New Aspects of Haplotype Inference from SNP Fragments

Huimin Chen
Department of Electrical Engineering
University of New Orleans, USA

Zhiyu Zhao
Department of Computer Science
University of New Orleans, USA

Kun Zhang
Department of Computer Science
Xavier University of Louisana, USA

Dongxiao Zhu
Department of Computer Science
University of New Orleans, USA

1 Introduction

The genome of an eukaryotic organism is a complete DNA sequence of one set of chromosomes. DNA sequences can be derived from biological raw material through the DNA sequencing process. A DNA sequence is a succession of nucleotides representing the primary structure of a real or hypothetical DNA molecule or strand with the capacity to carry genetic information. The possible letters are A, C, G, and T, representing the four nucleotide subunits of a DNA strand: adenine, cytosine, guanine and thymine. Therefore, a genome can be considered a string over the alphabet of nucleotides { A, C, G, T }.

With the advances in high-throughput genome sequencing technology, we can now assess the DNA sequence variation at the population level. A Single Nucleotide Polymorphism (SNP) is a locus in the DNA sequence where an alternation of the nucleotide from other members of the same species occurs at a considerably high frequency. Alleles of a set of linked genetic markers located on a single DNA sequence is called haplotype. For diploid organisms, such as humans, haplotypes come in pairs, where the two haplotypes in the pair are not necessarily identical. Each pair of haplotypes can also be combined to form a genotype. When a pair of alleles at an SNP site is made of identical type (either being both wild, which is denoted by 0, or both mutant, which is denoted by 1), the site is called a homozygous site. Otherwise, it is called a heterozygous site (denoted by 2).

Haplotypes are generally believed to contain more genetic information than individual SNPs in disease association studies. They are substantially more difficult to determine compared with genotypes or individual SNPs through experiments however. Due to its potential in genetics research, haplotype inference has drawn

significant attention from both statistical and computational research domains.

1.1 Inferring Haplotypes from Genotypes

Computational methods for haplotype inference can be largely categorized into two different categories. One concerns with obtaining compatible haplotypes from the genotyped samples in a population. Computationally, the problem can be defined as follows. Given m genotypes, where each genotype is an n-dimensional vector and each site of a genotype takes its value from $0, 1, 2$, find one pair of haplotypes for each genotype, where a haplotype is an n-dimensional vector with each site taking its value from $0, 1$. If a site in a genotype is homozygous, that is, 0 or 1, respectively, then the same site in both haplotypes should also be 0 or 1, respectively. If a site in a genotype is heterozygous, that is, 2, then the same site should be 0 in one haplotype and 1 in another. This computational challenging problem is also known as genotype phasing.

The pioneering work by Clark [Clark, 1990] demonstrated an effective haplotype inference from genotypes using the parsimony method. Later, various frequentist and Bayesian methods have been proposed for large genotype data sets as well as large number of subjects under different assumptions on the underlying biology systems [Excoffier & Slatkin, 1995, Gusfield, 2001, Niu et al., 2002, Brinza & Zelikovsky, 2008].

A parsimonious solution to the haplotype inference problem is the attempt to minimize the total number of different haplotypes. It first identifies genotypes with only homozygous sites or a single heterozygous site. In the first case, the two haplotypes of a genotype are exactly the same with the genotype itself. In the second case, the two haplotypes differ only at the heterozygous site. Such genotypes and their haplotypes are considered "resolved," and all other genotypes are considered "ambiguous." For each ambiguous genotype G and a resolved haplotype H compatible with G, we then infer another haplotype based on G and H, add the inferred haplotype to the set of resolved haplotypes, and remove G. The procedure is repeated until either all genotypes are resolved or no more of them can be resolved. Clark's method requires an initial set of resolved haplotypes. In addition, as applying different orders of resolved haplotypes to ambiguous genotypes gives different solutions, Clark recommended that this method be executed for many times with random reordering of the initial genotypes and output the solution that resolves most of the genotypes.

Clark's method led to the Maximum Resolution (MR) problem, an algorithmic problem on the maximum number of ambiguous genotypes that can be resolved using Clark's method or on the specific order of the initial input that maximizes the number of resolved genotypes. The MR problem has been proven to be NP-hard and Max-SNP complete [Gusfield, 2001]. Gusfield et al. cast the MR problem in the graph theoretic framework and proposed an integer linear programming solution, as described in [Gusfield, 2001].

A statistical approach based on the Expectation-Maximization (EM) algorithm was proposed by Excoffier et al. in [Excoffier & Slatkin, 1995]. The algorithm was used to calculate Maximum-Likelihood (ML) estimates of haplotype frequencies under the assumption of the Hardy-Weinberg Equilibrium (HWE). The HWE states that if two haplotypes occur in nature with frequencies f_1 and f_2, respectively, then the haplotype pair occurs in nature with a probability of $f_1 f_2$. Given a set of genotypes, the algorithm first randomly initializes the frequencies of all possible haplotypes and then it estimates and updates those frequencies to maximize the log-likelihood function.

Niu et al. proposed a Bayesian approach using Gibbs sampler to solve the haplotype inference problem [Niu et al., 2002]. They first partitioned the whole haplotype into smaller segments. They followed it up by a Gibbs sampler both to construct the partial haplotypes of each segment and to assemble all the segments together. The Bayesian approach elegantly handled the missing data.

Brinza et al. presented a method called 2SNP [Brinza & Zelikovsky, 2008]. The method first explored the phasing of genotypes with two ambiguous SNPs when both sites are heterozygous. Two phasings are possible. For a 2-SNP genotype 22, one possible phasing is to represent the two haplotypes as 00 and 11,

respectively. This is called the cis-phasing. Another possible phasing is trans-phasing, which represents the two haplotypes as 01 and 10, respectively. It is assumed that the true phasing chooses the most frequent pair of haplotypes observed in the population sample. Complete haplotypes for a given genotype were inferred based on the maximum spanning tree of a complete graph, with vertices corresponding to heterozygous sites and edge weights given by inferred 2-SNP frequencies.

Liang et al. presented a deterministic sequential Monte Carlo method in [Liang & Wang, 2008]. Given G, which is a set of N genotypes, and Z, which is a set of all the compatible haplotypes, they attempted to infer N pairs of haplotypes without knowing the frequencies of haplotypes in Z. The algorithm performs in a sequential manner, where the haplotype pairs are inferred in the order in which their genotype vectors are arranged.

Currently, the total number of SNPs in the human genome reported in public databases already exceeds 9 million, making both genotyping techniques and the associated inference algorithms fall much behind in solving the real-world problem [Kim & Misra, 2007]. Recent techniques use short genome fragments with SNPs coming from DNA shotgun sequencing or some other re-sequencing procedures to reconstruct the haplotypes directly. A specific problem herein called single individual haplotyping reconstructs a pair of long haplotype sequences from short fragments of SNPs [Cilibrasi et al., 2007].

1.2 Single Individual Haplotyping

If SNP fragments are well aligned, then the problem becomes how to partition these SNP fragments into two sets and then to use each set of fragments to reconstruct each haplotype sequence. Suppose that there are m SNP fragments from a pair of chromosomes, and the length of the corresponding haplotypes is n. Define an $m \times n$ SNP matrix M whose entry m_{ij} has value 0, 1, or –. The symbol – means a missing or skipped base, which is called a gap. Let h be a haplotype sequence whose entry can be either 0 or 1. Let $\Theta = (M_1, M_2)$ be a partition of of M that divides the rows of M into two disjoint sets. The haplotype reconstruction problem can be written in terms of the ML estimation

$$(\hat{h}_1, \hat{h}_2, \hat{\Theta}) = \arg \max_{h_1, h_2, \Theta} \Lambda(M_1|h_1)\Lambda(M_2|h_2),$$

assuming that the SNP fragments from each haplotype sequence are conditionally independent. The above method entails the knowledge of likelihood function of any potential haplotype sequence, which is closely related to the reading error and gap probabilities for each site and may not be available from the raw SNP data set. Alternatively, one can apply minimum fragment removal [Rizzi et al., 2002], minimum SNP removal [Rizzi et al., 2002], minimum error correction [Wang et al., 2005], or longest haplotype reconstruction [Dondi, 2009] as the optimality criterion in reconstructing the pair of haplotype sequences without the need for the likelihood function model $\Lambda(\cdot)$. Most of these criteria are inherently based on parsimony and are combinatorial in nature: finding exact optimal solution is often an NP hard problem [Lancia et al., 2001], although efficient suboptimal methods exist, such as SpeedHap [Genovese et al., 2008].

Haplotype reconstruction from SNP fragments as an optimization problem entails finding a minimum number of operations (e.g., removing fragments, SNPs, or errors) on an SNP matrix, such that it becomes feasible, that is, the rows of the SNP matrix can be divided into two disjoint sets of pairwise compatible fragments, with each set determining a haplotype. Lancia et al. defined the Minimum Fragment Removal (MFR), Minimum SNP Removal (MSR), and Longest Haplotype Reconstruction (LHR) models in [Lancia et al., 2001] for the above problem. Given an SNP matrix, the MFR model removes the minimum number of fragments (rows) to make the resulting matrix feasible; the MSR model removes the minimum number of SNPs (columns) to make the resulting matrix feasible; and the LHR model removes a set of fragments (rows) to make the resulting matrix feasible and to maximize the sum of the lengths of the derived hap-

lotypes is maximized. In [Rizzi et al., 2002], Rizzi et al. proposed practical algorithms and fixed-parameter tractability for the MFR and MSR models. The LHR model was further proven by Dondi [Dondi, 2009] to be NP-hard even when the SNP matrix was error-free. He also gave a fixed-parameter algorithm to the LHR problem, wherein the parameter is the size of the reconstructed haplotypes. Lippert et al. proposed a Minimum Error Correction (MEC) model and proved its NP-hardness in [Lippert et al., 2002]. The MEC model is as follows. Given an SNP matrix, correct a minimum number of elements (0 into 1 and vice versa) to make the resulting matrix feasible. An exact algorithm based on this model and a heuristic method were presented in [Wang et al., 2005] by Wang et al. In [Chen et al., 2008], Chen et al. proposed some linear time random algorithms to solve the problem based on a probabilistic model, where each element in the SNP matrix has probability α_1 to be an error, independently, probability α_2 to be missing, and the dissimilarity between the two haplotypes to be inferred is β. Their algorithms are able to solve the problem when the parameters, that is, the probability and dissimilarity values are either known or unknown.

In this chapter, we provide a different viewpoint on the haplotype reconstruction from the SNP fragments M. The problem is treated as decoding with noisy observations over a discrete memoryless channel. The parsimony-based criteria, albeit computationally demanding, do not necessarily yield the best reconstruction rate. There is a significant performance gap between the error correction capability with SNP fragments and the theoretical limit delineated by the channel capacity. To improve the error correction rate, we propose a new computational model that can also utilize the genotype information, if available, for haplotype reconstruction. The resulting algorithm has low complexity and is shown to guarantee the desired error rate by increasing the number of SNP fragments sequentially. The advantage of using genotype information is quantified by exploiting a simplified statistical model for haplotype sequences.

The rest of this chapter is organized as follows. Section 2 formulates the haplotype reconstruction problem from the information theoretic perspective. Section 3 introduces a low complexity sequential algorithm using genotype information. Section 4 compares the reconstruction rate between the cases without and with genotype information and quantifies performance gain. Concluding remarks are presented in Section 5.

2 Haplotype Reconstruction As A Decoding Problem

2.1 Decoding Haplotype Sequences

We consider the problem of inferring a pair of haplotype sequences from inconsistent and possibly gap-abundant SNP fragments. At any SNP site, there are three possible readings denoted by 0, 1, and – for convenience. For $x, y \in \{0, 1, -\}$, we define the distance

$$d(x,y) = \begin{cases} 1 & \text{if } x \neq y, x \neq -, y \neq - \\ 0 & \text{otherwise} \end{cases}$$

which is an extension of the Hamming distance defined for $x, y \in \{0, 1\}$. Similarly, we define the distance between two sequences of length n by

$$d(h_1, h_2) = \sum_{i=1}^{n} d(h_{1i}, h_{2i}).$$

The dissimilarity between h_1 and h_2 is measured by

$$\beta = d(h_1, h_2)/n$$

and thus $\beta \in [0,1]$. Note that this is a conservative measure, as a gap does not imply that the site is heterozygous. Nevertheless, we can treat each site of a haplotype as a bit, and the above distance will later become useful when we formulate the haplotype inference problem as a decoding problem over a certain discrete channel.

We assume that for any given bit of a haplotype h, each symbol of the corresponding column of the SNP matrix is generated according to a probability distribution $P(y|x)$. Specifically, we can write $P(y|x)$ as a transition probability matrix given by:

$P(y\|x)$	$x=0$	$x=1$
$y=0$	$1-\alpha_{01}-\alpha_{02}$	α_{10}
$y=1$	α_{01}	$1-\alpha_{10}-\alpha_{12}$
$y=-$	α_{02}	α_{12}

If we further assume that $\alpha_{01}=\alpha_{10}=e_1$ and $\alpha_{02}=\alpha_{12}=e_2$, then the SNP fragments can be viewed as a haplotype sequence being transmitted repetitively over a binary symmetric erasure channel [MacKay, 2003]. Clearly, e_1 is the probability of bit flip, and e_2 is the probability of bit erasure.

Without knowing the statistical model $P(y|x)$, one can still try to find and correct the errors in the SNP data to reconstruct a maximally consistent pair of haplotypes by a certain parsimonious argument, such as minimum SNP removal or minimum error correction. However, we will provide convincing evidence later that the best solution corresponding to the minimum error correction criterion does not necessarily lead to the most likely pair of haplotypes even when e_1 and e_2 are reasonably small.

2.2 Benefits of Inferring Long Haplotype Sequences

We first discuss the issue of associating the SNP fragment with the correct origin of the haplotype sequence. Note that any incorrect association owing to the gap and reading error may result in the model mismatch when inferring the individual haplotype sequence based on the associated subset of the SNP fragments. Fortunately, when the haplotype sequence is fairly long, the probability of associating the SNP fragment to the incorrect haplotype sequence diminishes if the dissimilarity of the haplotype pair exceeds the reading error probability of DNA sequencing by a constant factor.

Claim 1: Assume that the SNP matrix M is obtained by transmitting m_1 times on each bit of h_1 and m_2 times on each bit of h_2 through the binary symmetric erasure channel ($m_1+m_2=m$). If $\beta > 4e_1$, then $P(\hat{\Theta}=\Theta) \to 1$ as $n \to \infty$.

Proof: Denote by y_i the i-th row of M_1 generated from h_1 and by z_j the j-th row of M_2 generated from h_2. Asymptotically, we have

$$d(y_i,h_1)/n \sim \mathcal{N}(e_1, e_1(1-e_1)/n),$$

$$d(z_j,h_2)/n \sim \mathcal{N}(e_1, e_1(1-e_1)/n)$$

for $i=1,...,m_1$, $j=1,...,m_2$. Thus

$$\frac{d(y_i,y_j)}{n} \le \frac{d(y_i,h_1)}{n} + \frac{d(y_j,h_1)}{n}$$

$$\sim \mathcal{N}\left(2e_1, \frac{2e_1(1-e_1)}{n}\right).$$

On the other hand, we have

$$\begin{aligned}\frac{d(y_i,z_j)}{n} &\geq \frac{d(h_1,h_2)}{n} - \left[\frac{d(y_i,h_1)}{n} + \frac{d(z_j,h_2)}{n}\right] \\ &\sim \mathcal{N}\left(\beta - 2e_1, \frac{2e_1(1-e_1)}{n}\right).\end{aligned}$$

As $n \to \infty$, it is clear that

$$P\left(\frac{d(y_i,y_j)}{n} \leq 2e_1, \frac{d(y_i,z_k)}{n} \geq 2e_1\right) \to 1.$$

Note that for $\forall i \neq j$ and $\forall k$, the condition for equality to hold has zero probability measure.
Claim 2: If $e_1 > 0$, then for any finite m_1, the ML estimate of the haplotype sequence yields $P(\hat{h}_1 \neq h_1) > 0$ and $P(\hat{h}_1 \neq h_1) \to 1$ as $n \to \infty$.
Proof: Without loss of generality, we assume that the haplotype sequence is independent and identically distributed, allowing us to focus on the decoding of each bit of the haplotype sequence based on the corresponding column of M_1. Assume that k_1 0s and k_2 1s are observed with $m_1 - k_1 - k_2$ erasures in a particular column of M_1. The log-posterior ratio is

$$\log\left(\frac{P(x=0|k_1,k_2)}{P(x=1|k_1,k_2)}\right) = \log\left(\frac{P(x=0)}{P(x=1)}\right) + (k_1-k_2)\log\left(\frac{1-e_1-e_2}{e_1}\right).$$

Assuming equal prior probability and $1-e_2 > 2e_1$, then the decision rule becomes declaring $x=0$ when $k_1 - k_2 > 0$, that is, $k_1 > k_2$ and $x=1$ when $k_1 < k_2$. The probability of decision error is given by

$$P(\hat{x} \neq x) = \sum_{k_1<k_2} \frac{m_1!}{k_1!k_2!(m_1-k_1-k_2)!}(q)^{k_1} e_1^{k_2} e_2^{m_1-k_1-k_2} > 0$$

where $q = 1-e_1-e_2$. Thus, the probability of correctly decoding an n-bit haplotype is

$$P(\hat{h}_1 = h_1) = (1 - P(\hat{x} \neq x))^n \to 0$$

as $n \to \infty$.

2.3 Performance Limit of using SNP Fragments

From the computational viewpoint, we want to find the most probable pair of haplotypes among all other alternatives given the SNP fragments. However, quantifying how accurate the reconstructed haplotypes are when reading errors and gaps are abundant in the SNP fragments is generally difficult. An important theoretical question is whether we can reconstruct the haplotype sequences with an arbitrary small error based on the existing DNA sequencing techniques that can only provide unreliable SNP fragments. In fact, it is closely related to a communication problem where one needs to transmit the bit streams of haplotype information through the unreliable sequencing channel. Thus, we will borrow the idea from the classical information theory to address the theoretical performance limit of the SNP fragment-based haplotype inference problem.
Claim 3: If $C = \log_2 3 - H(e_1,e_2,1-e_1-e_2) > 0$ and $m_1 > n/C$, then for a large enough n, there exists a procedure to generate the SNP matrix, such that $P(\hat{h}_1 \neq h_1)$ can be made arbitrarily small.
Proof: If we assume that a haplotype is an independent and identically distributed stochastic sequence with each bit having a distribution $P(x)$, then for a large n, there are about $2^{nH(x)}$ typical sequences. For each row of M, there are about $2^{nH(y)}$ typical sequences. As there are only $2^{nH(x,y)}$ jointly typical sequences, we can

see that not all pairs of the typical haplotype and the typical fragment sequence are jointly typical. For the discrete memoryless channel $P(y|x)$, the probability that any randomly chosen pair is jointly typical is about 2^{-nC}, where the channel capacity is [Cover & Thomas, 1991]

$$C = I(x;y) = \log_2 3 - H(e_1, e_2, 1-e_1-e_2).$$

Hence, for a fixed fragment sequence, we can consider about 2^{nC} such pairs before we are likely to come across a jointly typical pair. This suggests that there are about 2^{nC} distinguishable haplotypes we can handle with M_1. From Shannon's channel coding theorem [Cover & Thomas, 1991], if the communication rate is $n/m_1 < C$, then there exists a code book that by examining the joint typicality, the decoding error can be made arbitrarily small.

In summary, finding the correct partition Θ has an overwhelmingly large probability asymptotically if the two haplotype sequences are long and dissimilar enough compared with the reading error probability. However, as the standard SNP matrix M is constructed by a repetition code, which has zero rate asymptotically [MacKay, 2003], there exists a significant performance gap between the information content conveyed by the SNP fragments M and the actual information needed to make the decoding error probability arbitrarily small.

3 Efficient Haplotype Reconstruction with Genotype Information

3.1 Limitations of Haplotype Inference from Genotype Data

From a pair of haplotypes, the genotype information is relatively easy to obtain. As mentioned in Section 1, this has resulted in many computational methods for inferring the haplotypes from genotype data. However, genotype data alone can not fully resolve the haplotype sequences without additional biological insight. For example, if the observed genotype fragment is 0212, then there are two feasible explanations by the haplotype pairs given by $\{0110, 0011\}$ and $\{0010, 0111\}$. In general, for an individual genotype with T heterozygous sites, there are 2^{T-1} possible haplotype pairs that are consistent with the observation.

The Clark's method, which was proposed in [Clark, 1990] and was modified by many other followers, is a parsimony criterion. In essence, for a set of genotype sequences, one seeks the minimum number of haplotype pairs compatible with all the observations. This criterion is based on the fact that in natural populations, the number of distinct haplotypes is vastly smaller than the number of combinatorially possible haplotypes. The above resolution however assumes that there is no reading error in all genotype sequences, which can be unrealistic when inferring long haplotype sequences. In addition, multiple solutions are unavoidable when any site of all genotype sequences is heterozygous. Hence, it is naturally to consider jointly using SNP fragments and genotype data in the process of haplotyping. Genotype data is indeed complementary to the SNP fragments especially when both types of data are unreliable.

3.2 Sequential Haplotype Reconstruction Algorithm with both SNP Fragments and Genotype Data

Consider a long haplotype sequence of length n being measured by L aligned SNP fragments, each of which has length $s = n/L$. For each SNP block with s columns, we assume that the genotype information is available. For a genotype $g = [g_1 g_2 \ldots g_s]$ without reading error, when the i-th SNP site is wild-type homozygous, $g_i = 0$; when it is mutant-type homozygous, $g_i = 1$; and when it is heterozygous, $g_i = 2$. Clearly, a pair of identical haplotypes cannot yield a genotype sequence, with any site being 2 unless it is a reading error. If we screen each SNP block and observe 2 on certain sites, the information will be very useful to infer Θ. Thus, we can view the genotype information as some form of parity check [MacKay, 2003].

With the help of genotype information, we can design an efficient haplotype reconstruction algorithm that controls the decoding error of each site by sequentially taking an additional SNP block. The algorithm runs in two stages.
Given an initial SNP matrix $M = [B_1 B_2 ... B_s]$ with known bit flip probability e_1, erasure probability e_2, and the corresponding genotype sequence $\{g_1, ..., g_s\}$.
Output a pair of haplotypes h_1 and h_2 with error probability for each site below the desired level e.
Algorithm

- Partition: Start with an arbitrary initial partition $\Theta = (M_1, M_2)$. Identify all sites, where $g_i = 2$, and set $h^*_{1i} = 0$, $h^*_{2i} = 1$ if there are more 0s in the i-th column of M_1 than 1s in the i-th column of M_2. Otherwise, set $h^*_{1i} = 1$, $h^*_{2i} = 0$.

  For the remaining sites, set $h^*_{1j} = h^*_{2j} = g_j$. Use the majority vote through the corresponding column vector in M_1 to decide h^*_{1j} and the corresponding column vector in M_2 to decide h^*_{2j} when g_j is a gap.

  Cluster M into two groups, with the initial centers given by h^*_1 and h^*_2 and the generalized Hamming distance defined over each pair of row vectors in M. Standard k-means algorithm [MacKay, 2003] will converge in less than 10 iterations in practice.

- Decoding: For each column l in B_i ($i = 1, ..., s$), do the following.

  If $g_{il} = 0$, then count the number of 0s k_1 and the number of 1s k_2 in the whole column l. Declare $h_{il1} = h_{il2} = 0$ if $k_1 - k_2 > c_0$; declare $h_{il1} = h_{il2} = 1$ if $k_2 - k_1 > c_0$; and request one more piece of SNP block appending to B_i if $|k_1 - k_2| \leq c_0$.

  If $g_{il} = 1$, then count the number of 0s k_2 and the number of 1s k_1 in the whole column l. Declare $h_{il1} = h_{il2} = 1$ if $k_1 - k_2 > c_0$; declare $h_{il1} = h_{il2} = 0$ if $k_2 - k_1 > c_0$; and request one more piece of SNP block appending to B_i if $|k_1 - k_2| \leq c_0$.

  If $g_{il} = 2$, then count the number of 0s k_{11} and the number of 1s k_{12} in column l belonging to group 1, and count the number of 0s k_{21} and the number of 1s k_{22} in column l belonging to group 2. Declare $h_{il1} = 0$, $h_{il2} = 1$ if $k_{11} - k_{12} > c_1$ or $k_{22} - k_{21} > c_1$; declare $h_{il1} = 1$, $h_{il2} = 0$ if $k_{12} - k_{11} > c_1$ or $k_{22} - k_{21} > c_1$; otherwise, request one more piece of SNP block appending to B_i.

  If g_{il} is a gap, then count the number of 0s k_{11} and the number of 1s k_{12} in column l belonging to group 1, and count the number of 0s k_{21} and the number of 1s k_{22} in column l belonging to group 2. Declare $h_{il1} = 0$ if $k_{11} - k_{12} > c_1$; declare $h_{il2} = 1$ if $k_{22} - k_{21} > c_1$; declare $h_{il1} = 1$ if $k_{12} - k_{11} > c_1$; $h_{il2} = 0$ if $k_{21} - k_{22} > c_1$; otherwise, request one more piece of SNP block appending to B_i.

Note that the above algorithm is applicable to the case without genotype information by setting g_{il} to be gap $\forall l, i$.
Claim 4: Denote by p the probability that the genotype reading of an SNP site is correct. If $\beta > 4e_1$ or $\beta > 2(1-p)$ and

$$c_0 = \left\lceil \frac{\log\left(\frac{e(1-p)}{(1-e)p}\right)}{\log\left(\frac{(1-e_1-e_2)}{e_1}\right)} \right\rceil, c_1 = \left\lceil \frac{\log\left(\frac{e}{1-e}\right)}{\log\left(\frac{(1-e_1-e_2)}{e_1}\right)} \right\rceil,$$

then $P(\hat{h}_{jl} \neq h_{jl}) < e$, for $j = 1, 2$ and $l = 1, ..., n$, as $n \to \infty$.
Proof: Similar to the distance argument in the proof of Claim 1, we can show that $P(\hat{\Theta} = \Theta) \to 1$ as $n \to \infty$. Assuming the conditional independence between M and $\{g_1, ..., g_s\}$, the decoding rule for each haplotype sequence is the sequential probability ratio test (SPRT), with the upper and lower limit given by Wald's

fundamental approximation [Wald, 1947]. As the test statistic $(k_1 - k_2)$ is discrete, we have to increase the threshold to the nearest integer, which in principle reduces the decoding error, that is, $P(\hat{h}_{il\,j} \neq h_{il\,j}) < e$ with strict inequality $\forall i, l, j$. However, the expected number of samples to reach a decision will not be minimized with the exact error constraint e as the original SPRT, which requires randomized decision rule switching between the threshold c and $c-1$ [Wald, 1947].

4 Experiment on Synthetic Data

4.1 Comparison with the Parsimony-based Criterion

We consider one chromosome data set used in [Wang et al., 2005] for haplotype reconstruction through the MEC criterion. It contains a pair of haplotypes with a length of 95 after removing 8 missing sites. For the SNP matrix with fixed sample size $m = 40$, we generate random fragments with a 0.75 gap rate of fragments. Among the non-gap elements in the SNP matrix, the error rate is 0.3. The reconstruction rate given by

$$1 - \frac{\min\{r_{11} + r_{22}, r_{12} + r_{21}\}}{2n}$$

, where $r_{ij} = d(h_i, \hat{h}_j)$, $i = 1, 2$, and $j = 1, 2$, is used to evaluate algorithm performance. The expected number of errors that needs to be corrected is 285, while the MEC method using the branch-and-bound algorithm in [Wang et al., 2005] only needs to correct 215 errors to make the reconstructed haplotypes compatible with the SNP matrix. The reported reconstruction rate is 0.705 [Wang et al., 2005], while the expected decoding error rate is 0.188. Thus, the MEC method does not provide the best reconstruction rate. To apply the proposed sequential haplotype reconstruction algorithm, we set $L = 5$; thus, $s = 19$. The initial sample size is 10, and we set $c = 2$. The expected number of SNP fragments used for reconstruction is 32.7 in 10 Monte Carlo runs. Ideally, the Wald's SPRT only needs 30 samples on average to reach a decision, which is smaller than the fixed sample size $m = 40$. On the other hand, the reconstruction rate is 0.806, which is significantly higher than that by the MEC method. It is also close to the expected decoding error rate even in the above non-asymptotic regime.

Next, we consider that one has genotype information associated with the SNP fragments, with each SNP site having a gap probability of 0.25. For a non-gap genotype reading, the error probability is 0.1 ($p = 0.9$). We want to have $e < 0.01$ and end up with $c = 2$, as in the case without genotype information. In 10 Monte Carlo runs, the reconstruction rate becomes 0.998, which is higher than the desired rate. This is due to the conservative design of the SPRT procedure, wherein the actual bit error rate is lower than the desired level. If one seeks the minimum number of error corrections to make the reconstructed haplotypes compatible with the modified genotype, then the reconstruction rate is reduced to 0.924. Clearly, genotype information helps improve the reconstruction rate. Correcting the minimum number of errors however does not lead to optimal decoding due to the fact that typical errors will be unlikely minimal ones when the sequence length n and reading error e_1 are large enough. Similar observations have also been confirmed using other parsimony-based criteria. Parsimony-based methods, which are attractive mainly due to their general applicability without knowing the error rate, can be quite suboptimal compared with the best achievable reconstruction rate with the knowledge on the statistical model of the SNP fragments.

4.2 Asymptotic Error for Haplotype Inference with SNP Fragments and Genotype Data

To quantify the improvement by using genotype information, we consider an idealized scenario where $p \to 1$, and every SNP site in the genotype is heterozygous. In this case, one cannot infer any site of each haplotype

purely from the genotype.

Claim 5: Assume that the decoding error rate using the proposed sequential algorithm is e for large n without genotype information. The decoding error rate is then $O(e^2)$ when the genotype information is under the idealized scenario.

Proof: We assume that under both cases, the inference on Θ is perfect. Without loss of generality, consider decoding the j-th site of the haplotypes, where $h_{1j} = 0$ and $h_{2j} = 1$. Let k_{11} be the number of 0s of the j-th column in M_1 and k_{12} be the number of 1s of the j-th column in M_1. Similarly, let k_{21} and k_{22} be the corresponding number of 0s and 1s in M_2, respectively. Without haplotype information, an error will occur either when $k_{11} - k_{12} < -c$ or when $k_{22} - k_{21} < -c$. Both events have been designed to have probability smaller than e. With haplotype information, an error will occur when both $k_{11} - k_{12} < -c$ and $k_{22} - k_{21} \leq c$ are true or both $k_{11} - k_{12} \leq c$ and $k_{22} - k_{21} < -c$ are true. As under the same sample size $k_{22} - k_{21} > c$ implies the correct decoding probability to be at least $1 - e$, the event that $k_{22} - k_{21} \leq c$ is true prior to the stop time when $k_{11} - k_{12} < -c$ occurs has a probability of $O(e)$ at most. Thus, the overall decoding error rate is at most $O(e^2)$.

5 Concluding Summary

When SNP fragments from unreliable DNA sequencing technology are primarily used for haplotyping, one has to be cautious when applying any parsimonious criterion. We have shown that haplotype reconstruction based on aligned SNP fragments can be treated as decoding over a discrete memoryless channel. There exists a nontrivial performance gap between the error correction capability by parsimony-based methods and that given by the channel capacity. On the other hand, genotype information, if available, is very useful to improve the haplotype reconstruction accuracy. A new sequential haplotype reconstruction algorithm using genotype information has been proposed that guarantees the desired reconstruction rate with a smaller expected number of SNP fragments than what is needed with a fixed sample size. The advantage of using genotype information is theoretically quantified by exploiting a simplified statistical model with nearly perfect genotype information.

Acknowledgment

H. Chen was supported in part by the US Army Research Office under contract number W911NF-08-1-0409, the Louisiana Board of Regents NSF(2009)-PFUND-162, and the Office of Research and Sponsored Programs, the University of New Orleans. K. Zhang was supported by a NIH RCMI grant (1G12RR026260-01) and a Louisiana BOR award (LEQSF(2008-11)-RD-A-32). D. Zhu was supported by a NIH R21 grant (1R21LM010137). The authors gratefully acknowledge the stimulating discussions they had with Dr. Bin Fu of University of Texas – Pan American.

References

[Brinza & Zelikovsky, 2008] Brinza, D. & Zelikovsky, A. (2008). 2SNP: scalable phasing method for trios and unrelated individuals. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 5(2), 313–318.

[Chen et al., 2008] Chen, Z., Fu, B., Schweller, R., Yang, B., Zhao, Z., & Zhu, B. (2008). Linear time probabilistic algorithms for the singular haplotype reconstruction problem from snp fragments. *Journal of Computational Biology*, 15(5), 535–546.

[Cilibrasi et al., 2007] Cilibrasi, R., van Iersel, L., Kelk, S., & Tromp, J. (2007). The complexity of the single individual snp haplotyping problem. *Algorithmica*, 13(1), 13–36.

[Clark, 1990] Clark, A. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular biology and evolution*, 7(2), 111.

[Cover & Thomas, 1991] Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience.

[Dondi, 2009] Dondi, R. (2009). The Longest Haplotype Reconstruction Problem Revisited. In *Fundamentals of Computation Theory* (pp. 109–120).: Springer.

[Excoffier & Slatkin, 1995] Excoffier, L. & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular biology and evolution*, 12(5), 921.

[Genovese et al., 2008] Genovese, L., Geraci, F., & Pellegrini, M. (2008). SpeedHap: An accurate heuristic for the single individual SNP haplotyping problem with many gaps, high reading error rate and low coverage. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, (pp. 492–502).

[Gusfield, 2001] Gusfield, D. (2001). Inference of haplotypes from samples of diploid populations: Complexity and algorithms. *Journal of computational biology*, 8(3), 303–323.

[Kim & Misra, 2007] Kim, S. & Misra, A. (2007). Snp genotyping: Technologies and biomedical applications. *Annual Review of Biomedical Engineering*, 9(1), 289–320.

[Lancia et al., 2001] Lancia, G., Bafna, V., Istrail, S., Lippert, R., & Schwartz, R. (2001). SNPs problems, complexity, and algorithms. *AlgorithmsESA 2001*, (pp. 182–193).

[Liang & Wang, 2008] Liang, K. & Wang, X. (2008). A deterministic sequential monte carlo method for haplotype inference. *IEEE Journal Selected Topics in Signal Processing*, 2(3), 322–331.

[Lippert et al., 2002] Lippert, R., Schwartz, R., Lancia, G., & Istrail, S. (2002). Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Briefings in Bioinformatics*, 3(1), 23.

[MacKay, 2003] MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge Press.

[Niu et al., 2002] Niu, T., Qin, Z., Xu, X., & Liu, J. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 70(1), 157–169.

[Rizzi et al., 2002] Rizzi, R., Bafna, V., Istrail, S., & Lancia, G. (2002). Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem. *Algorithms in Bioinformatics*, (pp. 29–43).

[Wald, 1947] Wald, A. (1947). *Sequential Analysis*. New York: Wiley and Sons.

[Wang et al., 2005] Wang, R., Wu, L., Li, Z., & Zhang, X. (2005). Haplotype reconstruction from SNP fragments by minimum error correction. *Bioinformatics*, 21(10), 2456.