

# Haplotypes histories as pathways of recombinations

## Proofs of lemmas and proposition

Nadia El-Mabrouk \*      Damian Labuda †

We prove Lemma 2 after Lemmas 1. This is possible as the two lemmas are independent.

### Lemma 2

*Proof:* Consider a canonical pathway of  $p$  recombinations generating  $H$  from the associated haplotype set  $\text{HAP}_k$ .

**I.** If  $p = 1$ , then just one recombination is necessary to generate the unitary haplotype  $H$ , and as  $H$  is different from any of the haplotypes of  $\text{HAP}$ , it is clearly a greedy recombination.

Consider now  $p = 2$ . Then,  $H$  is generated from two recombinations:

$$\begin{aligned} H_1, H_2 &\longrightarrow^{r_1} R \\ R, H_3 &\longrightarrow^{r_2} H \end{aligned}$$

We consider all the possibilities for  $r_1$  and  $r_2$  as depicted in Figure 1.

1. The two recombinations can be replaced by the unique recombination  $r_2$ , that is clearly greedy.
2. As  $H$  is the unitary haplotype,  $r_2$  is greedy. Suppose  $r_1$  is not greedy. Then the strip prefix of  $R$  should be shorter than the strip prefix of  $H_2$  or  $H_1$ . But the strip prefix of  $R$  ends on  $H_2$ . Thus, it cannot be shorter than a strip prefix of  $H_2$ . Suppose it is shorter than the strip prefix of  $H_1$ . Then the two recombinations can be replaced by a unique greedy recombination  $r'_2$ .
3.  $r_2$  is clearly greedy. Suppose  $r_1$  is not greedy. As the strip prefix of  $R$  ends on  $H_1$ , it cannot be shorter than the strip prefix of  $H_1$ . Suppose it is shorter than the strip prefix of  $H_2$ , then the two recombinations can be replaced by the unique recombination  $r'_1$ .
4.  $r_1, r_2$  can be replaced by  $r'_1, r'_2$ .  $r'_2$  is clearly greedy. The strip prefix of  $R$  ends on  $H_2$ , and thus it cannot be shorter than the strip prefix of  $H_2$ . If it is shorter than the strip prefix of  $H_3$ , then the two recombinations can be replaced by a unique greedy recombination on  $H_3, H_1$ .

**II.** Suppose the lemma is satisfied for  $i < p$ . We should prove by induction that it is still true for  $p$ . Let  $r_1, \dots, r_p$  be a recombination pathway. Consider the  $p - 1$  first recombinations, and replace each haplotype  $X$  by  $X' = \neg(X \oplus R_p)$ , in particular  $R'_p$  is unitary. From the induction hypothesis, there exists at most  $p - 1$  greedy recombinations, obtained from the haplotypes  $\{H'_1, H'_2, \dots, H'_p\}$ , that lead to a unitary haplotype. Suppose these greedy recombinations are:

---

\*Département d'informatique et de recherche opérationnelle, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, QC, Canada, H3C 3J7.

†Centre de recherche, Hopital Sainte-Justine, Département de Pédiatrie, Université de Montréal, 3175 Côte-Sainte-Catherine, Montréal, QC, Canada, H3T 1C5.

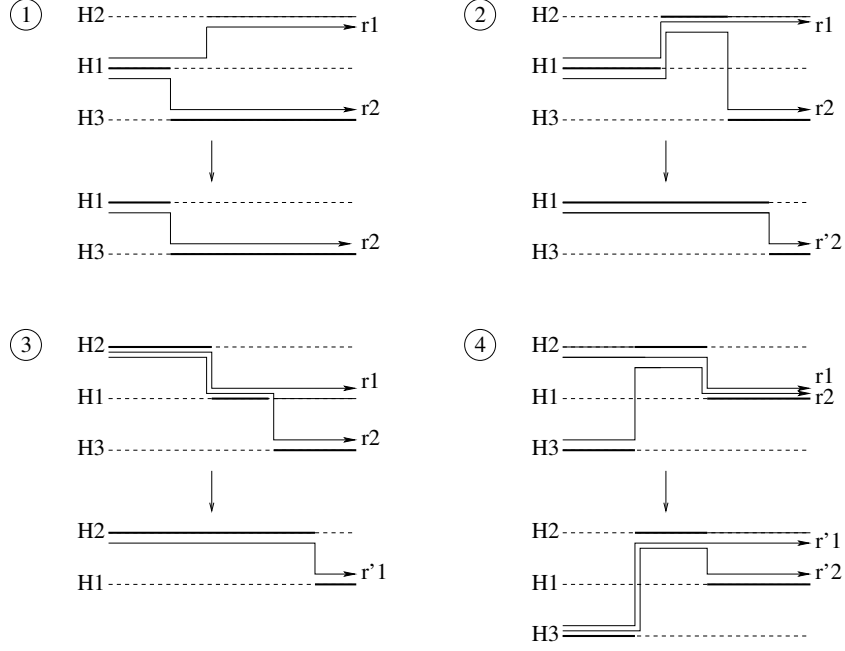


Figure 1: Four different possibilities for the recombinations  $r_1, r_2$ . In each case, we show how to replace  $r_1, r_2$  by greedy recombinations (bottom diagram for each case). Arrows represent recombinations. The recombination  $r_2$  acts on the recombinant resulting from  $r_1$ , and  $H_3$ . Thick lines represent the unitary segments of the haplotypes  $H_1, H_2, H_3$  that are segments of the final resulting haplotype  $H$ .

$$\begin{aligned}
H'_{j_1}, H'_{j_2} &\longrightarrow R'_{l_2} \\
R'_{l_2}, H'_{j_3} &\longrightarrow R'_{l_3} \\
&\vdots \\
&\vdots \\
R'_{l_{p-1}}, H'_{j_p} &\longrightarrow R'_p
\end{aligned}$$

Let us now replace each haplotype  $X'$  by the initial haplotype  $X$  (the one such that  $X' = \neg(X \oplus R_n)$ ):

$$\begin{aligned}
H_{j_1}, H_{j_2} &\xrightarrow{r'_1} R_{l_2} \\
R_{l_2}, H_{j_3} &\xrightarrow{r'_2} R_{l_3} \\
&\vdots \\
&\vdots \\
R_{l_{p-1}}, H_{j_p} &\xrightarrow{r'_{p-1}} R_p
\end{aligned}$$

Each recombination  $r'_i: R_{l_i}, H_{j_{i+1}} \longrightarrow R_{l_{i+1}}$  is such that  $P(R_{l_{i+1}}, R_p) \geq P(R_{l_i}, R_p)$  and  $P(R_{l_{i+1}}, R_p) \geq P(H_{j_i}, R_p)$ , where  $P(X, Y)$  is the size of the longest common prefix of  $X$  and  $Y$ .

Consider now the last recombination:  $r_p: R_p, H_{p+1} \longrightarrow H$ .

1. Suppose  $r_p$  is of the form:  $R_p^5 R_p^T, H_{p+1}^5 H_{p+1}^T \longrightarrow R_p^5 H_{p+1}^T$ .

Suppose there exists a haplotype  $H_k$  such that  $P(H_k, R_p) \geq |R_p^5|$ . Then the recombination pathway  $(r'_1, \dots, r'_p)$  can be replaced by the unique greedy recombination:

$$H_k^5 H_k^T, H_{p+1}^5 H_{p+1}^T \longrightarrow H_k^5 H_{p+1}^T$$

Otherwise, in the sequence  $(r'_1, \dots, r'_{p-1})$  there exists a recombination  $r'_i : R_{l_i}, H_{j_{i+1}} \longrightarrow R_{l_{i+1}}$  such that  $P(R_{l_i}, R_p) \leq |R_p^5|$  and  $P(R_{l_{i+1}}, R_p) > |R_p^5|$ . Then if we define the new recombination:

$$r''_{i+1} : R_{l_{i+1}} = R_{l_{i+1}}^5 R_{l_{i+1}}^T, H_{p+1} = H_{p+1}^5 H_{p+1}^T \longrightarrow R_{l_{i+1}}^5 H_{p+1}^T$$

then  $R_{l_{i+1}}^5 H_{p+1}^T = H$ , and  $(r'_1, r'_2, \dots, r'_i, r''_{i+1})$  is a greedy recombination pathway leading to the unitary haplotype of size  $n$ , that is to  $H$ .

2. Suppose  $r_p$  is of the form:  $R_p^5 R_p^T, H_{p+1}^5 H_{p+1}^T \longrightarrow H_{p+1}^5 R_p^T$ .

In the sequence  $(r'_1, \dots, r'_{p-1})$  there exists a first recombination  $r'_i : R_{l_i}, H_{j_{i+1}} \longrightarrow R_{l_{i+1}}$  such that  $P(R_{l_{i+1}}, R_p) \geq |R_p^5|$ . If we define the new recombination:

$$r''_i : H_{p+1} = H_{p+1}^5 H_{p+1}^T, H_{j_{i+1}} = H_{j_{i+1}}^5 H_{j_{i+1}}^T \longrightarrow H_{p+1}^5 H_{j_{i+1}}^T$$

then  $(r''_i, r'_{i+1}, \dots, r'_{p-1})$  represents a greedy recombination pathway leading to  $H$   $\square$

We are now ready to show that any minimal pathway generating  $R$  from  $\mathcal{C}$  can be reordered into a canonical pathway.

## Lemma 1

*Proof:* A recombination pathway  $\mathcal{R}$  on a subset  $\mathcal{H}_R$  that is not canonical should contain a first recombination of the form  $R_1, R_2 \longrightarrow R$ , where neither  $R_1$  nor  $R_2$  is in HAP. Consider the three last recombinations of  $\mathcal{R}$  leading to  $R$ :

$$\begin{aligned} R_3 &= R_3^5 R_3^T, H_1 = H_1^5 H_1^T \xrightarrow{r_1} R_1 = R_3^5 H_1^T \text{ or } H_1^5 R_3^T \\ R_4 &= R_4^5 R_4^T, H_2 = H_2^5 H_2^T \xrightarrow{r_2} R_2 = R_4^5 H_2^T \text{ or } H_2^5 R_4^T \\ R_1 &= R_1^5 R_1^T, R_2 = R_2^5 R_2^T \xrightarrow{r} R = R_1^5 R_2^T \text{ or } R_2^5 R_1^T \end{aligned}$$

As  $r$  is the first non-canonical recombination,  $R_1$  and  $R_2$  should have been generated by  $r$  recombinations of form  $X, Y \longrightarrow Z$ , where  $X$  and/or  $Y$  is in HAP. Moreover, the recombination pathway  $\mathcal{R} \setminus \{r, r_1, r_2\}$  can be subdivided into two disjoint subsequences of recombinations: the pathway  $\mathcal{R}_3$  leading to  $R_3$  and the pathway  $\mathcal{R}_4$  leading to  $R_4$ . As these are canonical recombination pathways, from Lemma 2, they can be replaced by greedy recombination pathways. Therefore, we can assume that  $\mathcal{R}_3$  and  $\mathcal{R}_4$  are greedy recombination pathways leading to  $R_3$  and  $R_4$ .

We have now to consider each possibility for  $R_1, R_2$  and  $R$ . As all possibilities are treated in the same way, we choose one of them. Suppose, for example, that  $R_1 = R_3^5 H_1^T, R_2 = H_2^5 R_4^T$  and  $R = R_1^5 R_2^T$  (Figure 2.(a)). Consider the new recombination pathway obtained by performing all the recombinations of  $\mathcal{R}_3$  plus the recombination  $r_1$ , and followed by a recombination of form  $R_1, H_2 \longrightarrow R_s$  where  $H_2^T$  is a suffix of  $R_s$ , and the remaining prefix of  $R_s$  being a prefix of  $R$  (Figure 2.(b)). To reach  $R$ , the suffix  $H_2^T$  should be replaced by  $R_4^T$ . Let  $k = |R_4^T|$ .

For each recombination  $X, Y \xrightarrow{r} Z$  of  $\mathcal{R}_4$ , consider the induced recombination  $X', Y' \xrightarrow{r'} Z'$  on the suffixes of size  $k$  of  $X, Y$  and  $Z$ . With this restriction,  $Z'$  may be equal to  $X'$  or  $Y'$ . We remove from  $\mathcal{R}_4$  the recombinations that have an induced recombination satisfying  $Z' = X'$  or  $Z' = Y'$ . Let the new recombination pathway be of form:

$$\begin{aligned} R_f, H_f &\xrightarrow{r^f} R_{f+1} \\ R_{f+1}, H_{f+1} &\xrightarrow{r^{f+1}} R_{f+2} \\ &\vdots \\ &\vdots \\ R_{f+n}, H_{f+n} &\xrightarrow{r^{f+1}} R \end{aligned}$$

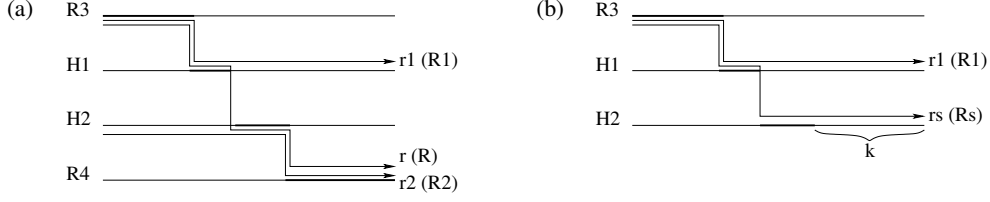


Figure 2: (a) The recombination pathway  $(r_1, r_2, r)$ . Thick lines are the segments of  $R_3, H_1, H_2$  and  $R_4$  that are segments of  $R$ . (b)  $r_1$  followed by  $r_s$ .

Consider the first recombination  $r_f$ . If the prefix of  $R_{f+1}$  is taken from  $H_f$ , then define the new recombination  $R_s, H_f \xrightarrow{r'_f} R'_f = R_s^5 H_f^T$ , where  $|H_f^T| = k$ . If the prefix of  $R_{f+1}$  is taken from  $R_f$ , consider the haplotype  $H'_f$  that has the same suffix of size  $k$  than  $R_f$ . The existence of such a haplotype is deduced from the fact that  $r_f$  is the first recombination of  $\mathcal{R}_4$ . In that case, define the new recombination  $R_s, H'_f \xrightarrow{r'_f} R'_f = R_s^5 H_f'^2$ , where  $|H_f'^T| = k$ .

We then replace the recombination  $r_f$  by  $r'_f$ . Now, we recursively replace each recombination  $r_{f+p}$  by the recombination  $r'_{f+p}$ , by replacing  $R_{f+p}$  by  $R'_{f+p}$ . This recombination pathway clearly leads to  $R$ . Let  $\mathcal{R}'_4 = (r'_f, r'_{f+1}, \dots, r'_{f+p})$ .

To summarize, our new recombination pathway is formed by  $\mathcal{R}_3$ , followed by  $r_1$ , followed by  $\mathcal{R}'_4$ . This is clearly a canonical recombination pathway. And as  $\mathcal{R}'_4$  has the same or fewer recombinations than  $\mathcal{R}_4$ , this new recombination pathway has at most the same number of recombinations than the original one  $\square$

## Proposition 1

*Proof:* “ $\Rightarrow$ ” We first prove that any solution is necessarily a haplotype table associated to a path of  $G(V, E)$ . Let  $(H_1, H_2, \dots, H_{min}, H_{min+1})$  be a solution. Then, there is a perfectly greedy recombination pathway of form:

$$\begin{aligned}
H_1^5 H_1^3, H_2^5 H_2^3 &\xrightarrow{r_1} R_2 = H_1^5 H_2^3 \\
R_2 = R_2^5 R_2^3, H_3^5 H_3^3 &\xrightarrow{r_2} R_3 = R_2^5 H_3^3 \\
&\vdots \\
&\vdots \\
R_{min-1} = R_{min-1}^5 R_{min-1}^3, H_{min}^5 H_{min}^3 &\xrightarrow{r_{min-1}} R_{min} = R_{min}^5 H_{min}^3 \\
R_{min} = R_{min}^5 R_{min}^3, H_{min+1}^5 H_{min+1}^3 &\xrightarrow{r_{min}} H = R_{min-1}^5 H_{min}^3
\end{aligned}$$

where for each  $i$ ,  $|R_i^5| = P(R_i)$ , and  $P(H_i^3) > 0$ . Consider the sequence of vertices  $(0, v_1, v_2, \dots, v_{min}, v_{min+1})$  such that, for any  $i$ ,  $v_i = P(H_i^3)$ . Then, it is easy to verify that, for any  $i$ ,  $v_{i-1} \rightarrow v_i$  is an edge of  $E$ , and thus,  $(0, v_1, v_2, \dots, v_{min}, v_{min+1})$  is a path of  $G(V, E)$ .

“ $\Leftarrow$ ” Let  $(0, v_1, v_2, \dots, v_p, v_{p+1} = n)$  be a path of  $G(V, E)$  and let  $(H_1, H_2, \dots, H_p, H_{p+1})$  be an associated haplotype table leading to  $R_{p+1}$ . It is evident, from the graph construction, that  $R_{p+1}$  is the unitary haplotype of size  $n$ . It remains to prove that we have a minimal recombination pathway. From the breadth-first construction of the graph, any path of  $G(V, E)$  has the same length. On the other hand, from the first part of the proof, any minimal recombination pathway leading to  $H$  is associated to a path of the graph. Therefore, any haplotype table associated to a path of  $G(V, E)$  is a solution  $\square$