Haplotyping populations by pure parsimony based on compatible genotypes and greedy heuristics

I-Lin Wang[a,*], Hui-E Yang[a]

[a] Department of Industrial and Information Management, National Cheng Kung University
No.1 University Rd, Tainan, 701, Taiwan

Abstract

The population haplotype inference problem based on the pure parsimony criterion (HIPP) infers an $m \times n$ genotype matrix for a population by a $2m \times n$ haplotype matrix with the minimum number of distinct haplotypes. Previous integer programming based HIPP solution methods are time-consuming, and their practical effectiveness remains unevaluated. On the other hand, previous heuristic HIPP algorithms are efficient, but their theoretical effectiveness in terms of optimality gaps have not been evaluated, either. We propose two new heuristic HIPP algorithms (MGP and GHI) and conduct more complete computational experiments. In particular, MGP exploits the compatible relations among genotypes to solve a reduced integer linear programming problem so that a solution of good quality can be obtained very quickly; GHI exploits a weight mechanism to selects better candidate haplotypes in a greedy fashion. The computational results show that our proposed algorithms are efficient and effective, especially for solving cases with larger recombination rates.

Keywords: Haplotype inference; Integer Programming; Bioinformatics; Compatibility Graph; Heuristics

1. Introduction

In the Post-Genomic Era, the development of a full Haplotype Map has high priority since Helmuth [1] suggests that the knowledge about the genetic constitution of an individual chromosome called haplotypes can be applied in linkage disequilibrium, inference of population evolutionary history, disease diagnosis, and customization of treatment for each individual. The haplotype is a sequence of closely linked single nucleotide polymorphisms (SNPs) in one copy of a chromosome. There are two haplotypes in a pair of chromosomes in all diploid organisms. Since the collection of haplotypes data requires a huge amount of cost and time, the data for genotypes rather than haplotypes are collected. A genotype is the description of one conflated pair of haplotypes. Clark [2] first brings up the population haplotype inference (PHI) problem to infer haplotypes from genotypes for a group of individuals.

* Corresponding author: Tel: +886-6-2757575-53123, Fax: +886-6-2362162,
E-mail address: ilinwang@mail.ncku.edu.tw (I.-L. Wang)

He also gives an inference rule, assuming the genotypes with zero or one ambiguous sites contain the frequently observed haplotypes in the population. After that, many PHI solution methods and problems are proposed, including the estimation-maximization (EM) algorithm by Excoffier and Slatkin [3], Long et al. [4], and Hawley and Kidd [5], Bayesian method by Stephens et al. [6], Niu et al. [7] and Stephens and Donnelly [8], maximum resolution (MR) problem by Gusfield [9], perfect phylogeny haplotype (PPH) problem by Gusfield [10], and PHI problem that satisfies the pure parsimony criterion (i.e. called the HIPP problem) proposed by Gusfield [11]. In this paper, we focus on solving the HIPP problem, where the number of distinct haplotypes for resolving a given genotype matrix is minimized.

Figure 1.

Suppose we are given m genotype vectors, each has length n, and denote this set of vectors as an $m \times n$ genotype matrix $G = [g_{i,j}]$. Each row in G corresponds to a genotype data for one individual, while each column stands for one SNP. The element $g_{i,j}$ is the genotype data for individual i at locus j. Let $G = \{g_1, g_2, ..., g_m\}$ be the set of all rows in G and $g_i = \left(g_{i,1}, g_{i,2}, ..., g_{i,n}\right)$ denote the genotype data for individual i. Every element $g_{i,j}$ has value 0, 1, or 2 depending on whether the ith individual is homozygous wild type, homozygous mutant, or heterozygous, respectively, at the jth SNP. Any element $g_{i,j}$ in G is said to be resolved if its value is 0 or 1, and ambiguous if its value is 2. Suppose h_a and h_b are two $1 \times n$ haplotype vectors. We say h_a and h_b resolve a genotype g_i, denoted by $g_i = h_a \otimes h_b$, if and only if $h_{a,j} = h_{b,j} = g_{i,j}$ for each $g_{i,j} \in \{0,1\}$ and $h_{a,j} + h_{b,j} = 1$ for each $g_{i,j} = 2$. See Figure 1 for an example. In this case, we say h_a (or h_b) a candidate (or explaining) haplotype for g_i, $\{h_a, h_b\}$ a candidate (or explaining) haplotype pair for g_i, and h_a (or h_b) the conjugate (or complementary) haplotype for h_b (or h_a) in the pair $\{h_a, h_b\}$.

By definition, the number of candidate haplotype pairs equals to $2^{\kappa-1}$ for a genotype that contains κ ambiguous sites (i.e. heterozygous SNPs). Although there are many candidate haplotype pairs for resolving a given genotype matrix, the real-world haplotype pairs are induced by a very few amount of distinct haplotypes. For example, Drysdale et al. [12] identify 13 SNPs in the human $\beta_2 AR$ gene, which has 2^{13} = 8192 possible combinations. However, only 10 among those 8192 candidate haplotypes are related to asthmatic cohort. Thus Gusfield [11] suggests a combinatorial optimization problem called the haplotype inference based on pure

parsimony (HIPP) which seeks the minimum amount of distinct haplotypes to resolve a given genotype matrix.

The objective of the HIPP problem is to find a $2m \times n$ haplotype matrix, in which the i th row (i.e. g_i) in the genotype matrix is resolved by the $(2i-1)$ th and the $2i$ th rows (i.e. $g_i = h_{2i-1} \otimes h_{2i}$) in the haplotype matrix, and the number of distinct haplotypes is minimized. See Figure 1 for an example. Given the genotype matrix G ={202,021,212}, there are 2, 1, and 2 candidate haplotype pairs to resolve genotype 1, 2, and 3, respectively. Furthermore, there are 6, 5, 5, or 4 distinct haplotypes if we select (p_1, p_3, p_4), (p_1, p_3, p_5), (p_2, p_3, p_4) or (p_2, p_3, p_5) to resolve G, respectively. Using the pure parsimony criterion, (p_2, p_3, p_5) will be selected to resolve all the genotypes since this combination induces the minimum number of distinct candidate haplotypes.

The HIPP problem is APX-hard, as shown by Lancia et al. [13]. The HIPP solution methods in the literature are either based on the mathematical programming techniques such as integer linear programming (ILP) by Gusfield [11] and Brown and Harrower [14] and quadratic integer programming by Huang et al. [15] and Kalpakis and Namjoshi [16], or the heuristic algorithm by Li [17]. In particular, Gusfield [11] gives the first ILP formulation called RTIP to model the HIPP problem. For each genotype, RTIP first enumerates all the candidate haplotype pairs and then gives constraints to ensure that exactly one among all the possible pairs for each genotype is selected. The formulation contains potentially exponential number of variables and constraints with respect to the number of SNPs. Although RTIP is potentially exponential-sized, it is practically faster than PolyIP, the polynomial-sized ILP formulation proposed by Brown and Harrower [14]. Wang and Xu [18] give a branch and bound algorithm called HAPAR. HAPAR also takes a lot of computational time since it tries out all the combinations to solve the HIPP problem. RTIP, PolyIP, and HAPAR all calculate the exact optimal solution for the HIPP problem, but they usually consume a lot of computational resources and time, and may not be suitable for solving large-scale HIPP problems.

Recently, Boolean Satisfiability (SAT) has been proposed for solving the HIPP problem with success. According to Lynce and Marques-Silva [19, 20], their SAT-based method called SHIPs can calculate exact optimal solutions for large-scale HIPP problems with sizes up to 50 genotypes and 100 sites, which are often unsolvable by

RTIP, PolyIP, and HAPAR. Techniques such as local search (see Lynce et al. [21]) have also been used to improve the efficiency of SHIPs. Furthermore, Graca, et al. [22, 23] propose Pseudo-Boolean Optimization (PBO) models called PolyPB and its reduced version called RPoly, based on the PolyIP model. RPoly has smaller size and consistently better efficiency than PolyPB. Their experiments indicate RPoly is the state-of-the-art method since it can solve more large-scale HIPP cases and is also often faster than SHIPs. Although these optimal HIPP algorithms give the theoretical optimal HIPP solutions, whether these optimal solutions also have good practical effectiveness in terms of small error rates (i.e. the proportion of genotypes whose inferred haplotype pairs are different from the original ones) requires further evaluation.

Heuristics are important in solving the PHI problems since the methodologies for obtaining the exact optimal solutions are not only theoretically difficult but also time-consuming in practice. Based on different Integer Quadratic Programming formulations, Huang et al. [15] give an approximation algorithm called SDPHapInfer, and Kalpakis and Namjoshi [16] propose another heuristic algorithm to solve their Integer Quadratic Programming HIPP formulation. SDPHapInfer performs well for smaller cases but its error rates increase dramatically for larger cases (see Section Computational Experiments and Discussion for details). The heuristic algorithm by Kalpakis and Namjoshi [16] requires further evaluation. On the other hand, the parsimonious tree growing heuristic algorithm called PTG by Li et al. [17] is very fast, but its theoretical effectiveness in terms of optimality gap (i.e. the difference in the number of distinct candidate haplotypes used, compared with the optimal solution) remains to be evaluated.

To design an efficient algorithm that can give a good solution to solve the HIPP problem, this paper investigates the compatible relations between genotypes. Based on the compatible relations between genotypes, we give estimates on the upper and lower bounds for the optimal HIPP objective value. Such compatible relations can also be used to reduce the solution space of Gusfield's formulation (RTIP) and obtain good solutions faster. Our first heuristic algorithm called MGP identifies common haplotypes for any two compatible genotypes, and then formulates an integer linear program to resolve the genotypes by those common haplotypes. MGP effectively reduces the solution space of the RTIP formulation so that it can obtain a good solution in a much shorter time than the original RTIP.

Since the HIPP problem seeks the minimum number of haplotypes to resolve a given genotype matrix, among all the candidate haplotypes, one that can resolve more genotypes should be more likely to appear in the solution since it may reduce the number of required haplotypes. To select such haplotypes, we design our second greedy heuristic algorithm called GHI by exploiting some weight mechanisms based on biological intuitions for each candidate haplotype and candidate haplotype pair so that the algorithm will select a candidate haplotype that seems to have higher probability to be in the optimal solution.

The remaining of this paper is organized as follows. Section 2 introduces the notations and compatibility graph used for our algorithms. In Section 3, we propose two new heuristic algorithms to solve the HIPP problem based on mathematical properties and biological insights. Several numerical experiments on both real-world and simulated genotype data for different formulations and algorithms are conducted and summarized in Section 4. Section 5 concludes the paper and suggests future research directions.

2. Preliminaries

Suppose ζ is a $1\times n$ row vector whose elements are either 0, 1, or 2 (e.g. a genotype or a haplotype). We say ζ_a and ζ_b are compatible to each other, denoted as $\zeta_a \sim \zeta_b$, if and only if $(\zeta_{a,j}, \zeta_{b,j}) \notin \{(0,1),(1,0)\}$ for each column $j = 1,...,n$. On the other hand, ζ_a and ζ_b are conflicting (or incompatible) to each other, if and only if there exists a column j such that $\zeta_{a,j} + \zeta_{b,j} = 1$. The compatible relation is reflexive and symmetric, but not transitive.

Figure 2.

A compatibility graph G_c (see Figure 2(a)) for a given genotype matrix G can consist of a set of m nodes and $\tilde{q}$ arcs where each node corresponds to a genotype in G and each arc connects two compatible genotypes. Similarly, one may define a conflicting graph $\overline{G_c} = K_m \setminus G_c$ (see Figure 2(b)), where K_m represents a complete graph of m nodes.

It is clear that two conflicting genotypes will not share common candidate haplotypes, which suggests to reduce an HIPP problem to several independent subproblems where each subproblem corresponds to a smaller HIPP problem whose genotypes form a component in G_c. Furthermore, the compatible relations can help to reduce the solution space of the RTIP model proposed by Gusfield [11]. In particular, one only requires enumerating those candidate haplotypes compatible with more than one genotype for the RTIP model, since those conflicting candidate haplotypes will increase the objective value and thus will never appear in the optimal solution, unless that genotype corresponds to an isolated node in G_c, in which case we can solve it separately.

The compatible relations also provide a base for designing some intriguing heuristic algorithms that give good HIPP solutions. For example, one may improve a trivial upper bound of $2m$ by solving an arc covering problem on G_c which seeks the minimum number of arcs, denoted as $A(G_C)$, whose selection covers all the nodes. In particular, selecting an arc (g_a, g_b) can be thought as resolving genotypes g_a and g_b using one of their common candidate haplotypes. Therefore, a selected arc in G_c will incur at most 3 haplotypes which provide a feasible solution that resolves all the genotypes with an estimated upper bound no more than $3A(G_C)$ on the optimal objective value. Similarly, one may use $\overline{G_C}$ to estimate the lower bound for the HIPP objective value. In particular, the minimum number of node colors required for $\overline{G_C}$ such that adjacent nodes in $\overline{G_C}$ have different colors, denoted by $\chi(\overline{G_C})$, provides an estimated lower bound $2\chi(\overline{G_C})$ for the HIPP objective value, since the nodes of the same color may at least require 2 candidate haplotypes. These estimated upper and lower bounds for the HIPP objective value may still be too loose. Similar compatible relations are also independently discussed by Lynce and Marques-Silva [20] and Lynce et al. [21], where they propose several techniques, different from ours, to improve the lower bounding procedures for their SHIPs method. In next section, we introduce two heuristic algorithms that seek a good solution in practice based on the compatible relations.

3. Proposed Heuristics

In this section, we give two heuristics to solve the HIPP problem. The first one exploits the compatible relations between genotypes and the second one selects popular haplotypes that can resolve more genotypes in a greedy fashion.

3.1 The method of merged genotype pairs (MGP)

Based on the compatible genotype pair relations, the common haplotypes for any two compatible genotypes can be easily identified and used to give a feasible solution to the HIPP problem. First, we construct a merged genotype pair, denoted as $mg_{\tilde{k}}$, $\tilde{k} = 1,...,\tilde{q}$, for each compatible genotype pair (g_a, g_b) as follows:

1. $mg_{\tilde{k},j} := g_{a,j}$, if $g_{a,j} = g_{b,j} \in \{0,1,2\}$, for $j = 1,...,n$,
2. $mg_{\tilde{k},j} := 1$ or 0, if $(g_{a,j}, g_{b,j}) \in \{(1,2),(2,1)\}$ or $(g_{a,j}, g_{b,j}) \in \{(0,2),(2,0)\}$, respectively, for $j = 1,...,n$.

After conducting the pairwise comparisons, all the $\tilde{q}$ distinct merged genotype pairs $MG(G) := \{mg_{\tilde{k}} : \tilde{k} = 1,...,\tilde{q}\}$ for a given genotype matrix G can be identified in $O(m^2 n)$ time. For each g_i, we use $IVMG(i) := \{mg_{\tilde{k}} : g_a \otimes g_i = mg_{\tilde{k}} \ \forall g_a \in G\}$ to record the set of its compatible merged genotype pairs.

Similar to the minimum arc covering technique for estimating the upper bound that selects the minimum number of arcs (i.e. compatible relations) in G_C to cover all the nodes (i.e. genotypes), our first heuristic algorithm, called as MGP, selects the minimum number of merged pairs in G_C to cover all the nodes. Since a merged pair may contain more than one arc (i.e. compatible relation) in G_C, the objective value of MGP tends to be better than the objective value obtained by the minimum arc covering heuristics for upper bound.

In particular, for each component $\hat{c}$ in G_C, we formulate an integer linear programming problem, denoted as $IP_{MGP1}(\hat{c})$, to select a set of minimal number of merged genotype pairs in G_C that covers all the nodes in component $\hat{c}$. Suppose these selected merged genotype pairs form a subgraph of G_C, denoted as $G_{mg\hat{c}}$. Then, we enumerate all the candidate haplotypes for each selected merged genotype pair, and

formulate another integer linear programming problem, denoted as $IP_{MGP2}(\hat{c})$, which corresponds to the Gusfield's formulation (RTIP) on the subgraph $G_{mg\hat{c}}$. Thus the optimal solution of $IP_{MGP2}(\hat{c})$ gives a set of minimum number of distinct haplotypes that resolves all the genotypes covered in complement $\hat{c}$ from those candidate haplotypes that can resolve the selected merged genotype pairs defined in $G_{mg\hat{c}}$. Since $G_{mg\hat{c}}$ is a spanning subgraph of $\hat{c}$ in G_C, in a sense MGP tries to compute a good feasible solution from a reduced solution space of the original problem. The final solution is obtained by the union of the solutions obtained for each component in G_C. The steps of MGP are described as follows:

Step 1. Conduct pairwise comparison to obtain $MG(G)$ and $IVMG(i)$ for each genotype g_i, $i = 1,...,m$.

Step 2. For each component $\hat{c}$ in G_C, formulate $IP_{MGP1}(\hat{c})$, and then use its optimal solution to formulate $IP_{MGP2}(\hat{c})$

Step 3. Obtain the solution of MGP by taking the union of the optimal solution of $IP_{MGP2}(\hat{c})$ for each component $\hat{c}$.

Without loss of generality, suppose G_C only contains one component $\hat{c}$. For each merged genotype pair $mg_{\tilde{k}}$, assign a binary variable $\tilde{x}_k$ to represent whether $mg_{\tilde{k}}$ is selected (i.e. $\tilde{x}_{\tilde{k}} = 1$) to resolve a genotype that it covers, or not (i.e. $\tilde{x}_{\tilde{k}} = 0$). The $IP_{MGP1}(\hat{c})$ can be formulated as follows:

$$\begin{aligned} \min \ & \sum_{\tilde{k}=1}^{\tilde{q}} \tilde{x}_{\tilde{k}} \\ s.t. \ & \sum_{\tilde{k} \in \arg IVMG(i)} \tilde{x}_{\tilde{k}} \geq 1, \ \forall i = 1,...,m \\ & \tilde{x}_{\tilde{k}} \in \{0,1\}, \ \forall \tilde{k} = 1,...,\tilde{q} \end{aligned} \qquad (IP_{MGP1}(\hat{c}))$$

The optimal solution for $IP_{MGP1}(\hat{c})$ can be used to form a solution space for $IP_{MGP2}(\hat{c})$. In particular, we enumerate all the candidate haplotypes for the selected

merged genotype pairs obtained from $IP_{MGP1}(\hat{c})$, and then use them to construct $IP_{MGP2}(\hat{c})$, which is a reduced RTIP formulation.

Figure 3.

Take the genotype $G = \{012, 221, 210, 220\}$ in Figure 3 for example. After conducting $4(4-1)/2 = 6$ pairwise genotype comparisons, we obtain 4 compatible relations and 3 merged genotypes: mg_1 = 011, mg_2 = 010 and mg_3 = 210. The compatibility graph for this example contains exactly one component and we can formulate $IP_{MGP1}(\hat{c})$ as follows:

$$\begin{array}{ll} \min & \tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 \\ s.t. & \tilde{x}_1 + \tilde{x}_2 \geq 1, \quad \tilde{x}_1 \geq 1, \quad \tilde{x}_2 + \tilde{x}_3 \geq 1 \\ & \tilde{x}_1, \tilde{x}_2, \tilde{x}_3 \in \{0,1\} \end{array}$$

Multiple optimal solutions $(\tilde{x}_1^*, \tilde{x}_2^*, \tilde{x}_3^*) \in \{(1,1,0),(1,0,1)\}$ exist for this example, which may lead to different results. The first optimal solution shows that G can be resolved by five distinct candidate haplotypes {010,011,101,110,100} using mg_1 and mg_2, since $(g_1, g_2, g_3, g_4) = (010 \otimes 011, 011 \otimes 101, 010 \otimes 110, 010 \otimes 100)$. On the other hand, the second optimal solution shows that mg_1 and mg_3 can be further expanded to a set of three haplotypes {011,010,110}, and then a set of six candidate haplotypes {010,011,101,110,100,000} can be derived to resolve G. Let $\{p_1, p_2, p_3, p_4, p_5\}$ denote the set of candidate haplotype pairs $\{010 \otimes 011, 011 \otimes 101, 010 \otimes 110, 010 \otimes 100, 110 \otimes 000\}$ derived from those six candidate haplotypes. Since g_4 can be resolved by more than one haplotype pair (i.e. p_4 and p_5), we have to form the second ILP $IP_{MGP2}(\hat{c})$ to decide the candidate genotype pairs that use fewest number of distinct haplotypes as follows:

$$\begin{array}{ll} \min & x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \\ s.t. & y_{1,1} = 1, \ y_{2,1} = 1, \ y_{3,1} = 1, \ y_{4,1} + y_{4,2} = 1 \\ & y_{1,1} \leq x_1, \ y_{1,1} \leq x_2, \ y_{2,1} \leq x_2, \ y_{2,1} \leq x_3, \ y_{3,1} \leq x_1, \\ & y_{3,1} \leq x_4, \ y_{4,1} \leq x_1, \ y_{4,1} \leq x_5, \ y_{4,2} \leq x_4, \ y_{4,2} \leq x_6, \\ & x_1, x_2, x_3, x_4, x_5, x_6, y_{1,1}, y_{2,1}, y_{3,1}, y_{4,1}, y_{4,2} \in \{0,1\} \end{array}$$

where the binary variables x_1, x_2, x_3, x_4, x_5 and x_6 represent whether the candidate haplotype 010, 011, 101, 110, 100, and 000 will be selected (i.e. $x=1$) in the optimal solution or not (i.e. $x=0$), and $y_{1,1}$, $y_{2,1}$, $y_{3,1}$, $y_{4,1}$, and $y_{4,2}$ represent whether the candidate haplotype pair p_1, p_2, p_3, p_4, and p_5 is selected (i.e. $y=1$) to resolve its corresponding genotype or not (i.e. $y=0$).

In summary, MGP is a heuristic algorithm that exploits the compatible relations of genotypes. A set of merged genotypes are produced by merging any two compatible genotypes. An ILP is formulated which seeks the minimum number of merged genotypes that can resolve all the genotypes. A second ILP based on Gusfield's RTIP formulation is then formulated using the candidate haplotypes for the selected merged genotype pairs to resolve all the genotypes. Since MGP can reduce the solution space of HIPP, it can give a good feasible solution for a large-scale HIPP problem that would be unsolvable in a reasonable time using the original RTIP formulation. We have implemented MGP and evaluated its performance in next section.

3.2 The greedy heuristic inference method (GHI)

The pure parsimony criterion implies a popular candidate haplotype that can resolve more genotypes should be more likely to appear in the HIPP solution, since its appearance reduces the number of required haplotypes. To select popular haplotypes, our second greedy heuristic algorithm, called as GHI, gives larger weight for a candidate haplotype compatible with more genotypes, and then selects those haplotypes of larger weights.

In addition to the popularity intuition, we also take the inference rule proposed by Clark [2] into consideration to design the weight mechanism for each candidate haplotype. Since Clark's rule tries to infer heterozygote (i.e. genotypes containing more than one ambiguous site) from homozygote (i.e. genotypes with zero or one ambiguous site), we think those genotypes with fewer ambiguous sites should contain candidate haplotypes that are more likely to be used for resolving the genotype matrix. Therefore, the candidate haplotypes associated with a genotype containing fewer ambiguous sites should be assigned larger weights. Based on these weight

mechanisms for each haplotype, GHI selects candidate haplotype pairs of larger weights to resolve all the genotypes.

Denote $H(G)$ and $HP(G)$ to be the set of candidate haplotypes and the set of candidate haplotype pairs that can resolve G, respectively. Let $|H(G)| = q$ and $|HP(G)| = \hat{q}$. For each g_i, we define $CH(i) := \{h_a : h_a \sim g_i, \forall h_a \in H(G)\}$ to be the set of its candidate haplotypes, and $CHP(i) := \{(h_a, h_b) : g_i = h_a \otimes h_b, \forall h_a, h_b \in H(G)\}$ to be the set of its candidate haplotype pairs. Therefore, $\bigcup_{i=1,...,m} CH(i) = H(G)$ and $\bigcup_{i=1,...,m} CHP(i) = HP(G)$ for each $CH(i) \subset H(G)$ and $CHP(i) \subset HP(G)$. For each candidate haplotype h_k, $k = 1,...,q$, we define $IG(k) := \{g_i : h_k \sim g_i, \ \forall g_i \in G\}$ to be the set of genotypes compatible with h_k.

Let $ntwo(i)$ and $gw(i)$ denote the number of ambiguous sites and the weights associated with g_i, respectively for each $i = 1,...,m$. By setting $gw(i) = (\sum_{j=1}^{m} ntwo(j)) \Big/ ntwo(i)$ for each $i = 1,...,m$, we assign a larger weight for a genotype with fewer ambiguous sites. Then, we calculate the weight for each candidate haplotype $h_k, k = 1,...,q$, by $hw(k) = \sum_{i \in \arg IG(k)} gw(i)$. The weight for each haplotype pair $hpw(\hat{k})$ equals to $hw(\hat{k}_a) \times hw(\hat{k}_b)$ for each $\hat{k} = 1,...,\hat{q}$, if the $\hat{k}$ th candidate haplotype pair is composed by the $\hat{k}_a$ th and the $\hat{k}_b$ th candidate haplotypes. Finally, our GHI algorithm selects the candidate haplotype pair that has the maximum pair weight to be the explaining pair for each g_i.

Besides the proposed weight mechanism for $gw(i)$, $hw(k)$, and $hpw(\hat{k})$, we have also tested other weight mechanisms such as equal weight, or linear incremental weight for $gw(i)$; or using addition rather than multiplication on the weights of the constituted haplotypes for $hpw(\hat{k})$. However, it turns out our proposed weight mechanism has the best performance. We give two points to explain the advantages of our proposed weight mechanism. First, the haplotype pair weight $hpw(\hat{k})$ can be thought as the frequency used in the EM algorithms by Excoffier and Slatkin [3], Long et al. [4], and Hawley and Kidd [5], where the idea of frequency is treated as a concept of probability. Therefore, the multiplication on the weights of the constituted haplotypes represents the probability for that candidate haplotype pair using those two

constituted haplotypes. Similar to the EM algorithm which finds the haplotype probabilities to optimize the probability of the entire population, here we select the haplotype combinations with the largest estimated probability. Second, using $hw(\hat{k}_a) \times hw(\hat{k}_b)$ instead of $hw(\hat{k}_a) + hw(\hat{k}_b)$ to calculate $hpw(\hat{k})$ for each haplotype pair provides a better tie-breaking strategy.

GHI algorithm is straightforward and contains five procedures:

Step 1. Construct $H(G)$, $HP(G)$ by $CH(i)$, $CHP(i)$ for each $g_i \in G$. Construct $IG(k)$ for each candidate $h_k \in H(G)$.

Step 2. Calculate $gw(i) = (\sum_{j=1}^{m} ntwo(j)) \Big/ ntwo(i)$ for each $g_i \in G$ using $ntwo(i)$.

Step 3. Calculate $hw(k) = \sum_{i \in \arg IG(k)} gw(i)$ for each candidate haplotype $h_k \in H(G)$.

Step 4. Calculate $hpw(\hat{k}) = hw(\hat{k}_a) \times hw(\hat{k}_b)$ for each candidate haplotype pair $hp_{\hat{k}} = (hp_{\hat{k}_a}, hp_{\hat{k}_b}) \in HP(G)$

Step 5. Select the candidate haplotype $hp_{i*} \in CHP(i)$ such that $i* = \arg\max_{hp_{\hat{k}}} \{hpw(\hat{k})\}$ for each genotype $g_i \in G$

Take Figure 4 for example. Given $G = \{202, 021, 212\}$, three candidate haplotype pairs: one of p_1 and p_2, p_3 itself, and one of p_4 and p_5, have to be selected to resolve g_1, g_2, and g_3, respectively. Since there are totally five ambiguous sites in G, $(gw(1), gw(2), gw(3)) = (2.5, 5, 2.5)$. The weights for haplotypes 000, 101, 001, 100, 011, 010, 111, and 110 are 2.5, 2.5, 7.5, 2.5, 7.5, 2.5, 2.5, and 2.5, respectively. The weighs for candidate haplotype pairs p_1, p_2, p_3, p_4 and p_5 become 6.25, 18.75, 56.25, 6.25, and 18.75, respectively. Then, GHI will select p_2, p_3, and p_5 to resolve g_1, g_2, and g_3, respectively.

Figure 4.

Suppose there are totally $|H(G)| = q$ candidate haplotypes and $|HP(G)| = \hat{q} = \sum_{i=1}^{m} 2^{ntwo(i)-1}$ candidate haplotype pairs. GHI takes $O(\hat{q} + mn + qm)$

time and $O(2^n)$ storage space using array to store $H(G)$ and $HP(G)$. Or, it takes $O(\hat{q}^3 + mn + qm)$ time and $O(\hat{q})$ storage space using linked list to store $H(G)$ and $HP(G)$. The efficiency of GHI thus depends very much in the total number of ambiguous sites that G contains. Note that the RTIP formulation by Gusfield [11] and the SDPHapInfer algorithm by Huang et al. [15] also require to enumerate all the necessary candidate haplotypes as does in the Step 1 of GHI, thus these three methods all consume similar time and storage space in their first step. However, considering the computational efforts afterwards, GHI only conducts simple calculation on weights, whereas RTIP has to solve an ILP problem and SDPHapInfer has to solve an SDP (i.e. semi-definite programming) problem. Therefore GHI should consume less computational time than RTIP and SDPHapInfer. Although GHI gives no theoretical guarantee on its solution quality, it includes the concept of Clark's rule and simple greedy intuition to obtain good solutions efficiently. Moreover, the practical performance of GHI has been shown to be good (see section Computational Experiments and Discussion), especially for cases with recombination.

4. Computational Experiments and Discussion

This section summarizes the computational experiments for several HIPP algorithms. After introducing the experimental settings, one biological genotype dataset (i.e. $\beta_2 AR$ gene) and several simulated genotype datasets will be used for the experiments.

4.1 Settings for our computational experiments

All the computational experiments are conducted on a Pentium 4 PC with 3.2 GHz CPU, 1GB RAM and Windows XP operating system. Six algorithms are implemented and evaluated: (1) our merged genotype pair heuristic, denoted as "MGP", is implemented in C++, compiled by Visual C++, and linked with CPLEX 9.0 callable library; (2) our greedy heuristic, denoted as "GHI", is implemented in C++ and compiled by g++ compiler; (3) the heuristic algorithm directly imported from Li et al. [17], denoted as "PTG", is implemented using Borland Delphi 5.0 in Pascal; (4) the semidefinite programming relaxation algorithm by Huang et al. [15],

denoted as “SDP”, is implemented using MATLAB; (5) the integer linear programming model by Gusfield [11], denoted as “RTIP”, is implemented in C++, compiled by Visual C++, and linked with CPLEX 9.0 callable library; and (6) the Clark inference rule algorithm directly imported from Clark [2], denoted as “HAP”, is implemented using Fortran. Note that although HAP is not designed for the pure parsimony criterion, we include it into our experiments since it is one of the haplotyping algorithms based on statistical models widely used in the community, and thus the error rates of HAP can serve as a base to compare the practical effectiveness of our proposed HIPP algorithms. Besides these six PHI algorithms, we have also implemented the basic PolyIP model (i.e. the one without branch and cut techniques) of Brown and Harrower [14]. Although PolyIP is an integer programming formulation with polynomial size which is theoretically better than the exponential-sized RTIP formulation by Gusfield [11], the results of our implementation indicates that PolyIP takes much more time than RTIP, even when solving small or medium sized HIPP problems. In fact, it performs the slowest in all of our experiments. Therefore, we do not include the results of PolyIP in this paper. We have not included the SAT-based HIPP methods such as SHIPs and RPoly into our computational tests, since their source codes are not publicly available for further modifications to meet our needs. Moreover, SHIPs and RPoly rely on some state-of-the-art SAT solver (e.g. MiniSAT by Eén and Sörensson, [24]) and PBO solver (e.g. MiniSAT+ by Eén and Sörensson, [25]), which are very different from the ILP solver (CPLEX) used by RTIP, PolyIP and MGP. Here in this paper, we focus on the performance of those ILP-based HIPP methods, and leave the experiments on those SAT-based HIPP methods for future research.

We record the computational time for several PHI algorithms in our experiments. Note that the amount of time spent by an algorithm can not really indicate its actual efficiency, since these algorithms are not implemented using the same programming language. However, for the practical purpose, a faster PHI algorithm is usually preferred.

We use the error rate, defined as the proportion of genotypes whose inferred haplotype pairs are different from the original ones, to evaluate the practical effectiveness for several classes of the HIPP algorithms. Since the original inferred haplotype pairs usually can not be obtained in advance, the error rate is highly intractable, case-dependent, and may easily grow to a scale of 30% or more, even for

those haplotyping algorithms based on statistical models by Stephens et al. [8] and Niu et al. [7]. Intuitively, an algorithm that gives smaller error rates is usually considered to be more effective, although such an algorithm may not be the most effective one at all times.

For those non-optimal HIPP algorithms, we are the first to use the optimality gap, defined as the ratio of the difference in the number of selected inferred haplotypes obtained by an HIPP algorithm to the minimum number of inferred haplotypes (i.e. the optimal objective value of an HIPP problem), as a parameter to evaluate their theoretical effectiveness. Although the HIPP problem is an optimization problem used to model the haplotyping problem, whether an optimal HIPP algorithm does always give a more effective (e.g. of smaller error rate) solution than a non-optimal HIPP algorithm or not has never been evaluated. Here we first calculate the optimality gap for the solution of each non-optimal HIPP algorithm, and then check whether an HIPP algorithm that gives smaller optimality gap is indeed more effective (i.e. has a smaller error rate).

4.2 Experiments on β_2AR gene

We use a sample of the human $\beta_2 AR$ gene identified by Drysdale et al. [12] for our testing. Similar experiments have also been conducted in Stephens and Donnelly [8], Huang et al. [15], Li et al. [17], and Wang and Xu [18]. It contains 18 different genotypes and 13 SNPs from 121 individuals, and should be resolved by the 10 haplotypes related to asthmatic cohort. Among all algorithms we have tested, RTIP, MGP, GHI, and HAP all obtain 10 distinct haplotypes on human $\beta_2 AR$ gene data, thus they are all optimal with respect to the pure parsimony criterion. The error rate of MGP, GHI, and HAP are all 0. However, RTIP has an error rate 6% due to the existence of multiple optimal solutions. Both SDP and PTG conduct randomized mechanism which may result in different solutions even for the same genotype matrix. Thus we apply SDP and PTG to solve this genotype matrix for three times. SDP obtains 12 haplotypes and has an error rate 11% on average. On the other hand, PTG obtains 10, 11, and 12 haplotypes, which correspond to the error rates of 0%, 17%, and 17%, respectively.

4.3 Experiments on simulated data

We use the program by Hudson [26] to simulate a $2m \times n$ haplotype matrix, and then randomly pair two haplotypes from these $2m$ haplotypes to produce an $m \times n$ genotype matrix in a way that none of the $2m$ haplotypes is repeatedly paired. Similar simulated techniques can also be found in Stephens et al. [6], Niu et al. [7], Gusfield [11], Brown and Harrower [14], Huang et al. [15], and Wang and Xu [18].

Recombination is a process during the formation of gametes where portions from the paternal and maternal genome are exchanged. Recombination may cause the haplotypes in offspring to be different from those in their parents, and higher recombination rate cause more different haplotypes in offspring. We simulate the input genotypes and haplotypes with different recombination rate r to evaluate the effectiveness for several PHI algorithms. In particular, we have tested three recombination rates: $r = 0$ (no recombination), $r = 4$ and $r = 16$.

For each tested recombination rate (i.e. r = 0, 4, or 16), nine problem sets of different genotype matrix sizes (10×10, 20×10, 30×10, 10×20, 20×20, 30×20, 10×30, 20×30, and 30×30) are simulated, where 30 random test cases for each problem set have been generated. The size of the largest genotype matrix tested in our experiments (i.e. 30×30) is fairly large, compared with the largest cases, 40×10 and 25×10, tested in Wang and Xu [18] and Huang et al. [17], respectively. In fact, the number of SNP affects the running time more than the number of genotypes, especially for some algorithms (e.g. RTIP and GHI) that require enumeration of all candidate haplotype pairs.

Two parameters, optimality gap and error rate, for each algorithm in each test case are recorded for evaluation. For each parameter, we use their average over the 30 test cases of the same problem set to represent its overall performance. Note that we do not list the optimality gap for the algorithms RTIP and HAP, since RTIP is already optimal and HAP may not resolve all the genotypes which makes its optimality gap meaningless.

Although the computational time (see Table 1, Table 2, and Table 3) in our experiments are not for the purpose of efficiency comparison, the results indicate that all the heuristics (MGP, GHI, and PTG) solve the HIPP very quickly. Among these three heuristics, PTG is the most efficient algorithm. MGP is the second efficient

since it solves two ILPs with sizes much smaller than the size of RTIP. GHI requires more time than the other two, since it sorts a set of exponential-sized candidate haplotype pairs to select the best one. On the other hand, both RTIP and SDP require much more time than the others. For our convenience, we only record the cases solvable within two hours for each algorithm. Although SDP can be terminated in two hours in our tests, RTIP may take longer. Thus we list the number of cases solvable within two hours by RTIP. For example, 10×30 (26/30) in the first column of Table 1, Table 4, and Table 7 indicates that only 26 out of the 30 test cases for the 10×30 problem sets are solvable by RTIP within two hours for the cases without recombination.

Table 1.

Table 2.

Table 3.

Table 4 and Table 7 record the average performance in the optimality gap and error rate for different algorithms when there is no recombination (i.e. $r = 0$). The results for the cases with recombination rate $r = 4$ and $r = 16$ are summarized in Table 5, Table 8, and Table 6, Table 9, respectively.

Table 4.

Table 5.

Table 6.

For the cases without recombination (i.e. $r = 0$), among those non-optimal algorithms, the optimality gap in Table 4 shows that PTG performs the best, GHI performs the second, MGP performs the third, and SDP performs the fourth. Moreover, the optimality gap of SDP increases dramatically for larger cases. In terms of error rate, RTIP has the smallest error rate than all other algorithms in all the cases without recombination (see Table 7), then PTG performs the second, GHI performs the third, MGP performs the fourth, HAP performs the fifth, and finally SDP performs the sixth. All the algorithms have consistent performance on both the optimality gaps and error rates for these cases without recombination. This indicates the pure

parsimony criterion does serve its purpose for cases without recombination, so that algorithms that have smaller optimality gaps tend to have smaller error rates.

Table 7.

Table 8.

Table 9.

For the cases with recombination (i.e. $r = 4$ or $r = 16$), among those non-optimal algorithms, the optimality gap in Table 5 and Table 6 shows that MGP performs the best, GHI performs slightly worse than MGP, PTG performs the third, whereas SDP has smaller optimality gap for cases up to 10 SNPs but the gap increases dramatically for larger cases. In terms of error rate (see Table 8 and Table 9), RTIP again has the smallest error rate than all other algorithms in all the cases with recombination. For algorithm A and B, let $A \succ B$ represent A has smaller error rate than B. In general, $GHI \succ PTG \succ MGP \succ HAP \succ SDP$ for the cases of $r = 4$, and $GHI \succ MGP \succ HAP \succ PTG \succ SDP$ for the cases of $r = 16$. From Table 7, Table 8, and Table 9, we also find that the error rates are larger for cases with the same number of SNPs but fewer genotypes, due to the lack of sufficient information to infer haplotypes (similar conclusion can also be found in Huang et al. [15] and Wang and Xu [18]).

Table 10.

Although these algorithms do not perform consistently on the optimality gaps and error rates for cases with recombination, our heuristics GHI and MGP do perform very well consistently, and have better effectiveness in most cases, especially for cases with recombination rates. In particular, GHI and MGP have better error rates than HAP in most cases. Since HAP is a widely used haplotyping algorithm in the community, this indicates our proposed algorithms are also practically useful and competitive to the other algorithms based on statistical models. On the other hand, PTG performs very well for cases without recombination rate, but has consistently worse theoretical and practical effectiveness than GHI and MGP, especially when the recombination rate increases.

In summary, Table 10 lists the comparative performance for these six PHI algorithms, and ranks their relative performance along a given parameter (time, optimality gap, or error rate); as the relative performance worsens, the ranking gets higher. From the results, we recommend RTIP for researchers who can bear with longer running time but seek a solution with the smallest error rates. On the other hand, PTG is recommended for researchers seeking a very quick solution with promising quality. For researchers seeking a better solution within a promising schedule, we recommend both GHI and MGP.

5. Conclusions

This paper focuses on issues in solving the PHI problem based on pure parsimony criterion (HIPP). The major contributions of this paper are in two folds. First, we propose two new heuristic algorithms (MGP and GHI) based on the concept of compatible genotype pairs and a weight mechanism that takes the Clark's inference rule into consideration for selecting better candidate haplotypes in a greedy fashion. Approximated lower and upper bounds for the objective value of the HIPP problem can also be derived from the properties of the compatibility graph. Second, we conduct extensive computational experiments for six PHI algorithms on one biological dataset and several simulated datasets. Our experiments are more complete than others in the literature since we evaluate the optimality gap and error rate, which cover both the theoretical and practical effectiveness of HIPP algorithms at the same time. Our results show the pure parsimony criterion does serve its purpose in providing a practically good HIPP solution, since a HIPP algorithm that produces smaller optimality gaps does usually give smaller error rates. Since seeking the exact optimal HIPP solution is too time-consuming, our results encourage the pursuit of developing more efficient and effective HIPP algorithms. To this end, our proposed heuristics (MGP and GHI) can successfully give very good solutions in a shorter time, and are especially effective for cases with larger recombination rates.

For future research, we suggest to evaluate the error rates for those SAT-based HIPP methods such as SHIPs and RPoly. We also encourage researchers to investigate or improve their HIPP methods, based on the compatibility graph and weight mechanism used by MGP and GHI, as proposed in this paper.

Acknowledgements

I-Lin Wang was partly supported by the National Science Council of Taiwan under Grant NSC95-2221-E-006-268.

References

[1] L. Helmuth, Map of the human genome 3.0. *Science* 293, 583-585 (2001).

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

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

[4] J.C. Long, R.C. Williams, M. Urbanek, An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* 56, 799-810 (1995).

[5] M.E. Hawley, K.K. Kidd, HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.* 86, 409-411 (1995).

[6] M. Stephens, N.J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978-989 (2001).

[7] T. Niu, Z.S. Qin, X. Xu, J.S. Liu, Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 70, 157-169 (2002).

[8] M. Stephens, P. Donnelly, A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* 73, 1162-1169 (2003).

[9] D. Gusfield, Inference of haplotypes from samples of diploid populations: complexity and algorithms. *J. Comput. Biol.* 8, 305-323 (2001).

[10] D. Gusfield, Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In: Annual Conference on Research in Computational Molecular Biology: Proceedings of the sixth annual international conference on computational biology, pp. 166-175 (2002).

[11] D. Gusfield, Haplotype inference by pure parsimony. In: Combinatorial Pattern Matching: 14th Annual Symposium, pp. 144-155 (2003).

[12] C. Drysdale, D. McGraw, C. Stack, J. Stephens, R. Judson, K. Nandabalan, K. Arnold, G. Ruano, S. Liggett, Complex promoter and coding region β_2 -

adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *P. Natl. Acad. Sci. USA.* 97, 10483-10488 (2000).

[13] G. Lancia, C.M. Pinotti, R. Rizzi, Haplotyping populations by pure parsimony: Complexity, exact and approximation algorithms. *INFORMS J. Comput.* 16, 348-359 (2004).

[14] D.G. Brown, I.M. Harrower, Integer programming approaches to haplotype Inference by pure parsimony. *IEEE ACM T. Comput. Bi.* 3, 141-154 (2006).

[15] Y.T. Huang, K.M. Chao, T. Chen, An approximation algorithm for haplotype inference by maximum parsimony. *J. Comput. Biol.* 12, 1261-1274 (2005).

[16] K. Kalpakis, P. Namjoshi, Haplotype phasing using semidefinite programming. In: Proceedings of the fifth IEEE symposium on bioinformatics and bioengineering, pp. 145-152 (2005).

[17] Z. Li, W. Zhou, X. Zhang, L. Chen, A parsimonious tree-grow method for haplotype inference. *Bioinformatics* 21, 3475-3481 (2005).

[18] L. Wang, Y. Xu, Haplotype inference by maximum parsimony. *Bioinformatics* 19, 1773-1780 (2003).

[19] I. Lynce, J. Marques-Silva, SAT in Bioinformatics: Making the Case with Haplotype Inference. In: International Conference on Theory and Applications of Satisfiability Testing, August 2006, Seattle, USA.

[20] I. Lynce, J. Marques-Silva, Haplotype Inference with Boolean Satisfiability. *Int. J. Artif. Intell. T.* 17 (2), 355-387 (2008).

[21] I. Lynce, J. Marques-Silva, S. Prestwich, Boosting Haplotype Inference with Local Search. Constraints: An International Journal, 13 (1-2) (2008).

[22] A. Graca, J. Marques-Silva, I. Lynce, A. Oliveira, Efficient Haplotype Inference with Pseudo-Boolean Optimization. In: Algebraic Biology, July 2007, Hagenberg, Austria.

[23] A. Graca, J. Marques-Silva, I. Lynce, A. Oliveira, Efficient Haplotype Inference with Combined CP and OR Techniques. In: International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, May 2008, Paris, France. (In Press)

[24] N. Eén, N. Sörensson, An extensible SAT-solver. In International Conference on Theory and Applications of Satisability Testing (SAT), 502-518 (2003).

[25] N. Eén, N. Sörensson, Translating pseudo-Boolean constraints into SAT. *Journal on Satisability, Boolean Modeling and Computation* 2, 1-26 (2006).

[26] R.R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337-338 (2002).