

Trellis BMA: Coded Trace Reconstruction on IDS Channels for DNA Storage

Sundara Rajan Srinivasavaradhan*, Sivakanth Gopi[†], Henry D. Pfister^{‡†}, and Sergey Yekhanin[†]

*University of California, Los Angeles, Electrical and Computer Engineering, sundar@ucla.edu

[†]Microsoft Research, Redmond, WA, {sigopi,yekhanin}@microsoft.com

[‡]Duke University, Electrical and Computer Engineering, Durham, NC, henry.pfister@duke.edu

Abstract—Sequencing a DNA strand, as part of the read process in DNA storage, produces multiple noisy copies which can be combined to produce better estimates of the original strand; this is called *trace reconstruction*. One can reduce the error rate further by introducing redundancy in write sequence and this is called *coded trace reconstruction*. In this paper, we model the DNA storage channel as an insertion-deletion-substitution (IDS) channel and design both encoding schemes and *low-complexity* decoding algorithms for coded trace reconstruction.

We introduce Trellis BMA, a new reconstruction algorithm whose complexity is linear in the number of traces, and compare its performance to previous algorithms. Our results show that it reduces the error rate on both simulated and experimental data. The performance comparisons in this paper are based on the Clustered Nanopore Reads Dataset publicly released with this paper. Our hope is that this dataset will enable research progress by allowing objective comparisons between candidate algorithms.

I. INTRODUCTION

DNA storage is an exciting area because of its potential to provide both high information density and long-term stability [1]. To achieve a good trade-off between efficiency and reliability, DNA storage systems use error-correcting codes [2]–[9]. This paper considers the design and decoding of error-correction codes for the DNA storage channel (see Figure 1).

In this paper, the DNA storage channel is modeled as an insertion-deletion-substitution (IDS) channel and we focus on the case where a single encoded message is transmitted and multiple independent *traces* are observed [2]–[4], [7]. Sequence reconstruction methods for this problem date back to the 1980s [10]. This is closely related to the trace reconstruction problem in CS literature which asks how many traces (from a deletion channel) are needed to perfectly reconstruct the input message sequence, in the average or worst case. Many algorithms exist for trace reconstruction [11]–[17], a few of which (such as *Bitwise Majority Alignment* (BMA) [12]) can be modified for the IDS channel and have been used in DNA data storage systems [18], [19].

In practical systems, outer codes are used to code across multiple DNA strands in order to recover missing sequences and correct substitutions of individual symbols. Thus, we focus primarily on approximate reconstruction, as opposed to exact reconstruction. For IDS-like channels, one can compute exact posterior marginals by combining ideas from multiple-sequence alignment [10] and the BCJR algorithm [20] (e.g., see [21]–[23]). Using these posterior marginals, it is easy to compute estimates that minimize additive distortion measures.

If the outer code uses hard-decision decoding, then a reasonable goal is to construct a practical estimator that, given a small number of traces, minimizes the expected Hamming distance to the input message. Strands may also use an inner code that is designed to provide additional protection [24], [25]. The inner code constraints can also be included in channel trellis [26] so that trellis-based methods can still be used for inference. In particular, for convolutional codes, it is possible to build a multidimensional trellis and perform symbolwise maximum-a-posteriori (MAP) reconstruction, as observed in [23]. But, the complexity grows exponentially with the number of traces making exact inference infeasible.

A. Contributions¹

- A low-complexity heuristic dubbed Trellis BMA is proposed that allows multiple single-trace trellis decoders to interact and estimate the input message on-the-fly. This is different from the approaches in [23], [27], [28] because each single-trace trellis decoder is influenced by the other decoders but has some connection to the factor graph method in [22]. Our idea marries BCJR inference for IDS channels with the consensus approach of BMA, hence the name *Trellis BMA*.
- A dataset of short strand DNA reads is generated that can be used to compare algorithms with actual DNA reads. This dataset will be released publicly to serve as a benchmark for coded trace reconstruction algorithms.
- A new construction for the multi-trace IDS trellis is provided where the number of edges grows at a lower exponential rate (with the number of traces) than previous approaches. Using BCJR inference to compute the symbolwise posterior probabilities for multiple traces is exponentially faster with this formulation.

II. BACKGROUND

A. DNA sequencing channel

The observed noise in DNA storage is a complicated combination of synthesis errors, amplification errors, and sequencing noise [29]. Even if we ignore the first two elements, the exact error profile of the noisy observations is dependent on the DNA sequencing technology used. However, exactly modeling this error profile is tedious and often impractical. Moreover, DNA sequencing technologies are evolving at a rapid pace and

¹The majority of this work was completed while the first author was an intern at Microsoft Research and was presented there on Sept. 11th, 2020.

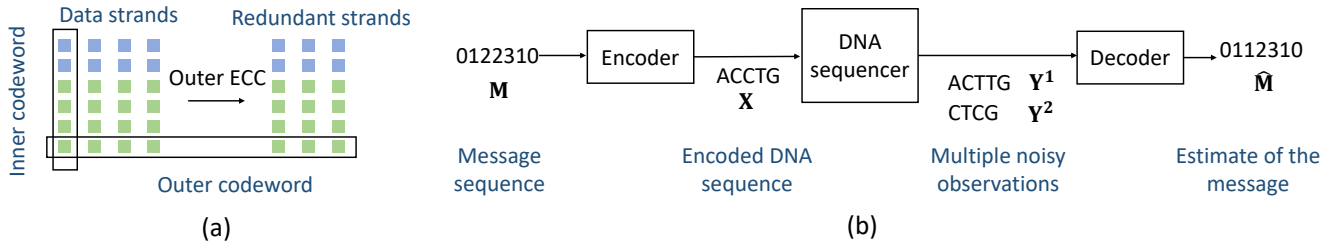


Fig. 1: (a) The interplay between inner and outer code in a DNA storage system. Data strands are first encoded using an outer code (to correct for missing sequences) and then using an inner code which corrects IDS errors. (b) The inner code architecture for DNA storage. Encoded DNA strands are read or “sequenced” using a sequencing technology, such as Illumina/Nanopore sequencers, and this outputs many noisy copies of the DNA sequence, from which the message vector in the data strand is recovered.

focusing on a particular error profile does not provide a future-proof approach to the problem. Instead, one typically considers a simplistic approximation and models the sequencing channel as an IDS channel (defined in the next subsection). Our ideas also extend naturally to more complex approximations for the channel model. For instance, insertions and deletions often occur in “bursts” and such events can be captured by a first-order Markov model; our decoder can easily be modified to accommodate for such variations.

Due to the difficulty of synthesizing and sequencing long DNA strands, DNA storage systems typically encode a single file into many different short strands. The Poisson nature of sampling short strands from the pool means that many of these strands will not be sequenced. Thus, an outer code is required and sequence numbers must be included for disambiguation [30]. This detail is sometimes neglected in simulation-based experiments (e.g., it seems a single long strand is used in [23]).

B. Insertion Deletion Substitution Channel

The insertion deletion substitution (IDS) channel is defined by its input/output alphabet Σ and four non-negative parameters p_{ins} , p_{del} , p_{sub} , p_{cor} with $p_{\text{ins}} + p_{\text{del}} + p_{\text{sub}} + p_{\text{cor}} = 1$. Given an N -length input sequence $\mathbf{X} = X_1 X_2 \dots X_N \in \Sigma^N$, the IDS channel sequentially takes in one input symbol at a time and constructs a variable length output $\mathbf{Y} = Y_1 Y_2 \dots \in \Sigma^*$ sequentially². Let the input pointer be i and the output pointer be j . Starting from $i = j = 1$, sample from the following events i equals $N + 1$:

- **Insertion** (probability p_{ins}): choose Y_j uniformly at random from Σ , increase j by 1, and leave i unchanged.
- **Deletion** (probability p_{del}): increase i by 1 and leave j unchanged.
- **Substitution** (probability p_{sub}): choose Y_j uniformly at random from $\Sigma \setminus \{X_i\}$. Increase both i and j by 1.
- **Correct**: (probability p_{cor}): Set $Y_j = X_i$. Increase i and j by 1.

C. Trace reconstruction with and without coding

As discussed in the introduction, the trace reconstruction (TR) problem has been formalized in the CS literature as the question, “How many traces are required to exactly reconstruct \mathbf{X} ?” [12]–[17]. However, exact reconstruction is typically

²Here $\Sigma^* \triangleq \cup_{m=0}^{\infty} \Sigma^m$ denotes the set of finite strings over Σ .

impossible from only a few traces [31]. Thus, we use the term *TR algorithm* for any algorithm that uses multiple independent traces $\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^K \in \Sigma^*$ of \mathbf{X} to construct an estimate $\hat{\mathbf{X}}(\mathbf{Y}^{1:K})$, where $\mathbf{Y}^{1:K}$ is shorthand for $(\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^K)$.

A more general formulation is to consider a code that maps a message sequence $\mathbf{M} = M_1 M_2 \dots M_L \in \mathcal{M}^L$ to a codeword $\mathbf{X} = X_1 X_2 \dots X_N \in \Sigma^N$. The goal of coded TR is to compute an estimate $\hat{\mathbf{M}}(\mathbf{Y}^{1:K})$ of the message sequence \mathbf{M} from the K multiple independent traces $\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^K$ of \mathbf{X} . This setup naturally fits the DNA storage architecture in Fig. 1.

D. Error-correcting codes

For the inner code, this work considers marker repeat (MR) codes with the addition of a random scrambling vector to prevent shift invariance. Marker codes are synchronization codes where a short *marker sequence* is inserted periodically [32]. MR codes are a new variation where, periodically, a single input symbol is transmitted multiple times. For example, a length- N MR code with r length-2 repeats satisfies $x_{n+1} = x_n$ when $n = \lfloor iN/(r+1) \rfloor$ for $i = 1, \dots, r$. Results are given for MR codes with $N = 110$ and $r = 6, 10$. Rate-1/2 quaternary convolutional codes with memory 3-5 and puncturing were tested and found to be inferior to MR codes above rate 3/4.

While this work focuses on the efficient decoding of the inner code when multiple traces are received, our analysis also assumes there will be an outer code. In particular, we target schemes where the inner codes are decoded first followed by the outer code. In contrast to [23], we do not consider iteration between the inner and outer decoder nor do we estimate the error rate after decoding of the outer code.

E. Performance metrics and information rates

The choice of performance metric for BCJR inference depends crucially on how the outputs will be used. Different decoding methods for the outer code lead to different achievable rates. Any rate loss due to inner MR codes is included in these computations whereas rate loss due to sequence numbers, which are typically required by outer codes, is neglected.

For general trace reconstruction (or detection before hard-input decoding of an outer code defined over Σ), one typically chooses $\hat{\mathbf{X}}$ to minimize the expected Hamming distance

$$\mathbb{E} \left[d(\hat{\mathbf{X}}, \mathbf{X}) \right] = \sum_{n=1}^N \Pr(\hat{X}_n \neq X_n), \quad (1)$$

and the optimal \widehat{X} is given by the symbolwise MAP estimate. Choosing $\widehat{\mathbf{X}}$ to minimize the edit distance has also been considered in [28], [33], [34]. For hard-decision decoding of an outer code defined by M symbols, the expected Hamming error rate is likewise minimized by choosing $\widehat{\mathbf{M}}$ to be the symbolwise MAP estimate of \mathbf{M} .

For soft-decision decoding, the outer decoder uses the posterior marginals, $Q_l(m) \triangleq \Pr(M_l = m | \mathbf{Y}^{1:K})$, whose uncertainty is quantified by the average symbolwise entropy

$$H = \frac{1}{L} \sum_{l=1}^L \mathbb{E} \left[\frac{1}{\log Q_l(M_l)} \right] \leq \frac{1}{L} \sum_{l=1}^L \mathbb{E} \left[\log \frac{1}{\widehat{Q}_l(M_l)} \right]. \quad (2)$$

Here, \widehat{Q}_l is any approximate posterior marginal (e.g., due to channel mismatch or suboptimal processing) satisfying $\sum_{m \in \Sigma} \widehat{Q}_l(m) = 1$ for all l . For i.i.d. equiprobable inputs into a rate- R inner code, the quantity $(2 - H)R$ (bits/base) is an overall achievable information rate (AIR) for separate detection and decoding, called the BCJR-once rate [35]–[37]. If a random outer code is used with joint decoding, then the AIR is the mutual information rate $\frac{1}{N} I(\mathbf{M}; \mathbf{Y}^{1:K}) = \frac{R}{L} I(\mathbf{M}; \mathbf{Y}^{1:K})$ which can be estimated using the BCJR algorithm [38]–[40].

In actual DNA storage systems, the number of traces K will be a random variable that is different for each observed cluster. In that case, a particular AIR for random K is given by averaging that AIR over the distribution of K .

III. DATASET

The performance comparisons in this paper are based on a new dataset of 269,709 traces of 10,000 uniform random DNA sequences of length 110 that is now publicly available at:

<https://github.com/microsoft/clustered-nanopore-reads-dataset>

Our hope is that this dataset will enable further research progress by allowing objective comparisons between the algorithms. DNA sequences were synthesized by Twist Bioscience and amplified using polymerase chain reaction. The amplified products were ligated to Oxford Nanopore Technologies (ONT) sequencing adapters by following the manufacturer’s protocol (LQK-LSK 109 kit). Finally, ligated samples were sequenced using ONT MinION. Clusters of noisy reads have been recovered using the algorithm from [41]. The insertion, deletion, and substitution rates for this dataset are roughly $p_{\text{ins}} = 0.017$, $p_{\text{del}} = 0.02$, and $p_{\text{sub}} = 0.022$.

Using the dataset for coded TR: The dataset is a collection of (\mathbf{x}, \mathbf{y}) pairs allowing one to estimate the expected performance of TR algorithms for uniform random DNA sequences. For coded TR, the problem is that one cannot estimate an expectation over codewords because the randomly generated DNA sequences are unlikely to be codewords in the code.

One can estimate the expected performance for a coded system with random scrambling. Assume Σ has an abelian group structure and let the code $\mathcal{C} \subseteq \Sigma^N$ be a subset with encoder $\mathcal{E}: \mathcal{M}^L \rightarrow \mathcal{C}$. Consider estimating a performance measure $\phi = \mathbb{E}[\Phi(\mathbf{Y}^{1:K}; \mathbf{M}, \mathbf{Z})]$ for the scrambled encoder defined by $\mathbf{X} = \mathcal{E}(\mathbf{M}) + \mathbf{Z}$, where $\mathbf{Z} \in \Sigma^N$ is a uniform random scrambling sequence. Since this induces a uniform

distribution on \mathbf{X} , the dataset can be used to estimate ϕ . For an \mathbf{x} in the dataset, let $U(\mathbf{x})$ denote the set of \mathbf{y} traces generated by \mathbf{x} . Samples can be drawn as follows:

- Let \mathbf{x} be the result of drawing a uniform random DNA sequence from the dataset, \mathbf{m} be the result of choosing a uniform random message, and then compute $\mathbf{z} = \mathbf{x} - \mathcal{E}(\mathbf{m})$.
- Compute the sample value $\Phi(\mathbf{y}^{1:K}; \mathbf{m}, \mathbf{z})$ for K traces sampled randomly from $U(\mathbf{x})$ without replacement.

To summarize, for an encoder \mathcal{E} , we estimate the average of Φ over \mathbf{Z} . Hence, there is a \mathbf{z} that performs this well or better.

IV. ALGORITHMS FOR TR AND CODED TR

A. Multi-trace trellis via hidden Markov model

Our discussion of algorithms begins with a brief description of a hidden Markov model (HMM) associated with the problem. The state diagram of this HMM implies a natural multi-trace IDS trellis that is different from previous methods [21]–[23]. This trellis has significantly fewer edges and this reduces the complexity of BCJR inference. However, the resulting trellis and BCJR definitions are a bit different from those typically used in coding theory. Due to space constraints, some details are neglected and we refer interested readers to an extended version of the paper [42].

In essence, our construction of the trellis describing the joint distribution of $(\mathbf{M}, \mathbf{X}, \mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^K)$ avoids local exponential blow-up in the number of edges by

- modeling insertion events as vertical edges, thereby sequentially accounting for insertions.
- modeling events in each trace sequentially.

Consider a message sequence $\mathbf{M} = M_1 M_2 \dots M_L$, where $M_i \in \mathcal{M}$, which is mapped onto a codeword $\mathbf{X} = X_1 X_2 \dots X_N$, where $X_i \in \mathcal{X}$, using a (possibly time-varying) deterministic FSM encoder. Such an encoder takes as input a message symbol M_i , transitions to state Q_i and emits u codeword symbols $X_{u(i-1)+1} X_{u(i-1)+2} \dots X_{ui}$. The transition and codeword symbols emitted only depend on M_i and its state Q_{i-1} before accepting input symbol M_i . For simplicity, assume that the number of emitted symbols u is fixed for all i ($N = Lu$); our trellis can also account for cases where u varies with i .

Suppose we observe K independent traces $\mathbf{y}^1, \dots, \mathbf{y}^K$ generated from \mathbf{X} . Let $\mathbf{y}^k = y_1^k y_2^k \dots y_{R_k}^k$, therefore the length of the k -th trace is R_k . The trellis is a directed acyclic graph (DAG) with weighted edges where the vertices are ordered by “stages” – edges connect two vertices in the same stage or connect a vertex at stage t to a vertex at stage $t + 1$. At stage t , vertex v_t is defined by $(q_t, p_t^1, p_t^2, \dots, p_t^K, m_t, x_t)$ where

- $q_t \in \mathcal{Q}$, is the state of the encoder at stage t .
- $p_t^k \in \mathcal{P}_t^k$ where $\mathcal{P}_t^k \subseteq \{1, 2, \dots, R_k\}$ is the output pointer for the k -th trace at stage t .
- $m_t \in \mathcal{M}_t$ with $\mathcal{M}_t = \mathcal{M} \cup \{\star\}$, is the on-deck message symbol.
- $x_t \in \mathcal{X}_t$ with $\mathcal{X}_t = \mathcal{X} \cup \{\star\}$, is the on-deck codeword symbol.

Therefore, $v_t \in \mathcal{Q} \times \mathcal{P}_t^1 \times \mathcal{P}_t^2 \dots \times \mathcal{P}_t^K \times \mathcal{M}_t \times \mathcal{X}_t$, where \times denotes the Cartesian product. For clarity, we construct the trellis stage-by-stage, describing the stages corresponding to the first message symbol.

Modeling the input. An edge connects vertex $v_1 = (q_{init}, 1, 1, \dots, 1, \star, \star)$ at stage 1 to $v_2 = (q, 1, 1, \dots, 1, m, x)$ at stage 2, where q_{init} is the initial state of the encoder and encoder makes the transition $q_{init} \rightarrow q$ when presented with input m , emitting first codeword symbol x . The edge weight is equal to $\Pr(M_i = m)$ to model the input distribution.

Modeling IDS events. An edge connects a vertex $v_2 = (q, p^1, p^2, \dots, p^K, m, x)$ to $v_3 = (q, p^1, p^2, \dots, p^K, m, x)$ with a weight equal to p_{del} modeling a deletion event in the first trace. An edge connects a vertex $v_2 = (q, p^1, p^2, \dots, p^K, m, x)$ to $v_3 = (q, p^1 + 1, p^2, \dots, p^K, m, x)$ with a weight equal to p_{cor} if $y_{p^1}^1 = x$ and $\frac{p_{sub}}{|\mathcal{X}|-1}$ otherwise. This models a substitution/correct event in the first trace. An edge connects a vertex $v_2 = (q, p^1, p^2, \dots, p^K, m, x)$ to $v'_2 = (q, p^1 + 1, p^2, \dots, p^K, m, x)$ in the same stage with a weight equal to $\frac{p_{ins}}{|\mathcal{X}|}$, modeling an insertion event in the first trace. Notice how only the output pointer to the first trace changes in all cases. We construct K such stages for K traces.

Updating on-deck codeword symbol. We have only considered the events corresponding to the first codeword symbol so far. Next, we update the output buffer to replace the first codeword symbol x by the second x' , followed by K stages of IDS event modeling for the second codeword symbol.

Transitioning to the next input. The above two steps of modeling the IDS events and updating the output buffer are repeated until all codeword symbols for a given input symbol are processed. Then, the input and output buffer are cleared and the next message symbol is accepted.

The above steps comprise one input cycle. These steps are repeated until all message symbols are exhausted. Each path connecting $(q_{init}, 1, 1, \dots, 1, \star, \star)$ at the first stage to $(q_{end}, R_1, R_2, \dots, R_K, \star, \star)$ at the final stage correspond to a message sequence and a sequence of events that resulted in the observed traces $\mathbf{Y}^1 = \mathbf{y}^1, \dots, \mathbf{Y}^K = \mathbf{y}^K$. The weight of this path is the joint probability of observing the message, the sequence of events and the traces. For this setup, one can use BCJR inference to compute the posterior probability that the true system passed through a given vertex at a particular stage. Then, one can compute $\Pr(M_l = m | \mathbf{Y}^1 = \mathbf{y}^1, \dots, \mathbf{Y}^K = \mathbf{y}^K)$ by summing the posterior probabilities of all vertices associated with message symbol m in the input cycle of stage l .

Time Complexity. Assuming the length of the traces $R_k = O(N) \forall k$, and \mathcal{Q} is the state-space of the encoder FSM, the total number of edges in the trellis is $O(N^{K+1}K|\mathcal{Q}|)$, which is the time complexity to exactly compute the APPs. In practice, it is reasonable to assume that the output pointer does not drift too far from the input pointer for each IDS channel, i.e., at a given stage one assumes that $|\mathcal{P}_t| = \Delta < N$ [21], [22], [28]. Using this assumption, the complexity to compute APPs is roughly $O(NK|\mathcal{Q}|\Delta^K)$. Note that, for large K , this is significantly smaller compared to the complexity of computing APPs in [23] (which is at least $\Omega(NK|\mathcal{Q}|\Delta^K u^K)$).

B. Trellis BMA

Given the exponential growth of the multi-trace IDS trellis with the number of traces, we next describe a heuristic that

combines IDS trellises for individual traces to sequentially construct an estimate $\widehat{\mathbf{M}} = \widehat{M}_1 \widehat{M}_2 \dots \widehat{M}_I$ of the message.

a) Initialization: Following the steps outlined in the previous subsection, we first construct K independent trellises: one for each trace \mathbf{y}^k with $k \in [K]$. Then, we run BCJR inference on each of the K trellises with the corresponding traces as observations and compute $F^k(v)$ and $B^k(v)$, the forward and backward values of each vertex v in the trellis corresponding to trace k , for all k – these values will be updated using a consensus across traces.

b) Decoding: We now compute $\widehat{\mathbf{M}}$ by iterating through the following two steps. Working inductively, we assume that we have already computed $\widehat{M}_1, \widehat{M}_2, \dots, \widehat{M}_{l-1}$ and we would like to compute \widehat{M}_l .

• **Combining beliefs from each trellis.** First, we use the current values of $F^k(v)$ and $B^k(v)$ to compute a “belief” about symbol M_l for each trellis, denoted by $V^k(M_l = m)$. Recall that each M_l is part of the trellis state in some stages (e.g., stages corresponding to input cycle l). Then, pick one of these (e.g., the last stage), call it stage t , and define

$$V^k(M_l = m) \triangleq \sum_{v \in \mathcal{V}_t(M_l = m)} F^k(v) (B^k(v))^{\beta_b}, \quad (3)$$

where the sum is over stage- t vertices with on-deck message symbol $M_l = m$ and $\beta_b \geq 0$ reweights the backward values.

The channel outputs are conditionally independent given \mathbf{M} , so we have $\Pr(\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^K | \mathbf{M}) = \prod_k \Pr(\mathbf{Y}^k | \mathbf{M})$. The RHS likelihoods can theoretically be combined to compute the true posterior. However, BCJR inference outputs the marginals and multiplying them only gives the approximation

$$V(M_l = m) \triangleq \prod_{k=1}^K V^k(M_l = m).$$

• **Updating the forward values.** For trellis k , the idea is to combine information from the other trellises to help maintain the correct synchronization on this trellis. To do this, the forward BCJR values in stage t are updated using the rule

$$F^k(v) \leftarrow \gamma^k(m(v)) F^k(v),$$

where $m(v)$ is value of M_l associated with vertex v and $\gamma^k(m)$ acts as a “new prior” for M_l in trellis k due to the other trellises. We also note that the sum $\sum_m \gamma^k(m)$ does not affect the answer and, thus, $\gamma^k(\cdot)$ is like an unnormalized probability.

To define $\gamma^k(\cdot)$, we use the parametrized expression

$$\gamma^k(m) \triangleq (V^k(M_l = m))^{\beta_i} \prod_{j \neq k} (V^j(M_l = m))^{\beta_e}.$$

This is motivated by the idea of extrinsic information processing [43], [44]. The parameter $\beta_e \geq 0$ controls the dependence induced between the separate strand detectors and $\beta_i \geq 0$ controls the intrinsic bias in each strand. While $\beta_e = 1$ is a natural choice, smaller values of β_e reduce the dependence between strands and larger values push the $\gamma^k(\cdot)$ distribution towards a hard decision. Similarly, $\beta_i = 0$ is a natural choice but larger values can sometimes improve performance.

For the posterior estimate of M_l given $\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^K$, we define $\widehat{Q}_l(m) \triangleq c_l V(M_l = m)^{\beta_o}$ for some $\beta_o > 0$ and

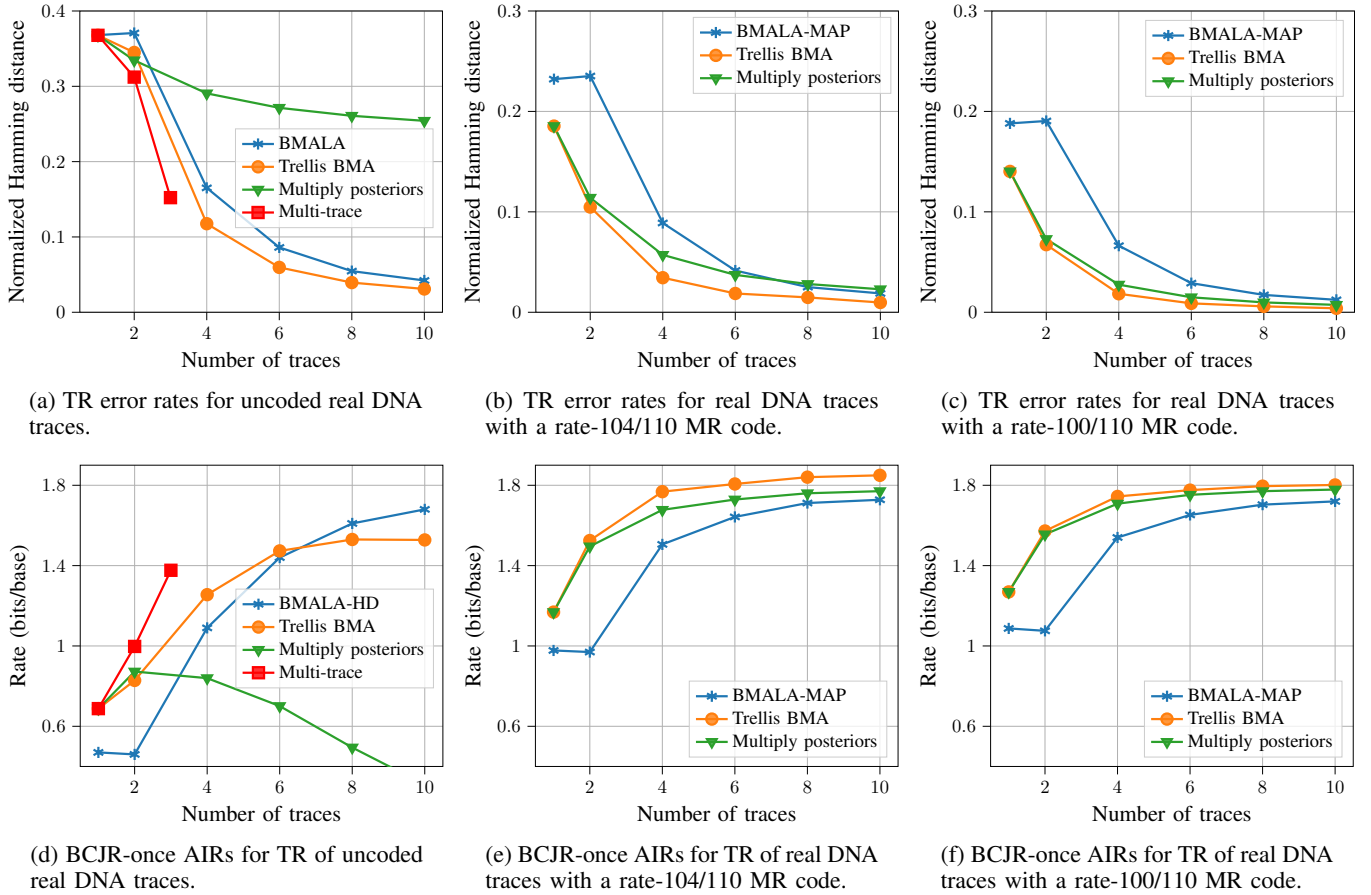


Fig. 2: Experimental results. Note that Subfigures 2e and 2f include the rate loss of their MR codes.

choose c_l so the sum over m equals 1. To lower bound the AIR, we apply the RHS of (2) to \hat{Q}_l . Choosing $\beta_o < 1$ may mitigate overconfidence and increase this lower bound. Using the updated forward values at input cycle l , we the forward pass continues to input cycle $l+1$ and computes $V^k(M_l = m)$. Then, this is used to update the forward values for the vertices of input cycle $l+1$. This process repeats until the midpoint.

c) Estimating each half: Using this updating approach, we sequentially compute the estimates $\hat{M}_1 \hat{M}_2 \dots \hat{M}_{L/2}$. Analogously, we start from the end of the trellis and update the backward values to compute an estimate $\hat{M}_{L/2+1} \hat{M}_2 \dots \hat{M}_L$ which proceeds in the reverse order. For the reverse estimate, (3) should use the first stage with M_l in the state.

d) Time Complexity: The time complexity is K times the complexity of computing APPs using the multi-trace trellis with one trace, which is equal to $O(KN|Q|\Delta)$.

V. EXPERIMENTAL RESULTS

In Fig. 2, we provide experimental results, with and without coding, for the algorithm introduced in this paper. We also compare to previous approaches such as “separate decoding” using “multiply posteriors” from [23], BMALA (BMA with Look Ahead) from [45, pp. 6–7] [19], and to BCJR on the multi-trace IDS trellis from Section IV-A (see also [23]). Note that BMALA is a TR algorithm and does not give soft output. Hence, the BMALA-HD curve in Fig. 2(d) maps the hard-

decision symbol error rate into an AIR. We note that BMALA-HD beats Trellis BMA for more than 6 traces even though Trellis BMA has a lower error rate. This is because the soft outputs of Trellis BMA are not ideally calibrated. In future work, we will investigate learning-based methods to see if they can generate better calibrated output probabilities.

For coded TR, we use BMALA to give a hard estimate of the DNA sequence and treat this estimate as an observed trace for IDS trellis decoding of the message symbols; we call this BMALA-MAP. We also report the numbers for the multi-trace trellis only for TR with 3 or fewer traces; other experiments with the multi-trace trellis are computationally infeasible.

The 10000 clusters of DNA sequences (and corresponding traces) in the dataset are divided into training (clusters 1-2000), validation (clusters 2001-2500) and test sets (clusters 2501-10000). Training is used to learn the IDS channel (p_{ins} , p_{del} , p_{sub}), validation is used to tune the hyperparameters (β_e , β_o , etc.) for Trellis BMA, and the test set is used for the reported results. We remark that multiply posteriors is an instance of Trellis BMA when $\beta_b = \beta_o = 1$ and $\beta_e = \beta_i = 0$.

ACKNOWLEDGMENT

We thank Karin Strauss, Yuan-Jyue Chen, and the Molecular Information Systems Laboratory (MISL) for providing the DNA dataset released with this paper and useful discussions on this topic.

REFERENCES

- [1] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.
- [2] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.
- [3] S. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, 2015.
- [4] S. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific reports*, vol. 5, p. 14138, 2015.
- [5] R. Heckel, I. Shomorony, K. Ramchandran, and N. David, "Fundamental limits of DNA storage systems," in *Proc. IEEE Int. Symp. Inform. Theory*, 2017, pp. 3130–3134.
- [6] S. M. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific reports*, vol. 7, no. 1, pp. 1–6, 2017.
- [7] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen *et al.*, "Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.
- [8] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakohi, "Achieving the capacity of the DNA storage channel," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2020, pp. 8846–8850.
- [9] P. L. Antkowiak, J. Lietard, M. Z. Darestani, M. M. Somoza, W. J. Stark, R. Heckel, and R. N. Grass, "Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction," *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020.
- [10] H. Carrillo and D. Lipman, "The multiple sequence alignment problem in biology," *SIAM J. Appl. Math.*, vol. 48, no. 5, pp. 1073–1082, 1988.
- [11] V. I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 2–22, 2001.
- [12] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, 2004, pp. 910–918.
- [13] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results," in *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, 2008, pp. 389–398.
- [14] A. De, R. O'Donnell, and R. Servedio, "Optimal mean-based algorithms for trace reconstruction," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2017, pp. 1047–1056.
- [15] F. Nazarov and Y. Peres, "Trace reconstruction with $\exp(o(n^{1/3}))$ samples," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2017, pp. 1042–1046.
- [16] N. Holden, R. Pemantle, and Y. Peres, "Subpolynomial trace reconstruction for random strings and arbitrary deletion probability," in *Proceedings of the Conference On Learning Theory (COLT)*, 2018, pp. 1799–1840.
- [17] Z. Chase, "New upper bounds for trace reconstruction," *ArXiv preprint : 2009.03296*, 2020.
- [18] P. Gopalan, S. Yekhanin, S. Dumas Ang, N. Jovic, M. Racz, K. Strauss, and L. Ceze, "Trace reconstruction from noisy polynucleotide sequencer reads," 2018, US Patent application : US 2018 / 0211001 A1.
- [19] M. Racz and S. Yekhanin, "Trace reconstruction from reads with indeterminate errors," 2020, US Patent application : US 2020/0057838 A1.
- [20] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. 20, no. 2, pp. 284–287, March 1974.
- [21] M. C. Davey and D. J. C. MacKay, "Reliable communication over channels with insertions, deletions and substitutions," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 687–698, Feb. 2001.
- [22] R. Sakogawa and H. Kaneko, "Symbolwise MAP estimation for multiple-trace insertion/deletion/substitution channels," in *Proc. IEEE Int. Symp. Inform. Theory*. IEEE, 2020, pp. 781–785.
- [23] A. Lenz, I. Maarouf, L. Welter, A. Wachter-Zeh, E. Rosnes *et al.*, "Concatenated codes for recovery from multiple reads of DNA sequences," *arXiv preprint arXiv:2010.15461*, 2020.
- [24] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6084–6103, 2020.
- [25] J. Brakensiek, R. L. Li, and B. Sprang, "Coded trace reconstruction in a constant number of traces," in *Proceedings of the 61st IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2020, pp. 482–493.
- [26] S. Chandak, J. Neu, K. Tatwawadi, J. Mardia, B. Lau, M. Kubit, R. Hulett, P. Griffin, M. Wootters, T. Weissman, and H. Ji, "Overcoming high nanopore basecaller error rates for DNA storage via basecaller-decoder integration and convolutional codes," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8822–8826.
- [27] S. R. Srinivasavaradhan, M. Du, S. Diggavi, and C. Fragouli, "Symbol-wise map for multiple deletion channels," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 181–185.
- [28] —, "Algorithms for reconstruction over single and multiple deletion channels," *IEEE Transactions on Information Theory*, 2020.
- [29] W. Mao, S. N. Diggavi, and S. Kannan, "Models and information-theoretic bounds for nanopore sequencing," *IEEE Trans. Inform. Theory*, vol. 64, no. 4, pp. 3216–3236, 2018.
- [30] I. Shomorony and R. Heckel, "DNA-based storage: Models and fundamental limits," *arXiv preprint arXiv:2001.06311*, 2020.
- [31] Z. Chase, "New lower bounds for trace reconstruction," 2019, arXiv preprint arXiv:1905.03031.
- [32] E. A. Ratzner, "Marker codes for channels with insertions and deletions," in *Annales des télécommunications*, vol. 60, no. 1, 2005, pp. 29–44.
- [33] O. Sabary, A. Yucovich, G. Shapira, and Y. Eitan, "Reconstruction algorithms for DNA-storage systems," *BioRxiv preprint : https://doi.org/10.1101/2020.09.16.300186*, 2020.
- [34] S. Davies, M. Rácz, C. Rashtchian, and S. Benjamin, "Approximate trace reconstruction," *ArXiv preprint : 2012.06713*, 2020.
- [35] A. Kavčić, X. Ma, and M. Mitzenmacher, "Binary intersymbol interference channels: Gallager codes, density evolution and code performance bounds," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1636–1652, July 2003.
- [36] R. R. Müller and W. H. Gerstacker, "On the capacity loss due to separation of detection and decoding," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1769–1778, Aug. 2004.
- [37] J. B. Soriaga, H. D. Pfister, and P. H. Siegel, "Determining and approaching achievable rates of binary intersymbol interference channels using multistage decoding," *IEEE Trans. Inform. Theory*, vol. 53, no. 4, pp. 1416–1429, April 2007.
- [38] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite state ISI channels," in *Proc. IEEE Global Telecom. Conf.*, San Antonio, TX, USA, Nov. 2001, pp. 2992–2996.
- [39] D. Arnold, H. A. Loeliger, P. O. Vontobel, A. Kavčić, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3498–3508, Aug. 2006.
- [40] A. Kavcic and R. Motwani, "Insertion/deletion channels: Reduced-state lower bounds on channel capacities," in *Proc. IEEE Int. Symp. Inform. Theory*, 2004, p. 229.
- [41] C. Rashtchian, K. Makarychev, M. Rácz, S. Ang, D. Jevdjic, S. Yekhanin, L. Ceze, and K. Strauss, "Clustering billions of reads for DNA data storage," in *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 3360–3371.
- [42] S. R. Srinivasavaradhan, S. Gopi, H. D. Pfister, and S. Yekhanin, "Trellis BMA: Coded trace reconstruction on IDS channels for DNA storage," 2021, to appear on arXiv.
- [43] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in *Proc. IEEE Int. Conf. Commun.*, vol. 2. Geneva, Switzerland: IEEE, May 1993, pp. 1064–1070.
- [44] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 498–519, 2001.
- [45] R. Lopez, Y.-J. Chen, S. Ang, S. Yekhanin, K. Makarychev, M. Racz, G. Seeling, K. Strauss, and L. Ceze, "Dna assembly for nanopore data storage readout," *Nature Communications*, vol. 10, 2019.