2	2	2
3)	1	0	3	-	-	2	0	3
4)	0	1	3	-	-	2	1	2
5)	1	0	2	-	-	3	1	2
6)	0	1	2	-	-	3	2	1
7)	0	0	2	-	-	2	2	3
8)	2	0	2	-	-	2	0	3
9)	0	2	2	-	-	2	2	1
10)	1	1	2	-	-	2	1	2

	AB		C		D		E	
	L	R	L	-	-	R	L	R
11)	0	0	3	-	-	2	1	3
12)	0	0	2	-	-	4	2	1
13)	0	0	4	-	-	2	0	3
14)	0	0	2	-	-	3	2	2
15)	0	0	3	-	-	3	1	2

In orange are highlighted the cells in which 1 or more fathers have more than 1 child. The case 1|3 can only be broken down into 2|1|1 the same for 2|2. Case 2|3 can instead be broken down into 3|1|1 and 2|2|1.

In green and yellow are highlighted the cells where we have to partition between the fathers. If there are 2 fathers and 3 children then the 3 must be divided into 2|1. The number 4 instead can be divided into 3|1 and 2|2.

In pink are marked the solution that require partitions. The 2 children can be divided between 3 fathers into 2|0 and 1|1.

Case I_2 and I_3

$c = 0$

With $c = 0$ we assign all the characters present in C and D to the already known fathers present. We cannot add new fathers therefore C, D, E get 0 and therefore there are only 2 categories to assign the children to, *i.e.* A, B. Since the first column, A is of type I_2 then categories A and B cannot be merged into one single category AB (like for I_1). With $k = 0$ then there is no B sub-block.

We compute the partition of 9 chlidren among the 2 sub-blocks:

$\mathbb{Q}(9,2) =$

9	0
8	1
7	2
6	3
5	4

and

$\tilde{\mathbb{Q}}(9,2) =$

	A	B
1)	9	0
2)	8	1
3)	7	2
4)	6	3
5)	5	4
6)	0	9
7)	1	8
8)	2	7
9)	3	6
10)	4	5

case 1)

The ancestor in the first column takes all 9 links. We must calculate $Q(9,2)$, the possible partition of 9 links into 2 parts that respect $L = 4$ and $R = 5$. There is only 1 solution.

The corresponding likelihood is:

$$\mathrm{lk}_1 = P_3(t_0)\Big(P_1(t_1)P_1(t_2)\Big)^2 P_{1+4}(t_1)P'_{0+5}(t_2)$$

if the first column instead of being type I_2 is type I_3 the likelihood gets

$$\mathrm{lk}_1 = P_3(t_0)\Big(P_1(t_1)P_1(t_2)\Big)^2 P'_{0+4}(t_1)P_{1+5}(t_2)$$

case 2)

The first column takes 8 links, sub-block B takes 1:

	A		B	
	L	R	L	R
i)	3	5	1	0
ii)	4	4	0	1

if the first column is of type I_2 the likelihood is:

$$\mathrm{lk}_{2,i} = P_3(t_0)P_1(t_1)P_1(t_2)P_2(t_1)P_1(t_2)P_{1+3}(t_1)P'_{0+5}(t_2)$$

and

$$\mathrm{lk}_{2,ii} = P_3(t_0)P_1(t_1)P_1(t_2)P_1(t_1)P_2(t_2)P_{1+4}(t_1)P'_{0+4}(t_2)$$

if the first column is of type I_3 the likelihood is:

$$\mathrm{lk}_{2,i} = P_3(t_0)P_1(t_1)P_1(t_2)P_2(t_1)P_1(t_2)P'_{0+3}(t_1)P_{1+5}(t_2)$$

and

$$\mathrm{lk}_{2,ii} = P_3(t_0)P_1(t_1)P_1(t_2)P_1(t_1)P_2(t_2)P_{0+4}(t_1)P_{1+4}(t_2)$$

case 4)

In this case 6 children are assigned to sub-block A and 3 children to sub-block B.

Then we compute $Q(6,2)$:

$Q(6,2) =$

6	0
5	1
4	2
3	3

and

$\tilde{Q}(6,2) =$

A	
L	R
6	0
5	1
4	2
3	3
0	6
1	5
2	4
3	3

in sub-block B are assigned 3 links, so we compute

$\tilde{Q}(3,2) =$

B	
l_1	l_2
3	0
2	1
1	2
0	3

and partitioning between v_1 and v_2 we get

l_1		l_2	
L	R	L	R
3	0	0	0
2	1	0	0
1	2	0	0
0	3	0	0
2	0	1	0
2	0	0	1
1	1	1	0
1	1	0	1
0	2	1	0
0	2	0	1

l_1		l_2	
L	R	L	R
1	0	2	0
0	1	2	0
1	0	1	1
0	1	1	1
1	0	0	2
0	1	0	2
0	0	3	0
0	0	2	1
0	0	1	2
0	0	0	3

The cells highlighted in red are to be deleted because the 2 characters in column B are indistinguishable from each other.

By crossing all combinations related to block A and block B and filtering the solutions in which $L \neq 4$ and $R \neq 5$ we obtain

A		B			
l_1		l_1		l_2	
L	R	L	R	L	R
4	2	0	3	0	0
4	2	0	2	0	1
3	3	1	2	0	0
3	3	1	1	0	1
3	3	0	2	1	0
1	5	3	0	0	0

A		B			
l_1		l_1		l_2	
L	R	L	R	L	R
1	5	2	0	1	0
2	4	2	1	0	0
2	4	2	0	0	1
2	4	1	1	1	0
3	3	1	2	0	0
3	3	1	1	0	1
3	3	0	2	1	0

Let's see the likelihood of one of the lines, for example 3|3|0|2|1|0

for case l_2

$$\mathrm{lk} = P_3(t_0)\left(P_{1+3}(t_1)P'_{0+3}(t_2)\right)\left(P_{1+0}(t_1)P_{1+2}(t_2)\right)\left(P_{1+1}(t_1)P_{1+0}(t_2)\right)$$

for case l_3

$$\mathrm{lk} = P_3(t_0)\left(P'_{0+3}(t_1)P_{1+3}(t_2)\right)\left(P_{1+0}(t_1)P_{1+2}(t_2)\right)\left(P_{1+1}(t_1)P_{1+0}(t_2)\right)$$

$c = 7$

Let's analyse the case with $c = 7$, which means that there are in total 10 candidate parent links to be splitted into 3 sub-block types, *i.e.* C, D, E. Using a previous result we get

$\tilde{\mathbb{Q}}(7,3) =$

C	D	E
7	0	0
0	7	0
0	0	7
6	1	0
1	6	0
6	0	1
0	6	1
0	1	6
1	0	6
5	1	1
1	1	5
1	5	1

C	D	E
4	2	1
2	4	1
2	1	4
1	2	4
1	4	2
4	1	2
3	3	1
3	1	3
1	3	3
5	2	0
5	0	2
2	5	0

C	D	E
0	5	2
2	0	5
0	2	5
3	2	2
2	3	2
2	2	3
4	3	0
0	4	3
0	3	4
3	4	0
4	0	3
3	0	4

Highlighted in red are the configurations that are not possible constrained by $L = 4$ or $R = 5$.

case 2|2|3

2 random columns in sub-block C now have an ancestor in the their column, 2 columns in sub-block D also have their father in their respective column. 3 new columns have been added in sub-block E. It means that now $L^* = L - 2 = 2$, $R^* = R - 2 = 3$ and $W = 3$. The remaining $L^* + R^* = 5$ characters must be assigned to the sub-blocks B, AC, D, E when A is of type I_2 or to the sub-block B, C, AD, E when of type I_3.

The algorithm proceeds with the same logic described above.

Case II

Case A II requires more attention (see Figure 7).

Ω	#
v_0	#/-
v_1	-
v_2	-

II

Figure 7. First column of type II. The internal state in this column can be either a link or a gap.

The reason is that the state in v_0 in this column can be either a link or a gap, so this case requires marginalizing the two possibilities.

- In case there is a character in v_0 the process continues as described for cases $I_{1,2,3}$. If $k = 0$ (there are no completely full columns) we always have at least one link that may have generated the characters in L and R.
- If the internal state in the column in v_0 is a gap and $k = 0$ then we don't have any character in v_0 (in the columns of the canonical block) that could have generated the characters in C and D. So the loop that adds the characters must start from c=1 instead of $c = 0$ up to a total of $L + R$ new parents. Beyond this it makes no sense because it is already contained in $P^\star(t_0, t_1, t_2)$.

First block (immortal link)

The first block is generated by the immortal link. If the first column of the MSA contains a character at the root then the first block is empty, meaning that it contains only an immortal (not observable) character (see Figure 8).

Ω	●
v_0	●
v_1	●
v_2	●

Figure 8: First block generated by the immortal link. In this case it's represented an empty block.

In this case the likelihood gets:

$$\mathrm{lk} = P_1''(t_0)P_1''(t_1)P_1''(t_2)$$

If the first column does not have a character in the root Ω then the first block extends to the first character, the first block will be part of the first "mortal" block (Figure 9).

Ω	●	-	-	-	-	-
v_0	●	-	#	-	-	#
v_1	●	#	-	-	#	-
v_2	●	-	#	#	-	-

Figure 9: First block generated by the immortal link. In this case it's represented a non empty block. The last column, separated from the canonical block, represents the sub-block E.

The "immortal" block is first converted by transforming the immortal links, normally not visible, into visible characters in this way (Figure 10):

Ω	(#)	-	-	-	-	-
v_0	(#)	-	#	-	-	#
v_1	(#)	#	-	-	#	-
v_2	(#)	-	#	#	-	-

Figure 10: First block of Figure 9 where the immortal link is converted into a visible character, represented as (#).

This allows us to use the algorithm described above almost unchanged. The only difference is that you will use the equations related to the immortal link ($P''_N(t)$).

Substitution model

Once we calculate the likelihood component $P(\alpha' | \theta)$ for the insertion and deletion process (α' denotes the structural MSA) we can calculate the likelihood component $P(\alpha | \alpha', \theta)$ due to the substitution process.

For each character present at the root Ω, whose status is σ, we have a factor $\pi(\sigma)$ in the likelihood. The same is valid for each new character born below node v_0 and whose status in v_1 and v_2 is known.

For all the other states that pass through the v_0 state, of which we do not know the state, they require marginalization over all possibilities.

For instance, using the alphabet $\mathscr{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, the probability for a column of this type gets:

Ω	A
v_0	?
v_1	C
v_2	T

$$\mathrm{lk} = \left(e^{Qt_1} \cdot [0,1,0,0]^{\mathrm{T}} \circ e^{Qt_2} \cdot [0,0,1,0]^{\mathrm{T}}\right) \circ e^{Qt_0} \cdot [1,0,0,0]^{\mathrm{T}}$$

where the symbol '$\circ$' denotes the Hadamard product. The matrix-array product has the priority over the Hadamard multiplication.