The Number of Recombination Events in a Sample History: Conflict Graph and Lower Bounds

Vineet Bafna and Vikas Bansal

Abstract—We consider the following problem: Given a set of binary sequences, determine lower bounds on the minimum number of recombinations required to explain the history of the sample, under the infinite-sites model of mutation. The problem has implications for finding recombination hotspots and for the Ancestral Recombination Graph reconstruction problem [29]. Hudson and Kaplan [15] gave a lower bound based on the four-gamete test. In practice, their bound R_m often greatly underestimates the minimum number of recombinations. The problem was recently revisited by Myers and Griffiths [22], who introduced two new lower bounds R_h and R_s which are provably better, and also yield good bounds in practice. However, the worst-case complexities of their procedures for computing R_h and R_s are exponential and super-exponential, respectively. In this paper, we show that the number of nontrivial connected components, R_c, in the *conflict graph* [4] for a given set of sequences, computable in time $O(nm^2)$, is also a lower bound on the minimum number of recombination events. We show that in many cases, R_c is a better bound than R_h. The conflict graph was used by Gusfield et al. [4] to obtain a polynomial time algorithm for the galled tree problem, which is a special case of the Ancestral Recombination Graph (ARG) reconstruction problem. Our results also offer some insight into the structural properties of this graph and are of interest for the general Ancestral Recombination Graph reconstruction problem.

Index Terms—Recombination, phylogenetic networks, ancestral recombination graph, haplotypes, lower bounds, conflict graph, NP-completeness.

1 Introduction

In the postgenomic era [16], [28], several projects (see e.g., [12]) have emerged which seek to characterize the genetic diversity of entire populations of individuals. Two of the major evolutionary forces that shape this diversity are *Mutation* and *Recombination*. A reconstruction of the likely historical mutation and recombination events that explain the divergence of the extant population of sequences from a single ancestral sequence is a challenging unsolved problem in population genetics. Clearly, all variation must start from some mutation. In the absence of recombination, each individual inherits the mutations from a single parent and adds some new ones. Each diverging site needs to mutate at least once. Under the *infinite-sites* assumption which states that each site mutates at most once, the history can be explained by a *perfect phylogeny*, which is a tree that explains the sample allowing exactly one mutation event per site. For binary characters, there is a linear time algorithm to reconstruct a perfect phylogeny, or to determine that no such phylogeny is possible [9].

It is not surprising that the problem of reconstructing evolutionary histories, even from haplotype data, becomes much harder when recombinations are considered. A recombination event results in the inheritance of genetic material from two individuals. Therefore, the history must be expressed, not by a tree, but by a general network, which is often referred to as the *Ancestral Recombination Graph* (ARG). The ARG is a directed acyclic graph with a single root (the *Most Recent Common Ancestor*), and the indegree of each node being restricted to being one or two. A node with indegree two is called a *recombinant node*. A natural parsimonious criterion for reconstructing the Ancestral Recombination Graph is to minimize the number of recombinant nodes. The problem of reconstructing Parsimonious ARGs has seen only sporadic action in the Computational Biology community. Some of the early work [13] focused on reasonable heuristics, but with no explicit performance guarantees. Wang et al. [29] considered the rooted case and showed the problem of Parsimonious ARG reconstruction to be NP-hard under the infinite sites assumption. They proposed a polynomial time algorithm for the special case of constructing *Galled trees*, which are ARGs in which every recombinant node is in its own edge disjoint cycle. Gusfield et. al. [4] showed that the algorithm of Wang et. al. is incomplete and proposed an $O(nm + n^3)$ time algorithm to solve the Galled tree problem when the root is known. They exploit the structure of the *conflict graph* on the sites for their algorithm and also show that if there is a galled tree for a set of sequences, it is also an ARG with the minimum number of recombinations. Recently, Gusfield [10] gave an algorithm for the case where the root is unknown with the same time complexity. Song and Hein [26] provide an algorithm for the Parsimonious ARG problem that involves enumerating labeled trees and computing their subtree transfer distance. While they make no attempt to describe the complexity of their algorithm, it appears to be $\Omega(n!!)$, where n is the number of sequences in M. Therefore, a

• *The authors are with the Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093-0114. E-mail: {vbafna, vibansal}@cs.ucsd.edu.*

Manuscript received 6 July 2004; revised 24 Sept. 2004; accepted 27 Sept. 2004.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0075-0704.

natural, but unresolved question is whether there is an *exponential* time algorithm to construct a Parsimonious ARG.

There is renewed interest in the ARG reconstruction problem due to the emergence of genome-wide diversity studies, such as the HapMap project [12]. These large-scale population studies have made it possible to investigate the nature and causes of recombination rate variation in the human genome. Recombination rates are known to vary greatly across the length of the genome. Initial analysis of genotype data from the human genome (Gabriel et. al. [7]), revealed an interesting haplotype block structure of the human genome, where long stretches of the genome have had little or no recombination, and the recombination events are said to be clustered in so called *recombination hotspots*. However, the arguments for this are usually based on an indirect measure of recombination, described by (loss of) *Linkage Disequilibrium*, or correlation between sites as a function of distance. Several statistical-based methods for the quantitative estimation of recombination rates from polymorphism data have recently been developed. (e.g. [8], [5], [14], [20], [19]) There is increasing evidence for recombination rate variation over small physical distances and the presence of recombination hotspots in the human genome (see, e.g., [3], [17], [24]). Recently, McVean et. al. [21] and Crawford et. al. [2] have provided model-based evidence for fine-scale variation in recombination rates throughout the human genome.

In this paper, we approach the possibly simpler problem of computing the minimum number of recombinations required to explain the history of a sample of binary sequences. Under the infinite sites assumption, lower bounds on the minimum number of recombination events can provide estimates of recombination rates and possibly detect recombination hotspots. While not as informative as ARG reconstruction, lower bounds on recombination events provide valuable information. First, tight lower bounds may prove to be useful in a branch and bound reconstruction of a Parsimonious ARG. Furthermore, the problem is useful in its own right, because the number provides a direct estimate for recombination hotspots. Fearnhead et. al. [6] applied the R_h method of Myers and Griffiths [22] to detect recombination events in the β-globin gene cluster which has a well-characterized recombination hotspot. They found that the results obtained using this method were consistent with their estimates obtained using a full likelihood method and moreover gave a better evidence for the recombination hotspot than the pairwise Linkage Disequilibrium summaries.

The problem of determining lower bounds on the number of recombination events has its own history in population genetics research. Hudson and Kaplan [15] introduced a lower bound R_m under the infinite-sites model. However, in practice, the bound often greatly underestimates the actual number of recombinations. The problem was recently revisited by Myers and Griffiths [22], who proposed two new lower bounds, R_h and R_s. They also gave a general framework in which local lower bounds on smaller subregions can be combined to get a global lower bound for longer regions. Their bound R_h used in this framework of *combining local recombination bounds* yields lower bounds that are much better in practice than R_m. They also show that $R_s \geq R_h \geq R_m$. However, their procedures for computing R_h and R_s are exponential and super-exponential, respectively. We have shown that natural formulations of the problems they are solving are NP-complete and, therefore, unlikely to admit polynomial time solutions (to be described elsewhere). A set theoretic bound was recently proposed by Song and Hein [27]. They also showed that their method obtains better bounds for some examples, but the time complexity of their procedure is not clear. Thus, the problem of computing lower bounds efficiently remains open. In this paper, we exploit the idea of the *conflict graph* introduced by Gusfield et. al. [4] to obtain an efficient recombination lower bound. In particular, we show that

1. The number of nontrivial connected components R_c in the conflict graph of a set of haplotypes is a lower bound on the number of recombinations required to explain the history of the sample under the infinite-sites model. This bound can be computed in $O(nm^2)$ time for a matrix of size $n \times m$. It was brought to our attention that this bound has been independently obtained by Gusfield and Hickerson [11].
2. The number of recombinations required does not increase if sites are deleted. Based on this idea, we introduce the **Max-NTCC** problem of deleting sites so that the number of nontrivial connected components is maximized (to give improved lower bounds). We show that this problem is NP-complete using a reduction from the Maximum Independent Set problem.
3. We show that for any given data set, $R_s \geq R_c$. Although R_c is generally weaker than R_h, there are many instances where it offers improvement over R_h. We show that for any matrix M, $R_h \geq \frac{2}{3}R_c - \frac{1}{3}$ and also provide infinitely many examples for which $R_h = \frac{2}{3}R_c$. Additionally, we show how R_c can be combined with the other bounds to reduce the computation time in practice, and show a real example where it offers improvements.

In addition, our methods for computing lower bounds employ the structure between the nonconflicting sites to provide an insight into the ARG reconstruction problem, and are of independent interest.

2 Definitions and Previous Work

We make the common infinite sites assumption which says that a polymorphic site mutates exactly once. Two polymorphic sites can be separated by few kilobases and can have multiple recombination events in between them (see, e.g., [23], [1], and [17]). As there are only two alleles at every site (the ancestral and the mutant), the extant data is represented by a binary matrix M with n rows (individuals or haplotypes) and m columns. Each column or site in the haplotype represents an SNP (single nucleotide polymorphism) which is a single base substitution of one nucleotide for another and both versions are observed in the population with frequency above a certain threshold. Very few polymorphic sites (about 0.1 percent) in humans have been found to be triallelic, i.e., having more than two

different bases at the given site. In fact, triallelic polymorphism detection is often used to flag possible experimental error (see, e.g., [25]). Since many of the SNP loci are neutral, a significant violation of the infinite-sites assumption should result in a much higher fraction of triallelic polymorphic sites. Therefore, it is reasonable to make the "infinite-sites" or no-homoplasy assumption while dealing with human polymorphism data. Moreover, as every site has at most two nucleotides, they can arbitrarily be renamed 0 and 1. Hence, all our results on binary character data are applicable to real haplotype data. Next, we give a formal definition of a recombination event.

A haplotype of length n is simply a binary string of length n. A recombination event at site p, between two haplotypes A and B, produces a recombinant sequence C, which is either a concatenation of sites $A[1 \dots p-1]$ with $B[p \dots m]$ or $B[1 \dots p-1]$ with $A[p \dots m]$.

A phylogenetic network or an Ancestral Recombination graph G for a set M of n sequences is a directed acyclic graph with a root. The root has no incoming edges. Each node in G is labeled by a m-length binary sequence where m is the number of sites. Each leaf of this graph is labeled by a sequence in M. Each node other than the root has either one or two incoming edges. A node with two incoming edges is called a *recombination* node. Some of the edges are labeled by the columns (sites) of M which correspond to a mutation event at that site. For a nonrecombination node v, let e be the single incoming edge into v. The sequence labeling v can be obtained from the sequence labeling v's parent by changing the value at the sites which label the edge e from 0 to 1 (assuming that the root sequence is all-0). Each recombination node v is associated with an integer r_v (in the range $[2, m]$), called the recombination point for v. Corresponding to the recombination at node v, one of the two sequences labeling the parents of v is denoted as P and the other one as S. The sequence labeling node v is a concatenation of the first $r_v - 1$ characters of P with the last $m - r_v + 1$ characters of P. The sequences labeling the leaves of the phylogenetic network are referred to as *extant* sequences.

A recombination graph G explains a set M of n sequences iff each sequence labels exactly one of the leaves of G. For brevity, we denote an ancestral recombination graph by *ARG*. An ancestral recombination graph is a stochastic process which generates populations given appropriate parameters, however, given a population that was generated through an ancestral recombination graph, the problem of reconstructing the ARG refers to the reconstruction of the digraph. In this paper, we refer to the digraph itself as the ancestral recombination graph.

We recall some standard definitions below:

Definition 1. *Two columns (sites) i and j in M are said to be in conflict if there is set of four rows with the pairs $\{00, 01, 10, 11\}$ in these two columns. If the ancestral type at each site is known, the presence of three rows with the values $\{01, 10, 11\}$ in these two columns implies a conflict, since we can infer the existence of the ancestral type 00. A pair of columns (i, j) is said to be compatible if i and j do not conflict.*

Definition 2. *We define m_M to be the minimum number of recombinations required to explain M, i.e., there exists a ARG with m_M number of recombinations which explains M and there is no ARG with fewer recombinations that explains M.*

Definition 3 (from [4]). *The conflict graph $G_C(M) = (V, E)$ for a given set M of n sequences is a graph with vertex set $V = \{i |\ i$ is a column of $M\}$ and $E = \{(i, j) |$ columns i and j conflict $\}$. Note that matrix M defines an ordering for the vertices of $G_C(M)$. We define two edges (a, b) and (c, d) in $G_C(M)$ to be* noninterleaving *if $\max\{a, b\} < \min\{c, d\}$ or $\max\{c, d\} < \min\{a, b\}$.*

Definition 4. *We define $R_c(M)$ to be the number of non-trivial connected components (components of size more than one) in the conflict graph of a set of sequences M.*

2.1 Previous Lower Bounds

R_m: The lower bound R_m of Hudson and Kaplan [15] is based on the simple fact that if there is a conflict between sites i and j, $(i < j)$, then one can infer a recombination event in the interval (s_i, s_j). In terms of the conflict graph, this bound can be viewed as finding a maximum set of noninterleaving edges in the conflict graph. The number of edges gives a lower bound on the number of recombinations.

R_h: The bound R_h is based on the observation that if we have a set of n sequences and m sites, then at least $n - m - 1$ sequences must have been created through recombination and since a single recombination event can create at most one new sequence, at least $n - m - 1$ recombination events are required. For a given matrix M, let $M(S)$ be the submatrix created by choosing a subset S of columns in a matrix M. Let $D(M(S))$ denote the set of distinct rows in $M(S)$. Then,

$$R_h(M) = \max_S(|D(M(S))| - |S| - 1).$$

If the number of segregating sites is S, the number of subsets to be considered in order to compute the bound R_h is 2^S. Myers and Griffiths propose a simple heuristic of considering only subsets of size at most S' where S' is an input parameter. Hence, the worst-case complexity of their procedure is exponential in the number of sites.

R_s: Myers and Griffiths [22] only provided a procedural definition of the bound R_s. Their algorithm performs three kinds of operations on a given matrix: row deletion, column deletion, and nonredundant row removal. A row deletion can be performed if the row in the matrix is identical to another row. Such a row is also referred to as a *redundant* row. A column deletion can be done if the column (site) is noninformative. A nonredundant row removal is a row removal when there are no noninformative sites in the matrix and no duplicate rows. In this paper, we also refer to a nonredundant row removal as simply row removal. Informally, the algorithm for computing the bound R_s performs a sequence of column deletions, row deletions, and nonredundant row removals until there is no row left in the matrix M. A concise description of the algorithm is given in Fig. 1.

The above procedure for computing R_s considers all possible orderings for the n rows in the matrix and, hence, the worst-case complexity of this procedure is $O(n!)$ which is superexponential in n. Apart from proposing the bounds

procedure *Compute_*R_s(M)
if M is empty
  return 0;
else if column s is non-informative (* Minor allele appears at most once*)
  return (*Compute_*R_s(M-{s}));
else if row h is redundant (* Equal to some other row *)
  return (*Compute_*R_s(M-{h}));
else
  return $(1 + \min_{h \in M}\{$*Compute_*$R_s(M - \{h\})\})$
end if

Fig. 1. $Compute\_R_s$ is a procedural definition of the bound R_s as described in Myers and Griffiths [22]. In the algorithm, h (haplotype) refers to a row, and s (site) refers to a column in matrix M.

R_h and R_s, Myers and Griffiths [22] presented a general framework for combining local recombination bounds on continuous subregions of a larger region to obtain recombination bounds for the larger parent region. Consider a matrix M with S sites labeled 1 to S and suppose we have a local bound B_{ij} on the number of recombination events in the region (i, j), $1 \leq i < j \leq S$. Then, their method, which is essentially a dynamic programming algorithm, can use these local bounds to compute the minimum number of recombination events between every pair of sites in the sample. This new bound for an interval (i, j) is denoted as R_{ij} and, in particular, R_{1S} gives the bound for the whole region. Note that the bound obtained using the dynamic programming algorithm for a region can be better than the local bound for that region, i.e., R_{ij} can be larger than B_{ij}.

2.2 A Road Map

In Section 3, we prove our main result, i.e., the number of nontrivial connected components in the conflict graph is a lower bound on the number of recombinations. In Section 3.1, we propose some extensions and ideas for improving this bound, and show a hardness result for the Maximum Nontrival Connected Components (Max-NTCC) problem. Finally, in Section 4, we compare this new bound R_c with the bounds R_m and R_h, both theoretically, and with some examples, and discuss possible scenarios in which the bound R_c can be used.

3 Connected Components in the Conflict Graph

We begin by showing that removing sites from the matrix corresponding to the given set of sequences does not increase the required number of recombination events.

Definition 5. *Let M be an $n \times m$ matrix, and $S = \{1, 2, \ldots, m\}$ denote the set of sites in M. For a subset $S' \subseteq S$, let $M(S')$ denote the submatrix obtained by restricting columns to be in S', and removing all redundant rows. For a site s, denote $M(S - \{s\})$ by M_{-s}.*

Lemma 1. *For any subset $S' \subseteq \{1, 2, \ldots, m\}$ of a matrix M, $m_{M(S')} \leq m_M$ and, therefore, any lower bound on the number of recombinations for the matrix $M(S')$ is also a lower bound on m_M.*

Proof. Consider an ancestral recombination graph $G(M)$ explaining M. For any arbitrary site $s \in S$, we show how to transform $G(M)$ into a ancestral recombination graph explaining M_{-s} without increasing the number of recombination cycles in $G(M)$. Consider the edge $e = (u, v)$ labeled with s in $G(M)$. If the edge e is labeled with other sites as well, we simply delete the label s and keep the ancestral recombination graph unchanged. Therefore, the interesting case is when s is the only label on the edge e. If v is a leaf node, we simply remove e and the node v. Clearly, this can only happen if the sequence labeling the leaf node was the only sequence in M with a mutation on c, and hence is not present in M_{-s}.

In the case where v is an internal node in $G(M)$, we collapse the edge e and make u to be the starting vertex of all outgoing edges from v. It is easy to see that collapsing the edge e does not induce a cycle in the graph $G(M)$. Hence, the modified graph is an ARG for the matrix M_{-s}. Using an inductive argument, it is easy to see that

$$m_{M(S')} \leq m_M \quad \forall S' \subseteq S.$$

□

Corollary 1. *Let S' denote a subset of the sites S in M. Then, $\max_{S' \subseteq S} m_{M(S')}$ is a lower bound on m_M.*

Lemma 2. *For a matrix M and a set of sites $S' \subseteq S$, $R_s(M(S')) \leq R_s(M) \leq m_M$.*

Proof. Observe that any sequence of noninformative column deletions, row deletions, and nonredundant row removal (from the definition of bound R_s) operations that reduces the matrix M to an empty matrix, also reduces the matrix $M(S')$ to an empty matrix. Hence, $R_s(M(S'))$ which is the number of row removal operations in a sequence which uses the minimum number of row removal operations, is at most $R_s(M)$. □

The basic idea behind the proof of the connected components lower bound is based on the computation of R_s. In the R_s computation, we delete rows and columns, but we only charge to a recombination event when we are deleting a nonredundant row. We will show essentially that a row that is nonredundant when restricted to sites in a connected component MUST be redundant when restricted to sites of any other connected component. In order to do this, we must prove a structural property of two connected components described in the *2-edge theorem (Theorem 1)*. The

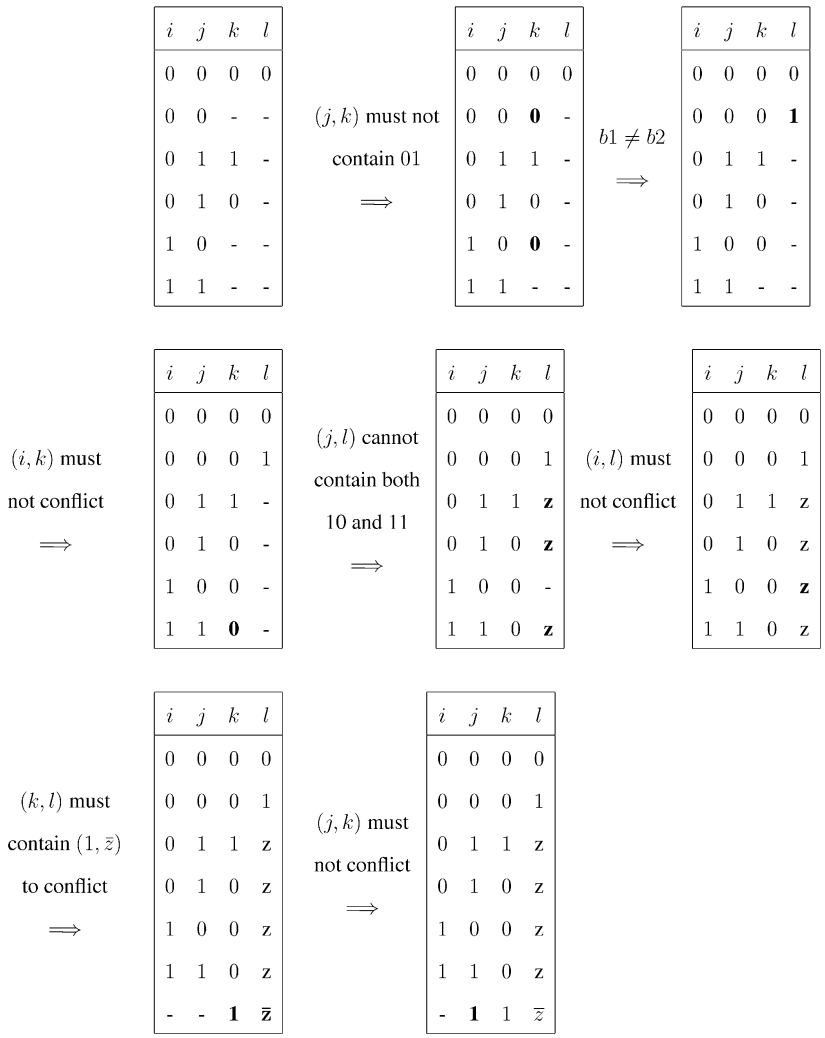

Fig. 2. xxxxx.

proof of this theorem depends on two technical lemmas which we prove next.

Lemma 3. *Let $S = \{i, j, k, l\}$ be four columns in a matrix M such that (i, j) and (k, l) conflict and (i, k), (i, l), (j, k), and (j, l) are compatible. Then, there is at most one pair $a \in \{00, 01, 10, 11\}$ such that $[ab_1]$ and $[ab_2]$ are distinct rows in $M(S)$, where $b_1, b_2 \in \{00, 01, 10, 11\}$.*

Proof. The proof is by contradiction. Suppose there are four distinct rows $[a_1b_1], [a_1b_2], [a_2b_3], [a_2b_4]$ in $M(S)$. Without loss of generality, we can assume that $a_1 = b_1 = 00$ (we can relabel the columns without changing the conflict graph). We now consider two cases where $a_2 = 01$ and $a_2 = 11$. Since the ordering of the columns is not important, the case where $a_2 = 10$ is the same as $a_2 = 01$.

Case $\mathbf{a_2 = 01}$. Since $b_3 \neq b_4$, they differ in at least one of the sites $\{k, l\}$. We can assume that they differ in the column k (as the order of the columns is not relevant). Also, as (i, j) conflict, the rows $a_3 = 10$, and $a_4 = 11$ exist. We show that the remaining values in this matrix are forced and lead to a contradiction (see Fig. 2).

Now (j, l) contains $00, 01, 1z$, and $1\bar{z}$, a contradiction!

Case $\mathbf{a_2 = 11}$. The argument for this case proceeds along the same lines as the previous one. Since the only conflicts possible are between the sites (i, j) and (k, l), the submatrix is constrained to contain the following distinct set of rows:

i	j	k	l
0	0	0	0
0	0	0	1
1	1	1	z
1	1	0	z
1	0	0	z
0	1	0	z
x	y	1	$\bar{z}$

If $y = 0$, then (j, k) conflict. If $y = 1$, then (j, l) conflict, a contradiction! □

Lemma 4. *Let $S = \{i, j, k, l\}$ be four columns in a matrix M such that (i, j) and (k, l) conflict and (i, k), (i, l), (j, k), and (j, l) are compatible. There, the submatrix $M(S)$ does not have three distinct rows of the form $[a_1b_1], [a_2b_2], [a_3b_3]$, where $a_1 \neq a_2 \neq a_3$ and $b_1 \neq b_2 \neq b_3$.*

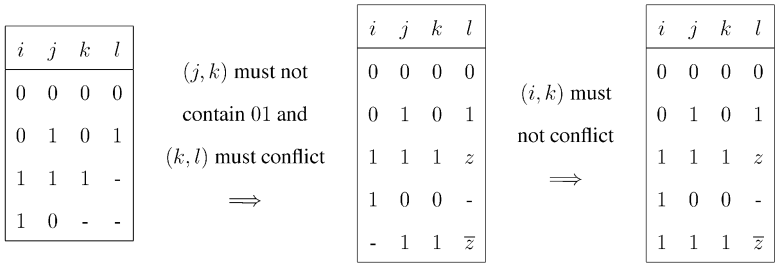

Fig. 3. xxxxx.

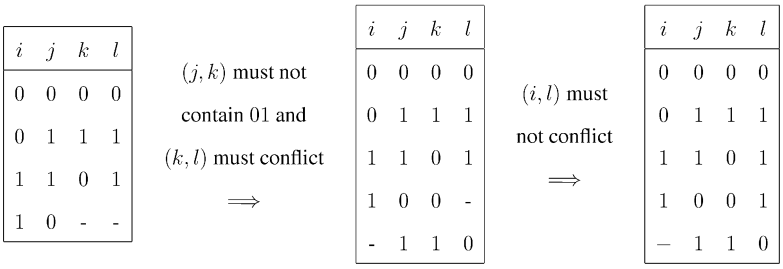

Fig. 4. xxxxx.

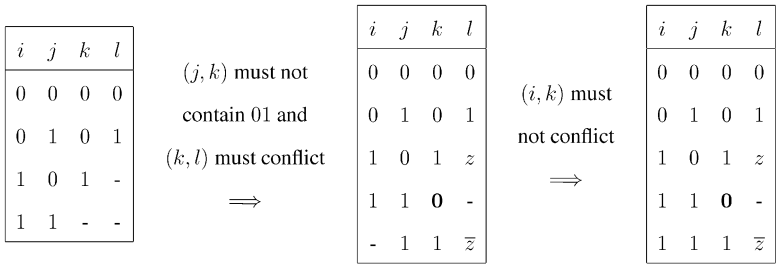

Fig. 5. xxxxx.

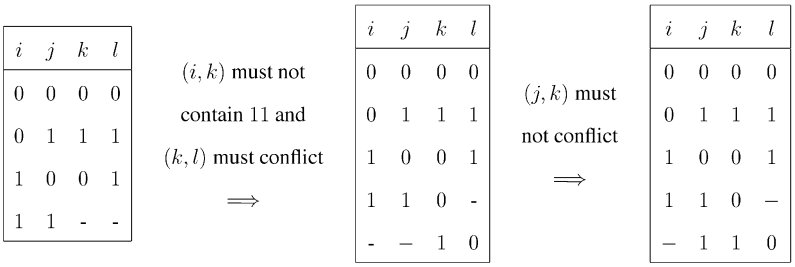

Fig. 6. xxxxx.

Proof. Suppose, to the contrary, we have three distinct rows $[a_1b_1], [a_2b_2], [a_3b_3]$ in $M(S)$. Without loss of generality, we can assume that $a_1 = b_1 = 00$ and $a_2 = 01$ (we can relabel the columns without changing the conflict graph and the ordering of the rows and columns is not important). As before, we proceed using a case-by-case analysis.

Case: $\mathbf{a_3 = 11}$ and $\mathbf{b_2 = 01}$ (see Fig. 3).

But, now, (i,l) contains $00, 01, 1z$, and $1, \bar{z}$, a contradiction!

Case: $\mathbf{a_3 = 11}$ and $\mathbf{b_2 = 11}$ (see Fig. 4).

Now, the pair of columns (j,l) conflict, which is a contradiction.

Note that, since the ordering of the columns does not matter, the cases where $b_2 = 10$ is identical to the case where $b_2 = 01$.

Case: $\mathbf{a_3 = 10}$ and $\mathbf{b_2 = 01}$ (see Fig. 5).

But, now, (i,l) contains $00, 01, 1z$, and $1, \bar{z}$, hence a conflict.

Case: $\mathbf{a_3 = 10}$ and $\mathbf{b_2 = 11}$ (see Fig. 6).

Now, the pair of columns (j,l) conflict, which is a contradiction.

This completes the proof of the lemma. □

Theorem 1 (2-edge theorem). *Let $S = \{i,j,k,l\}$ be four columns in a matrix M such that (i,j) and (k,l) conflict and (i,k), (i,l), (j,k), and (j,l) are compatible. Then, there exists pairs a_{ij} and b_{kl}, such that every row in the submatrix $M(S)$ is of the form $[a_{ij}b]$ or $[ab_{kl}]$, where $a_{ij}, b_{kl}, a, b \in \{00, 01, 10, 11\}$.*

Proof. Consider three distinct rows in the submatrix $M(S)$ of the form $[a_1b_1], [a_2b_2], [a_3b_3]$, where $a_1 \neq a_2 \neq a_3$. From Lemma 4, it follows that it cannot be the case that $b_1 \neq b_2 \neq b_3$.

First, we consider the special case where $b_1 = b_2 = b_3$. Then, consider the three rows $[a_4b_4], [a_5b_5], [a_6b_6]$, where $b_4 \neq b_5 \neq b_6 \neq b_1$. Since the columns (k, l) conflict, three such rows exist. Now, if $a_4 = a_5 = a_6$, then there cannot be a row of the form $[ab]$ where $a \in \{a_1, a_2, a_3\}$ and $b \neq b_1$ since, then, we would have four distinct rows $[ab_1], [ab], [a_4b_4], [a_4b_5]$ which violate Lemma 3. Hence, the only other row we can have is $[a_4b_1]$ and, therefore, the lemma holds with $a_{ij} = a_4$ and $b_{kl} = b_1$. If we have $a_4 = a_5 \neq a_6$, then we cannot have another row of the form $[ab]$ where $a \neq a_4$, since then we would again have four distinct rows violating the constraint of Lemma 3. Hence, the lemma is satisfied with $a_{ij} = a_4$ and $b_{kl} = b_1$.

In the case where $b_1 = b_2 \neq b_3$, Lemma 3 enforces that there is exactly one row of the form $[a_4b_4]$ where b_4 is different from both b_1 and b_3 and $a_4 \in \{00, 01, 10, 11\}$. Similarly, there is one and only one row of the form $[a_5b_5]$ where b_5 is different from either of b_1, b_3, or b_4 and $a_5 \in \{00, 01, 10, 11\}$. If $a_4 \neq a_3$ and $a_4 \neq a_2$, then the three rows $[a_2b_1], [a_3b_3], [a_4b_4]$ violate Lemma 4. Applying Lemma 4 to the three rows $[a_1b_1], [a_3b_2], [a_4b_4]$, we get that either $a_4 = a_1$ or $a_4 = a_3$. But, both the previous constraints can be satisfied if and only if $a_4 = a_3$, since $a_1 \neq a_2$. Using similar arguments, it follows that $a_5 = a_3$. Hence, the lemma is true with $a_{ij} = a_3$ and $b_{kl} = b_1$. □

It has been brought to our attention that the above theorem appears in a different form in a paper by Song and Hein (see [27]: Lemma 3).

Definition 6. *Let $\{i, j, k, l\}$ be four sites in a matrix M such that the pairs (i, j) and (k, l) conflict and $(i, k), (i, l), (j, k), (j, l)$ are compatible. Then, we denote by a_{ij} and b_{kl} the pairs, such that every row in the submatrix $M(S)$ is of the form $[a_{ij}b]$ or $[ab_{kl}]$.*

The next lemma explains how a nonredundant row removal (a row removal is said to be nonredundant when a column deletion or a row deletion cannot be done) can only destroy one connected component. The connected component theorem will follow from a simple application of this lemma and Lemma 1.

Lemma 5. *Consider a matrix M such that each connected component in the conflict graph $G_C(M)$ has size 2, i.e., it consists of two sites which are in conflict. Then, $R_s(M) \geq R_c(M)$.*

Proof. We show that every possible sequence of nonredundant row removal events which reduces a matrix M (whose conflict graph has the structure described above) to the empty matrix, requires at least $R_c(M)$ nonredundant row removal operations. Since the bound R_s is the minimum number of row removal operations performed for some sequence of nonredundant row removal events, it follows that $R_s(M) \geq R_c(M)$. We claim that any column or row deletion cannot remove an edge in the conflict graph (or, equivalently, destroy a connected component). A site is deleted only when it is noninformative, i.e., there is either only a single haplotype with a 1 at that site or a single haplotype with a 0 at that site. Clearly, such a site cannot be involved in a conflict with another site, since one needs at least 2 ones and 2 zeroes at a site for it to be involved in a conflict. Similarly, a row is deleted when there are two identical haplotypes. Clearly, a removal of one of them cannot remove a conflict between two sites.

Now, consider a nonredundant row removal operation which destroys a conflict between two sites (i, j), i.e., it removes one of pairs $\{00, 10, 01, 11\}$ from the two columns. Denote the pairs removed by (ab). Consider a pair of conflicting sites (k, l) and the submatrix $M(S)$ restricted to the four sites $S = \{i, j, k, l\}$. Clearly, $ab \neq a_{ij}$ (where a_{ij} is as defined above), since a_{ij} is present in more than one row of $M(S)$. From the 2-edge lemma, it follows that the pair in the columns (k, l) in the row that was removed is also present in other rows in $M(S)$. Hence, the removal of the row containing (ab) in the columns (i, j) cannot destroy a conflict between the sites (k, l). Hence, a nonredundant row removal operation can destroy at most one connected component (or conflict) in the conflict graph. Therefore, by induction, any sequence of column deletions, row deletions, or nonredundant row removal events requires at least $R_c(M)$ nonredundant row removals to reduce the matrix to the empty matrix. □

Theorem 2. *For every matrix M, $R_c(M) \leq R_s(M) \leq m_M$.*

Proof. For every nontrivial connected component in $R_c(M)$, we remove sites such that only two conflicting sites remain in each connected component. Clearly, we can do this for every connected component with two or more sites. Hence, after removal of a subset of sites, we have a reduced matrix $M(S')$ where S' is the set of remaining sites. From Lemma 2, $m_M \geq R_s(M(S'))$. Also from Lemma 5, it follows that $R_s(M(S')) \geq R_c(M)$, which proves the required result.

Note that the 2-edge theorem (Theorem 1) imposes a strong structure on the underlying matrix. We can extend this theorem to the general case where we have a connected component instead of an edge. The theorem below, shows that the rows conferring haplotype diversity to a connected component are disjoint for each connected component. □

Theorem 3. *Let A, B be disjoint subsets of columns representing distinct connected components in the conflict graph. Let $a_1, \ldots, a_k$ and $b_1, \ldots, b_l$ be the distinct rows (haplotypes) in $M(A)$ and $M(B)$, respectively. There exist haplotypes a_i and b_j, such that all distinct rows of the matrix $M[A \cup B]$ are of the type $[a_i\ b]$ for some $b \in \{b_1, \ldots, b_l\}$, or $[a\ b_j]$ for some $a \in \{a_1, \ldots, a_k\}$.*

Proof. We prove by induction on the total number of columns in M. As the two components are nontrivial, M has at least four columns containing an edge in each component.

Base case (four columns): Each component has four distinct haplotypes $00, 01, 10$, and 11. The base case follows directly from Theorem 1.

Induction step ($k + 1$ columns): Assume that the hypothesis is true for all matrices containing two nontrivial components with a total of k columns. Let A be the component with the larger number of columns.

Remove a column i from A to get A', such that the columns in A' still form a single connected component.[1]

By the induction hypothesis, there exist haplotypes a'_i and b_j, such that all distinct rows of $M[A' \cup B]$ are of the type $[a'_i\ b]$ for some $b \in \{b_1, \ldots, b_l\}$, or $[a\ b_j]$ for some $a \in \{a_1, \ldots, a_k\}$. Thus, the distinct rows of $M(A' \cup B)$ are:

A'	B
a'_i	b_1
a'_i	b_2
$\vdots$	
a'_i	b_l
a'_1	b_j
a'_2	b_j
$\vdots$	
a'_k	b_j

Now, add the ith column back to get $M(A \cup B)$. Consider all the rows containing a'_i. We claim that all rows containing a'_i must have the same value z in the ith column. If this is true, the rows of $M(A \cup B)$ are

i	A'	B
z	a'_i	b_1
z	a'_i	b_2
	$\vdots$	
z	a'_i	b_l
	a'_1	b_j
	a'_2	b_j
	$\vdots$	
	a'_k	b_j

Let $a_i = [z\ a'_i]$. Then, each row is of the form $[a_i\ b]$, or $[a\ b_j]$, and we are done.

Next, consider the case when the rows containing a'_i have instances of z and $\bar{z}$ in the ith column. Without loss of generality, rename the haplotypes of B so that the rows of $M(A \cup B)$ contain

i	A'	B
z	a'_i	b_1
$\bar{z}$	a'_i	b_2
	$\vdots$	
	a'_i	b_l
	a'_1	b_j
	a'_2	b_j
	$\vdots$	
	a'_k	b_j

1. Such a column always exists since one can remove a vertex from a connected graph such that the remaining vertices still form a connected graph.

Next, consider an arbitrary column $j \in A'$ such that i, j conflict, and denote the value of row a'_i in column j as x. Consider the connected component $X = \{i, j\}$, and B. As (i, j) conflict, all four rows $zx, \bar{z}x, z\bar{x}, \bar{z}\bar{x}$ must appear. On the other hand, as a'_i only contains x, all rows containing $\bar{x}$ must line up against b_j. The columns of $M(X \cup B)$ contain

X	B
zx	b_1
$\bar{z}x$	b_2
$\vdots$	
	b_l
$z\bar{x}$	b_j
$\bar{z}\bar{x}$	b_j
$\vdots$	
	b_j

It is easy to verify that the components X and B violate the inductive hypothesis even though they $X \cup B$ has fewer than k columns, a contradiction! □

Definition 7. *For a pair of nontrivial connected components (A, B), we denote the common haplotype of A with respect to B as $h(A, B)$.*

3.1 Extensions to the R_c Lower Bound

In this section, we show how the connected component lower bound can be combined with previous bounds. We also introduce the MAX-NTCC problem for finding the subset of columns in a given matrix with the maximum number of nontrivial connected components.

We begin by proving that we can apply the lower bound R_s independently to each connected component of the conflict graph of a matrix M to obtain a recombination lower bound for M.

Lemma 6.

$$\sum_{C \in \mathcal{CC}} R_h(M(C)) \leq \sum_{C \in \mathcal{CC}} R_s(M(C)) \leq R_s(M) \leq m_M.$$

Proof. Consider an optimal history for R_s for the matrix M, i.e., a sequence of column deletions, row deletions, and nonredundant row removal events which reduces a matrix M to the empty matrix and requires $R_s(M)$ nonredundant row removal operations. This history can be used to obtain R_s histories for each connected component as follows: Consider any nonredundant row removal operation in the optimal history and let r denote the row removed. There is at least one connected component C such that the row r is nonredundant in the submatrix restricted to the sites in C. Let C' be another connected component in the conflict graph of M. We claim that the row r is redundant is the submatrix restricted to the sites in this component C'. If this was not the case, it would contradict Theorem 3. Hence, the row r is nonredundant in the submatrix restricted to the sites of exactly one connected component. Therefore, every

nonredundant row removal can be assigned to the R_s history of one connected component. For all other components, this row removal corresponds to a row deletion event in the history. Therefore, the sequence of column deletion, row deletion, and row removal operations in the history of a connected component is identical to that in the optimal R_s history for M. Moreover, the total number of nonredundant row removal operations summed over histories of all connected components is exactly $R_s(M)$. It follows that $\sum_{C\in\mathcal{CC}} R_s(M(C)) \leq R_s(M)$. Since $R_s(S) \geq R_h(S)$ for any set of sites S, the sum of R_h bounds on the connected components is also a valid lower bound. □

Note that computing R_s is intractable in general. However, we can possibly speed up the computation of the R_s bound by computing the bound independently on each connected component of the conflict graph. Moreover, in practice, the above lemma can be used to obtain improved bounds as follows: For each connected component, it is easy to check in polynomial time whether $R_s = 1$ or $R_s > 1$. If it is the case that $R_s > 1$, then one can infer a recombination bound of two for the sites in the connected component instead of one. We illustrate this on a real data set in Section 4.1. Moreover, the above lemma also allows us to combine the bounds R_h and R_c. For a given set of sites, one can get a bound that is at least as good as the maximum of the bounds R_h and R_c.

We can use Lemma 1 in conjunction with the connected component lower bound to obtain a somewhat stronger lower bound on m_M. For a subset S' of the sites in M, let $\mathcal{CC}(S')$ denote the nontrivial connected components in the conflict graph for $M(S')$. For every matrix M,

$$\max_{S'\subseteq S} |\mathcal{CC}(S')| \leq m_M. \tag{1}$$

Based on this observation, we define the Max-NTCC problem for finding a subset of sites whose conflict graph has the maximum number of nontrivial connected components. Unfortunately, we show that this problem is NP-complete.

Max-NTCC problem:

Input: A matrix M with n sequences and a set S of s sites.

Output: $S' \subseteq S$, such that the number of nontrivial connected components in the conflict graph of $M(S')$ is $\geq k$.

Theorem 4. *The Max-NTCC problem is NP-complete.*

Proof. It is easy to see that the problem is in NP. To prove the NP-completeness, we give a reduction from the Independent Set problem. The independent set problem is defined as follows: Given an undirected graph $G = (V, E)$, is there a subset V' of V of cardinality $\geq k$ such that there is no edge between any pair of vertices in V'.

We construct a matrix M with $2|V|$ sites (columns) and $3|V| + 3|E|$ rows as follows: Label the nodes in V arbitrarily from 1 to $|V|$. For every vertex v_i in V, we define two sites v_i and v_i'. We initially start with no rows and add new rows to the matrix keeping the number of columns fixed as $2|V|$. For every vertex v_i, we add three rows with the pairs $\{01, 10, 11\}$ in the columns $\{v_i, v_i'\}$ and with value 0 in all other columns. Hence, we obtain a matrix with $3|V|$ rows, such that there is a conflict between the sites $\{v_i, v_i'\}, 1 \leq i \leq |V|$ and no other conflicts. Now, for every edge $(v_i, v_j) \in E$, we add three new rows with the pairs $\{01, 10, 11\}$ in the columns $\{v_i, v_j\}$ and with value 0 in all other columns. As a result, we obtain a matrix M with $2|V|$ columns and $3|V| + 3|E|$ rows. We claim that the only edges in the conflict graph for this matrix are of the form $\{v_i, v_i'\}$, $v_i \in V$, or $\{v_i, v_j\}$, where $(v_i, v_j) \in E$. This is true since the only pairs of columns for which there is a row with pair $\{11\}$ are $\{v_i, v_i'\}, v_i \in V$, and $\{v_i, v_j\}$, where $(v_i, v_j) \in E$.

Suppose that there exists an independent set $V' \subseteq V$ of cardinality k in G. Consider the conflict graph for $M(S')$ where $S' = \cup_{u\in V'}\{u, u'\}$. Each pair of sites $\{u, u'\}, u \in V'$ forms a connected component of size 2, since there is no conflict between a pair of sites (u, v) where $u, v \in V'$. Hence, the conflict graph $G_C(M(S)'))$ has k nontrivial connected components.

Now, let $S' \subseteq S$ be such that the conflict graph for $M(S')$ has k nontrivial connected components. It is easy to see that every nontrivial connected component has at least one nonprimed vertex $u \in V$. For each connected component, we choose one nonprimed vertex to form the set $I \subseteq V$. Now, I is an independent set in G since if there was an edge between two vertices in I then they would have been in the same connected component in $\mathcal{CC}(M(S'))$ and, therefore, not both in the set I. Also, the independent set I has cardinality k. Hence, there is an independent set $V' \subseteq V$ of cardinality at least k iff there exists a subset S' of S such that the conflict graph of $M(S')$ has k nontrivial connected components. □

4 Comparison of R_c with Other Bounds

In this section, we compare R_c to the bounds R_m and R_h. We have already proved (see Theorem 2) that the history-based bound R_s is at least as good as the bound R_c, however, it is not feasible to compute R_s for a set of 10 or more haplotypes (see [22]). First, we observe that if we apply the R_c method to subsets of continuous columns, and compute the best bound using dynamic programming on the local lower bounds, then R_c can never be worse than R_m. Note that the running times for computing the bounds R_h and R_s are exponential and superexponential, respectively. In general, the best lower bound that can be obtained using the connected component approach is $m/2$, where m is the number of columns. If the number of distinct haplotypes $n > m + m/2$, then the bound R_h is trivially better than the connected component lower bound. For example, for a set of 2^k haplotypes with k columns, the R_h bound is exponentially better than the connected component lower bound. For regions of low diversity in haplotype data, the connected component lower bound can possibly offer better bounds that the haplotype diversity bound R_h. Here, we provide an example of a matrix M for which $R_h = \frac{2}{3}|\mathcal{CC}|$, where $\mathcal{CC}$ is the set of the nontrivial connected components in the conflict graph for M. Although this example is not real, it serves to illustrate the kind of haplotype data for which the bound R_c could offer improvements over the bound R_h.

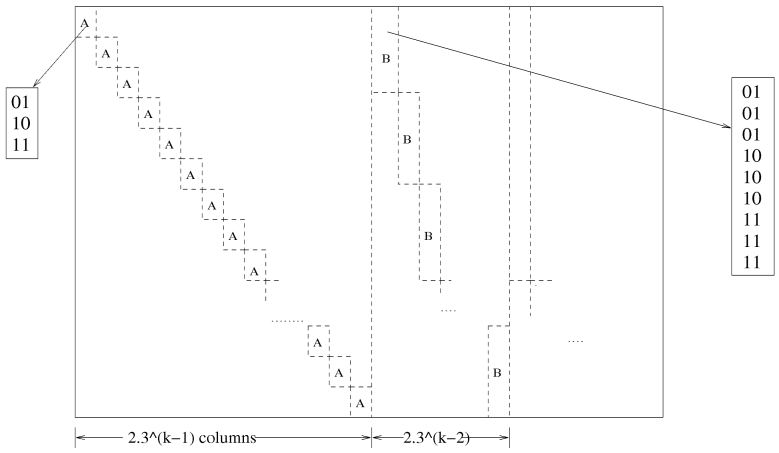

Fig. 7. The structure of the matrix M for which $R_h = \frac{2}{3}R_c$.

Theorem 5. *For all n_0, there exists a matrix M with $n \geq n_0$ rows such that $R_h(M) = \frac{2}{3}R_c(M)$.*

Proof. We choose the number of rows for the matrix M to be $3^k \geq n_0$, where $k \geq 2$. Starting from an empty matrix, we add new columns keeping the number of rows fixed. We add columns in groups of 2, which represent a connected component in the conflict graph. The following procedure defines the matrix M (depicted in Fig. 7):

1. **for** $j = 1$ to $k-1$
2. **for** $i = 0$ to $3^{k-j} - 1$ do
3. add two new columns with the following values:
4. 01 in 3^{j-1} rows starting from row $3^{j-1}(3i+1)$
5. 10 in 3^{j-1} rows starting from row $3^{j-1}(3i+2)$
6. 11 in 3^{j-1} rows starting from row $3^{j-1}(3i+3)$
7. 00 in the remaining rows

Claim. *Every column conflicts with only one other column.*

Proof. Consider a column i where $0 \leq i \leq 2 \cdot 3^{k-1} - 1$. There are only two rows with a 1 at this site and every other site (apart from the one site this site conflicts with) has the same value in these two rows. Hence, every site $i, 0 \leq i \leq 2 \cdot 3^{k-1} - 1$ is involved in only one conflict. Next, consider a column j where $2 \cdot 3^{k-1} \leq j \leq 2 \cdot 3^{k-1} + 2 \cdot 3^{k-2} - 1$. There are six rows with a 1 at this site and every other site (except the column with which j conflicts) has the same value in these rows. Hence, every site j in the range $[2 \cdot 3^{k-1}, 2 \cdot 3^{k-1} - 1 + 2 \cdot 3^{k-2} - 1]$ is involved in only one conflict. Similar arguments are applicable to each submatrix added in steps 2-6 of the procedure above which describes the matrix M. Hence, the required property holds for every column in the matrix M. □

Lemma 7. *R_h for the matrix M (constructed above) is exactly $\frac{2}{3}\mathcal{CC}$.*

Proof. After adding new rows as defined by this procedure, the matrix M has $2(3^{k-1} + 3^{k-2} + \ldots + 3) = 3^k - 3$ columns. Also, there is a conflict between any two new columns added. Therefore, we have $\left[\frac{3^k-3}{2}\right]$ nontrivial connected components. Hence, $|\mathcal{CC}| = \left[\frac{3^k-3}{2}\right]$.

Observe that the first two columns are the only columns that can distinguish rows 1, 2, and 3. The next two columns are the only columns that can distinguish rows 4, 5, and 6. In general, columns $2i$ and $2i+1$ are the only columns that can distinguish between rows $3i$, $3i+1$, and $3i+2$, $0 \leq i \leq 3^{k-1} - 1$. Let I be the set of the first 3^{k-1} sites. Observe that $|D(M_I)| = 3^k$. Restricting the matrix to the first $2 \cdot 3^{k-1}$ sites, we obtain $R_h \geq 3^k - 2 \cdot 3^{k-1} - 1 = 3^{k-1} - 1$. Now, we need to show that for every subset of rows S, $(|D(M_S)| - |S| - 1) \leq 3^{k-1} - 1$. Suppose, on the contrary, there is a subset S for which $(|D(M_S)| - |S| - 1) > 3^{k-1} - 1$. If S does not contain a pair of columns $(2i, 2i+1)$, $0 \leq i < 3k-1$, then we can add the pair of columns S to obtain a set of columns S' such that $(|D(M'_S)| - |S'| - 1) > (|D(M_S)| - |S| - 1)$. In the other case, where S does not contain one of the columns $(2i, 2i+1)$, we can add that column to get a set S', for which $(|D(M'_S)| - |S'| - 1) = (|D(M_S)| - |S| - 1)$. Inductively, we can add columns to S to obtain a set of columns $S^* = S \cup I$ such that $(|D(M_{S^*})| - |S^*| - 1) \geq |D(M_S)| - |S| - 1) > 3^{k-1} - 1$. We know that $|S^*| \geq |I| = 2 \cdot 3^{k-1}$. Hence, we obtain $|D(M_{S^*})| > 3^{k-1} + 2 \cdot 3^{k-1} = 3^k$, which is a contradiction since we only have 3^k rows. Therefore, it follows that $R_h = 3^{k-1} - 1$. Hence,

$$R_h = 3^{k-1} - 1 = \frac{2}{3}\left(\frac{3^k - 3}{2}\right) = \frac{2}{3}|\mathcal{CC}|.$$

This shows that $R_h = \frac{2}{3}R_c$ for the matrix M. For this particular example, one can also show that $R_s = |\mathcal{CC}|$. □

This completes the proof of Theorem 5. □

Although the R_h bound for the matrix M is $3^{k-1} - 1$, by obtaining local bounds using R_h on subregions of the matrix and using the framework of Myers and Griffiths [22] to combine these local bounds, the overall bound for the whole matrix can be improved to $|\mathcal{CC}|$. However, by permuting columns appropriately, overall bound obtained by combining the local R_h bounds can be forced to be $3^{k-1} - 1$. while the connected component bound is unchanged. The next theorem shows that the above example is in fact a worse-case scenario.

Theorem 6. *For any matrix M, $R_h(M) \geq \frac{2}{3}R_c(M) - \frac{1}{3}$.*

For a given matrix M, we can remove columns such that every nontrivial connected component is of size 2 and the number of nontrivial connected components does not decrease. Therefore, it suffices to prove the above theorem for a matrix M in which every connected component has size 2. Next, we prove a series of lemmas for a matrix M in which every connected component is of size 2. For such a matrix M with n distinct rows, we show that the number of nontrivial connected components cannot exceed $n/2$. For a matrix with n rows, we denote the number of nontrivial connected components by $C(n)$. Since every component is of size 2, the number of sites is $2 \cdot C(n)$.

Lemma 8. *For a matrix M in which every connected component is of size 2, there exists a connected component (pair of columns), such that three of the four pairs $\{(00, 01, 10, 11\}$ appear exactly once.*

Proof. Consider a connected component C. Let $C(ab)$ denote the set of rows with value ab in the columns of C, where $ab \in \{00, 01, 01, 11\}$. Let $C(2)$ denote the set of rows corresponding to the second largest among the four values: $\{|C(ab)| : ab \in \{00, 01, 01, 11\}\}$. and xy denote the pair in the component C in the rows $C(2)$. Let C_{min} be the connected component for which $|C_{min}(2)| = \min_{C \in \mathcal{CC}}\{|C(2)|\}$.

If $|C_{min}(2)| = 1$, then clearly $|C_{min}(2)| = |C_{min}(3)| = |C_{min}(4)| = 1$ and, therefore, three of the four pairs appear exactly once in the connected component C_{min}. If $|C_{min}(2)| > 1$, since all the rows in the matrix M are distinct, there exists a connected component C' such that the pairs in the component C' in the set of rows $C_{min}(2)$ are not all equal. Hence, $h(C_{min}, C') = xy$ (here, $h(A, B)$ denotes the common haplotype of component A with respect to B) and, therefore, three out of four pairs in the component C' are present in the rows $C_{min}(2)$. Let $C'(2)$ denote the set of rows corresponding to the second largest among the four values: $\{|C'(ab)| : ab \in \{00, 01, 01, 11\}\}$. Clearly, $|C'(2)| < |C_{min}(2)|$. However, $|C_{min}(2)| = \min_{C \in \mathcal{CC}}\{|C(2)|\}$ which leads to a contradiction. Therefore, $|C_{min}(2)| = 1$ and there is a connected component such that three of the four pairs $\{00, 01, 10, 11\}$ appear exactly once. □

Lemma 9. *For a matrix M in which every connected component is of size 2, $R_h(M) \geq n/3 - 1$.*

Proof. The proof is by induction on the number of rows. For a matrix with at most six rows and at least one connected component (pair of columns), we can restrict the matrix to a single connected component for computing R_h and, hence, $R_h \geq 4 - 2 - 1 = 1 \geq 6/3 - 1$. This proves the base case. Suppose the induction hypothesis is true for $k < n$, i.e., for every matrix M with k rows ($k < n$) and in which every connected component is of size 2, $R_h(M) \geq k/3 - 1$. Now, consider a matrix with n rows, where $n > 6$. From the previous lemma, there is a pair of conflicting columns (connected component), such that three of the four pairs appear exactly once. Let M' denote the matrix after removing the two columns with three of the four pairs occurring only once and the three rows corresponding to the three pairs. Note that removing the two columns does not cause any other rows to become identical, since all rows apart from the three removed had the same value in the two columns (see Lemma 3). One can write $R_h(M) \geq R_h(M') + 3 - 2 = R_h(M') + 1$. From the induction hypothesis, we have $R_h(M') \geq (n-3)/3 - 1$. Combining, we obtain $R_h(M) \geq n/3 - 1$, which proves the lemma. □

Lemma 10. *For a matrix M in which every connected component is of size 2, $C(n) \leq n/2 - 1$.*

Proof. Consider a matrix M with n rows in which every connected component has size 2. From Lemma 8, there exists a connected component C, such that three of the four pairs occur exactly once in C. Denote the three rows containing these pairs as $R(C)$. Moreover, applying the 2-edge theorem, we also have the property that the pairs in the rows $R(C)$, in every component apart from C are identical. Hence, if we remove the columns in C, three rows in M become identical. Therefore, we have the equation: $C(n) \leq 1 + C(n-2)$. We also have $C(4) = 1$ and, hence, $C(n) \leq n/2 - 1$. □

From the above two lemmas, it follows that $R_h \geq n/3 - 1 \geq \frac{2}{3}C(n) - \frac{1}{3}$, which completes the proof of Theorem 14. Note that the theorem still holds if $R_c(M)$ is replaced by $\max_{S' \subseteq S} R_c(M(S'))$.

4.1 Application to a Real Data Set

Next, we consider a real data set taken from the alcohol dehydrogenase locus from 11 chromosomes of Drosophila melanogaster [18]. The original data set had 11 haplotypes and 2,800 sites. We coalesce two identical haplotypes and remove all sites that are not incompatible with any other sites and sites that are identical to an adjacent site. This leaves us with the following reduced set of nine haplotypes typed at 16 sites (see Fig. 8). From Lemma 1, a lower bound on a subset of sites is also a lower bound for the complete set of sites. Hence, we consider the haplotypes to be restricted to the sites: $\{1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}$. We denote a recombination lower bound for the sites between sites a and b by B_{ab}. For this smaller data set, the recombination lower bound that we would get using R_m is 4. However, the conflict graph for the subset $\{1, 4, 5, 6\}$ has two connected components, which implies a local recombination lower bound of 2 between the sites 1 and 6, i.e., $B_{1,6} = 2$. Hence, the connected component lower bound for the smaller data set is 5. However, for the set of sites between 7 and 15, there are two conflicting pair of sites: $(7, 15)$ and $(14, 15)$. One can check that the removal of one sequence does not destroy both these conflicts. Hence, one can infer a recombination bound of 2 for this subset (see Lemma 6), i.e., $B_{7,15} = 2$. Therefore, $B_{1,16} = B_{1,6} + B_{6,7} + B_{7,15} + B_{15,16} = 2 + 1 + 2 + 1 = 6$ which gives an overall lower bound of 6. For this data set, Song and Hein [27] showed that the minimum number of recombination events is 7. This example illustrates that the connected component bound can provide improvements over the bound R_m.

5 Discussion and Future Work

The technical part of this paper has been devoted to establishing that the number of nontrivial connected

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
D	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
E	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	1
F	0	1	0	0	0	1	0	0	0	1	0	1	0	1	1	1
G	0	1	0	0	0	1	0	0	1	1	1	1	1	1	0	1
H	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1
I	1	1	1	1	0	1	0	0	1	1	1	1	1	1	0	1

Fig. 8. xxxxx.

components is a lower bound on the number of recombination events. How good is this lower bound? The problem turns out to be difficult because the recombination events can occur independently between sites, but also multiple times for different sequences at the same site. The evidence for the first case is accumulated by partitioning across columns as shown by R_m, and for the second case by looking at possible histories and increase in diversity due to recombination. The bounds R_h, and R_s use the increase in diversity to deduce past recombination events and fall in the second category. The R_c bound falls in the first category as well. It uses the genealogical information more carefully than the bound R_m. More importantly, the computation time of R_s can be improved by computing the bound separately on each connected component of the conflict graph. Finding a polynomial time algorithm for computing a lower bound with some approximation guarantees still remains an outstanding open problem. In a subsequent paper, we will present theoretical formulations of the two bounds given by Myers and Griffiths [22]. Using these formulations, we also obtain complexity results for computing these bounds, and an algorithm for computing R_s that is exponential instead of superexponential in the worst case.

ACKNOWLEDGMENTS

The authors would like to thank Dan Gusfield for many valuable comments and for introducing many of the techniques used in this paper. The authors would also like to thank the anonymous referees for their insightful comments and several pointers to related literature that have helped to improve the presentation of this work.

REFERENCES

[1] A.G. Clark, K.M. Weiss, D.A. Nickerson, S.L. Taylor, A. Buchanan, J. Stengard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C.F. Sing, "Haplotype Structure and Population Genetic Inferences from Nucleotide-Sequence Variation in Human Lipoprotein Lipase," *Am. J. Human Genetics,* vol. 63, pp. 595-612, 1998.

[2] D.C. Crawford, T. Bhangale, N. Li, G. Hellenthal, M.J. Rieder, D.A. Nickerson, and M. Stephens, "Evidence for Substantial Fine-Scale Variation in Recombination Rates across the Human Genome," *Nature Genetics,* vol. 36, pp. 700-706, 2004.

[3] M.J. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander, "High-Resolution Haplotype Structure in the Human Genome," *Nature Genetics,* vol. 29, pp. 229-232, 2001.

[4] D. Gusfield, S. Eddhu, and C. Langley, "Efficient Reconstruction of Phylogenetic Networks with Constrained Recombination," *Proc. IEEE Computer Soc. Bioinformatics Conf.,* pp. 363-374, 2003.

[5] P. Fearnhead and P. Donnelly, "Estimating Recombination Rates from Population Genetic Data," *Genetics,* vol. 159, pp. 1299-1318, 2001.

[6] P. Fearnhead, R.M. Harding, J.A. Schneider, S. Myers, and P. Donnelly, "Application of Coalescent Methods to Reveal Fine-Scale Rate Variation and Recombination Hotspots," *Genetics,* vol. 167, pp. 2067-2081, 2004.

[7] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altschuler, "The Structure of Haplotype Blocks in the Human Genome," *Science,* vol. 296, no. 5576, pp. 2225-2229, 2002.

[8] R.C. Griffiths and P. Marjoram, "Ancestral Inference from Samples of DNA Sequences with Recombination," *J. Computational Biology,* vol. 3, no. 4, pp. 479-502, 1996.

[9] D. Gusfield, "Efficient Algorithms for Inferring Evolutionary Trees," *Networks,* vol. 21, pp. 19-28, 1991.

[10] D. Gusfield, "Optimal, Efficient Reconstruction of Root-Unknown Phylogenetic Networks with Constrained and Structured Recombination," technical report, Univ. of California at Davis, 2004.

[11] D. Gusfield and D. Hickerson, "A Fundamental, Efficiently-Computed Lower Bound on the Number of Recombinations Needed in Phylogenetic Networks," technical report, Univ. of California at Davis, 2004.

[12] The Int'l HapMap Project, http://www.hapmap.org/. year?

[13] J. Hein, "A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination," *J. Molecular Evolution,* vol. 20, pp. 402-411, 1993.

[14] R.R. Hudson, "Two-Locus Sampling Distributions and Their Applications," *Genetics,* vol. 159, pp. 1805-1817, 2001.

[15] R.R. Hudson and N.L. Kaplan, "Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences," *Genetics,* vol. 111, pp. 147-164, 1985.

[16] Int'l Human Genome Sequencing Consortium, "Initial Sequencing and Analysis of the Human Genome," *Nature,* vol. 409, pp. 860-921, 2001.

[17] A.J. Jeffreys, A. Ritchie, and R. Neumann, "High Resolution Analysis of Haplotype Diversity and Meiotic Crossover in the Human Tap2 Recombination Hotspot," *Human Molecular Genetics,* vol. 9, pp. 725-733, 2000.

[18] M. Kreitman, "Nucleotide Polymorphism at the Alcohol Dehydrogenase Locus of Drosophila Melanogaster," *Genetics,* vol. 11, pp. 147-164, 1985.

[19] N. Li and M. Stephens, "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data," *Genetics,* vol. 165, pp. 2213-2233, 2003.

[20] G.A. McVean, T.P. Awadella, and P. Fearnhead, "A Coalescent Method for Detecting Recombination from Gene Sequences," *Genetics,* vol. 160, pp. 1231-1241, 2002.
[21] G.A. McVean, S.R. Myers, S. Hunt, P Deloukas, D.R. Bentley, and P. Donnelly, "The Fine-Scale Structure of Recombination Rate Variation in the Human Genome," *Science,* vol. 304, pp. 581-584, 2004.
[22] S.R. Myers and R.C. Griffiths, "Bounds on the Minimum Number of Recombination Events in a Sample History," *Genetics,* vol. 163, pp. 375-394, 2003.
[23] D.A. Nickerson, S.L. Taylor, S.M. Fullerton, K.M. Weiss, A.G. Clark, J.H. Stengaard, V. Salomaa, E. Boerwinkle, and C.F. Sing, "Sequence Diversity and Large-Scale Typing of SNPS in the Human Apolipoprotein E Gene," *Genome Research,* vol. 10, pp. 1532-1545, 2000.
[24] T.D. Petes, "Meiotic Recombination Hot Spots and Cold Spots," *Nature Rev. Genetics,* vol. 1, pp. 360-369, 2001.
[25] SNP Discovery and Genotyping, http://cardiogenomics.med.harvard.edu/component-detail?project_id=240. year?
[26] Y.S. Song and J. Hein, "Parsimonious Reconstruction of Sequence Evolution and Haplotype Blocks: Finding the Minimum Number of Recombination Events," *Proc. Conf. Algorithms in Bioinformatics, WABI 2003,* pp. 287-302, 2003.
[27] Y.S. Song and J. Hein, "On the Minimum Number of Recombination Events in the Evolutionary History of DNA Sequences," *J. Math. Biology,* vol. 48, pp. 160-186, 2004.
[28] G. Venter et al., "The Sequence of the Human Genome," *Science,* vol. 291, pp. 1304-1351, 2001.
[29] L. Wang, K. Zhang, and L. Zhang, "Perfect Phylogenetic Networks with Recombination," *J. Computational Biology,* vol. 8, no. 1, pp. 69-78, 2001.

Vineet Bafna is an assistant professor in computer science and engineering at the University of California, San Diego (UCSD). Prior to joining UCSD in 2003, he worked at Celera Genomics, ultimately, as director of informatics research. He also held positions at SmithKline Beecham and at The Center for Advancement of Genomics. His research is focused on bioinformatics, including population genetics, gene-finding, and computational proteomics.

Vikas Bansal received the BTech degree in computer science and engineering from the Indian Institute of Technology, Delhi, India, in August 2003. He is currently a PhD student in computer science at the University of California, San Diego. His research interests lie in the general areas of algorithms and computational biology.