Haplotype Assembly: An Information Theoretic View

Hongbo Si, Haris Vikalo, and Sriram Vishwanath
Department of Electrical and Computer Engineering
The University of Texas at Austin
1 University Station, C0806, Austin, TX 78712
Email: hongbosi@utexas.edu, {hvikalo, sriram}@ece.utexas.edu

***Abstract*—This paper studies the haplotype assembly problem from an information theoretic perspective. A haplotype is a sequence of nucleotide bases on a chromosome, conveniently represented by a binary string, that differ from the bases in the corresponding positions on the other chromosome in a homologous pair. Information about the order of bases in a genome is readily inferred using short reads provided by high-throughput DNA sequencing technologies. Here, the recovery of the target pair of haplotype sequences using short reads is rephrased as a joint source-channel coding problem. Two messages, representing haplotypes and chromosome memberships of reads, are encoded and transmitted over an erasure channel, where the channel model reflects salient features of high-throughput sequencing. In the absence of sequencing noise, both the necessary and sufficient conditions are presented with order-wise optimal bounds for perfect haplotype recovery. A brief discussion of the erroneous scenario is also included in the paper.**

I. Introduction

Diploid organisms, including humans, have homologous pairs of chromosomes. One chromosome in a pair is inherited from the mother and the other from the father. The two chromosomes in a pair are similar and essentially carry the same type of information but are not identical. In particular, chromosomes in a pair differ at a small fraction of positions (i.e., loci). Such variations are referred to as single nucleotide polymorphisms (SNPs). A haplotype is the collection of SNPs on a single chromosome, of a chromosome pair that are associated with one another. It is believed that the knowledge of each haplotype for each individual in the species helps to understand the genetics of common diseases. However, direct measurement and identification of the whole haplotype is generally expensive and inefficient. Therefore, haplotype determination is more commonly and efficiently done by computational inference in practice. Data is generally provided in diploid form known as a genotype, where contributions from both chromosomes are conflated. Subsequently, the goal is to determine a good set of haplotypes that resolve a given collection of genotypes. This body of work has resulted in the development of haplotype inference algorithms from genotypes, also known as haplotype phasing [1].

Haplotype assembly (also called "individual haplotyping") is rapidly emerging as an alternative to haplotype phasing. In this approach, haplotypes are reconstructed from numerous short haploid segments of DNA. In next generation sequencing methods, the length of each fragment (called "read") is approximately 100 to 1000 base pairs [2]. This length is comparable to the distance between SNPs on chromosomes. To this end, a single read is statistically incapable of covering more than one variant site. Further, a single read is unable to furnish information that can aid in the determination of which chromosome the read is originally sampled from [3]. Therefore, the approach of paired-end sequencing [4], also known as mate-paired sequencing [5] is used. This process produces pairs of short reads that are derived from two different segments of a DNA sequence, where the insert size between one and the other is known. These mate-pairs make it possible to combine information on SNPs within the same haplotype over a long region, and to further assemble them together to recover the original haplotype. Fig. 1 illustrates the procedure of generating paired-end reads from original pair of chromosomes, where each read may cover two or more variant sites. The goal of haplotype assembly is to identify the correct source chromosome from which fragments are sampled, and to reconstruct both haplotype sequences. A fragment conflict graph [6] interpretation converts the original problem into partitioning a bipartite if the data is error free. Subsequently, for erroneous data, it formulates haplotyping into a optimization problem of minimizing the number of transformation steps to generate a bipartite graph [7].

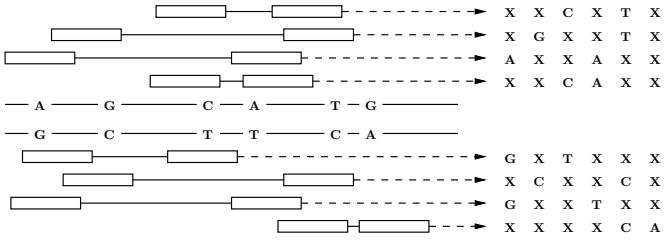

Fig. 1: Paired-end fragments from paired chromosomes. Rectangles linked by lines above and below the target chromosome pair represent paired-end reads, and their relative positions comparing to chromosome pair show their coverage regions.

In this paper, we develop an information theoretic understanding of haplotype assembly. In this process, we determine necessary and sufficient conditions for assembly, both in the absence and presence of noise. The rest of paper is organized as follows. The next section describes the formulation

of haplotype assembly problem. In Section III, we present an information theoretic view of haplotype assembly in the absence of sampling errors, and the erroneous case is discussed in Section IV. Finally, Section V concludes the paper.

II. Problem Formulation

A SNP is a single base pair position where nucleotides from both chromosomes differ from each other. Usually, there are only two variants at one SNP site in diploid organisms. The more frequent variant is referred to as the major allele (which we denote as 0), and the less frequent one as the minor allele (which we denote as 1). With this notation, a haplotype consisting of all information of SNP sites on one of the paired chromosome can be represented as a binary sequence. Specifically, we focus on a pair of target haplotypes with complementary relationship, i.e. $\boldsymbol{h} = \bar{\boldsymbol{h}}$, where

$$\boldsymbol{h} = (h_1, h_2, \ldots, h_n),$$

and n is the length of haplotypes.

Each paired-end read contains partial information about either of these two haplotypes. Consider an indicator random variable c_i, $i \in \{1, \ldots, m\}$ (m is the number of reads), which corresponds to the chromosome membership for read i. Here, c_i equals 0 if read i is sampled from $\boldsymbol{h}$ and equals 1 otherwise. Due to a limitation of read length, only a small fraction of entries are observed. In other words, a paired-end read could be considered as a sequence drawn from the alphabet $\{0, 1, \times\}$, where "$\times$" refers to a lack of information at this site due to the absence of coverage. Typically, only 2 entries are numerical in one paired-end read, if the effect of burst mutations is ignored. The collection of all reads forms a matrix $\boldsymbol{R}$, whose rows correspond to m paired-end reads, and whose columns correspond to n SNP sites. The ith row of $\boldsymbol{R}$ is denoted as $\boldsymbol{r}_i$, and the jth element of $\boldsymbol{r}_i$ is denoted as r_{ij}.

In the absence of sampling noise, every observed element in matrix $\boldsymbol{R}$ is obtained as an exclusive-OR (XOR) of the corresponding SNP and associated membership information. Formally, this relationship is given by

$$r_{ij} = h_j \oplus c_i. \tag{1}$$

At this point, the observed matrix $\boldsymbol{R}$ could be understood to be obtained from a rank 1 matrix $\boldsymbol{S}$, whose row is either $\boldsymbol{h}$ or $\bar{\boldsymbol{h}}$ based on the value of c_i, and most of entries are erased due to reading process. Hence, the task of haplotype assembly is to recover haplotype $\boldsymbol{h}$ and chromosome membership vector $\boldsymbol{c}$, or equivalently the matrix $\boldsymbol{S}$, from the observation matrix $\boldsymbol{R}$.

An example, illustrated by Fig. 1, corresponds to the scenario of 6 SNP sites and 8 paired-end reads. Since only the first 4 reads are (shotgun) sequenced from chromosome 1, we obtain the chromosome membership vector $\boldsymbol{c} = (0,0,0,0,1,1,1,1)$. If denoting the haplotype from chromosome 1 as $\boldsymbol{h} = (1,1,0,1,0,0)$, then the observed reads

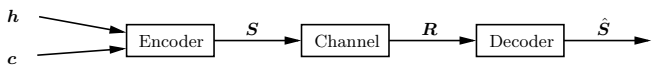

Fig. 2: Information theoretic view of the haplotype assembly problem.

matrix, without the influence of error, is given by

$$\boldsymbol{R} = \begin{bmatrix} \times & \times & 0 & \times & 0 & \times \\ \times & 1 & \times & \times & 0 & \times \\ 1 & \times & \times & 1 & \times & \times \\ \times & \times & 0 & 1 & \times & \times \\ 0 & \times & 1 & \times & \times & \times \\ \times & 0 & \times & \times & 1 & \times \\ 0 & \times & \times & 0 & \times & \times \\ \times & \times & \times & \times & 1 & 1 \end{bmatrix}. \tag{2}$$

III. Information Theoretic View

From a joint source-channel coding perspective, haplotype assembly comprises of aiming to recover two sources being communicated through an erasure channel (see Fig. 2). The first source is haplotype information, $\boldsymbol{h}$, and the second source is the chromosome membership vector $\boldsymbol{c}$. Both these vectors are assumed to originate from a uniform distribution, i.e. each entry obeys Bernoulli distribution with parameter $1/2$. These two sources are encoded jointly using the function: $f : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}^{m \times n}$ such that the encoded codeword $\boldsymbol{S} = f(\boldsymbol{h}, \boldsymbol{c})$. In particular, each entry in $\boldsymbol{S}$ is given by $s_{ij} = h_j \oplus c_i$, which implies the encoder is a bijection. After receiving the output from channel, the decoder uses the decoding function to map its channel observations into an estimate of the message. Specifically, we consider the decoder (corresponds to a recovery algorithm for haplotype assembly) given by $g : \{0,1\}^{m \times n} \to \{0,1,\times\}^{m \times n}$, such that $\hat{\boldsymbol{S}} = g(\boldsymbol{R})$, where $\hat{\boldsymbol{S}}$ represents the estimate. We define the error probability of decoding as

$$P_{\mathrm{e}} \triangleq \Pr\{\hat{\boldsymbol{S}} \neq \boldsymbol{S} | \boldsymbol{R}\}. \tag{3}$$

As in conventional analysis, we consider this probability over all possible choices of matrix $\boldsymbol{S}$ (denote the assemble as $\mathcal{S}$), and desire m to be large enough such that there exists at least one decoding function g with small probability of error.

The channel model reflects particular reading technique. More precisely, for paired-end reading, the channel $W : \{0,1\}^{m \times n} \to \{0,1,\times\}^{m \times n}$ considered for the moment is described as follows.

1) Erasures happen independently across rows.
2) In each row, only 2 entries remain and their positions are random.

To this end, we observe only 2 entries of each row from $\boldsymbol{S}$, and observations are independent across different rows. A straightforward inference from this channel model is that for each column of $\boldsymbol{R}$, the number of remaining entries approximately obeys Poisson distribution, which is consistent with the basic observation of paired-end reading. Another point is the expected length of insert size between these 2

sampled entries within a row is given by $(n-2)/3$, which also accords with the practical consideration that this value could not be made arbitrarily large due to the limitation of reading technology.

Based on this model for haplotype assembly, we consider the necessary and sufficient conditions for recovery.

Theorem 1. *Given* 2 *arbitrary observations in each row, the original haplotype matrix* $\boldsymbol{S}$ *could be reconstructed only if the number of reads satisfies*

$$m = \Omega(n),$$

where n *is the length of target haplotype. Moreover, if* $m = \Theta(n \log n)$*, a reconstruction algorithm, erasure decoding, could determine* $\boldsymbol{S}$ *accurately with high probability. Specifically, given a target small constant* ε*, there exists an* n *large enough such that by choosing* $m = \Theta(n \ln n)$ *the probability of error* $P_e \leq \varepsilon$*.*

A. Necessary Condition for Recovery

Using Fano's inequality [8], we find that:

$$H(\boldsymbol{S}|\boldsymbol{R}) \leq P_{\mathrm{e}} \log |\mathcal{S}| \leq P_{\mathrm{e}}(m+n), \tag{4}$$

where $\mathcal{S}$ is the assemble of all possible $\boldsymbol{S}$, and its size is upper bounded by 2^{m+n}.

Denote the matrix $\boldsymbol{T}$ as indicators of locations where $\boldsymbol{S}$ is observed, i.e. $t_{ij} = 1$ if $r_{ij} \neq \times$, and $t_{ij} = 0$ otherwise. Then, $\boldsymbol{T}$ is independent of $\boldsymbol{S}$, and its rows are independent due to our channel assumption. Therefore, we have

$$\begin{aligned}
H(\boldsymbol{S}) &\overset{(a)}{=} H(\boldsymbol{S}|\boldsymbol{T}) \\
&= I(\boldsymbol{S};\boldsymbol{R}|\boldsymbol{T}) + H(\boldsymbol{S}|\boldsymbol{T},\boldsymbol{R}) \\
&= I(\boldsymbol{S};\boldsymbol{R}|\boldsymbol{T}) + H(\boldsymbol{S}|\boldsymbol{R}) \\
&\overset{(b)}{\leq} I(\boldsymbol{S};\boldsymbol{R}|\boldsymbol{T}) + P_{\mathrm{e}}(m+n) \\
&= H(\boldsymbol{R}|\boldsymbol{T}) - H(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{T}) + P_{\mathrm{e}}(m+n) \\
&\overset{(c)}{=} H(\boldsymbol{R}|\boldsymbol{T}) + P_{\mathrm{e}}(m+n) \\
&\overset{(d)}{=} \sum_{i=1}^{m} H(\boldsymbol{r}_i|\boldsymbol{t}_i) + P_{\mathrm{e}}(m+n) \\
&\overset{(e)}{=} 2m + P_{\mathrm{e}}(m+n)
\end{aligned}$$

where (a) follows from independence between $\boldsymbol{S}$ and $\boldsymbol{T}$; (b) from Fano's inequality, i.e. equation (4); (c) from the fact $\boldsymbol{R}$ is deterministic if $\boldsymbol{S}$ and $\boldsymbol{T}$ are both known; (d) from the row independence assumption of our channel model; (e) from the assumption that every row has exactly 2 entries observed.

Finally, by noting that $H(\boldsymbol{S}) = m+n$, we need

$$m \geq \frac{(1-P_{\mathrm{e}})n}{1+P_{\mathrm{e}}}. \tag{5}$$

for accurate recovery. More precisely, roughly we need $m = \Omega(n)$ for recovery with arbitrary small probability of decoding error.

Remark 2. *Note that in this proof, channel model is only utilized when bounding* $H(\boldsymbol{R}|\boldsymbol{T})$*. To this end, similar proof could be generalized to more types of channel models. For instance, deterministic choice of reading sites, and paired-end reading with fixed insert size. As long as the number of observed entries is sparse, more precisely in proportion to n, the lower bound* $m = \Omega(n)$ *still holds by tracing similar steps in this proof.*

B. Sufficient Condition for Recovery

The target of a decoding algorithm is to recover $\boldsymbol{S}$ (or equivalently $\boldsymbol{h}$ and $\boldsymbol{c}$) from $\boldsymbol{R}$ with high confidence. Here, we show a simple and effective algorithm, called "erasure decoding", for haplotype assembly. Detailed steps of this algorithm are described as follows:

1) Choose the "seed" s as arbitrary non-erased entry in the first row, i.e. $s = r_{1j}$, where j is randomly chosen such that $r_{1j} \neq \times$. Evaluate the membership of first row as $c_1 = 0$.
2) Find all other rows with position j not erased, i.e.
$$\mathcal{A} = \{k | r_{kj} \neq \times, k \neq 1\}. \tag{6}$$
3) Evaluate the membership of all rows with indices in $\mathcal{A}$ as
$$c_k = \begin{cases} 0, & \text{if } r_{kj} = r_{1j}, \\ 1, & \text{otherwise}, \end{cases} \tag{7}$$
for every $k \in \mathcal{A}$.
4) Decode SNPs in the first row by
$$r_{1l} = r_{kl} \oplus c_k, \tag{8}$$
for every $k \in \mathcal{A}$ and $r_{kl} \neq \times$.
5) Delete all rows with indices in $\mathcal{A}$.
6) Arbitrarily choose another non-erased entry in the first row as the new seed $s = r_{1j}$, which has not been chosen as seed in any of the former steps. Repeat Step 2) to 6) until no row could be further erased.
7) If the first row is the only remaining one and its entries are all decoded, claim $\boldsymbol{h} = \boldsymbol{r}_1$; otherwise claim a failure.

Remark 3. *In this algorithm, we arbitrarily evaluate a chromosome membership for the first row, but it may not be the correct one. In fact, if the algorithm successfully decodes both* $\boldsymbol{h}$ *and* $\boldsymbol{c}$*, then all elements could be flipped due to an incorrect choice of initial membership. However, the recovered matrix* $\boldsymbol{S}$ *remains the same, due to the particular exclusive-OR operation to generate* $\boldsymbol{S}$*. At this point, the choice of initial membership does not influence the decoding performance.*

Remark 4. *Erasure decoding is closely connected to the bipartite partition interpretation [3]. Note that if our algorithm successfully recovers the message matrix* $\boldsymbol{S}$*, we can realign its rows such that the matrix could be partitioned into two sub-matrices with different chromosome memberships. To this end, the erasure decoding provides a practical algorithm to fulfill partition a bipartite for haplotype assembly.*

Fig. 3 shows the details of decoding procedures for the example illustrated in Fig. 1, where the read matrix is given by (2).

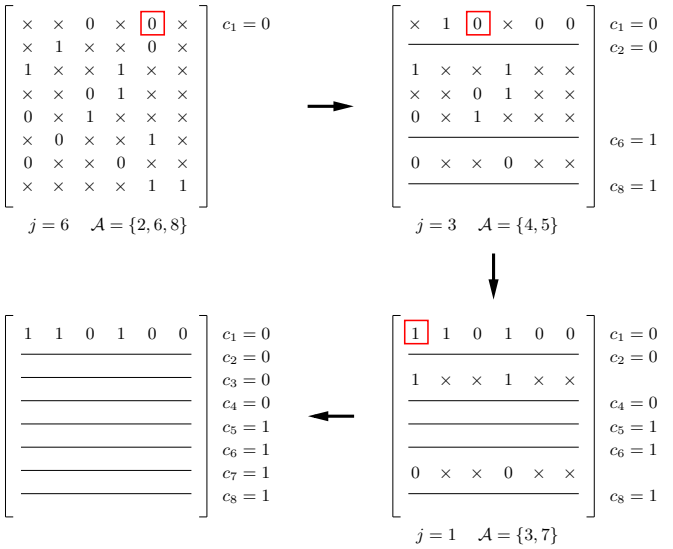

Fig. 3: Erasure decoding of the example illustrated in Fig. 1. In every round (Step 2) to 6)), the seed is marked in a rectangle, with its column index given by j. Rows that share the same positions observed as the seed are collected in assemble $\mathcal{A}$. A straight line crossing a whole row of the matrix represents a deletion.

Here, we analyze the performance of this proposed algorithm. More precisely, we show that if the number of reads sample is large enough, i.e. $m=\Theta(n\ln n)$, the source matrix $\boldsymbol{S}$ could be recovered correctly with high probability. Observe that in the absence of sampling errors, the erasure decoding algorithm ensures the output to be the correct haplotype if both of the following conditions are satisfied.

1) All rows except for the first one are deleted.
2) All entries in the first row are decoded

At this point, decoding error occurs if at least one of the following events happen.

1) The event E_1: at least one of the columns in $\boldsymbol{R}$ are erased, such that the corresponding SNP could not be decoded;
2) The event E_2: there exist a partition of row indices $\{1,\ldots,m\}=\mathcal{U}_1\cup\mathcal{U}_2$, and a partition of column indices $\{1,\ldots,n\}=\mathcal{V}_1\cup\mathcal{V}_2$, such that $|\mathcal{V}_1|\geq 2$ and $|\mathcal{V}_2|\geq 2$ (to make sure 2 entries could be sampled from each row), and $r_{ij}=\times$ for any $(i,j)\in(\mathcal{U}_1\times\mathcal{V}_2)\cup(\mathcal{U}_2\times\mathcal{V}_1)$. In other words, the sampled entries could be considered as originated from two disjoint subsets of target haplotypes, then, there is no hope to recover due to the lack of information bridging these subsets.

Here gives the details to bound the probability of each error events. First, note that by coupon collector effect, if $m=\Theta(n\log n)$, every column is at least covered by a read with high probability. More precisely, by taking $m=n\ln n$, the error event (or equivalently the tail distribution for coupon collector problem) is given by

$$\begin{aligned}\Pr\{E_1\}&=\frac{\sum_{i=1}^{n-2}\binom{n}{i}\binom{n-i}{2}^m}{\binom{n}{2}^m}\\&=\sum_{i=1}^{n-2}\binom{n}{i}\left[\frac{(n-i)(n-i-1)}{n(n-1)}\right]^m\\&\leq\sum_{i=1}^{n-2}n^i e^{-m\frac{2in-i(i+1)}{n(n-1)}}\\&=\sum_{i=1}^{n-2}O(n^{-i})\\&=O(n^{-1}).\end{aligned}\tag{9}$$

On the other hand, the second error event E_2 could be further decomposed into sub-events $E_2^{u,v}$, which represents the type 2 error event with particular $u=|\mathcal{U}_1|$ and $v=|\mathcal{V}_1|$. Then, we have

$$\Pr\{E_2^{u,v}\}=\frac{\binom{n}{v}\binom{m}{u}\binom{v}{2}^u\binom{n-v}{2}^{m-u}}{\binom{n}{2}^m}.\tag{10}$$

Observe that right hand side of (10) is maximized by two extreme points on the feasible $(u,v)-$region, i.e. for any u and v, $\Pr\{E_2^{u,v}\}\leq\Pr\{E_2^{1,2}\}=\Pr\{E_2^{m-1,n-2}\}$. In particular, we have

$$\begin{aligned}\Pr\{E_2^{1,2}\}&=\frac{\binom{n}{2}\binom{m}{1}\binom{2}{2}^1\binom{n-2}{2}^{m-1}}{\binom{n}{2}^m}\\&=\frac{m[(n-2)(n-3)]^{m-1}}{[n(n-1)]^{m-1}}\\&\leq n\ln n\left(1-\frac{4n-6}{n(n-1)}\right)^{n\ln n-1}\\&\leq n\ln n e^{-\frac{4n-6}{n(n-1)}(n\ln n-1)}\\&=O(n^{-3}\ln n).\end{aligned}$$

Hence, the probability of the second error event is upper bounded by

$$\begin{aligned}\Pr\{E_2\}&=\sum_{u=1}^{m-1}\sum_{v=2}^{n-2}\Pr\{E_2^{u,v}\}\\&\leq(m-2)(n-4)\Pr\{E_2^{1,2}\}\\&\leq n^2\ln n O(n^{-3}\ln n)\\&=O(n^{-1}(\ln n)^2).\end{aligned}\tag{11}$$

Combining these two pieces together, we obtain

$$P_{\mathrm{e}}\leq\Pr\{E_1\}+\Pr\{E_2\}=O(n^{-1})+O(n^{-1}(\ln n)^2)<\varepsilon,$$

for arbitrary $\varepsilon>0$ with large enough n.

Remark 5. *Note that there is a log-factor gap between the lower and upper bounds. As analyzed in [9], this log-factor ensures enough entries sampled from each column for accurate recovery. If a more systematic reading method could be adopted to generate the observation matrix, this log-factor may not be essential for reconstruction.*

Remark 6. *An alternative interpretation for quantifying the minimum number of entries needed to recover a rank-*1 *matrix is using optimality. In this branch of work [10] [11] [12], an optimization approach is utilized to determine the necessary condition, and the recovery is claimed to be possible by solving this convex program.*

IV. Discussion of Erroneous Case

When sequencing error happens, the binary entry of $\boldsymbol{R}$ is flipped. Here, we assume errors are independent and identical. More precisely, from the view of information theory, the generation of errors could be modeled as messages passing through a set of independent binary symmetric channels with parameter p, where p is the probability of flipping. To this end, denoting the noises as a matrix $\boldsymbol{N}$, where n_{ij} are i.i.d. $\mathrm{Ber}(p)$ distributed, then the final observed matrix

$$\tilde{\boldsymbol{R}} = \boldsymbol{R} \oplus \boldsymbol{N}. \tag{12}$$

Hence, the system model for the erroneous case could be considered as the one for error-free case cascated with binary symmetric channels. Then, for perfect recovery, we want to reconstruct $\boldsymbol{S}$ from $\tilde{\boldsymbol{R}}$ with high probability. More precisely, if denoting the estimate from a recovery algorithm as $\hat{\boldsymbol{S}}$, we define the probability of error as

$$P_{\mathrm{e}} = \Pr\{\hat{\boldsymbol{S}} \neq \boldsymbol{S} | \tilde{\boldsymbol{R}}\},$$

to evaluate the recovery accuracy. We desire this probability to be arbitrary small, on an average across all possible implementation of $\boldsymbol{S}$.

From the perspective of necessary condition, Fano's equality still holds in this case, i.e.

$$H(\boldsymbol{S}|\tilde{\boldsymbol{R}}) \leq P_{\mathrm{e}}(m+n).$$

Using this, we obtain

$$H(\boldsymbol{S}) \leq H(\tilde{\boldsymbol{R}}|\boldsymbol{T}) - H(\tilde{\boldsymbol{R}}|\boldsymbol{S},\boldsymbol{T}) + P_{\mathrm{e}}(m+n).$$

In this case, $H(\tilde{\boldsymbol{R}}|\boldsymbol{S},\boldsymbol{T})$ does not vanish due to the influence of noise. In particular, by noting noises are assumed to be i.i.d., we have

$$H(\tilde{\boldsymbol{R}}|\boldsymbol{S},\boldsymbol{T}) = \sum_{i=1}^{m} H(\tilde{\boldsymbol{r}}_i|\boldsymbol{s}_i,\boldsymbol{t}_i) = 2mH(p).$$

Combining with the observations that $H(\tilde{\boldsymbol{R}}|\boldsymbol{T}) = 2m$ and $H(\boldsymbol{S}) = m+n$, we have

$$(1 + P_{\mathrm{e}} - 2H(p))m \geq (1 - P_{\mathrm{e}})n.$$

Thus, in order to obtain arbitrary small P_{e}, we need $H(p) < 1/2$ for further requirement of noise, otherwise there where is no hope for perfect recovery. Then, it is evident to have

$$m \geq \frac{(1-P_{\mathrm{e}})n}{1+P_{\mathrm{e}}-2H(p)}, \tag{13}$$

which is still an $m = \Omega(n)$ scale lower bound, and this bound is consistent with error-free sampling by letting $p = 0$.

On the other hand, from the perspective of sufficient condition, if errors happen, erasure decoding algorithm may not apply. In fact, an effective algorithm with small number of reads for haplotype assembly remains open. Most algorithms in use are based on an optimization formulation, by adopting different objective criteria [6] [7], and their algorithms basically consider the number of reads as a known parameter for analysis [13], rather than regarding it as the essential measurement to argue sufficient condition for haplotyping.

V. Conclusion

In this paper, we consider the haplotype assembly problem from an information theoretic perspective. In order to determine the chromosome membership and reconstruct paired haplotypes, we consider them as messages to encode and transmit over a particular erasure channel. This channel model reflects the characters of paired-end reading, such that every row observes only two entries with random site positions. In the case of error-free sampling, we show that the necessary condition for the number of reads to reconstruct is at least the same order of the length of haplotypes, and the erasure decoding algorithm ensures to implement reconstruction, with the optimal order regardless of a log-factor gap. The necessary condition for erroneous sampling case is analogue, and the sufficient condition remains to be an open problem.

References

[1] B. V. Halldórsson, V. Bafna, N. Edwards, R. Lippert, S. Yooseph, and S. Istrail, "Combinatorial problems arising in SNP and haplotype analysis," in *Discrete Mathematics and Theoretical Computer Science*. Springer, 2003, pp. 26–47.

[2] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature*, vol. 200, no. 8, 2007.

[3] R. Schwartz, "Theory and algorithms for the haplotype assembly problem," *Communications in Information & Systems*, vol. 10, no. 1, pp. 23–38, 2010.

[4] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy *et al.*, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nature genetics*, vol. 40, no. 6, pp. 722–729, 2008.

[5] A. Edwards, H. Voss, P. Rice, A. Civitello, J. Stegemann, C. Schwager, J. Zimmermann, H. Erfle, C. T. Caskey, and W. Ansorge, "Automated DNA sequencing of the human HPRT locus," *Genomics*, vol. 6, no. 4, pp. 593–608, 1990.

[6] G. Lancia, V. Bafna, S. Istrail, R. Lippert, and R. Schwartz, "SNPs problems, complexity, and algorithms," in *Algorithms-ESA 2001*. Springer, 2001, pp. 182–193.

[7] R. Lippert, R. Schwartz, G. Lancia, and S. Istrail, "Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem," *Briefings in bioinformatics*, vol. 3, no. 1, pp. 23–31, 2002.

[8] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1991.

[9] S. Vishwanath, "Information theoretic bounds for low-rank matrix completion," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1508–1512.

[10] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *Information Theory, IEEE Transactions on*, vol. 56, no. 5, pp. 2053–2080, 2010.

[11] B. Recht, "A simpler approach to matrix completion," *The Journal of Machine Learning Research*, vol. 7, pp. 3413–3430, 2011.

[12] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in neural information processing systems*, 2009, pp. 2080–2088.

[13] R. Rizzi, V. Bafna, and G. Lancia, "Practical algorithms and fixed-parameter tractablility for the single individual SNP haplotyping problem." in *Proc. of the Workshop on Algorithms in Bioinformatics*, 2002, pp. 29–43.