Reconstruction from Substrings with Partial Overlap

Yonatan Yehezkeally†, Daniella Bar-Lev‡, Sagi Marcovich‡, and Eitan Yaakobi‡
†Institute for Communications Engineering, Technical University of Munich, 80333 Munich, Germany
‡Department of Computer Science, Technion—Israel Institute of Technology, Haifa 3200003, Israel

Abstract—This paper introduces a new family of reconstruction codes which is motivated by applications in DNA data storage and sequencing. In such applications, DNA strands are sequenced by reading some subset of their substrings. While previous works considered two extreme cases in which all substrings of some fixed length are read or substrings are read with no overlap, this work considers the setup in which consecutive substrings are read with some given minimum overlap. First, an upper bound is provided on the attainable rates of codes that guarantee unique reconstruction. Then, we present efficient constructions of asymptotically optimal codes that meet that upper bound.

I. INTRODUCTION

String reconstruction refers to a large class of problems where information about a string can only be obtained in the form of multiple, incomplete and/or noisy observations. Examples of such problems are the reconstruction problem by Levenshtein [11], the trace reconstruction problem [3], and the k-deck problem [13].

Notably, when observations are comprised of unordered consecutive substrings, the reconstruction from substring-compositions problem [1], [4], [9], [14], [16], [23] and the torn-paper problem [2], [17], [18], [20] (closely related to the shuffling channel [8], [10], [22]) have received significant interest in the past decade due to applications in DNA- or polymer-based storage systems, resulting from contemporary sequencing technologies [4], [7], [16]. The former arises from an idealized assumption of full overlap (and uniform coverage) in read substrings, while the latter results from an assumption of no overlap; in applications, this models the question of whether the complete information string may be replicated and uniformly segmented for sequencing, or if segmentation occurs adversarially in the medium prior to sequencing.

Motivated by these two paradigms, we study in this paper a generalized (or intermediate) setting where an information string is observed through an arbitrary collection of its substrings, where the minimum length \( L_{\text{min}} \) of each retrieved substring, as well as the length \( L_{\text{over}} \) of overlap between consecutive substrings, are bounded from below. A similar setting was recently studied in [19], where both substrings’ lengths and overlap were assumed to be random; we study the problem in the aforementioned worst-case regime. We ask what the minimum value of \( L_{\text{min}} \) is for which there exist codes of length-\( n \) strings allowing for unique reconstruction in this channel with asymptotically non-vanishing rates, and then what is the asymptotically optimal obtainable rate given the value of \( L_{\text{over}} \). Having answered both questions, we demonstrate that in these regimes it is possible to efficiently encode and decode information for unique reconstruction attaining asymptotically optimal rates.

The rest of this paper is organized as follows. In Section II we present notation and definitions used throughout the paper. In Section III we bound from above the asymptotically attainable rate of codes for unique reconstruction as a function of \( L_{\text{min}} \), \( L_{\text{over}} \), and then in Section IV we develop efficient encoding and decoding algorithms for such codes, asymptotically meeting this bound. Due to space limitations, some proofs have been delegated to an extended version in preprint on arXiv.org.

II. DEFINITIONS AND PRELIMINARIES

Let \( \Sigma \) be a finite alphabet of size \( q \). Where advantageous, we assume \( \Sigma \) is equipped with a ring structure, and in particular identify elements \( 0, 1 \in \Sigma \). For a positive integer \( n \) we denote the set \( [n] \triangleq \{0, 1, \ldots, n-1\} \).

For two non-negative functions \( f, g \) of a common variable \( n \), denoting \( L \triangleq \limsup_{n \to \infty} \frac{f(n)}{g(n)} \) (in the wide sense) we say that \( f = o(n) \) if \( L = 0 \), \( f = \Omega(n) \) if \( L > 0 \), \( f = \Theta(n) \) if \( L < \infty \), and \( f = \omega(n) \) if \( L = \infty \). We say that \( f = \Theta_n(g) \) if \( f = \Omega_n(g) \) and \( f = O_n(g) \). If \( f \) is not positive, we say \( f = O_n(g) \) (\( f = o_n(g) \)) if \( |f| = O_n(g) \) (respectively, \( |f| = o_n(g) \)). Interchange of context, we omit the subscript from aforementioned notations.

Let \( \Sigma^* \) denote the set of all finite strings over \( \Sigma \). The length of a string \( x = (x_0, x_1, \ldots, x_{n-1}) \in \Sigma^* \) is denoted by \( |x| = n \). For strings \( x, y \in \Sigma^* \), we denote their concatenation by \( x \circ y \). We say that \( v \) is a substring of \( x \) if there exist strings \( u, w \) such that \( x = u \circ v \circ w \). If \( u \) (respectively, \( w \)) is empty, we say that \( v \) is a prefix (suffix) of \( x \). If the length of \( v \) is \( \ell \), we specifically say that \( v \) is an \( \ell \)-substring of \( x \) (similarly, an \( \ell \)-pre/suffix). For \( I \subseteq [|x|] \), we let \( x_I \) denote the subsequence of \( x \) obtained by restriction to the coordinates of \( I \); specifically, for \( i \in [|x| - \ell + 1] \) we denote by \( x_{i+\ell} \) the \( \ell \)-substring of \( x \) at location \( i \) (we reserve the term index for a different use), where \( i + \ell = \{i + j : j \in \ell \} \).

We consider in this paper the problem of string reconstruction from substrings with partial overlap. That is, we assume that a message \( x \in \Sigma^* \) is observed only through a multiset of its substrings, without order, with the following restrictions: (i) all observed substrings are of length at least \( L_{\text{min}} \); and (ii) preceding substrings overlap with length at least \( L_{\text{over}} \) (in particular, every symbol of \( x \) is observed in some substring).
More formally, a substring-trace of \( x \in \Sigma^n \) is a multiset \( \{ (x_{ij} + t_j) : 1 \leq j \leq m \} \), for some \( m \in \mathbb{N} \), such that \( i_1 < i_2 < \cdots < i_m \) and \( \ell_j \in [n - i_j + 1] \). A substring-trace is complete if \( \ell_1 = 0 \), \( i_j + 1 < i_j + \ell_j \) for all \( j < m \), and \( i_m + \ell_m - 1 = n \). A complete substring-trace of \( x \in \Sigma^n \) is called \((L_{\min}, L_{\overline{\text{over}}})\)-trace if \( \ell_j \geq L_{\min} \geq L_{\overline{\text{over}}} \) for all \( j \), and \( i_j + \ell_j - i_j + 1 \geq L_{\overline{\text{over}}} \) for all \( j < m \). For example, for \( x = 111011101111 \)

- \{11101111, 111011010, 110111\}\} is a \((6, 2)\)-trace of \( x \);
- \{11101111, 110101, 11011\} is a complete substring-trace of \( x \) which is not a \((6, 2)\)-trace; and
- \{110111, 110101, 11011\}\} is a substring-trace of \( x \) which is not complete (since \( i_1 > 0 \)).

The \((L_{\min}, L_{\overline{\text{over}}})\)-trace spectrum of \( x \in \Sigma^n \), denoted \( T_{L_{\min}}(x) \), is the set of all \((L_{\min}, L_{\overline{\text{over}}})\)-traces of \( x \). Our channel accepts \( x \in \Sigma^n \) and outputs some \((L_{\min}, L_{\overline{\text{over}}})\)-trace of \( x \) error free.

For all \( C \subseteq \Sigma^n \) we denote the rate, redundancy of \( C \) by \( R(C) \triangleq \frac{1}{n} \log|C| \), red\((C) \triangleq n - \log|C| \), respectively. Throughout the paper, we use the base-\( q \) logarithms. Motivated by the above channel definition, a code \( C \subseteq \Sigma^n \) is called an \((L_{\min}, L_{\overline{\text{over}}})\)-trace code if for all \( x_1 \neq x_2 \in C \), \( T_{L_{\min}}(x_1) \cap T_{L_{\overline{\text{over}}}}(x_2) = \emptyset \). The main goal of this work is to find, for \( L_{\min}, L_{\overline{\text{over}}} \), as functions of \( n \), the maximum asymptotic rate of \((L_{\min}, L_{\overline{\text{over}}})\)-trace codes. We will also be interested in efficient constructions of \((L_{\min}, L_{\overline{\text{over}}})\)-trace codes with rate asymptotically approaching that value.

For convenience of analysis, we denote by \( L_{L_{\min}}^2(x) \in T_{L_{\min}}(x) \), for \( x \in \Sigma^n \), the \((L_{\min}, L_{\overline{\text{over}}})\)-trace of \( x \) containing specifically its \( L_{\min}\)-prefix, and subsequent \( L_{\min}\)-substrings overlapping in precisely \( L_{\overline{\text{over}}} \)-coordinates. For example, if \( x = 1110110101111 \) then

\[
L_{L_{\min}}^2(x) = \{1110, 1011, 1110, 1010, 1011, 1111\}.
\]

(Here, if \( L_{\min} - L_{\overline{\text{over}}} \) does not divide \( n - L_{\min} \) we allow the \( L_{\min}\)-suffix to contain a longer overlap with its preceding \( L_{\min}\)-substring; in the sequel we assume for ease of presentation that this does not occur, a fact that again shall not affect our asymptotic analysis.) Since \( L_{L_{\min}}^2(x) \in T_{L_{\min}}(x) \) for all \( x \in \Sigma^n \), observe that any \((L_{\min}, L_{\overline{\text{over}}})\)-trace code \( C \subseteq \Sigma^n \) satisfies

\[
|C| \leq \left| \left\{ L_{L_{\min}}^2(x) : x \in \Sigma^n \right\} \right|.
\]

Applying the profile-vectors argument [5], we count the incidences of each possible \( u \in \Sigma^{L_{\min}} \) in \( L_{L_{\min}}^2(x) \) and observe that the sum of incidences equals \( 1 + \left[ \frac{n-L_{\min}}{L_{L_{\min}}-L_{\overline{\text{over}}}} \right] = \left[ \frac{n-L_{\min}}{L_{\min}-L_{\overline{\text{over}}}} \right] \); thus, we have an embedding of \( \{ L_{L_{\min}}^2(x) : x \in \Sigma^n \} \) into

\[
\left\{ f : \Sigma^{L_{\min}} \to \mathbb{N} : \sum_{u \in \Sigma^{L_{\min}}} f(u) = \left[ \frac{n-L_{\overline{\text{over}}}}{L_{\min}-L_{\overline{\text{over}}}} \right] \right\},
\]

and therefore

\[
|C| \leq \left( \left[ \frac{n-L_{\overline{\text{over}}}}{L_{\min}-L_{\overline{\text{over}}}} \right] + q \left[ \frac{L_{\min}-L_{\overline{\text{over}}}}{q} \right] - 1 \right)
\]

\[
\leq \left( \left[ \frac{n-L_{\overline{\text{over}}}}{L_{\min}-L_{\overline{\text{over}}}} \right] + q L_{\min} \right).
\]

Before concluding this section we discuss the pertinent notion of repeat-free strings [6], which we denote herein for all \( l \leq n \) by

\[
\mathcal{RF}_l(n) \triangleq \{ x \in \Sigma^n : x_{i+l} \neq x_{j+l}, \forall i < j \in [n - l + 1] \}.
\]

It was observed in [21] that for \( x \in \mathcal{RF}_l(n) \), \( L_{L_{\min}}^2(x) \neq L_{L_{\min}}^2(y) \) for all \( y \in \Sigma^n \setminus \{ x \} \). A straightforward generalization of the arguments therein demonstrates the following lemma.

**Lemma 1** Given \( L_{\min} > L_{\overline{\text{over}}} \), for all \( x \in \mathcal{RF}_{L_{\min}}(n) \), there exists an efficient algorithm reconstructing \( x \) from any \((L_{\min}, L_{\overline{\text{over}}})\)-trace of \( x \).

**Proof:** Let \( T \) be any \((L_{\min}, L_{\overline{\text{over}}})\)-trace of \( x \). For any \( u \in T \), suppose by negation that there exist \( v_1, v_2 \in T \), \( v_1 \neq v_2 \), such that the \( L_{\min}\)-prefix of \( v_i \) equals the \( L_{\min}\)-prefix of \( u \), where \( \ell_i \geq L_{\overline{\text{over}}} \), for \( i \in \{1, 2\} \). Since \( v_1 \neq v_2 \), they occur in distinct locations in \( x \), and in particular their min\( \{\ell_1, \ell_2\}\)-suffix occurs in distinct locations; this in contradiction to \( x \in \mathcal{RF}_{L_{\min}}(n) \). The same argument proves that there do not exist \( v_1, v_2 \in T \), \( v_1 \neq v_2 \), such that the \( L_{\min}\)-prefix of \( v_i \) equals the \( L_{\min}\)-prefix of \( u \), where again \( \ell_i \geq L_{\overline{\text{over}}} \), for \( i \in \{1, 2\} \).

Hence, matching prefix to suffix, of lengths at least \( L_{\overline{\text{over}}} \), one reconstructs \( x \) from \( T \). Equivalently, for each \( u \in T \), finding the unique \( v \in T \) that contains the \( L_{\overline{\text{over}}} \)-prefix of \( u \) as a substring (which exists unless \( u \) itself is a prefix of \( x \)) results with complete reconstruction. A naive implementation requires \( O(n^2 L_{\overline{\text{over}}} \) run-time.

Further, we note that if \( \lim \inf L_{\overline{\text{over}}}/\log(n) > 1 \), then [6] showed that \( \mathcal{RF}_{L_{\min}}(n) \) forms a rate \( 1 - o(1) \) code in \( \Sigma^n \) with an efficient encoder/decoder pair; we summarize their result in the following lemma.

**Lemma 2** [6, Sec. V] For \( k = \lceil 2 \log(\log(n) \rceil + 5(k + 2) \lceil n/2k \rceil \) there exists an efficient encoder into \( \mathcal{RF}_{l \log(n) + 5(k + 2) \lceil n/2k \rceil} \), which in addition produces strings containing no \( (2k + 2) \)-length run of zeros. Redundancy is introduced only in the initialization phase encoding into

\[
Z(n, k) \triangleq \{ u \in \Sigma^n : u \text{ has no length-}k \text{ run of zeros} \}
\]

(this is the well-understood Run-length-limited (RLL) constraint; see, e.g., [15, Sec. 1.2]). Further, an efficient decoder exists for the provided encoder.

Analysis of the asymptotic rate achieved by the encoder of Lemma 2 is given in the following lemma.

**Lemma 3** There exist efficient encoders into \( Z(n, k) \) requiring \( \left[ \frac{2k n}{q} \right] \) redundant symbols for \( q > 2 \) [23, Lem. 5], or \( \left[ \frac{2k n}{q} \right] \) for \( q = 2 \) [12, Sec. III].
For our purposes, however, it will be beneficial to observe that the arguments used in [6, Sec. VI] apply without change to any other choice of $k$, resulting in the following corollary. It demonstrates that $R F_{\ell} (n)$ forms a rate $1 - o_n (1)$ code in $\Sigma^n$ as long as $\ell - \log(n) = \omega (1)$.

**Corollary 4** For $\ell (n) = [\log (n)] + \omega_n (1)$ and any $t \leq 2 \left( \ell (n) - \left[ \log (n) \right] \right)/5 - 3$

there exists an efficient encoder/decoder pair into $R F_{\ell (n)} (n)$, producing strings containing no $t$-length run of zeros, and requiring at most $\frac{n^2}{q^2 \cdot 2^{\ell (n)/2}}$ redundant symbols for $q > 2$ or $2 \left[ 4 \cdot 2^{-\ell (n)/2} \right]$ symbols for $q = 2$, i.e., rate $1 - O_n (q^{-t/2})$.

Proof: For any $s \leq \left( \ell (n) - \left[ \log (n) \right] \right)/5 - 2$, letting $k = s$ in Lemma 2 produces $\ell'$-repeat-free strings, for some $\ell' \leq \ell (n)$, hence in particular also $\ell (n)$-repeat-free strings, containing no $(2s + 2)$-length run of zeros.

Letting $s \triangleq \left[ \ell (n)/2 \right] - 1$ we encode $\ell (n)$-repeat-free strings containing no $t$-length run of zeros. Finally, the encoders of Lemma 3 then require the claimed redundancy.

Given Corollary 4 and the preceding discussion, we focus in the sequel on the complement, unsolved case of $\limsup L_{\text{over}} / \log(n) < 1$.

III. BOUNDS

In this section we demonstrate an upper bound on the achievable asymptotic rate of $(L_{\min}, L_{\overline{\text{over}}})$-trace codes.

**Lemma 5** If $L_{\min} = a \log (n) + o_n (1)$ and $L_{\overline{\text{over}}} = \gamma L_{\min} + O_n (1)$, for some $a > 1$ and $0 < \gamma \leq \frac{1}{a}$, then any $(L_{\min}, L_{\overline{\text{over}}})$-trace code $C \subseteq \Sigma^n$ satisfies

$$R (C) \leq \frac{1 - 1/a}{1 - \gamma} + O \left( \frac{\log (n)}{\log (n)} \right).$$

(Not that $\frac{1 - 1/a}{1 - \gamma} \leq 1$ if and only if $\lim L_{\overline{\text{over}}} / \log (n) = \gamma a \leq 1$.)

Proof: Observe for $v \geq u \geq 0$

$$\log \left( \frac{u + v}{u} \right) \leq \log \left( \frac{u + v}{u} \right)^{u} < \log \left( \frac{u + v}{u} \right)^{u}$$

$$= u \log (e) + \log \left( 1 + \frac{v}{u} \right)$$

$$= u \log (e) + \log \left( 1 + \frac{v}{u} \right) + \log \left( \frac{u}{u} \right)$$

$$\leq u \left( 1 + \frac{v}{u} \right) \log (e) + \log \left( \frac{u}{u} \right)$$

$$\leq u \left( 2 \log (e) + \log \left( \frac{u}{u} \right) \right).$$

Letting $u \triangleq q^{L_{\min}}$ and $v \triangleq \left[ \frac{a}{L_{\min} - L_{\overline{\text{over}}}} \right] \leq v$, we observe that $\log \left( \frac{u}{v} \right) = O (\log (n))$ and by Eq. (2), $|C| \leq \left( \frac{u + v}{u} \right)^{u}$, hence we have

$$\log |C| \leq n \left( \frac{L_{\min} - \log (n)}{L_{\min} - L_{\overline{\text{over}}}} + \log (L_{\min} - L_{\overline{\text{over}}}) + 2 \log (\epsilon) \right)$$

$$+ O (\log (n))$$

$$= n \left( \frac{L_{\min} - \log (n)}{L_{\min} - L_{\overline{\text{over}}}} + O \left( \frac{\log (\log (n))}{\log (n)} \right) \right)$$

$$= n \left( \frac{1 - 1/a}{1 - \gamma} + O \left( \frac{\log (\log (n))}{\log (n)} \right) \right).$$

In particular, Lemma 5 implies the following lower bound on $L_{\min}$ for the existence of codes with asymptotically non-vanishing rates.

**Corollary 6** If $\limsup_n L_{\min} / \log (n) \leq 1$, then $R (C) = o_n (1)$ for any $(L_{\min}, L_{\overline{\text{over}}})$-trace code $C \subseteq \Sigma^n$.

IV. CONSTRUCTION

In this section we present an efficient encoder for $(L_{\min}, L_{\overline{\text{over}}})$-trace codes, achieving asymptotically optimal rate, for the case $\limsup L_{\overline{\text{over}}} / \log (n) \leq 1$. Throughout the section, we let $L_{\min} \triangleq a \log (n)$ and $L_{\overline{\text{over}}} \triangleq \gamma L_{\min}$, for some $a > 1$ and $0 < \gamma \leq 1/a$. Further, we let $f$ be any integer function satisfying $f (n) = \omega (1)$ and $f (n) = o (\log (n))$, and finally $I \triangleq \left[ \frac{1 - 1/a}{1 - \gamma} \right] \log (n) + (\log (n))^{0.5 + \epsilon}$ for some small $\epsilon > 0$. In the sequel we tacitly assume that $q^I L_{\min}$ divides $n$.

Construction A The construction is based on the following two ingredients:

- **Index generation.** Let $(c_i)_{i \in [F]}$, $c_i \in \Sigma^I$ be indices in ascending lexicographic order. We encode each $c_i$ independently as follows (see Figure 1). Denoting $F \triangleq \left[ f / f (n) \right] - 1$, we partition $c_i$ into $F + 1$ non-overlapping segments; more formally, define $\left\{ c_i^{(k)} \right\}_{k \in [F]} \subseteq \Sigma^f (n)$, and $c_i^{(F)} \in \Sigma^{f (n)}$, by

$$c_i^{(0)} \circ c_i^{(1)} \circ \cdots \circ c_i^{(F)} = c_i.$$

Lastly, denote $c_i^{(k)} \triangleq 1 \circ c_i^{(k)} \circ 1 \in \Sigma^f (n) + 2$ (similarly, $c_i^{(F)}$). We refer to $c_i$ (or simply $i$) as an index in the construction, and to $\left\{ c_i^{(k)} \right\}_{k \in [F + 1]}$ as segments of an encoded index.

- **Repeat-free (RF) encoding.** For $\ell \in \mathbb{N}$, we use an RF encoder which receives strings of length $m$ and returns strings of length $n_{m, \ell} (m)$ which contain no repeated substrings of length $\ell$, and no zero-runs of length $f (n)$ (see Corollary 4). We denote such an encoder by $E_{m, \ell}^{R, F}$.
The constructed code, denoted by $C_A(n)$, is carried as follows. Denote
\[ r \triangleq f(n) + 3 + (F + 1)(f(n) + 2), \]
\[ \ell \triangleq \left\lfloor \frac{L_{over} - 2f(n) - 5}{1 + (f(n) + 2)\left\lfloor \frac{L_{min} - r}{2L_{min}} \right\rfloor} \right\rfloor. \]

Let $m$ be an integer such that
\[ N_{n,\ell}(m) = q^{-1}n(1 - r / L_{min}). \]

(See the proof of Theorem 8 for an explanation of why such $m$ exists for the choices of $n, L_{min}, f(n), I, r, \ell$, and the observation that $m \geq N_{n,\ell}(m)(1 - O(q^{-f(n)/2}))$.) An information string $x \in \Sigma^q m$ is first partitioned into $q^I$ non-overlapping segments $x = x_0 \circ \cdots \circ x_{q^I - 1}$, where $x_i \in \Sigma^m$ for each $i \in [q^I]$. Each $x_i$ is then independently encoded into a string $z_i$ of length $q^{-1}n$ as explained below; the motivation for this step, which we later describe more formally, is to satisfy two properties: (i) the index $i$ can be decoded from any $L_{min}$-substring of $z_i$; and (ii) the string $z_i$ can be uniquely reconstructed from an $(L_{min}, L_{over})$-trace of $z_i$. Lastly, we let
\[ \text{Enc}_A(x) \triangleq z = z_0 \circ \cdots \circ z_{q^I - 1}. \]

The encoding $x \mapsto z_i$, for all $i \in [q^I]$, is performed as follows (see Fig. 2).

1. Let $y_i \triangleq E_{m,\ell}^{RF}(x_i)$, where $y_i \in \Sigma^N_{n,\ell}(m)$.

2. Partition $y_i$ into $n/(q^IL_{min})$ non-overlapping segments of length $L_{min} - r$ (recall $(L_{min} - r)n/(q^I L_{min}) = N_{n,\ell}(m)$) by denoting
\[ y_i = y_{i,0} \circ y_{i,1} \circ \cdots \circ y_{i,n/(q^I L_{min}) - 1}. \]

3. For all $j \in [n/(q^I L_{min})]$:
   a) Partition each $y_{i,j}$ into $F + 1$ non-overlapping segments of equal lengths (up to $\pm 1$, if necessary)
   \[ y_{i,j} = y_{i,j}^{(0)} \circ y_{i,j}^{(1)} \circ \cdots \circ y_{i,j}^{(F)}, \]
   b) Combine $\{y_{i,j}^{(k)} : k \in [F + 1]\}$ with segments of the encoded index $i$, as follows. Define for all $k \in [F + 1]$
   \[ z_{i,j}^{(k)} \triangleq y_{i,j}^{(k)} \circ c_i^{(k)}, \]
   then
   \[ z_{i,j} \triangleq \left\{ \begin{array}{ll}
   10^f(n)11 \circ z_{i,j}^{(0)} \circ \cdots \circ z_{i,j}^{(F)}, & j = 0; \\
   10^f(n)01 \circ z_{i,j}^{(0)} \circ \cdots \circ z_{i,j}^{(F)}, & j > 0
   \end{array} \right. \]
   (where we refer to the substrings $10^f(n)01, 10^f(n)11$ as synchronization markers).

4. Concatenate
\[ z_i \triangleq z_{i,0} \circ \cdots \circ z_{i,n/(q^I L_{min}) - 1}. \]

Before we prove the correctness of Construction A, we first analyze $R(C_A)$.

Lemma 7  
1) $r = I + 1 + \frac{2I}{f(n)} + 2f(n) + O(1)$. 
2) Denoting $\lambda \triangleq 1 - \frac{r}{L_{min}}$, we have $\ell = \lambda L_{over} - O(f(n))$.
By part 1 of Lemma 7,
\[ r = \frac{1}{L_{\text{min}}} = \frac{1}{1 - \gamma} + \frac{1}{\log(n)^{0.5}} + O\left(\frac{1}{\sqrt{\log(n)}}\right), \]
hence noting \( R(C_A(n)) = q^\mu m/n \) concludes the proof. \( \blacksquare \)

We note that the choice \( f(n) = \sqrt{\log(n)} \) in Theorem 8 is optimal, since \( \epsilon \) in Construction A must satisfy \( \epsilon \geq \max\left\{ \frac{\log(f(n))}{\log(n)} - 1, -\log(f(n)) \right\} - 0.5 \).

Finally, we prove the correctness of Construction A. We begin with two technical lemmas.

**Lemma 9** Every \( L_{\text{min}} \)-substring \( u \) of \( z \) contains as subsequences at least an \( (I - \mu) \)-suffix of \( c_i \) and an \( \mu \)-prefix of either \( c_i \) or \( c_{i+1} \), for some \( i \in [q^\mu] \) and \( \mu \in [I] \), in identifiable locations.

**Lemma 10** Every \( L_{\text{over}} \)-substring \( v \) of \( z \) contains at least \( \ell \) consecutive symbols of \( y_1 \otime y_{q^\mu} \).

Combining both lemmas, we have the following theorem.

**Theorem 11** For all admissible values of \( n \), the code \( C_A(n) \) is an \( (L_{\text{min}}, L_{\text{over}}) \)-trace code.

**Proof:** Take \( z \in C_A(n) \) and let \( T \subseteq L_{\text{over}}(z) \) be any \( (L_{\text{min}}, L_{\text{over}}) \)-trace of \( z \).

For every \( u \in T \), we extract the \( (I - \mu) \)-suffix of \( c_i \) and an \( \mu \)-prefix of either \( c_i \) or \( c_{i+1} \), for some \( i \), guaranteed by Lemma 9. Observe that if this prefix belongs to \( c_{i+1} \), then \( u \) also contains a complete synchronization marker \( 10^I(n)11 \) (the instance appearing as prefix of \( z_{i+1} \)), hence these two cases may be distinguished. Further, note that the \( \mu \)-prefix of \( c_{i+1} \) equals the \( \mu \)-prefix of \( c_i \), unless the \( (I - \mu) \)-suffix of \( c_i \) is the greatest element of \( \Sigma^I - \mu \) (in lexicographic order). In both cases, one can correctly deduce that the location of \( u \) in \( z \) begins in the segment \( z_i \). It is therefore possible to partition \( T \) by index \( i \) (corresponding to the starting location of each substring).

For each substring \( u \) of index \( i \), intersecting both \( y_i \times y_{i+1} \), \( u \) must contain a complete synchronization marker \( 10^I(n)11 \) (the instance appearing as prefix of \( z_{i+1} \)); hence its location in \( u \) implies the exact location of \( u \) in \( z \). For all other substrings of index \( i \), it holds by Lemma 10, and since each \( y_i \) is \( \ell \)-repeat-free, that there exist a unique way to concatenate these substrings (excluding overlap) as shown in Lemma 1.

Finally, once \( z \) is reconstructed we may extract \( \{y_i\}_{i \in [\ell]} \) with the decoder of \( E_{m, \ell}^{RF} \).

**REFERENCES**

[1] J. Acharya, H. Das, O. Milenkovic, A. Orlitsky, and S. Pan, “String reconstruction from substring compositions,” *SIAM J. Discrete Math.*, vol. 29, no. 3, pp. 1340–1371, 2015.

[2] D. Bar-Lev, S. Marcovich, E. Yaakobi, and Y. Yehezkeally, “Adversarial torn-paper codes,” in *Proceedings of the 2022 IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, Jun. 2022, pp. 2934–2939.

[3] T. Batu, S. Kannan, S. Khanna, and A. McGregor, “Reconstructing strings from random traces,” in *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA’04)*, New Orleans, LA, USA. Society for Industrial and Applied Mathematics, Jan. 2004, pp. 910–918.

[4] G. Bresler, M. Bresler, and D. Tse, “Optimal assembly for high throughput shotgun sequencing,” *BMC Bioinformatics*, vol. 14, no. 5, p. S18, Jul. 2013.

[5] Z. Chang, J. Chrisnata, M. F. Ezeren, and H. M. Kiah, “Rates of DNA sequence profiles for practical values of read lengths,” *IEEE Trans. on Inform. Theory*, vol. 63, no. 11, pp. 7166–7177, Nov. 2017.

[6] O. Elshio, R. Gabrys, M. Mâdâ, and E. Yaakobi, “Repeat-free codes,” *IEEE Trans. on Inform. Theory*, vol. 67, no. 9, pp. 5749–5764, Sep. 2021.

[7] R. Gabrys and O. Milenkovic, “Unique reconstruction of coded strings from multiset substring spectra,” *IEEE Trans. on Inform. Theory*, vol. 65, no. 12, pp. 7682–7696, Dec. 2019.

[8] R. Heckel, I. Shomorony, K. Ramachandran, and D. N. C. Tse, “Fundamental limits of DNA storage systems,” in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 3130–3134.

[9] H. M. Kiah, G. J. Puleo, and O. Milenkovic, “Codes for DNA sequence profiles,” *IEEE Trans. on Inform. Theory*, vol. 62, no. 6, pp. 3125–3146, Jun. 2016.

[10] A. Lenz, P. H. Siegel, A. Wachtler-Zeh, and E. Yaakobi, “An upper bound on the capacity of the DNA storage channel,” in *Proceedings of the 2019 IEEE Information Theory Workshop (ITW)*, Visby, Sweden, Aug. 2019.

[11] V. I. Levenshtein, “Efficient reconstruction of sequences from their subsequences or supersequences,” *J. Comb. Theory Ser. A*, vol. 93, no. 2, pp. 310–332, Feb. 2001.

[12] M. Levy and E. Yaakobi, “Mutually uncorrelated codes for DNA storage,” *IEEE Trans. on Inform. Theory*, vol. 65, no. 6, pp. 3671–3691, Jun. 2019.

[13] B. Manvel, A. Meyerowitz, A. Schwenk, K. Smith, and P. Stockmeyer, “Reconstruction of sequences,” *Discrete Mathematics*, vol. 94, no. 3, pp. 209–219, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0012365X77900449

[14] S. Marcovich and E. Yaakobi, “Reconstruction of strings from their substrings spectrum,” *IEEE Trans. on Inform. Theory*, vol. 67, no. 7, pp. 4369–4384, Jul. 2021.

[15] B. H. Marcus, R. M. Roth, and P. H. Siegel, “An introduction to coding for constrained systems,” Oct. 2001, unpublished Lecture Notes. [Online]. Available: www.math.ubc.ca/~marcus/Handbook

[16] A. S. Motahari, G. Bresler, and D. N. C. Tse, “Information theory of DNA shotgun sequencing,” *IEEE Trans. on Inform. Theory*, vol. 59, no. 10, pp. 6273–6289, Oct. 2013.

[17] S. Nassirpour, I. Shomorony, and A. Vahid, “Reassembly codes for constrained systems,” Oct. 2001, unpublished Lecture Notes. [Online]. Available: www.math.ubc.ca/~marcus/Handbook

[18] S. Ravi, A. Vahid, and I. Shomorony, “Capacity of the torn paper channel with lost pieces,” in *Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT)*, Melbourne, Victoria, Australia, Jul. 2021, pp. 1937–1942.

[19] ——, “Coded shotgun sequencing,” *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 1, pp. 147–159, Mar. 2022.

[20] I. Shomorony and A. Vahid, “Torn-paper coding,” *IEEE Trans. on Inform. Theory*, vol. 67, no. 12, pp. 7904–7913, Dec. 2021.

[21] E. Ukkonen, “Approximate string-matching with g-grams and maximal matches,” *Theoretical Computer Science*, vol. 92, no. 1, pp. 191–211, 1992.

[22] N. Weinberger and N. Merhav, “The DNA storage channel: Capacity and error probability,” *arXiv preprint arXiv:2109.12549*, Sep. 2021. [Online]. Available: https://arxiv.org/abs/2109.12549

[23] Y. Yehezkeally, S. Marcovich, and E. Yaakobi, “Multi-strand reconstruction from substrings,” in *Proceedings of the 2022 IEEE Information Theory Workshop (ITW)*, Kanazawa, Japan, Oct. 2021.