Inference on Recombination and Block Structure Using Unphased Data

Carsten Wiuf[1]

Bioinformatics Research Center, University of Aarhus, 8000 Aarhus C, Denmark

Manuscript received December 16, 2002
Accepted for publication September 24, 2003

ABSTRACT

In this study compatibility with a tree for unphased genotype data is discussed. If the data are compatible with a tree, the data are consistent with an assumption of no recombination in its evolutionary history. Further, it is said that there is a solution to the *perfect phylogeny problem*; *i.e.*, for each individual a pair of haplotypes can be defined and the set of all haplotypes can be explained without invoking recombination. A new algorithm to decide whether or not a sample is compatible with a tree is derived. The new algorithm relies on an equivalence relation between sites that mutually determine the phase of each other. (The previous algorithm was based on advanced graph theoretical tools.) The equivalence relation is used to derive the number of solutions to the *perfect phylogeny problem*. Further, a series of statistics, R_M^j, $j \geq 2$, are defined. These can be used to detect recombination events in the sample's history and to divide the sample into regions that are compatible with a tree. The new statistics are applied to real data from human genes. The results from this application are discussed with reference to recent suggestions that recombination in the human genome is highly heterogeneous.

CURRENT efforts, initiated by the National Human Genome Research Institute (http://www.genome.gov), seek to accomplish a haplotype map of the human genome. One idea underlying these efforts is that recombination in the human genome happens mainly in localized regions, so-called hotspots, with little or virtually no recombination going on between the hotspots (*e.g.*, Gabriel *et al.* 2002, and references therein). This suggests a block-structured genome, where markers within the same block preferentially are inherited together. In consequence, one should be able to infer the location of hotspots (or, equivalently the boundaries of the nonrecombining blocks) from a detailed map of markers, *e.g.*, single-nucleotide polymorphisms (SNPs), spread throughout the genome. This has been attempted by various groups; among these are Daly *et al.* (2001), Jeffreys *et al.* (2001), Johnson *et al.* (2001), and Gabriel *et al.* (2002). Unfortunately, when markers are spread with long distances between them, it is experimentally difficult and time-consuming to obtain information about phase, *i.e.*, whether a marker allele has paternal or maternal origin. One must then rely on unphased data. Unphased data, in contrast to phased data, contains less information about the evolutionary history of a sample and increases the risk of inferring nonexisting hotspots or, oppositely, failing to infer existing hotspots and actual recombination events. To be concrete, consider the following sample with three individuals genotyped for two markers,

Individual 1:	2	2
Individual 2:	0	0
Individual 3:	1	1

Here 0 and 1 denote that an individual is homozygous for the 0 and 1 allele, respectively, and 2 denotes that an individual is heterozygous. Depending on how the phase of the double heterozygote, 2 2, is assigned, the inferred haplotypes are indicative of recombination (or gene conversion) in the sample's history or consistent with an assumption of no recombination. The presence of the four possible gametes in two sites is taken as evidence of recombination (*cf.* the four-gamete test; Hudson and Kaplan 1985), which is a reliable indicator as long as recurrent mutations are rare or absent. In the following all mutation events are assumed to be unique.

Recently, Gusfield (2002) showed that it can be determined efficiently whether a sample of unphased genotypes (*e.g.*, the example given above) is consistent with the assumption of no recombination. If it is consistent, then the sample is said to be compatible with a tree: A pair of haplotypes can be defined for each individual and the genealogical history of all these haplotypes can be illustrated with a tree. It is said that there is a solution to the *perfect phylogeny haplotype* (PPH) problem (Gusfield 2002, and references therein). Potentially pairs of haplotypes can be defined in various ways resulting in multiple solutions to the PPH problem. Gusfield's algorithm to determine whether or not the PPH problem has a solution can be used to screen the data and divide markers into disjoint blocks that are all compatible with a tree. (This relies essentially on insight in Wiuf 2002.) The blocks can be seen as estimating the (supposed) block structure of the genome and/

[1] *Address for correspondence:* Bioinformatics Research Center, Department of Computer Science, University of Aarhus, Ny Munkegade, Bldg. 540, 8000 Aarhus C, Denmark. E-mail: wiuf@daimi.au.dk

or the number of regions with different evolutionary histories.

This work extends and adds to Gusfield's work. He applied advanced graph theoretical tools to derive the algorithm. In this article a simpler and more intuitive algorithm is developed, on the basis of an equivalence relation between sites that mutually determine each other's phase. Using the equivalence relation, one can derive analytically the number of different solutions to the PPH problem and when a unique solution exists. This has not been done in previous work.

There has been some work on the related problem of inferring recombination from phased data, *i.e.*, from haplotype data. It is a considerably simpler problem because compatibility with a tree can be characterized in terms of the four-gamete test (Estabrook *et al.* 1975; Gusfield 1991). Thus, whether a sample of phased genotypes is compatible with a tree can be decided by comparing sites pairwise. For unphased data such a simple characterization does not exist.

One commonly applied statistic for inferring recombination from phased data is Hudson and Kaplan's (1985) R_M, a lower bound to the number of recombination events in the evolutionary history of the sample. It is based on the four-gamete test. Wiuf (2002) showed that the sample can be divided into $R_M + 1$ disjoint blocks, such that each block is compatible with a tree, and that this cannot be done with fewer than $R_M + 1$ blocks. Thus, R_M can be seen as an estimator of the number of blocks between hotspots in the genome or as an estimator of the number of regions with different evolutionary histories. Myers and Griffiths (2003) extended Hudson and Kaplan's (1985) work in various ways; in particular they developed a general method or framework for inferring recombination. In this framework R_M is just one of many possible statistics for this purpose. Their framework is not restricted to phased data, but applies equally to unphased data. Here, it is used to define an increasing series of statistics, R_M^1, $R_M^2, \ldots, R_M^m$ (m is the number of variable sites), on the basis of the equivalence relation, which utilize an increasing amount of the information in the sample. R_M^m is similar to R_M and it is shown that the (unphased) sample can be divided into $R_M^m + 1$ disjoint blocks, all compatible with a tree, and that this cannot be done with fewer than $R_M^m + 1$ blocks. The statistics $R_M^2, R_M^3, \ldots, R_M^{m-1}$ are approximations of R_M^m. It turns out that in general R_M^3 is a very good approximation of R_M^m and much simpler to compute. R_M^2 is considerably poorer. The new statistics are applied to simulated and real data and compared to the "ideal" statistic R_M.

The next section introduces the setting and the following section (examples) gives examples to motivate further theoretical development. In results general analytical results about compatibility for unphased genotypes and the equivalence relation are presented. The results are applied in applications and in the discussion: In applications the new statistics are applied to simulated data, and in the discussion they are applied to haplotype data from two human genes, the *APOE* gene and the β-globin gene. Last, in the discussion the presented work is discussed and direction for future research pointed out. All proofs are derived in the appendix.

Haplotypes	Genotypes
0 0 0	2 2 0
1 1 0	
0 1 0	2 2 0
1 0 0	

Figure 1.—Shown are four haplotypes with three sites from two individuals. All four haplotypes are different, but the genotypes are identical. Assume that the genotypes are known, but not the haplotypes. A double heterozygote, a 2 2 in a row, can be resolved in either of two ways, as illustrated. The phase of the first heterozygote in a row can be assigned arbitrarily, *i.e.*, whether it is 0 on top of 1, denoted (0, 1), or 1 on top of 0, denoted (1, 0), for reasons of symmetry.

SETTING AND DEFINITIONS

Let S be a matrix of m biallelic unphased genotypes with alleles 0 and 1, sampled from n individuals, *i.e.*, $S = (s_{ij})_{i,j}$, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$. The columns are sites, and the rows are pairs of unphased chromosomes, one pair for each of n individuals. The matrix S has entries 0, 1, and 2. The entry s_{ij} is 0 if both copies of the allele are 0, $s_{ij} = 1$ if both copies are 1, and $s_{ij} = 2$ if one copy is 0 and the other is 1. Thus, individual i is homozygous for site j, if $s_{ij} = 0$ or $s_{ij} = 1$, and heterozygous if $s_{ij} = 2$. Note that 0 and 1 are used to denote two different things, sometimes denoting a single allele, sometimes a genotype. It will be clear from the context which of the two denotations is referred to. Column j is denoted $s_j = (s_{1j}, s_{2j}, \ldots, s_{nj})$. The notation is illustrated in Figure 1.

The haplotypes determine the genotypes uniquely, and the opposite statement is not true. Phase can be assigned to a double heterozygote in either of two ways (see Figure 1). In some cases, one or both of them give rise to an incompatibility, in other cases none of them do. Throughout, "to resolve a heterozygote, a double heterozygote, a site, a pair of sites, S etc.," is used in the sense "to assign phase to the genotype(s) of a heterozygote, a double heterozygote, a site, a pair of sites, a row, S etc.," and "resolution" as the resolved (phased) genotypes (in that Clark 1990 is followed). A compatible resolution is a resolution for which the set of inferred haplotypes is compatible with a tree. Thus, a compatible resolution is a solution to the PPH problem.

Two sites, i and j, can have identical columns, $s_i = s_j$, or identical columns after interchanging 0's and 1's, leaving 2's unchanged. This is denoted $i \approx j$.

A number of definitions are required to carry on.

Definition 1. S is said to be compatible if there is a compatible resolution of S. If S is not compatible, S is said to be incompatible.

Definition 2. S is said to be k-compatible if all subsets of size $k \leq m$ are compatible. S is said to be k-incompatible if S is $(k - 1)$-compatible, but not k-compatible.

If S is k-compatible it is also k'-compatible, $k' < k$. However, S can be k-incompatible for at most one k. Furthermore, 2-incompatibility has a property that is not shared by k-incompatibility, $k > 2$. For S to be 2-incompatible there must be two sites from which all four gametes can be inferred, irrespective of how double heterozygotes are resolved.

Definition 3. Let (i, j) be a pair of sites. Define het (i, j) by het$(i, j) = 1$, if there is a double heterozygote in $(s_i\ s_j)$, and het$(i, j) = 0$, otherwise.

Obviously, if het$(i, j) = 0$, then $S = (s_i\ s_j)$ is unambiguously resolved. If het$(i, j) = 1$, this is not the case.

Definition 4. Let (i, j) be a pair of sites. The pair is said to be resolvable, $i \overset{r}{\sim} j$, if het$(i, j) = 1$ and there exists a unique compatible resolution of $(s_i\ s_j)$. S is said to be resolvable, if for any pair of sites (i, j) either (1) het$(i, j) = 0$ or (2) $i \overset{r}{\sim} j$.

Gusfield (2002) studied the submatrix, S_{01}, of columns with at least one instance of 1 and one instance of 0. If S_{01} is compatible, then S_{01} is resolvable. However, Definition 4 does not require that 0 and 1 are present in all sites. Also note that "resolvability" is defined on pairs of sites, rather than on double heterozygotes.

Definition 5. Let (i, j) be a pair of sites. The pair is said to be weakly resolvable, $i \overset{w}{\sim} j$, if het$(i, j) = 1$ and either (1) $i \overset{r}{\sim} j$ or (2) a site k exists, such that $i \overset{w}{\sim} k$ and $j \overset{w}{\sim} k$.

Figure 2 illustrates Definition 5 through three examples. To show that $i \overset{w}{\sim} j$ potentially involves sites other than i and j. In contrast, to show $i \overset{r}{\sim} j$ involves only i and j. If $i \overset{w}{\sim} j$ (or $i \overset{r}{\sim} j$), then $\overset{w}{\sim}$ (or $\overset{r}{\sim}$) is said to impose a resolution on $(s_i\ s_j)$. If $\overset{w}{\sim}$ imposes resolutions on $(s_i\ s_k)$ and $(s_j\ s_k)$, then $\overset{w}{\sim}$ also imposes a resolution on $(s_i\ s_j)$. The resolution might not be unique, however, as will become clear later. The implications of Definition 5 are discussed further in examples and in results.

A key observation is the following: If $i \overset{w}{\sim} j$ then i_i, i_2, . . . , i_k exist, such that $i \overset{r}{\sim} i_1$, $i_1 \overset{r}{\sim} i_2$, . . . , $i_{k-1} \overset{r}{\sim} i_k$, $i_k \overset{r}{\sim} j$ (i_1, . . . , i_k need not be different). Thus, if there is a resolution of $(s_i\ s_j)$, compatible with a resolution of the sites i_1, i_2, . . . , i_k, then it can be found by repeated application of $\overset{r}{\sim}$. The relations "resolvable" and "weakly resolvable" are obviously symmetric, but in general not reflexive; *e.g.*, if site i is a single heterozygote, $s_i = (2)$ $(n = m = 1)$, then the relations $i \overset{r}{\sim} i$ and $i \overset{w}{\sim} i$ fail. However, if a site j exists, such that $i \overset{w}{\sim} j$ and het$(i, j) = 1$, then also $i \overset{w}{\sim} i$ by definition. (The same holds if $i \approx j$. Then $i \overset{r}{\sim} j$ and $i \overset{w}{\sim} j$ might fail.) Also transitivity might fail, because het$(i, j) = 1$ does not in general follow from $i \overset{r}{\sim} k$ and $j \overset{r}{\sim} k$ (and similarly for $\overset{w}{\sim}$).

i	j	k	i	j	k	i	j	k
2	2	2	2	2	2	2	2	0
2	2	0	2	2	0	2	0	2
1	0	0	0	0	1	0	2	2
0	0	1						

Figure 2.—Examples of weakly resolvable sites. The sites, i and j, in the first two examples are weakly resolvable, $i \overset{w}{\sim} j$, but not resolvable. In the first two examples there is a row of three 2's, 2 2 2; in the last example there is not, and it transpires that $i \overset{r}{\sim} j$, irrespective of whether some of the 0's are replaced by 1's.

In the next section a few examples that relate to the definitions are given.

EXAMPLES

Below is an example, S_1, that is 2-compatible, but not 3-compatible. Hence S_1 is 3-incompatible. This shows that compatibility for genotypes cannot be characterized similarly to compatibility for haplotypes. The matrix S_1 has $n = 4$ individuals and $m = 3$ sites:

1	2	3
2	2	2
1	1	0
1	0	1
0	0	0

For a 3-incompatibility to occur there must be a row of three 2's (*cf.* Theorem 1 in the next section). All pairs of sites in this example are resolvable, but the order in which they are resolved with $\overset{r}{\sim}$ affects the result, *e.g.*,

$1 \overset{r}{\sim} 2$	$1 \overset{r}{\sim} 3$
00	00
11	11
11	10
10	11
00	00

$1 \overset{r}{\sim} 2$	$2 \overset{r}{\sim} 3$
00	01
11	10
11	10
10	11
00	00

(the two rows above the lower line represent the resolved heterozygotes). In either case, an incompatibility occurs, and the relation $\overset{r}{\sim}$ cannot be applied consistently without creating an incompatibility.

Note that this is the smallest possible example of a 3-incompatibility, in terms of both the number of sites and the number of individuals. There are examples of k-incompatibilities for all k.

As a second example consider S_2 given by

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

Here the sites 1, 2, and 3 are mutually weakly resolvable and compatible, and so are 4, 5, and 6. No other pairs are weakly resolvable. If site 4 is resolved as (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (listed top down), where (x, y) denotes a phased genotype, then two resolutions exist of site 1 compatible with site 4: (A) (0, 1), (0, 0), (0, 1), (0, 0), (0, 0), (0, 0); and (B) (1, 0), (0, 0), (1, 0), (0, 0), (0, 0), (0, 0). After choosing either A or B, a compatible resolution of S_2 is uniquely imposed by $\overset{w}{\sim}$; *e.g.*, if site 1 is resolved as A, then site 2 is resolved as (1, 0), (1, 0), (0, 0), (0, 0), (0, 0), (0, 0), etc. The second phased genotype cannot be flipped without creating an incompatibility. In consequence, there are two compatible resolutions of S_2 or, equivalently, two solutions to the PPH problem for S_2.

However, if the sites 1, 2, and 3 are given by

1	2	3
2	2	0
0	2	2
2	0	2
0	0	0
0	0	0
0	0	0

then the sites 1, 2, and 3 are still mutually weakly resolvable and compatible, but not compatible with the sites 4, 5, and 6. The sites 3 and 4 become incompatible.

RESULTS

First, a couple of special cases, where compatibility can be characterized in simple ways, are presented.

THEOREM 1. *Assume that there are at most two* 2*'s in a row. Then S is compatible if and only if S is* 2*-compatible.*

If the condition in Theorem 1 is fulfilled and S is compatible, then a compatible resolution might not be unique. The simplest example of this kind is a double heterozygote (2 2) that can be resolved in either of two ways. It is important to note that there is no similar characterization of compatibility for $k > 2$. In general, 2-compatibility is a necessary condition for compatibility to hold, not a sufficient condition.

THEOREM 2. *Assume that S is resolvable. Then S is compatible if and only if S is* 3*-compatible. As a consequence, if S is compatible then any resolution of S is unique.*

Next, a number of results about $\overset{w}{\sim}$ are presented. To this end it is useful to develop $\overset{w}{\sim}$ into an equivalence relation, $\overset{e}{\sim}$.

DEFINITION 6. Define the equivalence relation, $\overset{e}{\sim}$, on sites by the following requirements: $i \overset{e}{\sim} j$, if (1) $i = j$, (2) $i \overset{w}{\sim} j$, or (3) a site k exists, such that $i \overset{e}{\sim} k$ and $j \overset{e}{\sim} k$.

In Definition 6 it is not required that $\text{het}(i, j) = 1$, contrary to the definition of $\overset{w}{\sim}$. Further, it follows from the definitions of $\overset{r}{\sim}$ and $\overset{w}{\sim}$ that $i \overset{e}{\sim} j$ implies that $i_1, i_2, \ldots, i_k$ exist, such that $i \overset{r}{\sim} i_1$, $i_1 \overset{r}{\sim} i_2, \ldots, i_k \overset{r}{\sim} j$. Thus, whether or not $i \overset{e}{\sim} j$ can be determined by repeated application of $\overset{r}{\sim}$.

LEMMA 1. *The relation $\overset{e}{\sim}$ is an equivalence relation.*

LEMMA 2. *If $i \overset{e}{\sim} j$ and $het(i, j) = 1$, then $i \overset{w}{\sim} j$.*

The relation $i \overset{e}{\sim} j$ (or $i \overset{w}{\sim} j$) does not imply that $(s_i\ s_j)$ is 2-compatible. This is implied only by $i \overset{r}{\sim} j$. Consider,

1	2	3
2	2	2
0	2	2
1	2	0
2	0	2

Here, $1 \overset{r}{\sim} 3$ and $2 \overset{r}{\sim} 3$, so $1 \overset{e}{\sim} 2$ (and $1 \overset{w}{\sim} 2$), but $(s_1\ s_2)$ is obviously 2-incompatible. If S is 2-compatible and $i \overset{e}{\sim} j$ for all i and j, then S can still be incompatible.

Define E_0 by

$$E_0 = \{i | \text{het}(i, j) = 0 \text{ for all } j \neq i\},$$

and consider the equivalence classes of $\overset{e}{\sim}$. The sites i and j are in the same class if and only if $i \overset{e}{\sim} j$. All $i \in E_0$ form classes with single elements, namely i. Denote the remaining equivalence classes by $E_1, \ldots, E_M$, $M \geq 1$ (if $i \in E_k$ and $i \overset{e}{\sim} j$, then $j \in E_k$). Then $E_0, E_1, \ldots, E_M$ are disjoint and $\cup_{j=0}^{M} E_j = \{1, 2, \ldots, m\}$. Let F_1 and F_2 be two of the $M + 1$ classes (or M, if $E_0 = \emptyset$) and i and j two sites, such that $i \in F_1$ and $j \in F_2$. The relation $i \overset{r}{\sim} j$ fails, because the sites are in different classes. If S is 2-compatible, then the possible genotype patterns (s_i and s_j) of i and j are limited. This is shown in Figure 3. For easy notation the following is defined.

i	j	i	j
2	2	2	2
0	2	1	2
0	1	1	0
0	0	1	1

Figure 3.—Possible genotype patterns if het$(i, j) = 1$ and $i \in F_1$, $j \in F_2$ (the order of i and j can be reversed). The two patterns are symmetric; the first can be obtained from the second by interchanging 0's and 1's. At least 2 2 must be present, because het$(i, j) = 1$; 0 2, 0 1, and 0 0 (similarly, 1 2, 1 0, and 1 1) are optional.

Definition 7. Assume that $(s_i\ s_j)$ is 2-compatible. If het$(i, j) = 1$ and $i \overset{r}{\sim} j$ fails, write $i < j$ with s_i and s_j given as in Figure 3. If het$(i, j) = 0$, write $i \perp j$.

If i and j are distinct sites, but $i \approx j$, then it is possible that $i < j$ and $j < i$; *e.g.*, if $S = (s_i\ s_j) = (2\ 2)$, then $i < j$ and $j < i$. Another example of this kind consists of the two columns labeled i in Figure 3 (see also Corollary 1).

Theorem 3. *Assume that* S is 2-*compatible and let* F_1 *and* F_2 *be given. Then* $F \subseteq F_1 \times F_2$ *exists, possibly empty, such that* $i < j$ *for all* $(i, j) \in F$ *and* $i \perp j$ *for all* $(i, j) \in F_1 \times F_2 \backslash F$ *(or the same with* F_1 *and* F_2 *reversed). In consequence, either* (1) *the set of rows (individuals) with heterozygotes in* F_1 *and the set of rows with heterozygotes in* F_2 *are disjoint or* (2) *the set of rows with heterozygotes in* F_1 *is a subset of the rows with heterozygotes in* F_2, *according to whether* $F = \emptyset$ *or* $F \neq \emptyset$, *respectively.*

It is convenient to write $F_1 \perp F_2$, if $F = \emptyset$ in Theorem 3, and otherwise $F_1 < F_2$ (or $F_2 < F_1$).

Corollary 1. *Assume as in Theorem* 3. *If* $F_1 < F_2$ *and* $F_2 < F_1$, *then* $F_1 = \{i\}$, $F_2 = \{j\}$, *and* $i \approx j$ *for some* $i \neq j$.

Corollary 2. *Assume that* S *is* 2-*compatible. The classes* $E_0, E_1, \dots, E_M$ *form a hierarchy such that for all* $\alpha, \beta = 0, 1, \dots, M$, $E_\alpha \perp E_\beta$, $E_\alpha < E_\beta$, *or* $E_\beta < E_\alpha$. *Further, if* $E_\alpha < E_\beta$ *and* $E_\beta < E_\gamma$, *then* $E_\alpha < E_\gamma$ *(transitivity); and if* $E_\alpha < E_\beta$ *and* $E_\beta \perp E_\gamma$, *then* $E_\alpha \perp E_\gamma$. *In particular,* $E_0 \perp E_\alpha$ *for all* $\alpha > 0$.

Consider E_α, $\alpha > 0$. All double heterozygotes in E_α can be resolved using $\overset{r}{\sim}$ and the sites in E_α only (*cf.* Lemma 2). However, the resolution might depend on the order in which $\overset{r}{\sim}$ is applied (as illustrated in examples). Let $i, j \in E_\alpha$ and assume that $(s_i\ s_j)$ can be resolved in two ways: (A) $i \overset{r}{\sim} i_1, \dots, i_k \overset{r}{\sim} j$ and (B) $i \overset{r}{\sim} j_1, \dots, j_l \overset{r}{\sim} j$, such that application of $\overset{r}{\sim}$ to the sites in A (or B) in the given order eventually gives the phase of $(s_i\ s_j)$. If A and B give different resolutions, then E_α cannot be compatible. Thus, either all resolutions, A, B, . . . , of E_α imposed by $\overset{w}{\sim}$ are compatible or none of them are. In the former case, a resolution is necessarily unique. This proves the second part of the next lemma.

Lemma 3. *Assume that* S *is* 2-*compatible. Then there is a unique compatible resolution of* E_0. *The class* E_α, $\alpha > 0$, *is compatible if and only if all resolutions imposed by* $\overset{e}{\sim}$ *on* E_α *are compatible. If* E_α *is compatible, then any resolution of* E_α *is unique.*

Lemma 4. *Assume that* S *is* 2-*compatible and that* E_α *and* E_β *both are compatible. If* $E_\alpha \perp E_\beta$, *then there is a unique compatible resolution of* $E_\alpha \cup E_\beta$. *If* $E_\alpha < E_\beta$, *then* $E_\alpha \cup E_\beta$ *is compatible if and only if all* 2*'s in* s_i, $i \in E_\alpha$, *can be resolved* (0, 1), *or be resolved* (1, 0).

The matrix S_3 in examples illustrates the result of the lemma.

Definition 8. Let $\varepsilon \subseteq \{E_1, E_2, \dots, E_M\}$ be such that

1. if $E_\alpha, E_\beta \in \varepsilon$, then $E_\alpha \perp E_\beta$;
2. if $E_\alpha \in \{E_1, E_2, \dots, E_M\}$, then $E_\beta \in \varepsilon$ exists such that $E_\alpha < E_\beta$.

At set ε fulfilling 1 and 2 is called a set of terminals, and the elements in ε are terminals. Note that if ε_1 and ε_2 are two sets of terminals, then they have the same cardinality, $\#\varepsilon_1 = \#\varepsilon_2 = T$ for some T.

Theorem 4. *Assume that* S *is* 2-*compatible. Either* S *is incompatible or there are* 2^{M-T} *different compatible resolutions of* S. *In the latter case there are* 2^{M-T} *solutions to the PPH problem. If a solution is unique then* $M = T$.

APPLICATIONS

Simulations: Myers and Griffiths (2003) developed a general framework for inferring recombination from haplotype data. Their framework is readily applicable to genotype data as well. It consists of two steps. First, an $m \times m$ matrix, $B = (b_{ij})_{i,j=1,\dots,m}$ is filled out: The entry b_{ij} is a lower bound to the number of recombination events between the sites i and j. Such a bound can be obtained in many ways. For example, Hudson and Kaplan (1985) defined $b_{ij} = 1$, if i and j are incompatible sites, and 0 otherwise. Myers and Griffiths (2003) suggested different improvements of Hudson and Kaplan's (1985) bound. This is taken up in the discussion.

The second step is an algorithm for calculating a combined bound B_{ij} that respects all the bounds $b_{i'j'}$, $i \leq i' < j' \leq j$. Their algorithm is given by

$$B_{ij} = \max\{b_{ik} + B_{kj} | k = i + 1, \dots, j - 1\}, \qquad (1)$$

with boundary conditions $B_{ii} = 0$ and $B_{i,i+1} = b_{i,i+1}$. It follows that

$$B_{ij} = b_{i_1 i_2} + b_{i_3 i_4} + \ldots + b_{i_{k-1} i_k}, \qquad (2)$$

for some $i = i_1 < i_2 < \ldots < i_k = j$. In particular, the global bound B_{1m} is of interest. Let R_M denote the global bound for haplotypes obtained with Hudson and Kaplan's b_{ij}, as defined above.

In the context of unphased data, the bounds b_{ij} could be defined in various ways related to Hudson and Kaplan's. Of interest is

$$\begin{aligned} b_{ij}^k &= 1, \quad \text{if } b_{ij}^{k-1} = 1, \text{ or } i_1, i_2, \ldots, i_{k-2} \text{ exists,} \\ &\qquad \text{such that } (s_i s_{i_1} \ldots s_{i_{k-2}} s_j) \text{ is } k\text{-incompatible} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

Theorems 1 and 2 provide conditions when b_{ij}^1 and b_{ij}^2 are optimal. For $k \geq m$, the definition reduces to $b_{ij}^k = 1$, if $(s_i\ s_{i+1} \ldots s_{j-1}\ s_j)$ is incompatible, and $b_{ij}^k = 0$ otherwise. Denote the global bound based on b_{ij}^k by R_M^k.

Lemma 5. *For $k \geq 2$,*

$$R_M \geq R_M^{k+1} \geq R_M^k,$$

where the genotype bounds, R_M^k, are obtained by randomly pairing haplotypes to create individuals.

Let $[x, y]$ denote the interval of integers z, such that $x \leq z \leq y$.

Theorem 5. *Define $\mathcal{I}_k = (I_1^k, I_2^k, \ldots, I_{i_k}^k)$, $i = 1, \ldots, m$, recursively by*

1. $\mathcal{I}_1 = (I_1^1)$, *with* $I_1^1 = [1, 1]$ *and* $i_1 = 1$,
2. *if* s_{k+1} *is compatible with the sites in* $I_{i_k}^k$, *then* $i_{k+1} = i_k$ *and*
$$\begin{aligned} \mathcal{I}_{k+1} &= (I_1^{k+1}, \ldots, I_{i_{k+1}}^{k+1}) \\ &= (I_1^k, \ldots, I_{i_{k-1}}^k, I_{i_k}^k \cup [k+1, k+1]), \end{aligned}$$
3. *if* s_{k+1} *is incompatible with the sites in* $I_{i_k}^k$, *then* $i_{k+1} = i_k + 1$ *and*
$$\begin{aligned} \mathcal{I}_{k+1} &= (I_1^{k+1}, \ldots, I_{i_{k+1}}^{k+1}) \\ &= (I_1^k, \ldots, I_{i_k}^k, I_{i_{k+1}}^{k+1}), \quad \text{with } I_{i_{k+1}}^{k+1} = [k+1, k+1]. \end{aligned}$$

Then $\mathcal{I}_m = (I_1^m, \ldots, I_{i_m}^m)$ fulfills: The sites in I_j^m, $j = 1, \ldots, i_m$, are compatible and i_m is the smallest number of disjoint intervals with this property. In particular, $R_M^m = i_m - 1$.

Gusfield's (2002) algorithm to decide whether a matrix S of sites is compatible with a tree has a running time of the order of $O(nm)$. This implies the algorithm in Theorem 5 can be implemented with a running time of the order of $O(nm^2)$.

The bounds R_M^2, R_M^3, and R_M^m were compared to R_M via simulations. The neutral coalescent with recombination, constant population size, and infinite-site mutation (Hudson 1983) was used to generate samples of haplotypes from which genotype data were obtained by randomly pairing haplotypes. This approach makes it possible to calculate R_M and R_M^k on the same data sets. In all simulations, the scaled mutation rate per gene (or genomic region), θ, is fixed, θ = 10, and the ratio α = ρ/θ is varied. Here ρ is the scaled recombination rate per gene. Two models for the recombination process were used: (A) a model with one hotspot in the middle of the gene, *i.e.*, two blocks of equal size, and (B) a model with flat rate; *i.e.*, recombination happens uniformly along the gene. Table 1 gives a summary of the simulations.

TABLE 1

Simulated results

	n^a					
	10			50		
α	0.1	1	10	0.1	1	10
A. Two blocks						
$E(R_M)$	0.459	0.952	0.987	0.758	0.996	1.000
$E(R_m)$	0.437	0.926	0.970	0.753	0.994	1.000
$E(R_3)$	0.417	0.920	0.970	0.752	0.994	1.000
$E(R_2)$	0.342	0.849	0.923	0.705	0.988	0.999
B. Flat rate						
$E(R_M)$	0.520	2.724	7.867	1.046	4.486	12.129
$E(R_m)$	0.484	2.435	6.628	1.030	4.321	11.252
$E(R_3)$	0.464	2.376	6.587	1.024	4.310	11.248
$E(R_2)$	0.373	2.104	6.073	0.938	4.027	10.742

[a] n, number of individuals; there are $2n$ haplotypes.

Always, $R_M^j \leq R_M \leq 1$ under A. As estimators of the number of blocks the statistics underperform. If the recombination rate is high, blocks are more easily inferred. The expected number of segregating sites is $E(S_n) = \theta\Sigma_{i=1}^{2n-1} 1/i$ (Watterson 1975), which is 34.5 for $n = 10$ and 51.8 for $n = 50$. In consequence, $E(S_n)/2$ is the average number of variable sites in one block; *e.g.*, $E(S_{50})/2 = 25.9$, if $n = 50$.

The situation is different for the flat rate model. If ρ is high, then a chromosome becomes distributed onto many different ancestral genomes in the course of evolutionary time. The number of recombination events that cause the tree topology to change is ∼1.5ρ for $n = 10$ and 3.1ρ for $n = 50$ (Hudson and Kaplan 1985). For α = 0.1 (*i.e.*, ρ = 1), R_M and R_M^m find only about one-third of all topology changes.

It transpires that the gain by using R_M^2 instead of R_M^3 is in general much larger than the gain by using R_M^3 instead of R_M^m and also that there is a significant gain in knowing the haplotypes rather than just the unphased genotypes.

Gene data: Data from two genes were chosen. They were split into five data sets. The first data set is composed of 60 chromosomes sequenced at the β-globin

TABLE 2

The six data sets

Gene	kb	$2n$	H_n	S_n
β-Globin	3.1	60	17	18
APOE, R	5.5	48	18	13
APOE, N	5.5	48	13	13
APOE, J	5.5	48	16	14
APOE, C	5.5	48	8	7
APOE, All	5.5	192	47	21

kb, length of gene in kilobases, $2n$, number of chromosomes (n is the number of individuals); H_n, number of different haplotypes; S_n, number of SNPs; R, Rochester, Minnesota (European-American); N, North Karelia, Finland (European); J, Jackson, Mississippi (African-American); C, Campeche, Mexico (Hispanic); and All, R, N, J, and C together.

TABLE 3

Summary statistics for R_M^j

Gene	R_M	Range	R_M^m [a]		R_M^3:[a]	R_M^2:[a]
			Same[b]	Ave[b]	Ave[b]	Ave[b]
β-Globin	5	4–5	0.755	4.755	4.755	4.686
APOE, R[c]	6	2–6	0.036	3.427	3.425	3.073
APOE, N[c]	4	2–4	0.572	3.511	3.511	3.146
APOE, J[c]	3	1–3	0.212	1.928	1.927	1.791
APOE, C[c]	1	0–1	0.996	0.996	0.996	0.995
APOE, All[c]	9	4–9	0.008	6.158	6.158	5.518

[a] Summary of 1000 samples generated by randomly pairing haplotypes.

[b] Same, observed probability that R_M^j equals R_M; Ave, average.

[c] See Table 2.

locus (Fullerton *et al.* 1994). The other four data sets consist of chromosomes sequenced at the *APOE* gene sampled at four different locations around the world, each composed of 48 chromosomes (Fullerton *et al.* 2000). In addition, the four *APOE* samples were combined into one data set of 192 chromosomes. Table 2 provides a summary of the data.

To investigate the performance of R_M^2, R_M^3, and R_M^m, genotypes were generated 1000 times from the haplotypes by randomly pairing haplotypes. The statistics R_M^2, R_M^3, and R_M^m were calculated for each of the 1000 data sets and compared to R_M, calculated on the true haplotypes. Table 3 shows summaries of the results. For all data sets R_M^3 and R_M^m gave very similar results. However, R_M differs in some cases sharply from R_M^m, *e.g.*, for *APOE*, European-American, and *APOE*, All, whereas in other cases R_M^m is in close agreement with R_M, *e.g.*, for β-globin, *APOE*, European, and *APOE*, Hispanic. Overall phase information is very useful. It is surprising that the African-American sample showed less recombination than the European-American and European samples.

If R_M is calculated on only the common haplotypes different results are obtained (see Table 4). Less recombination break points are detected and some sort of block structure emerges. The same was observed in simulated data with a flat recombination rate (results not shown). It seems that a supposed block structure can be an artifact of how the data are analyzed. This is taken up further in the discussion.

DISCUSSION

In Gabriel *et al.* (2002) blocks are defined on the basis of a linkage disequilibrium (LD) measure. Roughly speaking, two sites are in the same block if LD between them is high. A similar procedure is applied in Daly *et al.* (2001). Basically, such a procedure tends to cluster sites with histories that differ by recent recombination and gene conversion events. In general, these types of events affect the history of a small fraction of the sample only. As a consequence, the LD measure also is affected only marginally. At least naively, this does not seem to be appropriate: Hotspot recombination increases the rate of recombination in the region around a hotspot, but should not impose constraints on the time of particular events.

The statistics, R_M^j, $j \geq 2$, discussed in this article do not discriminate between recent and old or sporadic and hotspot events. Some of the break points detected by R_M^j might be due to gene conversion, and others might be due to recent events affecting only a minority of the haplotypes. Still others might be due to recurrent mutation. Table 4 and the accompanying text showed that when only common haplotypes are taken into account, the results leave the impression of a block-structured genome. Thus, there is an obvious danger in overinterpretation of results in favor of block structure. Empirical results are somewhat ambivalent on this point. For example, in Gabriel *et al.* (2002) the same block structures did not show up in all populations, contrary to what one might expect if blocks are really hotspot delimited. However, due to different demo-

TABLE 4

Common haplotypes only

Gene	H_n	S_n	R_M
β-Globin	5	13	1
APOE, R[a]	4	7	0
APOE, N[a]	5	7	0
APOE, J[a]	4	4	0
APOE, C[a]	7	6	1
APOE, All[a]	8	4	3

Only haplotypes that appear in frequency 5% or higher are shown. H_n, number of distinct haplotypes; S_n, number of variable sites among common haplotypes.

[a] See Table 2.

graphic and genealogical histories of populations, evidence for hotspots might fail to show in some populations and sporadic recombination events might falsely be taken as evidence for (nonexisting) hotspots. In conclusion, there seem to be obstacles to overcome and more careful analyses to be done before the block-structured genome can be claimed as a solid fact.

The statistics R^j_M, $j \geq 2$, underestimate the number of break points compared to R_M calculated on haplotype data. Another drawback of the statistics R^j_M, $j \geq 2$, is that they are not able to take frequencies of haplotypes into account nor are they able to take the number of different haplotypes into account. (LD measures like r^2 and D' take SNP frequencies into account.) One of the haplotype measures proposed by Myers and Griffiths (2003) makes use of the number of haplotypes. However, it is computationally difficult to generalize their statistic to the case of unphased data. The statistic is denoted H_M and it is always true that $H_M \geq R_M$. Essentially, it compares the number of haplotypes defined by a subset, S, of sites to the number of sites in S. If H_M is applied to the gene data in this section, it is found that $H_M = 7$, 11, 5, 6, 2, and 32, for the data sets β-globin, *APOE*, R, N, J, C, and All, respectively. These values should be compared to those obtained by R_M: 5, 6, 4, 3, 1, and 9, respectively. Thus, there is a clear benefit in taking extra information into account. Regions where H_M is high are indicative of hotspots, or multiple recombination events, whereas regions with low H_M (but > 0) are indicative of gene conversion and sporadic events. Unfortunately, an efficient algorithm for calculation of H_M does not exist; therefore there also cannot be an efficient algorithm for calculation of an "unphased" H_M. However, approximations of H_M have proven useful, for example, using a sliding-window approach or restricting the number of sites that are considered at the same time (see Myers and Griffiths 2003 for details). Similar techniques might be useful in defining an unphased H_M. An unphased version of H_M might also be useful in addressing questions regarding sporadic events.

This article is for μ, my wife. L. Subrahmanyan is thanked for many useful discussions and suggestions relating to the topic of this article, for advice in matters regarding empirical data, and for useful computational shortcuts. S. R. Kimura is thanked for computational advice.

LITERATURE CITED

Clark, A., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7:** 111–122.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander, 2001 High-resolution haplotype structure in the human genome. Nat. Genet. **29:** 229–232.

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

Fullerton, S. M., R. M. Harding, A. J. Boyce and J. B. Clegg, 1994 Molecular and population genetic analysis of allelic sequence diversity at human β-globin locus. Proc. Natl. Acad. Sci. USA **91:** 1805–1809.

Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor *et al.*, 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of major human polymorphism. Am. J. Hum. Genet. **67:** 881–900.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

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

Gusfield, D., 2002 Haplotyping as perfect phylogeny: conceptual framework and efficient solutions, pp. 165–175 in *Proceedings of RECOMB 2002*, edited by G. Myers, S. Hannenhalli, D. Sankoff, S. Istrail, P. Peuzner *et al.* ACM Press, New York.

Hudson, R. R., 1983 Properties of the neutral allele model with intergenic recombination. Theor. Popul. Biol. **23:** 183–201.

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

Jeffreys, A. J., L. Kauppi and R. Neumann, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29:** 217–222.

Johnson, G. C. L., L. Esposito, B. J. Baratt, A. N. Smith, J. Heward *et al.*, 2001 Haplotype tagging for the identification of common disease genes. Nat. Genet. **29:** 233–237.

Myers, S., and R. C. Griffiths, 2003 Bounds on the number of recombination events in a sample history. Genetics **163:** 375–394.

Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wiuf, C., 2002 On the minimum number of topologies explaining a sample of DNA sequences. Theor. Popul. Biol. **62:** 357–363.

Communicating editor: S. Tavaré

APPENDIX

Proof of Theorem 1. The "if" part is trivial. The "only if" goes like this. Consider two sites i and j. There can at most be two 2's in a row. Thus, all double heterozygotes in s_i and s_j are in rows with no other 2's and the resolution of these does not affect the resolution of other sites. Further, a compatible resolution of $(s_1\ s_2)$ exists because S is 2-compatible. In conclusion, all pairs of sites are haplotype compatible and S is compatible. ■

Proof of Theorem 2. The last part is trivial because $\overset{r}{\sim}$ resolves sites uniquely. Also, the "only if" is trivial. Now, suppose S has no 3-incompatibilities. The proof is by induction. If $k = 1$, 2, or 3 the theorem is trivially true; *i.e.*, S is compatible and there is a unique compatible resolution. For the induction step assume that the proposition is true for $k' < k$. The induction basis assures that the first $k - 1$ columns are compatible and the resolution is unique. Consider any 2 in the kth column. If this is the only 2 in that row it cannot create an incompatibility and the phase can be assigned arbitrarily. If more than one 2 are in the row, choose a site j and resolve k using j. This can be done in only one way because $j \overset{r}{\sim} k$. If another 2 is in the same row, say, in column i, then it cannot create an incompatibility without creating a 3-incompatibility, because $i \overset{r}{\sim} j$ and $j \overset{r}{\sim} k$. Thus, all k sites are compatible. ■

Proof of Lemma 1. Reflexivity, symmetry, and transitivity hold by definition. ■

Proof of Lemma 2. The proof is by induction on the length ($k \geq 0$) of the sequence introduced below Definition 6: If $i \overset{e}{\sim} j$, then $i_1, i_2, \ldots, i_k$ exist, such that $i \overset{w}{\sim} i_1, i_1 \overset{w}{\sim} i_2, \ldots\ldots, i_{k-1} \overset{w}{\sim} i_k, i_k \overset{w}{\sim} j$. If $k = 0$, then the result is trivially true. Assume now that it is true for $k' < k$ and consider $k' = k$. Thus, $i_1, i_2, \ldots, i_k$ exist, such that $i \overset{w}{\sim} i_1, \ldots, i_k \overset{w}{\sim} j$. If $\text{het}(x, y) = 1$ for some $x, y \in \{i, j, i_1, \ldots, i_k\}$, $\{x, y\} \neq \{i, j\}$, that are not already joined by $\overset{r}{\sim}$ in the list above, then a smaller sequence can be extracted with the same property, $i \overset{w}{\sim} i_1, i_1 \overset{w}{\sim} i_2, \ldots, i_{k_1} \overset{w}{\sim} x, x \overset{w}{\sim} y, y \overset{w}{\sim} i_{k_2}, \ldots, i_k \overset{w}{\sim} j$, and in consequence $i \overset{w}{\sim} j$. This follows by application of the induction hypothesis twice. If $\text{het}(x, y) = 0$ for all x and y, then the sites in the sequence take the form

i	i_1	i_2	…	i_{k-1}	i_k	j
2	2		…	z	z	z
z	2	2	…	z	z	z
z	z	2	…	2	z	z
z	z	z	…	2	2	z
z	z	z	…	z	2	2
2	z	z	…	z	z	2

Here, the same row might appear several times, and z is either 0 or 1 (not necessarily the same value in all places). Consider the first and the sixth row. It follows that $i \overset{r}{\sim} j$, and thus $i \overset{w}{\sim} j$, trivially. The lemma is proved. ■

Proof of Theorem 3. If $\text{het}(i, j) = 0$ for all $i \in F_1$ and $j \in F_2$, then $i \perp j$. In consequence, the set of rows with heterozygotes in F_1 and the set of rows with heterozygotes in F_2 are disjoint; otherwise there would be i and j such that $\text{het}(i, j) = 1$.

If not $\text{het}(i, j) = 0$ for all $i \in F_1$ and $j \in F_2$, choose $i \in F_1$, such that $\text{het}(i, j) = 1$ and $i < j$ for some $j \in F_2$ (*cf.* Figure 3). (The case $j < i$ is treated similarly.) All $j' \in F_2$ belong to one of three sets: $A_< = \{j' | j' < i\}$, $A_> = \{j' | j' > i\}$, or $A_\perp = \{j' | j' \perp i\}$ (j is in $A_>$). If $i < j'$ and $j' < i$ for some j', then define j' to be in $A_>$ only. It will be proven that $A_<$ is empty. Assume, oppositely, that $A_<$ is nonempty and let $j' \in A_<$. Then $j_1, j_2, \ldots, j_k \in F_2$ exist, such that $j \overset{r}{\sim} i_1, j_1 \overset{r}{\sim} j_2, \ldots, j_{k-1} \overset{r}{\sim} j_k, j_k \overset{r}{\sim} j'$. Let j_{k_1} be the first element among $j_1, \ldots, j_k$, which is not in $A_> \cup A_\perp$. It follows that $j_{k_1} < i$, and either $i < j_{k_1-1}$ or $i \perp j_{k_1-1}$. According to Figure 3 this implies that the genotype pattern schematically takes one of the following two forms:

$i < j_{k_1-1}$

j_{k_1}	i	j_{k_1-1}
2	2	2
0	2	2
0	0	z

$i \perp j_{k_1-1}$

j_{k_1}	i	j_{k_1-1}
2	2	0
0	2	0
0	z'	z

Where all 0's in a column can be replaced by 1's, rows might be repeated, $z \in \{0, 1, 2\}$, and $z' \in \{0, 1\}$. In both cases it follows that $i_{k_1} \overset{r}{\sim} i_{k_1-1}$ cannot be true. In consequence, $A_< = \emptyset$ and the theorem is proved. ■

Proof of Corollary 1. Assume $F_1 < F_2$ and $F_2 < F_1$. Then $i \in F_1$ and $j \in F_2$ exist, such that $\text{het}(i, j) = 1$, $i < j$, and $j < i$. According to Figure 3 one must have $i \approx j$. Let $i' \in F_1$ be such that, $i \overset{e}{\sim} i'$, $i' \neq i$. Then also $j \overset{e}{\sim} i'$, which contradicts that F_1 and F_2 are disjoint. In conclusion, $F_1 = \{i\}$, $F_2 = \{j\}$, and $i \approx j$. ■

Proof of Corollary 2. The corollary follows from Theorem 3. ■

Proof of Lemma 3. Only the first part needs a proof. The second part is proved in the remark above the lemma. Since $i \in E_0$ only if $\text{het}(i, j) = 0$ for all $j \neq i$, $j \in \{1, \ldots, m\}$, then there can at most be one 2 in each row of S. The result now follows from Theorem 1. ■

Proof of Lemma 4. The first part ($E_\alpha \perp E_\beta$) follows easily from Lemma 3. To prove the second part ($E_\alpha < E_\beta$) note that all rows with 2's in E_α form a subset of the rows with 2's for at least one $i \in E_\beta$ (Theorem 3). The phase of s_i can be arranged such that all heterozygotes are (0, 1). Thus, if E_α is compatible with E_β, then the phase of E_α can be arranged such that all heterozygotes in $i \in E_\beta$ are resolved (0, 1) or are resolved (1, 0). This proves the lemma. ■

Proof of Theorem 4. Assume that all terminals are compatible with a tree. Then the resolution of terminals is uniquely given. According to Lemma 4, if E_α is compatible with E_β and $E_\alpha < E_\beta$, then E_α can be resolved in either of two ways for a given resolution of E_β. Since there are $M - T$ nonterminals, it follows there are 2^{M-T} compatible resolutions. ■

Proof of Lemma 5. Note that $b_{ij}^{k+1} \geq b_{ij}^k$, thus also $B_{ij}^{k+1} \geq B_{ij}^k$ from Equation 2, and the second inequality is proved. To prove the first inequality note that R_M remains unchanged if b_{ij} is defined by

$$\begin{aligned} b_{ij} &= 1, \quad \text{if sites } i \leq i' < j' \leq j \text{ exist,} \\ & \text{such that } i' \text{ and } j' \text{ are incompatible} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

It follows from the algorithm given in Hudson and Kaplan (1985). However, b_{ij} as defined above fulfills $b_{ij} \geq b_{ij}^m$, thus $R_M \geq R_M^m \geq R_M^k$, and the inequality is proved. ■

Proof of Theorem 5. Clearly, $i_m - 1 \geq R_M^m$. To prove the converse let $I_j^m = [i_j, i_{j+1} - 1]$, $i_1 = 1$, $i_{m+1} = m + 1$. Irrespective of how phase is assigned to the sites in $I_{j-1} \cup \{i_j\}$, $2 \leq j \leq m$, either i_j is haplotype incompatible with a site in I_{j-1} or two sites in I_{j-1} are haplotype incompatible. Thus $i_m - 1 \leq R_M$ and the theorem is proved. ■