BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–9

Fundamental Bounds and Approaches to Sequence Reconstruction from Nanopore Sequencers

Jaroslaw Duda, Wojciech Szpankowski, and Ananth Grama

[1]Department of Computer Science and Center for Science of Information, Purdue University.

ABSTRACT

Motivation: Nanopore sequencers are emerging as promising new platforms for high-throughput sequencing. As with other technologies, sequencer errors pose a major challenge for their effective use. In this paper, we present a novel information theoretic analysis of the impact of insertion-deletion (InDel) errors in nanopore sequencers. In particular, we consider the following problems: (i) for given InDel error characteristics and rate, what is the probability of accurate reconstruction as a function of sequence length; (ii) what is the number of 'typical' sequences within the distortion bound induced by InDel errors; (iii) using repeated extrusion through the nanopore, what is the number of repetitions needed to reduce the distortion bound so that only one typical sequence exists within the distortion bound.

Results: Our results provide a number of important insights: (i) the maximum length of a sequence that can be accurately reconstructed in the presence of InDel errors is relatively small; (ii) the number of typical sequences within the distortion bound is large; and (iii) repeated extrusion is an effective technique for unique reconstruction. In particular, we show that the number of repeats is a slow function (logarithmic) of sequence length – implying that through repeated extrusion, we can sequence large reads using nanopore sequencers. InDel errors are the primary error mode for nanopore sequencers. To this end, the results in this paper can be viewed as (tight) bounds on reconstruction lengths and repetitions for accurate reconstruction.

Contact: ayg@cs.purdue.edu

1 INTRODUCTION

The past few years have seen significant advances in sequencing technologies. Sequencing platforms from Illumina, Roche, PacBio and other vendors are commonly available in laboratories. Accompanying these hardware advances, significant progress has been made in statistical methods, algorithms, and software for tasks ranging from base calling to complete assembly. Among the key distinguishing features of these sequencing platforms are their read lengths and error rates. Short read lengths pose problems for sequencing high-repeat regions. Higher error rates, on the other hand, require oversampling to either correct, or discard erroneous reads without adversely impacting sequencing/ mapping quality. Significant research efforts have studied tradeoffs of read-length, error rates, and sequencing complexity. An excellent survey of these efforts is provided by Quail et al [9].

More recently, nanopores have been proposed as platforms for sequencing. Nanopores are fabricated either using organic channels (pore-forming proteins in a bilayer) or solid-state material (silicon nitride or graphene). An ionic current is passed through this nanopore by establishing an electrostatic potential. When an analyte simultaneously passes through the nanopore, the current flow is disrupted. This disruption of the current flow is monitored, and used to characterize the analyte. This general principle can be used to characterize nucleotides, small molecules, and proteins. Complete solutions based on this technology are available from Oxford Nanopore Technologies [7]. In this platform, a DNA strand is extruded through a protein channel in a membrane. The rate of extrusion must be slower than current measurement (sampling) for characterizing each base (or groups of small number of bases, up to four, in the nanopore at any point of time).

In principle, nanopores have several attractive features – long reads (beyond 100K bases) and minimal sample preparation. However, there are potential challenges that must be overcome – among them, the associated error rate. The extrusion rate of a DNA strand through a protein channel is controlled using an enzyme [8]. This rate is typically modeled as an exponential distribution. When a number of identical bases pass through the nanopore, the observed (non-varying) signal must be parsed to determine the precise number of bases. This results in the dominant error mode for nanopore sequences. Specifically, insertion-deletion errors in such sequencers are reported to be as high as 4% [8, 10].

The high InDel error rate can be handled using repeated reads for de-novo assembly, or through algorithmic techniques using reference genomes. The Oxford Nanosequencer claims a scalable matrix of pores and associated sensors using which repeats can be generated. Alternately, other technologies based on bi-directional extrusion have been proposed. In either case, two fundamental questions arise for de novo assembly: (i) for single reads, what is the bound on read length that can be accurately reconstructed using a nanopore sequencer with known InDel rates; and (ii) what is the number of repeats needed to accurately reconstruct the sequence with high probability (analytically defined).

In this paper, we present a novel information theoretic analysis of the impact of InDel errors in nanopore sequencers. We model the sequencer as a sticky insertion-deletion channel. The DNA sequence is fed into this channel and the output of the channel is used to reconstruct the input sequence. Using this

model, we solve the following problems: (i) for given InDel error characteristics and rate, what is the probability of accurate reconstruction as a function of sequence length; (ii) what is the number of 'typical' sequences within the distortion bound induced by InDels; and (iii) what is the number of repeats needed to reduce the distortion bound so that only one typical sequence exists within the distortion bound (unique reconstruction).

Our results provide a number of important insights: (i) the maximum length of sequence that can be accurately reconstructed in the presence of InDel errors is relatively small; (ii) the number of typical sequences within the distortion bound induced by InDels is large; and (iii) the number of repeats required for unique reconstruction is a slow function (logarithmic) of the sequence length – implying that through repeated extrusion, we can sequence large reads using nanopore sequencers. The bounds we derive are fundamental in nature – i.e., they hold for any resequencing/ processing technique. Furthermore, while InDels (deletion errors primarily) are the dominant error mode in nanopore sequencers, substitution errors may further limit their performance. In this sense, the results in the paper can be viewed as bounds on reconstruction lengths and repetitions.

2 APPROACH

In this section, we present our model and the underlying concepts in information theory that provide the modeling substrates. We define notions of a channel, reconstruction, an insertion-deletion channel, and distortion bound. We then describe how these concepts are mapped to the problem of sequence reconstruction in nanopore sequencers.

Our basic model for a nanopore sequencer is illustrated in Figure 1. A DNA sequence is input to the nanopore sequencer. This sequence is read and suitably processed to produce an output sequence. We view the input sequence as a sequence of blocks. Each block is comprised of a variable number (k) of identical bases. The nanopore sequencer potentially introduces errors into each block by altering the number of repeated bases. If the output block size k' is not the same as the input block size k, an InDel error occurs. Specifically, $k' < k$ corresponds to a deletion error, and $k' > k$ to an insertion error. Please note that this model does not account for substitution errors.

We model the sequencing process (both the sequencer and the associated processing) as a channel. A channel in information theory is a model (traditionally for a storage or communication device, but in our case, used more generally) for information transfer with certain error characteristics. The input sequence of blocks is sent into this channel. The error characteristics of the channel transform a block of k characters into a block of k' characters. This transformation is modeled as a distribution – $k' = G(k, p)$, where G is the distribution and p, the associated set of parameters tuned to the sequencing platform. In typical scenarios, the distribution peaks at k and decays rapidly on either side. The distribution may be asymmetric around k depending on relative frequency of insertion and deletion errors. We refer to such a channel as a sticky channel.

Insertion-Deletion Channels In ideal communication systems, one often assumes that senders and receivers are perfectly synchronized – i.e., each sent bit is read by the receiver. However, in real systems, such perfect synchronization is often not possible. This leads to sent bits missed by the receiver (a deletion error), or read more than once (an insertion error). Such communication systems are traditionally modeled as insertion-deletion channels. Formally, an independent insertion channel is one in which a single bit transmission is accompanied with the insertion of a random bit with a probability p. An independent deletion channel is one in which a transmitted bit can be deleted (omitted from the output stream) with a probability p'. An insertion-deletion channel contains both insertions and deletions [3]. Please note that a number of basic characteristics of insertion-deletion channels, such as their capacity, are as-yet unknown in information theory literature as well.

We consider a variant of the independent insertion-deletion model that is better suited to nanopore sequencers. In particular, we recognize the primary source of error in nanopore sequencers is associated with disambiguating the exact number of identical bases passing through the nanopore. We modify the independent insertion-deletion channel to the sticky insertion-deletion channel described above (Figure 1). Coincidentally, the analysis of this block modification insertion-deletion channel is easier – as we demonstrate in this paper.

Typical Sequences. There are 4^n distinct nucleotide sequences of n bases, each generated with a corresponding (idealized) probability of 4^{-n}. For convenience, we can also view this as probability as 2^{-2n}, with the understanding that if each base is equally likely, we would need two bits for each base. However, from asymptotic equipartition property (AEP), we know that there is a *typical set* such that the probability of generating a sequence belonging to this set approaches 1. In other words, while there may be sequences outside of this set whose individual probability may be high, their number is small enough that the total probability is dominated by the sequences in this typical set. Furthermore, we know that the probability of drawing a sequence of length n from this set is given by $2^{-H(X)n}$, where $H(X)$ is the entropy of the source. Comparing this expression with the sequence probability assuming each base is equally likely (2^{-2n}), we note the unsurprising conclusion that for our idealized case $H(X) = 2$. However, a number of studies have shown that the entropy of living DNA is in fact much lower, as low as 1.7 or below [6]. This suggests that the number of typical sequences is in fact much less than the number of total sequences. We use this notion of typical sequences and the *typical set* in our derivation of the performance bounds of a nanopore channel.

Distortion Bound and Unique Reconstruction. Viewing a DNA sequence passing through a nanopore as a point (in some very high dimensional space), passing it through a sequencer introduces an error. This error can be viewed as a hypersphere around the original sequence. Each point in the hypersphere corresponds to a possible input sequence with a probability that can be analytically quantified. We refer to this hypersphere as a distortion ball. Ideally, we want the radius of this distortion ball to be as small as possible – containing only a single point.

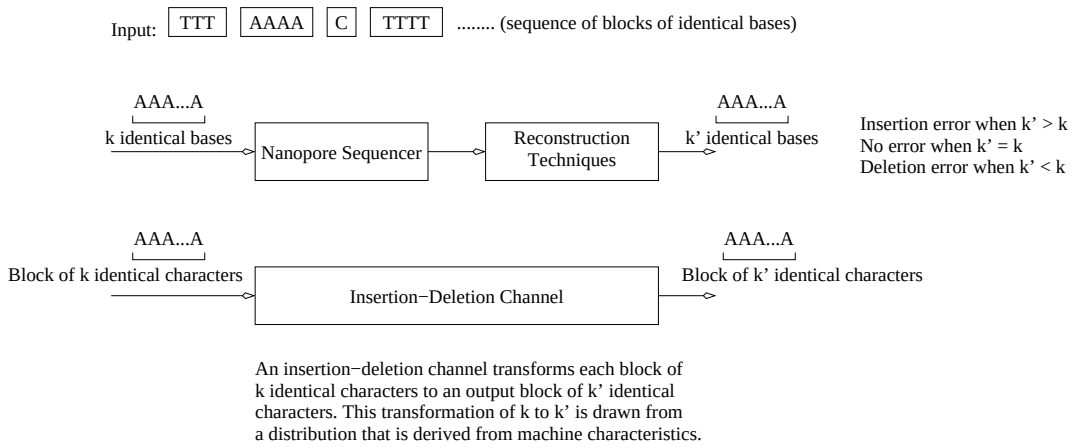

Figure 1. Overview of the proposed channel and its correspondence with a sequencer.

A weaker condition is that the distortion ball contains only a single typical sequence. We refer to the former as an accurate reconstruction and the latter as a unique reconstruction.

A single pass through the nanopore induces a distortion ball whose probability profile can be quantified. We show that the radius of this ball can be reduced by repeated extrusion through the nanopore. One of the key contributions of this paper is that the radius shrinks rapidly with the number of extrusions, thus enabling accurate and/or unique reconstruction with relatively small number of repeats.

3 METHODS

3.1 Notation and Theoretical Model

The input sequence to the channel/ sequencer is drawn from an alphabet $\mathcal{A}$. For DNA sequencing, $\mathcal{A} = \{A, T, C, G\}$. The alphabet size $|\mathcal{A}|$ is denoted by m. We assume that an n length input sequence X is independent and identically distributed (i.i.d.); i.e., each sequence has the same probability distribution as the others and all sequences are mutually independent. Mathematically, probability $\Pr(x_i = s) = p_s,\ \sum_s p_s = 1$.

3.1.1 Blocking Identical Symbols As mentioned, we view input and output sequences as sequences of blocks, with each block comprised of one or more identical symbols. If s^k denotes $k \geq 1$ repeats of symbol s, we can write sequence X as concatenation of $N \leq n$ blocks: $X = s_1^{k_1} \dots s_N^{k_N}$, such that $k_i > 0$, $s_{i+1} \neq s_i$. For example, a sequence X="AATATTAA" is represented in the block form as $A^2TAT^2A^2$.

We initiate our discussion by enumerating basic statistical properties of block sequences. We see that:

$$\Pr(s_{i+1} = s' | s_i = s \neq s') = \frac{p_{s'}}{1 - p_s}. \tag{1}$$

Since we have a block of symbols s, the block must start with at least one appearance of symbol s. The probability that the block will have length $k \geq 1$ is given by:

$$P_{sk} := \Pr(k_i = k | s_i = s) = p_s^{k-1}(1 - p_s) \tag{2}$$

$$\text{satisfying} \quad \sum_{k \geq 1} P_{sk} = 1.$$

The expected length a block of symbols s is given by:

$$\overline{k}_s := \sum_{k \geq 1} k P_{sk} = \sum_{k \geq 0} k p_s^{k-1}(1 - p_s) = \frac{1}{1 - p_s}.$$

The expected number of symbols s in the entire sequence of length n is np_s. Therefore, the expected number of blocks of type s, and of all blocks is, respectively,

$$N_s := np_s / \overline{k}_s = np_s(1 - p_s), \tag{3}$$

$$N := \sum_s N_s = n \left(1 - \sum_s p_s^2\right). \tag{4}$$

3.2 Sticky Insertion-Deletion Channel (InDel)

We model the sequencer using a sticky insertion-deletion channel. In this channel, a block of k consecutive identical symbols, with probability q_{kl}, is transformed into a block of l copies of the symbol:

$$\Pr(s^l | s^k) = q_{kl}(s) \qquad \text{where} \quad \sum_l q_{kl}(s) = 1. \tag{5}$$

The term sticky is used to imply that the block structure remains unchanged; i.e., $q_{k0}(s) = q_{0l}(s) = 0$.
In this model, $\{q_{kl}(s)\}$ specifies the block length change probabilities. We can assume that for given k, $q_{kl}(s)$ has a maxima at $l = k$; i.e., the probability that there is no error in a block exceeds any other (erroneous) transformation. In Section 4, we consider two specific choices for function q: an exponential distribution and independent insertion-deletions. In this model, the probability of observing an output sequence Y from our InDel channel is given by:

$$\Pr(Y = s_1^{l_1} s_2^{l_2} \dots s_N^{l_N} | X = s_1^{k_1} s_2^{k_2} \dots s_N^{k_N}) = \prod_{i=1}^{N} q_{k_i l_i}(s_i). \tag{6}$$

For example, true sequence X="AAAATTAAA"=$A^4T^2A^3$ is read by the sequencer as Y="AATTTAA"=$A^2T^3A^2$ with probability $q_{42}(A) \cdot q_{23}(T) \cdot q_{31}(A)$.
For simplicity, we assume that q is identical for different symbols s:

$$q_{kl}(s) \equiv q_{kl}. \tag{7}$$

This implies that insertion-deletion errors in sequencing are independent of the bases. Please note that this assumption is not a limitation of

our framework. Assuming $q_{k0} = q_{0l} = 0$, the number of blocks and their order remain the same while applying the channel. However, the lengths of blocks may change. The sequencing problem can then be reduced to the following task: determine the original set of block counts $(k_i)_{i=1..N}$ from the following two cases under consideration: (i) the single extrusion case – a single read sequence Y: $(l_i)_{i=1..N}$; and (ii) the multiple extrusion case – each sequence is repeated $c \in \mathbb{N}$ times: $(l_i^j)_{i=1..N}$ for $j = 1, \ldots c$.

3.3 Accurate Reconstruction from Nanopore Sequencers

We now consider the problem of constructing a true sequence (from the channel or the sequencer) from an observed sequence (from the ionic current measurement). This problem is one of reconstructing a true block sequence from an observed block sequence at the output of the sticky channel. Since we assume that the block structure is not changed by the channel, this problem can be solved one block at a time. Specifically, we observe a block at the output of the channel of length l and we must infer the block length of the corresponding input, k. We refer to this as the problem of finding the most probable reconstruction.

3.3.1 Single Run Estimation: Inferring k from a Single l In the first instance of the problem, we do not consider any repeats – i.e., a single block passes through the channel (sequencer) only once. From this single observation of l, we must infer k. The probability that an output block of length l was observed from an input block of length k is given by:

$$Pr(s^k|s^l) = \frac{P_{sk}q_{kl}}{\sum_{k'} P_{sk'}q_{k'l}} \tag{8}$$

where $P_{sk} = \Pr(k_i = k|s_i = s) = p_s^{k-1}(1-p_s)$.

In this case, the most likely input block length k for observed output block length l is the one that maximizes $p_s^{k-1}q_{kl}$ for given l. Let us denote this k by k_l. We then have:

$$p_s^{k_l-1}q_{k_l l} = \max_k p_s^{k-1}q_{kl} \tag{9}$$

We would expect that $k_l = l$. However, for a general distribution q, this is not necessarily the case. The natural condition for q, given by $\max_k q_{kl} = q_{ll}$, turns out not to be sufficient for this purpose. Even with this condition, a single input block length k may correspond to multiple corresponding observed block lengths l, and some input block lengths k might not have corresponding values of l at all. Consequently, we need a stronger condition. It is easy to see that:

$$\forall_{l,i,s}\ q_{l-i,l} < (p_s)^i \cdot q_{l,l} \quad \Rightarrow \quad k_l = l \tag{10}$$

We can now determine the probability that an observed block is properly corrected as:

$$m_k := \sum_{l:k_l=k} q_{kl} \qquad (0 \text{ if } k \text{ cannot be obtained}),$$

which reduces to $m_k = q_{kk}$ if (10) is satisfied. The expected number of blocks of symbol s is $N_s = np_s(1-p_s)$. Their expected total length is np_s. Therefore, the probability that we accurately correct all blocks is asymptotically given by:

$$\prod_s \left(\sum_k P_{sk}m_k\right)^{N_s} = 2^{n\sum_s p_s(1-p_s)\lg(\sum_k P_{sk}m_k)} \tag{11}$$

It is easy to see that this probability decreases exponentially with the length of the sequence. Stated otherwise, this result shows that the probability that we accurately reconstruct the entire sequence decreases exponentially in the length of the sequence.

3.3.2 Multiple runs: estimating k from multiple l: We now investigate how reading the same block multiple times can help infer the input block length. We assume that each block is read c times. This can be done by extruding the same sequence through an array of nanopores. In this case, we have c observed values of l_i (the ith block length), $(l_i^j)_{i=1..N}$ for $j = 1, \ldots c$. The probability that the corresponding input block length is k is given by:

$$Pr(s^k|s^{l^1}, .., s^{l^c}) = \frac{P_{sk}q_{kl^1}...q_{kl^c}}{\sum_{k'} P_{sk'}q_{k'l^1}...q_{k'l^c}}. \tag{12}$$

Let us analogously define $k_{l^1,..,l^c}$ as the input block length k that maximizes $p_s^{k-1} \cdot q_{kl^1}...q_{kl^c}$, given by:

$$p_s^{k_{l^1,..,l^c}-1}q_{k_{l^1,..,l^c}l^1}...q_{k_{l^1,..,l^c}l^c} = \max_k p_s^{k-1}q_{kl^1}...q_{kl^c}. \tag{13}$$

The probability that input block length k is accurately determined is given by:

$$m_{kc} := \sum_{l_1,..,l_c:k_{l_1,..,l_c}=k} q_{kl^1}..q_{kl^c}. \tag{14}$$

As before, the probability that we accurately infer all block sizes (accurate sequencing) in analogy to (11) decreases exponentially as:

$$2^{n\sum_s p_s(1-p_s)\lg(\sum_k P_{sk}m_{kc})}. \tag{15}$$

Unfortunately the problem of finding $k_{l^1,..,l^c}$ is a complex estimation procedure, and therefore finding the required number of repeats, c, for unique reconstruction appears difficult. However, as we show next, using tools from information theory, we can estimate this repeat rate. More importantly, we show that this repeat rate is a slow function of sequence length.

3.4 Fundamental Bounds for Unique Reconstruction

We rely on an information theoretic approach to computing the minimum number of repeats c required for accurate reconstruction. We do this by modifying the original problem somewhat. Recall that the noise model of the channel introduces a distortion ball around the output sequence. It is possible that multiple input sequences belong in this distortion ball – leading to the problem of identification of the most probable reconstruction. However, if we could use repeated extrusion of blocks to shrink the distortion radius to the point where only one sequence belongs in the ball, we have unique reconstruction. We use this principle to focus on the problem of number of *typical sequences* that belong in the distortion ball, and find the number of repeats c for which this number approaches one. Please note that this problem is slightly distinct from the problem of most probable reconstruction.

3.4.1 Difference Between the Most Probable and Typical Reconstruction We begin our discussion by highlighting the differences between the *most probable* and *typical* reconstructions. To gain an intuition about the difference between these two types of reconstructions, let us briefly look at error correction of the basic binary symmetric channel (BSC): we send N bits, each of them has independent probability $\epsilon < 1/2$ of being flipped. Observe that we could write this in the formalism we have introduced for blocks as: $k, l \in \{0,1\},\ q_{00} = q_{11} = 1-\epsilon,\ q_{01} = q_{10} = \epsilon$.

Obtaining from the output sequence $Y \in \{0,1\}^N$, the most probable input sequence X is simple – it is simply $X = Y$. The probability that this input sequence X is the correct sequence is given by: $(1-\epsilon)^N = 2^{N\lg(1-\epsilon)}$. However, for large N, we expect that approximately ϵN bits are flipped. There are

$$\binom{N}{\epsilon N} \approx 2^{Nh(\epsilon)}$$

$$(\text{where} \quad h(p) = -p\lg(p) - (1-p)\lg(1-p))$$

different ways of doing it. These are all *typical corrections*. In this case, it is easy to see that relative entropy $H(X|Y) = N \cdot h(\epsilon)$. Having

no additional information, all of these typical corrections are equally probable. Consequently, the probability of choosing the correct one is given by $2^{-H(X|Y)} = 2^{-Nh(\epsilon)}$. Conversely, the definition of the channel says that the probability that a given typical correction (flipped ϵN bits) is the correct one is asymptotically given by:

$$\epsilon^{N\epsilon} \cdot (1-\epsilon)^{N(1-\epsilon)} = 2^{-Nh(\epsilon)} = 2^{-H(X|Y)}$$

– exactly the same as before.

To summarize, typical corrections correspond to a Hamming sphere $S(Y, \epsilon N)$. The most probable correction corresponds to its center and for unique reconstruction, this sphere should reduce to a point; i.e., $H(X|Y) \approx 0$.

3.4.2 Entropy in the Framework of Blocks of Identical Symbols Returning to our original problem, by definition of a typical sequence, the number of typical sequences of length n, X^n, grows asymptotically as $\exp(H(X^n))$, where

$$H(X^n) = nh^x \qquad \text{and} \qquad h^x := -\sum_s p_s \ln(p_s). \quad (16)$$

We now derive this entropy formula in the block framework for the asymptotic case (large n). This analysis is analogous to the result of Mitzenmacher et al. [3], where the non-asymptotic case is presented. We must deal with two additional considerations here: our alphabet is not binary; and the distribution among input symbols is not necessarily uniform.

The information contained in the input sequence in the block framework: $X^n = s_1^{k_1} \dots s_N^{k_N}$ can be split into two parts – the sequence of symbols (s_i), and corresponding block lengths(k_i). $H(X^n)$ is sum of the two entropies. Using results from Section (3.1.1), the entropy of selecting the symbol for succeeding block ($s_{i+1} \neq s_i$) is given by:

$$\begin{aligned} h_s \equiv H(s_{i+1}|s_i = s) &= -\sum_{s' \neq s} \frac{p_{s'}}{1-p_s} \ln\left(\frac{p_{s'}}{1-p_s}\right) \\ &= \frac{1}{1-p_s} \sum_{s' \neq s} p_{s'}(\ln(1-p_s) - \ln(p_{s'})) \\ &= \frac{1}{1-p_s}\left((1-p_s)\ln(1-p_s) + h^x + p_s \ln(p_s)\right) \\ &= \frac{h^x - h(p_s)}{1-p_s}. \end{aligned} \quad (17)$$

The entropy of choosing block length for symbol s is given by:

$$\begin{aligned} h_s^x \equiv H(k_i|s_i = s) &= -\sum_{k \geq 1} P_{sk} \ln(P_{sk}) \\ &= -\sum_{k \geq 1} p_s^{k-1}(1-p_s) \ln\left(p_s^{k-1}(1-p_s)\right) \\ &= -\frac{p_s}{1-p_s}\ln(p_s) - \ln(1-p_s) \\ &= \frac{h(p_s)}{1-p_s}. \end{aligned} \quad (18)$$

because $\sum_{k \geq 0} k z^k = z/(1-z)^2$.

We can now express (16) in the block framework. The source entropy $H(X^n)$ is the sum of entropy of symbol order (h_s) and block lengths (h_s^x):

$$\begin{aligned} H(X^n) &= \sum_s N_s(h_s + h_s^x) = n \sum_s p_s(1-p_s)\frac{h^x}{1-p_s} \\ &= nh^x \sum_s p_s = nh^x. \end{aligned} \quad (19)$$

3.4.3 InDel Channel with a Single Extrusion Assuming the sticky insertion-deletion channel (InDel) described above, the sequence of symbols s_i is unmodified: $Y^n = s_1^{l_1} \dots s_N{}^{l_N}$. Only the block lengths are changed in accordance with the noise model ($k_i \to l_i$). Analogously, as in the previous section, we can determine the entropy of joint distribution $H(X^n, Y^n)$ as:

$$H(X^n) + H(Y^n|X^n) = H(X^n, Y^n) = \sum_s N_s(h_s + h_s^{xy}) \quad (20)$$

where h_s^{xy} is entropy of pair lengths for input (k) and output (l) type s blocks:

$$\begin{aligned} h_s^{xy} &= -\sum_{k,l \geq 1} P_{sk} q_{kl} \ln(P_{sk} q_{kl}) \\ &= h_s^x - \sum_{k \geq 1} P_{sk} \sum_{l \geq 1} q_{kl} \ln(q_{kl}) \\ &= h_s^x + \sum_{k \geq 1} P_{sk} h_k^q \end{aligned} \quad (21)$$

and $h_k^q := -\sum_{l \geq 1} q_{kl} \ln(q_{kl})$.

The entropy of output sequence Y and mutual information $I(X^n; Y^n) = H(X^n) + H(Y^n) - H(X^n, Y^n)$ are given by:

$$H(Y^n) = \sum_s N_s(h_s + h_s^y)$$

$$\text{for} \qquad h_s^y = -\sum_{l \geq 1}\left(\sum_{k \geq 1} P_{sk} q_{kl}\right) \ln\left(\sum_{k \geq 1} P_{sk} q_{kl}\right) \quad (22)$$

$$\begin{aligned} I(X^n; Y^n) &= \sum_s N_s\left((h_s + h_s^x) + (h_s + h_s^y) - (h_s + h_s^{xy})\right) \\ &= \sum_s N_s(h_s + h_s^x + h_s^y - h_s^{xy}) \end{aligned} \quad (23)$$

$$\begin{aligned} H(X^n|Y^n) &= H(X^n) - I(X^n; Y^n) \\ &= \sum_s N_s(h_s^{xy} - h_s^y) \\ &= n \sum_s p_s(1-p_s)(h_s^{xy} - h_s^y). \end{aligned} \quad (24)$$

Asymptotically, the number of typical corrections is given by $2^{H(X^n|Y^n)}$, which grows exponentially in the length of the sequence n. This directly implies the exponentially decreasing probability of accurate reconstruction. We now discuss how the number of typical corrections can be reduced (approaching 1) by reading the input sequence multiple times (c). The goal of this analysis is to estimate the number of extrusions we should perform for unique reconstruction.

3.4.4 InDel Channel with Multiple Extrusions We consider the case of c repeats on each block: $(l_i^j)_{i=1..N}$ for $j = 1, \dots c$. In

this case, we have:

$$h_s^{xyc} = -\sum_{k,l^1,..,l^c\geq 1} P_{sk}q_{kl^1}..q_{kl^c}\ln\left(P_{sk}q_{kl^1}..q_{kl^c}\right)$$
$$= h_s^x + c\sum_{k\geq 1} P_{sk}h_k^q$$

$$h_s^{yc} =$$
$$-\sum_{l^1..l^c\geq 1}\left(\sum_{k\geq 1} P_{sk}q_{kl^1}\cdot..\cdot q_{kl^c}\right)\ln\left(\sum_{k\geq 1} P_{sk}q_{kl^1}\cdot..\cdot q_{kl^c}\right)$$

$$h_s^{xyc} - h_s^{yc} =$$
$$\sum_{k\geq 1} P_{sk}\sum_{l^1..l^c\geq 1} q_{kl^1}..q_{kl^c}\ln\left(1+\frac{\sum_{k'\neq k}P_{sk'}q_{k'l^1}..q_{k'l^c}}{P_{sk}q_{kl^1}..q_{kl^c}}\right).$$

Grouping q_{ki} corresponding to the same i by assigning $\tilde{l}^i = \#\{j : l^j = i\}$, using $n! \approx (n/e)^n$, the distribution becomes ($\sum_i \tilde{l}^i = c$):

$$\sum_{l^1..l^c\geq 1} q_{kl^1}..q_{kl^c} = \sum_{\tilde{l}^1,\tilde{l}^2,...}\binom{c}{\tilde{l}^1,\tilde{l}^2,...}\prod_{i\geq 1} q_{ki}^{\tilde{l}^i}$$
$$\approx \sum_{\tilde{l}^1,..}\prod_{i\geq 1}\left(\frac{c}{\tilde{l}^i}\right)^{\tilde{l}^i} q_{ki}^{\tilde{l}_i}$$
$$= \sum_{\tilde{l}^1,..}\exp\left(-c\sum_{i\geq 1}\frac{\tilde{l}^i}{c}\ln\left(\frac{\tilde{l}^i/c}{q_{ki}}\right)\right).$$

The sum in bracket is the Kullback-Leibler (asymmetric) distance: $D_{KL}\left(\left\{\tilde{l}^i/c\right\}_i \mid\mid \{q_{ki}\}_i\right)$ - the exponent is asymptotically (large c) dominated by $\tilde{l}^i/c = q_{ki}$ distribution. To find an approximation of the formula $h_s^{xyc} - h_s^{yc}$, we focus only on these distributions:

$$h_s^{xyc} - h_s^{yc} \approx \sum_{k\geq 1} P_{sk}\ln\left(1+\frac{\sum_{k'\neq k}P_{sk'}\left(\prod_{l\geq 1}q_{k'l}^{q_{kl}}\right)^c}{P_{sk}\left(\prod_{l\geq 1}q_{kl}^{q_{kl}}\right)^c}\right)$$
$$\approx \sum_{k\geq 1}\sum_{k'\neq k} P_{sk'}\left(\prod_{l\geq 1}\left(\frac{q_{k'l}}{q_{kl}}\right)^{q_{kl}}\right)^c.$$

Since $H(X|Y) = H(X,Y) - H(Y)$ and $P_{sk} = p_s^{k-1}(1-p_s)$, we can write:

$$H(X^n|Y^{nc}) = n\sum_s p_s(1-p_s)(h_s^{xyc} - h_s^{yc})$$
$$\approx n\sum_s (1-p_s)^2\sum_{k\geq 1}\sum_{k'\neq k} p_s^{k'}\exp(-c\cdot d_{kk'})$$
$$=: n\cdot D(c) \qquad (25)$$

where $d_{kk'}$ is the Kullback-Leibler distance:

$$d_{kk'} := -\ln\left(\prod_{l\geq 1}\left(\frac{q_{k'l}}{q_{kl}}\right)^{q_{kl}}\right)$$
$$= \sum_{l\geq 1} q_{kl}\ln\left(\frac{q_{kl}}{q_{k'l}}\right)$$
$$= D_{KL}\left(\{q_{kl}\}_l \mid\mid \{q_{k'l}\}_l\right)$$

and $D(c)$ is asymptotically dominated by the $\overline{d} := \min_{k'\neq k} d_{kk'} = d_{k_0k'_0}$ term (if the minimum exists):

$$D(c) = \sum_s (1-p_s)^2\sum_{k\geq 1}\sum_{k'\neq k} p_s^{k'}\exp(-c\cdot d_{kk'})$$
$$\approx \exp(-c\cdot\overline{d})\cdot\sum_s (1-p_s)^2 p_s^{k'_0}$$

The distance $d_{kk'}$ describes the similarity between the results of reading blocks having original lengths k and k'. It quantifies the likelihood of mistakenly identifying a k length block as a k' length block. The smaller it is, the faster is the growth of number of typical corrections ($2^{H(X^n|Y^{nc})}$) with k erroneously replaced by k'. The smallest distance ($\overline{d}$) corresponds to the most likely mistake, and it asymptotically dominates the growth of typical corrections.

The use of multiple extrusions (c) allows us to reduce the exponent in the number of typical corrections $2^{H(X^n|Y^{nc})}$. For unique reconstruction, the number of typical corrections must approach 1. Consequently, we choose c such that $n\cdot\exp(-c\cdot\overline{d})$ is of order of 1. From this, we see that the number of extrusions should grow logarithmically in sequence length: $c \approx \ln(n)/\overline{d}$. This important result establishes the feasibility of low-overhead sequencing using nanopore sequencers.

4 EXPERIMENTAL RESULTS

We present a simulation study of the implications of our analysis on real-world sequencing experiments. We consider two models for our channel – the first model is a sticky channel with exponential distribution and the second, an independent insertion-deletion channel. In each case, we examine the bound on length of sequence for accurate reconstruction, the number of repeats needed for larger reconstructions, and the Kullback-Leibler distance. The goal of these studies is to demonstrate that for a wide class of channel characteristics: (i) the length of sequence that can be accurately reconstructed in single read is small; (ii) the number of required repeats for longer reconstructions is a slowly growing function (logarithmic); and (iii) even in the presence of high InDel error rates, nanopore sequencers can accurately reconstruct sequences with required number of repeats.

We consider a binary equi-probable input – $m = 2$, $p_0 = p_1 = 1/2$. The behavior of relative entropy $H(X|Y)$ is primarily determined by the q_{ks} distribution. The results in this section can be naturally generalized to any alphabet and probability distribution.

Exponential Insertion-Deletion Error Model. We will first consider a sticky channel with exponential distribution for the error probabilities – for some $0 < q < 1$:

$$q_{kl} = q^{|k-l|}\frac{1-q}{1+q-q^k}$$

The second term in the product is for normalizing the probability.

We first consider the single run case. Equation 24 allows us to calculate relative entropy $H(X^n|Y^n)$, describing the growth in the number of typical corrections: $2^{H(X^n|Y^n)}$. The left Panel of Figure 2 presents values of relative entropy for various values of q. Assuming correction procedure as taking a random typical correction, the Right Panel of this figure presents the probability of obtaining the right correction. This probability drops exponentially with the length of the sequence, making a single run approach impractical for longer sequences.

This limitation can be handled by performing multiple extrusions of the same sequence. We use Equation (25) to find

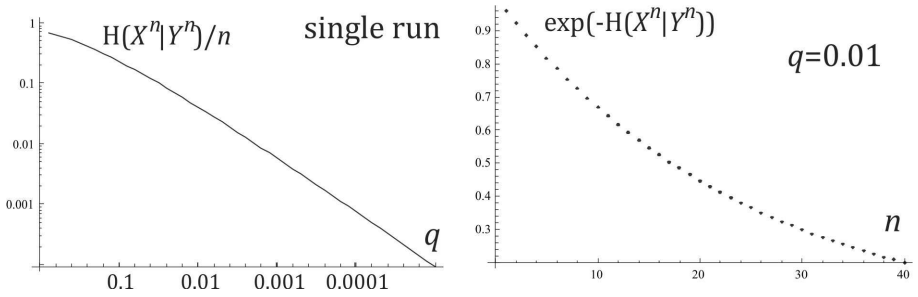

Figure 2. Left Panel: Relative entropy for single extrusion between input sequence (input to the nanopore sequencer) and the output sequence (observed sequence). This is derived from Equation (24). Right Panel: The probability that we select the correct typical correction for $q = 0.01$ and increasing value of n.

$k' \rightarrow$ $k \downarrow$	1	2	3	4	5	6	7
1	0	0.223	0.665	1.123	1.857	2.512	3.194
2	0.193	0	0.234	0.694	1.270	1.905	2.568
3	0.567	0.220	0	0.233	0.696	1.274	1.909
4	1.054	0.644	0.227	0	0.232	0.695	1.273
5	1.621	1.181	0.671	0.229	0	0.232	0.694

Table 1. Kullback-Leibler distances for $q = 0.5$, exponential distribution, and various original block lengths: k, k'.

$k' \rightarrow$ $k \downarrow$	1	2	3	4	5	6	7
1	0	1.316	3.147	5.118	7.166	9.262	11.390
2	∞	0	0.896	2.445	4.212	6.094	8.050
3	∞	∞	0	0.668	1.986	3.570	5.299
4	∞	∞	∞	0	0.527	1.666	3.094
5	∞	∞	∞	∞	0	0.434	2.730

Table 2. Similarity of q_{kl} for different values of k for the independent insertion-deletion channel.

relative entropy in this case. This requires finding Kullback-Leibler distances between $\{q_{kl}\}_l$ distributions for different original block lengths k. Table 1 presents some of these values for $q = 0.5$.

The minimal distance is $\overline{d} = d_{21} \approx 0.193$, and corresponds to misinterpreting original $k = 2$ sequence as $k' = 1$. This intuitively stronger overlap of the first two distribution can be observed in the Left Panel of Figure 3, containing $\{q_{kl}\}_l$ for the first 15 values of k. The distance between farther neighboring distributions is nearly the same; i.e., misreading a block of length five nucleotides as six, is as likely as an input block of 100 nucleotides being read as 101.

The right panel of Figure 3 shows relative entropy as a function of number of repeats c, for $q = 0.5$. There are important observations drawn from this figure: (i) the nearly linear nature of the curve shows that the number of repeats is almost logarithmic in the read length ($H(X^n|Y^n)/n$ asymptotically behaves as $\exp(-c \cdot \overline{d})$); and (ii) for realistic reconstruction lengths (say, 100K bases), the number of repeats is relatively small (less than 60). These are important results that establish the feasibility of nanopore sequencers for accurate low-cost construction of long reads.

Independent Insertion-Deletion Channel Model. To demonstrate the robustness of our results we now consider a different channel model – the independent insertion-deletion channel. In this model, for each nucleotide, there is probability $\epsilon > 0$, that the symbol is deleted. There is also an identical probability that the symbol is duplicated. It follows that there is a probability $1 - 2\epsilon$ that the symbol is sequenced without an error. For this model, q_{kl}, for given k, is the convolution of k such random variables, additionally truncated to enforce that it is a sticky channel; i.e., $q_{k0} = 0$. For large k, q_{kl} approaches the Gaussian distribution with standard deviation $\sqrt{2\epsilon k}$.

Figure 4 shows the first 20 distributions (q_{kl}) for this model. It is illustrative to note that unlike the previous models, larger block lengths have higher insertion-deletion error rates. Table 2 presents the approximated first values of similarities of q_{kl} for different values of k, for $\epsilon = 0.1$. The infinite values in the table correspond to difference in support (one of values is zero). We note that the distributions become closer to their own neighbors as k grows: $\lim_{k\to\infty} d_{kk+1} = 0$, the minimal nonzero distance $\overline{d}$ does not exist. In other words, distinguishing between k and $k+1$ becomes more difficult with increasing k, and their difference vanishes asymptotically. Consequently, as we can observe from the right panel of Figure 4, the required number of repeats grows slower than logarithm of sequence length. Figure 4 shows the relative entropy of the input and output sequences for different numbers of repeats (c). long sequences (100K nucleotides), we need even fewer repeats (less than 30, in this case, compared to 60 for the exponential model).

5 RELATED RESEARCH

Technologies underlying nanopore sequencers have been investigated for over a decade [2, 1]. Commercial platforms based on these technologies have only recently been announced – with Oxford Nano being the leading platform. An excellent introduction to this platform is available at: `https://www.nanoporetech.com/technology/analytes-and-applications-dna-rna-proteins/`

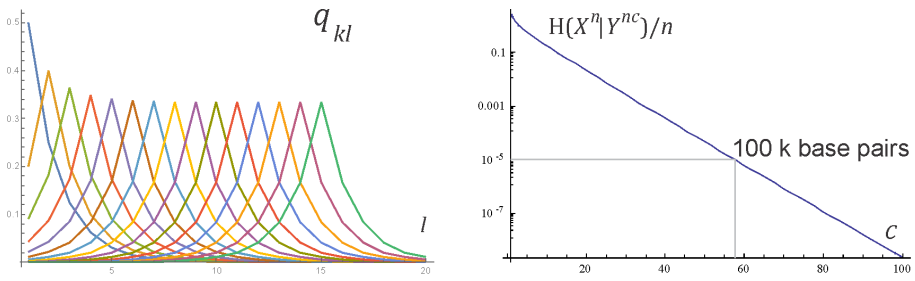

Figure 3. Left Panel: $\{q_{kl}\}_l$ distributions for $q = 0.5$ and $k = 1$ to 15. Right Panel: (25) approximation for $q = 0.5$ and different numbers of copies c.

`dna-an-introduction-to-nanopore-sequencing`. There have been preliminary efforts aimed at characterizing the performance of nanopore sequencing platforms in terms of error rate, error classification, and run lengths [7, 8, 4, 10]. A consensus emerges from these studies that the primary error mode in nanopore sequencers is deletion errors and that the error rate is approximately 4% with a read length of over 150K bases. These studies provide important data that is used to build our insertion-deletion channel.

Error characteristics and models for nanopore sequencers have been recently studied by O'Donnell et al [8]. In this study, the authors investigate error characteristics, and build a statistical model for errors. They use this model to show, through a simulation study, that repeated extrusion can be used to improve error characteristics. In particular, they show that using their model, it is possible to achieve 99.99% accuracy by repeating the read 140 times. This empirical study provides excellent context for our analytical study, which provides rigorous bounds and required repeat rates.

There has been significant work on different channels, their capacities, and error characteristics over the past five decades since the work of Shannon. Of particular relevance to our results is the work in deletion channels [5]. As mentioned, the capacity of independent deletion channels is as-yet unknown. There have been efforts aimed at error correction in insertion-deletion channels in the context of communication, storage, and RFID systems [11]. We are, however, unaware of any results aimed at the use of insertion-deletion channels to establish fundamental bounds on performance of sequencers. Our channel itself is novel, its analysis is new, and the associated bounds on reconstruction length, and required repeat rates are presented for the first time.

6 DISCUSSION AND CONCLUSION

In this paper, we present a novel modeling methodology based on a channel representation of a nanopore sequencer. We use this methodology to show a number of important results: (i) the high deletion error rate of the nanopore sequencer limits the sequence length that can be accurately reconstructed; (ii) repeated extrusion through the nanopore is an effective technique for increasing the accurate reconstruction length; (iii) the number of repeats is a slow function of the sequence length (logarithmic in sequence length), enabling nanopore sequencers to accurately reconstruct long sequences at low cost.

We demonstrate our results for a wide class of error models and show that our analyses is robust. We note that our analyses only considers insertion-deletion errors, and not substitution errors. This is justified by the fact that deletion errors constitute the primary error mode in nanopore sequencers. In the presence of substitution errors, our analyses can be viewed as providing bounds on performance and required repeats for accurate reconstruction.

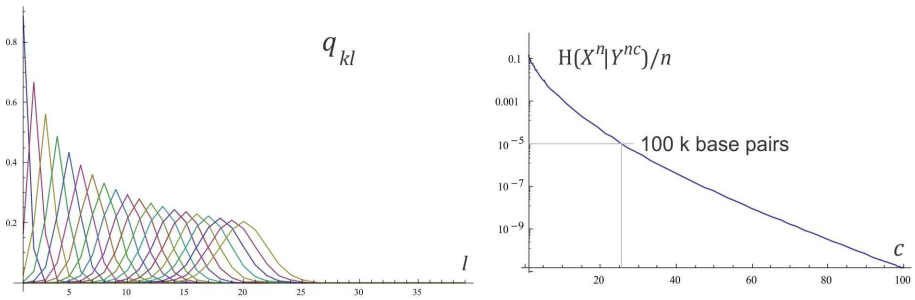

Figure 4. Left Panel: The first 20 q distributions for $\epsilon = 0.1$ for the independent insertion-deletion channel. Right Panel: Approximation of joint entropy for $\epsilon = 0.1$ and different numbers of copies c (from Equation (25).

REFERENCES

[1]T. Butler, M. Pavlenok, I. Derrington, M. Niederweis, and J. Gundlach. Single-molecule dna detection with an engineered mspa protein nanopore. *Proceedings of the National Academy of Science*, 105(52):20647–20652, 2008.

[2]D. Deamer and D. Branton. Characterization of nucleic acids by nanopore analysis. *Acc Chem Res*, 35(10):817–825, 2002.

[3]E. Drinea and M. Mitzenmacher. Improved lower bounds for the capacity of i.i.d. deletion and duplication channels. *IEEE Transactions on Information Theory*, 53:8:2693–2714, 2007.

[4]E. Hayden. Nanopore genome sequencer makes its debut. *Nature News*, Feb. 2012.

[5]Ian A. Kash, Michael Mitzenmacher, Justin Thaler, and Jonathan Ullman. On the zero-error capacity threshold for deletion channels. *CoRR*, abs/1102.0040, 2011.

[6]J. Kevin Lanctot, Ming Li, and En-hui Yang. Estimating dna sequence entropy. In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '00, pages 409–418, Philadelphia, PA, USA, 2000. Society for Industrial and Applied Mathematics.

[7]A. Mikheyev and M. Tin. A first look at the oxford nanopore minion sequencer. *Molecular Ecology Resources*, 14(6):1097–1102, Nov. 2014.

[8]C. O'Donnell, H. Wang, and W. Dunbar. Error analysis of idealized nanopore sequencing. *Electrophoresis*, 34(15):2137-44, 2013.

[9]M. Quail, M. Smith, P. Coupland, T. Otto, S. Harris, T. Connor, A. Bertoni, H. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC Genomics*, 13:341, 2012.

[10]J. Schreiber, Z. Wescoe, R. Abu-Shumays, J. Vivian, B. Baatar, K. Karplus, and Mark Akeson. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual dna strands. *Proceedings of the National Academy of Science*, 110(47), Nov. 19, 2013.

[11]Guang Yang, Angela I. Barbero, Eirik Rosnes, and Yvind Ytrehus. Error correction on an insertion/deletion channel applying codes from rfid standards. *ITA*, 2012.