Reconstructing Contiguous Regions of an Ancestral Genome (Supplement)

Jian Ma, Louxin Zhang, Bernard B. Suh, Brian J. Raney, Richard C. Burhans, W. James Kent, Mathieu Blanchette, David Haussler, Webb Miller

In this supplementary document, we provide a detailed account of the algorithm for reconstructing CARs and of the reconstruction accuracy analysis.

1 The algorithm of inferring CARs

Given information about adjacencies between conserved segments in each modern species, our goal is to infer segment order in the ancestral genome. To get a clean and precise statement of the problem we formalize it using graph theory. The algorithm identifies a most-parsimonious scenario for the history of each individual adjacency, though the whole-genome prediction is not guaranteed to optimize traditional measures like the number of breakpoints. We introduce weights to the graph edges to model the reliability of each adjacency. Finally, we use a greedy heuristic algorithm to find a set of paths in the graph that cover maximum total weights. These paths correspond to contiguous ancestral regions (CARs).

Here, we explain the algorithm using a detailed example.

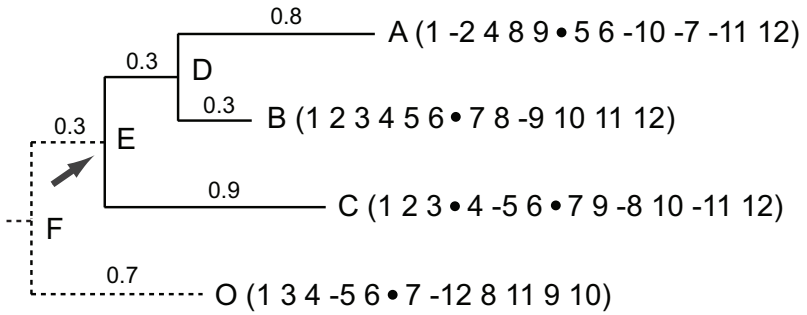

Figure 1: The phylogeny of genomes A, B, C. Our target ancestor is E, and O is the outgroup. The bullet symbol, •, separates chromosomes. Branch lengths are above each branch.

Figure 1 shows the phylogeny of genomes A, B, and C. We want to reconstruct CARs in E with O as an outgroup.

The predecessor graphs of A, B, and C can be obtained directly from the leaf genomes; see Figure 2, 3, and 4. There are two special nodes representing the beginning and the end of a chromosome. The predecessor graphs of internal nodes D and E are as shown in Figure 5 and 6. The predecessor graph of root F is shown in Figure 7. Figure 8 is the result after E being adjusted by F. The corresponding successor graph for E is shown in Figure 9. Then we create the intersection of the predecessor and successor graphs in 8 and 9, giving the graph in Figure 10. Note that in this step we do not intersect edges connecting the beginning or the end of the chromosome.

In Figure 10, there are ambiguous cases for node 7, 8, 9, 10. We then assign weights to edges recursively using the approach discussed in the Method section. For example, $w(7,8) = w(-8,-7) = 0.54$. We have $w_A(7,8) = 0$, $w_B(7,8) = 1$, $w_C(7,8) = 0$, and $w_D(7,8) = \frac{0.8}{0.3+0.8} = 0.72$. So $w_E(7,8) = \frac{0.72\times 0.9}{0.3+0.9} = 0.54$. Note that edges of weight 1 are not shown in the picture.

Then we sort all the edges by weight and add them to the graph until every node in the graph has a unique predecessor and successor. The final edges are indicated by the dark edges in Figure 11. The paths come in pairs, which corresponds to the two orientations of each CAR. We select one path from each pair, obtaining for example CARs (1 2 3 4 -5 6) and (7 8 -9 10 -11 12).

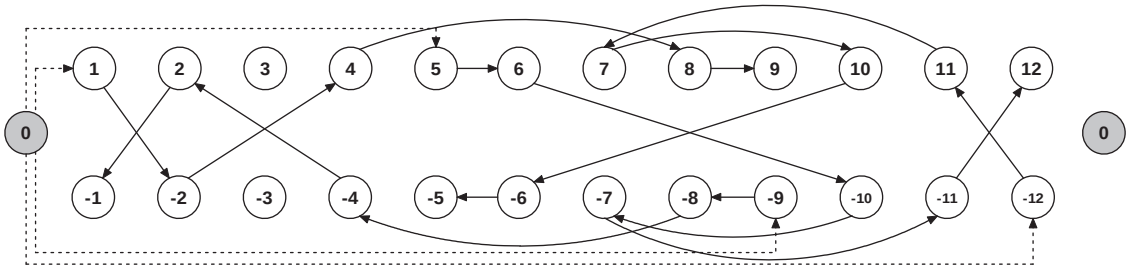

Figure 2: Predecessor graph of A

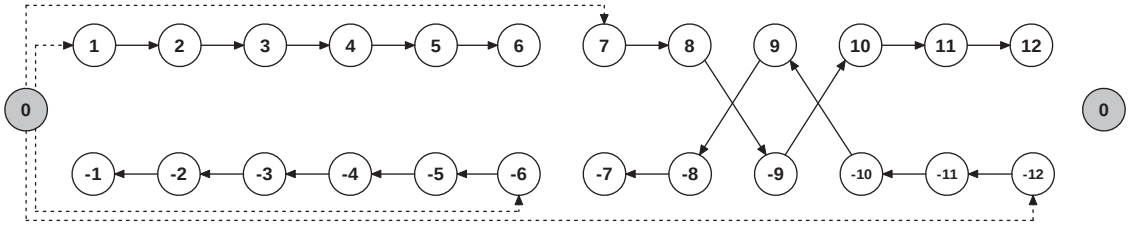

Figure 3: Predecessor graph of B

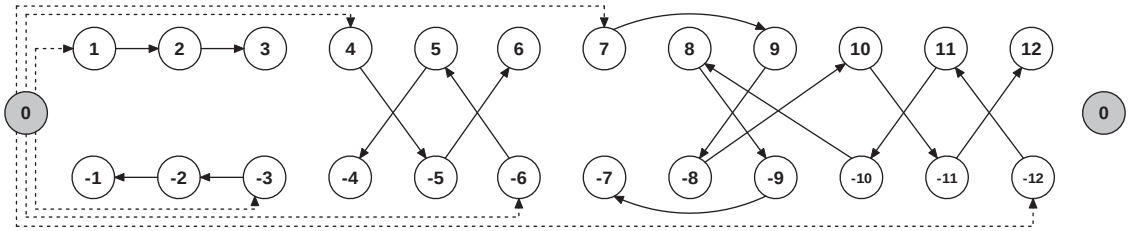

Figure 4: Predecessor graph of C

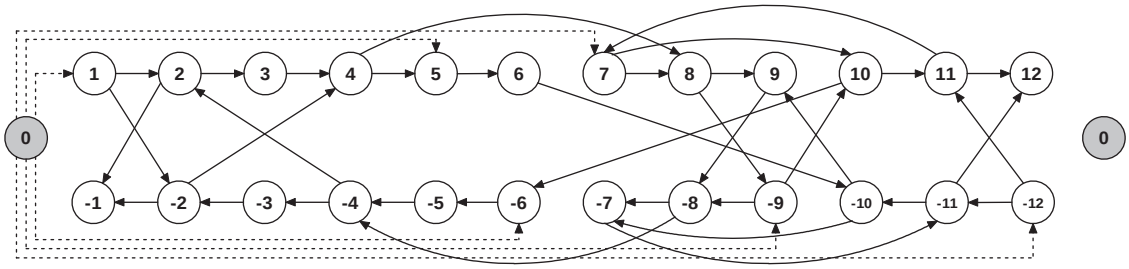

Figure 5: Predecessor graph of D

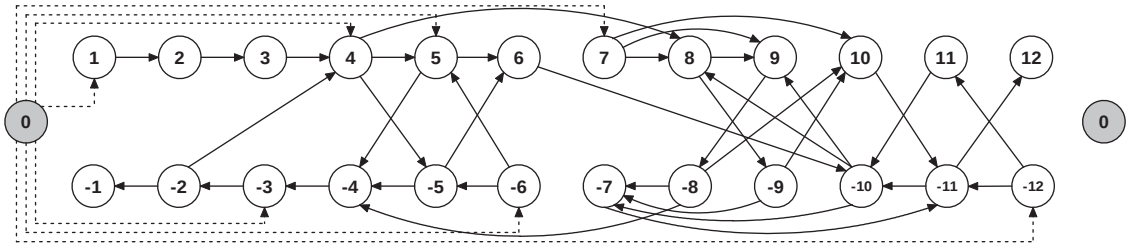

Figure 6: Predecessor graph of E

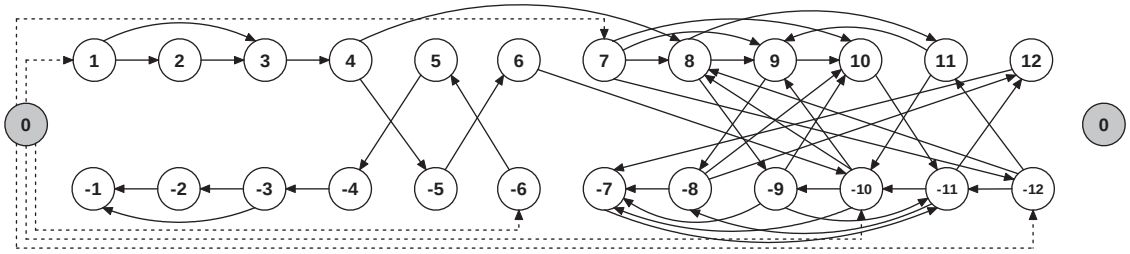

Figure 7: Predecessor graph of F

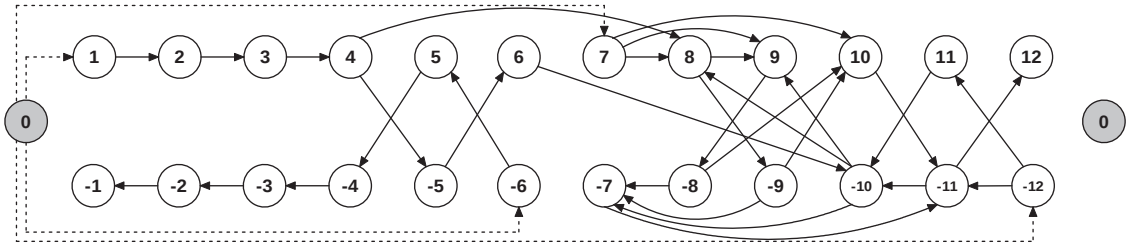

Figure 8: Predecessor graph of E after being adjusted by F

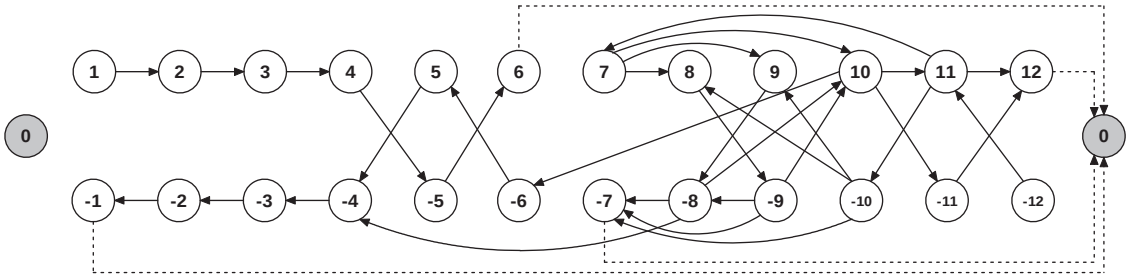

Figure 9: Successor graph of E

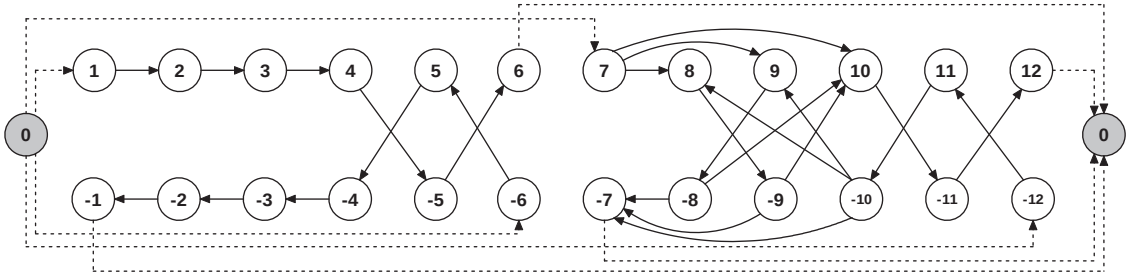

Figure 10: Intersection of the predecessor graph and successor graph of E

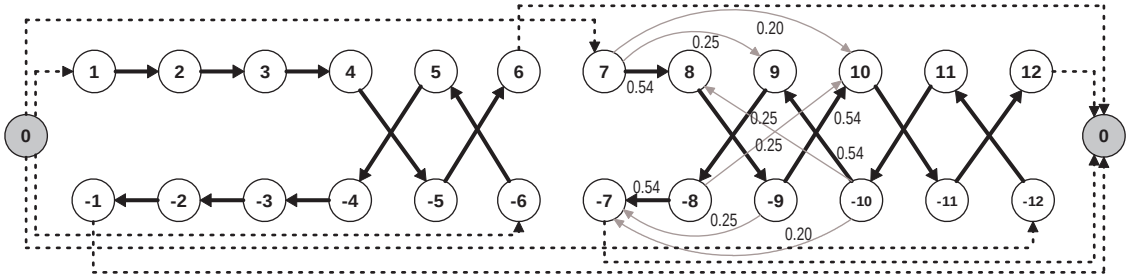

Figure 11: The resulting CARs

2 Probabilistic reconstruction accuracy analysis

Breakpoint distance between multichromosomal genomes

Suppose genomes A and B share n conserved elements, located in p chromosomes, and q chromosomes, respectively. Then, there are a total of $n+p$ adjacencies in A and $n+q$ adjacencies in B. Using ϕ to denote the beginning and end of a chromosome, we assign a score c_k $(k = 1, \ldots, n+p)$ to every adjacency $(a_i a_j)$ in A:

$$c_k = c(a_i, a_j) = \begin{cases} 0 & \text{if } (a_i a_j) \text{ or } (-a_j - a_i) \text{ is in } B; \\ \frac{1}{2} & \text{if } a_i = \phi \text{ or } a_j = \phi \text{ and both } (a_i a_j) \text{ and } (-a_j - a_i) \text{ are not in } B; \\ 1 & \text{otherwise.} \end{cases}$$

Then the breakpoint distance between A and B is defined as:

$$d(A, B) = \sum_{k=1}^{n+p} c_k \tag{1}$$

For example, $A = [1\ 2 \bullet 3\ -4\ 5]$ and $B = [5\ 3 \bullet 1 \bullet 2\ 4]$ (the bullet symbol, $\bullet$, separates chromosomes). In A, we have $c(\phi, 1) = 0$, $c(1, 2) = 1$, $c(2, \phi) = 0.5$, $c(\phi, 3) = 0.5$, $c(3, -4) = 1$, $c(-4, 5) = 1$, $c(5, \phi) = 0.5$, therefore $d(A, B) = 4.5$.

Assume each conserved element i (except ϕ) in the genome g has a predecessor $p_g(i)$ and a successor $s_g(i)$. We set $P(A, B)$ to be the number of i where $p_A(i) \neq p_B(i)$, and $S(A, B)$ to be the number of i where $s_A(i) \neq s_B(i)$. We can see that:

$$P(A, B) + S(A, B) = \frac{1}{2} d(A, B)$$

Since $P(A, B) = P(B, A)$ and $S(A, B) = S(B, A)$, it follows that $d(A, B) = d(B, A)$.

Furthermore, using the breakpoint distance, we can estimate the probability that the successor (or predecessor) of i is different between genome A and B, i.e.

$$Pr[s_A(i) \neq s_B(i)] \approx \frac{d(A, B)}{n} \tag{2}$$

The estimation will be used in the analysis of reconstruction accuracy.

The extended Jukes-Cantor model

We extend the Jukes-Cantor model for analyzing breakpoints. Here we assume that a genome π with n elements has evolved through a series of rearrangement events with unknown proportions. Then, for any element f in the genome $\pi = ...fg...$, its successor g is changed to h over a time unit with the same probability α for all $h \neq f, -f, g$

[Sankoff and Blanchette, 1999]. Hence, there are $2n-3$ such changes possible. The probability that g remains as the successor is $1-(2n-3)\alpha$

Suppose π evolves into τ along a branch with time t. We use $Pr_i[s(f)=g]$ to denote the probability that $s(f)=g$, i.e. g is the successor of f after time i, for $g \neq f, -f$, $i=0,1,\ldots,t$. Then for any i, we have,

$$Pr_{i+1}[s(f)=g] = (1-(2n-3)\alpha)Pr_i[s(f)=g] + \alpha(1-Pr_i[s(f)=g])$$

Equivalently,

$$Pr_{i+1}[s(f)=g] - Pr_i[s(f)=g] = \alpha - \alpha(2n-2)Pr_i[s(f)=g]$$

If we approximate the discrete-time process by a continuous model, we can rewrite the above equation as:

$$\frac{dPr_y[s(f)=g]}{dy} = \alpha - \alpha(2n-2)Pr_y[s(f)=g]$$

We solve the above first-order linear differential equation,

$$Pr_i[s(f)=g] = \frac{1}{2n-2} + \left(Pr_0[s(f)=g] - \frac{1}{2n-2}\right)e^{-(2n-2)\alpha i}$$

Therefore, using $s_\pi(f)=g$ to denote the event that the successor of f is g in π, we have,

$$Pr[s_\tau(f)=g|s_\pi(f)=g] = \frac{1}{2n-2} + \frac{2n-3}{2n-2}e^{-(2n-2)\alpha t},$$

$$\text{since } Pr_0[s(f)=g] = Pr[s_\pi(f)=g] = 1.$$

Similarly, for any $h \neq f, -f, g$ in genome τ,

$$Pr[s_\tau(f)=h|s_\pi(f)=g] = \frac{1}{2n-2} - \frac{1}{2n-2}e^{-(2n-2)\alpha t}$$

Reconstruction accuracy analysis

We reconstruct the CARs in the boreoeutherian ancestor using genomes of human, mouse, rat, dog, opossum, and chicken. The reconstruction is based on the phylogeny shown in Figure 12, in which chicken and opossum are the outgroups. The ancestor genome we want to reconstruct corresponds to E in the phylogeny. This phylogeny is derived from the phylogeny in Figure 8 in the manuscript.

Under the extended Jukes-Cantor model for breakpoints, the probability of correctly reconstructing a join in the boreoeutherian ancestor E is equivalent to the probability

$$\mathcal{P} = Pr[g \text{ is predicted to be the successor of } f | g \text{ is the successor of } f \text{ in } E]$$

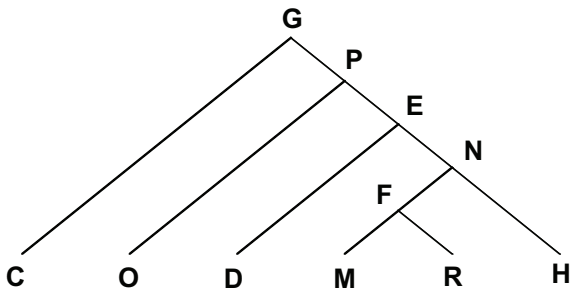

Figure 12: Phylogeny over human (H), mouse (M), rat (R), dog (D), opossum (O), and chicken (C). The branch lengths are t_{GC}=0.453876, t_{GP}=0.190537, t_{PO}=0.365507, t_{PE}=0.271214, t_{ED}=0.206169, t_{EN}=0.023260, t_{NF}=0.260327, t_{FM}=0.072818, t_{FR}=0.081244, t_{NH}=0.140987.

On a branch xy, we set

$$p_{xy} = Pr[s_y(f) = g | s_x(f) = g] = \frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2)t_{xy}\alpha_{xy}} \tag{3}$$

$$q_{xy} = Pr[s_y(f) = g | s_x(f) \neq g] = \frac{1}{2n-2} - \frac{1}{2n-2} e^{-(2n-2)t_{xy}\alpha_{xy}} \tag{4}$$

where t_{xy} denotes the length of branch xy and α_{xy} the successor substitution rate on the branch xy. Note that $(2n-3)q_{xy} = 1 - p_{xy}$.

The algorithm, without the help of an outgroup, will uniquely connect f and g in the output if and only if the following conditions are satisfied:

(U1) g is the successor of f in D; and

(U2) g is the successor of f in H, M, or R. *Restriction*: if the successor of f in H is not g, it should not be identical to the successor of f in both M and R.

The restriction above is to exclude the cases that are caused by parallel neighbor changing. For example, in common ancestor E, $s_E(f) = g$, and in leaf genomes, $s_D(f) = g$, $s_R(f) = h$, $s_M(f) = g$, $s_H(f) = h$, then, the algorithm will not reconstruct g correctly at E.

The probability that **(U1)** holds is

$$p_{ED} = \frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2)t_{ED}\alpha_{ED}}$$

The probability that **(U2)** holds is

$$\begin{aligned}
&Pr[s_H(f) = g \vee s_M(f) = g \vee s_R(f) = g | s_E(f) = g] \\
&\quad - Pr[s_M(f) = g \wedge s_R(f) = s_H(f) \neq g | s_E(f) = g] \\
&\quad - Pr[s_R(f) = g \wedge s_M(f) = s_H(f) \neq g | s_E(f) = g] \\
&\approx \quad 1 - Pr[s_H(f) \neq g \wedge s_M(f) \neq g \wedge s_R(f) \neq g | s_E(f) = g] \\
&= \quad 1 - ((1 - p_{EN})(1 - q_{NH}) (q_{NF}(1 - p_{FM})(1 - p_{FR}) + (1 - q_{NF})(1 - q_{FM})(1 - q_{FR})) \\
&\qquad + p_{EN}(1 - p_{NH}) (p_{NF}(1 - p_{FM})(1 - p_{FR}) + (1 - p_{NF})(1 - q_{FM})(1 - q_{FR})))
\end{aligned}$$

since parallel neighbor changing rarely occurs. In fact, numerical analysis shows that, when n is larger than 1000 as in our case,

$$Pr[s_M(f) = g \wedge s_R(f) = s_H(f) \neq g | s_E(f) = g]$$

and

$$Pr[s_R(f) = g \wedge s_M(f) = s_H(f) \neq g | s_E(f) = g]$$

are about 0.00001.

Overall, the accuracy without the outgroup information is the product of the probability that **(U1)** holds and the probability that **(U2)** holds. If we define

$$P_{\neq g} = Pr[s_H(f) \neq g \wedge s_M(f) \neq g \wedge s_R(f) \neq g | s_E(f) = g] \tag{5}$$

then the overall probability of **(U1)** and **(U2)** can be written as:

$$p_{ED}(1 - P_{\neq g}) \tag{6}$$

For $Z = P, E, N, F$, we use $CS_Z(f)$ to denote the set of the successor candidates constructed at Z by the algorithm. With the outgroups C and O, the algorithm reconstructs g correctly if the following conditions are true:

(X1) The successor of f is g in chicken (C) or opossum (O); and

(X2) The set $CS_E(f)$ of the successor candidates constructed at E is a multiple set containing g.

In the following discussion, we also ignore the probability that parallel changes occur.

The probability for **(X1)** is, since our model is reversible,

$$\begin{aligned}
&Pr[s_C(f) = g \vee s_O(f) = g | s_E(f) = g] \\
&= Pr[s_O(f) = g | s_E(f) = g] + Pr[s_O(f) \neq g \wedge s_C(f) = g | s_E(f) = g] \\
&= Pr[s_O(f) = g | s_E(f) = g] \\
&\quad + Pr[s_O(f) \neq g \wedge s_C(f) = g | s_P(f) = g] Pr[s_P(f) = g | s_E(f) = g] \\
&\quad + Pr[s_O(f) \neq g \wedge s_C(f) = g | s_P(f) \neq g] Pr[s_P(f) \neq g | s_E(f) = g] \\
&= p_{EO} + p_{EP}(1 - p_{PO})p_{PC} + (1 - p_{EP})(1 - q_{PO})q_{PC}.
\end{aligned}$$

Define

$$\begin{aligned}
P_{R=H\neq g} &= Pr[s_M(f) = g \wedge s_R(f) = s_H(f) \neq g | s_E(f) = g], \\
P_{M=H\neq g} &= Pr[s_R(f) = g \wedge s_M(f) = s_H(g) \neq g | s_E(f) = g].
\end{aligned}$$

The probability that **(X2)** holds is:

$$\begin{aligned}
&Pr[s_D(f) = g \wedge g \notin CS_N(f) | s_E(f) = g] + Pr[s_D(f) \neq g \wedge g \in CS_N(f) | s_E(f) = g] \\
&\quad - Pr[s_D(f) \neq g \wedge s_D(f) \in CS_N(f) \wedge g \in CS_N(f) | s_E(f) = g] \\
&\approx Pr[s_D(f) = g \wedge g \notin CS_N(f) | s_E(f) = g] + Pr[s_D(f) \neq g \wedge g \in CS_N(f) | s_E(f) = g] \\
&= p_{ED}(P_{\neq g} + P_{R=H\neq R} + P_{M=H\neq g}) + (1 - p_{ED})(1 - P_{\neq g} - P_{R=H\neq g} - P_{M=H\neq g}) \\
&= 1 - p_{ED} - P_{\neq g} - P_{M=H\neq g} - P_{R=H\neq g} + 2p_{ED}(P_{\neq g} + P_{M=H\neq g} + P_{R=H\neq g}) \\
&\approx 1 - p_{ED} - P_{\neq g} + 2p_{ED}P_{\neq g}
\end{aligned}$$

In the above calculation, we also ignore $P_{R=H\neq g}$ and $P_{M=H\neq g}$ because both of them tend to be extremely small. So the probability that both **(X1)** and **(X2)** are true is:

$$[p_{EO} + p_{EP}(1 - p_{PO})p_{PC} + (1 - p_{EP})(1 - q_{PO})q_{PC}]\,(1 - p_{ED} - P_{\neq g} + 2p_{ED}P_{\neq g}) \quad (7)$$

From (6) and (7), the overall probability of accurately reconstructing a join in the ancestor is:

$$\begin{aligned}
\mathcal{P} \approx\ & p_{ED}(1 - P_{\neq g}) \\
&+ [p_{EO} + p_{EP}(1 - p_{PO})p_{PC} + (1 - p_{EP})(1 - q_{PO})q_{PC}] \\
&\times (1 - p_{ED} - P_{\neq g} + 2p_{ED}P_{\neq g})
\end{aligned} \quad (8)$$

In order to calculate $\mathcal{P}$, we also need to estimate α_{xy} for each branch in the phylogenetic tree. For simplicity, we assume that $\alpha_{FM} = \alpha_{FR}$ and $\alpha_{ED} = \alpha_{EN} = \alpha_{NH} = \alpha_{PE} =$

$\alpha_{PO} = \alpha_{GP} = \alpha_{GC}$. Since

$$\begin{aligned}
Pr[s_M(f) = s_R(f) = g] &= \sum_{h \neq f, -f} Pr[s_M(f) = g | s_F(f) = h] Pr[s_R(f) = g | s_F(f) = h] \\
&= p_{FM} p_{FR} + (2n-3) q_{FM} q_{FR} \\
&= \left(\frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2) t_{FM} \alpha_{FM}} \right) \left(\frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2) t_{FR} \alpha_{FR}} \right) \\
&\quad + (2n-3) \left(\frac{1}{2n-2} - \frac{1}{2n-2} e^{-(2n-2) t_{FM} \alpha_{FM}} \right) \left(\frac{1}{2n-2} - \frac{1}{2n-2} e^{-(2n-2) t_{FR} \alpha_{FR}} \right) \\
&= \frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2)(t_{FM} + t_{FR}) \alpha_{FM}}
\end{aligned}$$

Thus,

$$Pr[s_M(f) \neq s_R(f)] = 1 - Pr[s_M(f) = s_R(f)] = \frac{2n-3}{2n-2} \left(1 - e^{-(2n-2)(t_{FM} + t_{FR}) \alpha_{FM}} \right)$$

This implies

$$\begin{aligned}
\alpha_{FM} &= -\frac{1}{(2n-2)(t_{FM} + t_{FR})} \ln \left(1 - \frac{2n-2}{2n-3} Pr[s_M(f) \neq s_R(f)] \right) \\
&\approx -\frac{1}{(2n-2)(t_{FM} + t_{FR})} \ln \left(1 - \frac{2n-2}{2n-3} \cdot \frac{d(M,R)}{n} \right) \qquad (9)
\end{aligned}$$

Similarly, we have

$$Pr[s_D(f) = s_H(f)] = \frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2)(t_{ED} + t_{EN} + t_{NH}) \alpha_{ED}}$$

and

$$\alpha_{ED} \approx -\frac{1}{(2n-2)(t_{ED} + t_{EN} + t_{NH})} \ln (1 - \frac{2n-2}{2n-3} \cdot \frac{d(D,H)}{n}) \qquad (10)$$

Also,

$$\begin{aligned}
Pr[s_H(f) = s_M(f)] &= \sum_{h \neq f, -f} Pr[s_H(f) = g | s_N(f) = h] Pr[s_M(f) = g | s_N(f) = h] \\
&= p_{NH} p_{NM} + (2n-3) q_{NH} q_{NM} \\
&= \left(\frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2) t_{NH} \alpha_{NH}} \right) \left(\frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2)(t_{NF} \alpha_{NF} + t_{FM} \alpha_{FM})} \right) \\
&\quad + (2n-3) \left(\frac{1}{2n-2} - \frac{1}{2n-2} e^{-(2n-2) t_{NH} \alpha_{NH}} \right) \times \\
&\qquad \left(\frac{1}{2n-2} - \frac{1}{2n-2} e^{-(2n-2)(t_{NF} \alpha_{NF} + t_{FM} \alpha_{FM})} \right) \\
&= \frac{1}{2n-2} + \frac{2n-3}{2n-2} e^{-(2n-2)(t_{NH} \alpha_{NH} + t_{NF} \alpha_{NF} + t_{FM} \alpha_{FM})}
\end{aligned}$$

Hence, we have

$$\alpha_{NF} \approx -\frac{1}{t_{NF}} \left(\frac{1}{2n-2} \ln \left(1 - \frac{2n-2}{2n-3} \cdot \frac{d(H,M)}{n} \right) + t_{NH}\alpha_{NH} + t_{FM}\alpha_{FM} \right) \tag{11}$$

In our reconstruction, we have $n = 1338$ conserved segments. According to (1) we also have $d(M,R) = 727.5$, $d(D,H) = 452.5$, $d(H,M) = 564.5$. Based on (9)(10)(11), we can calculate $\alpha_{FM} = 19.0576 \times 10^{-4}$, $\alpha_{ED} = 4.1693 \times 10^{-4}$, $\alpha_{NF} = 0.2875 \times 10^{-4}$. Using the branch lengths shown in Figure 12 we can calculate p_{xy} and q_{xy} following equations (3) and (4). Finally, the overall probability in (8) is estimated as $\mathcal{P} \approx 0.9018$.

References

[Sankoff and Blanchette, 1999] Sankoff, D. and Blanchette, M. (1999). Probability models for genome rearrangement and linear invariants for phylogenetic inference. In *RECOMB*, pages 302–309.