A PROBABILISTIC ANALYSIS OF SHOTGUN SEQUENCING FOR METAGENOMICS

MARLEE HERRING

Project Advisor: Kevin McGoff

Abstract. Genome sequencing is the basis for many modern biological and medicinal studies. With recent technological advances, metagenomics has become a problem of interest. This problem entails the analysis and reconstruction of multiple DNA sequences from different sources. Shotgun genome sequencing works by breaking up long DNA sequences into shorter segments called reads. Given this collection of reads, one would like to reconstruct the original collection of DNA sequences. For experimental design in metagenomics, it is important to understand how the minimal read length necessary for reliable reconstruction depends on the number and characteristics of the genomes involved. Utilizing simple probabilistic models for each DNA sequence, we analyze the identifiability of collections of M genomes of length N in an asymptotic regime in which N tends to infinity and M may grow with N. Our first main result provides a threshold in terms of M and N so that if the read length exceeds the threshold, then a simple greedy algorithm successfully reconstructs the full collection of genomes with probability tending to one. Our second main result establishes a lower threshold in terms of M and N such that if the read length is shorter than the threshold, then reconstruction of the full collection of genomes is impossible with probability tending to one.

1. Introduction

Genome sequencing is the basis for many modern biological and medicinal studies. Recently, the study of metagenomics has risen to the forefront of DNA sequencing technologies. Metagenomics is the study of communities of genomes, meaning that multiple sequences from different species are analyzed at once. For example, metagenomics is often used to determine bacterial populations within the human body, as well as in the soil, and it is important for understanding the overall health of many biological systems.

Shotgun genome sequencing works by breaking up long DNA sequences into shorter reads. This set of reads is then separated, identified, and reconstructed (see Figure 1). To accomplish these tasks, there is a simple greedy algorithm that works by choosing a random read and assigning overlap values to each of the other reads. The reads with the greatest overlap values are combined, picking those that are tied at random. This process is repeated until all reads have been combined. This algorithm has been previously analyzed in the context of the reconstruction of a single genome [3], and we note that there are other more sophisticated algorithms used in practice (see [1] for further discussion of other algorithms). We also note that shotgun reconstruction problems can be generalized to other settings, including graphs [2] and groups [4].

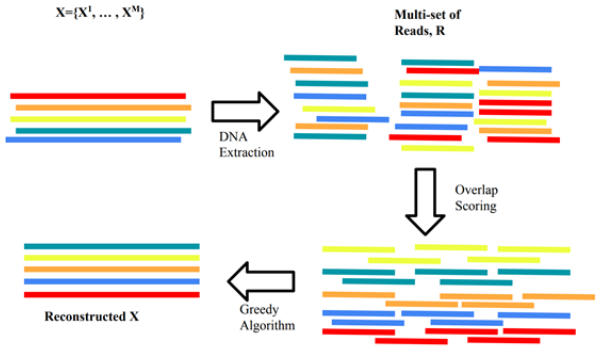

FIGURE 1. The process of shotgun sequencing for metagenomics. A collection of DNA sequences (upper left) is sampled, producing a multi-set of reads (upper right). Then the overlap properties of the reads are used to assemble them into longer segments (lower right), eventually resulting an a reconstructed collection of sequences (lower left).

The present paper contains three main contributions. First, we formulate the shotgun genome sequencing problem within the context of metagenomics. Second, we establish sufficient conditions on the biological parameters under which reconstruction is almost certainly successful (Theorem 1). Lastly, we establish conditions on the parameters that guarantee that successful reconstruction is almost certainly impossible (Theorem 2).

The rest of this paper is structured as follows. In Section 2 we define the metagenomics problem in precise mathematical language and present our main results. In Section 3 we introduce the preliminary results needed to proceed with the proofs of the main results, which will be provided in sections 4 and 5. Section 6 discusses open questions and possible directions for future work.

2. Problem formulation and main results

2.1. **Probabilistic formulation of metagenomics.** The purpose of this section is to introduce the metagenomics problem and our probabilistic model. In this context, we would like to do a blind reconstruction of a set of reads from multiple genomes and find the parameters that make reliable reconstruction possible.

Let $\mathcal{A}$ be the finite alphabet of all possible nucleotides (often abbreviated by A, C, G, and T). We assume that there are M different DNA sequences in the sample, denoted by $X = (X^1, \dots, X^M)$. Furthermore, we assume that each genome X^m is a sequence of N symbols from $\mathcal{A}$, denoted by $X^m = x_1^m \dots x_N^m \in \mathcal{A}^N$. Next we suppose that the shotgun sequencing technology provides us access to the multi-set $\mathcal{R} = \mathcal{R}(X)$ of all possible reads of length L from X:

$$\mathcal{R}(X) = \left\{ x_n^m \dots x_{n+L-1}^m \in \mathcal{A}^L \mid 1 \le m \le M,\ 1 \le n \le N - L + 1 \right\}. \tag{2.1}$$

The goal of shotgun metagenomics sequencing is to carry out a blind reconstruction of X given access to the multi-set of reads. Successful reconstruction is only possible if $\mathcal{R}$ is identifiable, meaning that there is only one collection of genomes (up to a permutation of the labels) that gives rise to the given read-set $\mathcal{R}$. It is important

to note that it is impossible to determine the precise order of genomes within the sample, and therefore it is more accurate to say that we are seeking to reconstruct the multi-set of genomes $\{X^1, \ldots, X^M\}$ from the sample. In order to state this idea precisely, let S_M be the group of permutations of the indices $\{1, \ldots, M\}$, let $\sigma \in S_M$, and let $\sigma(X) = (X^{\sigma(1)}, \ldots, X^{\sigma(M)})$. Then the equivalence class of X is denoted by $[X] = \{\sigma(X) : \sigma \in S_M\}$. Finally, we say that X is *identifiable* if $\mathcal{R}(X) = \mathcal{R}(X')$ implies that $X' \in [X]$.

The main question of this work is as follows: if a collection of genomes is drawn at random, then what is the probability that it will be identifiable? When approaching this problem, we suppose that the sample of genomes X is drawn randomly from some probability distribution. We then analyze the probability of reconstructing X from its multi-set of reads, $\mathcal{R}(X)$, as the genome length N tends to infinity. To construct our probabilistic model, we consider a collection $P = (p^1, \ldots, p^M)$ of M probability distributions on the finite set $\mathcal{A}$ of possible base pairs. We assume that each genome X^m is composed of N symbols drawn in an IID fashion with distribution p^m, and we suppose that all the sequences in X are generated independently of each other. We'll refer to the tuple (M, N, L, P) as specifying a *metagenomics problem*. In this context, we let I be the event that X is identifiable.

Within this work, we denote sequences of metagenomics problems as follows. For every $n \in N$, we assume that we take reads of length L_n from M_n genomes of length N_n that are constructed according to the probability distributions in $P_n = (p_n^1, \ldots, p_n^{M_n})$. We consider the probability of reliable reconstruction in terms of the asymptotic behavior of the above parameters. We let I_n denote the identifiability event for the problem (M_n, N_n, L_n, P_n).

We recognize that the model given here can be generalized and made more realistic in a variety of ways (see Section 6), but we believe it represents a necessary first step in the mathematical analysis of metagenomics problems.

2.2. **Identifiability result.** In this section, we give a sufficient condition for reliable reconstruction of X with high probability. Our sufficient condition states that the read length must be long enough relative to the number of genomes, the length of the genomes, and the probability distributions involved. Before we state the result, we require the definition of the second order Rényi entropy: for a probability distribution p on $\mathcal{A}$, we have

$$H_2(p) = -\log \sum_{a \in \mathcal{A}} (p(a))^2. \tag{2.2}$$

We may now state our first main result.

Theorem 1. *Suppose $\{(M_n, N_n, L_n, P_n)\}_{n=1}^{\infty}$ is a sequence of metagenomics problems such that*

(1) $\lim_{n\to\infty}(N_n) = \infty$,
(2) there exists $H_* > 0$ such that for all n and all $p_n^m \in P_n$ we have $H_2(p_n^m) \geq H_*$, and
(3) there exists $\epsilon > 0$ such that for all large n, we have

$$L_n \geq \frac{2(1+\epsilon)}{H_*} \log(M_n N_n).$$

Then

$$\lim_{n\to\infty} \mathbb{P}(I_n) = 1.$$

Roughly speaking, we interpret Theorem 1 to mean that by taking reads of length greater than $\frac{2}{H_*}\log(M_nN_n)$, the probability of identifiability tends to one as N_n tends to infinity. When $M_n = 1$ for all n, we recover the result of [3] up to a factor of 2. Furthermore, our proof shows that, under the hypotheses of the theorem, a simple greedy algorithm will reliably reconstruct X given $\mathcal{R}(X)$, with probability tending to one.

Observe that by rearranging the threshold for reconstruction to solve for M, we see that the number of sequences in the sample should satisfy

$$M \leq \frac{1}{N} e^{LH_*/2}. \tag{2.3}$$

Interpreting Theorem 1 from this point of view, we see that the number of sequences in the sample should not be too large relative to the length of the genomes, reads, and probability distributions.

2.3. **Non-identifiability result.** In this section, we present our main result on non-identifiability. Indeed, we give a threshold in terms of M_n, N_n, and P_n such that if L_n is less than the threshold, then the probability of identifiability tends to zero. Before stating this result, we define a type of cross-entropy between distributions: for two probability distributions p and p' on $\mathcal{A}$, let

$$F(p, p') = -\log\left(\sum_{a\in\mathcal{A}} p(a)\cdot p'(a)\right).$$

Now we are ready to state our second main result.

Theorem 2. *Suppose that $\{(M_n, N_n, L_n, P_n)\}_{n=1}^{\infty}$ is a sequence of metagenomics problems such that*

1. *for all n, we have $M_n \geq 2$,*
2. *$\lim_{n\to\infty}(N_n) = \infty$,*
3. *$\lim_{n\to\infty}\log(M_n)/N_n = 0$.*
4. *there exists $F_*, F^* \in (0,\infty)$ such that for all n and all $1 \leq m, m' \leq M_n$, we have $F_* \leq F(p_n^m, p_n^{m'}) \leq F^*$,*
5. *there exists $\delta \in (0,1)$ such that $L_n \leq \delta N_n$ for all large enough n, and*
6. *letting $C = 1/(2F^* - F_*)$, there exists $\epsilon > 0$ such that for all large enough n, we have*

$$L_n \leq C(1-\epsilon)\log(M_nN_n).$$

Then

$$\lim_{n\to\infty} P(I_n) = 0.$$

Let us discuss each of the hypotheses in this result. In (1) we simply assume that we are truly in the metagenomics scenario: indeed, the case $M_n = 1$ is the standard shotgun sequence problem with a single genome, for which bounds of this type already exist [3]. With hypothesis (2) we assume that we are in the asymptotic regime in which the length of the genomes tends to infinity, and in hypothesis (3) we assume that the number of genomes is subexponential in N_n. In (4) we assume some uniform bounds on the cross-entropy of the generating distributions, and in (5) we assume that the ratio of the read length to the genome length is bounded away from one. We view (6) as the key hypothesis of the theorem, as it gives a bound on the read length in terms of the number and length of the genomes involved. In short, we interpret Theorem 2 to mean that by taking reads

of length $L_n \leq C(1-\epsilon)\log(M_n N_n)$, the probability of identifiability tends to zero. In algorithmic terms, under the hypotheses of the theorem, for any fixed algorithm, the probability of successful reconstruction by that algorithm will tend to zero.

By rearranging the threshold for reconstruction to solve for M_n, we observe that the number of sequences allowed in X is related to the length of the sequences as follows:

$$M \geq \frac{e^{L/C}}{N}.$$

This inequality suggests that the if there are too many sequences in the sample relative to the length of the genomes and the read length, then the probability of successful reconstruction is small.

3. Background, notation, and preliminary results

The identifiability of a sample is highly dependent on the presence of repeated segments of length approximately $L-1$ where L is the length of the reads. For this reason, we begin our analysis by considering the event that certain strings are repeated within the sample. To begin, for a genome X^m and two indices $1 \leq i \leq j \leq N$, let $X^m[i,j] = x_i^m \dots x_j^m$. Next, let us say that the sample X has a repeated pattern of length ℓ if there exists $(m,i) \neq (m',i')$, with $1 \leq m, m' \leq M$ and $1 \leq i, i' \leq N-\ell+1$, such that $X^m[i, i+\ell-1] = X^{m'}[i', i'+\ell-1]$. In this case, we refer to the pair (m,i), (m',i') as a *repeat*. Additionally, we let $S_\ell^{m,m'}(i,i')$ denote the event that (m,i), (m',i') is a repeat of length ℓ for X. If X contains no repeated pattern of length ℓ, then we say that all ℓ-segments are distinct. We note that if $m \neq m'$ or $|i-i'| > \ell$, then the strings $X^m[i, i+\ell-1]$ and $X^{m'}[i', i'+\ell-1]$ are independent, and therefore

$$\mathbb{P}\big(S_\ell^{m,m'}(i,i')\big) = \left(\sum_{a \in \mathcal{A}} p^m(a)\, p^{m'}(a)\right)^\ell = e^{-\ell F(p^m, p^{m'})}. \tag{3.1}$$

Our identifiability result relies on the following combinatorial lemma. For a proof of this result in the language of graphs, see [2, Lemma 2.4]. We note that the proof extends immediately to the metagenomics situation.

Lemma 3.1. *Let (M,N,L,P) be a metagenomics tuple. If the $(L-1)$-segments of the sample X are all distinct, then X is identifiable.*

We now present several lemmas that are used in our proof of the non-identifiability result. First, we require a factor relating the cross-entropy to the entropy.

Lemma 3.2. *Let p and p' be two distributions on $\mathcal{A}$. Then*

$$F(p,p') \geq \frac{1}{2}H_2(p) + \frac{1}{2}H_2(p').$$

Consequently, we also have $F(p,p') \geq \min(H_2(p), H_2(p'))$.

Proof. Recall that

$$F(p,p') = -\log\left(\sum_{i \in \mathcal{A}} p(i)\cdot p'(i)\right).$$

Viewing the distributions p and p' as vectors in $\mathbb{R}^{|\mathcal{A}|}$, we see that $F(p,p') = -\log(\langle p, p'\rangle)$, where $\langle\cdot,\cdot\rangle$ denotes the inner product. By applying the Cauchy-Schwartz inequality and properties of the logarithm, we get

$$F(p,p') \geq \frac{1}{2}H_2(p) + \frac{1}{2}H_2(p'),$$

as desired. □

Now let $2 \leq j \leq N - L$, and suppose that $m_1, \ldots, m_4$ are mutually distinct integers in $[1, M]$. Let $T_j(m_1, m_2, m_3, m_4)$ denote the event that there exists a word w of length L such that

$$\begin{aligned} X^{m_1} &= awb \\ X^{m_2} &= cwd \\ X^{m_3} &= awd \\ X^{m_4} &= cwb, \end{aligned}$$

where a and c have length $j-1$.

Lemma 3.3. *Suppose that for all $1 \leq m \leq M$, we have $H_2(p^m) \geq H_*$. Then*

$$\mathbb{P}(T_j(m_1, m_2, m_3, m_4)) \leq e^{-2NH_*}.$$

Proof. For notation, recall that $X^m[i,k] = x_i^m \ldots x_k^m$. Then we have

$$\begin{aligned} \mathbb{P}\big(T_j(m_1, m_2, m_3, m_4)\big) \leq\ & \mathbb{P}\big(X^{m_1}[1, j+L-1] = X^{m_3}[1, j+L-1]\big) \\ & \cdot \mathbb{P}\big(X^{m_2}[1, j+L-1] = X^{m_4}[1, j+L-1]\big) \\ & \cdot \mathbb{P}\big(X^{m_1}[j+L, N] = X^{m_4}[j+L, N]\big) \\ & \cdot \mathbb{P}\big(X^{m_2}[j+L, N] = X^{m_3}[j+L, N]\big). \end{aligned}$$

Since $m_1, \ldots, m_4$ are all distinct, we may apply (3.1), and we see that

$$\begin{aligned} \mathbb{P}\big(T_j(m_1, m_2, m_3, m_4)\big) \leq\ & e^{-(j+L-1)F(p^{m_1}, p^{m_3})} \\ & \cdot e^{-(j+L-1)F(p^{m_2}, p^{m_2})} \\ & \cdot e^{-(N-j-L+1)F(p^{m_1}, p^{m_4})} \\ & \cdot e^{-(N-j-L+1)F(p^{m_2}, p^{m_3})}. \end{aligned}$$

Then by Lemma 3.2 and our hypothesis that $H_2(p^m) \geq H_*$ for all m, we see that

$$\mathbb{P}\big(T_j(m_1, m_2, m_3, m_4)\big) \leq e^{-2NH_*}.$$

□

The following combinatorial lemma underlies our non-identifiability result. For a word $w \in \mathcal{A}^k$ of length k, we let $|w| = k$.

Lemma 3.4. *Let (M, N, L, P) be a metagenomics tuple. Suppose there exists $2 \leq j \leq N - L$ and $m \neq m'$ such that $X^m = awb$ and $X^{m'} = cwd$, where $|a| = |c| = j-1$, $|w| = L$, $a \neq c$, and $b \neq d$. Further suppose that X is not in $\bigcup_{m_3, m_4} T_j(m, m', m_3, m_4)$. Then X is not identifiable.*

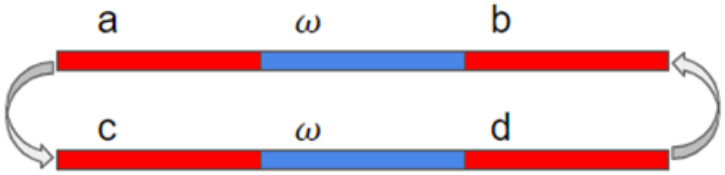

FIGURE 2. Non-Identifiability Configuration. By swapping the segments labeled a and c or b and d, a unique set of genomes, $\tilde{X}$, arises from the set of reads produced by X.

Proof. Consider a sample X satisfying the hypotheses of the lemma. Define $\tilde{X}$ to be a sample such that $\tilde{X}^m = awd$ and $\tilde{X}^{m'} = cwb$. Since $|w| = L$, the samples X and $\tilde{X}$ have the same read set. However, by the hypotheses that $a \neq c$, $b \neq d$, and X is not in $\bigcup_{m_3,m_4} T_j(m, m', m_3, m_4)$, there is no permutation of the genomes in X that will yield $\tilde{X}$. Hence X is not identifiable. □

At this point, let us recall Chebychev's inequality: if Z is any real-valued random variable with mean μ and standard deviation $\sigma > 0$, then for any $k > 0$ we have

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{3.2}$$

In the proof of our non-identifiability result, we make use of the second moment method, which we formalize in the following lemma. We include a proof of this result for completeness.

Lemma 3.5. *Let $\{Z_n\}_{n=1}^{\infty}$ be a sequence of random variables taking values in the nonnegative integers. If*

- $\lim \mathbb{E}[Z_n] = \infty$, *and*
- $\lim \frac{\mathrm{Var}[Z_n]}{\mathbb{E}[Z_n]^2} = 0$,

then

$$\lim \mathbb{P}\big(Z_n \geq 1\big) = 1.$$

Proof. By applying Chebychev's inequality, we get:

$$\begin{aligned} 0 \leq \mathbb{P}(Z_n = 0) &= \mathbb{P}(\mathbb{E}[Z_n] - Z_n \geq \mathbb{E}[Z_n]) \\ &= \mathbb{P}\left(\mathbb{E}[Z_n] - Z_n \geq \frac{\mathbb{E}[Z_n]}{\mathrm{Var}[Z_n]^{1/2}} \mathrm{Var}[Z_n]^{1/2}\right) \\ &\leq \frac{\mathrm{Var}[Z_n]}{\mathbb{E}[Z_n]^2} \to 0, as n \to \infty. \end{aligned}$$

Thus, we obtain that $\mathbb{P}(Z_n \geq 1)$ tends to one. □

4. Proof of identifiability result

This section is devoted to proving Theorem 1. We begin with two lemmas in which we estimate the probability of having at least one repeat of length $L - 1$ in a random sample.

Lemma 4.1. *Suppose (M, N, L, P) is a metagenomics problem such that for all $1 \leq m \leq M$, we have $H_2(p^m) \geq H_*$. Let E_1 denote the event that there exists an $(L-1)$-repeat $(m, i), (m, j)$ with $i < j \leq i + (L-2)$ (meaning that the repeated string overlaps with itself). Then*

$$\mathbb{P}(E_1) \ \leq \ (MNL)e^{-(L-1)H_*/2}. \tag{4.1}$$

Proof. For $1 \leq m \leq M$ and $1 \leq i < j \leq i + L - 2$ and $j \leq N - L + 1$, recall that $S_{L-1}^{m,m}(i, j)$ is the event that the reads at positions i and j on genome X^m are equal, i.e., $x_i^m \dots x_{i+L-2}^m = x_j^m \dots x_{j+L-2}^m$. The probability of this event occurring can be estimated as follows (see [3, Lemma 11] for a proof):

$$\mathbb{P}\big(S_{L-1}^{m,m}(i, j)\big) \ \leq \ e^{-(L-1)H_2(p^m)/2}.$$

Note that $E_1 = \bigcup_m \bigcup_{i,j} S_{L-1}^{m,m}(i, j)$ with $1 \leq m \leq M$ and $1 \leq i < j \leq i + L - 2$ and $j \leq N - L + 1$. Then by the previous display, the hypothesis that each $H_2(p^m) \geq H_*$, and the union bound, we estimate the probability of E_1 as follows:

$$\mathbb{P}(E_1) \ \leq \ (MNL)e^{-(L-1)H_*/2}.$$

□

A repeat satisfying the hypotheses of this lemma is said to be *overlapping*. Any other repeat is said to be *non-overlapping*.

Lemma 4.2. *Suppose (M, N, L, P) is a metagenomics problem such that for all $1 \leq m \leq m' \leq M$, we have $F(p^m, p^{m'}) \geq F_*$. Let E_2 denote the event that there is a non-overlapping repeat of length $L - 1$. Then*

$$\mathbb{P}(E_2) \leq M^2 N^2 (e^{-(L-1)F_*}). \tag{4.2}$$

Proof. For $1 \leq m \leq m' \leq M$ and $1 \leq i, j \leq N - L$, recall that $S_{L-1}^{m,m'}(i, j)$ is the event that the $(L-1)$-strings at position i on X^m and position j on $X^{m'}$ are equal, i.e., $x_i^m \dots x_{i+L-2}^m = x_j^{m'} \dots x_{j+L-2}^{m'}$. These strings are non-overlapping precisely when either $m \neq m'$ or $m = m'$ and $|i - j| > L - 1$. Suppose that one of these conditions holds. Since the strings are non-overlapping, we may apply (3.1), and we have that

$$\mathbb{P}\left(S_{L-1}^{m,m'}(i, j)\right) = e^{-(L-1)F(p^m, p^{m'})}.$$

Next notice that $E_2 = \bigcup_{m,m'} \bigcup_{i,j} S_{L-1}^{m,m'}(i, j)$, where the union is taken over the non-overlapping indices. Then by our hypothesis that $F(p^m, p^{m'}) \geq F_*$ for all m and m' and by the union bound, we have

$$\mathbb{P}(E_2) \leq M^2 N^2 e^{-(L-1)F_*}$$

□

Now we are ready to proceed with the proof of Theorem 1.

Proof of Theorem 1. To begin, we fix n and consider the metagenomics problem $(M, N, L, P) = (M_n, N_n, L_n, P_n)$. Recall that by Lemma 3.1, if all $L - 1$ subsequences of X are unique, then identifiability holds. Let E^c denote the event that all $(L-1)$-subsequences are unique, and let E be the complement of this event. By Lemma 3.1, we have that $E^c \subset I$. Thus, $I^c \ \subset \ E$. Then by monotonicity,

we have $\mathbb{P}(I^c) \leq \mathbb{P}(E)$, making $\mathbb{P}(E)$ an upper bound on the probability of non-identifiability. In the remainder of this proof, we estimate $\mathbb{P}(E)$ and then show that it tends to zero under the hypotheses of the theorem.

First, note that E is the event that there exists an $(L-1)$-repeat. Then we have $E = E_1 \cup E_2$, where E_1 and E_2 are defined in the previous two lemmas. By Lemma 4.1, we have that

$$\mathbb{P}(E_1) \leq (C_1 MNL)e^{-LH_*/2},$$

where $C_1 = e^{H_*/2} > 1$ is a constant (independent of n). By Lemma 4.2, we have that

$$\mathbb{P}(E_2) \leq C_2 M^2 N^2 \, e^{-LF_*}$$

where $C_2 = e^{F_*}/4$ is a constant (independent of n). Then by the union bound, we see that

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(E_1 \cup E_2) \\ &\leq C_1(MNL)e^{-LH_*/2} + C_2 M^2 N^2 \, e^{-LF_*}.\end{aligned}$$

By Lemma 3.4, we have that $F_* \geq H_*$. Thus,

$$\mathbb{P}(E) \leq (C_1 MNL)e^{-LH_*/2} + C_2 M^2 N^2 \, e^{-LH_*}. \tag{4.3}$$

By our hypothesis, we have

$$L \geq \frac{2(1+\epsilon)}{H_*} \log(MN).$$

Combining this estimate with (4.3), we now let n tend to infinity and observe that $\mathbb{P}(E)$ tends to zero. Hence we have that $\mathbb{P}(I^c) \leq \mathbb{P}(E)$ tends to zero, and therefore $\mathbb{P}(I)$ tends to one. □

5. Proof of non-identifiability result

In this section we prove our second main result, Theorem 2. The main proof relies on a second moment argument, which establishes that certain "bad" configurations will arise with probability tending to one. Before getting to the main proof, we require some notations and lemmas regarding these bad configurations.

Throughout this section we suppress the dependence on n in our notation. For $j \in \{2, \ldots, N-L\}$ and $1 < m < m' < M$, we let and let $B_j^{m,m'}$ be the event that there is an L repeat at location j on the genomes X^m and $X^{m'}$ (i.e., $B_j^{m,m'} = S_L^{m,m'}(j,j)$ in the notation of Section 3). As we show below, if $B_j^{m,m'}$ occurs, then it is likely that X is non-identifiable. Our first goal is to use the second moment method to give conditions under which there are some indices for which $B_j^{m,m'}$ occurs with probability tending to one. Towards that end, suppose that $\delta \in (0,1)$ is fixed and $L \leq \delta N$ (as in Theorem 2). Choose $\eta > 0$ such that $\eta < \min(\delta/2, (1-\delta)/2)$, and let

$$Z_n = \sum_{1 \leq m < m' \leq M_n} \sum_{\eta N \leq j \leq (1-\eta)N - L} \chi_{B_j^{m,m'}}, \tag{5.1}$$

where χ_E denotes the indicator function of the event E.

Lemma 5.1. *Under the hypotheses of Theorem 2, there exists a constant $C_3 > 0$ such that for all large n, we have*

$$\mathbb{E}[Z_n] \geq C_3 M^2 N e^{-LF^*}.$$

In particular,

$$\lim \mathbb{E}[Z_n] = \infty.$$

Proof. First, we note that the number of pairs (m, m') such that $1 \leq m < m' \leq M$ is equal to $\binom{M}{2}$, and we have $\binom{M}{2} \geq \frac{1}{2}M^2$ for all $M \geq 2$. Next, we observe that there are $(1-2\eta)N - L - 1$ values of j such that $\eta N \leq j \leq (1-\eta)N - L$, and for $L \leq \delta N$, we have $(1-2\eta)N - L - 1 \geq (1-2\eta-\delta)N$ for all large enough N. Now let $C_3 = \frac{1-\delta-2\eta}{2}$. Then we have

$$\begin{aligned}
\mathbb{E}[Z_n] &= \sum_{1\leq m<m'\leq M_n} \sum_{\eta N \leq j \leq (1-\eta)N-L} \mathbb{E}[\chi_{B_j^{m,m'}}] \\
&= \sum_{1\leq m<m'\leq M_n} \sum_{\eta N \leq j \leq (1-\eta)N-L} e^{-LF(p_n^m, p_n^{m'})} \\
&\geq C_3 M^2 N e^{-LF^*}.
\end{aligned}$$

By our hypothesis that

$$L \leq \frac{1-\epsilon}{F^*} \log(MN),$$

we see that $\lim \mathbb{E}[Z_n] = \infty$. □

Lemma 5.2. *Under the hypotheses of Theorem 2,*

$$\lim \frac{\operatorname{Var}[Z_n]}{\mathbb{E}[Z_n]^2} = 0.$$

Proof. First, note that $B_j^{m,m'}$ is independent of $B_k^{m_0,m_0'}$ whenever $\{m, m'\} \cap \{m_0, m_0'\} = \varnothing$ or $|j-k| > L$. Thus we have

$$\begin{aligned}
\operatorname{Var}[Z_n] &= \sum_{m,m'} \sum_{j} \sum_{m_0,m_0'} \sum_{k} \mathbb{E}\big[(\chi_{B_j^{m,m'}} - \mathbb{E}[\chi_{B_j^{m,m'}}])(\chi_{B_k^{m_0,m_0'}} - \mathbb{E}[\chi_{B_k^{m_0,m_0'}}])\big] \\
&= \sum_{\{m,m'\}\cap\{m_0,m_0'\}\neq\varnothing} \sum_{|j-k|\leq L} \mathbb{E}\big[(\chi_{B_j^{m,m'}} - \mathbb{E}[\chi_{B_j^{m,m'}}])(\chi_{B_k^{m_0,m_0'}} - \mathbb{E}[\chi_{B_k^{m_0,m_0'}}])\big].
\end{aligned}$$

By Cauchy-Schwartz, 3.1, and our hypothesis that $F(p^m, p^{m'}) \geq F_*$ for all m and m', we have that

$$\mathbb{E}\big[(\chi_{B_j^{m,m'}} - \mathbb{E}[\chi_{B_j^{m,m'}}])(\chi_{B_k^{m_0,m_0'}} - \mathbb{E}[\chi_{B_k^{m_0,m_0'}}])\big] \leq \left(\operatorname{Var}[\chi_{B_j^{m,m'}}] \operatorname{Var}[\chi_{B_k^{m_0,m_0'}}]\right)^{1/2} \leq e^{-LF_*}.$$

By the previous two displays, we see that

$$\operatorname{Var}[Z_n] \leq C_4 M^3 N L e^{-LF_*},$$

where $C_4 > 0$ is a constant (independent of n). Also, by Lemma 5.1, there exists a constant $C_3' > 0$ such that

$$\mathbb{E}[Z_n] \geq C_3' M^2 N e^{-LF^*}.$$

Combining the previous two inequalities, we see that

$$\frac{\operatorname{Var}[Z_n]}{\mathbb{E}[Z_n]^2} \leq \frac{C_4 M^3 N L e^{-LF_*}}{(C_3')^2 M^4 N^2 e^{-2LF^*}} = \frac{C_4 L}{(C_3')^2 MN} e^{-L(F_* - 2F^*)}. \tag{5.2}$$

By our hypothesis, we have that

$$L \leq \frac{1-\epsilon}{2F^* - F_*} \log(MN),$$

and as a result we obtain that the right-hand side of (5.2) tends to zero as n tends to infinity. □

Let $1 \le m < m' \le M$ and $2 \le j \le N - L$. Let $E_j^{m,m'}$ be the event that there exist $a, c \in \mathcal{A}^{j-1}$ and $b, d \in \mathcal{A}^{N-j-L+1}$ and $w \in \mathcal{A}^L$ satisfying $a \ne c$, $b \ne d$ such that $X^m = awb$ and $X^{m'} = cwd$ and X is not in $\bigcup_{m_3,m_4} T_j(m, m', m_3, m_4)$.

Lemma 5.3. *Under the hypotheses of Theorem 2,*

$$\mathbb{P}\left(B_j^{m,m'} \setminus E_j^{m,m'}\right) \le e^{-(j-1)F_*} + e^{-(N-(j+L)+1)F_*} + M^2 e^{-2NH_*}.$$

Proof. First, note that

$$B_j^{m,m'} \setminus E_j^{m,m'} \subset \left(S_{j-1}^{m,m'}(1,1)\right) \cup \left(S_{N-(j+L)+1}^{m,m'}(j+L, j+L)\right) \cup \left(\bigcup_{m_3,m_4} T_j(m, m', m_3, m_4)\right).$$

Since $m \ne m'$, we may apply (3.1) and obtain

$$\mathbb{P}\left(S_{j-1}^{m,m'}(1,1)\right) \le e^{-(j-1)F_*},$$

and

$$\mathbb{P}\left(S_{N-(j+L)+1}^{m,m'}(j+L, j+L)\right) \le e^{-(N-(j+L)+1)F_*}.$$

Also, by Lemma 3.3 and the union bound, we have

$$\mathbb{P}\left(\bigcup_{m_3,m_4} T_j(m, m', m_3, m_4)\right) \le M^2 e^{-2NH_*}.$$

By the union bound and these individual estimates, we obtain the desired bound. □

Lemma 5.4. *Under the hypotheses of Theorem 2,*

$$\lim \mathbb{P}\left(\left(\bigcup B_j^{m,m'}\right) \setminus \left(\bigcup E_j^{m,m'}\right)\right) = 0,$$

where the unions are taken over the same set of indices defining Z_n (in (5.1)).

Proof. Observe that

$$\left(\bigcup B_j^{m,m'}\right) \setminus \left(\bigcup E_j^{m,m'}\right) \subset \bigcup \left(B_j^{m,m'} \setminus E_j^{m,m'}\right).$$

Then by Lemma 5.3 and the fact that $j - 1 \ge \eta N$ and $N - (j + L) \ge (1 - \delta - \eta)N$, we have that

$$\begin{aligned}
\mathbb{P}\left(\left(\bigcup B_j^{m,m'}\right) \setminus \left(\bigcup E_j^{m,m'}\right)\right) &\le \sum_{m,m'} \sum_j \mathbb{P}\left(B_j^{m,m'} \setminus E_j^{m,m'}\right) \\
&\le M^2 N\left(e^{-(j-1)F_*} + e^{-(N-(j+L)+1)F_*} + M^2 e^{-2NH_*}\right) \\
&\le M^2 N e^{-\eta N F_*} + M^2 N e^{-(1-\delta-\eta)NF_*} + M^4 N e^{-2NH_*}.
\end{aligned}$$

By the hypothesis that $\log(M)/N$ tends to zero, the right-hand side of the previous display tends to zero, and we obtain the desired limit. □

We now turn our attention to the main proof of Theorem 2.

Proof of Theorem 2. By Lemmas 5.1 and 5.2, the second moment method (Lemma 3.5) applies, and thus $Z_n \ge 1$ holds with probability tending to one. Note that the event $Z_n \ge 1$ is the same as $\bigcup B_j^{m,m'}$ (where the indices in the union are

the same as in (5.1)), and thus the probability of $\bigcup B_j^{m,m'}$ tends to one. Then by Lemma 5.4, we see that $\bigcup E_j^{m,m'}$ occurs with probability tending to one. By Lemma 3.4, we have $\bigcup E_j^{m,m'} \subset I^c$, and therefore non-identifiability occurs with probability tending to one.

□

6. Discussion and future work

This research paper is designed to explore the basic requirements for shotgun DNA reconstruction in the metagenomics case. We obtain these results by considering a blind reconstruction of a set of genomes under an IID model. The model we explore in this paper serves as a base for future, more realistic work. We consider the following generalizations to be interesting directions for future research.

Negative Result. There is a gap between the identifiability and non-identifiability result studied in this work. Future work may analyze a different repeat configuration in order to get a tighter bound in the negative direction.

Short Tandem Repeats. Large sequences, on the order of several thousand base pairs, made up of short repeated segments occur naturally within biological systems. The IID model fails to account for this phenomenon, and the greedy algorithm falls apart upon the introduction of these short tandem repeats. However, the length and composition of these repeats are unique to different organisms and this knowledge can be utilized to simplify real world sequencing problems.

Markov Statistical Model. Within biological systems, the probability of the occurrence of a specific nitrogenous base is not independent of the bases that have occurred previously in the sequence. Due to the nature of DNA transcription and translation, base pair occurrences are heavily influenced by the previous three base pairs as well as lightly influenced by up to several hundred previous base pairs. Information on how strings of tetramers can be utilized to reconstruct specific DNA sequences may be found in biological databases.

Noisy Data. DNA shotgun sequencing technology inadvertently introduces noise into the system via base pair insertions and deletions. The model studied in this paper fails to account for noise that may be present within the DNA sequences. Future work may study how this noise shifts the threshold for reconstruction and find more realistic parameters for a successful experiment.

Acknowledgments

This research was conducted during a summer research program for undergraduates that was supported by the National Science Foundation through grant DMS-1847144. The author wishes to thank Robert Bland and Jacob Raymond for their help and Dr. Kevin McGoff for supervising this project.

References

[1] Guy Bresler, Ma'ayan Bresler, and David Tse. Optimal assembly for high throughput shotgun sequencing. In *BMC bioinformatics*, volume 14, pages 1–13. BioMed Central, 2013.

[2] Elchanan Mossel and Nathan Ross. Shotgun assembly of labeled graphs. *IEEE Transactions on Network Science and Engineering*, 6(2):145–157, 2017.

[3] Abolfazl S Motahari, Guy Bresler, and NC David. Information theory of DNA shotgun sequencing. *IEEE Transactions on Information Theory*, 59(10):6273–6289, 2013.

[4] Jacob Raymond, Robert Bland, and Kevin McGoff. Shotgun identification on groups. *Involve, a Journal of Mathematics*, 14(4):631–682, 2021.

Marlee Herring, School of Engineering and Applied Sciences, Columbia University, 400 W 119th St., New York, NY 10027
Email address: mh4273@columbia.edu

Kevin McGoff, Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223
Email address: kmcgoff1@uncc.edu