On the Minimum Number of Recombination Events in the Evolutionary History of DNA Sequences

Yun S. Song[a,b,1] and Jotun Hein[a]

[a]*Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK*

[b]*Mathematical Institute, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK*

Abstract

In representing the evolutionary history of a set of binary DNA sequences by a connected graph, a set theoretical approach is introduced for studying recombination events. We show that set theoretical constraints have direct implications on the number of recombination events. We define a new lower bound on the number of recombination events and demonstrate the usefulness of our new approach through several explicit examples.

1 Introduction

DNA sequence data, much of which is for studying populational variations, is presently accumulating at an increasing rate. It is of interest to investigate the evolutionary history of sampled sequences, but such investigation is complicated by the fact that DNA sequences from a species have often been subjected to recombination events. An important problem in this regard is to determine how many recombination events must have occurred in the evolutionary history of the sampled sequences. In [HK] Hudson and Kaplan have constructed an algorithm which gives a lower bound on the number of recombination events. The primary goal of this paper is to construct an improved lower bound.

Let $S = \{s_\alpha\}$ be a set of binary DNA sequences

$$s_\alpha = c_1^\alpha, c_2^\alpha, \ldots, c_h^\alpha \quad ,$$

where $c_i^\alpha \in \{0, 1\}$ for every $\alpha \in \{\underline{1}, \underline{2}, \ldots, \underline{n}\}$ and $i \in I_S := \{1, 2, \ldots, h\}$. Note that each sequence is of fixed length h. c_i^α is called the i^{th} *character* of sequence s_α, and the index i denotes the character *site*. A mutation event at the i^{th} character site of a sequence, say $s = c_1, c_2, \ldots, c_h$, is a map $\mathbf{m}_i : \{0,1\}^h \to \{0,1\}^h$ which sends s to $\mathbf{m}_i(s) = c_1, c_2, \ldots, c_{i-1}, \bar{c}_i, c_{i+1}, \ldots, c_h$, where $\bar{c}_i = 1$ if $c_i = 0$ and $\bar{c}_i = 0$ if $c_i = 1$. In other words, $\mathbf{m}_i$ changes the i^{th} character of the binary sequence which undergoes mutation. Throughout this paper we make the following assumption:

[1]Corresponding author. E-mail: song@stats.ox.ac.uk . Preferred postal address: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.

Assumption 1.1 *There are no reverse or recurrent mutations.*

Under this working assumption Hudson and Kaplan have constructed an algorithm for determining a lower bound $\mathcal{R}_{\text{HK}}(S)$ on the number of recombination events necessary to represent S by a connected graph [HK]. In this paper, the minimum number of recombination events is denoted by $\mathcal{R}_{\min}(S)$ and is defined by the property that there exists no graphical representation of S with less than $\mathcal{R}_{\min}(S)$ recombination events. The term "graph" refers to a finite graph which may have cycles but no loops. In our way of drawing graphs, time runs from top to bottom. To any point on the graph except for a finite set Γ of points, there corresponds a unique DNA sequence of length h. Every point in Γ corresponds to either a mutation event or a recombination event. More precisely, every point-like mutation event occurs on an edge of the graph, and to every recombination event, there corresponds a unique trivalent vertex which belongs to a cycle. Let v be a trivalent vertex where a recombination event occurs. We call such a vertex a *recombination vertex*. Let e_1, e_2, e_3 be the edges incident with v and let s_1, s_2, s_3 be their corresponding DNA sequences in the immediate neighbourhood of v where neither a mutation event nor any other recombination event occurs. Then, there exists a cycle C such that e_1 and e_2 belong to C, and s_3 is the by-product of a recombination event between s_1 and s_2. In fact, every cycle in the graph contains at least one such recombination vertex. Lastly, we remark that for each recombination event the position of the break-point in the involved DNA sequences must be specified. The reader should refer to Figure 1 and the explanation therein for a specific example.

Although Hudson and Kaplan's estimate $\mathcal{R}_{\text{HK}}(S)$ certainly is a lower bound, in many cases of S, $\mathcal{R}_{\text{HK}}(S)$ is in fact less than the minimum number $\mathcal{R}_{\min}(S)$ of necessary recombination events. In this paper we improve upon the ideas in [HK] to construct a sharper lower bound on the number of recombination events. In so doing, we show why Hudson and Kaplan's algorithm does not give the minimum $\mathcal{R}_{\min}(S)$ for a general data set S. An important observation is that the total number of recombination events depends on which sequences have undergone recombination. Hudson and Kaplan's algorithm, however, does not take that point into account, thus often leading to an underestimation of the minimum number of recombination events. As we illustrate in this paper, the gist of our new approach is that, in order to obtain a lower bound which is sharper than $\mathcal{R}_{\text{HK}}(S)$, we need to choose judiciously a proper subset of S for each pair of incompatible character sites. More precisely, the proper subset must be chosen from what we call "the quadpartition" of S associated to each pair of incompatible character sites.

The organisation of this paper is as follows. In §2, we establish some set theoretical results and discuss their implications on the number of recombination events. Our new lower bound $\mathcal{R}_{\text{Q}}(S)$ is defined in §3, where we also provide an algorithm for simplifying the determination of $\mathcal{R}_{\text{Q}}(S)$. In the subsequent section we discuss applications of our result to some specific examples, including Kreitman's 1983 data of the alcohol dehydrogenase locus from 11 chromosomes of *Drosophila melanogaster* [K]. We conclude with some remarks in

§5, and for the ease of reference, rephrase in Appendix A Hudson and Kaplan's algorithm for determining $\mathcal{R}_{\mathrm{HK}}(S)$. Proofs of the lemmas from §2 are provided in Appendix B.

NOTATIONS: We here summarise our notations to be used throughout the paper:

S	A set $\{s_\alpha\}$ of binary DNA sequences of fixed length h.
2^S	The set of all subsets of S.
$X \setminus A$	The complement of A relative to X, where A is a subset of the set X.
c_i^α	The i^{th} character of sequence s_α.
$\mathbf{m}_i$	A mutation event at the i^{th} character site.
(i, j)	A pair of character sites.
$I(i, j)$	The open interval $\{x \in \mathbb{R} \mid i < x < j\}$.
$\{B_i, B_i^c\}$	The bipartition of S associated to the character site i.
$\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$	The quadpartition of S associated to the pair (i, j) of incompatible character sites. See Definition 2.2
$R^{i,j}$	The *recombining subset* corresponding to the pair (i, j) of incompatible character sites. See Definition 2.3.
$\mathscr{P}(S)$	For $1 \leq i < j \leq h$, the set of all pairs (i, j) of incompatible character sites in S.
$\mathscr{R}(S)$	The set of $R^{i,j}$, where $(i, j) \in \mathscr{P}(S)$
$\mathcal{A}$	An assignment. See §3.1.
$\omega(X; \mathcal{A})$	The weight of $X \in 2^S$ in the assignment $\mathcal{A}$. See §3.1.
$\mathfrak{A}$	The set of all assignments which satisfy Properties (P1) and (P2) of in §3.1.
$\mathcal{R}_{\min}(S)$	The minimum number of recombination events necessary for representing S by a connected graph.
$\mathcal{R}_{\mathrm{HK}}(S)$	Hudson and Kaplan's lower bound on the number of recombination events.
$\mathcal{R}_{\mathrm{Q}}(S)$	The new lower bound defined in the present paper.

2 Set Theoretical Results

To each character site i, there corresponds a bipartition, also known as a split, of S into disjoint subsets B_i and B_i^c, where ^{c} denotes complement relative to S, such that s_α and s_β belong to the same subset if and only if $c_i^\alpha = c_i^\beta$.

Definition 2.1 (Informative Character Site)
The i^{th} character site is called non-informative *if*

(a) $c_i^\alpha = 0, \forall\, \alpha$,
(b) $c_i^\alpha = 1, \forall\, \alpha$,
(c) $c_i^\alpha = 0$ *for exactly one value of* α,
or (d) $c_i^\alpha = 1$ *for exactly one value of* α.

In other words, if the i^{th} character site is non-informative, then either B_i or B_i^c has cardinality ≤ 1. If a character site is not non-informative, it is called informative.

Non-informative character sites are of little importance in our analysis, since no additional recombination event is needed to explain such a site. Clearly, if the i^{th} character site is informative, then the corresponding bipartition $\{B_i, B_i^c\}$ consists of proper subsets of S.

Definition 2.2 (Incompatibility & Quadpartition)
A pair (i,j) of character sites is called compatible *if at least one of the following intersections is empty:*

$$B_i \cap B_j \quad , \; B_i \cap B_j^c \quad , \; B_i^c \cap B_j \quad , \; B_i^c \cap B_j^c \; .$$

If none of the above intersections is empty, then the pair (i,j) is called incompatible, *and there exists a corresponding* quadpartition

$$\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\} = \{B_i \cap B_j \,, B_i \cap B_j^c \,, B_i^c \cap B_j \,, B_i^c \cap B_j^c\} \tag{2.1}$$

of S into four pair-wise disjoint proper subsets.

To each pair of incompatible character sites i and j, where $i < j$, we assign an open interval of the form

$$I(i,j) := \{x \in \mathbb{R} \,|\, i < x < j\}.$$

As discussed in [HK], Assumption 1.1 implies that recombination events are necessary to represent incompatible characters sites by a connected graph. For example, consider the data set on the right where we denote each sequence s_α by its index value. There are four distinct sequences, each of length 2. But, there are only two character sites and hence only two possible mutation events. Therefore, at least one recombination event is necessary to represent S graphically. A possible graphical representation of S is given in Figure 1. Notice that all four possible combinations—0 0, 0 1, 1 0, and 1 1—of characters are present. Looking for the presence of all four such combinations is commonly known as the "four gamete" test of incompatibility.

$$S = \{\underline{1}, \underline{2}, \underline{3}, \underline{4}\}$$
$$\underline{1} = 1\;0$$
$$\underline{2} = 1\;1$$
$$\underline{3} = 0\;1$$
$$\underline{4} = 0\;0$$

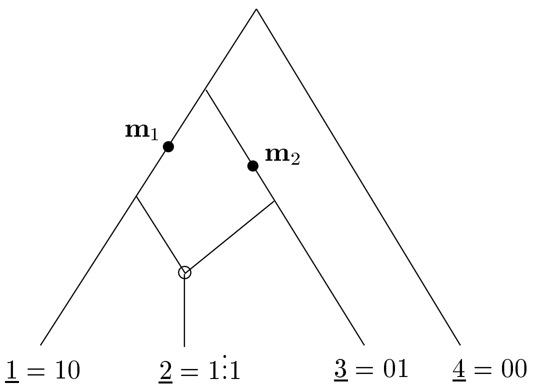

Figure 1: $\mathbf{m}_i$ denotes a mutation event at the i^{th} character site, and the symbol $\bullet$ the location of the mutation event. Throughout this paper, recombination vertices are denoted by the symbol $\circ$. The symbol $\vdots$ in a sequence is used to indicate that the sequence is obtained by a recombination event, in which the part of the sequence to the left (right) of the symbol comes from the left (right) edge incident on the recombination vertex. The length of each edge has been chosen arbitrarily and does not carry any significance. Time runs from top to bottom.

For the convenience of discussion, we introduce the following definition:

Definition 2.3 (Recombining Subset) *Let (i,j) be a pair of incompatible character sites. A recombining subset $R^{i,j} = Q_a^{ij} \in \{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$, associated to the incompatible pair (i,j), is defined as the set of sequences in S whose character sites i and j are supposed to be explained by recombination.*

For each pair of incompatible sites, we need to make a choice of recombining subset. An important observation is that the total number of recombination events depends on the choice of the set

$$\mathscr{R}(S) := \left\{ R^{i,j} \in \{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\} \,\middle|\, i < j\,, (i,j) \text{ an incompatible pair in } S \right\}$$

of recombining subsets. Moreover, there exists a preferred choice of $\mathscr{R}(S)$ which leads to the minimum number $\mathcal{R}_{\min}(S)$ of recombination events. To make this fact more transparent, we provide the following two examples, which well illustrate the dependence of the number of recombination events on $\mathscr{R}(S)$.

(**Remark**: Henceforward, when we discuss examples, we adapt the following conventions: For every incompatible pair (i,j), $Q(ab)$ denotes the proper subset of S whose elements are the sequences s_α with $c_i^\alpha = a$ and $c_j^\alpha = b$. To further simplify the notation, we denote the sequence s_α, $\alpha \in \{\underline{1}, \underline{2}, \ldots, \underline{n}\}$, by the value of the index α.)

Example 2.1 Let S be a set of binary DNA sequences of length h and assume that we have constructed a graph $\mathcal{G}$ which represents S with the minimum number of recombination events. Now, consider adding a new character site to the data, so that the length of each sequence becomes $h+1$. Moreover, assume that the newly added character site is compatible with the original h character sites. Let S' denote the newly constructed data set.

Let $\mathbf{m}_{h+1}$ denote a mutation event corresponding to the new character site. In contrast to common expectation, one might not be able to modify the graph $\mathcal{G}$ by a single mutation $\mathbf{m}_{h+1}$ alone to produce a graph which represents the new data set S'. In fact, if one insists on using $\mathcal{G}$ to represent the original data S, one might have to introduce an additional recombination event to $\mathcal{G}$ in order to represent S'. This does not necessarily mean, however, that even a compatible character site can affect the the minimum number of recombination events. Rather, the point is that one needs to choose $\mathscr{R}(S')$ judiciously to achieve the minimum. A concrete example would be timely.

Consider the set $S = \{\underline{1}, \underline{2}, \ldots, \underline{6}\}$, where the sequences $\underline{1}, \underline{2}, \ldots, \underline{6}$ are shown below:

$\underline{1} = 1\ 1\ 0\ 1$
$\underline{2} = 1\ 1\ 0\ 0$
$\underline{3} = 1\ 1\ 1\ 0$
$\underline{4} = 0\ 1\ 1\ 1$
$\underline{5} = 1\ 0\ 1\ 1$
$\underline{6} = 0\ 0\ 1\ 1$

Incompatible Pairs	$Q(00)$	$Q(01)$	$Q(10)$	$Q(11)$
$(1,2)$	$\{\underline{6}\}$	$\{\underline{4}\}$	$\{\underline{5}\}$	$\{\underline{1},\underline{2},\underline{3}\}$
$(3,4)$	$\{\underline{2}\}$	$\{\underline{1}\}$	$\{\underline{3}\}$	$\{\underline{4},\underline{5},\underline{6}\}$

In the table we have listed the pairs of incompatible character sites and their corresponding quadpartitions. It is clear that $\mathcal{R}_{\min}(S) = 2$ and there are several ways to achieve this minimum. Let us choose the recombining subsets to be $R^{1,2} = \{\underline{4}\}$ and $R^{3,4} = \{\underline{1}\}$, for which case a graphical representation of S is shown in Figure 2a.

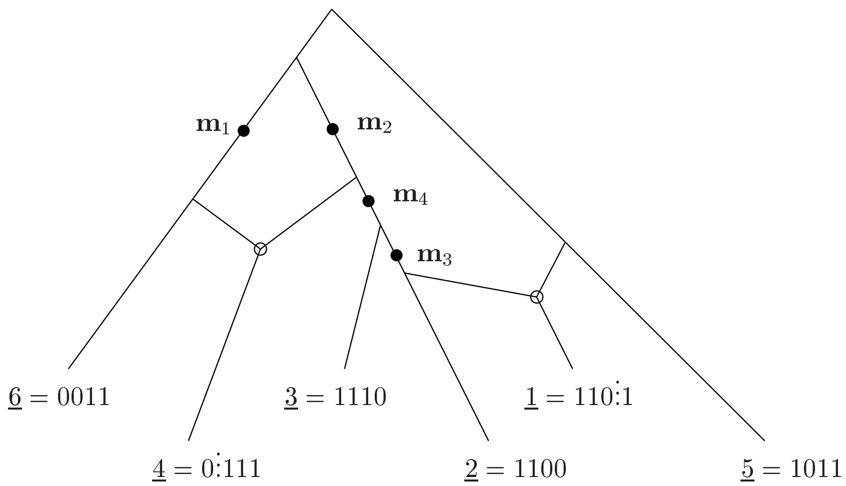

Figure 2a: A graphical representation of S in Example 2.1 with $R^{1,2} = \{\underline{4}\}$ and $R^{3,4} = \{\underline{1}\}$. The same notations as in Figure 1 are used here.

Now construct a new data set S' by introducing a fifth character site to S as follows:

$$c_5^{\underline{1}} = 0 \ , \ c_5^{\underline{2}} = 0 \ , \ c_5^{\underline{3}} = 0 \ , \ c_5^{\underline{4}} = 1 \ , \ c_5^{\underline{5}} = 1 \ , \ c_5^{\underline{6}} = 1 \ . \tag{2.2}$$

Note that this new character site is compatible with every one of the first four sites. But, since the part to the right of the symbol $\vdots$ in sequence $\underline{1}$ comes from sequence $\underline{5}$, it is clear that the new data set S' cannot be represented by simply modifying Figure 2a with an additional mutation $\mathbf{m}_5$. In fact, an additional recombination event, as well as $\mathbf{m}_5$, must be introduced to Figure 2a to represent S'. The same conclusion would follow if the sequence $\underline{1}$ were obtained from $\underline{2}$ and $\underline{4}$ or from $\underline{2}$ and $\underline{6}$, instead of from $\underline{2}$ and $\underline{5}$. If we had chosen the recombining subsets of S to be $R^{1,2} = \{\underline{4}\}$ and $R^{3,4} = \{\underline{2}\}$, however, we would not need an additional recombination event. This fact is illustrated in Figure 2b, where $\mathbf{m}_5$ alone is enough to yield a graph which represents the fifth character site as well.

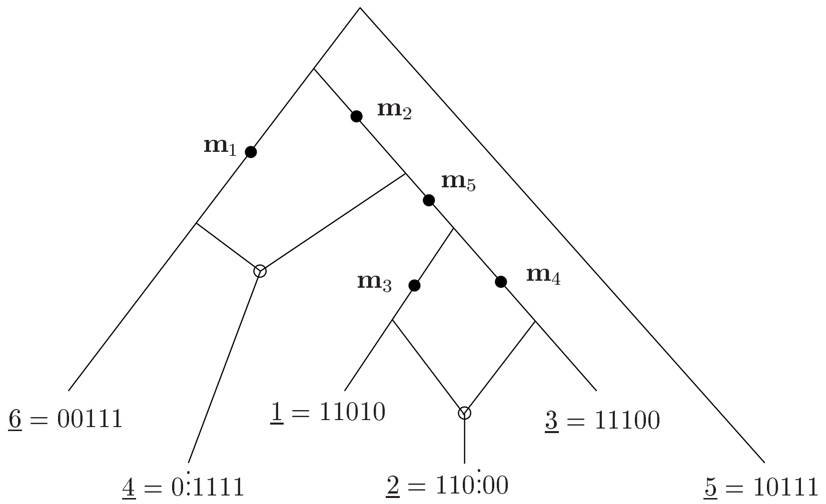

Figure 2b: A graphical representation of S' in Example 2.1 with $R^{1,2} = \{\underline{4}\}$ and $R^{3,4} = \{\underline{2}\}$.

Example 2.2 Consider the data set $S = \{\underline{1}, \underline{2}, \underline{3}, \underline{4}, \underline{5}\}$, where the sequences and the quad-partitions are as follows.

$\underline{1} = 1\ 1\ 0$
$\underline{2} = 0\ 1\ 0$
$\underline{3} = 0\ 1\ 1$
$\underline{4} = 1\ 0\ 1$
$\underline{5} = 0\ 0\ 1$

Incompatible Pairs	$Q(00)$	$Q(01)$	$Q(10)$	$Q(11)$
$(1, 2)$	$\{\underline{5}\}$	$\{\underline{2}, \underline{3}\}$	$\{\underline{4}\}$	$\{\underline{1}\}$
$(1, 3)$	$\{\underline{2}\}$	$\{\underline{3}, \underline{5}\}$	$\{\underline{1}\}$	$\{\underline{4}\}$

In this example, the minimum number $\mathcal{R}_{\min}(S)$ of recombination event is 1, which is obtainable if and only if we choose either $R^{1,2} = R^{1,3} = \{\underline{1}\}$ or $R^{1,2} = R^{1,3} = \{\underline{4}\}$. Furthermore, the location of that single recombination event must be between the first and

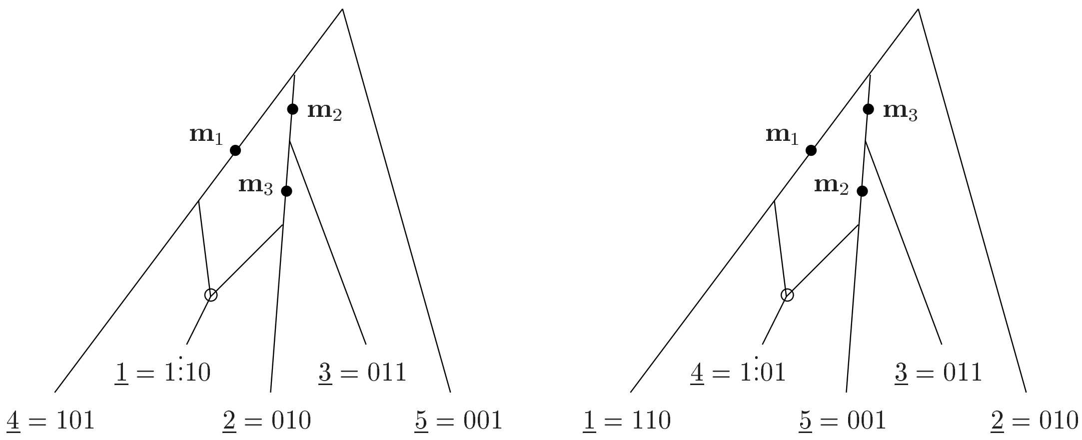

Figure 3: Graphical representations of S in Example 2.2 with the minimum number of recombination event.

the second character sites. For other choices of $R^{1,2}$ and $R^{1,3}$, we need more than one recombination event to represent the data graphically. Two graphical representations of S each involving only one recombination event are shown in Figure 3.

In contrast to the above case, the following situation requires at least two recombination events:

$\underline{1} = 1\ 1\ 0$
$\underline{2} = 0\ 1\ 0$
$\underline{3} = 0\ 1\ 1$
$\underline{4} = 1\ 0\ 1$
$\underline{5} = 0\ 0\ 1$
$\underline{6} = 1\ 1\ 1$

$S' = \{\underline{1}, \underline{2}, \underline{3}, \underline{4}, \underline{5}, \underline{6}\}$

Incompatible Pairs	$Q(00)$	$Q(01)$	$Q(10)$	$Q(11)$
$(1,2)$	$\{\underline{5}\}$	$\{\underline{2},\underline{3}\}$	$\{\underline{4}\}$	$\{\underline{1},\underline{6}\}$
$(1,3)$	$\{\underline{2}\}$	$\{\underline{3},\underline{5}\}$	$\{\underline{1}\}$	$\{\underline{4},\underline{6}\}$

Notice that S and S' have the same form of incompatibilities; that is, incompatible pairs are $(1,2)$ and $(1,3)$ in both cases. But, whereas in the case of S we can choose $R^{1,2} = R^{1,3}$, that is not possible in the case of S', for the quadpartitions $\{Q_a^{1,2}\}$ and $\{Q_a^{1,3}\}$ have no matching elements.

As mentioned before, the most important lesson to be learned from the above examples is that the number of recombination events depends on which sequences undergo recombination for which pair of incompatible character sites. For instance, if there are two pairs (i,j), (k,l) of incompatible character sites[2], with $I(i,j) \cap I(k,l) \neq \varnothing$, then it might be possible to account for both incompatibilities using only one recombination event. A *necessary*

[2]Here, i, j, k, l may not all be distinct.

condition[3] for such a case is that there exists a common choice of recombining subset, i.e $R^{i,j} = R^{k,l}$. As we show presently, however, this is not always possible.

In the next section, we define a new lower bound motivated by our observation, focusing on which sequences undergo recombination and where in the sequences recombination events occur. As for now, we turn to a rigorous reason why Hudson and Kaplan's algorithm fails to give the minimum $\mathcal{R}_{\min}(S)$ for a general data set S. The reader unfamiliar with Hudson and Kaplan's algorithm may refer to Appendix A of the present paper.

Proposition 2.1 *Let ρ denote the minimum number of recombination events necessary to represent graphically the character sites i_1, i_2, j_1, j_2, where $i_1 < j_1$ and $i_2 < j_2$. If (i_1, j_1) and (i_2, j_2) are incompatible pairs, whereas $(i_1, i_2), (i_1, j_2), (j_1, i_2)$ and (j_1, j_2) are compatible pairs, then $\rho > 1$.*

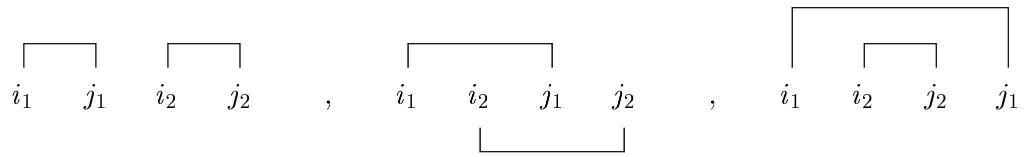

Figure 4: A schematic representation of the incompatibilities considered in Proposition 2.1. There are three inequivalent situations and, in each case, incompatibility between a pair of character sites is represented by a connecting line.

As shown in Figure 4, there are three possible situations. To prove the above Proposition, we need to establish several facts and we have decomposed them into the following set of successive lemmas:

Lemmas 2.1 & 2.2 together imply Lemma 2.3, which in turn implies Lemma 2.4. Finally, Lemma 2.5 follows from Lemma 2.4. Proofs of the lemmas are provided in Appendix B. So that the reader can gain some intuition[4] behind the lemmas, we have drawn a couple of Venn diagrams in Figure 5. We would like to encourage the reader to translate the claims made in the lemmas into graphical statements in the context of Venn diagrams.

Lemma 2.1 *Let i, j, k denote informative character sites, of which (i, j) is an incompatible pair, whereas (i, k) and (j, k) are compatible pairs. If either $B_k \cap B_i = \varnothing$ or $B_k \cap B_i^c = \varnothing$, then*

$$B_k^c \cap B_j \neq \varnothing \qquad \textit{and} \qquad B_k^c \cap B_j^c \neq \varnothing.$$

Proof: See Appendix B.

[3] We emphasise that, in general, $R^{i,j} = R^{k,l}$ is not a *sufficient* condition for having a single recombination event which explains both incompatible pairs (i, j) and (k, l). If other incompatible pairs are present in the data S, then more than one recombination event may be required to explain the sequences in $R^{i,j} \subset S$.

[4] In fact, the main idea running through the present paper first stemmed from thinking about such simple diagrams.

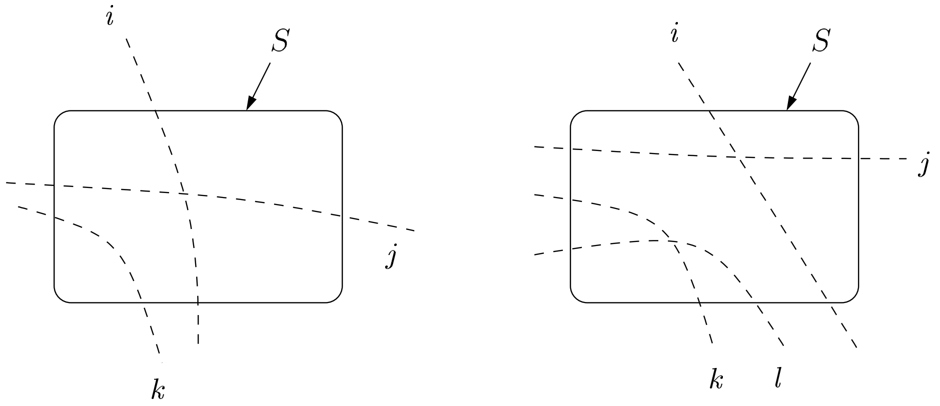

Figure 5: Venn diagrams pertaining to the situations described in the lemmas. In each diagram, the enclosed object drawn in solid line denotes the set S. Dashed lines are used to represent how an informative character site bipartitions S. When two informative character sites are incompatible, their corresponding lines intersect, partitioning the enclosed object into four disjoint regions.

Lemma 2.2 *Let i, j, k be defined as in Lemma 2.1. Then, exactly one of the following intersections is empty:*

$$B_i \cap B_k \quad , \; B_i \cap B_k^c \quad , \; B_i^c \cap B_k \quad , \; B_i^c \cap B_k^c \; .$$

Proof: See Appendix B.

Lemma 2.3 *Let i, j, k be as in Lemma 2.1. Then, the quadpartition $\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ of S contains exactly one Q_a^{ij} such that either*

$$\begin{array}{ll} & \text{(a)} \;\; B_k \cap Q_a^{ij} \neq \varnothing \; \textit{and} \; B_k \cap (\, S \setminus Q_a^{ij} \,) = \varnothing, \\ \textit{or} & \text{(b)} \;\; B_k^c \cap Q_a^{ij} \neq \varnothing \; \textit{and} \; B_k^c \cap (\, S \setminus Q_a^{ij} \,) = \varnothing. \end{array}$$

In other words, up to a choice of relabelling, there exists a unique $Q_a^{ij} \in \{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ such that either $B_k \subseteq Q_a^{ij}$ or $B_k^c \subseteq Q_a^{ij}$.

Proof: See Appendix B.

Lemma 2.4 *Let i, j, k, l be informative character sites such that (i, j) and (k, l) are incompatible pairs, whose corresponding quadpartitions are $\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ and $\{Q_1^{kl}, Q_2^{kl}, Q_3^{kl}, Q_4^{kl}\}$, respectively; and $(i, k), (i, l), (j, k)$, and (j, l) are compatible pairs. Then, there exists a subset Q_a^{ij} which is unique, up to a choice of relabelling, such that*

$$S \setminus Q_b^{kl} \subseteq Q_a^{ij}$$

for exactly one index value $b \in \{1, 2, 3, 4\}$.

Proof: See Appendix B.

Lemma 2.5 *Let i, j, k, l be defined as in Lemma 2.4. Then, there exist no $a, b \in \{1, 2, 3, 4\}$ such that*

$$Q_a^{ij} = Q_b^{kl} .$$

PROOF: See Appendix B.

Having established the above lemmas, we are now in a position to provide the desired proof of Proposition 2.1.

Proof of Proposition 2.1: Since incompatible pairs are present, $\rho > 0$. If $I(i_1, j_1) \cap I(i_2, j_2) = \varnothing$, then it is clear that more than one recombination event is required, and therefore $\rho > 1$. If $I(i_1, j_1) \cap I(i_2, j_2) \neq \varnothing$, then a necessary condition for $\rho = 1$ is that $R^{i_1,j_1} = R^{i_2,j_2}$. But, that is never possible according to Lemma 2.5. ■

A reason why Hudson and Kaplan's algorithm for determining the minimum number of recombination events may not work is now apparent. In their algorithm, one is to ignore an incompatible pair (k, l), $k < l$, if there exists another incompatible pair (i, j), $i < j$, such that either $I(i, j) \subset I(k, l)$ or $I(i, j) \cap I(k, l) \neq \varnothing$ and $i < k < j$ [HK]. But, Proposition 2.1 implies that, in general, such an algorithm may lead to a lower bound which is less than the minimum $\mathcal{R}_{\min}(S)$.

As we have seen thus far, set theoretical constraints provide interesting information which can be used to determine under what conditions a single recombination event can explain several incompatibilities. Furthermore, it is possible to obtain more set theoretical results regarding various situations, thus putting more constraints on the problem of counting recombination events. For example, one can easily obtain the following result:

Lemma 2.6 *Let i, j, k denote informative character sites. Let (i, j) and (j, k) be incompatible pairs, and let (i, k) be a compatible pair. Then, there exists a proper subset $Q_a^{ij} \in \{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ and a proper subset $Q_b^{jk} \in \{Q_1^{jk}, Q_2^{jk}, Q_3^{jk}, Q_4^{jk}\}$ such that*

$$Q_b^{jk} \subset Q_a^{ij}.$$

PROOF: See Appendix B.

We do not pursue here establishing more set theoretical lemmas, however. We have seen in the present section that set theoretical constraints have direct implications on recombination events. We can consider various patterns of incompatibilities, establish more set theoretical results similar to Proposition 2.1 and construct local lower bounds based on them. Myers and Griffiths [MG] have proposed a dynamic programming algorithm which uses a set of local lower bounds to construct a global lower bound. It would be interesting to apply their algorithm on the local lower bounds obtained from set theoretical results.

In the following section, we proceed to construct a new lower bound using our newly gained intuition. The underlying theme of our approach is to understand how a set of incompatibilities can share common recombination events.

3 A New Lower Bound

The goal of this section is to construct a new lower bound on the number of recombination events. In addition to taking the pattern—for example, overlapping open intervals $I(i,j)$—of incompatibilities into account, we also pay particular attention to recombining subsets $R^{i,j}$. In §3.1, we define $\mathcal{R}_{\mathrm{Q}}(S)$ and show that it satisfies the inequality $\mathcal{R}_{\mathrm{HK}}(S) \leq \mathcal{R}_{\mathrm{Q}}(S) \leq \mathcal{R}_{\min}(S)$, where $\mathcal{R}_{\min}(S)$ is the minimum number of recombination events and $\mathcal{R}_{\mathrm{HK}}(S)$ is Hudson and Kaplan's lower bound. We provide in §3.2 a recursive algorithm for simplifying the evaluation of $\mathcal{R}_{\mathrm{Q}}(S)$.

3.1 Definition of the New Lower Bound $\mathcal{R}_{\mathrm{Q}}(S)$

Recall that $S = \{s_\alpha\}$ denotes a set of binary DNA sequences

$$s_\alpha = c_1^\alpha, c_2^\alpha, \ldots, c_h^\alpha,$$

where $c_i^\alpha \in \{0,1\}$ for every $\alpha \in \{\underline{1}, \underline{2}, \ldots, \underline{n}\}$ and $i \in I_S := \{1, 2, \ldots, h\}$. We denote an open interval by $I(i,j) := \{y \in \mathbb{R} \mid i < y < j\}$.

Henceforth, we denote the set of all pairs of incompatible character sites by

$$\mathscr{P}(S) = \{(i,j) \in I_S \times I_S \mid i < j,\ (i,j) \text{ incompatible}\}\ .$$

In Definition 2.3 we have defined recombining subset $R^{i,j}$ and subsequently defined the set $\mathscr{R}(S)$ of recombining subsets. We note that if $\mathscr{P}(S)$ has cardinality $|\mathscr{P}|$, then there are $4^{|\mathscr{P}|}$ distinct choices of $\mathscr{R}(S)$. As this number can be very large and our new lower bounds are obtained by searching through all choices of $\mathscr{R}(S)$, our approach has a possibility of becoming intractable. In the following sub-section, we provide a reduction algorithm which simplifies the evaluation of our new lower bounds. From now on, we denote the set of all choices of $\mathscr{R}(S)$ by $\mathfrak{R}$ and say that $\mathscr{R}(S) \in \mathfrak{R}$.

Before we proceed to define our new lower bound, we wish to lay out some underlying intuition behind the formalism of our definition. Let 2^S denote the set of all subsets of S. Given a data set S, let $\mathcal{G}$ be a graphical representation of S with $\mathcal{R}_{\min}(S)$ recombination events, i.e. $\mathcal{G}$ contains exactly $\mathcal{R}_{\min}(S)$ recombination vertices[5]. For each recombination vertex v in $\mathcal{G}$, we define $Y(v) \in 2^S$ as the set of all sequences in S whose history at one of its character sites can trace back to v. Consider an incompatible pair $(i,j) \in \mathscr{P}(S)$ and let $\mathcal{H} \subset \mathcal{G}$ be the subgraph corresponding to the history of the character sites i and j. Since (i,j) is an incompatible pair, the subgraph $\mathcal{H}$ must contain at least one recombination vertex v whose corresponding recombination break-point is somewhere in the open interval $I(i,j)$. Furthermore, we know that every sequence in the set $R^{ij} \in \{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ must have histories at i and j which trace back to at least one recombination vertex in $\mathcal{H}$ with break-point in $I(i,j)$. We therefore have the following two facts, which form the guiding principles of our approach:

[5]Recombination vertex is defined in Introduction.

(F1) In the subgraph $\mathcal{H}$, there exists a finite set of recombination vertices $v_1, v_2, \dots, v_{n_{ij}}$ each with recombination break-point somewhere in the open interval $I(i,j)$, such that

(F2) using their associated sets $Y(v_1), Y(v_2), \dots, Y(v_{n_{ij}})$ one must be able to construct $R^{i,j}$.

We now describe the construction of our new lower bound. An assignment $\mathcal{A}$ is defined as follows: We assign to every $X \in 2^S$ a set of the form[6]

$$\mathcal{A}(X) := I(r_1, r_1+1) \cup I(r_2, r_2+1) \cup \cdots \cup I(r_{\omega(X;\mathcal{A})}, r_{\omega(X;\mathcal{A})}+1) \;\subset I(1,h)$$

and call $\omega(X;\mathcal{A})$ the weight of X in the assignment $\mathcal{A}$. If $\mathcal{A}(X) = \varnothing$, then $\omega(X;\mathcal{A}) := 0$. Roughly speaking, to each open interval of the form $I(r, r+1)$ in the assignment, we wish to associate a recombination event in the history of the sequences in X. We emphasise at the outset that, as we later demonstrate in the examples in §4, in practice most of the elements X_a in 2^S get assigned $\mathcal{A}(X_a) = \varnothing$ and therefore do not contribute to evaluating our new lower bound.

We now introduce constraints on the assignment $\mathcal{A}$. The assignment must be performed to satisfy the following properties:
For every $R^{i,j} \in \mathscr{R}(S)$, $(i,j) \in \mathscr{P}(S)$, there exist $X_1, X_2, \dots, X_{m_{ij}} \in 2^S$ such that

(P1) for all $a \in \{1, 2, \dots, m_{ij}\}$, $\mathcal{A}(X_a) \cap I(i,j) \neq \varnothing$, and

(P2)
- either $R^{i,j} = X_1 \cup X_2 \cup \cdots \cup X_{m_{ij}}$, or
- $R^{i,j} = X_1 \cup X_2 \cup \cdots \cup X_l \setminus (X_{l+1} \cup X_{l+2} \cup \cdots \cup X_{m_{ij}})$, where for every $a \in \{l+1, l+2, \dots, m_{ij}\}$, $X_a \subset X_b$ for some $b \in \{1, 2, \dots, l\}$.

Then, we define $\mathcal{R}_{\mathrm{Q}}(S)$ as the minimum possible value of the sum of the weights $\omega(X_a;\mathcal{A})$ over all choices of $\mathscr{R}(S)$; that is, if we denote by $\mathfrak{A}$ the set of all possible assignments satisfying the above properties, then

$$\mathcal{R}_{\mathrm{Q}}(S) := \min_{\substack{\mathcal{A} \in \mathfrak{A}, \\ \mathscr{R}(S) \in \mathfrak{R}}} \left(\sum_{X \in 2^S} \omega(X;\mathcal{A}) \right) . \tag{3.1}$$

That $\mathcal{R}_{\mathrm{Q}}(S)$ gives a lower bound on the number of recombination events is shown in Proposition 3.1, where we also show that $\mathcal{R}_{\mathrm{Q}}(S)$ is in general sharper than Hudson and Kaplan's lower bound $\mathcal{R}_{\mathrm{HK}}(S)$. The reader is strongly encouraged to go through the proof of the proposition to understand the meaning of Properties (P1) and (P2).

[6] To avoid cluttering notation, we have suppressed the dependence on X and $\mathcal{A}$ in writing r_n.

Proposition 3.1 *Let S be a set of binary DNA sequences of fixed length. In representing S by a connected graph as described in this paper, $\mathcal{R}_{\mathrm{Q}}(S)$ defined in (3.1) satisfies the inequality*

$$\mathcal{R}_{\mathrm{HK}}(S) \le \mathcal{R}_{\mathrm{Q}}(S) \le \mathcal{R}_{\min}(S),$$

where $\mathcal{R}_{\min}(S)$ is the minimum number of recombination events and $\mathcal{R}_{\mathrm{HK}}(S)$ is the lower bound given in [HK].

Proof: We first show that $\mathcal{R}_{\mathrm{Q}}(S) \le \mathcal{R}_{\min}(S)$. As described in the beginning of this section, consider the graphical representation $\mathcal{G}$ with exactly $\mathcal{R}_{\min}(S)$ recombination vertices. We again denote by $\mathcal{H}$ the subgraph of $\mathcal{G}$ which corresponds to the history of incompatible character sites i and j. In addition, we let $\mathcal{V}^{ij}$ be the set of all recombination vertices in $\mathcal{H}$ whose corresponding recombination break-points are contained in the open interval $I(i,j)$.

Without any loss of generality, we suppose that the sequences in $Q(c_i c_j) = Q(00) \subset S$ are explained by recombination events; that is, $R^{i,j} = Q(00)$. Then, every sequence $s_\alpha \in R^{ij}$ has undergone at least one recombination event with break-point in $I(i,j)$ to attain $c_i^\alpha = 0$ and $c_j^\alpha = 0$. Moreover, R^{ij} must be obtained from $Y(v_1), Y(v_2), \ldots, Y(v_{n_{ij}})$, where $v_1, v_2, \ldots, v_{n_{ij}}$ are some recombination vertices in $\mathcal{V}^{ij}$. Before we proceed to elaborate this point, to avoid long-winded phrases, we introduce coarse classifications of recombination vertices and of their associated Y-sets, i.e. $Y(v)$. A recombination vertex v in $\mathcal{V}^{ij}$ can be of one of the following two types:

Type A : The recombination event at v produces $c_i c_j = 0\,0$.

Type B : The recombination event at v produces $c_i c_j \neq 0\,0$.

The associated Y-sets can be classified into three types:

Type I : Every sequence s_α in $Y(v)$ has $c_i^\alpha c_j^\alpha = 0\,0$.

Type II : Every sequence s_α in $Y(v)$ has $c_i^\alpha c_j^\alpha \neq 0\,0$.

Type III : Some sequences in $Y(v)$ have $c_i c_j = 0\,0$, whereas others have $c_i c_j \neq 0\,0$.

We have divided our discussion into two main cases as follows:

• Case 1) *Every recombination vertex in $\mathcal{V}^{ij}$ is of type A.*

We note that both mutation events $\mathbf{m}_i$ and $\mathbf{m}_j$ must have already occurred before the earliest recombination vertex of type A. That is because the earliest recombination event must produce $c_i c_j = 0\,0$ from $c_i c_j = 0\,1$ and $c_i c_j = 1\,0$, and mutation events $\mathbf{m}_i$ and $\mathbf{m}_j$ must have occurred between $0\,1$ and $1\,0$. Hence, by Assumption 1.1, no further mutation events occur at sites i and j after the first recombination event which produces $c_i c_j = 0\,0$. Now, suppose there exists a type A recombination vertex $v \in \mathcal{V}^{ij}$ with $Y(v)$ *not* of type I, i.e. $Y(v)$ contains at least one sequence with $c_i c_j \neq 0\,0$. Then, since no further mutation events are allowed, $\mathcal{V}^{ij}$ must contain a recombination vertex of type B so that the sequences in $Y(v)$ with $c_i c_j \neq 0\,0$ are represented by the graph. This contradicts the condition that every recombination vertex v in $\mathcal{V}^{ij}$ is of type A. Hence, we conclude that if every recombination vertex in $\mathcal{V}^{ij}$ is of type A, then $Y(v)$ must be of type I for every $v \in \mathcal{V}^{ij}$. It now follows that every $s_\alpha \in Y(v)$, where $v \in \mathcal{V}^{ij}$, is also contained in $R^{ij} = Q(00)$. Also, it is clear that, to every sequence $s_\alpha \in R^{ij}$, there must correspond a recombination vertex $v \in \mathcal{V}^{ij}$ whose associated set $Y(v)$ contains s_α. These two facts together imply that there exists a set of recombination vertices $v_1, v_2, \ldots v_{n_{ij}} \in \mathcal{V}^{ij}$ such that $R^{ij} = Y(v_1) \cup Y(v_2) \cup \cdots \cup Y(v_{n_{ij}})$.

- Case 2) $\mathcal{V}^{ij}$ *contains at least one recombination vertex of type B.*

  The present case can be further divided into two sub-cases.

  - Case 2.1) *There exists no type A recombination vertex v with $Y(v)$ of type III.*

  In this case $\mathcal{V}^{ij}$ must contain some type A recombination vertices with associated Y-sets of type I. Moreover, following a similar line of reasoning as above, we can conclude that $R^{ij} = Y(v_1) \cup Y(v_2) \cup \cdots \cup Y(v_{n_{ij}})$, where for every $a \in \{1, 2, \ldots, n_{ij}\}$, $v_a \in \mathcal{V}^{ij}$ is of type A and $Y(v_a)$ of type I.

  - Case 2.2) *There exists at least one type A recombination vertex v with $Y(v)$ of type III.*

  Let $v_a \in \mathcal{V}^{ij}$ be a type A recombination vertex with $Y(v_a)$ of type III. Then, $Y(v_a)$ contains some sequences $s_{\alpha_1}, \ldots, s_{\alpha_p}$ which are not contained in R^{ij}—i.e. $c_i c_j \neq 00$ for $s_{\alpha_1}, \ldots, s_{\alpha_p}$— and Assumption 1.1 implies that the sequences $s_{\alpha_1}, \ldots, s_{\alpha_p}$ must have undergone further recombination with break-point in $I(i,j)$. More precisely, there must exist a subsequent recombination vertex $v_b \in \mathcal{V}^{ij}$ of type B such that $Y(v_b) \subset Y(v_a)$ and $\{s_{\alpha_1}, \ldots, s_{\alpha_q}\} \subseteq Y(v_b)$ for some $q \in \{1, 2, \ldots, p\}$. $Y(v_b)$ cannot be of type I since $Y(v_b)$ contains at least one sequence which is not in $R^{ij} = Q(00)$. If $Y(v_b)$ is of type II, then $\{s_{\alpha_1}, \ldots, s_{\alpha_q}\} = Y(v_b)$. If $Y(v_b)$ is of type III—i.e. contains at least one sequence in R^{ij}—then there must exist a subsequent type A recombination vertex $v_c \in \mathcal{V}^{ij}$ such that $Y(v_c) \subset Y(v_b)$ and $Y(v_c)$ contains at least one sequence which is also contained in R^{ij}. $Y(v_c)$ can be either of type I or of type III, and we can continue in a similar vein with the analysis as we have just described. Returning to the recombination vertex v_a, we note that the same logic as sketched above applies to the remaining sequences in $s_{\alpha_1}, \ldots, s_{\alpha_p}$. More generally, what we have just described implies that, in the present case of our consideration, there exists a set of recombination vertices $v_1, v_2, \ldots, v_{n_{ij}}$ such that $R^{ij} = Y(v_1) \cup Y(v_2) \cup \cdots \cup Y(v_k) \setminus Y(v_{k+1}) \cup Y(v_{k+2}) \cup \cdots \cup Y(v_{n_{ij}})$; where for every $a \in \{k+1, k+2, \ldots, n_{ij}\}$, $Y(v_a) \subset Y(v_b)$ for some $b \in \{1, 2, \ldots, k\}$; each of $Y(v_1), \ldots, Y(v_k)$ is either of type I or of type III; and each of $Y(v_{k+1}), \ldots, Y(v_{n_{ij}})$ is either of type II or of type III.

It remains to draw a connection with the definition of $\mathcal{R}_{\mathrm{Q}}(S)$. There exists an isomorphism between what we have described in the preceding paragraphs and a particular assignment $\mathcal{A}_0$. Namely, for each incompatible pair $(i,j) \in \mathscr{P}(S)$, we take $m_{ij} = n_{ij}$ and $X_a = Y(v_a)$ for every $a \in \{1, 2, \ldots, n_{ij}\}$. It is easy to see that Fact (F1) simply translates to Property (P1). To each recombination vertex w in $\mathcal{G}$, we can associate an open interval of the form $I(r, r+1)$ which contains the corresponding recombination break-point of w. Hence, we can associate an open interval of the form $I(r, r+1)$ to each $Y(w)$ and have $I(r, r+1)$ included in $\mathcal{A}_0(Y(w))$. Therefore, if v is a recombination vertex which appears in (F1), then it follows that $\mathcal{A}_0(Y(v)) \cap I(i,j) \neq \varnothing$. If a certain subset $X \in 2^S$ satisfies $X = Y(v_1), X = Y(v_2), \ldots, X = Y(v_d)$, for d distinct recombination vertices $v_1, v_2, \ldots, v_d$ in $\mathcal{G}$, then $\mathcal{A}_0(X) = I(r_1, r_1+1) \cup I(r_2, r_2+1) \cup \cdots \cup I(r_d, r_d+1)$. $\mathcal{R}_{\min}(S)$ is equal to the total number of recombination vertices in $\mathcal{G}$ and therefore $\mathcal{R}_{\min}(S) = \sum_{X \in 2^S} \omega(X; \mathcal{A}_0)$, where $\omega(X; \mathcal{A}_0)$ is the weight of X in the assignment $\mathcal{A}_0$ described above. Moreover, it follows from previous paragraphs that the assignment $\mathcal{A}_0$ with $X_a = Y(v_a)$ satisfies Properties (P1) and (P2) for a particular $\mathscr{R}(S)$, and therefore we conclude that $\mathcal{A}_0 \in \mathfrak{A}$. In equation (3.1), we define $\mathcal{R}_{\mathrm{Q}}(S)$ by taking the minimum over all possible assignments $\mathcal{A} \in \mathfrak{A}$ and all possible choices of $\mathscr{R}(S)$. We therefore conclude that $\mathcal{R}_{\mathrm{Q}}(S) \leq \mathcal{R}_{\min}(S)$.

We now show that $\mathcal{R}_{\mathrm{HK}}(S) \leq \mathcal{R}_{\mathrm{Q}}(S)$. Recall that Hudson and Kaplan's algorithm in [HK] gives a set $\mathscr{D}$ which consists of certain incompatible pairs such that if $(i,j),(k,l) \in \mathscr{D}$, then $I(i,j) \cap I(k,l) = \varnothing$. Moreover, $\mathcal{R}_{\mathrm{HK}}(S)$ is defined to be equal to the cardinality $|\mathscr{D}|$ of $\mathscr{D}$. According to the definition of $\mathcal{R}_{\mathrm{Q}}(S)$, since $\mathscr{D}$ is a subset of $\mathscr{P}(S)$, for every $(i,j) \in \mathscr{D}$, we need to take $X_1, X_2, \ldots, X_{m_{ij}} \in 2^S$ and $\mathcal{A}(X_1), \mathcal{A}(X_2), \ldots, \mathcal{A}(X_{m_{ij}})$ which satisfy the specified properties. In particular, since there are $|\mathscr{D}|$ disjoint open intervals associated to the elements in $\mathscr{D}$, Property (P1) implies that

$$|\mathscr{D}| \leq \sum_{X \in 2^S} \omega(X; \mathcal{A})$$

for every possible assignment $\mathcal{A} \in \mathfrak{A}$. It thus follows that $\mathcal{R}_{\mathrm{HK}}(S) = |\mathscr{D}| \leq \mathcal{R}_{\mathrm{Q}}(S)$. ∎

3.2 Reduction Algorithm for Evaluating $\mathcal{R}_{\mathrm{Q}}(S)$

In the definition of $\mathcal{R}_{\mathrm{Q}}(S)$, we take the minimum over all choices of $\mathscr{R}(S)$ in $\mathfrak{R}$. But, as remarked in the previous section, $\mathfrak{R}$ has cardinality $4^{|\mathscr{P}|}$, which can be very large, thus making the evaluation of $\mathcal{R}_{\mathrm{Q}}(S)$ intractable. Hence we would like to construct a method of evaluating $\mathcal{R}_{\mathrm{Q}}(S)$ more efficiently. Instead of considering all incompatible pairs in $\mathscr{P}(S)$, our goal is to isolate a subset, hopefully of small cardinality, and only consider the incompatible pairs in that subset to evaluate $\mathcal{R}_{\mathrm{Q}}(S)$. In the following we describe a recursive method.

(**Remark**: We drop the notation (S) when writing sets of incompatible pairs. For example, we just write $\mathscr{P}$ instead of $\mathscr{P}(S)$.)

Our method recursively decomposes $\mathscr{P}$ into

$$\{\mathscr{E}_0, \mathscr{E}_1, \ldots, \mathscr{E}_\ell\}\ ,\ \{\mathscr{X}_0, \mathscr{X}_1, \ldots, \mathscr{X}_\ell\}\ ,\ \{\mathscr{Y}_0, \mathscr{Y}_1, \ldots, \mathscr{Y}_\ell\}\ ,$$

where $\mathscr{E}_n, \mathscr{X}_n, \mathscr{Y}_n$ are disjoint sets of incompatible pairs which satisfy certain conditions. Only the incompatible pairs in $\{\mathscr{E}_0, \mathscr{E}_1, \ldots, \mathscr{E}_\ell\}$ are used in computing $\mathcal{R}_{\mathrm{Q}}(S)$. Note that the properties used in defining $\mathscr{X}_n$ are a special case of that used in defining $\mathscr{Y}_n$. We have divided them into two separate cases just to make bookkeeping easier. In order to understand wherefore the forthcoming constrains are imposed, the reader should refer back to Properties (P1) and (P2) of §3.1. The main idea is that if one can find an assignment $\mathcal{A}$ which satisfies Properties (P1) and (P2) for every incompatible pair in $\{\mathscr{E}_0, \mathscr{E}_1, \ldots, \mathscr{E}_\ell\}$, then $\mathcal{A}$ satisfies (P1) and (P2) also for every incompatible pair in $\{\mathscr{X}_0, \mathscr{X}_1, \ldots, \mathscr{X}_\ell\} \cup \{\mathscr{Y}_0, \mathscr{Y}_1, \ldots, \mathscr{Y}_\ell\}$.

Definition 3.1 (Essential Incompatibility in $\mathscr{F}$) *Let $\mathscr{F}$ be a set which contains incompatible pairs and let (i,j) be an incompatible pair in $\mathscr{F}$. We say that (i,j) is an* essentially *incompatible pair in $\mathscr{F}$ if there exists no incompatible pair $(k,l) \in \mathscr{F}$ satisfying $I(k,l) \subset I(i,j)$.*

For non-negative integers n, we define

$$\mathscr{E}_n = \{\ (i,j) \in (\mathscr{P} \setminus \mathscr{H}_{n-1})\ \mid\ (i,j) \text{ essentially incompatible in } \mathscr{P} \setminus \mathscr{H}_{n-1}\}\ ,$$

where $\mathscr{H}_{n-1}$ are defined presently. For each incompatible pair $(i,j) \in \mathscr{E}_n$, we make a choice of recombining subset $R^{i,j}$. We then let $\mathscr{X}_n = \bigcup_{(i,j)\in\mathscr{E}_n} \mathscr{X}_n^{i,j}$, where

$$\mathscr{X}_n^{i,j} := \{(k,l) \in \mathscr{P} \setminus (\mathscr{H}_{n-1} \cup \mathscr{E}_n) \,\big|\, \text{properties (a) and (b) are satisfied }\}.$$

(a) $I(i,j) \subset I(k,l)$, and

(b) given a choice of $R^{i,j}$ for the incompatible pair $(i,j) \in \mathscr{E}_n$, there exists $b \in \{1,2,3,4\}$ such that $Q_b^{kl} = R^{i,j}$.

It is straightforward to see that, if there exists an assignment $\mathcal{A}$ and subsets $X_1, X_2, \ldots, X_{m_{i,j}} \in 2^S$ such that properties (P1) and (P2) of §3.1 are satisfied for $(i,j) \in \mathscr{E}_n$, then the same assignment $\mathcal{A}$ and subsets $X_1, X_2, \ldots, X_{m_{i,j}} \in 2^S$ satisfy (P1) and (P2) also for the incompatible pair $(k,l) \in \mathscr{X}_n^{i,j}$ with $R^{k,l} = Q_b^{kl}$, where Q_b^{kl} is from property (b) shown above. Therefore, we can ignore the pair $(k,l) \in \mathscr{X}_n^{i,j}$ as long as $(i,j) \in \mathscr{E}_n$ is taken into account when computing $\mathcal{R}_{\mathrm{Q}}(S)$.

The set $\mathscr{Y}_n$ is defined as

$$\mathscr{Y}_n = \{(k,l) \in \mathscr{P} \setminus (\mathscr{H}_{n-1} \cup \mathscr{E}_n \cup \mathscr{X}_n) \,|\, \text{properties (a}'\text{) and (b}'\text{) are satisfied } \} \ .$$

(a$'$) There exists a set of incompatible pairs $(i_1, j_1), (i_2, j_2), \ldots, (i_{d_{k,l}}, j_{d_{k,l}}) \in \bigcup_{s=0}^{n} \mathscr{E}_s$ such that, for all $p \in \{1, 2, \ldots, d_{k,l}\}$, $I(i_p, j_p) \subset I(k,l)$, and

(b$'$) there exists $b' \in \{1,2,3,4\}$ such that $Q_{b'}^{kl} = R^{i_1,j_1} \cup R^{i_2,j_2} \cup \cdots \cup R^{i_{d_{k,l}}, j_{d_{k,l}}}$.

Similar to the case of $\mathscr{X}_n$, it is easy to see that if properties (P1) and (P2) are satisfied for each of $(i_1, j_1), (i_2, j_2), \ldots, (i_{d_{k,l}}, j_{d_{k,l}}) \in \bigcup_{s=0}^{n} \mathscr{E}_s$, then the same also holds true for $(k,l) \in \mathscr{Y}_n$ with $R^{k,l} = Q_{b'}^{kl}$, where $Q_{b'}^{kl}$ is from property (b$'$). Hence, we can ignore the pair $(k,l) \in \mathscr{Y}_n$ as long as $(i_1, j_1), (i_2, j_2), \ldots, (i_{d_{k,l}}, j_{d_{k,l}}) \in \bigcup_{s=0}^{n} \mathscr{E}_s$ are taken into account when computing $\mathcal{R}_{\mathrm{Q}}(S)$.

The recursion begins with the initial condition $\mathscr{H}_{-1} = \varnothing$, so that

$$\mathscr{E}_0 = \{(i,j) \in \mathscr{P} \,|\, (i,j) \text{ essentially incompatible in } \mathscr{P}\}.$$

In each succeeding step, we define

$$\mathscr{H}_n := \mathscr{H}_{n-1} \cup \mathscr{E}_n \cup \mathscr{X}_n \cup \mathscr{Y}_n \,.$$

Let ℓ be an index value that satisfies $\mathscr{E}_\ell \neq \varnothing$ and $\mathscr{E}_{\ell+1} = \varnothing$, and define $\mathscr{E} := \bigcup_{s=0}^{\ell} \mathscr{E}_s$. At that point we would have $\mathscr{H}_\ell = \mathscr{P}$, and $\mathcal{R}_{\mathrm{Q}}(S)$ can be determined as follows:

1. Determine the set $\mathscr{P}$ of all pairs (i,j), $i < j$, of incompatible character sites.

2. Determine the quadpartitions corresponding to the incompatible pairs in $\mathscr{P}$.

3. Choose $R^{i,j}$ for each $(i,j) \in \mathscr{E}_n$ and recursively construct

$$\mathcal{M} := \{\{\mathscr{E}_0, \mathscr{E}_1, \ldots, \mathscr{E}_\ell\}, \{\mathscr{X}_0, \mathscr{X}_1, \ldots, \mathscr{X}_\ell\}, \{\mathscr{Y}_0, \mathscr{Y}_1, \ldots, \mathscr{Y}_\ell\}\} \ .$$

   Denote by $\mathfrak{M} \ni \mathcal{M}$ the set of all such constructions.

4. For every $(i,j) \in \mathscr{E}$, find an assignment $\mathcal{A}$ so that Properties (P1) and (P2) in §3.1 are satisfied

5. Let $\widetilde{\omega}(X; \mathcal{A}; \mathcal{M})$ be the weight of $X \in 2^S$ determined by the above steps. Then,

$$\mathcal{R}_{\mathrm{Q}}(S) = \min_{\substack{\mathcal{A} \in \mathfrak{A}, \\ \mathcal{M} \in \mathfrak{M}}} \left(\sum_{X \in 2^S} \widetilde{\omega}(X; \mathcal{A}; \mathcal{M}) \right) \ .$$

4 Applications

In this section, we consider some specific examples, the last of which being Kreitman's 1983 data of the alcohol dehydrogenase locus from 11 chromosomes of *Drosophila melanogaster* [K], and demonstrate how one can go about determining the new lower bounds $\mathcal{R}_{\mathrm{Q}}(S)$. The reduction algorithm in §3.2 pays off the most when we analyse Kreitman's data. In each of the examples considered, the new lower bound $\mathcal{R}_{\mathrm{Q}}(S)$ coincides with the minimum $\mathcal{R}_{\min}(S)$, whereas $\mathcal{R}_{\mathrm{HK}}(S) < \mathcal{R}_{\min}(S)$.

We have written a computer program which partially implements the recursive algorithm described in the previous section, and have applied it to the following examples.

(**Remark**: The same notations as in §2 are used here.)

Example 4.1 $S = \{\underline{1}, \underline{2}, \ldots, \underline{7}\}$

$\underline{1} = 0\ 0\ 0\ 1$
$\underline{2} = 0\ 0\ 1\ 1$
$\underline{3} = 0\ 1\ 0\ 0$
$\underline{4} = 1\ 0\ 0\ 1$
$\underline{5} = 0\ 1\ 1\ 0$
$\underline{6} = 1\ 1\ 0\ 0$
$\underline{7} = 1\ 0\ 1\ 0$

$(i,j) \in \mathscr{P}$	$Q(00)$	$Q(01)$	$Q(10)$	$Q(11)$
$(1,2)$	$\{\underline{1},\underline{2}\}$	$\{\underline{3},\underline{5}\}$	$\{\underline{4},\underline{7}\}$	$\{\underline{6}\}$
$(1,3)$	$\{\underline{1},\underline{3}\}$	$\{\underline{2},\underline{5}\}$	$\{\underline{4},\underline{6}\}$	$\{\underline{7}\}$
$(1,4)$	$\{\underline{3},\underline{5}\}$	$\{\underline{1},\underline{2}\}$	$\{\underline{6},\underline{7}\}$	$\{\underline{4}\}$
$(2,3)$	$\{\underline{1},\underline{4}\}$	$\{\underline{2},\underline{7}\}$	$\{\underline{3},\underline{6}\}$	$\{\underline{5}\}$
$(3,4)$	$\{\underline{3},\underline{6}\}$	$\{\underline{1},\underline{4}\}$	$\{\underline{5},\underline{7}\}$	$\{\underline{2}\}$

Essentially incompatible pairs in $\mathscr{P}$ are $\mathscr{E}_0 = \{(1,2),(2,3),(3,4)\}$. Choosing

$$R^{1,2} = \{\underline{6}\}\ ,\ R^{2,3} = \{\underline{5}\}\ ,\ R^{3,4} = \{\underline{2}\}$$

gives $\mathscr{X}_0 = \varnothing$, $\mathscr{Y}_0 = \varnothing$. In the next step of recursion $\mathscr{E}_1 = \{(1,3)\}$, and we can choose $R^{1,3} = \{\underline{7}\}$ to obtain $\mathscr{X}_1 = \varnothing$ and $\mathscr{Y}_1 = \{(1,4)\}$, for $(1,4)$ then has $Q(10)$ equal to $R^{1,2} \cup R^{1,3}$ and $I(1,4)$ contains both $I(1,2)$ and $I(1,3)$. The recursion terminates here and we have $\mathscr{E} = \mathscr{E}_0 + \mathscr{E}_1 = \{(1,2),(2,3),(3,4),(1,3)\}$. We make the following assignment $\mathcal{A}$ to the elements of 2^S:

$$\mathcal{A}(\{\underline{6}\}) = I(1,2)\quad,\quad \mathcal{A}(\{\underline{7}\}) = I(2,3)\quad,\quad \mathcal{A}(\{\underline{5}\}) = I(2,3)\quad,\quad \mathcal{A}(\{\underline{2}\}) = I(3,4)\quad,$$

and $\mathcal{A}(X) = \varnothing$ for all other $X \in 2^S$.

Hence, for the present construction $\mathcal{M}$ and assignment $\mathcal{A}$, the weights are

$$\widetilde{\omega}(X;\mathcal{A};\mathcal{M}) = \begin{cases} 1, & X = \{\underline{2}\},\{\underline{5}\},\{\underline{6}\},\{\underline{7}\}\ , \\ 0, & \text{otherwise,} \end{cases}$$

thus giving $\sum_{X\in 2^S} \widetilde{\omega}(X;\mathcal{A};\mathcal{M}) = 4$. We have checked that every other possible pair $(\mathcal{A},\mathcal{M}) \in \mathfrak{A} \times \mathfrak{M}$ gives

$$\sum_{X\in 2^S} \widetilde{\omega}(X;\mathcal{A};\mathcal{M}) \geq 4,$$

and we thus conclude that $\mathcal{R}_{\mathrm{Q}}(S) = 4$, which in fact is equal to $\mathcal{R}_{\min}(S)$. A graphical representation of S with four recombination vertices is shown in Figure 6. Note that each of $\underline{2}, \underline{5}, \underline{6}, \underline{7}$ has undergone a single recombination event, and in each case the recombination break-point is in the open interval specified by the assignment $\mathcal{A}$ we have made above.

In contrast, if we apply Hudson and Kaplan's algorithm to the same data S, we end up the set $\mathscr{E}_0$, which contains three pairs of incompatible sites whose associated open intervals $I(i,j)$ are pair-wise disjoint. Thus, we obtain $\mathcal{R}_{\mathrm{HK}}(S) = 3$.

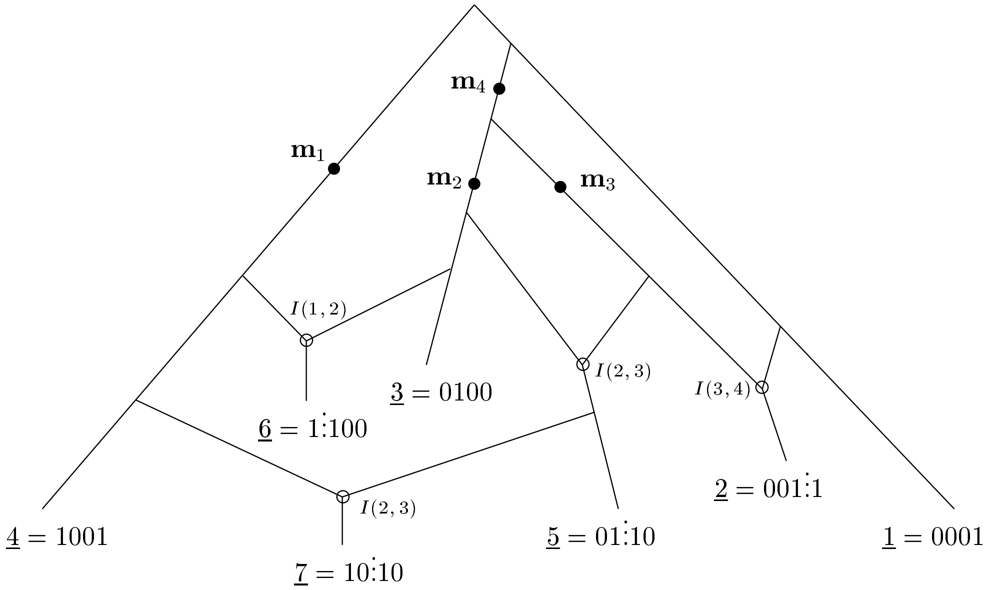

Figure 6: A graphical representation of S from Example 4.1 with the minimum number of recombination vertices. The have used the same conventions as that explained in Figure 1. In addition, we have indicated where recombination break-points occur by writing an open interval $I(r, r+1)$ next to each recombination vertex.

Example 4.2 $S = \{\underline{1}, \underline{2}, \ldots, \underline{10}\}$

$\underline{1} = 0\,0\,0\,0\,0$
$\underline{2} = 1\,0\,0\,0\,0$
$\underline{3} = 1\,1\,0\,0\,0$
$\underline{4} = 1\,1\,1\,0\,0$
$\underline{5} = 0\,0\,0\,0\,1$
$\underline{6} = 1\,0\,0\,0\,1$
$\underline{7} = 1\,1\,0\,0\,1$
$\underline{8} = 1\,1\,1\,0\,1$
$\underline{9} = 1\,1\,1\,1\,0$
$\underline{10} = 1\,1\,1\,1\,1$

$(i,j) \in \mathscr{P}$	$Q(00)$	$Q(01)$	$Q(10)$	$Q(11)$
$(1,5)$	$\{\underline{1}\}$	$\{\underline{5}\}$	$\{\underline{2}, \underline{3}, \underline{4}, \underline{9}\}$	$\{\underline{6}, \underline{7}, \underline{8}, \underline{10}\}$
$(2,5)$	$\{\underline{1}, \underline{2}\}$	$\{\underline{5}, \underline{6}\}$	$\{\underline{3}, \underline{4}, \underline{9}\}$	$\{\underline{7}, \underline{8}, \underline{10}\}$
$(3,5)$	$\{\underline{1}, \underline{2}, \underline{3}\}$	$\{\underline{5}, \underline{6}, \underline{7}\}$	$\{\underline{4}, \underline{9}\}$	$\{\underline{8}, \underline{10}\}$
$(4,5)$	$\{\underline{1}, \underline{2}, \underline{3}, \underline{4}\}$	$\{\underline{5}, \underline{6}, \underline{7}, \underline{8}\}$	$\{\underline{9}\}$	$\{\underline{10}\}$

In this example, regardless of what we choose for $R^{4,5}$, $R^{3,5}$, $R^{2,5}$, $R^{1,5}$, we end up with

$$\mathscr{E}_0 = \{(4,5)\} \;\;,\;\; \mathscr{E}_1 = \{(3,5)\} \;\;,\;\; \mathscr{E}_2 = \{(2,5)\} \;\;,\;\; \mathscr{E}_3 = \{(1,5)\} \;\;,$$

and $\mathscr{X}_a = \mathscr{Y}_a = \varnothing$, for all $a \in \{0,1,2,3\}$. Hence, the reduction algorithm of §3.2 is of little use. For $R^{4,5} = \{\underline{9}\}$, $R^{3,5} = \{\underline{4}, \underline{9}\}$, $R^{2,5} = \{\underline{3}, \underline{4}, \underline{9}\}$, $R^{1,5} = \{\underline{2}, \underline{3}, \underline{4}, \underline{9}\}$, a possible assignment is

$$\mathcal{A}(\{\underline{9}\}) = I(4,5) \quad , \quad \mathcal{A}(\{\underline{4}\}) = I(3,4) \quad , \quad \mathcal{A}(\{\underline{3}\}) = I(2,3) \quad , \quad \mathcal{A}(\{\underline{2}\}) = I(1,2) \quad ,$$

and $\mathcal{A}(X) = \varnothing$ for all other $X \in 2^S$. Hence, $\sum_{X \in 2^S} \omega(X; \mathcal{A}) = 4$ for the above assignment. We have checked that every choice of $\mathscr{R}(S)$ and possible assignment $\mathcal{A}$ gives $\sum_{X \in 2^S} \omega(X; \mathcal{A}) \geq 4$. Thus, $\mathcal{R}_{\mathrm{Q}}(S) = 4$. In fact, it is possible to construct a graphical representation of S with four recombination vertices, and therefore we have again reproduced the minimum $\mathcal{R}_{\min}(S) = 4$ using our set theoretical approach.

On the other hand, Hudson and Kaplan's algorithm only takes the pair $(4,5)$ into account and gives $\mathcal{R}_{\mathrm{HK}}(S) = 1$.

Example 4.3 This example concerns Kreitman's 1983 data of the alcohol dehydrogenase locus of *Drosophila melanogaster* [K]. Ignoring insertions and deletions, there are 43 character sites in the data. We have transformed the data into binary sequences as follows:

$$\begin{array}{rcl}
1S & = & 0000000011000000001101110111100000000000000 \\
2S & = & 0010000000000000001101110111100000000000000 \\
3S & = & 0000000000000000000000000000000000010000101 \\
4S & = & 0000000000000000110000000000000000010011000 \\
5S & = & 0001100010110011110000000000000000001000000 \\
6S & = & 0010000000000000100000000000000010101110000 10 \\
1F & = & 0010000000000000100000000000001111110100000 0 \\
2F & = & 1111100010111001000000000000011111101100000 \\
3F & = & 1111100010111001000000000000011111101100000 \\
4F & = & 1111100010111001000000000000011111101100000 \\
5F & = & 1111111110000101000010001000011111101000000
\end{array}$$

Since $2F = 3F = 4F$, in our analysis we only need to consider 9 distinct sequences. Hence, we relabel the sequences as follows:

$$\begin{array}{lll}
\underline{1} := 1S & \underline{2} := 2S & \underline{3} := 3S \\
\underline{4} := 4S & \underline{5} := 5S & \underline{6} := 6S \\
\underline{7} := 1F & \underline{8} := 2F = 3F = 4F & \underline{9} := 5F
\end{array}$$

There are a total of 83 pairs of incompatible character sites in S. Incompatible pairs of sites and their corresponding quadpartitions are summarised in the following table:

$(i,j) \in \mathscr{P}$	$Q(00)$	$Q(01)$	$Q(10)$	$Q(11)$
$(1,11)$	$\{1,2,3,4,6,7\}$	$\{5\}$	$\{9\}$	$\{8\}$
$(1,12)$	$\{1,2,3,4,6,7\}$	$\{5\}$	$\{9\}$	$\{8\}$
$(2,11)$	$\{1,2,3,4,6,7\}$	$\{5\}$	$\{9\}$	$\{8\}$
$(2,12)$	$\{1,2,3,4,6,7\}$	$\{5\}$	$\{9\}$	$\{8\}$
$(3,4)$	$\{1,3,4\}$	$\{5\}$	$\{2,6,7\}$	$\{8,9\}$
$(3,5)$	$\{1,3,4\}$	$\{5\}$	$\{2,6,7\}$	$\{8,9\}$
$(3,9)$	$\{3,4\}$	$\{1,5\}$	$\{2,6,7\}$	$\{8,9\}$
$(3,11)$	$\{1,3,4\}$	$\{5\}$	$\{2,6,7,9\}$	$\{8\}$
$(3,12)$	$\{1,3,4\}$	$\{5\}$	$\{2,6,7,9\}$	$\{8\}$
$(3,16)$	$\{1,3,4\}$	$\{5\}$	$\{2\}$	$\{6,7,8,9\}$
$(3,19)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,20)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,22)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,23)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,24)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,26)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,27)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,28)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,29)$	$\{3,4,5\}$	$\{1\}$	$\{6,7,8,9\}$	$\{2\}$
$(3,36)$	$\{1,5\}$	$\{3,4\}$	$\{2,7,8,9\}$	$\{6\}$
$(3,37)$	$\{1,3,4\}$	$\{5\}$	$\{2\}$	$\{6,7,8,9\}$
$(4,17)$	$\{1,2,3,6,7\}$	$\{4\}$	$\{8,9\}$	$\{5\}$
$(4,18)$	$\{1,2,3,6,7\}$	$\{4\}$	$\{8,9\}$	$\{5\}$
$(4,30)$	$\{1,2,3,4,6\}$	$\{7\}$	$\{5\}$	$\{8,9\}$
$(4,31)$	$\{1,2,3,4\}$	$\{6,7\}$	$\{5\}$	$\{8,9\}$
$(4,32)$	$\{1,2,3,4,6\}$	$\{7\}$	$\{5\}$	$\{8,9\}$
$(4,33)$	$\{1,2,3,4\}$	$\{6,7\}$	$\{5\}$	$\{8,9\}$
$(4,34)$	$\{1,2,3,4,6\}$	$\{7\}$	$\{5\}$	$\{8,9\}$
$(4,35)$	$\{1,2,3,4\}$	$\{6,7\}$	$\{5\}$	$\{8,9\}$
$(5,17)$	$\{1,2,3,6,7\}$	$\{4\}$	$\{8,9\}$	$\{5\}$
$(5,18)$	$\{1,2,3,6,7\}$	$\{4\}$	$\{8,9\}$	$\{5\}$
$(5,30)$	$\{1,2,3,4,6\}$	$\{7\}$	$\{5\}$	$\{8,9\}$
$(5,31)$	$\{1,2,3,4\}$	$\{6,7\}$	$\{5\}$	$\{8,9\}$
$(5,32)$	$\{1,2,3,4,6\}$	$\{7\}$	$\{5\}$	$\{8,9\}$
$(5,33)$	$\{1,2,3,4\}$	$\{6,7\}$	$\{5\}$	$\{8,9\}$
$(5,34)$	$\{1,2,3,4,6\}$	$\{7\}$	$\{5\}$	$\{8,9\}$
$(5,35)$	$\{1,2,3,4\}$	$\{6,7\}$	$\{5\}$	$\{8,9\}$
$(9,16)$	$\{2,3,4\}$	$\{6,7\}$	$\{1\}$	$\{5,8,9\}$
$(9,17)$	$\{2,3,6,7\}$	$\{4\}$	$\{1,8,9\}$	$\{5\}$
$(9,18)$	$\{2,3,6,7\}$	$\{4\}$	$\{1,8,9\}$	$\{5\}$
$(9,19)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$
$(9,20)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$

$(i,j) \in \mathscr{P}$	$Q(00)$	$Q(01)$	$Q(10)$	$Q(11)$
$(9,22)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$
$(9,23)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$
$(9,24)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$
$(9,26)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$
$(9,27)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$
$(9,28)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$
$(9,29)$	$\{3,4,6,7\}$	$\{2\}$	$\{5,8,9\}$	$\{1\}$
$(9,30)$	$\{2,3,4,6\}$	$\{7\}$	$\{1,5\}$	$\{8,9\}$
$(9,31)$	$\{2,3,4\}$	$\{6,7\}$	$\{1,5\}$	$\{8,9\}$
$(9,32)$	$\{2,3,4,6\}$	$\{7\}$	$\{1,5\}$	$\{8,9\}$
$(9,33)$	$\{2,3,4\}$	$\{6,7\}$	$\{1,5\}$	$\{8,9\}$
$(9,34)$	$\{2,3,4,6\}$	$\{7\}$	$\{1,5\}$	$\{8,9\}$
$(9,35)$	$\{2,3,4\}$	$\{6,7\}$	$\{1,5\}$	$\{8,9\}$
$(9,37)$	$\{2,3,4\}$	$\{6,7\}$	$\{1\}$	$\{5,8,9\}$
$(11,17)$	$\{1,2,3,6,7,9\}$	$\{4\}$	$\{8\}$	$\{5\}$
$(11,18)$	$\{1,2,3,6,7,9\}$	$\{4\}$	$\{8\}$	$\{5\}$
$(11,30)$	$\{1,2,3,4,6\}$	$\{7,9\}$	$\{5\}$	$\{8\}$
$(11,31)$	$\{1,2,3,4\}$	$\{6,7,9\}$	$\{5\}$	$\{8\}$
$(11,32)$	$\{1,2,3,4,6\}$	$\{7,9\}$	$\{5\}$	$\{8\}$
$(11,33)$	$\{1,2,3,4\}$	$\{6,7,9\}$	$\{5\}$	$\{8\}$
$(11,34)$	$\{1,2,3,4,6\}$	$\{7,9\}$	$\{5\}$	$\{8\}$
$(11,35)$	$\{1,2,3,4\}$	$\{6,7,9\}$	$\{5\}$	$\{8\}$
$(12,17)$	$\{1,2,3,6,7,9\}$	$\{4\}$	$\{8\}$	$\{5\}$
$(12,18)$	$\{1,2,3,6,7,9\}$	$\{4\}$	$\{8\}$	$\{5\}$
$(12,30)$	$\{1,2,3,4,6\}$	$\{7,9\}$	$\{5\}$	$\{8\}$
$(12,31)$	$\{1,2,3,4\}$	$\{6,7,9\}$	$\{5\}$	$\{8\}$
$(12,32)$	$\{1,2,3,4,6\}$	$\{7,9\}$	$\{5\}$	$\{8\}$
$(12,33)$	$\{1,2,3,4\}$	$\{6,7,9\}$	$\{5\}$	$\{8\}$
$(12,34)$	$\{1,2,3,4,6\}$	$\{7,9\}$	$\{5\}$	$\{8\}$
$(12,35)$	$\{1,2,3,4\}$	$\{6,7,9\}$	$\{5\}$	$\{8\}$
$(16,17)$	$\{1,2,3\}$	$\{4\}$	$\{6,7,8,9\}$	$\{5\}$
$(16,18)$	$\{1,2,3\}$	$\{4\}$	$\{6,7,8,9\}$	$\{5\}$
$(16,36)$	$\{1,2\}$	$\{3,4\}$	$\{5,7,8,9\}$	$\{6\}$
$(17,36)$	$\{1,2,7,8,9\}$	$\{3,6\}$	$\{5\}$	$\{4\}$
$(17,37)$	$\{1,2,3\}$	$\{6,7,8,9\}$	$\{4\}$	$\{5\}$
$(18,36)$	$\{1,2,7,8,9\}$	$\{3,6\}$	$\{5\}$	$\{4\}$
$(18,37)$	$\{1,2,3\}$	$\{6,7,8,9\}$	$\{4\}$	$\{5\}$
$(31,36)$	$\{1,2,5\}$	$\{3,4\}$	$\{7,8,9\}$	$\{6\}$
$(33,36)$	$\{1,2,5\}$	$\{3,4\}$	$\{7,8,9\}$	$\{6\}$
$(35,36)$	$\{1,2,5\}$	$\{3,4\}$	$\{7,8,9\}$	$\{6\}$
$(36,37)$	$\{1,2\}$	$\{5,7,8,9\}$	$\{3,4\}$	$\{6\}$

We now apply our reduction algorithm and demonstrate its usefulness for the present example. Essentially incompatible pairs in $\mathscr{P}$ are

$$\mathscr{E}_0 = \{(3,4),(9,16),(16,17),(35,36),(36,37)\}.$$

Note that their corresponding open intervals are pair-wise disjoint and hence we need at least one recombination event for each incompatible pair in $\mathscr{E}_0$. So, we know that we need at least 5 recombination events. Upon choosing

$$R^{3,4} = \{\underline{5}\} \ , \ R^{9,16} = \{\underline{1}\} \ , \ R^{16,17} = \{\underline{5}\} \ , \ R^{35,36} = \{\underline{6}\} \ , \ R^{36,37} = \{\underline{6}\} \ ,$$

we obtain

$$\begin{aligned}
\mathscr{X}_0^{3,4} &= \{(1,11),(1,12),(2,11),(2,12),(3,5),(3,11),(3,12),(3,16),(3,37)\} \ , \\
\mathscr{X}_0^{9,16} &= \{(3,19),(3,20),(3,22),(3,23),(3,24),(3,26),(3,27),(3,28),(3,29),(9,19),(9,20), \\
&\quad (9,22),(9,23),(9,24),(9,26),(9,27),(9,28),(9,29),(9,37)\} \ , \\
\mathscr{X}_0^{16,17} &= \{(4,17),(4,18),(4,30),(4,31),(4,32),(4,33),(4,34),(4,35),(5,17),(5,18), \\
&\quad (5,30),(5,31),(5,32),(5,33),(5,34),(5,35),(9,17),(9,18),(11,17),(11,18), \\
&\quad (11,30),(11,31),(11,32),(11,33),(11,34),(11,35),(12,17),(12,18),(12,30), \\
&\quad (12,31),(12,32),(12,33),(12,34),(12,35),(16,18)\} \ , \\
\mathscr{X}_0^{35,36} &= \{(3,36),(16,36),(31,36),(33,36)\} \ , \\
\mathscr{X}_0^{36,37} &= \varnothing \ ,
\end{aligned}$$

and $\mathscr{X}_0 = \mathscr{X}_0^{3,4} \cup \mathscr{X}_0^{9,16} \cup \mathscr{X}_0^{16,17} \cup \mathscr{X}_0^{35,36} \cup \mathscr{X}_0^{36,37}$. Furthermore, we have

$$\mathscr{Y}_0 = \{(9,30),(9,31),(9,32),(9,33),(9,34),(9,35)\} \ ,$$

since the above pairs have $Q(10) = \{\underline{1},\underline{5}\} = R^{9,16} \cup R^{16,17}$ and their corresponding open intervals contain $I(9,16)$ and $I(16,17)$. In the next step of recursion, we have $\mathscr{E}_1 = \{(3,9),(18,36)\}$, and upon choosing $R^{3,9} = \{\underline{1},\underline{5}\}$ and $R^{18,36} = \{\underline{5}\}$ we obtain

$$\mathscr{X}_1 = \{(17,36),(17,37),(18,37)\} \ , \ \mathscr{Y}_1 = \varnothing \ .$$

Finally, $\mathscr{E}_2 = \varnothing$ and therefore the recursion terminates here. Note that $\mathscr{H}_2 = \bigcup_{n=0}^{1}(\mathscr{E}_n \cup \mathscr{X}_n \cup \mathscr{Y}_n) = \mathscr{P}$. To recapitulate, we have

$$\mathscr{E} = \mathscr{E}_0 + \mathscr{E}_1 = \{(3,4),(9,16),(16,17),(35,36),(36,37),(3,9),(18,36)\} \ ,$$

and

$$\begin{aligned}
&R^{3,4} = \{\underline{5}\} \ , \ R^{9,16} = \{\underline{1}\} \ , \ R^{16,17} = \{\underline{5}\} \ , \ R^{35,36} = \{\underline{6}\} \ , \ R^{36,37} = \{\underline{6}\} \ , \\
&R^{3,9} = \{\underline{1},\underline{5}\} \ , \ R^{18,36} = \{\underline{5}\} \ .
\end{aligned}$$

An assignment $\mathcal{A}$ that satisfies Properties (P1) and (P2) of §3.1 is

$$\begin{aligned}
\mathcal{A}(\{\underline{1}\}) &= I(8,9) \cup I(15,16) \ , \\
\mathcal{A}(\{\underline{5}\}) &= I(3,4) \cup I(16,17) \cup I(18,19) \ , \\
\mathcal{A}(\{\underline{6}\}) &= I(35,36) \cup I(36,37) \ , \\
\mathcal{A}(X) &= \varnothing \ \text{ for all other } X \in 2^S \ ,
\end{aligned}$$

from which we obtain

$$\sum_{X \in 2^S} \widetilde{\omega}(X;\mathcal{A};\mathcal{M}) \;=\; \widetilde{\omega}(\{\underline{1}\};\mathcal{A};\mathcal{M}) + \widetilde{\omega}(\{\underline{5}\};\mathcal{A};\mathcal{M}) + \widetilde{\omega}(\{\underline{6}\};\mathcal{A};\mathcal{M}) \;=\; 2+3+2=7\ .$$

We have checked that there exists no $(\mathcal{A},\mathcal{M}) \in \mathfrak{A} \times \mathfrak{M}$ which gives $\sum_{X \in 2^S} \widetilde{\omega}(X;\mathcal{A};\mathcal{M})$ less than 7. Hence, our new lower bound is $\mathcal{R}_{\mathrm{Q}}(S) = 7$. In fact, it is possible to construct an explicit graphical representation with only 7 recombination events.

Hudson and Kaplan's algorithm gives the incompatible pairs in $\mathscr{E}_0$ and therefore yields $\mathcal{R}_{\mathrm{HK}}(S) = |\mathscr{E}_0| = 5$.

5 Concluding Remarks

In this paper we have used set theoretical ideas to study recombination events. We have seen that the total number of recombination events depends on which subsets of S undergo recombination. For instance, we have demonstrated in Example 2.1 that even a character site which is compatible with all other character sites can affect the number of recombination events. We have shown that it is important to investigate under what conditions a particular recombination event can explain more than one incompatibility in the data. More generally, it is important to understand how a set of incompatibilities can share common recombination events.

In our approach, counting the number of recombination events can be translated to counting certain subsets of S with weights. In defining $\mathcal{R}_{\mathrm{Q}}(S)$ we have given simple weights to the subsets of S, i.e. the weight of $X \in 2^S$ is set equal to the number of times it is assigned on the interval $I(1,h)$. As indicated in the proof of Proposition 3.1, to obtain a more refined lower bound on the number of recombination events, we need to assign weights which better reflect the number of necessary recombination events.

A computer implementation of finding $\mathcal{R}_{\mathrm{Q}}(S)$ can be slow and inefficient when there are too many incompatibilities. We can, however, obtain *local* lower bounds using our definition of $\mathcal{R}_{\mathrm{Q}}$ and apply Myers-Griffiths algorithm [MG] to produce a *global* lower bound. That combined method should work very well. Furthermore, as we have discussed in §2, we can also use set theoretical results similar to Proposition 2.1 to find local lowers bounds and then apply Myers-Griffiths algorithm.

We stress that, although the present paper provides lower bounds $\mathcal{R}_{\mathrm{Q}}(S)$ which are sharper than previous lower bounds, $\mathcal{R}_{\mathrm{Q}}(S)$ might still be less than the minimum $\mathcal{R}_{\min}(S)$ for some cases of S. Constructing an algorithm which finds the minimum $\mathcal{R}_{\min}(S)$ for arbitrary S still remains an interesting open problem. Also, it would be interesting to construct an algorithm, however slow, which attempts to find a minimal evolutionary history—with the minimum number of recombination events—for a set of DNA sequences. Hein has provided a heuristic algorithm to achieve that goal, but has introduced restrictions on the problem to make it tractable [H]. As a consequence his method proposes histories which have internal contradictions and which therefore cannot not be realised by any set of sequences.

Acknowledgments
We gratefully acknowledge Carsten Wiuf for valuable discussions and suggestions. We also thank Gerton Lunter, Rune Lyngsø and Simon Myers for helpful comments on the manuscript. Y.S.S. would like to acknowledge the Mathematical Institute at the University of Oxford for hospitality while this work was carried out.

A Hudson and Kaplan's Algorithm

We here describe the algorithm given in [HK] for obtaining the lower bound $\mathcal{R}_{\text{HK}}(S)$. The algorithm begins with a linearly ordered set of pairs of incompatible character sites and removes from it certain elements which satisfy a set of specified conditions. $\mathcal{R}_{\text{HK}}(S)$ is defined to be the cardinality of the set obtained after all the required removals. More precisely, the algorithm can be rephrased as follows:

1. Given a set S of binary sequences, to each pair (i, j) of incompatible character sites i and j, where $i < j$, associate an open interval $I(i, j) = \{x \in \mathbb{R} \mid i < x < j\}$.

2. Order the above pairs of incompatible character sites and define the linearly ordered set
$$\mathcal{P} = \{(i_1, j_1), (i_2, j_2), \ldots, (i_p, j_p)\},$$
where p is the total number of distinct pairs of incompatible sites. The ordering is defined so that $i_1 \leq i_2 \leq \cdots \leq i_p$, and if $i_a = i_b$, then $j_a < j_b$.

3. Let $\mathcal{C} = \{(i, j) \in \mathcal{P} \mid I(i, j) \supsetneq I(k, l) \text{ for some } (k, l) \in \mathcal{P}\}$.

4. Let $\mathcal{P}_0 := \mathcal{P} \setminus \mathcal{C}$ and recursively remove certain elements from $\mathcal{P}_0$ as described below. For $n \in \mathbb{Z}^+$, $\mathcal{P}_n$ and $\mathcal{A}_n$ are recursively defined as follows:
Let $(\alpha_n, \beta_n) \in \mathcal{P}_{n-1}$ be the first element that satisfies

   (i) $I(\alpha_n, \beta_n) \cap I(i, j) = \varnothing$ for all $(i, j) \in \mathcal{P}_{n-1}$ with $i < \alpha_n$, and
   (ii) $I(\alpha_n, \beta_n) \cap I(k, l) \neq \varnothing$ for at least one $(k, l) \in \mathcal{P}_{n-1}$ with $k > \alpha_n$.

   Then, we define
$$\mathcal{A}_n := \{(i, j) \in \mathcal{P}_{n-1} \mid \alpha_n < i < \beta_n\}$$
and $\mathcal{P}_n := \mathcal{P}_{n-1} \setminus \mathcal{A}_n$.

   The recursion terminates at $n = r$ when $I(i, j) \cap I(k, l) = \varnothing$ for every distinct pair of elements $(i, j), (k, l) \in \mathcal{P}_r$.

5. Define $\mathcal{R}_{\text{HK}}(S) := |\mathcal{P}_r|$.

In effect $\mathcal{R}_{\text{HK}}(S)$ gives the maximum number of pairs (i, j) of incompatible sites whose associated open intervals $I(i, j)$ are pair-wise disjoint.

B Proofs of Set Theoretical Lemmas

In this appendix we provide proofs of the lemmas stated in §2.

Lemma 2.1 *Let i, j, k denote informative character sites, of which (i, j) is an incompatible pair, whereas (i, k) and (j, k) are compatible pairs. If either $B_k \cap B_i = \varnothing$ or $B_k \cap B_i^c = \varnothing$, then*

$$B_k^c \cap B_j \neq \varnothing \quad \text{and} \quad B_k^c \cap B_j^c \neq \varnothing.$$

Proof: Assume that $B_k \cap B_i = \varnothing$. If $B_k^c \cap B_j = \varnothing$, then it implies that $B_j \subseteq B_k$. Since by assumption $B_k \cap B_i = \varnothing$, we then conclude that $B_i \cap B_j = \varnothing$. But, this contradicts the condition that (i, j) is an incompatible pair. Hence, $B_k^c \cap B_j \neq \varnothing$ if $B_k \cap B_i = \varnothing$. The same argument applies to the remaining cases. ■

Lemma 2.2 *Let i, j, k be defined as in Lemma 2.1. Then, exactly one of the following intersections is empty:*

$$B_i \cap B_k \ , \ B_i \cap B_k^c \ , \ B_i^c \cap B_k \ , \ B_i^c \cap B_k^c \ . \tag{B.1}$$

Proof: Since (i, k) is a compatible pair, by definition at least one of (B.1) is empty. We need to show that there cannot be more than one empty intersection. Appropriately name the bipartitions of i and k so that $B_i \cap B_k = \varnothing$. Then, since i and k are informative character sites, we must have $B_i \cap B_k^c \neq \varnothing$ and $B_i^c \cap B_k \neq \varnothing$. Now, observe that $B_i \cap B_k = \varnothing$ implies $B_k \subseteq B_i^c$, while $B_i^c \cap B_k^c = \varnothing$ implies $B_k^c \subseteq B_i$. Hence, if we have both $B_i \cap B_k = \varnothing$ and $B_i^c \cap B_k^c = \varnothing$, then $B_i = B_k^c$ and $B_i^c = B_k$. But, this contradicts the assumption that (j, i) is an incompatible pair, whereas (j, k) is a compatible pair. ■

Lemma 2.3 *Let i, j, k be as in Lemma 2.1. Then, the quadpartition $\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ of S contains exactly one Q_a^{ij} such that either*[7]

(a) $B_k \cap Q_a^{ij} \neq \varnothing$ *and* $B_k \cap (\ S \setminus Q_a^{ij}\ ) = \varnothing$,

or (b) $B_k^c \cap Q_a^{ij} \neq \varnothing$ *and* $B_k^c \cap (\ S \setminus Q_a^{ij}\ ) = \varnothing$.

In other words, up to a choice of relabelling, there exists a unique $Q_a^{ij} \in \{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ such that either $B_k \subseteq Q_a^{ij}$ or $B_k^c \subseteq Q_a^{ij}$.

Proof: By Lemma 2.2, exactly one of the following intersections is empty:

$$B_i \cap B_k \ , \ B_i \cap B_k^c \ , \ B_i^c \cap B_k \ , \ B_i^c \cap B_k^c \ . \tag{B.2}$$

Likewise, exactly one of

$$B_j \cap B_k \ , \ B_j \cap B_k^c \ , \ B_j^c \cap B_k \ , \ B_j^c \cap B_k^c \tag{B.3}$$

is empty. Combining these facts with the result in Lemma 2.1, we conclude that if either $B_k \cap B_i = \varnothing$ or $B_k \cap B_i^c = \varnothing$, then either $B_k \cap B_j = \varnothing$ or $B_k \cap B_j^c = \varnothing$. Similarly, if either

[7] Let A be a subset of the set X. Then, the notation $X \setminus A$ denotes the complement of A relative to X.

$B_k^c \cap B_i = \varnothing$ or $B_k^c \cap B_i^c = \varnothing$, then either $B_k^c \cap B_j = \varnothing$ or $B_k^c \cap B_j^c = \varnothing$.

Now, up to simple change of notation, we can assume that the empty intersections in (B.2) and (B.3) are $B_k \cap B_i^c = \varnothing$ and $B_k \cap B_j^c = \varnothing$, respectively. Then, with $Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}$ defined as

$$Q_1^{ij} := B_i \cap B_j \ , \ Q_2^{ij} := B_i \cap B_j^c \ , \ Q_3^{ij} := B_i^c \cap B_j \ , \ Q_4^{ij} := B_i^c \cap B_j^c, \tag{B.4}$$

we see that

$$B_k \cap Q_2^{ij} = \varnothing \ , \ B_k \cap Q_3^{ij} = \varnothing \ , \ B_k \cap Q_4^{ij} = \varnothing \ .$$

Therefore, $B_k \subseteq Q_1^{ij} = S \setminus (Q_2^{ij} \cup Q_3^{ij} \cup Q_4^{ij})$ and we thus have $B_k \cap Q_1^{ij} \neq \varnothing$. Furthermore, up to simple renaming, the uniqueness of Q_a^{ij} with the desired properties follows from the uniqueness of empty intersections in (B.2) and (B.3). ■

Lemma 2.4 *Let i, j, k, l be informative character sites such that (i, j) and (k, l) are incompatible pairs, whose corresponding quadpartitions are $\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ and $\{Q_1^{kl}, Q_2^{kl}, Q_3^{kl}, Q_4^{kl}\}$, respectively; and $(i, k), (i, l), (j, k)$, and (j, l) are compatible pairs. Then, there exists a subset Q_a^{ij} which is unique, up to a choice of relabelling, such that*

$$S \setminus Q_b^{kl} \subseteq Q_a^{ij}$$

for exactly one index value $b \in \{1, 2, 3, 4\}$.

Proof: Suppose Q_a^{ij} is the unique subset such that $B_k \subseteq Q_a^{ij}$ after appropriate relabelling (c.f. Lemma 2.3). If $B_l \cap Q_a^{ij} = \varnothing$, then it implies that $B_l \subseteq (Q_a^{ij})^c$, which in turn implies that $B_k \cap B_l = \varnothing$. But, this cannot be since (k, l) is an incompatible pair. Hence, $B_k \subseteq Q_a^{ij}$ implies $B_l \cap Q_a^{ij} \neq \varnothing$. In a similar vein, one obtains that if $B_k \subseteq Q_a^{ij}$, then $B_l^c \cap Q_a^{ij} \neq \varnothing$. One reaches the same conclusion if B_k is replaced by B_k^c in the above argument. Hence, if either $B_k \subseteq Q_a^{ij}$ or $B_k^c \subseteq Q_a^{ij}$, then $B_l \cap Q_a^{ij} \neq \varnothing$ and $B_l^c \cap Q_a^{ij} \neq \varnothing$. The latter statement means, according to Lemma 2.3, that either $B_l \subseteq Q_a^{ij}$ or $B_l^c \subseteq Q_a^{ij}$.

Let $\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ and $\{Q_1^{kl}, Q_2^{kl}, Q_3^{kl}, Q_4^{kl}\}$ be defined as in (B.4). Furthermore, let $\{B_k, B_k^c\}$ and $\{B_l, B_l^c\}$ be appropriately named so that $B_k \subseteq Q_1^{ij}$ and $B_l \subseteq Q_1^{ij}$. Then, we immediately conclude that

$$Q_1^{kl} \subset Q_1^{ij} \ , \ Q_2^{kl} \subset Q_1^{ij} \ , \ Q_3^{kl} \subset Q_1^{ij} \ ,$$

and therefore that $S \setminus Q_4^{kl} \subseteq Q_1^{ij}$. Since $Q_1^{kl}, Q_2^{kl}, Q_3^{kl}$ are all non-empty and disjoint, they are proper subsets of Q_1^{ij}. Up to a choice of notation, the uniqueness of Q_1^{ij} with the desired properties follows from Lemma 2.3. ■

Lemma 2.5 *Let i, j, k, l be defined as in Lemma 2.4. Then, there exist no $a, b \in \{1, 2, 3, 4\}$ such that*

$$Q_a^{ij} = Q_b^{kl} \ .$$

PROOF: Lemma 2.4 implies that there exists an index value c such that $(S \setminus Q_b^{kl}) \subset Q_c^{ij}$. Suppose that there exist a and b such that $Q_a^{ij} = Q_b^{kl}$. Then, we must $S = Q_a^{ij} \cup Q_c^{ij}$, from which we conclude that there exists an index value d such that $Q_d^{ij} = \varnothing$. This contradicts the fact that the quadpartition $\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ associated to the incompatible pair (i,j) consists of non-empty pair-wise disjoint subsets of S. ■

Lemma 2.6 *Let i, j, k denote informative character sites. Let (i,j) and (j,k) be incompatible pairs, and let (i,k) be a compatible pair. Then, there exists a proper subset $Q_a^{ij} \in \{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ and a proper subset $Q_b^{jk} \in \{Q_1^{jk}, Q_2^{jk}, Q_3^{jk}, Q_4^{jk}\}$ such that*

$$Q_b^{jk} \subset Q_a^{ij}.$$

PROOF: Appropriately name the bipartitions $\{B_i, B_i^c\}$ and $\{B_k, B_k^c\}$ so that $B_i \cap B_k = \varnothing$. This is always possible since (i,k) is a compatible pair. If we define $\{Q_1^{ij}, Q_2^{ij}, Q_3^{ij}, Q_4^{ij}\}$ and $\{Q_1^{jk}, Q_2^{jk}, Q_3^{jk}, Q_4^{jk}\}$ as in (B.4), then from $B_i \cap B_k = \varnothing$ we conclude that

$$\begin{array}{ccc} Q_1^{ij} \cap Q_1^{jk} = \varnothing & , & Q_1^{ij} \cap Q_3^{jk} = \varnothing \ , \\ Q_2^{ij} \cap Q_1^{jk} = \varnothing & , & Q_2^{ij} \cap Q_3^{jk} = \varnothing \ . \end{array}$$

Moreover, we have

$$Q_4^{ij} \cap Q_1^{jk} = \varnothing \quad , \quad Q_3^{ij} \cap Q_3^{jk} = \varnothing,$$

since $B_j \cap B_j^c = \varnothing$. Therefore, $Q_1^{jk} \subseteq Q_3^{ij} = S \setminus (Q_1^{ij} \cup Q_2^{ij} \cup Q_4^{ij})$ and $Q_3^{jk} \subseteq Q_4^{ij} = S \setminus (Q_1^{ij} \cup Q_2^{ij} \cup Q_3^{ij})$. ■

References

[H] Hein, Jotun, *A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination,* J.Mol.Evol. **20** (1993) 402-411.

[HK] Hudson, Richard R. and Kaplan, Norman L., *Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences*, Genetics **11** (1985) 147-164.

[K] Kreitman, M., *Nucleotide Polymorphism at the Alcohol Dehydrogenase Locus of Drosophila Melanogaster*, Nature **304** (1983) 412-417.

[MG] Myers, Simon R. and Griffiths, Robert C., *Bounds on the Minimum Number of Recombination Events in a Sample History,* to be published in Genetics.