Sequence assembly from corrupted shotgun reads

Shirshendu Ganguly * Elchanan Mossel † Miklós Z. Rácz ‡

January 27, 2016

Abstract

The prevalent technique for DNA sequencing consists of two main steps: shotgun sequencing, where many randomly located fragments, called reads, are extracted from the overall sequence, followed by an assembly algorithm that aims to reconstruct the original sequence. There are many different technologies that generate the reads: widely-used second-generation methods create short reads with low error rates, while emerging third-generation methods create long reads with high error rates. Both error *rates* and error *profiles* differ among methods, so reconstruction algorithms are often tailored to specific shotgun sequencing technologies. As these methods change over time, a fundamental question is whether there exist reconstruction algorithms which are *robust*, i.e., which perform well under a wide range of error distributions.

Here we study this question of sequence assembly from corrupted reads. We make no assumption on the *types* of errors in the reads, but only assume a bound on their *magnitude*. More precisely, for each read we assume that instead of receiving the true read with no errors, we receive a corrupted read which has edit distance at most ε times the length of the read from the true read. We show that if the reads are long enough and there are sufficiently many of them, then approximate reconstruction is possible: we construct a simple algorithm such that for almost all original sequences the output of the algorithm is a sequence whose edit distance from the original one is at most $O(\varepsilon)$ times the length of the original sequence.

1 Introduction

DNA sequencing is by now an essential element of a variety of biological and clinical studies. Current de novo sequencing technologies typically have two main stages. First, many randomly located fragments, called reads, are extracted from the DNA sequence in a process called shotgun sequencing. Next, an assembly algorithm aims to reconstruct the original sequence based on overlaps between the reads.

The rise of second-generation sequencing methods, such as Illumina, have resulted in many advances in the past decade because they generate high-throughput data cheaply and quickly. However, these methods produce short reads (a few hundred basepairs long) in order to have a low error rate (1-3%), which results in incomplete and fragmented assemblies [9]. Emerging technologies, such as PacBio's Single Molecule Real-Time sequencing technology and Oxford Nanopore Technologies, were developed in part to solve this problem. They produce long reads (over ten thousand basepairs long), but currently suffer from a high error rate (10-22%) (see, e.g., [2, 6, 5]).

Not only do these different shotgun sequencing methods produce reads with different error *rates*, they also have different error *profiles*, e.g., due to various systematic errors. Consequently assembly

*University of Washington; sganguly@math.washington.edu.

†University of Pennsylvania and University of California, Berkeley; mossel@wharton.upenn.edu.

‡Microsoft Research; miracz@microsoft.com.

algorithms are often tailored to specific sequencing technologies to exploit their unique properties. As these technologies will inevitably change and new ones will arise, a fundamental question is the *robustness* of reconstruction algorithms. Will the current ones still be useful a decade from now? Are there algorithms which perform well under a wide variety of error distributions? This is the question we study in this paper.

Several recent papers have taken an *information-theoretic* point of view to the sequence assembly problem. The basic question is: what are the fundamental limits to *any* assembly algorithm? Given a sequencing technology and the statistics of the DNA sequence, how long do the reads need to be and how many are required for reconstruction? Motahari, Bresler, and Tse [7] study this question assuming an i.i.d. DNA sequence and error-free reads, and show a sharp phase transition: if the reads are short enough to have repeats, then reconstruction is impossible, while as long as the reads are long enough to have no repeats, the necessary condition of having enough reads to cover the whole DNA sequence is essentially sufficient. Previously Dyer, Frieze, and Suen [3] obtained the same phase transition in the length of the reads assuming sequencing by hybridization, i.e., that a copy of every read is available. Several extensions and variations of this problem have been studied. Adding some amount of i.i.d. noise to the reads still allows for reconstruction of the perfect layout of the reads [8]. In [1] the authors give a sufficient condition for reconstruction for any sequence based on its repeat statistics, assuming error-free reads. This was later extended to allow the reads to come from an erasure error model [10]. These papers are discussed in more detail later.

In this paper we continue this line of work, assuming an i.i.d. DNA sequence as in [7, 8]. The main novelty in the model we consider is a strong *adversarial corruption/error model* on the reads. More precisely, for each read we assume that instead of receiving the true read with no errors, we receive a corrupted read with the edit distance between the true and the corrupted reads being at most ε times the length of the true read. Given such a strong adversarial error model, we relax our goal from perfect reconstruction to approximate reconstruction. Our main contribution is to show that if the reads are long enough and there are sufficiently many of them, then approximate reconstruction is possible: we present a simple sequential algorithm for which the edit distance between the original sequence and the output of the algorithm is at most $O(\varepsilon)$ times the length of the original sequence.

2 Problem setting

We are interested in approximately recovering a long sequence of interest from a set of randomly chosen shorter reads which are arbitrarily corrupted up to a certain extent. Consequently the problem has four main parameters: sequence length n, read length L, number of reads N, and error/corruption rate ε; a fifth parameter, δ, measures the probability of unsuccessful approximate reconstruction.

Before defining the problem precisely, we introduce some notation. Let Σ be a finite alphabet from which the entries of the sequence come from; in the case of DNA sequencing we have $\Sigma = \{A, C, G, T\}$. For a sequence $x = (x_1, x_2, \ldots, x_n) \in \Sigma^n$ and integers i and j, let $x[i, j]$ denote the substring $(x_i, x_{i+1}, \ldots, x_j)$. Let $\Sigma^* = \cup_n \Sigma^n$. For $x, y \in \Sigma^*$, let $\mathrm{ed}(x, y)$ denote the edit distance between x and y, i.e., the minimum number of deletion, insertion, or substitution operations necessary to go from x to y.

The approximate reconstruction problem with parameters $(n, L, N, \varepsilon, \delta)$ is then defined as follows; see Fig. 1 for an illustration.

- The sequence of interest, $X \in \Sigma^n$, is chosen uniformly at random among all possible sequences;

i.e., the entries of X are i.i.d. chosen uniformly from the alphabet Σ.[1]

- The data are corrupted reads of X, defined as follows. Let $\{T_i\}_{i=1}^N$ be i.i.d. uniform in $\{1, 2, \ldots, n-L+1\}$ (the starting positions of the reads), and let $R_i = X[T_i, T_i + L - 1]$ be the i^{th} (uncorrupted) read. Instead of receiving the (multi)set of (uncorrupted) reads $\mathcal{R} = \{R_1, R_2, \ldots R_N\}$, we receive a (multi)set of corrupted reads

$$\widetilde{\mathcal{R}} = \left\{\widetilde{R}_1, \widetilde{R}_2, \ldots, \widetilde{R}_N\right\},$$

where the only thing we know is that

$$\text{ed}\left(R_i, \widetilde{R}_i\right) \leq \varepsilon L, \qquad \text{for every } i \in [N], \tag{1}$$

but otherwise the $\widetilde{R}_i$ can be arbitrary.[2]

- The goal of an approximate reconstruction algorithm is to output a sequence $\widehat{X} \in \Sigma^*$ such that

$$\text{ed}\left(X, \widehat{X}\right) \leq C\varepsilon n \tag{2}$$

for some absolute constant C, with probability at least $1 - \delta$ (for all n large enough).

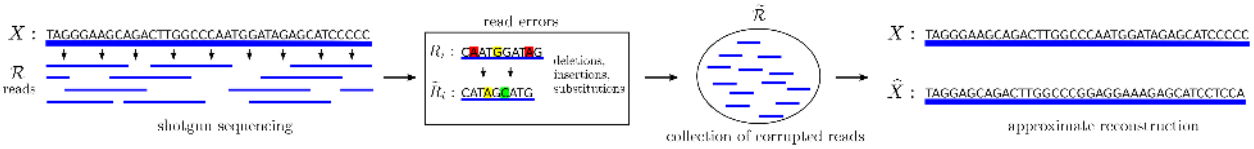

Figure 1: A schematic of the approximate reconstruction problem.

When $\varepsilon = 0$ (i.e., there are no errors in the reads), this amounts to exact reconstruction of the original sequence, which was studied and solved in [7].

If an algorithm achieves (2) for a given error rate $\varepsilon > 0$ with a given constant C, then we say that it is an *approximate reconstruction algorithm for error rate ε with approximation factor C*. Note that for the empty string $\emptyset$ we have $\text{ed}(X, \emptyset) = n$, so for a given approximation factor C the problem is interesting only for $\varepsilon < 1/C$; of course the goal is minimize the approximation factor C.

3 Results

Before presenting our results on the approximate reconstruction problem described above, we first recall the main results of [7] characterizing the case of $\varepsilon = 0$, i.e., error-free reads. In both cases the interesting regime is when the read length L scales as the logarithm of the sequence length n, so in the following we let $L = \overline{L}\ln(n)$, where $\overline{L}$ is constant.[3]

[1] We focus on the uniform distribution for simplicity; extensions to certain other distributions should also be tractable.

[2] The choice of edit distance arises naturally from DNA sequencing; one can consider the problem with other notions of distance as well.

[3] For the sake of readability, we refrain from rounding non-integer values that are meant to be integers, such as $\overline{L}\ln(n)$.

When there are no errors in the reads, then there are two obstructions to reconstruction. First, if the reads are too short, then there will be repeats coming from different parts of the original sequence, which create ambiguity in reconstruction, even if all substrings of length L of the sequence are given. This observation goes back to the work of Ukkonen [11], who characterized the patterns that preclude exact reconstruction of the sequence. This is often referred to as the *repeat-limited regime*.

Second, even if L is large enough, it is necessary to have enough reads to cover the original sequence; otherwise the data does not contain enough information for exact reconstruction. Define $N_{\text{cov}} = N_{\text{cov}}(n, L, \delta)$ as the minimum number of reads necessary such that with probability at least $1-\delta$ the randomly located reads cover the entire original sequence. Lander and Waterman [4] first studied the coverage properties of shotgun sequencing and showed that $N_{\text{cov}} \approx \frac{n}{L} \ln\left(\frac{n}{L\delta}\right)$. This is often referred to as the *coverage-limited regime*.

In the main result of [7], Motahari, Bresler, and Tse show that when there are no errors in the reads, these are the only two obstructions to reconstruction.

Theorem 3.1 (Exact reconstruction from error-free reads [7])**.** *If $\overline{L} < 2/\ln(|\Sigma|)$, then no algorithm can reconstruct the original sequence exactly with probability greater than $1/2 + o(1)$ (as $n \to \infty$).*

If $\overline{L} > 2/\ln(|\Sigma|)$, then exact reconstruction is possible, and the necessary condition of coverage is essentially sufficient. More precisely, let $N_{\min}(n, L, \delta)$ denote the minimum number of reads needed to reconstruct the original sequence exactly with probability at least $1-\delta$. If $\overline{L} > 2/\ln(|\Sigma|)$, then for every fixed $\delta \in (0, 1/2)$,

$$\lim_{n\to\infty, L=\overline{L}\ln(n)} \frac{N_{\min}(n, L, \delta)}{N_{\text{cov}}(n, L, \delta)} = 1.$$

When the reads are corrupted, the two obstructions to reconstruction discussed above remain. In fact, the repeat-limited regime is slightly larger in the presence of corruption. If $\overline{L} < 2/\{(1-\varepsilon)\ln(|\Sigma|)\}$ then consider the corruption process which deletes the last εL coordinates of every read. The reads then have normalized length $(1-\varepsilon)\overline{L} < 2/\ln(|\Sigma|)$, so by Theorem 3.1 exact reconstruction is impossible.

The picture describing successful algorithms is less clean when the reads are corrupted. This is primarily due to the desideratum of approximate reconstruction. When there are no errors, the success of an algorithm is binary: either it reconstructs the original sequence exactly or it does not. Here, however, an algorithm can have varying degrees of success based on the approximation factor C that it achieves in (2). Even for those parameters $\left(\overline{L}, N\right)$ for which an algorithm with finite approximation factor C exists, the best achievable C might depend on $\left(\overline{L}, N\right)$.

With this in mind, our goal in this paper is to show that when $\overline{L}$ and N are large enough, then there exists an approximate reconstruction algorithm with finite approximation factor.

Theorem 3.2 (Approximate reconstruction from corrupted reads)**.** *There exist constants C and $\overline{C}$, depending only on $|\Sigma|$, such that for every $\varepsilon > 0$, if $\overline{L} > \overline{C}/\varepsilon$ and $N > N_{\text{cov}}/\varepsilon$, then there exists an approximate reconstruction algorithm for error rate ε with approximation factor C.*

We show that a simple sequential algorithm achieves this result. Starting with an arbitrary read, at each step of the algorithm we find a read that overlaps with the current partially reconstructed sequence and extend the sequence using this read. If $\overline{L}$ and N are large enough, then in each step of this process we extend the sequence by at least cL for some positive constant c, while incurring an error in edit distance of at most $O(\varepsilon)L$. Estimates on the edit distance between random strings with some overlap are crucially used in the analysis. The algorithm terminates

when it has approximately reached both ends of the original sequence, and results in an estimate with the guarantee given by Theorem 3.2.

Several variants of such an algorithm can be considered, and it can be shown using one of them that the dependence of $\overline{L}$ and N on ε in Theorem 3.2 is not necessary. However, we decided to focus on one particular variant because it results in a small approximation factor (close to 3) when ε is small enough, as stated in the following theorem.

Theorem 3.3 (Approximate reconstruction from corrupted reads)**.** *For every $C > 3$ there exist constants $\overline{C} = \overline{C}(\Sigma)$, $\varepsilon_0 = \varepsilon_0(\Sigma, C)$ and $C' = C'(\Sigma, C)$ such that for every $\varepsilon \in (0, \varepsilon_0)$ if $\overline{L} \geq \overline{C}/\varepsilon$ and $N \geq C' N_{\mathrm{cov}}/\varepsilon$, then there exists an approximate reconstruction algorithm for error rate ε with approximation factor C.*

The closest results to Theorems 3.2 and 3.3 in the literature are those of [8] and [10]. In [8] the authors consider i.i.d. noise affecting the reads of an i.i.d. sequence, and show that perfect layout (where all the reads are mapped correctly to their true locations) is possible even when the noise level is relatively high. The main reason for this positive result is that the independent noise assumption allows error correction of the reads by averaging; in the adversarial error model considered here such averaging is not always possible, hence the weaker goal of approximate reconstruction and the results of Theorems 3.2 and 3.3.

In the recent follow-up work [10], the authors show positive results for a more realistic adversarial error model. They consider arbitrary sequences and give a bound on the read length, as a function of the repeat statistics of the sequence and the error rate, above which perfect assembly is possible. However, the model they consider simplifies several aspects of the problem. The authors mention several possible extensions as avenues for future work, two of which we consider in this paper: more general errors, and a shotgun read model. For one, they specifically consider *erasure errors*, where symbols in a read are erased, but the locations of the erasures are known. Furthermore they assume bounds not only on the number of erasures in a read, but also on the number of reads in which a given base is erased. This means that the reads contain information about the whole sequence. The error model we consider is much more adversarial, e.g., it can happen that even all the reads together contain no information at all about εn bases of the sequence due to deletions. Also, they consider a dense-read model, where all reads of length L of the original sequence are provided, therefore bypassing the question of coverage depth necessary for assembly. Here we instead consider the more realistic shotgun read model and provide a sufficient bridging condition for approximate reconstruction.

In summary, while our reconstruction results are weaker than previous ones, this is due to the much stricter adversarial error model we consider. Going forward, the main challenge is to bridge the gap between these models and results.

4 Sequential reconstruction algorithm and its analysis

We first present some results on the edit distance between random strings in Section 4.1 that then allow us to present a simple sequential reconstruction algorithm in Section 4.2. This algorithm is then analyzed and shown to have the desired performance in Section 4.3.

4.1 Results on the edit distance between random strings

Since the only information we have about the corrupted reads is that their edit distance from the actual reads is not too large (see (1)), it is essential for any reconstruction algorithm to have a

good understanding of and to make use of the edit distance between pairs of reads. Accordingly, we now present results on the edit distance between random strings with overlap; the proofs of these results are in Appendix A.

Consider two random strings with overlap. Simulations show (see Figure 2) that there is a phase transition in the edit distance between the two as a function of the overlap. If the overlap is above a certain threshold (which is linear in the length of the strings), then the edit distance is exactly twice the length of the overhang; if the overlap is below this threshold, then the edit distance between the two strings is as if they were completely independent. The results below rigorously verify certain aspects of this picture.

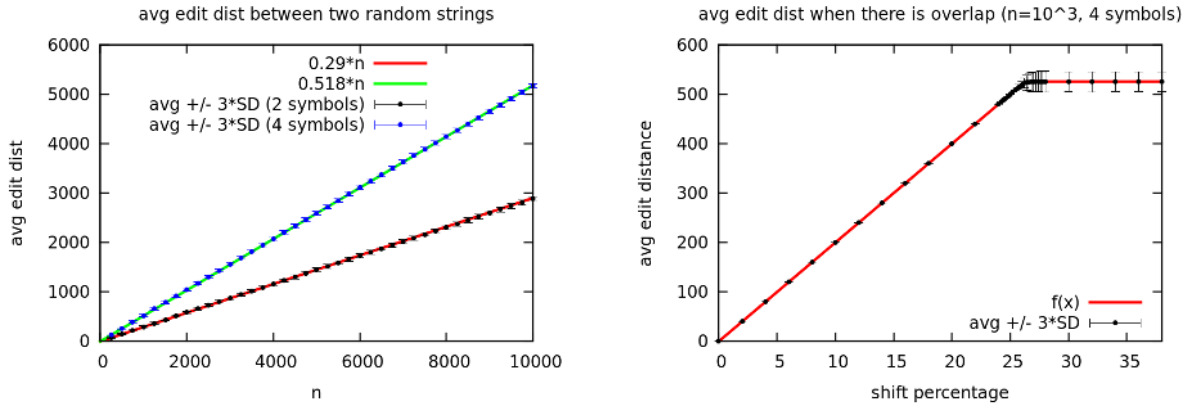

Figure 2: The left plot shows the empirical average edit distance between two independent strings of length n, where n ranges from 250 to 10^4, and $|\Sigma|$ is 2 (black) or 4 (blue). The lines $0.29n$ (for $|\Sigma| = 2$) and $0.518n$ (for $|\Sigma| = 4$) show a good fit to the data, though the limiting slope appears to be somewhat smaller than 0.29 and 0.518, respectively. The right plot shows the empirical average edit distance between two strings of length $n = 10^3$ and with $|\Sigma| = 4$, where one string is a shift of the other, as in the setup of Lemma 4.3. The function f is piecewise linear with two pieces: it is equal to twice the length of the shift when the shift percentage is less than 26.25%, and it is equal to the constant 525 (52.5% of the length $n = 10^3$) when the shift percentage is more than 26.25%. In both plots the average is over 10^3 runs and the error bars show plus/minus 3 times the empirical standard deviation.

Lemma 4.1. *Let $X_m, Y_m \in \Sigma^m$ be two independent uniformly random strings. There exists an absolute constant $c_{\text{ind}} = c_{\text{ind}}(\Sigma) > 0$ such that almost surely*

$$\lim_{m\to\infty} \frac{1}{m} \text{ed}(X_m, Y_m) = c_{\text{ind}}.$$

Determining the value of the limiting constant c_{ind} is a challenging open problem. When $|\Sigma| = 4$, as in the case of DNA sequencing, simulations suggest that $c_{\text{ind}} \approx 0.51$, while we show a simple lower bound of $c_{\text{ind}} > 0.338$.

Lemma 4.2. *Let $X \in \Sigma^{2m}$ be a uniformly random string. For every $d \in (0,1)$ there exist positive constants $\gamma = \gamma(d, \Sigma)$ and $c' = c'(\Sigma)$ such that*

$$\text{ed}(X[1,m], X[k+1, k+m]) \geq \gamma m$$

for all $k \geq dm$ with probability at least $1 - e^{-c'dm}$.

In particular, when $|\Sigma| = 4$, then we can take $\gamma = 0.338 \times d$.

Lemma 4.3. *Let $X \in \Sigma^{2m}$ be a uniformly random string. There exist positive constants $c = c(\Sigma)$ and $c' = c'(\Sigma)$ such that*

$$\mathrm{ed}(X[1,m], X[1+k, m+k]) = 2k$$

for all $k \leq cm$ with probability at least $1 - e^{-c'm}$.

We denote by $\kappa_{\mathrm{ed}} = \kappa_{\mathrm{ed}}(\Sigma)$ the supremum of all constants $c = c(\Sigma)$ for which Lemma 4.3 holds (with some $c' = c'(\Sigma)$). Figure 2 suggests that $\kappa_{\mathrm{ed}} = c_{\mathrm{ind}}/2$. We show that for $|\Sigma| = 4$, $\kappa_{\mathrm{ed}} > 0.0846$.

As a corollary of the lemmas we obtain the following result.

Corollary 4.4. *Let $X \in \Sigma^n$ be uniformly random and let $L = \overline{L} \ln(n)$. If the constant $\overline{L} = \overline{L}(\Sigma)$ is large enough, then there exists a positive constant $c = c(\Sigma)$ such that with probability going to 1 as $n \to \infty$ the following holds for all $i, j \in [n - L + 1]$:*

(a) if $|i - j| \leq cL$, then $\mathrm{ed}(X[i, i+L-1], X[j, j+L-1]) = 2|i-j|$;

(b) otherwise $\mathrm{ed}(X[i, i+L-1], X[j, j+L-1]) \geq 2cL$.

Proof. This follows directly from Lemmas 4.2 and 4.3, together with a union bound over all possible pairs $i, j \in [n - L + 1]$. The constant $\overline{L}$ needs to be chosen large enough so that the error probabilities in Lemmas 4.2 and 4.3 are $o\left(n^{-2}\right)$. □

4.2 Sequential reconstruction algorithm

We present now a simple sequential approximate reconstruction algorithm. The algorithm takes as a parameter the number of reads; this guarantees a certain amount of coverage. Let $N = C' N_{\mathrm{cov}}/\varepsilon$ and $c' = 1/C'$, where we assume that $C' \geq 1$; standard results [4] imply that then with probability at least $1 - \delta/2$ there is no gap greater than $1.1 \times c' \varepsilon L$ in between subsequent starting points of the (yet uncorrupted) reads. We fix $\alpha = 4/c(\Sigma)$, where $c(\Sigma)$ is given by Corollary 4.4. For the purposes of Theorems 3.2 and 3.3 we may and will assume that ε is small enough; in particular we assume that $\varepsilon \leq 1/(3\alpha)$.

Before specifying the algorithm we introduce further notation. Let negative integers denote counting coordinates from the opposite end of a sequence, e.g., for $x \in \Sigma^m$, $x[-k, -1]$ denotes the suffix of x of length k. Furthermore let $I : \widetilde{\mathcal{R}} \to [N]$ denote the map that takes a corrupted read to its index, i.e., $I\left(\widetilde{R}_i\right) = i$.

The algorithm is as follows:

Let $Y = \widetilde{R}_1$ and set $k = 1$.
while there exists $\widetilde{R} \in \widetilde{\mathcal{R}}$ such that $\mathrm{ed}\left(\widetilde{R}_k[-\alpha\varepsilon L, -1], \widetilde{R}[1, \alpha\varepsilon L]\right) \leq (2 + 2c')\varepsilon L$ **do**
    choose any such $\widetilde{R} \in \widetilde{\mathcal{R}}$;
    $Y \leftarrow$ the concatenation of Y and $\widetilde{R}[1 + \alpha\varepsilon L, -1]$;
    $k \leftarrow I\left(\widetilde{R}\right)$.
end while
Set $k = 1$.
while there exists $\widetilde{R} \in \widetilde{\mathcal{R}}$ such that $\mathrm{ed}\left(\widetilde{R}[-\alpha\varepsilon L, -1], \widetilde{R}_k[1, \alpha\varepsilon L]\right) \leq (2 + 2c')\varepsilon L$ **do**
    choose any such $\widetilde{R} \in \widetilde{\mathcal{R}}$;
    $Y \leftarrow$ the concatenation of $\widetilde{R}[1, -\alpha\varepsilon L - 1]$ and Y;

11: $k \leftarrow I\left(\widetilde{R}\right)$.
12: **end while**

In words: we take an arbitrary read, extend it to the right until we possibly can (first while loop), and then extend it to the left until we can (second while loop).

4.3 Analysis of the sequential algorithm

We now analyze the algorithm presented above and as a consequence prove our results: Theorem 3.2 follows by taking $C' = 1$, while Theorem 3.3 follows by taking C' large enough.

We first recall a fact that follows directly from the dynamic programming algorithm for computing the edit distance. For any m, sequences $x, y \in \Sigma^m$, and $i, j < m$, we have

$$\operatorname{ed}\left(x\left[1,i\right], y\left[1,j\right]\right) \leq \operatorname{ed}\left(x\left[1,i+1\right], y\left[1,j+1\right]\right). \tag{3}$$

In words: deleting a coordinate from the end of both x and y cannot increase their edit distance.

The first thing we have to understand is the set of corrupted reads that satisfy the conditions of the while loops in the algorithm; we focus on the first while loop as the second one is analogous. The following lemma says that if two corrupted reads are such that the length $\alpha\varepsilon L$ prefix of one and suffix of the other are close in edit distance, then the starting points of these reads are approximately $(1-\alpha\varepsilon)L$ apart.

Lemma 4.5. *Let $X \in \Sigma^n$ be a uniformly random string, let $L = \left(\overline{C}/\varepsilon\right)\ln(n)$, and let $\widetilde{R}_1, \widetilde{R}_2 \in \widetilde{\mathcal{R}}$ be two corrupted reads of X of length L. Suppose that $\operatorname{ed}\left(\widetilde{R}_1\left[-\alpha\varepsilon L, -1\right], \widetilde{R}_2\left[1, \alpha\varepsilon L\right]\right) \leq (2+2c')\varepsilon L$. If $\overline{C}$ is large enough, then with probability $1 - o\left(n^{-2}\right)$ we have that*

$$T_2 \in \left[T_1 + (1-\alpha\varepsilon)L - \left(2+2c'\right)\varepsilon L, T_1 + (1-\alpha\varepsilon)L + \left(2+2c'\right)\varepsilon L\right]. \tag{4}$$

Proof. Suppose that (4) does not hold. Then the overlap between $R_1\left[-\alpha\varepsilon L, -1\right]$ and $R_2\left[1, \alpha\varepsilon L\right]$ is less than $\left(\alpha - (2+2c')\right)\varepsilon L$. Recall the definition of $c = c(\Sigma)$ from Corollary 4.4. Since $c\alpha\varepsilon L = 4\varepsilon L \geq (2+2c')\varepsilon L$, we can apply Lemmas 4.2 and 4.3 to get that with probability $1 - o\left(n^{-2}\right)$ we have $\operatorname{ed}\left(R_1\left[-\alpha\varepsilon L, -1\right], R_2\left[1, \alpha\varepsilon L\right]\right) \geq 2(2+2c')\varepsilon L$. By the triangle inequality this implies that $\operatorname{ed}\left(\widetilde{R}_1\left[-\alpha\varepsilon L, -1\right], \widetilde{R}_2\left[1, \alpha\varepsilon L\right]\right) \geq (2+4c')\varepsilon L$, which is a contradiction. □

Since the probability that the conclusion of the lemma does not hold is $o\left(n^{-2}\right)$, we can take a union bound over all pairs of corrupted reads and have the conclusion of the lemma apply to all of them with probability $1 - o(1)$.

We are now ready to analyze the algorithm step by step. We show by induction that after each extension step the partially reconstructed sequence is a good approximation of a substring of the original sequence.

Lemma 4.6. *Let $X \in \Sigma^n$ be a uniformly random string, let $L = \left(\overline{C}/\varepsilon\right)\ln(n)$ where $\overline{C}$ is a large enough constant, and let $N = C' N_{\mathrm{cov}}/\varepsilon$. Let $\alpha = 4/c(\Sigma)$, where $c(\Sigma)$ is given by Corollary 4.4. Let Y_i be the state of the partially reconstructed sequence Y after i corrupted reads have been processed by the algorithm; we have $Y_1 = \widetilde{R}_1$. Let τ_1 be the number of reads processed in the first while loop of the algorithm, and let τ_2 be the number of reads processed in the second while loop. Also let $\tau = \tau_1 + \tau_2$, the total number of reads processed during the algorithm. With probability at least $1-\delta$ (over the choice of X and the starting points of the reads in $\mathcal{R}$) we have the following:*

(a) For every $i \leq \tau$ there exist $a_i, b_i \in [n]$ such that $|a_i - b_i| \geq (1 - (\alpha + 2 + 2c')\,\varepsilon)\, iL$ and

$$\operatorname{ed}(Y_i, X[a_i, b_i]) \leq (3 + 2c')\,\varepsilon i L; \tag{5}$$

(b) $\tau \leq \frac{n}{(1-(\alpha+2+2c')\varepsilon)L}$;

(c) $\operatorname{ed}(X[a_\tau, b_\tau], X) \leq 2L$.

Proof. Part (a) of the lemma holds for $i = 1$ by choosing $a_1 = T_1$ and $b_1 = T_1 + L - 1$. For larger i we prove the statement by induction on i.

Suppose we are in the first while loop of the algorithm, i.e., $i \leq \tau_1$. We set $a_i = a_1$ for all $i \leq \tau_1$ and only change b_i. Let $\widetilde{T}_i = T\left(\widetilde{R}_{k(i)}\right)$, where $k(i)$ is the index of the read chosen at the i^{th} round, and set $b_i := \widetilde{T}_i + L - 1$. As mentioned before, we may assume that there is no gap greater than $2c'\varepsilon L$ in between subsequent starting points of the reads. Therefore if $\widetilde{T}_i \leq n - 2L$, then there must exist $R \in \mathcal{R}$ such that $T(R) - \widetilde{T}_i \in [(1 - \alpha\varepsilon)L - c'\varepsilon L, (1 - \alpha\varepsilon)L + c'\varepsilon L]$. By the triangle inequality this implies that $\operatorname{ed}\left(\widetilde{R}_{k(i)}[-\alpha\varepsilon L, -1], \widetilde{R}[1, \alpha\varepsilon L]\right) \leq (2 + 2c')\,\varepsilon L$, i.e., $\widetilde{R}$ satisfies the condition of the while loop. Thus $\widetilde{T}_{\tau_1} > n - 2L$. Now take any $\widetilde{R} \in \widetilde{\mathcal{R}}$ that satisfies the condition of the while loop. By Lemma 4.5 we know that $T\left(\widetilde{R}\right) - \widetilde{T}_i - (1 - \alpha\varepsilon)L \in [-(2 + 2c')\,\varepsilon L, (2 + 2c')\,\varepsilon L]$. By subadditivity and the induction hypothesis we have that

$$\begin{aligned}\operatorname{ed}(Y_{i+1}, X[a_{i+1}, b_{i+1}]) &\leq \operatorname{ed}(Y_i, X[a_i, b_i]) + \operatorname{ed}\left(\widetilde{R}[1 + \alpha\varepsilon L, -1], X[b_i + 1, b_{i+1}]\right) \\ &\leq (3 + 2c')\,\varepsilon i L + \operatorname{ed}\left(\widetilde{R}[1 + \alpha\varepsilon L, -1], X[b_i + 1, b_{i+1}]\right),\end{aligned}$$

so it suffices to estimate the latter term. By (3) we have that $\operatorname{ed}\left(\widetilde{R}[1 + \alpha\varepsilon L, -1], R[1 + \alpha\varepsilon L, L]\right) \leq \operatorname{ed}\left(\widetilde{R}, R\right) \leq \varepsilon L$. Using the definition of b_i we have that

$$\operatorname{ed}(R[1 + \alpha\varepsilon L, L], X[b_i + 1, b_{i+1}]) = \left|\left(\widetilde{T}_i + L - 1\right) - (T(R) + \alpha\varepsilon L - 1)\right| \leq (2 + 2c')\,\varepsilon L,$$

and so by the triangle inequality we have that $\operatorname{ed}\left(\widetilde{R}[1 + \alpha\varepsilon L, -1], X[b_i + 1, b_{i+1}]\right) \leq (3 + 2c')\,\varepsilon L$, proving (5) for all $i \leq \tau_1$. The proof for $i \in [\tau_1, \tau_2]$ is similar, except now $b_i = b_{\tau_1}$ and a_i changes.

We proved that for all $i \leq \tau_1$ we have $\widetilde{T}_{i+1} - \widetilde{T}_i \geq (1 - (\alpha + 2 + 2c')\,\varepsilon)\,L$. A similar statement holds for $i \in [\tau_1, \tau_2]$, and together these imply part (b) of the lemma.

Since we have $a_\tau \leq L$ and $b_\tau \geq n - L$, this implies part (c) of the lemma. □

Putting everything together and using the triangle inequality we get that the algorithm outputs an estimate $\widehat{X}$ which satisfies

$$\operatorname{ed}\left(X, \widehat{X}\right) \leq \frac{3 + 2c'}{1 - (\alpha + 2 + 2c')\,\varepsilon}\varepsilon n + 2L,$$

which proves Theorems 3.2 and 3.3.

5 Discussion and future work

We introduced an adversarial error model for the problem of sequence assembly from shotgun reads. Our main result shows that if the reads are long enough and there is high enough coverage,

then approximate reconstruction of the original sequence is possible for almost all sequences. The main question our work leaves open is: what are the fundamental information-theoretic limits to approximate reconstruction? Given $\overline{L}$ and N, is approximate reconstruction possible? If so, what is the best approximation factor achievable? What is the best "strategy" for an adversary that can corrupt the reads?

The probabilistic model we consider for the sequence of interest is simplistic, and it would be worthwhile to consider more general distributions, such as a Markov chain model. However, in many genomes there are long repeats, which are not captured by a Markov model. A direction for future research is to understand the fundamental limits to approximate reconstruction for arbitrary sequences as a function of their (approximate) repeat statistics.

Our adversarial error model also contains a simplification: sequencing technologies typically do not have a uniform error rate. Instead, while the error rate is reasonably small for most reads, there are some where the error rate is large and the resulting reads are useless. Practitioners can often detect these bad reads and thus throw them away. A variant of our algorithm can also handle very bad reads if the quality of good and very bad reads are sufficiently separated: the very bad reads simply will not align anywhere and so will be thrown out. However, if there is a continuous spectrum of quality from good to very bad reads, the algorithm runs into issues due to the reads in the middle of the spectrum. We leave addressing this issue as a future challenge.

Acknowledgements

The research of E.M. is supported by NSF grant CCF-1320105, DOD ONR grant N00014-14-1-0823, and Simons Foundation grant 328025. M.Z.R. thanks Jasmine Nirody and Rachel Wang for helpful discussions.

References

[1] G. Bresler, M. Bresler, and D. Tse. Optimal Assembly for High Throughput Shotgun Sequencing. *BMC Bioinformatics*, 14(5):S18, 2013.

[2] C.-S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, and J. Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, 10(6):563–569, 2013.

[3] M. Dyer, A. Frieze, and S. Suen. The probability of unique solutions of sequencing by hybridization. *Journal of Computational Biology*, 1(2):105–110, 1994.

[4] E. S. Lander and M. S. Waterman. Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis. *Genomics*, 2(3):231–239, 1988.

[5] H. Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Preprint available at `http://arxiv.org/abs/1512.01801`, 2015.

[6] N. J. Loman, J. Quick, and J. T. Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12:733–735, 2015.

[7] A. Motahari, G. Bresler, and D. Tse. Information Theory of DNA Shotgun Sequencing. *IEEE Transactions on Information Theory*, 59(10):6273–6289, 2013.

[8] A. Motahari, K. Ramchandran, D. Tse, and N. Ma. Optimal DNA shotgun sequencing: Noisy reads are as good as noiseless reads. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1640–1644, 2013.

[9] S. L. Salzberg. Mind the gaps. *Nature Methods*, 7(2):105–106, 2010.

[10] I. Shomorony, T. Courtade, and D. Tse. Do Read Errors Matter for Genome Assembly? In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 919–923, 2015.

[11] E. Ukkonen. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92(1):191–211, 1992.

A Proofs of edit distance results

Proof of Lemma 4.1. For any m and n we clearly have

$$\begin{aligned}&\mathrm{ed}\left(X_{m+n}, Y_{m+n}\right)\\ &\qquad\leq \mathrm{ed}\left(X_{m+n}\left[1,m\right], Y_{m+n}\left[1,m\right]\right)+\mathrm{ed}\left(X_{m+n}\left[m+1,m+n\right], Y_{m+n}\left[m+1,m+n\right]\right).\end{aligned}$$

Thus Kingman's subadditive ergodic theorem implies that $\lim_{m\to\infty}\frac{1}{m}\mathrm{ed}\left(X_m, Y_m\right) =: c_{\mathrm{ind}}$ exists almost surely. Clearly $c_{\mathrm{ind}} \geq 0$; what remains to show is that $c_{\mathrm{ind}} > 0$. We do this via a volume argument; we first present a simple argument and then refine it to get a better lower bound on c_{ind}.

If $\mathrm{ed}\left(X_m, Y_m\right) \leq r$, then one can get from X_m to Y_m using at most r deletions, insertions and substitutions. The locations of the at most r deletions and substitutions can be chosen in at most $\binom{m}{r}$ ways, and the same holds for the locations of the at most r insertions and substitutions. Given the locations of these, there can be at most $|\Sigma|^r$ subsequences in these locations. That is, the edit distance ball of radius r around any point $x \in \Sigma^m$ contains at most $\binom{m}{r}^2 |\Sigma|^r$ points of Σ^m. For $r = \delta m$ we get

$$\binom{m}{\delta m}^2 |\Sigma|^{\delta m} \approx 2^{2H(\delta)m} |\Sigma|^{\delta m},$$

where $H\left(x\right) = -x\log_2\left(x\right) - \left(1-x\right)\log_2\left(1-x\right)$ is the binary entropy function. Note that the total number of sequences of length m is $|\Sigma|^m$. Let $\delta^* = \delta^*\left(\Sigma\right)$ be the unique solution in $(0,1)$ of $4^{H(\delta)}|\Sigma|^{\delta} = |\Sigma|$. By the volume argument above we have that for any $\delta < \delta^*$ the probability that $\mathrm{ed}\left(X_m, Y_m\right) \leq \delta m$ is exponentially small in m. Thus $c_{\mathrm{ind}} \geq \delta^* > 0$. In particular, for $|\Sigma| = 2$, we have $\delta^* \approx 0.09488$, and for $|\Sigma| = 4$, we have $\delta^* \approx 0.22709$.

We can obtain a better bound by a slightly more careful argument. Again, if $\mathrm{ed}\left(X_m, Y_m\right) \leq r$, then one can get from X_m to Y_m using at most r deletions, insertions and substitutions. Suppose that the number of deletions is D, the number of insertions is I, and the number of substitutions is S. Since X_m and Y_m have the same length, we have $D = I$ and also $D + I + S \leq r$, i.e., $S \leq r - 2D$. The locations of the D deletions can be chosen in at most $\binom{m}{D}$ ways, the locations of the I insertions can be chosen in at most $\binom{m}{I}$ ways, while the locations of the S substitutions can be chosen in at most $\binom{m}{S}$ ways. Given the locations of these, there can be at most $|\Sigma|^I$ subsequences in the locations of the insertions, and at most $|\Sigma|^S$ subsequences in the locations of the substitutions. Therefore the edit distance ball of radius r around any point $x \in \Sigma^m$ contains at most

$$\max_{0\leq D\leq r/2}\left\{\binom{m}{D}^2\binom{m}{r-2D}|\Sigma|^{r-D}\right\}$$

points of Σ^m. For $r = \delta m$ and $D = \delta_D m$ we have

$$\binom{m}{\delta_D m}^2 \binom{m}{(\delta - 2\delta_D)\, m} |\Sigma|^{(\delta-\delta_D)m} \approx 2^{(2H(\delta_D)+H(\delta-2\delta_D)-\delta_D \log_2(|\Sigma|))m} |\Sigma|^{\delta m}.$$

Let $\delta^{**} = \delta^{**}(\Sigma)$ be the unique solution in $(0,1)$ of

$$2^{\max_{0\le x\le \delta/2}\{2H(x)+H(\delta-2x)-x\log_2(|\Sigma|)\}} |\Sigma|^{\delta} = |\Sigma|.$$

The volume argument thus tells us that for every $\delta < \delta^{**}$ the probability that $\mathrm{ed}(X_m, Y_m) \le \delta m$ is exponentially small in m. Thus $c_{\mathrm{ind}} \ge \delta^{**}$. In particular, for $|\Sigma| = 2$, we have $\delta^{**} \approx 0.15776$, and for $|\Sigma| = 4$, we have $\delta^{**} \approx 0.33832$. □

Proof of Lemma 4.2. By (3) we have that

$$\mathrm{ed}(X[1,m], X[k+1,k+m]) \ge \mathrm{ed}(X[1,dm], X[k+1,k+dm]).$$

Since $k+1 > dm$, the strings $X[1,dm]$ and $X[k+1,k+dm]$ are independent uniformly random strings of length dm. Recall the definition of δ^{**} from the proof of Lemma 4.1 and let $\delta \in (0,\delta^{**})$. In the proof of Lemma 4.1 we showed that the probability that $\mathrm{ed}(X[1,dm], X[k+1,k+dm]) \le \delta dm$ is exponentially small in dm. By taking a union bound over $k \in [dm, m]$ we arrive at the desired result with, e.g., $\gamma(d,\Sigma) = 0.9 \times \delta^{**}(\Sigma)\, d$, and an appropriate constant c'. □

Before proving Lemma 4.3 we introduce a variant of the edit distance which is simpler to understand theoretically and for which we state and prove a result similar to Lemma 4.3. We denote by $\underline{\mathrm{ed}}(x,y)$ the minimum number of deletion or insertion operations necessary to go from x to y; that is, compared to the edit distance, substitutions are not allowed. Since a substitution can be simulated by a deletion followed by an insertion, we have that $\mathrm{ed}(x,y) \le \underline{\mathrm{ed}}(x,y) \le 2\mathrm{ed}(x,y)$ for all $x, y \in \Sigma^*$. The nice property of this distance is that $\underline{\mathrm{ed}}(x,y) = |x| + |y| - 2\mathrm{LCS}(x,y)$, where $\mathrm{LCS}(x,y)$ is the length of the longest common subsequence (LCS) of x and y. The following result is about the longest common subsequence of two random strings and is similar to Lemma 4.3.

Lemma A.1. *Let $X \in \Sigma^{2m}$ be a uniformly random string. There exist positive constants $c = c(\Sigma)$ and $c' = c'(\Sigma)$ such that*

$$\mathrm{LCS}(X[1,m], X[1+k,m+k]) = m - k$$

for all $k \le cm$ with probability at least $1 - e^{-c'm}$.

Proof. It is immediate that

$$\mathrm{LCS}(X[1,m], X[1+k,m+k]) \ge m - k,$$

since the last $m-k$ coordinates of $X[1,m]$ and the first $m-k$ coordinates of $X[1+k,m+k]$ are the same. What remains is to show that the probability of

$$\mathrm{LCS}(X[1,m], X[1+k,m+k]) > m - k \tag{6}$$

is exponentially small in m. Note that if a common subsequence of $X[1,m]$ and $X[1+k,m+k]$ is such that the ℓ^{th} coordinate of $X[1,m]$ is mapped to the $(\ell-k)^{\mathrm{th}}$ coordinate of $X[1+k,m+k]$ (the "trivial" map), then this subsequence can have length at most $m-k$, since in $X[1,m]$ there are only $m-\ell$ coordinates to the right of this coordinate, while in $X[1+k,m+k]$ there are only

$\ell - k - 1$ coordinates to the left of this coordinate. So if a common subsequence has length greater than $m - k$, then every coordinate is mapped "nontrivially". This then creates many constraints on the pair of sequences and so there will not be many of them, as we now argue.

Suppose that $\mathrm{LCS}\left(X\left[1,m\right], X\left[1+k, m+k\right]\right) = m - \ell$ for some $\ell \in \{0, 1, \ldots, k-1\}$. A noncrossing matching between $m - \ell$ coordinates of the two sequences is characterized by the ℓ coordinates in each sequence that are not part of the matching; the remaining pairs of the matching are determined due to the noncrossing property. Thus there are $\binom{m}{\ell}^2$ such matchings. The coordinates of a LCS between the two sequences corresponds to a noncrossing matching between the two, and it imposes conditions on the values of these coordinates. If $\mathrm{LCS}\left(X\left[1,m\right], X\left[1+k, m+k\right]\right) = m - \ell$ then all but ℓ coordinates of $X\left[1+k, m+k\right]$ are determined by $X\left[1,m\right]$. Furthermore, there are at least $m-k-\ell$ coordinates $j \in [1+k, m]$ such that the $(j-k)^{\text{th}}$ coordinate of $X\left[1+k, m+k\right]$ is in the matching, and thus the value of $X\left[j\right]$ is determined by a previous value $X\left[i\right]$ for some $i < j$. This means that at most $k+\ell$ coordinates of $X\left[1,m\right]$ are not determined by the value of a previous coordinate. So the probability of any given noncrossing matching being a common subsequence is at most $|\Sigma|^{(k+2\ell)-(m+k)} = |\Sigma|^{2\ell-m}$. We have thus shown that the probability of (6) is at most

$$\sum_{\ell=0}^{k-1} \binom{m}{\ell}^2 |\Sigma|^{2\ell-m} \leq k \binom{m}{k}^2 |\Sigma|^{2k-m}.$$

When $k = dm$, then this is approximately $(dm) \times 2^{\{2H(d)+(2d-1)\log_2(|\Sigma|)\}m}$, which goes to zero exponentially in m if $d \in (0,1)$ is small enough. Taking a union bound over k we have that this holds for all $k \in \{0, 1, \ldots, dm\}$ simultaneously. □

Proof of Lemma 4.3. The proof is very similar to the one above. It is again immediate that $\mathrm{ed}\left(X\left[1,m\right], X\left[1+k, m+k\right]\right) \leq 2k$, since one can obtain $X\left[1+k, m+k\right]$ from $X\left[1,m\right]$ by first deleting the first k coordinates of $X\left[1,m\right]$ and then inserting $X\left[m+1, m+k\right]$ at the end of the sequence. What remains is to show that the probability of

$$\mathrm{ed}\left(X\left[1,m\right], X\left[1+k, m+k\right]\right) < 2k \tag{7}$$

is exponentially small in m. If (7) holds then one can get from $X\left[1,m\right]$ to $X\left[1+k, m+k\right]$ by first performing S substitutions to get $X'\left[1,m\right]$, then performing D deletions and finally I insertions. These quantities have to satisfy $S + D + I \leq 2k - 1$ and $D = I$. We thus have that $\underline{\mathrm{ed}}\left(X'\left[1,m\right], X\left[1+k, m+k\right]\right) \leq 2k - 1 - S$, and so $\mathrm{LCS}\left(X'\left[1,m\right], X\left[1+k, m+k\right]\right) \geq m - k + (1+S)/2$. Let $m - \ell := \mathrm{LCS}\left(X'\left[1,m\right], X\left[1+k, m+k\right]\right)$. Again, the number of such noncrossing matchings is $\binom{m}{\ell}^2$. As in the proof of the previous lemma, the noncrossing matching corresponding to such a LCS has to map every coordinate "nontrivially". Every such matching imposes constraints on $X'\left[1,m\right]$ and $X\left[1+k, m+k\right]$, and while there are less constraints as before—since $X'\left[1,m\right]$ differs from $X\left[1,m\right]$ in S substitutions—we can still show that the probability of (7) is at most

$$k \binom{m}{2k} \binom{m}{k}^2 |\Sigma|^{3k-m}.$$

When $k = dm$, then this is approximately $(dm) \times 2^{\{2H(d)+H(2d)+(3d-1)\log_2(|\Sigma|)\}m}$, which goes to zero exponentially in m if $d \in (0,1)$ is small enough; in particular this happens when $d \leq 0.0846092$. Taking a union bound over k we have that this holds for all $k \in \{0, 1, \ldots, dm\}$ simultaneously. □