Far East Journal of Mathematical Sciences (FJMS)
Volume 28, Issue 2, 2008, Pages 319-339

ANALYSES OF HAPLOTYPE INFERENCE ALGORITHMS

SEAN CLEARY and KATHERINE ST. JOHN

Department of Mathematics
The City College of New York and The CUNY Graduate Center
New York, NY 10031, U. S. A.
e-mail: cleary@sci.ccny.cuny.edu

Department of Mathematics and Computer Science
Lehman College and Departments of Anthropology & Computer Science
The Graduate Center
City University of New York
Bronx, NY 10468, U. S. A.
e-mail: stjohn@lehman.cuny.edu

Abstract

We present experimental and theoretical analyses of data requirements for haplotype inference algorithms. Our experiments include a broad range of problem sizes under two standard models of tree distribution and were designed to yield statistically robust results despite the size of the sample space. Our results validate Gusfield's conjecture that a population size of $n \log n$ is required to give (with high probability) sufficient information to deduce the n haplotypes and their complete evolutionary history. The experimental results inspired our theoretical bounds on the population size. We also analyze the population size required to deduce some fixed fraction of the evolutionary history of a

---

2000 Mathematics Subject Classification: 68R05, 92C40, 05C05.

Keywords and phrases: combinatorial algorithms, molecular biology, trees.

Both authors acknowledge support from NSF grant #02-15942 and are grateful for hospitality of the Centre de Recerca Matemàtica.

Submitted by P. Mark Kayll

Received July 14, 2007; Revised December 17, 2007

set of n haplotypes and establish linear bounds on the required sample size. These linear bounds are also shown theoretically.

1. Introduction

Though DNA sequences of any two people are about 99.9% identical, the variations, however slight, may greatly affect an individual's risk for disease and response to different drug treatments [16]. Sites in the DNA sequence where individuals differ at a single DNA base are called *single nucleotide polymorphisms* (*SNPs*). The pattern of SNPs on a block (or continuous segment along the genome which is essentially non-recombinant) is a *haplotype*. DNA genotype data for an individual consists of the union of two sets of haplotype data, one from each parent. To understand the genetic makeup of a population fully, it is essential to understand not just the genotypes present in the population, but the set and distribution of haplotypes in the population and the evolutionary history of the haplotypes. The next important large scale project, following the success of the Human Genome Project is the Haplotype Mapping (HapMap) Project [17, 4]. The HapMap project is a data resource for genetic association studies, and as such seeks to build a map of these haplotype blocks, including the specific SNPs that identify the haplotypes. It is relatively inexpensive to get genotype data biochemically (in a "wet lab") but it is still very expensive to determine haplotype data this way, making this a suitable and attractive challenge for algorithm designers.

Many algorithms have been developed to computationally infer haplotypes, given only the genotype data, avoiding the great cost in biochemically determining haplotype information. Several statistical approaches have been developed to do this [6, 21], as well as parsimony approaches [3]. Gusfield et al. [2, 9, 11, 10] developed promising combinatorial algorithmic techniques for taking genotype information and computationally deducing the most likely set of haplotypes and their evolutionary history. To deduce the evolutionary history, he relies on the existence of a perfect phylogeny for the mutations in the population (basically, a binary tree), for which there is compelling biochemical evidence. Computational algorithms that give accurate estimates of haplotype avoid the great cost in biochemically determining haplotype information and hold a great deal of promise for efficiently yielding the evolutionary history of the set of haplotypes in the population. We focus on the lower bounds for the amount of data required for such algorithms to have a high probability of success. Gusfield [9] conjectured that the sample size required to infer the haplotype evolutionary history for a set of n haplotypes would be on the order of $n \log n$.

We present results of theoretical and experimental analyses of population size requirements for haplotype inference algorithms needed to determine the haplotypes with high accuracy and we establish Gusfield's conjecture. We also analyze theoretically and experimentally the data sufficiency requirements to determine a fixed fraction of the evolutionary history of a set of haplotypes. The paper is organized as follows: Section 2 covers background of haplotyping and evolutionary trees. In Section 3, we discuss path coverage of trees in the case where there is the minimal possible number of individuals necessary. In Section 4, we bound the expected sample size requirements for complete as well fractional coverage of trees. In Section 5, we describe the experimental results from simulation studies.

2. Background

2.1 Haplotypes. Each diploid individual has two sets of chromosomes, one from each parent, and thus two copies of every gene. The chromosome copies often differ only by a single base at a site in the DNA sequence; these are the SNPs described above. SNPs account for much of the variation in the human genome, and there are about 10 million SNPs that are common in human populations (see [8, 18, 4] for details). Genotyping of a person reveals the values at the SNPs from the two copies of the genes, but not from which parent it was inherited. For example, in Figure 1, the two copies of the gene for Person A differ at site 1. From the genotype data, we do not know if the "A" seen at site 1 came from the maternal or paternal copy of the gene. It is relatively quick and inexpensive to find genotypes in a wet lab, but difficult and expensive to find the haplotypes, or the "halves" that came from each parent. Figure 1 gives a possible set of haplotypes for the genotypes given. Note identical genotype data could come from more than one set of haplotype data, and collecting genotype data from relatives can facilitate haplotyping efforts [21]. In general, there can be as many as 2^{k-1} possible haplotype arrangements for haplotype data collected at k sites.

	Person A	Person B
Site 1	A:G	A:G
Site 2	T:C	T:T

A	maternal	paternal
	G	A
	T	C

B	maternal	paternal
	G	A
	T	T

Figure 1. The left table shows the genotype data from two people. The other tables show one possible haplotype data for the two parents. Note that for Person B the second site is the same from both parents. There are three other possible haplotype configurations for Person A and one other possible haplotype configuration for Person B.

2.2 Evolutionary trees. Trees are often used to model evolution between species in computational biology. The simplest models use rooted binary trees, with each leaf of the trees representing a different taxon. The root node represents the common ancestor of a collection of taxon, and mutations occur on edges, or branches, or the tree. A perfect phylogeny is an evolutionary history of the data in the case where each allele evolved at only one place in the tree.

Similarly, trees can be used to model evolution on haplotypes. Again, rooted binary trees are used to represent the ancestral history of a collection of haplotypes present in a population, with interior nodes representing mutations in the past. Gusfield's algorithm [9] for haplotype inference assumes a perfect phylogeny. It is far more efficient than the exhaustive enumeration of all possible haplotypes and evolutionary histories of those haplotypes (which has super-exponential running time) but has some limitations, as does any such computational inference. The relevant limitation we study is whether or not a haplotype mutation is detectable with a given amount of data. Specifically, in order to observe a mutation at an interior node, there must be an individual present in the population whose parents had haplotypes from the two different descendants of that node or from one descendant of that node and one non-descendant of that node. If there is a very large population with genotype data for many individuals, then it is likely that there will be sufficient information to determine the complete phylogenetic data for the haplotypes.

In order to identify the set of n haplotypes present and distinguish their phylogenetic relationships, we gather data from a set of m individuals. For there to be a chance of identifying and inferring the complete correct phylogeny for the haplotypes, of course, each haplotype must be present in some individual so thus necessarily $m \geq \frac{n}{2}$. If we consider the phylogenetic tree of haplotype ancestry, then we can regard each individual as being formed by a pair of haplotypes. The information about the phylogeny which may be inferred from that individual's data is only about nodes in the tree which lie along the path from the one haplotype to the other.

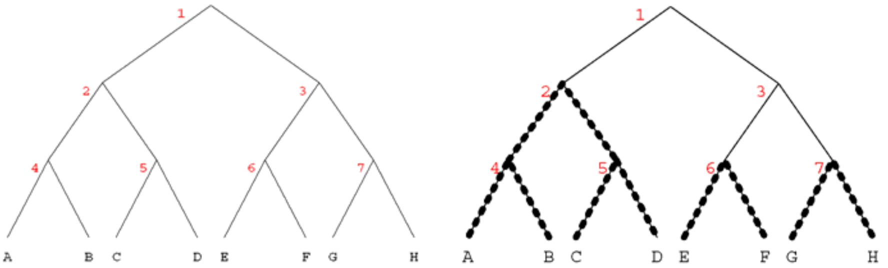

Figure 2. Example of relationship of 8 haplotypes $\{A, B, ..., H\}$. The dotted edges in the right tree represent the edges covered by the genotype data: (AC), (BD), (EF) and (GH).

For example, if the evolutionary relationship of haplotypes is given by the tree described as ((A, B), (C, D)), ((E, F), (G, H)) and pictured in Figure 2, and the genotype data collected for 4 individuals is (AC), (BD), (EF) and (GH), then there are 2 internal nodes numbered 1 and 3 which are not traversed by any path in the set of paths connecting the two haplotypes of each individual, so it would be impossible to deduce the evolutionary relationships represented by those internal nodes. In contrast, if we had 4 individuals with genotype data (A H), (B D), (C F) and (E G), then all edges (and thus all internal nodes) are traversed by the collection of paths and thus, in principle, it may be possible to deduce the complete evolutionary history of those 8 haplotypes using inference algorithms such as those described by Gusfield et al. below, we consider the question of how many individuals should be expected to be required to get complete path coverage of the tree, and also to get specified fractional coverage of the tree. The answer to this question depends upon the shape of the phylogenetic tree.

We note that if the set of haplotypes is known in advance, then it may be possible to deduce the complete phylogeny with less than full coverage of the tree. Also, even if the tree is completely covered by such paths, there may still be impossible to deduce the complete phylogenetic relationships of the haplotypes. However, full path coverage is a natural estimate for lower bounds of data sufficiency requirements for haplotype inference algorithms.

2.3 Tree distributions. We consider the combinatorial edge-covering properties of three families of trees under a number of assumptions about possible constructions of pairing of leaves. Natural models for generating trees to consider are the uniform random tree generation model, in which every possible tree on *n* leaves is equally likely, and also the Yule model for speciation (or birth-death trees) which constructs random trees based upon a sequence of bifurcations and results in a different random distribution of trees (see [1] for an overview of biologically-related tree distributions).

3. Exact Exhaustive Pairings

We first consider the simple case where each haplotype occurs in exactly one individual. We will necessarily have an even number of haplotypes since each individual has two haplotype constituents. We call the case where each haplotype occurs in exactly one individual an *exact exhaustive pairing*. Such pairings provide the absolute minimum possible information from which it might, in principle, be possible to deduce properties of the haplotype evolution. This contrived situation is not biologically realistic, but it is easy to observe and analyse behavior here that is characteristic of more realistic models considered later. We consider the question: given a population with $2n$ haplotypes and n individuals with each haplotype appearing in exactly one individual, what is the probability that there is sufficient coverage of the internal nodes? In more realistic settings where individuals are sampled randomly from a population, we expect to need even more individuals to give more complete coverage on average and we address those settings in later sections.

3.1 Balanced Tree Pairings. First, we consider perfectly balanced binary trees, and we begin by considering the probability that the root node is not traversed. That is, if we have a perfectly balanced tree with 2^l leaves, then the only way that the root node can be untraversed is if every leaf in the right subtree is paired with another leaf in the right subtree. We can see that this circumstance is unlikely for large trees below.

Proposition 3.1. *Given a balanced tree T with l levels having 2^l leaves, the fraction of all possible complete distinct pairings of leaves that do not cross the root is* $\frac{((2^{l-1}-1)!!)^2}{(2^l-1)!!}$.

Proof. The number of possible pairings of the 2^{l-1} leaves in the left subtree of the root is $(2^{l-1}-1)!!$, as is the number of possible pairings of the 2^{l-1} leaves in the right subtree of the root. The total number of pairings of all the leaves is $(2^l-1)!!$, so we get the fraction in the proposition. □

Proposition 3.2. *Given a balanced tree T with n levels having 2^l leaves, the fraction of all possible complete distinct pairings of leaves that do not cross an edge just above a particular node k levels down from the root is* $\frac{(2^{l-k-1}-1)!!(2^l-2^{l-k-1}-1)!!}{(2^l-1)!!}$.

Proof. Again we count all pairings, and we count the pairings which are contained totally within the descendants of the node and totally outside of the set of descendants of the node to get the desired fraction. □

We note that each of the above quantities goes to zero as the number of haplotypes increases, and these quantities go to zero particularly quickly for the root node and nodes close to the root node. Thus, even given a bare minimum of haplotype information, it is likely that information about mutations from early in the haplotype evolutionary history can be obtained algorithmically from the genotype data. For the lower levels, there are many more possible nodes to consider and the expected number of uncovered edges rises, although not dramatically as a fraction of the total nodes. For example, with a balanced tree of 18 levels, the expected value of uncovered edges at level 17 is 321 out of 131072 total.

3.2 The sibling pair obstruction. Experimental results show that in the distributions of balanced, uniform and Yule model trees, in the case of an exact exhaustive pairing which fails to cover the tree, the predominant means of failure is due to a sibling pair occurring as haplotypes in an individual, making it impossible to deduce where that sibling pair is attached to the remaining tree. We can estimate the likelihood of the failure to cover the tree as approximately the likelihood of there being a sibling pair matched together. As we saw in Section 3.1, the likelihood of being uncovered increases as we get further from the root, so the sibling pairs at the bottom of a balanced tree are the most likely to be uncovered.

For each sibling pair present in the tree, the probability that it is not paired in the matching is $1 - \frac{1}{n-1}$. If there are s sibling pairs, then we can calculate the probability that neither are paired in a matching as $1 - \left(\frac{2}{n-1} - \frac{1}{(n-1)(n-3)}\right)$. We can extend the case of s sibling pairs, which can be calculated exactly using the Inclusion-Exclusion Principle, and the probability is bounded below by the initial terms: $1 - \frac{s}{n-1}$ Work by Steel and McKenzie [14] computes the asymptotic number of sibling pairs (“cherries”) for the uniform distribution as $n/4$ and for the Yule distribution as $n/3$. Thus, we underestimate the probability that there is no sibling pair obstruction in a uniformly-selected random tree with an exact exhaustive matching as $1 - \frac{n}{4}\frac{1}{n-1} = 1 - \frac{1}{4}\frac{1}{1 - 1/n}\left(\frac{n-2}{n-1}\right)^{\frac{n}{4}}$, which converges to $3/4$ as n becomes large. Similarly, we have for the Yule distribution the probability of no sibling pair obstruction as at least $2/3$ as n becomes large. For the balanced tree case, where every leaf is part of a sibling pair and there are exactly $n/2$ such sibling pairs, we obtain the probability as n is large that there is no uncovered edge from the sibling pair obstruction is $1/2$. Thus, what appears to be the primary obstruction to incomplete coverage is increasingly present in the Yule, uniform and balanced tree families, but does not dominate the process and prevent many exact exhaustive pairings from having complete tree coverage.

The next most common failure for exact exhaustive pairings, from experimental experience and consistent with the analysis in the balanced case above, appears to be when two sets of adjacent sibling pairs are matched in such a way that the edge connecting them to the rest of the tree is uncovered. In that situation, though it may be possible to determine the evolutionary history of that group of four haplotypes, it will not be possible to determine where that group of four is connected to the remainder of the haplotype evolutionary tree. An example of this phenomenon is the uncovered node 1 from the earlier Figure 2 with the pairing which included individuals with haplotype pairs (AC) and (BD). Preliminary experimental data shows this behavior is significantly less probable than the simpler sibling pair obstruction phenomenon.

4. Sample Size Sufficiency

In general, the size of our sample of individuals from the population will be larger than the set of haplotypes, and we can ask the question of how many randomly selected individuals are needed to cover all of the edges (and thus all of the nodes) of the evolutionary tree. These individuals are selected at random, and it is likely that some haplotypes will appear repeatedly even before other haplotypes appear at all.

4.1 The sibling pair obstruction. We examined the preliminary experimental results in the distributions of uniform and Yule model trees, in the case of sets of pairings which cover the leaves but whose sets of paths fail to cover the tree. The predominant means of failure is due to a sibling pair of haplotypes occurring only as haplotypes in a single individual, making it impossible to deduce where that sibling pair of haplotypes is attached to the remaining tree.

4.2 Coupon collecting. A model for collecting haplotype data is to consider a population with n haplotypes and select individuals whose genotypes are given by two randomly selected haplotypes. If an individual has exactly the same haplotype for both parents, then this forms a homozygote – very useful for determining what the set of haplotypes is, but such a pairing can yield no information about the evolutionary history of the haplotype pool. To determine the evolutionary history of the set of haplotypes, we need heterozygotic individuals, which are formed from pairings of different haplotypes. We address the question – how many heterozygotic individuals are necessary for the paths of the pairings for those individuals to completely cover the tree? We use the simplifying assumption that each of the n haplotypes is equally likely to occur in the genotype of an individual in the population. Though this is not biologically realistic, the estimates obtained below can apply roughly to the non-uniform likelihood case by considering the uniform likelihood case with the number of haplotypes set at the reciprocal of the probability of the least likely haplotype in that distribution.

A lower bound to the number of individuals sampled (with replacement) needed to infer the haplotypes is given by the Coupon Collector Problem, also known as the "sequential occupancy problem." The Coupon Collector Problem is the following: suppose there is a large pool of coupons, equally distributed from n distinct types. We start with an empty collection and at each stage, a random coupon is drawn from the pool and added to the current collection. We would like to collect one of each type of coupon – typically, we will soon have many duplicates of those coupons already in our collection and will just keeping waiting for the last few types of coupons that we do not yet have. What is the waiting time to collect at least one of each coupon and thus have a complete set in our collection? The dominant term for the expected waiting time is given by $n \log n$; see Feller [7] for an excellent introduction to coupon collection methods. In the case where what is desired is a partial collection of coupons, given by a fraction p of the total number n of coupons, the dominant term in the expected waiting time for fractional coverage is given by $n \log \frac{1}{1-p}$.

4.3 Edge collection. To estimate the number of individuals required to get complete edge coverage in these tree models, we make several observations. We consider only *non-degenerate* leaf pairings – that is, heterozygotic individuals. A homozygotic individual whose parents possess the same haplotype (amounting to a leaf paired with itself) cannot yield any information about the haplotype ancestry, so we ignore those individuals and count only non-degenerate leaf pairings. Note, however, that such homozygotic individuals are very useful for identifying the set of haplotypes present in the population if it is not already known.

Theorem 4.1. *The expected number of non-degenerate individuals for complete coverage of a rooted binary evolutionary tree of a set of n haplotypes is at least $\frac{n}{2} \log n$ and bounded above by $\left(n - \frac{3}{2}\right) \log(2n-3)$.*

Proof. The lower bound for coverage comes from the coupon collection methods described above. In order to cover all of the pendant edges of the tree (those which lead to a leaf), there must be at least one pairing containing each individual. Thus, the standard coupon-collection lemma applies and we expect to be required to accumulate $n \log n$ leaves before having at least one of each pendant edge. Since there are two distinct haplotypes present in each nondegenerate individual, we obtain a lower bound of at least $\frac{n}{2} \log n$ individuals expected to infer the evolutionary history of n haplotypes.

The upper bound for coverage also comes from the coupon collection lemma. In this case, we consider the coupons to be the $2n-3$ edges of the tree. In this case, each non-degenerate pairing may give many edges in the path connecting the leaves of that pairing, but will always give at least 2 edges. The probability of an edge being crossed for a given pairing is at least $2p(1-p)$, where p is the fraction of the leaves closer to one end of the edge, and $1-p$ is the fraction of the leaves closer to the other end of the edge. That probability will vary according to the position of the edge in the tree and is highest when the edge approximately divides the tree into equal halves of leaves. That probability will always be at least that for the probability of a pendant edge being crossed, which is itself greater than $2/(2n-3)$. Thus, since the probability of being selected at each stage is greater than the $1/(2n-3)$ that would occur in the usual coupon collecting model with $2n-3$ coupons, the expected waiting time for a complete collection of coupons (edges) is no more than $(2n-3)\log(2n-3)$. Again, since there at least two edges selected for each individual, we expect the required number of individuals to be no more than $\left(n-\frac{3}{2}\right)\log(2n-3)$. □

Given the estimates used in the upper bound, this is clearly an overestimate, as borne out by the experimental data described below.

4.4 Fractional coverage. If instead of insisting on collecting enough data to determine the complete phylogenetic history, then we instead look only to collect enough data from individuals to determine a fixed fraction $p < 1$ of the evolutionary history, only a linear number of non-degenerate individuals are required.

Theorem 4.2. *The expected number of non-degenerate individuals to determine, from genotype data, a fixed fraction* $p < 1$ *of the rooted binary tree representing the evolutionary history of a set of n haplotypes is at least* $\frac{n}{2}\log\frac{1}{1-p}$ *and bounded above by* $n\log\frac{1}{1-p}$.

Proof. We use the coupon collecting lemma for fractional coverage. The expected waiting time for collecting pn distinct coupons has dominant term $n\log\frac{1}{1-p}$, so the lower bound comes from considering the expected waiting time to have fractional coverage of the n leaves, again halving the expected waiting time because the non-degenerate individuals will each give two leaves.

For an upper bound for the fractional coverage, we again consider the coupons to be the set of $2n-3$ edges of the tree, and we note that each individual given traverses at least two edges of the tree, giving the bound of $\frac{2n-3}{2}\log\frac{1}{1-p}$ for the expected waiting time for the desired fractional coverage of the set of edges, and thus the desired bound. □

These analyses do not depend upon the process to generate the shape of the tree. In cases where a particular process is used, for example in the Yule model for tree generation, we may be able to improve the analysis. For example, we note the mean path length for the path between two individuals selected at random in the Yule model, analyzed by Steel and McKenzie [20], is larger than 2 and grows slowly with the size of the tree and thus the upper bound can be improved in that case.

These estimates on sample size sufficiency are confirmed by the experimental data described in the next section.

5. Experimental Results

5.1 Experimental design. We generated synthetic datasets under two different tree distributions (uniform and Yule) and then randomly chose non-degenerate pairs of leaves and tallied the edges in the path induced by the pair. We continued to choose pairs until all edges had been crossed at least once. The pairs were chosen with equal probability of any distinct pair of leaves being selected. We also measured the number of pairs needed until all leaves were chosen and the height of the last edge crossed. We recorded the number of individuals at the earlier stages when we crossed thresholds of 70%, 80%, 90%, 95%, and 99% coverage of edges, as well as recording when the complete 100% coverage occurred.

5.1.1 Parameter space. We randomly generated model tree topologies from the uniform distribution on binary leaf-labelled trees, as well as the Yule distribution using Sanderson's r8s program [19]. The uniform trees were generated using tgen, a program written by Daniel Huson. We used the default settings for r8s under the YULE_C option to generate the trees from the Yule distribution. Due to the use of computational clusters, we were able to study a large number of tree sizes. We generated 100 random trees under each distribution for 100 different sizes of trees, ranging from ten haplotypes to 1000 haplotypes (increasing by increments of ten), for a total of 20,000 trees containing 10 million haplotypes.

5.1.2 Statistical considerations. Since the number of distinct rooted, leaf-labelled trees on n leaves is $(2n-2)!!$, it is not possible to take a fair sample of the entire input space. In order to obtain statistically robust results, we follow McGeoch [13] and Moret [15] and use a number of *runs*, each composed of a number of *trials* (a trial is the collecting of edges on a single tree), computed the mean outcome for each run, and studied the mean and standard deviation over the runs of these events. This approach is preferable to using the same total number of samples in a single run, because each of the runs is an independent pseudorandom stream. With this method, one can obtain estimates of the mean that are closely clustered around the true value, even if the pseudorandom generator is not perfect (see [13] for more details).

5.1.3 Methods. The experiments were performed on two computer clusters. In addition to the tree generation software described above, the authors' C programs and Perl scripts were used to calculate the paths generated and analyze the cumulative coverage. A set of 10 runs for 1000 haplotypes took about two hours of computation time on the slowest machines.

5.1.4 Measurements. Our focus was to determine the number of pairs of leaves (those heterozygotic individuals sampled in the population) needed to "witness" all of the edges (those mutations that form the haplotypes), confirming Gusfield's conjecture. In addition to measuring the number of pairs needed to cover all the edges, we also kept track of the number pairs until all the leaves were seen, as well as the height of the last edge seen. For each run of 10 trials, we retained only the mean values. Our results are composed of the means for each set of 10 runs.

5.2 Experimental results

5.2.1 Overview. For each distribution, we ran 10,000 trials consisting of trees ranging in size of 10 leaves to 1000 leaves. Due to the efficient generations of trees, we generated trees for each trial, instead of sampling subtrees of fewer large trees. We found similar results for the two tree distributions most commonly used in phylogeny. The number of samples, or steps, needed to cover all the edges approximated closely the expectation predicted by the coupon collecting lemma, $\frac{1}{2} n \log n$ for complete coverage and $\frac{1}{2} n \log \frac{1}{1-p}$ for fractional coverage (from Section 3, above). Interestingly, the samples needed to cover all the edges of a tree closely matched the numbers needed merely to cover all of the leaves. This correlation between the numbers needed to cover the edges and to cover the leaves is also seen when looking at the height of the last edge covered. The last edge covered was almost always an external edge (an edge connecting only a leaf to the rest of the tree). In both distributions, a fair fraction of the time was spent waiting for a final single pendant edge and its corresponding single leaf to be covered. This was seen in both distributions, with the very significant difference in the number of individuals required to ensure 99% coverage and 100% coverage indicating that a large fraction of the time was spent waiting for a few final edges and leaves to be covered.

5.2.2 Yule distribution. We present the results for the number of samples of pairs of distinct leaves needed to cover all the edges in trees under the Yule distribution. The left-hand graph in Figure 3 shows the number of samples from the population needed to see all the edges in the tree. The number is bounded by the theoretical results from Section 3 of $\frac{1}{2} n \log n$ and $\left(n - \frac{3}{2}\right) \log(2n-3)$ (shown by dotted lines in Figure 3). Interestingly, our results match the lower bound which is the expected number of samples, predicted by the coupon collecting lemma, needed to see all the leaves. This correlates well with the last edge covered almost always being a leaf, mentioned above. It is very rarely the case that we have all the leaves and are sampling just to cover remaining interior edges. Instead, it takes about the same time to cover the leaves as the edges. The experiments indicate that the upper bound could be sharpened.

The results for the number of samples needed to cover the leaves differed by less than 1 percent from the number needed to cover the edges and are omitted due to space constraints. As mentioned above, we ran 10 runs of 10 trials, averaged the results for each run, and then analyzed the averages. We present only the average of the averages. The standard deviations, for both the samples needed to cover the leaves and the edges were between 3 to 5 percent of the averages and were omitted to make the graph easier to read.

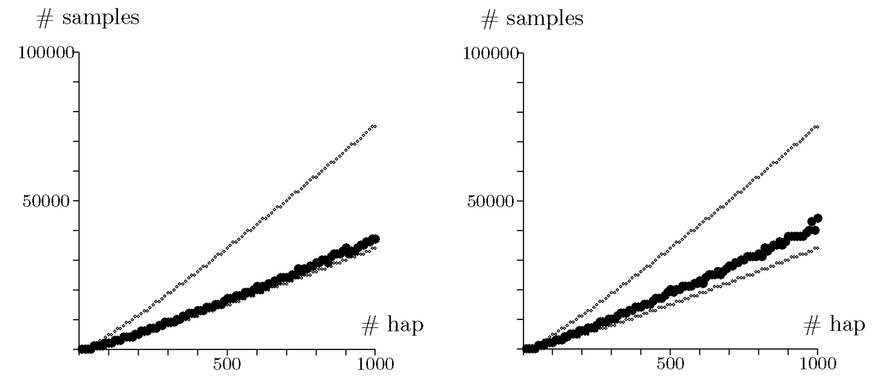

Figure 3. The graph on the left gives the number of samples from leaf pairs needed to cover all the edges of a tree chosen from the Yule distribution (dark line), as well as known lower and upper bounds of $\frac{1}{2} n \log n$ and $\left(n - \frac{3}{2}\right) \log(2n - 3)$ (dotted lines). On the right is the similar graph for the uniform distribution. The number of haplotypes ranges from 10 to 1000, in increments of 10. The results are the average of 10 runs of 10 trials. The standard deviations are 3 to 5 percent of the averages for the Yule distribution, 4 to 7 percent of the averages for the uniform distribution, and are omitted for readability.

5.2.3 Uniform distribution. We present the results for the number of samples of distinct pairs of leaves needed to cover all the edges in trees under the uniform distribution. The right-hand graph in Figure 3 shows the number of samples from the population needed to see all the edges in the tree. As in the Yule distribution, the number is bounded by $\frac{1}{2} n \log n$ and $\left(n - \frac{3}{2}\right) \log(2n - 3)$ (shown by dotted lines in Figure 3 and obtained in Section 3). Again, our results approximate the lower bound which is the expected number of samples, predicted by the coupon collecting lemma, needed to see all the leaves. It takes the same time to cover the leaves as the edges. The experiments indicate that there could be some improvement in the upper bound.

As in the Yule distribution case, the results for the number of samples needed to cover the leaves differed by less 1 percent from the number needed to cover the edges and are omitted due to space constraints. The standard deviations, for both the samples needed to cover the leaves and the edges were between 4 to 7 percent of the averages, slightly higher than was observed in the Yule distribution case.

Fractional Coverage

We present experimental results for the number of samples of distinct pairs of leaves needed to cover specified fractions of the edges in trees under the uniform random distribution of trees.

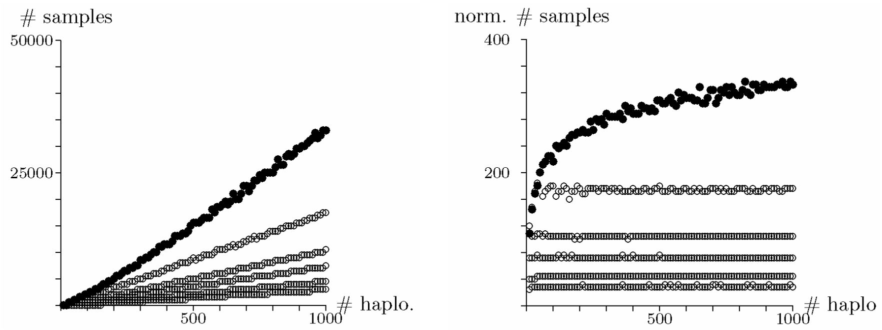

Figure 4. The left graph gives the number of samples from leaf pairs needed to cover 70%, 80%, 90%, 95%, 99% and 100% of the edges of a tree chosen from the Yule distribution. The right graph shows the number of samples versus the number of samples divided by *n*, the number of haplotypes. The number of haplotypes ranges from 10 to 1000.

In Figure 4, the left-hand graph shows the average number of individuals required to cover increasingly large fractions of the evolutionary history for an increasing large number of haplotypes, or mutations. As expected, the required population grows as the fraction and number of haplotypes grows. Furthermore, the right-hand graph shows the number of individuals required divided by the number of mutations.

Since the theoretical bounds established above in Theorems 4.1 and 4.2 give ranges on the order of $n \log \frac{1}{1-p}$ for $p < 1$ and $n \log n$ for the $p = 1$ case, we see clearly the constant linear coefficients emerging as linear coefficient terms for the fractional p and the logarithmically growing case from the complete coverage $p = 1$ case. We see similar behavior for the uniform process for random tree generation via similar graphs. In both of these processes, the standard deviations for the individuals required for a fraction $p < 1$ was small, with the standard deviations being less than 1% of the averages. For the $p = 1$ complete coverage case, there was much greater spread, with the standard deviation being about 5% of the average required population size.

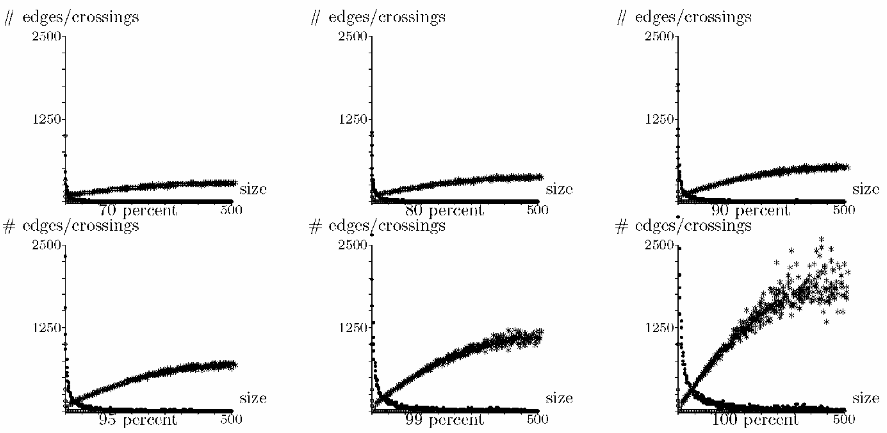

Figure 5. The distribution of edges covered, by size, for percent coverage (black dots) and the percentage normalized by the number of edges of that size (gray stars) under the Yule tree distribution. The tree size is fixed at 1000 leaves.

In Figure 5, the graphs show both the distribution of edge coverings and the total edge coverings, against the size of an edge, for the Yule distribution. Each edge divides the tree into two connected components; the size of an edge is the number of leaves on the smaller of those two connected components. In a tree with 1000 leaves, the largest possible edge size is obviously 500, which would usually occur near the root. The decreasing curve is the distribution of edges by size; we see, as expected, that due to the sheer number of small edges, we cross edges near the leaves of the tree more often than the larger edges close to the root. The increasing curve is the average number of edge crossings per edge (that is, the total number of crossings for each size divided by the total number of edges of that size), for edges of increasing size. We see from their shapes that the edges close to the leaves are crossed, on average, only a few times, but edges higher up and closer to the root are crossed hundreds to thousands of times. The wider spread in the number of crossings per edge for the complete coverage case of $p = 1$ corresponds to the wider spread in the required population size in that case and graphically shows the jump from linear to $n \log n$ waiting time when moving from a fixed fraction less than 1 to complete coverage, as seen in the theory. Similar behavior is seen for the uniform tree distribution, though not pictured here due to space constraints.

6. Future Work

When we chose pairs of leaves, we made the simplifying assumption that all haplotypes occur uniformly across the population. One line of future study is to look at more biologically realistic distributions of haplotypes and the effect on the number of individuals that need to be sampled to see various fractions of the mutations. An appropriate model to consider for distribution of haplotypes is the Fisher-Wright model [5, 12]. We also assume that the distribution of haplotypes does not change as a function of time and we would like to incorporate possible changes in haplotype distribution with time into our simulations. We also would like to get sharper theoretical bounds on the required steps by applying the work of McKenzie and Steel [20] which gives average path length between two randomly selected leaves in a tree generated by the Yule process.

Acknowledgements

The authors would like to thank Dan Gusfield and Joel Spencer for very helpful conversations, the Computational Research Facilities at the Graduate Center of the City University of New York and at the Center for Algorithms and Interactive Scientific Software (CAISS) at the City

College of New York for the use of their Linux computer clusters, and the Department of Mathematics at UC Davis and the Centre de Recerca Matemàtica in Barcelona for hosting them during some of this work. SC was supported in part by NSF-grant #02-15942 and PSC-CUNY grant #63438-0032, and KS was supported in part by NSF grants #00-96080/99-73874, #01-21651 and #02-15942.

References

[1] David Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, Statistical Science 16 (2001), (revision of "Visualizing Phylogenetic Tree Balance").

[2] V. Banfa, D. Gusfield, G. Lancia and S. Yooseph, Haplotyping as perfect phylogeny: A direct approach, Journal of Computational Biology 10(3) (2003), 323-340.

[3] A. G. Clark, Inference of haplotypes from pcr-amplified samples of diploid populations, Mol. Biol. Evol. 7 (1990), 111-122.

[4] The International HapMap Consortium, A haplotype map of the human genome, Nature 437 (2005), 1299-1320.

[5] W. J. Ewens, The sampling theory of selectively neutral alleles, Theoretical Population Biology 3 (1972), 87-112.

[6] L. Excoffier and M. Slatkin, Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, Mol. Biol. Evol. 12 (1995), 921-927.

[7] William Feller, An Introduction to Probability Theory and its Applications, Vol. I, John Wiley & Sons Inc., New York, 1968.

[8] D. B. Goldstein and M. E. Weale, Population genomics: Linkage disequilibrium holds the key, Curr. Biol. 11 (2001), R576-R579.

[9] Dan Gusfield, Haplotying as perfect phylogeny: Conceptual framework and efficient solutions, Proceedings of the Sixth Annual International Conference on Computational Biology, ACM, 2002, pp. 166-175.

[10] Dan Gusfield, An overview of combinatorial methods for haplotype inference, Computational Methods for SNPs and Haplotype Inference, Volume 2983 of Lecture Notes in Computer Science, Springer, 2004, pp. 9-25.

[11] Dan Gusfield and R. H. Chung, Empirical exploration of perfect phylogeny haplotyping and haplotypers, Proceedings of the 2003 Cocoon Conference, 2003.

[12] B. F. J. Manly, The Statistics of Natural Selection, Chapman and Hall, 1985.

[13] C. C. McGeoch, Analyzing algorithms by simulation: variance reduction techniques and simulation speedups, ACM Comp. Surveys 24 (1992), 195-212.

[14] Andy McKenzie and Mike Steel, Distributions of cherries for two models of trees, Math. Biosci. 164(1) (2000), 81-92.

[15] B. M. E. Moret, Towards a discipline of experimental algorithmics, In Data Structures, Near Neighbor Searches and Methodology: Fifth and Sixth DIMACS Implementation Challenges, DIMACS, 2002.

[16] National Institutes of Health Haplotype Map Project, Developing a haplotype map of the human genome for finding genes related to health and disease, 2001. Available at http://www.genome.gov/page.cfm?pageID=10001665.

[17] National Institutes of Health Haplotype Map Project, The haplotype map project, 2001. Available at http://www.genome.gov/page.cfm?pageID=10001688.

[18] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward and E. S. Lander, Linkage disequilibrium in the human genome, Nature 411 (2001), 199-204.

[19] Michael J. Sanderson, r8s version 1.06(beta): Analysis of rates (r8s) of evolution, 2002. Software available at http://ginger.ucdavis.edu/r8s/.

[20] Mike Steel and Andy McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models, Math. Biosci. 170(1) (2001), 91-112.

[21] Matthew Stephens, Nicholas J. Smith and Peter Donnelly, A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (2001), 978-989.