JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 3, 2011
© Mary Ann Liebert, Inc.
Pp. 415–428
DOI: 10.1089/cmb.2010.0270

# On the Genealogy of Asexual Diploids

FUMEI LAM,[1] CHARLES H. LANGLEY,[2] and YUN S. SONG[3]

## ABSTRACT

**Given molecular genetic data from diploid individuals that, at present, reproduce mostly or exclusively asexually without recombination, an important problem in evolutionary biology is detecting evidence of past sexual reproduction (i.e., meiosis and mating) and recombination (both meiotic and mitotic). However, currently there is a lack of computational tools for carrying out such a study. In this article, we formulate a new problem of reconstructing diploid genealogies under the assumption of no sexual reproduction or recombination, with the ultimate goal being to devise genealogy-based tools for testing deviation from these assumptions. We first consider the infinite-sites model of mutation and develop linear-time algorithms to test the existence of an asexual diploid genealogy compatible with the infinite-sites model of mutation, and to construct one if it exists. In this ideal case, our chance of detecting signatures of past sexual reproduction is maximized. Then, we relax the infinite-sites assumption and develop an integer linear programming formulation to reconstruct asexual diploid genealogies with the minimum number of homoplasy (back or recurrent mutation) events. If this number is substantially larger than that expected for typical asexual organisms, then it may suggest that sexual reproduction or recombination may have played an important role in the evolutionary history. We apply our algorithms on simulated data sets with sizes of biological interest.**

**Key words:** combinatorial libraries, genetic variation, haplotypes, protein structure, sequence analysis.

## 1. INTRODUCTION

Reproduction in asexual organisms usually is less costly than that in sexual organisms. Yet, sexual reproduction and genetic recombination are common to the majority of higher organisms in nature, and several different explanations have been put forward to address this intriguing phenomenon (Barton and Charlesworth, 1998; Keightley and Otto, 2006). Although it still remains debatable as to which precise evolutionary conditions and mechanisms maintain sex and recombination in natural populations, it is widely believed that sex and recombination are important for the long-term evolutionary success of an organism; that is, asexual organisms are believed to be much more susceptible to extinction than are their

[1] Department of Computer Science and [2] Section of Evolution and Ecology, University of California, Davis, California.
[3] Departments of EECS and Statistics, University of California, Berkeley, California.

sexual counterparts that undergo meiosis and mating (Maynard Smith, 1978). Contrary to this common belief, the phylum Rotifera, microscopic aquatic animals widespread throughout the world, contains a class—namely, Bdelloidea—that seems to have been reproducing asexually for tens of millions of years, diversifying into 360 known species that constitute four families and 18 genera. Fossil evidence suggests that bdelloid rotifers have been around for at least 35–40 million years (Waggoner and Poinar, 1993), while molecular genetic analysis suggests an age that is more than twice as large (Mark Welch and Meselson, 2000). Maynard Smith (1986) referred to the bdelloid rotifers as ''something of an evolutionary scandal,'' and it has been questioned in the past whether they indeed have remained asexual for all that while (Judson and Normark, 1996).

Recently, Mark Welch and Meselson (2000) analyzed molecular genetic data of four bdelloid species and provided evidence to support bdelloid rotifers' ancient, continuous asexuality. Their method was based on counting synonymous sequence differences between different copies of a gene within an individual, which, under neutrality, are expected to be over-represented in an old asexual organism (Birky, 2004). Mark Welch and Meselson showed that allelic sequence differences at synonymous sites are significantly greater in bdelloid rotifers than in their closest relative class monogonont rotifers, consisting of about 1500 species, which seem to reproduce mostly asexually, but with an occasional sexual reproduction. (More recent evidence in support of the ancient asexuality of bdelloid rotifers is provided in Fontaneto et al., 2007.) In contrast to this success, when a similar analysis was applied to other asexual organisms such as darwinulid ostrocods (Schön and Martens, 2003), of which morphological evidence strongly supports ancient asexuality (Martens et al., 2003), no significantly high level of sequence divergence was observed. In another study, a similar sequence divergence test applied to plant-parasitic worms (specifically, root-knot nematodes from the genus *Meloidogyne*) supported their ancient asexuality, while further analysis revealed that interspecific hybridization was involved in the history of this group (Lunt, 2008). From this study, the author concluded ''genetic signatures of ancient asexuality must be taken with caution due to the confounding effect of interspecific hybridization, which has long been implicated in the origins of apomictic species.'' As these cases illustrate, a more refined method that makes better use of DNA data is needed for studying asexuality.

In this article, we develop new methods to test asexuality by explicitly considering the evolutionary history of diploid individuals. We first consider the infinite-sites model of mutation, which corresponds to the ideal case in which mutations provide as much information about genealogy as possible. This ideal case should provide an upper bound on our chance of detecting signatures of past sexual reproduction. Given $n$ pairs of phased haplotypes or $n$ unphased genotypes, our goal is to test the existence of an $n$-leaved *diploid perfect phylogeny* (DPP)—an asexual diploid genealogy compatible with the infinite-sites model of mutation and no recombination—for the input individuals, and to construct one if it exists. We devise linear-time algorithms for both phased haplotypic and unphased genotypic input data, and show that a minimal DPP for a given data set is unique if it exists. If a DPP solution exists for unphased genotypic input data, our algorithm finds a phasing of the input genotypes into pairs of haplotypes compatible with the DPP, and the DPP serves as a data structure that encodes all such phasing solutions.

If the input sequences do not admit a DPP solution, then further analysis is required to understand the evolutionary forces such as homoplasy or recombination that are present in the genealogy. In the second part of this article, we relax the infinite-sites assumption and study the *diploid imperfect phylogeny* (DIP) problem, which is to reconstruct asexual diploid genealogies with the minimum number of homoplasy (recurrent or back mutation) events. If the minimum number of homoplasy events is significantly greater than that expected for typical asexual organisms, then it may indicate that other evolutionary forces such as recombination, hybridization, or sexual reproduction may have played a role in the genealogy of the organisms. We develop an integer linear programming formulation to tackle this problem and study the practicality of our approach by applying our algorithms on simulated data sets with sizes of current biological interest.

Our ultimate goal is to devise genealogy-based tools for testing deviation from asexual evolution. Given molecular genetic data from diploid individuals that, at present, reproduce mostly or exclusively asexually, an important open problem is to estimate the frequency of past sexual reproduction, as well as the amount of recombination (meiotic and mitotic crossovers and gene-conversions). Further, it will be important to estimate when sexuality was lost and how many independent times. The work described in this article is a modest step toward that general direction. The preliminary results described here suggest that genealogical approaches may provide new insights into the study of asexual evolution.

We carry out simulations to study how often DPPs exist for sample data sets when some of the underlying assumptions for DPP are violated. In particular, we introduce various amounts of sexual reproduction and mitotic gene-conversion in the evolutionary history of the population, and test the existence of DPP solutions for random samples. Our study suggests that the signature of historical sexual reproduction decays rather fast with time and that occasional sexual reproduction may be difficult to detect.

This article is organized as follows. We first define the DPP problem and formulate the computational problems to be addressed in this article. We consider the case with haplotypic input data and highlight several important properties satisfied by DPPs, and describe results for phasing genotypic input data as a DPP. We then extend the diploid phylogeny method to incorporate back and recurrent mutation events in the DIP model and formulate integer linear programming methods to solve the resulting problem. Simulation results are discussed, and we conclude with some remarks.

CloneTree, software that implements our algorithms, is available at `http://www.eecs.berkeley.edu/~yss/software.html`. It produces a graphical output that displays the diploid genealogy found by our algorithms.

## 2. DIPLOID PERFECT PHYLOGENY

We assume that the input data consist of either phased or unphased single nucleotide polymorphisms (SNPs) from $n$ diploid individuals with $m$ polymorphic sites. Each site has at most two phased alleles, denoted by $\{0, 1\}$. The data we consider are of the following two types:

**Definition 2.1** (Haplotype data). *A haplotype is a binary string of length m. Let $h_i$ and $\widetilde{h}_i$ denote the pair of haplotypes of individual i; $h_i$ and $\widetilde{h}_i$ are called* mates. *A collection of such pairs of haplotypes for n individuals is denoted by a $2n$-by-m binary matrix H, in which rows $2i$–1 and $2i$ correspond to the two haplotypes of individual i.*

**Definition 2.2** (Genotype data). *Let $g_i$ denote the genotype of individual i. The value of $g_i$ at site k is* 0 *if individual i has two copies of* 0 *at site k;* 1 *if individual i has two copies of* 1 *at site k; or* 2 *otherwise. A collection of genotypes for n individuals is denoted by G, with row i corresponding to $g_i$. A $2n$-by-m binary matrix H is said to be a phasing solution to an n-by-m ternary matrix G if, for all $i = 1, \ldots, n$, $g_i$ in G is the genotype consistent with the mates $h_i$ and $\widetilde{h}_i$ in H.*

In the infinite-sites model of mutation, at most one mutation may occur per site in the entire evolutionary history. Perfect phylogenies are trees representing evolutionary histories consistent with the infinite-sites model of mutation (for each site, there is at most one edge in the tree corresponding to mutation at that site) (Semple and Steel, 2003). We will refer to these as *haploid perfect phylogenies (HPP)* to distinguish them from *diploid perfect phylogenies* (*DPP*), a new concept defined as follows.

**Definition 2.3** (Diploid Perfect Phylogeny). *A diploid perfect phylogeny (DPP) for n diploid individuals is an n-leaved rooted tree T representing the evolutionary history of self-cloning (or asexually reproducing) individuals satisfying:*

1. *Mutations occur on edges and each site may mutate at most once in T. Time flows from the root (which has degree 2) to the leaves (which have degree 1), and each edge in T represents a diploid lineage. If site k mutates on an edge, only one of the two haplotypes gets modified at that site, and the newly arising allele (0 or 1) has never been seen before at that site.*
2. *Depending on whether the input data are pairs of haplotypes or genotypes, every vertex of a DPP is labeled by a pair of haplotypes or a genotype, respectively.*
3. *There is a one-to-one correspondence between the n leaves of T and the n input individuals.*

*A minimal DPP is a DPP in which the two ends of every interior edge have different labels.*

Note that a set of $2n$ haplotypes for $n$ individuals may admit an HPP solution while admitting no DPP solution. A DPP example is shown in the middle of Figure 1. In this article, we address the following two algorithmic questions:

**DPP for Haplotype Data:** Given a haplotypic data set $H$ for $n$ diploid individuals, determine whether $H$ admits a DPP solution, and find one if it exists.

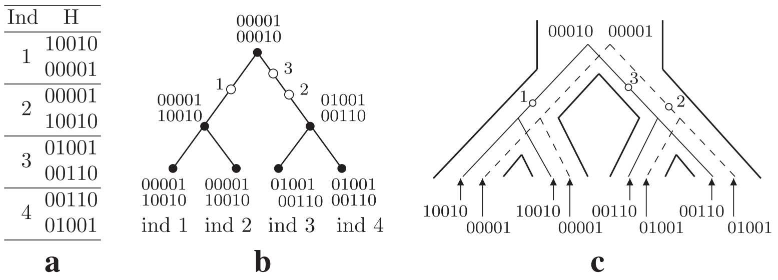

**FIG. 1.** **(a)** A haplotype data set $H$ for four individuals. **(b)** Its unique minimal diploid perfect phylogeny (DPP) $T$. **(c)** Evolutionary histories of the haplotypes embedded in $T$. We use $\tau_X$ and $\tau_Y$ to denote the solid and the dotted trees in $T$, respectively. An open circle labeled $k$ represents a mutation at site $k$.

**DPP Haplotyping for Genotype Data:** Given a genotypic data set $G$ for $n$ diploid individuals, determine whether $G$ can be phased to a haplotypic data set $H$ that admits a DPP solution, and if so, find such a phasing solution $H$.

## 3. DPP FOR HAPLOTYPE DATA

Gusfield (1991) devised a linear-time algorithm to test whether a given haplotypic input data set admits an HPP solution and to find one if it exists. In this section, we construct an analogous linear-time algorithm for DPP, making use of Gusfield's linear-time algorithm for HPP. First, we highlight several important properties satisfied by DPPs.

### *3.1. Properties of diploid perfect phylogenies*

Suppose there is an $n$-leaved minimal DPP $T$ for $H$. Let $x_r$ and $y_r$ denote the root haplotypes of $T$. Following the history of $x_r$ on $T$ leads to one haplotype per leaf in $T$. Denote this set of haplotypes $H_X$ and their history $\tau_X$. Similarly, follow the history of $y_r$ on $T$ to obtain $H_Y$ and $\tau_Y$ . Note that each diploid individual has exactly one of its two haplotypes in $H_X$ and the other in $H_Y$. The following properties are implied by the one-mutation-per-site condition:

**P1:** The set of mutations in $\tau_X$ and $\tau_Y$ are disjoint. (In Figure 1, sites 1 and 3 mutate in $\tau_X$ but not in $\tau_Y$. Similarly, site 2 mutates in $\tau_Y$ but not in $\tau_X$.)

**P2:** If $x_r[k] \neq y_r[k]$, then both 0 and 1 have already been seen, so part 1 of Definition 2.3 implies that neither $\tau_X$ nor $\tau_Y$ contains a mutation at site $k$. As a consequence, no individual in $T$ is homozygous at site $k$. (In Figure 1, sites 4 and 5 satisfy this property.)

For a given input data set $H$, the one-mutation-per-site condition imposes tight constraints on the possible root haplotypes of a DPP. In what follows, we use $\mathcal{E}(H)$ to denote the set of all sites in $H$ at which every individual is heterozygous.

**Lemma 3.1** (Constraints on the root). *The haplotypes $x_r$,$y_r$ of any possible root individual of a DPP satisfy the following properties:*

1. *For all $k$, $k \notin \mathcal{E}(H)$, there cannot be two distinct homozygous genotypes at site $k$. If any individual $i$ in $H$ has a homozygous genotype $h_i[k] = \widetilde{h}_i[k] = c$, then the root individual also has the same genotype $x_r[k] = y_r[k] = c$.*
2. *$H$ restricted to the sites in $\mathcal{E}(H)$ has exactly two distinct haplotypes, and those haplotypes are equal to the root haplotypes $x_r$,$y_r$ restricted to $\mathcal{E}(H)$. More precisely, for any particular site $j \in \mathcal{E}(H)$, let $H_X$ (respectively, $H_Y$ ) denote the set of $n$ haplotypes with a* 1 *(respectively,* 0*) at site $j$. Then, for all $k \in \mathcal{E}(H)$, both $H_X$ and $H_Y$ are non-polymorphic at site $k$, with $H_X$ and $H_Y$ having different alleles. Further, $x_r$ (respectively, $y_r$) restricted to the sites in $\mathcal{E}(H)$ is the same as any haplotype in $H_X$ (respectively, $H_Y$ ) restricted to $\mathcal{E}(H)$. So, for all $k \in \mathcal{E}(H)$, the root is heterozygous at site $k$.*

**Proof.** If there exists a DPP, Property P1 implies that no two distinct homozygous genotypes may exist at any site. Further, Property P2 implies that if $H$ contains an individual homozygous at site $k$, then the root individual of any DPP solution for $H$ must be homozygous at that site. The first part of this lemma then follows from these two facts.

Let $T$ denote an $n$-leaved DPP for the $n$ individuals in $H$, and suppose that the root $\rho$ of $T$ is homozygous at some site $j \in \mathcal{E}(H)$. Then, since every individual in $H$ is heterozygous at that site, $\rho$ is not in $H$. Now, the one-mutation-per-site condition implies that there is an edge that separates $\rho$ from all individuals in $H$, thus implying that $T$ contains a leaf not labeled by any individual in $H$, which in turn implies that $T$ is not an $n$-leaved DPP for $H$, a contradiction. Hence, the root individual of a DPP must be heterozygous at every site $j \in \mathcal{E}(H)$. This fact and the one-mutation-per-site condition together imply that there is no mutation event at any site $j \in \mathcal{E}(H)$ in $T$, and the second part of the lemma immediately follows. ■

Lemma 3.1 implies that if a DPP exists for $H$, there is a unique choice for the root. Using this lemma, we can show several useful results that hold if a DPP exists. First, we need two definitions.

**Definition 3.2** (Resolution of a vertex). *In a graph $\mathcal{G}$, resolution of a degree-$d$ vertex $v$ incident to edges $e_1, \ldots, e_d$ (with $d > 3$), is an operation that splits $v$ into two new vertices $v_1$ and $v_2$, such that (i) $v_1$ and $v_2$ are joined by a new edge, (ii) each of $e_1, \ldots, e_d$ is incident with either $v_1$ or $v_2$, (iii) both $v_1$ and $v_2$ have degree $\geq 3$, and (iv) the remaining vertices and edges of $\mathcal{G}$ remain the same.*

**Definition 3.3** ($\bowtie$, Join operation). *For two $k$-by-$l$ matrices $M_1$ and $M_2$, the $k$-by-$2l$ matrix $M_1 \bowtie M_2$ is obtained by appending row $i$ of $M_2$ to row $i$ of $M_1$.*

The following result provides a way to find the partition of $H$ into $H_X$ and $H_Y$ if a DPP solution exists.

**Proposition 3.4** (Partition of $H$ into $H_X$ and $H_Y$). *For $n > 1$, suppose there exists an $n$-leaved DPP $T$ for the $n$ individuals in $H$. Then, the $2n$ haplotypes in $H$ admit a unique $2n$-leaved minimal unrooted haploid perfect phylogeny $\tau$ satisfying the following:*

1. *If $\mathcal{E}(H) \neq \emptyset$, then there exists a unique edge in $\tau$ such that cutting that edge partitions $\tau$ into two $n$-leaved subtrees such that, for each individual $i$, haplotype $h_i$ appears as a leaf of one subtree while its mate haplotype $\widetilde{h}_i$ appears as a leaf of the other subtree (Fig. 2a).*
2. *If $\mathcal{E}(H) = \emptyset$, then there exists a unique vertex $v$ in $\tau$ with degree $d$, where $d > 3$, such that resolving $v$ and cutting the edge between the newly created vertices partitions $\tau$ into two $n$-leaved subtrees such that, for each individual $i$, haplotype $h_i$ appears as a leaf of one subtree while its mate haplotype $\widetilde{h}_i$ appears as a leaf of the other subtree (Fig. 2b).*

**Proof.** Define $\tau_X$, $\tau_Y$, $x_r$, and $y_r$ as in the beginning of this section. If $x_r$ and $y_r$ are not identical, then add a new edge between the root of $\tau_X$ and the root of $\tau_Y$, and add mutation events on that edge for all sites $k$ where $x_r[k] \neq y_r[k]$. If $x_r$ and $y_r$ are identical, combine $\tau_X$ and $\tau_Y$ by identifying the root vertex $\rho_X$ of $\tau_X$ with the root vertex $\rho_Y$ of $\tau_Y$, such that the new vertex $\rho_{X \oplus Y}$ obtained from identifying $\rho_X$ and $\rho_Y$ is incident with all the edges that were incident with $\rho_X$ or $\rho_Y$. Then, Properties P1 and P2 imply that the resulting tree $\tau_{X \oplus Y}$ is an unrooted HPP for the haplotypes in $H$. Now, contract interior edges in $\tau_{X \oplus Y}$ with no mutations and call the resulting tree $\tau$. Note that $\tau$ is a unique $2n$-leaved minimal unrooted HPP for the haplotypes in $H$.

If $\mathcal{E}(H) \neq \emptyset$, then $x_r$ and $y_r$ are not identical by definition of $\mathcal{E}(H)$, implying $\tau$ described above contains an edge between $x_r$ and $y_r$ with at least one mutation. Therefore, $\tau$ satisfies part 1 of the proposition. If $\mathcal{E}(H) = \emptyset$, then Lemma 3.1 implies that the root haplotypes $x_r$ and $y_r$ are identical, so $\tau$ has the vertex $\rho_{X \oplus Y}$ described above. Note that $\rho_{X \oplus Y}$ has degree $> 3$, and part 2 of the proposition is satisfied by construction. ■

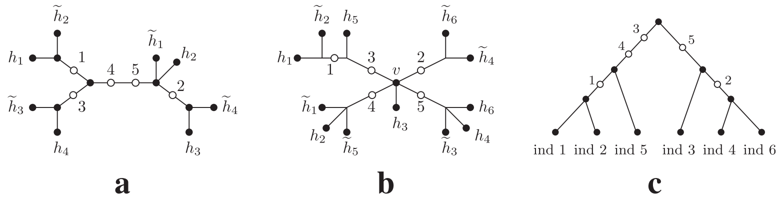

**FIG. 2.** Haploid perfect phylogeny (HPP) and diploid perfect phylogeny (DPP) examples for Propositions 3.4 and 3.5. **(a)** There is a unique edge (namely, the edge on which sites 4 and 5 mutate) satisfying the property described in part 1 of Proposition 3.4. **(b)** There is a unique vertex (labeled $v$ in the figure) satisfying the property described in part 2 of Proposition 3.4. **(c)** The unique minimal DPP consistent with the HPP shown in (b).

If a DPP exists for $H$ and $\mathcal{E}(H)=\emptyset$, part 2 of Proposition 3.4 indicates that there is a unique degree-$d$ vertex $v$, where $d>3$, such that a partition of $H$ into $H_X$ and $H_Y$ can be obtained from resolving that vertex. However, there can be more than one admissible resolution of that vertex (and hence more than one possible partition of $H$) that is consistent with the existence of a DPP. Below, we show that all such resolutions imply the same minimal DPP. For illustration, consider the HPP shown in Figure 2b. It admits two possible partitions of $H$ into $H_X$ and $H_Y$—either $H_X=\{h_1, \widetilde{h}_2, h_3, \widetilde{h}_4, h_5, \widetilde{h}_6\}$ and $H_Y=\{\widetilde{h}_1, h_2, \widetilde{h}_3, h_4, \widetilde{h}_5, h_6\}$, or $H_X=\{h_1, \widetilde{h}_2, \widetilde{h}_3, h_4, h_5, h_6\}$ and $H_Y=\{\widetilde{h}_1, h_2, h_3, \widetilde{h}_4, \widetilde{h}_5, \widetilde{h}_6\}$. It is easy to see that both cases lead to the same DPP (Fig. 2c). Now, the following result establishes the uniqueness of a minimal DPP solution:

**Proposition 3.5** (Uniqueness). *If a DPP exists for H, then H admits a unique minimal DPP.*

**Proof.** Suppose that $H$ admits a DPP. If $\mathcal{E}(H)\neq\emptyset$, then part 1 of Proposition 3.4 implies that there is a unique way to partition $H$ into $H_X$ and $H_Y$. The root haplotypes $x_r$ and $y_r$ are as described in Lemma 3.1. A minimal DPP for $H$ must be a minimal HPP for $H_X \bowtie H_Y$ with $x_r \bowtie y_r$ as the root sequence, and its uniqueness follows from the uniqueness of a rooted minimal HPP for a binary matrix with a given root.

Suppose that $\mathcal{E}(H)=\emptyset$, and let $\tau$ and $v$ be as in Proposition 3.4. Lemma 3.1 implies that the root haplotypes $x_r$ and $y_r$ of a DPP are identical. For ease of exposition, suppose that those haplotypes are all-zero. Then, the haplotypes assigned to $v$ are all-zero. To each mutation in $\tau$, one can associate a binary character for the $n$ individuals $\{1, 2, \ldots, n\}$ as follows. For each mutation occurring on some edge in $\tau$, imagine cutting that edge, and consider the subtree not containing $v$ that would be cut. Assign a 1 to every individual $i$ with either $h_i$ or $\widetilde{h}_i$ as a leaf in that subtree, and assign 0s to all other individuals. (Since there exists a DPP, no individual has both of its haplotypes in that subtree.) Now, the tree shape and the assignment of mutations to the edges of a minimal DPP for $H$ must be the same as that of a minimal HPP for the set of binary characters just described with the all-zero sequence as the root, and the uniqueness of that DPP is immediate. ■

### *3.2. A linear-time algorithm for haplotype data*

Using the above results and Gusfield's linear-time algorithm for HPP (Gusfield, 1991), we can devise the following $O(mn)$-time algorithm to find a DPP solution, if it exists:

1. Check that the conditions in Lemma 3.1 are satisfied. If not, there is no DPP solution. Otherwise, the root haplotypes $x_r$ and $y_r$ are uniquely determined.
2. Check whether there exists a $2n$-leaved minimal unrooted HPP $\tau$ for $H$. If not, there is no DPP solution. Otherwise, check whether a partition of the $2n$-by-$m$ input matrix $H$ into two $n$-by-$m$ matrices $H_X$ and $H_Y$ can be found as described in Proposition 3.4.
   (a) If $\mathcal{E}(H)\neq\emptyset$, the two ends of the edge in $\tau$ needed to be cut should be labeled by $x_r$ and $y_r$.
   (b) If $\mathcal{E}(H)=\emptyset$, then $x_r=y_r$. The vertex $v$ described in the second part of Proposition 3.4 is the one labeled by $x_r=y_r$. Determine whether there exists a resolution of $v$ such that cutting the newly created edge partitions $H$ into $H_X$ and $H_Y$. If not, there is no DPP solution.
3. Test whether there exists an $n$-leaved HPP for the $n$-by-$2m$ matrix $H_X \bowtie H_Y$ with $x_r \bowtie y_r$ as the root sequence. If so, then it corresponds to the unique minimal DPP for $H$. Otherwise, there is no DPP solution.

## 4. DPP HAPLOTYPING FOR GENOTYPE DATA

Gusfield (2002) considered phasing (or haplotyping) genotypic input data as an HPP and provided a nearly-linear-time algorithm for the problem. Simpler but slower solutions (Bafna et al., 2003; Eskin et al., 2003) were subsequently proposed for the problem, and linear-time algorithms were recently found (Ding et al., 2005; Vijayasatya and Mukherjee, 2005). The absence of recombination and homoplasy imposes stringent constraints on the genealogy of asexual diploid individuals. In this section, we exploit such constraints to devise a simple linear-time algorithm for the DPP Haplotyping Problem under the assumption of asexual reproduction. Our approach has two stages. First, for a given input genotype data set $G$, we find a DPP if it exists. Then, we use that DPP to find a phasing solution for $G$.

### 4.1. A linear-time algorithm for constructing a DPP for genotype data

When using 0, 1, and 2 to denote genotypic states, recall that 0 and 1 denote homozygotes, while 2 denotes a heterozygote. Lemma 3.1 implies that genotypic states 0 and 1 cannot both appear in any column in $G$. Further, the one-mutation-per-site condition implies that, when a mutation occurs at a site, it is either of type $0 \to 2$ or $1 \to 2$, but never $2 \to 0$ or $2 \to 1$. Using these facts, we devise the following linear-time algorithm for constructing a DPP for $G$, if it exists:

1. For every column $k = 1, \ldots, m$ in $G$ do the following:
   (a) Check whether both 0 and 1 appear in column $k$. If so, there is no DPP solution, so terminate the algorithm.
   (b) If column $k$ contains neither a 0 nor a 1, then set $z_r[k] = 2$.
   (c) Else, if column $k$ contains a 0 (respectively, 1), then set $z_r[k] = 0$ (respectively, $z_r[k] = 1$).
2. If the above step has succeeded, then there are at most two distinct genotypic states in each column of $G$. Viewing each column as a two-state character and each row as a haplotype, test whether $G$ admits an $n$-leaved HPP with $z_r$ as the root sequence, with mutations of type $0 \to 2$ or $1 \to 2$, depending on the root character state. If such an HPP exists, it corresponds to the unique minimal DPP for $G$.

With appropriate renaming of character states, the above algorithm can be carried out in $O(mn)$ time using Gusfield's (1991) linear-time HPP algorithm for binary matrices.

**Remark 4.1** (Uniqueness of the root genotype). *Note that if a DPP exists for an input genotype data, its root genotype $z_r$ is uniquely determined as described in the above algorithm.*

### 4.2. Solutions to the DPP haplotyping problem

If $G$ admits a unique minimal DPP $T$ with root $z_r$, a DPP haplotyping solution can be obtained using the following linear-time algorithm:

1. Let $x_r$, $y_r$ denote an arbitrary pair of haplotypes such that their corresponding genotype is equal to $z_r$.
2. Propagate $x_r$, $y_r$ down the tree $T$ as follows. Suppose that a mutation event, of type either $0 \to 2$ or $1 \to 2$, occurs at site $k$ on a directed edge $e$ in $T$, and let $h_d$, $\widetilde{h}_d$ denote the descendant haplotype mate pair labeling the vertex incident with the head of $e$. Then, either set $h_d[k] = 0, \widetilde{h}_d[k] = 1$ or set $h_d[k] = 1, \widetilde{h}_d[k] = 0$.

This procedure leads to pairs of haplotypes at the leaves consistent with $G$. As the above algorithm suggests, if there exists a DPP solution for $G$, then its unique minimal DPP $T$ serves as a data structure that encodes all DPP haplotyping solutions for $G$. There are two sources that lead to multiple DPP haplotyping solutions. First, if the root genotype $z_r$ contains $C$ heterozygous sites, then there are $2^{C-1}$ choices for $x_{\,r}$,$y_r$. Second, for a given pair $x_r$,$y_r$, the freedom described in step 2 of the above algorithm can lead to multiple haplotyping solutions.

The counting goes as follows. Suppose that there are $m_e$ mutations on an edge $e$, and let $h_a$, $\widetilde{h}_a$ denote the ancestor haplotype mate pair labeling the tail vertex of $e$. Then, the number $\mu(e)$ of choices for descendant haplotype mate pair $h_d$, $\widetilde{h}_d$, are given by

$$\mu(e) = \begin{cases} 2^{m_e - 1}, & \text{if } h_a \text{ and } \widetilde{h}_a \text{ are identical }, \text{and } m_e \geq 1, \\ 2^{m_e}, & \text{otherwise.} \end{cases} \tag{1}$$

Hence, if $G$ admits a minimal DPP $T$, the total number $\mu(G)$ of DPP haplotyping solutions is

$$\mu(G) = 2^{C-1} \prod_{e \in E(T)} \mu(e), \tag{2}$$

where $E(T)$ denotes the set of edges in $T$. All $\mu(G)$ haplotyping solutions can be obtained from $T$ by switching $h_d[k]$ with $\widetilde{h}_d[k]$ in step 2 of the above algorithm. An example is shown below.

Consider the genotype data set $G$ shown at the top of Figure 3. Using our DPP algorithm for genotype data, we can show that there exists a unique minimal DPP $T$ for $G$; that unique minimal DPP is shown in Figure 3. Then, it follows from (1) and (2) that there are exactly eight DPP haplotyping solutions for $G$. Those eight solutions, shown in Table 1, can easily be obtained from $T$ as described above.

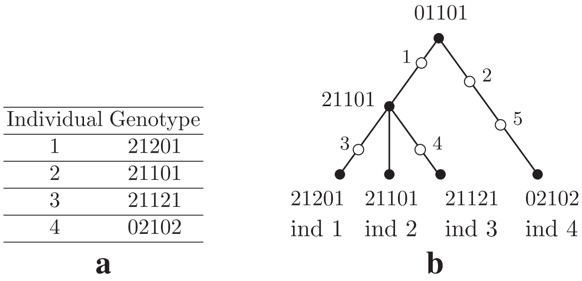

**FIG. 3.** **(a)** A genotype data set $G$ for four individuals. **(b)** Its minimal diploid perfect phylogeny (DPP) $T$. An open circle labeled $k$ represents a mutation event at site $k$. There are eight DPP haplotyping solutions, shown in Table 1.

| Individual | Genotype |
|---|---|
| 1 | 21201 |
| 2 | 21101 |
| 3 | 21121 |
| 4 | 02102 |

## 5. DIPLOID IMPERFECT PHYLOGENY

If a set of diploid sequences does not allow a DPP, then other forces must be present in the evolutionary history. These may include homoplasy or recombination events, and further analysis is necessary to distinguish between these possibilities. In this section, we relax the infinite-sites assumption and develop an integer linear programming formulation to reconstruct asexual diploid genealogies with the minimum number of homoplasy events.

**Definition 5.1** (Diploid Imperfect Phylogeny). *A diploid imperfect phylogeny (DIP) for n diploid individuals is an n-leaved rooted tree T satisfying conditions (2) and (3) in the definition of DPP, and satisfying condition (1) with the modification that multiple mutations are possible at each site.*

In order to measure the strength of evidence to distinguish between homoplasy and recombination events, we define a measure of deviation from a DPP. For a DIP $T$ displaying a set of sequences $S$, let $M_T(k)$ denote the number of edges in $T$ corresponding to mutation at site $k$.

**Definition 5.2.** *A diploid imperfect phylogeny T is q-imperfect (or q-near-perfect) if* $\sum_{k:M_T(k)\geq 1}(M_T(k)-1)=q$.

The DIP problem is to find a DIP $T$ displaying the input sequences which minimizes the imperfection $q$. In particular, if the sequences can be displayed in a DPP $T$, then $M_T(k)\leq 1$ for each site $k$, and $T$ satisfies $q=0$.

In the case of haploid input sequences, the problem of constructing imperfect haploid phylogenies has received much attention from both theoretical and practical points of view. Fernandez-Baca and Lagergren (2003), Halperin and Eskin (2004), and Sridhar et al. (2006) analyzed theoretical bounds for algorithms to solve this problem to optimality, while Sridhar et al. (2006) showed that the problem is fixed-parameter tractable in the imperfection of the resulting phylogeny. Further, it has been shown that linear programming approaches can efficiently handle data sets of biological interest (Sridhar et al., 2007). We now consider the case of constructing DIPs and introduce a problem which casts this problem in the framework of combinatorial optimization.

TABLE 1. A COMPLETE LIST OF DIPLOID PERFECT PHYLOGENY HAPLOTYPING SOLUTIONS FOR THE GENOTYPE DATA SET SHOWN IN FIGURE 3

| *Individual* | *Phase 1* | *Phase 2* | *Phase 3* | *Phase 4* | *Phase 5* | *Phase 6* | *Phase 7* | *Phase 8* |
|---|---|---|---|---|---|---|---|---|
| 1 | 11101 | 11001 | 11101 | 11001 | 11101 | 11001 | 11101 | 11001 |
| | 01001 | 01101 | 01001 | 01101 | 01001 | 01101 | 01001 | 01101 |
| 2 | 11101 | 11101 | 11101 | 11101 | 11101 | 11101 | 11101 | 11101 |
| | 01101 | 01101 | 01101 | 01101 | 01101 | 01101 | 01101 | 01101 |
| 3 | 11111 | 11111 | 11101 | 11111 | 11111 | 11111 | 11101 | 11101 |
| | 01101 | 01101 | 01111 | 01111 | 01101 | 01101 | 01111 | 01111 |
| 4 | 01101 | 01101 | 01101 | 01101 | 01100 | 01100 | 01100 | 01100 |
| | 00100 | 00100 | 00100 | 00100 | 00101 | 00101 | 00101 | 00101 |

There are eight solutions in total.

### *5.1. Group Steiner tree problem*

The problem of reconstructing phylogenies is closely related to the *Steiner Tree Problem*, a well-studied problem in combinatorial optimization. Given a graph $G=(V, E)$ with edge costs and a set of terminals $R \subseteq V$, a *Steiner tree* in $G$ is a subgraph of $G$ containing a path between any pair of terminals. The cost of a Steiner tree $T$ is the sum of the edge costs in $T$, and the Steiner Tree Problem is to find the minimum cost Steiner tree in $G$.

Let $H$ be a set of input sequences of length $m$, and let graph $G$ be the $m$-cube defined on vertices $V=\{0, 1\}^m$ and edges $E=\{(u, v) \in V \times V : \sum_i |u_i - v_i| = 1\}$. Let $R \subset V$ be the set of binary sequences corresponding to the rows of input $H$. The minimum (haploid) imperfect phylogeny problem is then equivalent to the minimum Steiner tree problem on underlying graph $G$ with terminal vertices $R$. Even in this restricted setting, the Steiner tree problem is NP-complete (Foulds and Graham, 1982).

To solve the DIP problem, we introduce the following more general Steiner tree problem. Let $G=(V, E)$ be an undirected graph, let $d$ be a non-negative cost function on edge set $E$, and let $R = R_1 \cup R_2 \ldots R_k \subset V$ be a partition of the terminal vertices into disjoint groups. A *group Steiner tree* of $G$ is a Steiner tree containing at least one vertex from each group $R_i$ and the Group Minimum Steiner Tree (GMST) Problem is to find the group Steiner tree of minimum cost.

The DIP problem can be transformed to an instance of the Group Steiner Tree problem as follows. Let $H=\{h_i, \widetilde{h}_i\}_{i=1}^n$ be the input set of paired haplotype sequences to the DIP problem, where $h_i$ and $\widetilde{h}_i$ are binary sequences of length $m$. Let graph $G$ be the $2m$-cube (where vertices are binary sequences of length $2m$ and edges are pairs of binary sequences with Hamming distance equal to one), and for each $i$, let terminal group $R_i$ be the pair of vertices $\{h_i\widetilde{h}_i, \widetilde{h}_ih_i\} \subseteq V(G)$. The GMST on this instance is then equivalent to the minimum DIP problem on $H$.

Because of its computational complexity, an important component of any computational approach for solving the Steiner Tree problem is to eliminate vertices that cannot be present in *any* optimal tree. Such a preprocessing step can substantially reduce the size of the input data without affecting the final output. In the haploid imperfect phylogeny problem, it has been shown that the *Buneman graph* of the input sequences contains all optimal trees (Bandelt et al., 1989; Buneman, 1971). Restricting the underlying graph of the problem in such a way has been shown to be efficient and practical on real data sets (Sridhar et al., 2007). The following proposition shows an analogous results holds for the diploid phylogeny problem:

**Proposition 5.3.** *Let $H$ be a set of $n$ pairs of haplotype sequences $\{h_i, \widetilde{h}_i\}$ and let $\mathcal{B}(H)$ denote the Buneman graph on $\cup_i\{h_i\widetilde{h}_i, \widetilde{h}_ih_i\}$. Then every minimum imperfect diploid phylogeny $T^*(H)$ is a subgraph of $\mathcal{B}(H)$.*

We prove this proposition using the following theorem of Bandelt et al. (1989) for haploid imperfect phylogeny construction:

**Theorem 5.4** *(Bandelt et al., 1989; Semple and Steel, 2003). For binary haplotype input sequences $H$, let $\mathcal{B}(H)$ denote the Buneman graph on $H$. Then every minimum imperfect phylogeny $T^*(H)$ is a subgraph of $\mathcal{B}(H)$.*

**Proof of Proposition 5.3.** Let $H=\{h_i, \widetilde{h}_i\}_{i=1}^n$ be a set of $n$ pairs of haplotype sequences of length $m$. Suppose $T^*(H)$ is a minimum GMST on the hypercube of dimension $2m$ with terminal groups $R_i=\{h_i, \widetilde{h}_i\}(1 \leq i \leq n)$. By definition, $T^*(H)$ must contain at least one terminal $t_i$ from each terminal group $R_i=\{h_i\widetilde{h}_i, \widetilde{h}_ih_i\}$. It follows that $T^*(H)$ is a minimum Steiner tree on terminal set $\{t_i\}_{i=1}^n$. By Theorem 5.4, $T^*(H)$ is a subgraph of the Buneman graph $\mathcal{B}(\{t_i\}_{i=1}^n)$. Since $t_i \in \{h_i\widetilde{h}_i, \widetilde{h}_ih_i\}$, it follows that $T^*(H)$ is a subgraph of the Buneman graph $\mathcal{B}(\{h_i\widetilde{h}_i, \widetilde{h}_ih_i\}_{i=1}^n) = \mathcal{B}(H)$. ■

Using this proposition, our problem can be reduced to constructing the Buneman graph on input $H$ and solving the problem on underlying graph $\mathcal{B}(H)$.

### *5.2. Integer linear programming formulation*

One approach for solving Steiner tree problems is to use integer linear programming (ILP) methods. We use the multicommodity flow formulation for the GMST problem, in which one unit of flow is sent from the root vertex to every group. For a subgraph $S$ of a graph $G$ with edge set $E$, associate a vector $x^S \in \mathbb{R}^E$, where edge variable $x_S^e$ takes value 1 if $e$ appears in $S$ and 0 otherwise. Each edge $(v, w) \in E$ has two binary variables $f^i_{v,w}$ and $s_{v,w}$: $f^i_{v,w}$ represents the amount of flow along edge ($v$, $w$) whose destination is group $G_i$ and variables $s_{v,w}$ are binary selection variables denoting the presence or absence of edge ($v$, $w$) in the group Steiner tree. The ILP is:

$$\min \sum_{v,\, w \in V} d_{v,\,w} s_{v,\,w} \tag{3}$$

$$\text{subject to} \quad \sum_{w \in V} f^i_{v,\,w} = \sum_{w \in V} f^i_{w,\,v} \text{ for all } v \in V \backslash (\cup_i R_i) \tag{4}$$

$$\sum_{v \in V} \sum_{t \in R_i} f^i_{v,\,t} = 1, \sum_{v \in V} \sum_{t \in R_i} f^i_{t,\,v} = 0, \sum_{v \in V} f^i_{r,\,v} = 1 \;\forall \text{ groups } R_i \tag{5}$$

$$0 \le f^i_{v,\,w} \le s_{v,\,w} \text{ for all } t \in T, \quad s_{v,\,w} \in \{0, 1\} \text{ for all } e \in E. \tag{6}$$

Constraints (4) impose flow conservation on all vertices not belonging to any group. Constraints (5) impose the inflow/outflow constraints on groups $R_i$. Finally, Constraints (6) impose the condition that there is positive flow on an edge only if the edge is selected. This ILP solves the DIP problem to optimality.

By Proposition 5.3, it is sufficient to solve this ILP on the Buneman graph $\mathcal{B}(H)$. The method of preprocessing using the Buneman graph and solving the resulting integer linear program has proved extremely efficient in practice, allowing the possibility of solving larger and more difficult datasets to optimality.

## 6. SIMULATION RESULTS

To mimic what was done in previous experimental studies (Mark Welch and Meselson, 2000; Schön and Martens, 2003), we considered only a single locus, where a locus is a collection of sites. No recombination was considered in our simulations. We implemented a forward simulator for a single locus in a diploid population of constant size $N$ undergoing discrete-time random mating with non-overlapping generations. Given $N$ diploid parents at generation $t-1$, individuals at generation $t$ were obtained as follows: With probability $1-p_s$, one parent was randomly chosen with replacement, and it produced a progeny via self-cloning. With probability $p_s$, a pair of parents was randomly chosen with replacement, and they produced a progeny via meiosis and mating. When producing a progeny, either by self-cloning or by sexual reproduction, new mutations were introduced according to a specified rate. This procedure was repeated until $N$ progenies were produced for generation $t$. With probability $g$, a gene-conversion event started between a pair of consecutive sites, and the tract length was assumed to be geometrically distributed with mean $\lambda$ (in gene-conversion, a new sequence is formed from a prefix of one sequence, followed by an internal segment of a second sequence, followed by a suffix of the first sequence). No crossovers were considered in our simulations.

Forward simulations are computationally intensive, so we used $N=1000$ to obtain simulations in a reasonable time. We started each simulation at generation 0 with $N$ identical diploid individuals, and then ran the simulation for $\tau = 4000$ generations with $p_s > 0$, followed by $\tau_A$ generations of asexual phase (i.e., with $p_s = 0$). Note that the average number of sexual reproductions in the history of the entire population is $\tau p_s N$. We took $n$ diploid samples at the end of each simulation. We performed the following simulations:

**S1:** Infinite-sites mutation model with the mutation rate fixed at $5 \times 10^{-3}$:

(i) In the first simulation, we set $g=0$, used varying values of $n$, $\tau_A$, and $p_s$, and performed 500 simulations for each parameter setting.

(ii) In the second simulation, we set $p_s=0$, gene-conversion rate $g>0$, and considered a locus consisting of 25000 sites.

**S2:** Finite-sites mutation model with homoplasy, using 25000 sites and mutation rate $u$ per locus; the per-site mutation rate $u$/25000. We fixed $n=40$ and used varying values of $u$, $\tau_A$, and $p_s$. We performed 50 simulations for each parameter setting.

**Infinite-sites case (S1):** Under this ideal toy model with no recombination or homoplasy, if the input data set does not admit a DPP solution, then it would indicate that sexual reproduction has played a role in

Table 2. Proportion of Data Sets Admitting DPP Solutions

| | | *Asexual phase* $\tau_A$ | | | | |
|---|---|---|---|---|---|---|
| *n* | $p_s$ | *0* | *100* | *500* | *1000* | *2000* |
| 10 | $1\times10^{-5}$ | 0.98 | 0.99 | 0.99 | 1.00 | 1.00 |
| 25 | $1\times10^{-5}$ | 0.96 | 0.99 | 0.99 | 0.99 | 1.00 |
| 50 | $1\times10^{-5}$ | 0.95 | 0.99 | 0.99 | 0.99 | 1.00 |
| 75 | $1\times10^{-5}$ | 0.94 | 0.98 | 0.99 | 0.99 | 1.00 |
| 10 | $1\times10^{-4}$ | 0.79 | 0.86 | 0.94 | 0.96 | 1.00 |
| 25 | $1\times10^{-4}$ | 0.67 | 0.79 | 0.93 | 0.96 | 1.00 |
| 50 | $1\times10^{-4}$ | 0.60 | 0.77 | 0.92 | 0.95 | 1.00 |
| 75 | $1\times10^{-4}$ | 0.56 | 0.75 | 0.92 | 0.95 | 1.00 |
| 10 | $1\times10^{-3}$ | 0.20 | 0.34 | 0.65 | 0.83 | 0.95 |
| 25 | $1\times10^{-3}$ | 0.08 | 0.20 | 0.57 | 0.80 | 0.95 |
| 50 | $1\times10^{-3}$ | 0.03 | 0.15 | 0.52 | 0.79 | 0.95 |
| 75 | $1\times10^{-3}$ | 0.01 | 0.14 | 0.50 | 0.78 | 0.95 |
| 10 | $1\times10^{-2}$ | 0.01 | 0.05 | 0.43 | 0.71 | 0.95 |
| 25 | $1\times10^{-2}$ | 0.00 | 0.01 | 0.29 | 0.65 | 0.94 |
| 50 | $1\times10^{-2}$ | 0.00 | 0.00 | 0.25 | 0.63 | 0.93 |
| 75 | $1\times10^{-2}$ | 0.00 | 0.00 | 0.24 | 0.62 | 0.93 |

Simulations performed a sexual phase (i.e., $p_s > 0$) of $\tau$ generations followed by an asexual phase (i.e., $p_s = 0$) of $\tau_A$ generations, with $g = 0$ throughout.

the evolutionary history. To assess our chance of detecting signatures of past sexual reproduction, we examined how often DPP solutions exist even if some amount of sexual reproduction actually took place in the evolutionary history of the population.

(i) The results of the first simulation are summarized in Table 2. These results suggest that infrequent sexual reproduction may be difficult to detect and that the signature of past sexual reproduction may decay rather quickly with time. However, note that the chance of detecting signatures of past sexual reproduction increases with the sample size *n*. Likewise, the chance increases with the number of segregating sites in the sample (results not shown).

(ii) The results of the second simulation are summarized in Table 3. These results indicate that the probability of the existence of a DPP solution decreases with increasing mean conversion tract length $\lambda$ and that even a moderate amount (i.e., comparable to the mutation rate) of gene-conversion can frequently lead to samples admitting no DPP solutions.

Instead of looking at one or a few genes at a time, as done in the past (Fontaneto et al., 2007; Mark Welch and Meselson, 2000; Schön and Martens, 2003), analyzing larger fractions of diploid genomes should increase the chance of detecting signatures of past sexual reproduction.

**Finite-sites case with homoplasy (S2):** To test the performance of the ILP described in Section 5.2, we analyzed data from the above-mentioned finite-sites simulation with homoplasy. We report the results obtained from solver CPLEX 12, but have also used the GNU Linear Programming Kit in order to release a free version of our software. We performed extensive testing to analyze the scaling behavior of our algorithms to larger number of sites and samples. While CPLEX is significantly faster for larger instances, GLPK is fast enough to illustrate the practicality of our algorithms on data sets of sizes of current biological interest. For each simulation instance, we used the ILP to find the minimum DIP and then calculated the imperfection $q$ (see Definition 5.2) of the resulting diploid genealogy. This number corresponds to the minimum number of back or recurrent mutations needed in addition to the number of mutations $\eta$ that would be present if the data admitted a DPP solution. As Table 4 shows, for each setting of the mutation rate $u$, increasing the probability $p_s$ of sexual reproduction or decreasing the number of generations in the asexual phase tends to increase the mean ratio $\frac{q}{\eta}$. This suggests that, for a given mutation rate, the amount of detected homoplasy may provide some information about past sexual reproduction. For most of the simulations, the solver CPLEX solved the resulting ILP in fractions of a second, with the largest instance taking 1.3 seconds.

Table 3. Simulations Performed with Asexual Reproduction with Gene-Conversion, Under a Finite-Sites Recombination Model with 25000 Sites

| | | *Mean tract length $\lambda$* | | | |
|---|---|---|---|---|---|
| $n$ | $g$ | *100* | *500* | *1000* | *1500* |
| 10 | $2\times10^{-9}$ | 0.99 | 0.95 | 0.93 | 0.91 |
| 25 | $2\times10^{-9}$ | 0.99 | 0.93 | 0.89 | 0.86 |
| 50 | $2\times10^{-9}$ | 0.98 | 0.92 | 0.88 | 0.83 |
| 75 | $2\times10^{-9}$ | 0.97 | 0.91 | 0.86 | 0.81 |
| 10 | $2\times10^{-8}$ | 0.93 | 0.69 | 0.58 | 0.47 |
| 25 | $2\times10^{-8}$ | 0.87 | 0.58 | 0.40 | 0.30 |
| 50 | $2\times10^{-8}$ | 0.84 | 0.49 | 0.32 | 0.21 |
| 75 | $2\times10^{-8}$ | 0.81 | 0.43 | 0.28 | 0.18 |
| 10 | $2\times10^{-7}$ | 0.40 | 0.07 | 0.02 | 0.01 |
| 25 | $2\times10^{-7}$ | 0.23 | 0.00 | 0.00 | 0.00 |
| 50 | $2\times10^{-7}$ | 0.16 | 0.00 | 0.00 | 0.00 |
| 75 | $2\times10^{-7}$ | 0.11 | 0.00 | 0.00 | 0.00 |
| 10 | $2\times10^{-6}$ | 0.01 | 0.00 | 0.00 | 0.00 |
| 25 | $2\times10^{-6}$ | 0.00 | 0.00 | 0.00 | 0.00 |
| 50 | $2\times10^{-6}$ | 0.00 | 0.00 | 0.00 | 0.00 |
| 75 | $2\times10^{-6}$ | 0.00 | 0.00 | 0.00 | 0.00 |

We used $N = 1000$, $\tau = 4000$, and $u = 5\times10^{-3}$ per locus.

## 7. CONCLUSION

In this article, we considered a new problem in phylogenetics. Reconstructing the genealogy of diploid individuals is not only an interesting problem, but also has important practical applications. We believe that such a genealogical approach offers much more than can existing tests based on counting sequence differences (Mark Welch and Meselson, 2000; Schön and Martens, 2003) or considering a single haplotype per individual (Frumkin et al., 2005). To gain intuition on this new problem, we have explored algorithmic aspects of reconstructing diploid genealogies. It remains an important open problem to develop a sound

Table 4. Average Ratio $Q/\eta$ of the Amount of Homoplasy to the Total Number of Mutating Sites in the Finite-Sites Simulation Study S2, with $N = 40$ and 25000 Sites

| | | *Asexual phase $\tau_A$* | | | |
|---|---|---|---|---|---|
| $u$ | $p_s$ | *100* | *500* | *1000* | *2000* |
| $1\times10^{-3}$ | $1\times10^{-5}$ | 0.159 | 0.134 | 0.126 | 0.122 |
| $1\times10^{-3}$ | $1\times10^{-4}$ | 0.212 | 0.181 | 0.096 | 0.091 |
| $1\times10^{-3}$ | $1\times10^{-3}$ | 0.265 | 0.261 | 0.247 | 0.213 |
| $1\times10^{-3}$ | $1\times10^{-2}$ | 0.559 | 0.290 | 0.242 | 0.112 |
| $2\times10^{-3}$ | $1\times10^{-5}$ | 0.162 | 0.157 | 0.159 | 0.129 |
| $2\times10^{-3}$ | $1\times10^{-4}$ | 0.179 | 0.169 | 0.132 | 0.124 |
| $2\times10^{-3}$ | $1\times10^{-3}$ | 0.445 | 0.254 | 0.241 | 0.161 |
| $2\times10^{-3}$ | $1\times10^{-2}$ | 0.469 | 0.241 | 0.229 | 0.165 |
| $3\times10^{-3}$ | $1\times10^{-5}$ | 0.098 | 0.092 | 0.099 | 0.091 |
| $3\times10^{-3}$ | $1\times10^{-4}$ | 0.115 | 0.105 | 0.119 | 0.069 |
| $3\times10^{-3}$ | $1\times10^{-3}$ | 0.344 | 0.209 | 0.139 | 0.109 |
| $3\times10^{-3}$ | $1\times10^{-2}$ | 0.496 | 0.231 | 0.158 | 0.134 |
| $4\times10^{-3}$ | $1\times10^{-5}$ | 0.074 | 0.119 | 0.087 | 0.083 |
| $4\times10^{-3}$ | $1\times10^{-4}$ | 0.147 | 0.136 | 0.109 | 0.098 |
| $4\times10^{-3}$ | $1\times10^{-3}$ | 0.301 | 0209 | 0.138 | 0.097 |
| $4\times10^{-3}$ | $1\times10^{-2}$ | 0.440 | 0.206 | 0.183 | 0.136 |

$n$, number of diploid individuals sampled; $p_{s,}$ probability of sexual reproduction; $u$, per-locus mutation rate.

statistical framework for studying the evolutionary history of asexual diploids, allowing for occasional sexual reproduction, recombination, and hybridization. Explicitly modeling the genealogy of asexual diploids will help to address a number of important questions in evolutionary biology: Could it be that sexual reproduction has actually occurred in the history of reputed ancient asexuals? If so, how big a role has sexual reproduction played in their long-term evolutionary success? If not, when was sexuality lost and how many independent times? For those species that are mainly asexual but occasionally reproduce sexually, how can we estimate the frequency of sexual reproduction? Can we distinguish the effects of mitotic recombination from that of past sexual reproduction? How does natural selection act on asexual diploids? The work described in this article is a modest step toward addressing such questions. If a DPP for two-state characters exists, we have shown that its root is uniquely determined and that there exists a unique minimal DPP. These results seem to be useful properties likely to have concrete applications in ancestral inference. Although we have focused on the single-locus case in this article, the algorithms developed here can easily be generalized to handle multiple loci, with unspecified relative phasing between the haplotypes at different loci. Depending on the distribution of variation within a locus, there could be much more power (to detect deviations from DPP) if fewer sites in many unlinked loci were surveyed. Identification of the optimal strategy to detect no DPP under various parameterizations is of practical importance, and the approach we have taken here can easily be extended to address this power and design issue.

As mentioned in the introduction, no significantly high level of sequence divergence was observed in the purportedly ancient asexual organism darwinulid ostrocods. It remains an open question whether this finding for ostrocods can be attributed to gene-conversion. It would be interesting to extend the work described in this article to develop a method of reconstructing parsimonious diploid genealogies that explicitly incorporate sexual reproduction and gene-conversion. As a first step, it will be interesting to investigate whether there exists an efficient algorithm for reconstructing diploid genealogies with constrained patterns of sexual reproduction and recombination, similar to the recent work on the so-called galled-trees (Gusfield, 2005; Gusfield et al., 2004; Huynh et al., 2005; Nakhleh et al., 2004). Although we have focused on DPP for two-state characters in this article, generalizing the work to handle multi-state characters and polyploidy seems possible. (For all fixed number of states, polynomial-time algorithms exist for the haploid perfect phylogeny problem [Agarwala and Fernandez-Baca, 1994; Kannan and Warnow, 1997].) These results may have interesting implications for phylogenetics and ancestral inference involving asexual organisms.

## ACKNOWLEDGMENTS

We thank Dan Gusfield for many helpful comments on a preliminary version of this manuscript. This research is supported in part by the National Institutes of Health (grant K99-GM080099 to Y.S.S., and grant R01-HG002942 to C.H.L.); by the National Science Foundation (grant IIS-0513910 to C.H.L., grant CCF-0515378 to F.L., and grant IIS-0803564 to F.L.); and by a Packard Fellowship for Science and Engineering (to Y.S.S.).

## DISCLOSURE STATEMENT

No competing financial interests exist.

## REFERENCES

Agarwala, R., and Fernández-Baca, D. 1994. A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed. *SIAM J. Comput.* 23, 1216–1224.

Bafna, V., Gusfield, D., Lancia, G., et al. 2003. Haplotyping as perfect phylogeny: a direct approach. *J. Comput. Biol.* 10, 323–340.

Bandelt, H.J., Forster, P., Sykes, B.C., et al. 1989. Mitochondrial portraits of human populations using median networks. *Genetics* 141, 743–753.

Barton, N.H., and Charlesworth, B. 1998. Why sex and recombination? *Science* 281, 1986–1990.

Birky, C.W., Jr. 2004. Bdelloid rotifers revisited. *Proc. Natl. Acad. Sci. USA* 101, 2651–2652.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity, 387–395. *In* Kendall, D.G., and Tautu, P., eds. *Mathematics in the Archeological and Historical Sciences.* Edinburgh University Press, Edinburgh, UK.

Ding, Z., Filkov, V., and Gusfield, D. 2005. A linear-time algorithm for the perfect phylogeny haplotyping. *Lect. Notes Bioinformatics* 3500, 585–600.

Eskin, E., Halperin, E., and Karp, R. 2003. Efficient reconstruction of haplotype structure via perfect phylogeny. *J. Bioinformatics Comput. Biol.* 1, 1–20.

Fernandez-Baca, D., and Lagergren, J. 2003. A polynomial-time algorithm for near-perfect phylogeny. *SIAM J. Comput.* 32, 1115–1127.

Fontaneto, D., Herniou, E.A., Boschetti, C., et al. 2007. Independently evolving species in asexual bdelloid rotifers. *PLoS Biol.* 5, e87.

Foulds, L., and Graham, R. 1982. The Steiner problem in phylogeny is NP-complete. *Adv. App. Math.* 3, 299.

Frumkin, D., Wasserstrom, A., Kaplan, S., et al. 2005. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* 1, e50.

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. *Proc. RECOMB 2002* 166–175.

Gusfield, D. 2005. Optimal, efficient reconstruction of Root-Unknown phylogenetic networks with constrained recombination. *J. Comput. Syst. Sci.* 70, 381–398.

Gusfield, D., Eddhu, S., and Langley, C. 2004. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinformatics Comput. Biol.* 2, 173–213.

Halperin, E., and Eskin, E. 2004. Haplotype reconstruction from genotype data using Imperfect Phylogeny. *Bioinformatics* 20, 1842–1849.

Huynh, T.N.D., Jansson, J., Nguyen, N.B., et al. 2005. Constructing a smallest refining galled phylogenetic network. *Proc. RECOMB, 2005* 265–280.

Judson, P.O., and Normark, B.B. 1996. Ancient asexual scandals. *Trends Ecol. Evol.* 11, 41–46.

Kannan, S., and Warnow, T. 1997. A fast algorithm for the computation and enumeration of perfect phylogenies when the number of character states is fixed. *SIAM J. Comput.* 26, 1749–1763.

Keightley, P.D., and Otto, S.P. 2006. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* 443, 89–92.

Lunt, D. 2008. Genetic tests of ancient asexuality in root knot nematodes reveal recent hybrid origins. *BMC Evol Biol.* 8, 194.

Mark Welch, D., and Meselson, M. 2000. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* 288, 1211–1215.

Martens, K., Rossetti, G., and Horne, D.J. 2003. How ancient are ancient asexuals? *Proc. R. Soc. Lond. B* 270, 723–729.

Maynard Smith, J. 1978. *The Evolution of Sex*. Cambridge University Press, Cambridge, UK.

Maynard Smith, J. 1986. Contemplating life without sex. *Nature* 324, 300–301.

Nakhleh, L., Warnow, T., and Linder, C. 2004. Reconstructing reticulate evolution in species—theory and practice. *Proc. RECOMB 2004* 337–346.

Schön I., and Martens, K. 2003. No slave to sex. *Proc. R. Soc. London B* 270, 827–833.

Semple, C., and Steel, M. 2003. *Phylogenetics*. Oxford University Press, Oxford, UK.

Sridhar, S., Blelloch, G.E., Ravi, R., et al. 2006. Optimal imperfect phylogeny reconstruction and haplotyping. *Proc. Comput. Syst. Bioinformatics* 199–210.

Sridhar, S., Dhamdhere, K., Blelloch, G.E., et al. 2006. Simple reconstruction of binary near-perfect phylogenetic trees. *Proc. Int. Workshop Bioinformatics Res. Appl.* 799–806.

Sridhar, S., Lam, F., Blelloch, G., et al. 2007. Efficiently finding the most parsimonious phylogenetic tree via linear programming. *Proc ISBRA* 37–48.

Vijayasatya, R., and Mukherjee, A. 2005. An efficient algorithm for perfect phylogeny haplotyping. *Proc. IEEE Comput. Syst. Bioinformatics Conf.* 103–110.

Waggoner, B.M., and Poinar, G.O. Jr. 1993. Fossil habrotrochid rotifers in Dominican amber. *Experientia* 49, 354–357.

Address correspondence to:<br>
*Dr. Fumei Lam*
*Department of Computer Science*
*University of California, Davis*
*Davis, CA 95616*

*E-mail:* flam@cs.ucdavis.edu