ACADEMIC PRESS

Theoretical Population Biology 62 (2002) 357–363

Theoretical Population Biology

www.academicpress.com

On the minimum number of topologies explaining a sample of DNA sequences

Carsten Wiuf

Variagenics, 60 Hampshire Street, Cambridge, Massachusetts 02139-1548 USA

Received 25 March 2002

Abstract

In this article I derive an alternative algorithm to Hudson and Kaplan's (*Genetics* **111**, 147–165) algorithm that gives a lower bound to the number of recombination events in a sample's history. It is shown that the number, T_M, found by the algorithm is the least number of topologies required to explain a set of DNA sequences sampled under the infinite-site assumption. Let $\mathscr{T} = (T_1, \ldots, T_r)$ be a list of topologies compatible with the sequences, i.e., T_k is compatible with an interval, I_k, of sites in the alignment. A characterization of all lists having T_M topologies is given and it is shown that T_M relates to specific patterns in the alignment, here called chain series. Further, a number of theorems relating general lists of topologies to the number T_M is presented. The results are discussed in relation to the true minimum number of recombination events required to explain an alignment.

Keywords: Algorithm; Recombination; SNP; Topology

1. Introduction

Currently, there is a lot of interest in estimating recombination rates from population samples of DNA sequences or haplotyped SNP data. Recombination events can be inferred with certainty from data if specific patterns of mutation occur. For example, with binary sequences, an instance of all four gamete types 00, 01, 10, and 11 in two columns in the alignment is indicative of at least one recombination event in the sample's history. Here and later the possibility of recurrent mutations is ignored. The pattern provides a simple way of testing for the presence of recombination and was termed the four-gamete test by Hudson and Kaplan (1985). They went further and derived, based on the four-gamete test, a lower bound T_M (their number R_M) on the number of recombination events in a sample's history.

Hudson and Kaplan's (1985) T_M does not give the true minimum in most cases and other statistics than T_M can be shown to perform better. For example, the haplotype statistic explored by Myers (2002) always gives a number not less than T_M. However, one big advantage of T_M is that it is easy to implement and compute, even for large sample sizes and many segregating sites. The true minimum, on the other hand, is hard to find. The trivial way to find the true minimum is by exhaustive trial and error. This is truly inefficient, but no efficient algorithm is known. Other very similar, though not entirely equivalent, problems in graph theory are known to be NP-hard and thus not solvable by any efficient means (see for example Allen and Steel, 2001).

In this paper I consider a set of binary sequences and derive an alternative algorithm to that of Hudson and Kaplan. To do so I consider a related problem, namely the problem of finding a minimal set of trees that explain the sequences, such that any two neighbor trees are incompatible with each other. If this is so, there is at least one recombination event between two neighbor trees. As an example consider a sample with $n = 6$ sequences and $m = 7$ sites,

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

The sequences can be explained by three trees, namely one tree explaining sites 1 and 2, one tree explaining 3, 4,

E-mail address: cwiuf@variagenics.com (Carsten Wiuf).

0040-5809/02/$ - see front matter
PII: S0040-5809(02)00004-7

and 5, and one explaining 6 and 7. For each tree only one substitution event is required in each column and we say that the sites are compatible with the tree. As we shall see the minimum number of trees is one larger than T_M, that is $T_M = 2$. The haplotype bound is 4, because there are six distinct haplotypes derived from the sites 1–3 and similarly 6 for the sites 4–6 (Myers, 2002). Each requires at least $6-3-1=2$ recombination events, where 3 is the number of sites. I will characterize the set of all possible solutions to the problem of finding the minimum number of trees and show how this number relates to certain patterns in the alignment, called chain series. In addition, the number T_M is related to general lists of topologies that explain the sequences. In the discussion I will comment in more detail on T_M's relation to the true minimum number of recombination events.

2. Results

Consider a sample of n sequences with m segregating sites. Assume we have no information about the root, branch lengths or about which of the two sites is ancestral. In that case a tree is just a topology. Throughout the paper topologies are thought of as bifurcating, consistent with standard stochastic models of sequence evolution. However, this assumption is not at all crucial to the results. Each site i in the sample induces a bipartition (B_0, B_1), where B_k, $k=0,1$, is the set of sequences with k in site i. Let S_i, $i=1,\ldots,m$, be the class of bifurcating topologies, or topologies for short, compatible with site i. Further, assume an infinite-site model, that is each site has mutated at most once.

The only prerequisite we need is the Compatibility Theorem (Estabrook et al., 1975, see also Gusfield, 1991). Say two sites or columns in the alignment, i_1 and i_2, are incompatible if all four gamete types, 00, 01, 10, and 11 are present in the two columns. Otherwise, the two sites are said to be compatible. In the previous example, sites 1 and 2 are compatible, whereas 2 and 3 are not. Note that this second concept of compatibility compares sites with sites, whereas the first relates sites to trees. The Compatibility Theorem provides the relation between the two concepts.

The compatibility theorem. *The whole sample is compatible with a single topology if and only if all pairs of sites i_1 and i_2, $i_1, i_2 \in \{1,\ldots,m\}$, are compatible.*

This theorem is very strong because it allows us to check for compatibility with a tree by checking for pairwise compatibility of sites, which is a lot easier to do. In the following, it will be demonstrated how useful this is. The next lemma is a consequence of the Compatibility Theorem.

Lemma 1. *Let G_1 and G_2 be two non-empty classes of topologies such that $G_k = \bigcap_{i\in I_k} S_i$, $I_k \subseteq [1,\ldots,m]$, $k=1,2$. Then G_1 and G_2, are incompatible, i.e., $G_1 \cap G_2 = \emptyset$, if and only if there exist $i_k \in I_k$, $k=1,2$ such that $S_{i_1} \cap S_{i_2} = \emptyset$.*

Proof. Only the 'only if' needs to be proved. Assume the statement is not true, then $S_{i_1} \cap S_{i_2} \neq \emptyset$ for all $(i_1, i_2) \in I_1 \times I_2$. This implies that $S_i \cap S_j \neq \emptyset$ for all $(i,j) \in (I_1 \cup I_2)^2$, and further according to the Compatibility Theorem that all sites in $I_1 \cup I_2$ can be related by a tree. This contradicts that G_1 and G_2 are incompatible. □

Definition 1. A list of topology classes, $\mathscr{T} = (T_1, T_2, \ldots, T_r)$ is said to be compatible with the sample if $T_k = \bigcap_{i\in I_k} S_i \neq \emptyset$ for $k=1,\ldots,r$, $I_k = [i_k, i_{k+1}-1]$, $i_k < i_{k+1}$, $i_1 = 1$, and $i_{r+1} = m+1$. The list $\mathscr{T}$ is said to be disjoint if $T_k \cap T_{k+1} = \emptyset$ for $k=1,\ldots,r-1$, and $\mathscr{T}$ is said to be minimal if r is as small as possible. The smallest r is denoted by T_M.

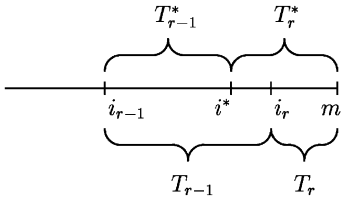
Obviously, $\mathscr{T}$ cannot be minimal without being disjoint. Also note that in the definition of disjointness it is only required that $T_i \cap T_j = \emptyset$ for neighbor classes $j = i+1$. In the following $\mathscr{T}$ is assumed to be compatible with the sample. Further, the notation $\mathscr{T}$ and $\mathscr{I} = (I_1, \ldots, I_r)$ are used interchangeably.

Definition 2. Let $\mathscr{T}$ be a disjoint list with r elements. A series $\alpha_k < \beta_k \leqslant \alpha_{k+1}$, $\alpha_k \in I_k$, $\beta_k \in I_{k+1}$, for $k=1,\ldots,r-1$, with the property that $S_{\alpha_k} \cap S_{\beta_k} = \emptyset$, i.e., sites α_k and β_k are incompatible, is called a chain series.

Theorem 1. *Let $\mathscr{T}$ be a disjoint list with r elements. Then $\mathscr{T}$ is minimal if and only if $\mathscr{T}$ admits a chain series.*

Proof. If $\mathscr{T}$ admits a chain series then there cannot be another compatible list with fewer elements than r, i.e., $T_M \geqslant r$, and further because $\mathscr{T}$ is compatible by assumption, $r \geqslant T_M$. That is $T_M = r$ and $\mathscr{T}$ is minimal. Assume now $\mathscr{T}$ is minimal. Induction will be used to prove that $\mathscr{T}$ admits a chain series. If $r=2$, then, there exists a chain series (Lemma 1). For the induction step assume $\mathscr{T}$ admits a chain series if there are at most $r-1$ elements in $\mathscr{T}$. If $\mathscr{T}$ has r elements construct a new list $\mathscr{T}^* = (T_1^*, \ldots, T_r^*)$ such that $T_k^* = T_k$ for $k < r-1$ and T_{r-1}^* and T_r^* are defined in the following way. Let i^* be the largest element in $I_{r-1} = [i_{r-1}, i_r - 1]$ such that S_{i^*} is incompatible with T_r. Such an element exists. Define T_{r-1}^* and T_r^* by $T_{r-1}^* = \bigcap_{i=i_{r-1}}^{i^*} S_i$ and $T_r^* = \bigcap_{i=i^*+1}^{m} S_i$. Note that T^* is compatible with the sample and that $(T_1^*, \ldots, T_{r-1}^*)$ is minimal for the sample restricted to the sites $1, \ldots, i^*$ (see illustration). Otherwise $\mathscr{T}$ could not be

minimal either.



By the induction hypothesis there exists a chain series, C, for $(T^*_1, \ldots, T^*_{r-1})$ and further there exists a site $\beta_{r-1} \in I_r$ such that i^* and β_{r-1} are incompatible (by definition of i^*). Extending C with $\alpha_{r-1} \equiv i^*$ and β_{r-1} gives a chain series for $\mathscr{T}$. □

Example 1. A list can be disjoint without being minimal and hence not admit a chain series. For example consider the following five sequences with four segregating sites, 0000, 0001, 0111, 1100, and 1111. The pairs of sites $(1,3)$, $(2,4)$ and $(1,4)$ are all incompatible, and one can construct a disjoint list of three classes, $T_1 = S_1$, $T_2 = S_2 \cap S_3$, and $T_3 = S_4$. However, there is no chain because 1 is only incompatible with 3 in T_2, and 4 only with 2. A minimal list is $\mathscr{T} = (S_1 \cap S_2, S_3 \cap S_4)$ which has three chains, $(\alpha^1_1, \beta^1_1) = (1,3)$, $(\alpha^2_1, \beta^2_1) = (2,4)$, and $(\alpha^3_1, \beta^3_1) = (1,4)$. It is clear that four is the least possible number of sites for which an example of this kind can be found. Neither can an example be found with four sequences because this would force sites 2 and 3 to have the same mutation pattern. Thus, an example with fewer sites or fewer sequences cannot be constructed.

Algorithms that find the minimum number of topology classes, T_M, or produce a minimal list are of interest. Two such algorithms are given here.

Algorithm 1. *Define* $\mathscr{T}_i = (T^i_1, T^i_2, \ldots, T^i_{r_i})$, $i = 1, 2, \ldots, m$, *recursively by*

(1) $\mathscr{T}_1 = (T^1_1) = (S_1)$ *and* $r_1 = 1$,
(2) *If* S_{i+1} *is compatible with* $T^i_{r_i}$ *then* $r_{i+1} = r_i$ *and*

$$\begin{aligned}\mathscr{T}_{i+1} &= (T^{i+1}_1, T^{i+1}_2, \ldots, T^{i+1}_{r_{i+1}}) \\ &= (T^i_1, T^i_2, \ldots, T^i_{r_i} \cap S_{i+1}),\end{aligned}$$

(3) *If* S_{i+1} *is incompatible with* $T^i_{r_i}$ *then* $r_{i+1} = r_i + 1$ *and*

$$\mathscr{T}_{i+1} = (T^{i+1}_1, T^{i+1}_2, \ldots, T^{i+1}_{r_{i+1}}) = (T^i_1, T^i_2, \ldots, T^i_{r_i}, S_{i+1}).$$

Then $\mathscr{T}_i$ *is minimal for the sample restricted to the !sites* $1, 2, \ldots, i$ *and the minimal number of classes is* r_i, *in particular*, $\mathscr{T}_m$ *is minimal for the sample and* $T_M = r_m$.

Note that $\mathscr{T}_m$ has the property that a new topology class is only postulated for site i if S_i is incompatible with the previous class.

Proof. $\mathscr{T}_r$ is obviously disjoint. Define $\beta_k = i_{k+1}$, $k = 1, \ldots, r_m - 1$, that is β_k is the left most site in topology class T_{k+1}. Choose $\alpha_k \in T_k$ such that α_k and β_k are incompatible. This is possible because β_k and T_k are incompatible (by construction). Clearly, $\beta_k \leqslant \alpha_{k+1}$ because β_k is the left-most site in T_{k+1} and $\alpha_k < \beta_k$ by construction. In conclusion, $\mathscr{T}_r$ admits a chain series and is thus minimal. □

Hudson and Kaplan (1985) provide a different algorithm. The following algorithm is due to them.

Algorithm 2. *Define* D_{ij} *to be one if sites* i *and* j *are incompatible, and zero otherwise. Order all intervals* $[i,j]$, $i<j$, *for which* $D_{ij} = 1$ *into a list,* Δ, *alphabetically. With a slight abuse of language,* $[i_1,j_1]$ *and* $[i_2,j_2]$ *are said to be disjoint if* $j_1 \leqslant i_2$. *Do the following*:

(1) *If* $[i_1,j_1]$, $[i_2,j_2] \in \Delta$ *and* $i_1 \leqslant i_2 < j_2 \leqslant j_1$, *then remove* $[i_1,j_1]$ *from the list, i.e.,* $\Delta := \Delta \setminus [i_1,j_1]$. *Continue until it is not possible to remove any more intervals.*
(2) *Let* $[i,j]$ *be the first interval in* Δ *not disjoint from all the other intervals. Remove all intervals,* $[i_1,j_1]$, *such that* $i < i_1 < j$. *Continue in the same fashion with the next interval that is not disjoint from all the other intervals in the updated list, and so forth.*
(3) *Stop when all intervals in* Δ *are disjoint.*

Let $\Delta^* = ([i_1,j_1], \ldots, [i_{r-1},j_{r-1}])$ *be the final disjoint list ordered alphabetically. Let* $j_0 = 1$, $j_r = m+1$ *and define* $T_k = \bigcap_{i=j_{k-1}}^{j_k - 1} S_i$, $k = 1, 2, \ldots, r$. *Then* $\mathscr{T} = (T_1, T_2, \ldots, T_r)$ *is minimal and* $T_M = |\Delta^*| + 1 = r$.

Comments. Two comments are appropriate: (i) The first step creates a unique list with the property that either $j_1 \leqslant i_2$ or $i_1 < i_2 < j_1 < j_2$ for any two intervals $[i_1,j_1]$ and $[i_2,j_2]$ in Δ with $i_1 \leqslant i_2$. If this was not the case one could continue to remove intervals. (ii) If an interval $[i,j]$ is removed from the list, then there is an interval in the final list, Δ^*, that is not disjoint from $[i,j]$. The comments are used in the proof below.

Proof. First, it will be shown that $T_k \neq \emptyset$ for $k = 1, \ldots, r$. Put $I_k = [j_{k-1}, j_k - 1]$. According to the Compatibility Theorem it suffices to prove that all sites in I_k are pairwise compatible. Assume oppositely that there exists a pair of sites $i,j \in I_k$, $i<j$, such that i and j are incompatible. Note that $i_k \in I_k$. Then there are three possibilities: Either (i) $i_k \leqslant i$, (ii) $i < i_k < j$, or (iii) $j \leqslant i_k$. If (i), then (i_k, j_k) should be removed from Δ, according to rule (1). If (ii), then again (i_k, j_k) should be removed from Δ because (i,j) occurs before (i_k, j_k) in application of rule (2). If iii) then (i,j) should appear in the final list, Δ^*, because (i,j) is disjoint from all the pairs listed in Δ^*. In all cases a contradiction is reached and it is concluded that the sites in I_k are compatible with a tree.

Next, it will be shown that $\mathscr{T}$ is disjoint and admits a chain series. If this is so, $\mathscr{T}$ is minimal and $T_M = r = |\Delta^*| + 1$. Define $\alpha_k = i_k$ and $\beta_k = j_k$ for $k = 1, \ldots, r-1$. Then, $\alpha_k < \beta_k \leqslant \alpha_{k+1}$ (according to the Comments) and (α_k, β_k) is an incompatible pair of sites, because $D_{\alpha_k \beta_k} = 1$. Further, note that $\alpha_k \in I_k$ and $\beta_k \in I_{k+1}$. Then, $\alpha_k, \beta_k,\ k = 1, \ldots, r-1$, form a chain series and $T_M = |\Delta^*| + 1 = r$. The proof is completed. □

It transpires by inspection of the rules in Algorithm 2 that the list $\mathscr{T}$ as defined in Algorithm 2 is identical to the list defined in Algorithm 1. This provides an alternative proof of the fact that Algorithm 2 produces a minimal list.

Algorithm 1 runs linearly in the number of segregating sites, m. There are at most $2n-3$ non-empty splits for a given topology with n leaves ($n-3$ if singletons are not counted) and it takes at most $O(n)$ operations to compare two splits. So, Algorithm 1 runs in time $O(mn \min\{m,n\})$ ($\min\{m,n\}$ because there are at most m different splits). Algorithm 2 runs quadratic in m, because there are $m(m-1)/2$ pairs of sites, and for each pair it takes at most $O(n)$ operations to check whether the pair is incompatible or not. The list can be sorted at the same time, in total in $O(nm^2)$. Reducing the list can be done in $O(m^2)$, going through the list twice, once for each of the steps 1 and 2. Thus, Algorithm 2 runs in time $O(nm^2)$. If $m \gg n$, Algorithm 1 results in a substantial reduction in computer time compared to Algorithm 2.

Gusfield (1991) founds an $O(nm)$ algorithm that decides whether the sample conforms to a single topology and if it does constructs this. It does not seem that his technique can improve the running time of the two algorithms discussed here.

Theorem 2. *Assume $\mathscr{T}_1 = (T_1^1, \ldots, T_r^1)$ is minimal with chain series $\alpha_k, \beta_k,\ k = 1, \ldots, 1-r$, and let $\mathscr{T}_2 = (T_1^2, \ldots, T_s^2)$, $s \geqslant r$, be another compatible list, not necessarily disjoint. Then $\alpha_1, \beta_1, \beta_2, \ldots, \beta_{r-1}$ fall in r different classes of $\mathscr{T}_2$. If $\mathscr{T}_2$ is minimal, i.e., $s = r$, then β_k and $\alpha_{k+1},\ k = 1, \ldots, r-2$, fall in the same class and $\alpha_k, \beta_k,\ k = 1, \ldots, r-1$, is a chain series of $\mathscr{T}_2$.*

Proof. Since $\alpha_k, \beta_k,\ k = 1, \ldots, r-1$, is a chain then

$$\alpha_1 < \beta_1 \leqslant \alpha_2 < \beta_2 \leqslant \cdots \leqslant \alpha_{r-1} < \beta_{r-1}.$$

The two sites, α_1 and β_1, fall in different classes because they are incompatible. Also β_k and $\beta_{k+1},\ k = 1, \ldots, r-2$, fall in different classes. Otherwise α_{k+1} would also fall in the same class as β_k and β_{k+1} because $\beta_k \leqslant \alpha_{k+1} < \beta_{k+1}$. But this is impossible because α_{k+1} and β_{k+1} are incompatible. Hence, $\alpha_1, \beta_1, \beta_2, \ldots, \beta_{r-1}$ fall in r different classes. If $\mathscr{T}_2$ is minimal, then there are exactly r classes in $\mathscr{T}_2$, implying that α_{k+1} falls in the same class as β_k. Hence, $\alpha_k,\ \beta_k,\ k = 1, \ldots, r-1$, is a chain series of $\mathscr{T}_2$. The proof is completed. □

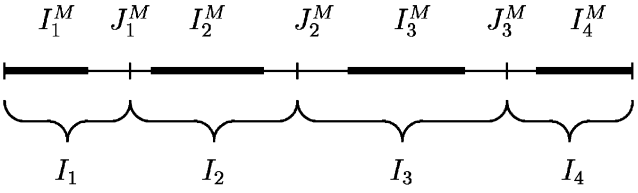
Assume that $T_M = r$. Define $\alpha_k^{\max}$ and $\beta_k^{\min}$, $k = 1, \ldots, r-1$, by

$$\alpha_k^{\max} = \max\{\alpha_k \mid \alpha_j, \beta_j,\ j = 1, \ldots, r-1,\ \text{is a chain}\} \quad (1)$$

and

$$\beta_k^{\min} = \min\{\beta_k \mid \alpha_j, \beta_j,\ j = 1, \ldots, r-1,\ \text{is a chain}\}. \quad (2)$$

Because the number of sites, m, is finite, $\alpha_k^{\max}$ and $\beta_k^{\min}$ both belong to a chain series (though not necessarily the same chain series, Example 1). Put $\beta_0^{\min} = 1$ and $\alpha_r^{\max} = m$, and let $I_k^M = [\beta_{k-1}^{\min}, \alpha_k^{\max}]$ and $J_k^M = [\alpha_k^{\max} + 1, \beta_k^{\min} - 1]$ (J_k^M can be empty), $k = 1, \ldots, r$. An illustration of the relation between a minimal list $\mathscr{T}$ (or $\mathscr{I}$) and I_k^M and J_k^M is useful. Below $r = 4$, thick lines represent I_k^M, thin lines J_k^M, and the boundary between I_k and I_{k+1} is marked with a vertical thin line. Note that Theorem 2 implies that $I_k^M \subseteq I_k$ for all k. The next lemma relates $\alpha_k^{\max}$ and $\beta_k^{\min}$ to the minimal list constructed as in Algorithm 1.



Lemma 2. *Let $\mathscr{T} = (T_1, \ldots, T_r)$ be the list constructed as in* Algorithm 1 *and let $\mathscr{T}^* = (T_1^*, \ldots, T_r^*)$ be the list constructed as in* Algorithm 1, *when applied to the reversed set of sequences (numbering of topologies is from left to right, i.e., not reversed). Define $\beta_k,\ k = 1, \ldots, r-1$, to be the first element in T_{k+1} and, similarly, define $\alpha_k,\ k = 1, \ldots, r-1$, to be the last element in T_k^* (the first element when going from right to left). Then $\beta_k = \beta_k^{\min}$ and $\alpha_k = \alpha_k^{\max}$.*

Proof. According to Theorem 2, $\beta_k^{\min} \in T_{k+1}$, because $\beta_k^{\min}$ is at least in one chain series. But β_k is the first element in T_{k+1}, and thus $\beta_k \leqslant \beta_k^{\min}$. However, by definition of $\beta_k^{\min}$ and construction of β_k, $\beta_k^{\min} \leqslant \beta_k$, hence $\beta_k^{\min} = \beta_k$. Obviously, $\mathscr{T}^*$ is also a minimal list with respect to the original left-to-right direction and a chain series can be constructed with α_k being the first element in an incompatible pair. Then, $\alpha_k \leqslant \alpha_k^{\max}$, because $\alpha_k^{\max} \in T_k^*$ according to Theorem 2. But also $\alpha_k^{\max} \leqslant \alpha_k$ because α_k is the last element in T_k^*. Hence equality, $\alpha_k = \alpha_k^{\max}$. □

Corollary 1. *The sets $I_k^M \cup J_k^M$ and $J_k^M \cup I_{k+1}^M$, $k = 1, \ldots, r-1$, are both compatible with a tree. In particular, J_k^M is compatible with a tree.*

Proof. According to Lemma 2, $I_k^M \cup J_k^M = [\beta_{k-1}^{\min}, \beta_k^{\min} - 1]$, which equals I_k in the list constructed as in Algorithm 1. Similarly, $J_k^M \cup I_{k+1}^M = [\alpha_k^{\max} + 1, \alpha_{k+1}^{\max}]$, which equals I_{k+1}^* in the list constructed as in Algorithm 1 when

applied to the reversed set of sequences. Both I_k and I_k^* are compatible with a tree by construction and the corollary is proved. □

Example 2. Consider the following $n = 5$ sequences with $m = 5$ sites: 00000, 00011, 01110, 11000, 11111. It can easily be seen that $I_1^M = [1]$, $I_2^M = [3]$, $I_3^M = [5]$, $J_1^M = [2]$ and $J_2^M = [4]$. A minimal list can either have (i) $I_1 = [1, 2]$ or (ii) $I_1 = [1]$. If (i), then one can choose to put site 4 in either I_2 or I_3. If (ii), then $I_2 = [2, 3]$ and $I_3 = [4, 5]$ because the sites 2 and 4 are incompatible. Thus, whether the sites in J_k^M are allocated to I_k or I_{k+1} potentially affects the allocation of the sites in J_{k+1}^M, and so forth. There cannot be an example of this kind with less than five sites, neither can there be one with four sequences.

In the next two lemmas general disjoint lists are related to the number T_M.

Theorem 3. *Let $\mathscr{T} = (T_1, \ldots, T_s)$ be any disjoint list. Then $T_M \leqslant s \leqslant 2T_M - 1$.*

Proof. That $T_M \leqslant s$ follows from the definition of T_M. To prove $s \leqslant 2T_M - 1$, let $\mathscr{T}^* = (T_1^*, \ldots, T_r^*)$ be a minimal list with intervals $I_1^*, \ldots, I_r^*$, and let $I_1, \ldots, I_s$ be the intervals of $\mathscr{T}$. The interval I_1^* can at most be overlapping with I_1 and I_2, otherwise I_1 and I_2 would be compatible with each other. Each interval I_k^*, $1 < k < r$, can at most be overlapping with two intervals, I_j, and I_{j+1} for some j, that are not also overlapping with any of $I_1^*, \ldots, I_{k-1}^*$. Finally, I_r^* can at most be overlapping with the interval, I_s, that is not also overlapping with any of $I_1^*, \ldots, I_{r-1}^*$. In conclusion, $s \leqslant 2(r-1) + 1 = 2T_M - 1$, as desired. □

If equality holds in Theorem 3 the list is called maximal. A maximal list always consists of an odd number of topology classes.

Example 3. The bound in Theorem 3 is as good as possible. Consider Example 1. Denote the sites patterns by a, b, c, and d, in the order they occur in Example 1, and define five sequences by the following series of site patterns: $a, b, c, d, d, c, b, a, a, b, \ldots$ (for as long as we want). A minimal list is obtained by grouping sites in the following way: $I_1 = [1, 2]$ (with patterns a, b), $I_2 = [3, 6]$ (c, d, d, c), $I_3 = [7, 10]$ (b, a, a, b), etc. If m is odd then a maximal list is $I_1 = [1]$ (a), $I_2 = [2, 3]$ (b, c), $I_3 = [4, 5]$ (d, d), $I_4 = [6, 7]$ (c, b), $I_5 = [8, 9]$ (a, a), etc., because a and c are incompatible, and b and d are incompatible.

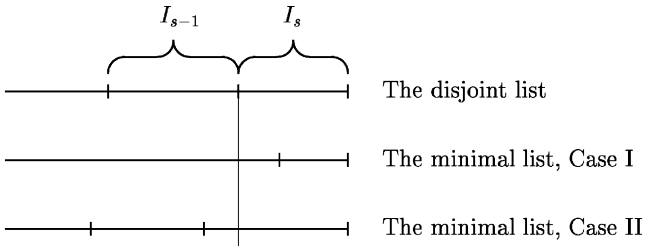
Example 4. A maximal list does not always exist. Consider the following two site patterns or columns in the alignment for four sequences: $a = 0011$ and $b = 0101$ (here, one string represents a column, not a sequence). Define four sequences by $a, b, a, b, a, b, a, b, a, \ldots$ (for as long as we want). Then there is only one disjoint (and minimal) list because a and b are incompatible.

Lemma 3. *For any disjoint list $\mathscr{T} = (T_1, \ldots, T_s)$ there exists $\alpha_k \in I_k$ and $\beta_k \in I_{k+1}$, $k = 1, \ldots, s-1$, such that α_k and β_k are incompatible, $\alpha_k < \alpha_{k+1}$, $\beta_k < \beta_{k+1}$, $\alpha_k < \beta_k$, and $\beta_k < \alpha_{k+2}$.*

Proof. Since $\mathscr{T}$ is disjoint, there exist $\alpha_k \in I_k$ and $\beta_k \in I_{k+1}$ such that α_k and β_k are incompatible. The inequalities follow from how α_k and β_k are chosen. □

Theorem 4. *Let $\mathscr{T} = (T_1, \ldots, T_s)$ be a disjoint list and let α_k, β_k, $k = 1, \ldots, s-1$, be a series fulfilling* Lemma 3. *Then there is at least $s - T_M$ instances of $\alpha_{k+1} < \beta_k$ for some $k = 1, \ldots, s-2$.*

Proof. Let $\mathscr{T}^* = (T_1^*, \ldots, T_r^*)$, $r = T_M$, be a minimal list with intervals $\mathscr{I}^* = (I_1^*, \ldots, I_r^*)$, constructed as in Algorithm 1. The proof is by induction on s. If $s = 1$ the result is obviously true since $s - T_M = 1 - 1 \geqslant 0$ (alternatively, $s = 2$ can be taken as the basis of induction: $s - T_M = 2 - 2 \geqslant 0$). Assume now the theorem is true for any disjoint list of length less than or equal to $s - 1$. There are two cases, I and II, illustrated below:



In case I, I_{r-1}^* ends after or at the same site as I_{s-1}. In case II, I_{r-1}^* ends before I_{s-1} and I_{r-2}^* before I_{s-2} (endpoints of intervals are marked with small vertical lines). That these are the only two cases follows from the construction of $\mathscr{T}^*$. Put $\mathscr{T}_j = (T_1, \ldots, T_j)$ and let $T_M(j)$ be the minimum number of topology classes for $\mathscr{T}_j$. Note that if α_k, β_k, $k = 1, \ldots, s-1$, is a series fulfilling Lemma 3 for $\mathscr{T}$, then α_k, β_k, $k = 1, \ldots, j$, fulfills Lemma 3 for $\mathscr{T}_j$. Further, let $c(j)$ be the number of times $\alpha_{k+1} < \beta_k$ for some k for $\mathscr{T}_j$. Clearly, $c(j+1) \geqslant c(j)$ for all j.

Consider case I. Then $T_M(s-1) = r - 1$ and $c(s) \geqslant c(s-1) \geqslant (s-1) - (r-1) = s - r$ by the induction hypothesis and the theorem holds for s.

Consider case II. Here $T_M(s-1) = r$. Note that $\alpha_{s-1} \in I_{s-1} \cap I_{r-1}^*$ and $\beta_{s-1} \in I_s$. The case splits into two subcases, A and B, according to which set β_{s-2} belongs to. Subcase A : $\beta_{s-2} \in I_{s-1} \cap I_r^*$, and subcase B : $\beta_{s-2} \in I_{s-1} \cap I_{r-1}^*$. If A, then clearly $\alpha_{s-1} < \beta_{s-2}$, and $c(s) = 1 + c(s-1) \geqslant 1 + (s-1) - r = s - r$, and the

theorem holds for s. If B, consider $\mathscr{T}^*_{s-1}$ with intervals defined by $\mathscr{I}^*_{s-1} = (I_1, \ldots, I_{s-2}, I_{s-1} \cap I^*_{r-1})$, i.e., $\mathscr{T}^*_{s-1}$ is $\mathscr{T}$ restricted to the sites up to and including those in I^*_{r-1}. The list $\mathscr{T}^*_{s-1}$ has $s-1$ elements, is disjoint because I_{s-2} and $I_{s-1} \cap I^*_{r-1}$ are incompatible ($\alpha_{s-2} \in I_{s-2}$ and $\beta_{s-2} \in I_{s-1} \cap I^*_{r-1}$ are incompatible by assumption), and α_k, β_k, $k = 1, \ldots, s-2$, is a series fulfilling Lemma 3 for $\mathscr{T}^*_{s-1}$. Further, the minimum number of topology classes of $\mathscr{T}^*_{s-1}$ is $r-1$, by construction. Using the induction hypothesis on $\mathscr{T}^*_{s-1}$ gives $c(s) \geqslant (s-1)-(r-1) = r-s$ and the theorem holds for s as well. The proof is completed.

Example 5. The inequality in Theorem 4 cannot be improved. Consider the sample in Examples 1 and 3 with eight sites given by the patterns a, b, c, d, d, c, b, a. Duplicate the five sequences to obtain 10 sequences in the following way: Add columns of five zeros to the first four sites, call these a_1, b_1, c_1, and d_1. As to the next four sites, prefix each column with five zeros. Call these d_2, c_2, b_2 and a_2. Then, the sample is given by $a_1, b_1, c_1, d_1, d_2, c_2, b_2, a_2$. For each $i = 1, 2$, the sites a_i, b_i, c_i, and d_i have the same incompatibility pattern as a, b, c, d, whereas any site with subscript 1 is compatible with a site with subscript 2, for example, a_1 is compatible with c_2, and so forth. The list $I^1_1 =$ [1] (a_1), $I^1_2 = [2, 3]$ (b_1, c_1), $I^1_3 = [4, 5]$ (d_1, d_2), $I^1_4 = [6, 7]$ (c_2, b_2), $I^1_5 = [8]$ (a_2) is disjoint and there is only one series, α_k, β_k, $k = 1, 2, 3, 4$, fulfilling Lemma 3: $(\alpha^1_1, \beta^1_1) = (1, 3)$, $(\alpha^1_2, \beta^1_2) = (2, 4)$, $(\alpha^1_3, \beta^1_3) = (5, 7)$, and $(\alpha^1_4, \beta^1_4) = (6, 8)$. Further, this series fulfills equality in Theorem 4, $s - T_M = 5 - 3 = 2$.

However, if d_1 and d_2 are interchanged such that the alignment now is $a_1, b_1, c_1, \boldsymbol{d_2}, \boldsymbol{d_1}, c_2, b_2, a_2$, then the list $I^2_1 =$[1] (a_1), $I^2_2 =$[2, 3] (b_1, c_1), $I^2_3 =$ [4, 5] (d_2, d_1), $I^2_4 = [6, 7]$ (c_2, b_2), $I^2_5 = [8]$ (a_2) has only one series fulfilling Lemma 3, namely: $(\alpha^2_1, \beta^2_1) = (1, 3)$, $(\alpha^2_2, \beta^2_2) = (2, 5)$, $(\alpha^2_3, \beta^2_3) = (4, 7)$, and $(\alpha^2_4, \beta^2_4) = (6, 8)$. But this series has three instances of $\alpha_{k+1} < \beta_k$ and $3 > s - T_M = 5 - 3 = 2$. In conclusion, the inequality in Theorem 4 cannot be improved.

3. Discussion

The number produced by Hudson and Kaplan's (1985) algorithm was shown to be the minimum number of topologies minus one required to explain a sample of sequences fulfilling the infinite-site assumption. However, as pointed out in the Introduction this number is rarely the true minimum number of recombinations, R_M, required to explain the data. It turns out that R_M can be found using a recursion of the form,

$$R(1, T) = w(1, T),$$

$$R(i, T) = \min\{R(i-1, T') + d(T, T') + w(i, T) \mid T' \text{ tree}\},$$

and

$$R = \min\{R(m, T) \mid T \text{ tree}\},$$

$i = 1, \ldots, m$, where $R(i-1, T')$ is the minimum for the first $i-1$ sites assuming the tree in the $(i-1)$th site is T' and $w(i, T)$ is 0 if T is compatible with site i (i.e., T is compatible with the partition (B_0, B_1) in site i), and infinite otherwise. Here a tree T is a topology with time points assigned to each node indicating when coalescence took place in the past and d is a metric derived from the coalescent process with recombination. In general d is difficult to compute. However, lower bounds to R_M can be obtained by bounding d by some metric d' such that $d \geqslant d'$. If $d' = 1\{T_1, T_2\}$ is an indicator variable that is zero if the two trees, T_1 and T_2, have the same topology and one otherwise, the bound is exactly T_M. Other possible bounding metrics d' are much more difficult to compute (Allen and Steel, 2001) and the algorithm easily becomes inefficient. This will be the subject of a subsequent paper. It is worth pointing out that the true minimum in general is much lower than the actual experienced number of recombination events in a sample's history. Hudson and Kaplan (1985) found, simulating under the neutral coalescent model, that for realistic values of the recombination rate the discrepancy between the true minimum and the actual number can be five-fold or more.

It transpires that the results presented here essentially depend on the Compatibility Theorem. The proofs are based on relating compatibility between sites to compatibility between sites and trees. By changing the definition of compatibility we change the results accordingly, but not the proofs. For example, ancestral states could be imposed, derived from chimp sequences or from some consensus rule, and two sites would be incompatible if all three gametes $(0, 1)$, $(1, 0)$, and $(1, 1)$ are found in two columns, assuming 0 is the ancestral state.

Acknowledgments

I am indebted to an anonymous reviewer who brought to my attention the connection between Lemma 2 and the proof of Corollary 1 (the original proof of Corollary 1 was tremendously more complicated). Y. Song and S. Myers are thanked for reading and commenting on the manuscript. Part of this work was done while the author was still at Oxford. In Oxford, the author was supported by the Carlsberg Foundation, Denmark, and by the Medical Research Council, UK.

References

Allen, B., Steel, M., 2001. Subtree transfer operations and their induced metrics on evolutionary trees. Ann. Comb. 5, 1–13.

Estabrook, G., Johnson, C., McMorris, F., 1975. An idealized concept of the true cladistic character. Math. Biosc. 23, 263–272.

Gusfield, D., 1991. Efficient algorithms for inferring evolutionary trees. Networks 21, 19–28.

Hudson, R.R., Kaplan, N., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111, 147–165.

Myers, S., 2002. Bounds on the minimum number of recombination events in a sample history, Genetics to appear.