On the Genealogy of Asexual Diploids

Fumei Lam[1], Charles H. Langley[2], and Yun S. Song[3,4]

[1] Department of Computer Science, University of California, Davis, CA 95616, USA
[2] Section of Evolution and Ecology, University of California, Davis, CA 95616, USA
[3] Computer Science Division, University of California, Berkeley, CA 94720, USA
[4] Department of Statistics, University of California, Berkeley, CA 94720, USA
flam@cs.ucdavis.edu, chlangley@ucdavis.edu, yss@eecs.berkeley.edu

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 paper, 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. 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. We apply our algorithms on simulated data sets with sizes of biological interest.

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 (see [4, 21] and references therein). 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 sexual counterparts that undergo meiosis and mating [25]. 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 4 families and 18 genera. Fossil evidence suggests that bdelloid rotifers have been around for at least 35 to 40 million years [34], while molecular genetic analysis suggests an age that is more than twice as large [23]. Maynard Smith [26] 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 [19].

Recently, Mark Welch and Meselson [23] 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 individual, which, under neutrality, are expected to be over-represented in an old asexual organism. (See [5] for a review.) 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 [10].) In contrast to this success, when a similar analysis was applied to other asexual organisms such as darwinulid ostrocods [28], of which morphological evidence strongly supports ancient asexuality [24], 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 [22]. 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 paper, 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.

In the second part of this paper, 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 evolutionary history. 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 paper 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.

CloneTree, software that implements our algorithms, will be made publicly 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 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 (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. Trees representing evolutionary histories consistent with the infinite-sites model of mutation are called perfect phylogenies [29]. We will refer to these as *haploid perfect phylogenies (HPP)* to distinguish them from *diploid perfect phylogenies*, a new concept defined as follows.

Ind	H
1	10010 00001
2	00001 10010
3	01001 00110
4	00110 01001

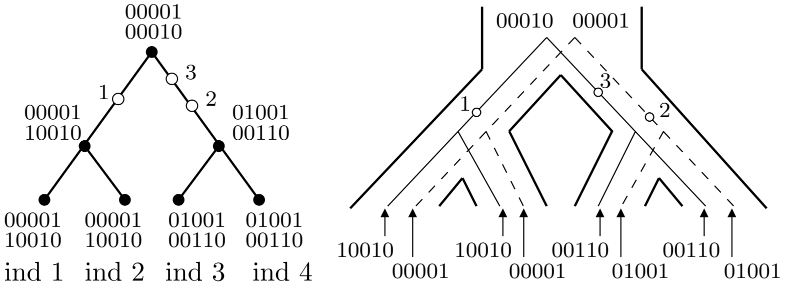

Fig. 1. From left to right, a haplotype data set H for four individuals, its unique minimal diploid perfect phylogeny T, and evolutionary histories of the haplotypes embedded in T. We use τ_X and τ_Y to denote the solid and the dotted trees in T, respectively. An open circle labeled k represents a mutation at site k.

Definition 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 paper, 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.

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

In [13], Gusfield 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 τ_X. Similarly, follow the history of y_r on T to obtain H_Y and τ_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 τ_X and τ_Y are disjoint. (In Figure 1, sites 1 and 3 mutate in τ_X but not in τ_Y. Similarly, site 2 mutates in τ_Y but not in τ_X.)

P2. If $x_r[k] \neq y_r[k]$, then both 0 and 1 have already been seen, so part 1 of Definition 3 implies that neither τ_X nor τ_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 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 $k \notin \mathcal{E}(H)$, 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 any at 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 ρ of T is homozygous at some site $j \in \mathcal{E}(H)$. Then, since every individual in H is heterozygous at that site, ρ is not in H. Now, the one-mutation-per-site condition implies that there is an edge that separates ρ 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 the entire T, and the second part of the lemma immediately follows. □

Lemma 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 4 (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 ≥ 3, and (iv) the remaining vertices and edges of $\mathcal{G}$ remain the same.*

Definition 5 ($\bowtie$, 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 1 (Partition of H into H_X and H_Y). *For $n > 1$, suppose there exists an n-leaved diploid perfect phylogeny T for the n individuals in H. Then, the $2n$ haplotypes in H admit a unique $2n$-leaved minimal unrooted haploid perfect phylogeny τ satisfying the following:*

1. *If $\mathcal{E}(H) \neq \varnothing$, then there exists a unique edge in τ such that cutting that edge partitions τ 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. (See Figure 2a.)*
2. *If $\mathcal{E}(H) = \varnothing$, then there exists a unique vertex v in τ with degree d, where $d > 3$, such that resolving v and cutting the edge between the newly created vertices partitions τ 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. (See Figure 2b.)*

Proof. Define τ_X, τ_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 τ_X and the root of τ_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 τ_X and τ_Y by identifying the root vertex ρ_X of τ_X with the root vertex ρ_Y of τ_Y, such that the new vertex $\rho_{X \oplus Y}$ obtained from identifying ρ_X and ρ_Y is incident with all the edges that were incident with ρ_X or ρ_Y. Then, Properties P1 and P2 implies that the resulting tree $\tau_{X \oplus Y}$ is an unrooted HPP for the haplotypes in H. Now, contract interior edges in $\tau_{X \oplus Y}$

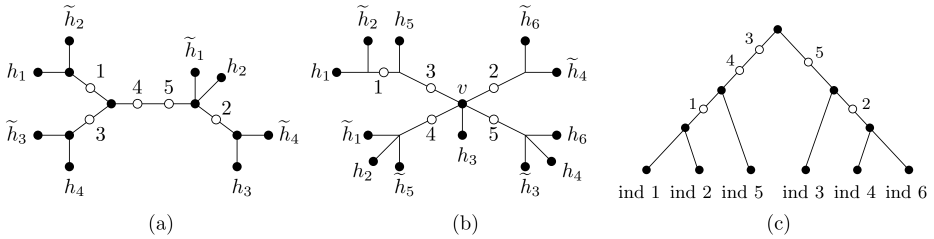

Fig. 2. HPP and DPP examples for Propositions 1 and 2. (a) There is a unique edge (namely, the edge on which sites 4 and 5 mutate) satisfying the property described in part 1 of Proposition 1. (b) There is a unique vertex (labeled v in the figure) satisfying the property described in part 2 of Proposition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

with no mutations and call the resulting tree τ. Note that τ is a unique $2n$-leaved minimal unrooted HPP for the haplotypes in H.

If $\mathcal{E}(H) \neq \varnothing$, then x_r and y_r are not identical, so τ described above contains an edge between x_r and y_r with at least one mutation. Therefore, τ satisfies part 1 of the Proposition. If $\mathcal{E}(H) = \varnothing$, then Lemma 1 implies that the root haplotypes x_r and y_r are identical, so τ 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. □

If a DPP exists for H and $\mathcal{E}(H) = \varnothing$, part 2 of Proposition 1 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, depicted in Figure 2c. Now, the following result establishes the uniqueness of a minimal DPP solution:

Proposition 2 (Uniqueness). *If a DPP exists for H, then H admits a unique minimal DPP.*

Proof. Suppose that H admits a DPP. If $\mathcal{E}(H) \neq \varnothing$, then part 1 of Proposition 1 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 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 $\mathcal{E}(H) = \varnothing$, and let τ and v be as in Proposition 1. Lemma 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 τ, one can associate a binary character for the n individuals $\{1, 2, \ldots, n\}$ as follows. For each mutation occurring on some edge in τ, 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 [13], we can devise the following $O(mn)$-time algorithm to find a DPP solution, if it exists:

1. Check that the conditions in Lemma 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 τ 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 1.
   (a) If $\mathcal{E}(H) \neq \varnothing$, the two ends of the edge in τ needed to be cut should be labeled by x_r and y_r.
   (b) If $\mathcal{E}(H) = \varnothing$, then $x_r = y_r$. The vertex v described in the second part of Proposition 1 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

In [14], Gusfield considered phasing (or haplotyping) genotypic input data as an HPP and provided a nearly-linear-time algorithm for the problem. Simpler but slower solutions [2, 8] were subsequently proposed for the problem, and linear-time algorithms were recently found [7, 33]. 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

Lemma 1 implies that genotypic states 0 and 1 (denoting homozygotes) 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 if both 0 and 1 appear in column k. If so, there is no DPP solution.
   (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 (1), then set $z_r[k] = 0$ ($z_r[k] = 1$).
2. If the above step has not failed, 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 linear-time HPP algorithm for binary matrices [13]. 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.

Due to space considerations, we omit the details of our algorithm for finding a DPP haplotyping solution.

5 Diploid Imperfect Phylogeny

If a set of diploid sequences does not allow a diploid perfect phylogeny, 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.

Definition 6 (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 Diploid Perfect Phylogeny, 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 diploid perfect phylogeny. For a diploid imperfect phylogeny 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 7. *A diploid imperfect phylogeny T for input H is q-imperfect (or q-near-perfect) if* $\sum_{k:M_T(k)\geq 1}(M_T(k)-1)=q$.

The diploid imperfect phylogeny 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 diploid *perfect* phylogeny 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 [9], Halperin and Eskin [17], and Sridhar et al. [31] analyzed theoretical bounds for algorithms to solve this problem to optimality, while Sridhar et al. [30] 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 [32]. We now consider the case of constructing diploid imperfect phylogenies and introduce a problem which casts this problem in the framework of combinatorial optimization.

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 [11].

To solve the *diploid* imperfect phylogeny 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 $R = R_1 \cup R_2 \dots R_k \subset V$ 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 diploid imperfect phylogeny 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 diploid imperfect phylogeny 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 $\{h_i\widetilde{h_i}, \widetilde{h_i}h_i\} \subseteq V(G)$. The GMST on this instance is then equivalent to the minimum diploid imperfect phylogeny 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. In the haploid imperfect phylogeny problem, it has been shown that the *Buneman graph* of the input sequences contains all optimal trees [3, 6]. Restricting the underlying graph of the problem in such a way has been shown to be efficient and practical on real data sets [32]. The following proposition shows an analogous results holds for the diploid phylogeny problem:

Proposition 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 $\cup_i\{h_i\widetilde{h_i}, \widetilde{h_i}h_i\}$. 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. for haploid imperfect phylogeny construction:

Theorem 1 (Bandelt et al.). *[3, 29] 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 (Proposition 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 $R_i = \{h_i\widetilde{h_i}, \widetilde{h_i}h_i\}$. It follows that $T^*(H)$ is a minimum Steiner tree on terminal set $\{t_i\}_{i=1}^n$. By Theorem 1, $T^*(H)$ is a subgraph of the Buneman graph $\mathcal{B}(\{t_i\}_{i=1}^n)$. Since $t_i \in \{h_i\widetilde{h_i}, \widetilde{h_i}h_i\}$, it follows that $T^*(H)$ is a subgraph of the Buneman graph $\mathcal{B}(\{h_i\widetilde{h_i}, \widetilde{h_i}h_i\}_{i=1}^n) = \mathcal{B}(H)$. □

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, associate a vector $x^S \in \mathbb{R}^E$, where edge variable x_e^S takes value 1 if e appears in S and 0 otherwise. Each edge $(v, w) \in E$ has two binary variables $f_{v,w}^i$ and $s_{v,w}$: $f_{v,w}^i$ 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:

$$\min \sum_{v,w \in V} d_{v,w} s_{v,w} \tag{1}$$

$$\text{subject to} \quad \sum_{w \in V} f^i_{v,w} = \sum_{w \in V} f^i_{w,v} \text{ for all } v \in V \setminus (\cup_i R_i) \tag{2}$$

$$\sum_{v \in V} \sum_{t \in R_i} f^i_{v,t} = 1, \sum_{v \in V} \sum_{t \in R_i} f^i_{t,v} = 0, \sum_{v \in V} f^i_{r,v} = 1 \ \forall \text{ groups } R_i \tag{3}$$

$$0 \leq f^i_{v,w} \leq s_{v,w} \text{ for all } t \in T, \quad s_{v,w} \in \{0,1\} \text{ for all } e \in E. \tag{4}$$

Constraints (2) impose flow conservation on all vertices not belonging to any group. Constraints (3) impose the inflow/outflow constraints on groups R_i. Finally, Constraints (4) impose the condition that there is positive flow on an edge only if the edge is selected. This ILP solves the diploid imperfect phylogeny problem to optimality.

6 Simulation Results

To mimic what was done in the past experimental studies [23,28], 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.

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 τ_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 two different types of simulation:

S1. Infinite-sites mutation model with the mutation rate fixed at 5×10^{-3} per locus. We used varying values of n, τ_A and p_s, and performed 500 simulations for each parameter setting.

S2. Finite-sites mutation model with homoplasy, using 25000 sites and mutation rate u per locus; the per-site mutation rate is $u/25000$. We fixed $n = 40$ and used varying values of u, τ_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 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. The results are summarized in Table 1(a). 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). Instead of looking at one or a few genes at a time, as done in the past [10,23,28], 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 7) 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 η that would be present if the data admitted a DPP solution. As Table 1(b) 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.

7 Discussion

In this paper, 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 [23, 28] or considering a single haplotype per individual [12]. 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 statistical framework for studying the evolutionary history of asexual diploids, allowing for occasional sexual reproduction, recombination, and hybridization.

Table 1. Simulation results discussed in Section 6. (a) Proportion of data sets admitting DPP solutions in the infinite-sites simulation study S1, with the mutation rate $= 5 \times 10^{-3}$ per locus. (b) Average ratio q/η 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.

(a)

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

(b)

		Asexual phase τ_A			
u	p_s	100	500	1000	2000
1×10^{-3}	1×10^{-5}	0.159	0.134	0.126	0.122
1×10^{-3}	1×10^{-4}	0.212	0.181	0.096	0.091
1×10^{-3}	1×10^{-3}	0.265	0.261	0.247	0.213
1×10^{-3}	1×10^{-2}	0.559	0.290	0.242	0.112
2×10^{-3}	1×10^{-5}	0.162	0.157	0.159	0.129
2×10^{-3}	1×10^{-4}	0.179	0.169	0.132	0.124
2×10^{-3}	1×10^{-3}	0.445	0.254	0.241	0.161
2×10^{-3}	1×10^{-2}	0.469	0.241	0.229	0.165
3×10^{-3}	1×10^{-5}	0.098	0.092	0.099	0.091
3×10^{-3}	1×10^{-4}	0.115	0.105	0.119	0.069
3×10^{-3}	1×10^{-3}	0.344	0.209	0.139	0.109
3×10^{-3}	1×10^{-2}	0.496	0.231	0.158	0.134
4×10^{-3}	1×10^{-5}	0.074	0.119	0.087	0.083
4×10^{-3}	1×10^{-4}	0.147	0.136	0.109	0.098
4×10^{-3}	1×10^{-3}	0.301	0209	0.138	0.097
4×10^{-3}	1×10^{-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.

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 paper is a modest step toward addressing such questions.

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 paper 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 [15, 16, 18, 27]. Although we have focused on diploid perfect phylogeny for two-state characters in this paper, 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. See [1, 20].)

Acknowledgment

We thank Dan Gusfield for many helpful comments on a preliminary version of this manuscript. This research is supported in part by NIH grants K99-GM080099 (YSS) and R01-HG002942 (CHL); by NSF grants IIS-0513910 (CHL), CCF-0515378 (FL), and IIS-0803564 (FL); and by a Packard Fellowship for Science and Engineering (YSS).

References

1. R. Agarwala and D. Fernández-Baca. A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed. *SIAM J. Computing*, 23:1216–1224, 1994.
2. V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: A direct approach. *J. Comput. Biol.*, 10:323–340, 2003.
3. H. J. Bandelt, P. Forster, B. C. Sykes, and M. B. Richards. Mitochondrial portraits of human populations using median networks. *Genetics*, page 743753, 1989.
4. N. H. Barton and B. Charlesworth. Why sex and recombination? *Science*, 281:1986–1990, 1998.
5. C. W. Birky, Jr. Bdelloid rotifers revisited. *Proc. Nat. Acad. Sci.*, 101:2651–2652, 2004.
6. P. Buneman. The recovery of trees from measures of dissimilarity. In F. H. et al., editor, *Mathematics in the Archeological and Historical Sciences*, pages 387–395. Edinburgh University Press, 1971.
7. Z. Ding, V. Filkov, and D. Gusfield. A linear-time algorithm for the perfect phylogeny haplotyping. In *Proc. 9th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, volume 3500 of *LNBI*, pages 585–600, 2005.
8. E. Eskin, E. Halperin, and R. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *J. Bioinf. Comput. Biol.*, 1:1–20, 2003.
9. D. Fernandez-Baca and J. Lagergren. A polynomial-time algorithm for near-perfect phylogeny. *SIAM Journal on Computing*, 32:1115–1127, 2003.
10. D. Fontaneto, E. A. Herniou, C. Boschetti, M. Caprioli, G. Melone, C. Ricci, and T. G. Barraclough. Independently evolving species in asexual bdelloid rotifers. *PLoS Biology*, 5(4):e87, 2007.
11. L. Foulds and R. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3(43-49):299, 1982.
12. D. Frumkin, A. Wasserstrom, S. Kaplan, U. Feige, and E. Shapiro. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.*, 1(5):e50, 2005.
13. D. Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21:19–28, 1991.

14. D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proc. 6th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, pages 166–175, 2002.
15. D. Gusfield. Optimal, efficient reconstruction of Root-Unknown phylogenetic networks with constrained recombination. *J. Comput. Sys. Sci.*, 70:381–398, 2005.
16. D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinf. Comput. Biol.*, 2:173–213, 2004.
17. E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using Imperfect Phylogeny. *Bioinformatics*, 20:1842–1849, 2004.
18. T. N. D. Huynh, J. Jansson, N. B. Nguyen, and W.-K. Sung. Constructing a smallest refining galled phylogenetic network. In *Proc. 9th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, pages 265–280, 2005.
19. P. O. Judson and B. B. Normark. Ancient asexual scandals. *Trends Ecol. Evol.*, 11:41–46, 1996.
20. S. Kannan and T. Warnow. A fast algorithm for the computation and enumeration of perfect phylogenies when the number of character states is fixed. *SIAM J. Computing*, 26:1749–1763, 1997.
21. P. D. Keightley and S. P. Otto. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443:89–92, 2006.
22. D. Lunt. Genetic tests of ancient asexuality in root knot nematodes reveal recent hybrid origins. *BMC Evolutionary Biology*, 8:194, 2008.
23. D. Mark Welch and M. Meselson. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science*, 288:1211–1215, 2000.
24. K. Martens, G. Rossetti, and D. J. Horne. How ancient are ancient asexuals? *Proc. R. Soc. London B*, 270:723–729, 2003.
25. J. Maynard Smith. *The Evolution of Sex*. Cambridge University Press, Cambridge, UK, 1978.
26. J. Maynard Smith. Contemplating life without sex. *Nature*, 324:300–301, 1986.
27. L. Nakhleh, T. Warnow, and C. Linder. Reconstructing reticulate evolution in species – theory and practice. In *Proc. 8th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, pages 337–346, 2004.
28. I. Schön and K. Martens. No slave to sex. *Proc. R. Soc. London B*, 270:827–833, 2003.
29. C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, UK, 2003.
30. S. Sridhar, G. E. Blelloch, R. Ravi, and R. Schwartz. Optimal imperfect phylogeny reconstruction and haplotyping. In *Proceedings of Computational Systems Bioinformatics*, pages 199–210, 2006.
31. S. Sridhar, K. Dhamdhere, G. E. Blelloch, E. Halperin, R. Ravi, and R. Schwartz. Simple reconstruction of binary near-perfect phylogenetic trees. In *Proceedings of International Workshop on Bioinformatics Research and Applications*, pages 799–806, 2006.
32. S. Sridhar, F. Lam, G. Blelloch, R. Ravi, and R. Schwartz. Efficiently finding the most parsimonious phylogenetic tree via linear programming. In *Proceedings of International Symposium on Bioinformatics Research and Applications (ISBRA)*, pages 37–48, 2007.
33. R. Vijayasatya and A. Mukherjee. An efficient algorithm for perfect phylogeny haplotyping. In *Proc. IEEE Comput. Syst. Bioinform. Conf.*, pages 103–10, 2005.
34. B. M. Waggoner and G. O. Poinar, Jr. Fossil habrotrochid rotifers in Dominican amber. *Experientia*, 49(4):354–357, 1993.