Motivation: Inference of haplotypes from genotype data is crucial and challenging for many vitally important studies. The first, and most critical step, is the ascertainment of a biologically sound model to be optimized. Many models that have been proposed rely partially or entirely on reducing the number of unique haplotypes in the solution.

Results: This article examines the parsimony of haplotypes using known haplotypes as well as genotypes from the HapMap project. Our study reveals that there are relatively few unique haplotypes, but not always the least possible, for the datasets with known solutions. Furthermore, we show that there are frequently very large numbers of parsimonious solutions, and the number increases exponentially with increasing cardinality. Moreover, these solutions are quite varied, most of which are not consistent with the true solutions. These results quantify the limitations of the Pure Parsimony model and demonstrate the imperative need to consider additional properties for haplotype inference models. At a higher level, and with broad applicability, this article illustrates the power of combinatorial methods to tease out imperfections in a given biological model.

Contact: weixiong.zhang@wustl.edu

1 INTRODUCTION

Diploid species, such as human, have pairs of chromosomes yielding 2 nt at most sites along the genome. For instance, there may be a ‘C’ and a ‘G’ at a given site for an individual. A sequencer will reveal the pair of nucleotides at the site. However, it is not known which chromosome in the chromosome pair has the ‘C’ and which has the ‘G’. If a pair of nucleotides at a site are different, or heterozygous, then methods are needed to infer which nucleotide state belongs to each chromosome. There are methods for solving this problem directly using laboratory procedures (Andrés et al., 2007; Orzack et al., 2003). However, they are only practical for small studies. Consequently, computational methods are normally used to resolve this problem, known as the haplotype inference problem.

A haplotype is an ordered set of single nucleotide polymorphisms (SNPs) in physically close proximity along a single chromosome, while a genotype is the meld of a pair of haplotypes. This broad definition permits discussions about sets of SNPs without a priori knowledge about linkage disequilibrium, recombination rates and/or frequencies in the population. For most SNPs, there exist only two different nucleotide states in a given population. For instance, ‘C’ and ‘G’ may be the only nucleotides at a given site for all of the individuals in the population. These SNPs are referred to as biallelic.

Most haplotype inference methods assume that the SNPs in the input are biallelic. With this assumption it follows that a genotype with \( k \) heterozygous sites could be resolved in \( 2^k - 1 \) different ways.

The haplotype inference problem can be stated as follows: given a set of \( n \) genotypes from a population of \( n \) individuals, find the set of unique haplotypes that exist in the sample set and, for each genotype, find the pairs of haplotypes that might exist in the given individual, along with their respective probabilities. This problem is also called phasing the genotypes.

What is the cardinality of the unique haplotypes for a sample set drawn from a population? Many factors contribute to this value. Recombinations, mutations and gene flow tend to increase the cardinality. On the other hand, natural selection and genetic drift tend to reduce it. Some regions of the genome are highly conserved while others, such as recombination ‘hot spots’ (Myers et al., 2005), may be highly unconserved. At the population level, the genotypes may have been drawn from a population that has evolved in high isolation or one that has been a historical ‘melting pot’. The interplay of these genetic forces tolerate a range of possible cardinality values. However, when the haplotype is composed of relatively closely spaced sites, linkage disequilibrium is higher and recombination is less likely to occur within the haplotype, with all other factors being equal. For these reasons, it is a common assumption that the number of unique haplotypes within a population is relatively small.

The first widely adopted haplotype inference method was Clark’s Subtraction Method (Clark, 1990). This method is a greedy approximation method that aims to minimize the number of unique haplotypes in the solution. Due to its greedy nature, Clark’s Subtraction Method is dependent on the ordering of the genotypes in the input. The number of genotypes that are resolved can vary for different orderings. There are \( n! \) possible orderings of \( n \) genotypes. Consequently, 60 genotypes can be ordered in \( 8 \times 10^{31} \) different ways. (Note that it has been estimated that there are approximately \( 10^{80} \) subatomic particles in the entire known universe!) For just 20 genotypes there are \( 2 \times 10^{18} \) possible orderings. Since there is an immense number of possibilities, only a very small sample of genotype orderings can be tried. Clark recommended using the ordering that resolved the most genotypes. In 2000, Gusfield defined the Maximum Resolution Problem, which seeks to identify an optimal ordering of genotypes such that Clark’s Subtraction Method resolves the maximum number of genotypes possible (Gusfield, 2000). Gusfield cast this problem as an integer linear program (IP). A basic premise of both of these models is that the number of unique
haplotypes that are used to resolve the genotypes is relatively small, far less than the upper bound of $2^n$.

The Pure Parsimony Haplotype Inference (PPHI) model directly addresses this premise. Given a set of genotypes, PPHI is to find a set of haplotypes that resolves all of the genotypes using the fewest number of unique haplotypes possible (Gusfield, 2003). Once the set of haplotypes are found, each genotype is resolved using a pair from this set. Sometimes multiple pairs of haplotypes can resolve a given genotype. When this situation occurs, the PPHI model does not address which pair should be selected and various methods have been used to make this selection. This article is focused on the PPHI model which is the problem of finding the set of haplotypes with minimum cardinality. Consequently, we do not address methods for the subsequent assignment of pairs to individual genotypes.

Earl Hubbell and Dan Gusfield independently derived the Pure Parsimony model after contemplating Clark’s Subtraction Method (personal communication). Gusfield (2003) published an integer program (IP) formulation and presented a reduction technique, referred to as RTIP, that reduces the problem size while preserving optimality. Since Gusfield’s seminal paper on this problem was published in 2003, there has been a great abundance of research on PPHI. It has been referred to by other names, such as the maximum parsimony problem (e.g. Wang and Xu 2003) and the optimal haplotype inference problem (e.g. Huang et al. 2005). This section provides a brief overview of some of this work, and is by no means comprehensive. The solution to the IP formulation of PPHI guarantees a globally optimal solution. However, solving this problem is NP-hard (Huang et al., 2005). Due to its complexity, there has been a great deal of research devoted to making the problem more tractable and analyzing the difficulty of the problem. This research generally falls into three different categories: theoretical advances, computational methods and approximation techniques.

Theoretical advances: the original IP formulation for PPHI yielded problem sizes that were exponential in the number of genotypes (Gusfield, 2003). Several new IP formulations have been derived in which the problem size is polynomial in the size of the input (Brown and Harrower, 2004; Halldorsson et al., 2003; Lancia et al., 2004). These new formulations made it possible for large PPHI instances to fit into memory. However, the reduction in problem size frequently resulted with increased computation time (Gusfield and Orzack, 2005). Brown and Harrower (2004) developed a hybrid formulation to combine their polynomial-sized IP with Gusfield’s original IP (Gusfield, 2003), improving their computation times while preserving minimal memory requirements (Brown and Harrower, 2006a). PPHI has also been cast as an integer quadratic program (Huang et al., 2005). Using a relaxation of this formulation, Huang et al. developed a matlab code called SDPHapInfer for finding approximate solutions. In addition to modifications of the mathematical formulation, there have been a number of additional theoretical advances, most of which have focused on the difficulty of the PPHI problem. Hubbell was the first to prove that PPHI is NP-hard by a reduction from max-clique (personal communication). Huang et al. (2005) later published a proof of this complexity. Lancia et al. (2004) proved that PPHI is APX-hard. Then Lancia and Rizzi (2006) proved that this problem can be solved in polynomial time if all genotypes have at most two heterozygous sites. Cilibrasi et al. (2005) derived a polynomial time algorithm for this special case. Sharan et al. (2006) identified ‘islands of tractability’ for PPHI based on the structure of haplotypes that are common to the genotypes. Finally, Brown et al. (2006b) examined the PPHI problem and observed a connection between the algebraic rank of an instance and the difficulty in solving it.

Computational methods: research on computational techniques for PPHI can be divided into two categories: studies in which completely different search strategies have been employed, and studies in which computation times have been decreased by making improvements to a given search strategy. Gusfield’s seminal paper on PPHI used a generic IP solver to compute the optimal solution (Gusfield, 2003). In addition to generic IP solvers, a variety of search strategies have been explored for this challenging problem. Wang and Xu (2003) implemented a branch-and-bound search, referred to as HAPAR, for solving PPHI instances. Blain et al. (2005) developed a graph theory formulation for PPHI, using bipartite graphs with labeled nodes. Lync and Marques-Silva (2006) recast PPHI as a Boolean satisfiability problem (SAT), allowing the use of highly refined SAT solvers for tackling PPHI. Finally, Graca et al. (2007) have cast PPHI as a ‘Pseudo-Boolean satisfiability’ problem. Several methods have been proposed to reduce the computation time for previously used search strategies. The first improvement was presented in Gusfield’s (2003) paper that introduced the PPHI model. He showed that any pair of haplotypes that could only resolve a single genotype could be omitted from the problem formulation without loss of optimality. This reduction method is referred to as RTIP. Wang and Xu (2003) introduced the equal column technique in which the number of sites in the genotypes are reduced when there are identical sets. Optimality is preserved by both of the methods for finding a single solution. However, in cases where there are multiple optimal solutions some of these solutions may be lost due to this reduction, so care must be taken when computing all optimal solutions. Recently, Di Gaspero and Roli (2008) implemented a reduction procedure using structural properties of compatibility graphs for use in an approximation implementation which does not necessarily preserve optimality even for a single solution. Branch-and-cut searches use cutting planes to tighten the relaxation that is solved at each node. Customized cutting planes for PPHI have been presented in at least two papers (Brown and Harrower, 2004, 2006a). Furthermore, tight upper and lower bounds are used in both branch-and-bound and branch-and-cut search. Tighter bounds frequently yield faster searches. Marques-Silva et al. (2006a) improved on upper bounds for PPHI, and Brown et al. (2006b) improved on lower bounds. Finally, in another study by Marques-Silva et al. (2006b), lower bounds were found for the SAT formulation of PPHI.

Approximation techniques: due to its computational difficulty, a number of approximation methods have been developed for the PPHI problem. Lancia et al. (2004) developed a $2^{k-1}$-approximation, where $k$ is the maximum number of heterozygous sites in any genotype. They also designed approximations not based on mathematical programming, which have nearly linear computation times. In another study, Lancia and Rizzi (2006) introduced combinatorial approximations. Wang et al. (2005) used a genetic algorithm to approximately solve the PPHI problem. Huang et al. (2005) developed an iterative semi-definite programming approximation method for PPHI. Li et al. (2005) designed the Parsimonious Tree Grow method that greedily resolves genotypes in a column-by-column manner. Wang and Yang (2006) approximated PPHI by using a greedy heuristic based on Gusfield’s model. Finally,
Di Gaspero and Roli (2008) implemented a method using local search metaheuristics. Our work in this article was motivated by several questions. Given a set of genotypes, what is the number of haplotypes that truly produced them? PPHI yields a lower bound on this number. The upper bound is \( 2^n \) for \( n \) genotypes. This yields a broad range of possible values. However, it is believed that the true value is close to the lower bound. Verifying this assumption is difficult as true haplotypes that have been experimentally derived are rare. We have found only seven sets of true haplotype data, i.e. not computationally inferred, but experimentally derived haplotypes, that can be used for this verification. Three of these seven datasets do not have parsimonious solutions. The least parsimonious is the solution for the largest dataset, having 32 true distinct haplotypes, when it can be resolved with only 28 haplotypes.

This article presents and justifies the following four propositions: (i) There frequently are a large number of optimal PPHI solutions. (ii) There tends to be a great deal of variation among these solutions. (iii) There is a good possibility that the true solution is not among these solutions. (iv) When considering solutions with cardinality of \( H+i \), where \( H \) is the cardinality of the PPHI solution, the number of solutions increases exponentially as \( i \) increases.

In the next section, the materials and methods used for this study are summarized. We have used genotype and haplotype data to evaluate the number and variance of PPHI solutions, as well as their accuracy. Our computational methods are also briefly summarized. Evidence supporting the four propositions is presented in the Section 3. This article is concluded with a brief discussion.

2 MATERIALS AND METHODS

2.1 Datasets

We have used two different types of data. The first type is composed of seven datasets for which the known haplotypes have been experimentally identified (Andrés et al., 2007; Orzack et al., 2003). The second type of data is drawn from the HapMap project (The International HapMap Consortium, 2003, 2005). While the true haplotypes for the second type of data are unknown, these data are useful for observing the structural characteristics of PPHI solutions for human genotypes.

2.1.1 Known haplotype data

We have tested the accuracy of PPHI using seven sets of known haplotype data (‘Known’ data). These data were derived experimentally (i.e. the individual haplotypes were identified, not the melded pairs). The data, whose characteristics are summarized in Table 1, came from two sources.

Table 1. Known haplotype datasets used for comparisons, including label for the dataset, nucleotide range in reference sequence, number of genotypes (\( n \)), number of sites in each genotype (\( m \)), number of ambiguous genotypes (genotypes that have at least two heterozygous sites, \# Amb.), percentage of all heterozygous sites (% Het.) and recombination rate (Rec.).

<table>
<thead>
<tr>
<th>Data</th>
<th>Nucleotide range</th>
<th>( n )</th>
<th>( m )</th>
<th># Amb.</th>
<th>% Het.</th>
<th>Rec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>17874–21 388</td>
<td>80</td>
<td>9</td>
<td>47</td>
<td>21.4</td>
<td>None</td>
</tr>
<tr>
<td>B</td>
<td>667–1464</td>
<td>39</td>
<td>5</td>
<td>13</td>
<td>22.6</td>
<td>None</td>
</tr>
<tr>
<td>C</td>
<td>32 107–34 389</td>
<td>39</td>
<td>17</td>
<td>27</td>
<td>20.5</td>
<td>None</td>
</tr>
<tr>
<td>D</td>
<td>12 867–13 729</td>
<td>39</td>
<td>8</td>
<td>15</td>
<td>16.7</td>
<td>Low</td>
</tr>
<tr>
<td>E</td>
<td>32 107–35 800</td>
<td>39</td>
<td>26</td>
<td>28</td>
<td>19.0</td>
<td>Low</td>
</tr>
<tr>
<td>F</td>
<td>8043–9256</td>
<td>39</td>
<td>22</td>
<td>26</td>
<td>21.7</td>
<td>High</td>
</tr>
<tr>
<td>G</td>
<td>33 117–38 365</td>
<td>39</td>
<td>47</td>
<td>33</td>
<td>14.0</td>
<td>High</td>
</tr>
</tbody>
</table>

The first source of the ‘Known’ data is a set of 80 human ApoE haplotype pairs, each with nine SNPs, that was experimentally found by Orzack et al. (2003). These SNPs are drawn from the apolipoprotein E locus. The individuals were unrelated and 18 were classified as Asians, 19 as Blacks and 43 as Caucasians. Templeton et al. (2005) found that there is no statistically significant recombination in this region. Dataset A in Table 1 is composed of these 80 pairs of haplotypes. The second source of data was experimentally collected by Andrés et al. (2007). It contains 39 pairs of human haplotypes, each with 411 sites, in a 48 kb region containing the KLK13 and KLK14 genes. There is a substantial amount of missing data in this set. Pure Parsimony requires complete genotype data. Six regions of complete data from this set are used for this study and correspond to datasets B through G in Table 1. They range from 5 sites to 47 sites in length. Recombination rates for this data have been found by T.Maxwell et al. (2008, personal communication). The 17 sites of set C have no recombination and are combined with nine additional sites, which have a low recombination rate, to make set E. All of the input files that were used for our tests are available online as supporting files.

2.1.2 HapMap data

The International HapMap Consortium was founded in 2002 to map genome-wide variation for four human populations: the Yoruba people of Ibadan, Nigeria (YRI), US residents with northern and western European ancestry (CEU), individuals from the Tokyo area (JPT) and from the Beijing area (CHB) (The International HapMap Consortium, 2003, 2005). The YRI and CEU datasets are each composed of 30 trios, where a trio consists of two parents and a child. The JPT and CHB datasets are each composed of 45 unrelated individuals. The HapMap Consortium provides the JPT and CHB data separately and also combined together as a single Asian population of 90 unrelated individuals (JPT+CHB). We have used JPT+CHB for our tests. We have also omitted the children from the YRI and CEU trio data. Using the assumption of random mating, the remaining individuals can be considered unrelated.

Phase I of the HapMap project has accurately identified genotypes representing approximately one SNP per 5 kb for all 269 individuals. We have randomly selected genotypes from this data for our tests. For each of 22 chromosomes, and for each of the three populations, we have randomly selected a starting point for \( m \) adjacent SNPs for \( n \) individuals, yielding 66 sets of data. We set \( n=m=20 \) and \( n=m=25 \), resulting with a total of 132 HapMap datasets. The PPHI model ignores duplicate genotypes, as they have no impact on the results. We have discarded duplicate genotypes when assembling our test data, so the \( n \) individuals in a set all have unique genotypes.

2.2 Computational methods

Finding a single PPHI solution is computationally demanding as this problem is \#P-complete (Huang et al., 2005). Finding all optimal solutions is \#P-complete (Garey and Johnson, 1979), which is the main computational task for the current study. Furthermore, reduction techniques, such as RTIP (Gusfield, 2003) and equal column elimination (Wang and Xu, 2003), reduce the problem size and ensure optimality when finding a single optimal solution. However, these reductions can eliminate alternate optimal solutions. To our knowledge, no method has been previously presented in which all optimal PPHI solutions are identified. We have identified several techniques for addressing this demanding problem and have compared performance of these methods in a companion study, which is currently under review. This section briefly overviews three of these methods: RTIP, equal column elimination and backbone identification.

In his seminal paper on PPHI, Gusfield (2003) presented the RTIP formulation for reducing problem sizes. This reduction is achieved by omitting certain pairs of candidate haplotypes. If a haplotype pair could resolve a genotype, but neither of the haplotypes could be used to partially resolve another genotype, then it can be omitted with no risk when finding a single solution. If the genotype has no candidate haplotypes due to this removal, the genotype can be removed from the problem and arbitrarily
assigned a pair of resolving haplotypes after the PPHI solution is found for the remaining genotypes. Otherwise, if the genotype still has remaining pairs of candidate haplotypes, the resultant solution has the fewest possible number of haplotypes as the omitted pair cannot help reduce this number.

Equal column elimination reduces the PPHI problem by omitting particular sites for all genotypes (Wang and Xu, 2003). Consider a matrix where each row represents a genotype and each column represents a site along the genotypes. If two columns are identical, one can be removed prior to solving the PPHI problem. After the problem is solved, the column can be reinserted using the same solution that was found for the identical column.

Both RTIP and equal column elimination are valid when finding a single PPHI solution. However, either method can reduce the number of optimal solutions. When these techniques are used to speed up computation, post-processing must be performed to capture the optimal solutions lost in the reductions. Sometimes the time required for post-processing outweighs the benefits of the reductions. Details about the post-processing techniques are described in our companion study, which is currently under review.

Even when there are thousands of optimal solutions, there are frequently some haplotypes that appear in every solution. We refer to these common haplotypes as backbones (Schneider et al., 1996). Some backbone haplotypes are explicit as they resolve unambiguous genotypes that have not more than one heterozygous site. These genotypes have only one possible pair of haplotypes that can resolve them. Explicit backbone haplotypes can be easily identified by scanning through the set of genotypes. The initial pool for Clark’s Subtraction Method is entirely composed of all explicit haplotypes.

The other backbone haplotypes are implicit as it is not immediately apparent that they must appear in all optimal solutions. Implicit backbone haplotypes can be identified after finding all optimal solutions. However, it is beneficial to identify all backbone haplotypes prior to enumerating all optimal solutions. A simple approach can be utilized to make these identifications in which at most \( H + 1 \) optimal solutions are computed, where \( H \) is the number of unique haplotypes in the PPHI solution. First a single optimal solution is found, identifying \( H \) haplotypes. These are the only candidate backbone haplotypes. Each of these haplotypes are removed from the problem one at a time and the problem is resolved. If the number of haplotypes in the solution is greater than \( H \), the removed haplotype is a backbone. Otherwise, it is not. In the latter case, some of the candidate backbone haplotypes that have not been tested yet can be disqualified if they do not appear in the current PPHI solution. Thus, the total number of problems solved is not more than \( H + 1 \). After finding the backbone haplotypes the computation time for finding all PPHI solutions is reduced by incorporating the backbone information in the formulation.

We experimented with many combinations of RTIP, equal column elimination and backbone identification using both an IP formulation and branch-and-bound search. We also experimented with novel hybrid formulations. While no method was the consistent winner for all of our trials, using equal column elimination with backbone identification for an IP formulation was the fastest overall. We used these methods for this study.

Backbones are of interest for several reasons. First, they can be found in a preprocessing step and used to speed up the computation time for finding all optimal solutions, as we have done for computing all optimal solutions in our experiments. They are also of interest as they have a correlation with the number of optimal solutions and the quality of these solutions. Indeed, when there is only one optimal solution, all of the haplotypes in the solution are backbones. In Section 3, we demonstrate the correlation between small numbers of backbone haplotypes and large numbers of optimal solutions. Finally, backbones are of interest due to the fact that they represent the part of the solutions that are common across all solutions, providing some insights into problem structures.

### 3 RESULTS

In this section, evidence supporting our four propositions is presented. First, we use all of our datasets to demonstrate the large numbers of PPHI solutions and their variations. Then we use known haplotypes that were experimentally derived to assess the accuracy of PPHI solutions. Finally, we present justification that the number of solutions increase exponentially as the cardinality of haplotypes is increased.

#### 3.1 Number of optimal solutions

The number of optimal solutions and how this number varies with implicit backbone size, number of individuals and number of sites are explored in this section.

Table 2 shows the properties and results for the ‘Known’ and HapMap datasets. The average number of optimal solutions is very large, ranging from 163 to 7968. For five of the datasets, these numbers are lower bounds on the true number as 14 of the instances required too much time to compute all optimal solutions.

As seen in the table, roughly half of the backbones are explicit backbones due to unambiguous genotypes. Only 9 of the 139 datasets had a single optimal solution. Therefore, all of the haplotypes in each solution were backbones for each of these nine instances.

The number of optimal solutions is dependent on the number of individuals and the number of sites in a dataset. First, we experimented by starting with only two individuals from a dataset and adding individuals one at a time. Let \( H_i \) equal the cardinality of the PPHI solution when there are \( i \) individuals in the dataset. If an additional individual is added to the set and \( H_{i+1} = H_i \), then the number of optimal solutions cannot increase. Furthermore, when the cardinality does increase, it can only increase by one or two

<table>
<thead>
<tr>
<th>Data</th>
<th>No. of sets</th>
<th>Ambiguous</th>
<th>Heterozyg (%)</th>
<th>Expl. BB</th>
<th>Total BB</th>
<th>H</th>
<th># of sols.</th>
<th>One Sol.</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Known’ Sets</td>
<td>7</td>
<td>27.0</td>
<td>19.4</td>
<td>5.7</td>
<td>10.0</td>
<td>14.6</td>
<td>≥1559</td>
<td>2</td>
</tr>
<tr>
<td>JPT+CHB 20 × 20</td>
<td>22</td>
<td>14.9</td>
<td>22.1</td>
<td>5.3</td>
<td>9.3</td>
<td>13.1</td>
<td>163</td>
<td>2</td>
</tr>
<tr>
<td>CEU 20 × 20</td>
<td>22</td>
<td>15.4</td>
<td>24.8</td>
<td>4.8</td>
<td>8.8</td>
<td>14.0</td>
<td>4394</td>
<td>0</td>
</tr>
<tr>
<td>YRI 20 × 20</td>
<td>22</td>
<td>15.9</td>
<td>22.9</td>
<td>4.5</td>
<td>9.3</td>
<td>13.7</td>
<td>1552</td>
<td>3</td>
</tr>
<tr>
<td>JPT+CHB 25 × 25</td>
<td>22</td>
<td>20.6</td>
<td>24.9</td>
<td>4.8</td>
<td>10.3</td>
<td>16.5</td>
<td>3076</td>
<td>0</td>
</tr>
<tr>
<td>CEU 25 × 25</td>
<td>22</td>
<td>20.7</td>
<td>23.5</td>
<td>4.8</td>
<td>11.0</td>
<td>16.8</td>
<td>≥959</td>
<td>1</td>
</tr>
<tr>
<td>YRI 25 × 25</td>
<td>22</td>
<td>21.1</td>
<td>22.9</td>
<td>4.5</td>
<td>10.0</td>
<td>17.9</td>
<td>≥7968</td>
<td>1</td>
</tr>
</tbody>
</table>

For five of the datasets, the number of optimal solutions shown are lower bounds as it required too much time to compute all optimal solutions.
haplotypes. However, the number of optimal solutions can increase dramatically. Figure 1a shows this spiky behavior for ‘Known’ dataset F. The vertical lines show where the cardinality of the PPHI solution increased. As expected, increases in the number of optimal solutions only occur when the cardinality increases. Note that when there are 19 individuals, the cardinality had stayed at the same value that it was for 17 individuals. This problem had become very constrained and there is a dramatic drop in the number of optimal solutions. Despite occasional drops, we saw a general trend of increasing number of optimal solutions for increasing numbers of individuals.

Figure 1b shows the results for ‘Known’ dataset F when starting with only two sites, but all individuals, and increasing the number of sites one at a time. We see the same pattern as when the number of individuals are increased. In this case, we occasionally observed extreme jumps. The most dramatic thing that we saw in the tests we ran was for the 19th CEU 20 × 20 HapMap dataset in which the number of solutions went from 60 to 22,283 when the number of sites increased from 13 to 14 (with all 20 sites there were 7,294). As with increasing the number of individuals, we again observed some dips in the number of solutions when sites were added, but overall, there was a general trend toward increasing numbers of optimal solutions with increasing numbers of sites.

These findings suggest that scaling up problem sizes, either in the number of individuals or the number of sites, will likely result with even greater numbers of optimal solutions. Unfortunately, the computational difficulty of this problem prevents us from verifying this intuitive perception on larger problem instances.

In our research, we have observed relationships between PPHI solutions and implicit backbone size. For each dataset, we computed the number of implicit backbone haplotypes divided by the total number of haplotypes to derive the *implicit backbone size*, which can take on values from 0 to 1. We grouped datasets into five different implicit backbone size categories. Figure 2a shows the number of datasets in each category. Figure 2b shows a lower bound on the average number of optimal PPHI solutions for each implicit backbone size category. As we intuitively expected, there is a clear correlation between small implicit backbone size and large numbers of optimal solutions.

To compare the variation of solutions, we used two different measures. First, we found the non-intersection of the solutions as follows. For each pair A and B of optimal solutions, let a be the number of haplotypes in A but not in B, and b the number of haplotypes in B but not in A. The non-intersection of A and B is equal to \(
\max(a, b)/H
\), where H is the cardinality of the PPHI solution. If none of the haplotypes match, this value would be 1.

The second measure computes a distance between each haplotype in solution A and each haplotype in solution B. This distance is set equal to the number of sites that are different for the two haplotypes, divided by the number of sites \(m\). If two haplotypes are identical, the distance is 0, and if they are complete opposites, the distance is 1. We computed the way to match each haplotype in set A to each in set B such that the total distance between the matches is minimized. This problem is the same as the minimum bipartite matching problem, which is also known as the Assignment Problem (Martello and Toth, 1987).

For both variation measurements, the average value is computed for all optimal solutions when there are 1000 or less optimal solutions. When there are more solutions, 1000 solutions are randomly selected, due to the high-computational demand of these measurements. Figures 2c and 2d show the variation amongst the optimal PPHI solutions for each of the implicit backbone size categories. The average variation amongst solutions tends to be greater for smaller implicit backbone size.

### 3.2 Accuracy

The true haplotype pairs are known for the seven ‘Known’ datasets A through G, offering the opportunity to assess the accuracy of PPHI, as shown in Table 3. While the number of unique haplotypes is relatively small, only four of the seven datasets have the very fewest possible number of haplotypes. Therefore, no matter what PPHI solution is chosen for the remaining three datasets, they are all incorrect.

Table 3 also shows the average non-intersection and average distance over all optimal PPHI solutions compared with the true solution as well as the number of PPHI backbones that are not in the true solution. For cases when the cardinality of the true solution is greater than the cardinality of the PPHI solution, the *Generalized Assignment Problem* is used for computing the average distance among PPHI solutions. In this case, each haplotype in the true solution is required to be matched exactly once, and each haplotype in the PPHI solution is required to be matched at least once.

Implicit backbones are common to all optimal PPHI solutions, so it might be thought that they are surely in the true solution. Unfortunately, the backbone haplotypes cannot be completely relied upon as two implicit backbones did not truly exist in the samples of individuals, as shown in Table 3.

### 3.3 The number of nearly parsimonious solutions

Let \(H\) be the cardinality of haplotypes in the PPHI solution. How many solutions would there be with cardinalities \(\leq H+i\) for some
The true solution does not always have the most parsimonious number of haplotypes, as seen for datasets A, F and G.

small value? Let us first consider the case where \( i = 1 \). We make the following definitions. A \emph{fat} genotype reduces the PPHI cardinality from \( H \) to \( H - 2 \) if it is removed from the set. A \emph{fringe} genotype reduces the number to \( H - 1 \) if removed. A \emph{common} genotype does not reduce the number if removed. We observed very few fat genotypes in our data, and a few more common genotypes than fringe genotypes. For each genotype, there exists \( 2^k - 1 \) pairs of resolving haplotypes, where \( k \) is the number of heterozygous sites. Consider a single optimal PPHI solution and one genotype. If the genotype is common, every pair of resolving haplotypes that have one of the haplotypes already in the solution will yield an \( H + 1 \) solution. If it is a fringe genotype, every resolving pair in which neither of the haplotypes appear in the solution will yield an \( H + 1 \) solution. The number of \( H + 1 \) solutions increases when considering the other genotypes. The number further increases when each of the PPHI solutions are considered. Furthermore, these \( H + 1 \) solutions are just a subset of the total number possible. There may be many more that are not merely a PPHI solution with one additional haplotype. For these reasons, we expect the number of solutions to increase exponentially with increasing \( i \). Our experiments provide empirical evidence supporting this expectation.

Computing all of the solutions having at most \( H + 1 \) haplotypes is even more difficult than computing all optimal PPHI solutions. The ‘Known’ datasets contain three short sequences for which we were able to compute all \( H + 1 \) solutions. While these results are limited by computational complexity, they agree with the intuition that the number of solutions increase exponentially with increasing cardinality and demonstrate the feasibility of computing all nearly parsimonious solutions for reasonably sized datasets.


discussion

What factors contribute to the cardinality of the haplotypes? From our study of the seven ‘Known’ datasets, we observe that the cardinality of the known haplotypes is much closer to the lower limit than the upper limit. On the other hand, three of the seven datasets do not have parsimonious solutions. The least parsimonious solution is for dataset G, which has 32 unique haplotypes in its true solution, and only 28 are necessary to resolve this dataset. Nevertheless, 32 is far closer to the PPHI lower bound than to the upper bound of 78.

PPHI is NP-hard, so computing even one solution is computationally demanding for moderate sized datasets. Even if it was feasible to get all nearly parsimonious solutions, the vast number of solutions and their variations would yield little reliable information. Even implicit backbones cannot be guaranteed to be in the true solution.

Clark’s Subtraction Method was the first widely used tool for phasing genotypes. PPHI essentially optimized the basic premise of the Subtraction Method. The concise embodiment of this property has facilitated evaluation of its merit. We are now able to surmise that parsimonious considerations alone are inadequate to identify
a small set of solutions that have a high probability of containing the true solution.

In general, many assumptions made by biological models can be difficult to verify, due to many sources of error. These sources include error in the data, error in the model and error in the computational method. If a single iteration of a statistical or approximation method is run, the confidence in the solution is low. If a trillion random restarts are performed, the confidence is high, despite the fact that the same errors in the data and model persist. When a globally optimal solution is found using a combinatorial method, the confidence in the solution is 1. No error from the computational method can cloud the other sources of error and the assumptions made by the model can be more closely examined for their validity. For this reason, the use of a combinatorial optimization method can be a very valuable tool for important biological problems even when the combinatorial method cannot scale up to large instances.

The results presented in this article are intuitive and, when brought to light, appear obvious. Many researchers have already recognized the need to incorporate additional biological insight in their models and have developed software that has higher accuracy than pure parsimony programs (e.g. Eskin et al., 2003; Gusfield, 2002; Niu et al., 2002; Stephens and Donnelly, 2003). However, substantial research efforts continue to be made to improve the computation of the Pure Parsimony model. In addition to providing insights to the solution structures of inferred haplotypes, the other goal of this article is to suggest a redirection of the investment of intellectual talent to more biologically meaningful models.

**Funding:** Olin Fellowship (to S.C., in part); two National Institutes of Health grants (P50-GM65509 and 2R01 GM0287192A2 to A.T.); Alzheimer’s Association; two National Science Foundation grants (IIS-0535257 and DBI-0743797 to W.Z.).

**Conflict of Interest:** none declared.

**REFERENCES**


