LDPC Codes for Portable DNA Storage

Peng Fei
Center for Pervasive Communications and Computing
University of California, Irvine
Email: pfei1@uci.edu

Zhiying Wang
Center for Pervasive Communications and Computing
University of California, Irvine
Email: zhiying@uci.edu

Abstract—DNA becomes an attractive storage medium in recent years for its ultra-high density, millennial-long retention, and efficient replication. In this work we consider DNA storage that uses the affordable and portable nanopore sequencing as the reading mechanics. Unlike traditional data storage systems, errors occur asymmetrically among the four types of nucleotide bases of DNA. Quaternary codes can be employed for error correction, but suffer from high complexity. In this paper, we design binary LDPC codes with a turbo-like decoder for the DNA storage channel. Simulation results show that our binary LDPC codes have a similar bit-error rate but with a speed up by an order of magnitude compared to quaternary codes.

I. INTRODUCTION

With the fast growth of data, DNA storage attracts much attention for its extremely large data densities, integrity in non-ideal conditions over millennial, and efficient data replication (e.g., [1]–[6]). In DNA storage, the writing process is called synthesis, which joins nucleotide symbols and produces the desired DNA string. Reading is completed through a DNA sequencer that reads the string and translates it to digital data. Among the DNA sequencing technologies, nanopore sequencing performed on the MinION device [7] is a promising one due to its low cost, scalability, and portability, and has been demonstrated suitable for DNA storage [5]. In nanopore sequencing, a DNA sequence is passed through a nano-sized hole and current variation is measured to estimate the sequence symbols.

Errors in nanopore sequencing [8] should be addressed in order to maintain data integrity. In [9] burst deletion errors are considered and associated codes are constructed. In [10] a channel model of nanopore sequencing is established capturing non-linear inter-symbol interference, deletions, and substitutions. In [11] the measured current is modeled as a quaternary amplitude modulation with additive white Gaussian noise. In [12] it is observed that substitution errors occur asymmetrically among the four bases of nucleotides (A, T, G, C), and the minimum asymmetric Lee distance are studied.

The goal of this work is to design LDPC codes for the asymmetric substitutions of nanopore sequencing. Instead of studying the minimum code distance as in [12], we focus on the performance in terms of bit error rate, and LDPC codes are chosen for its near-capacity performance and low encoding/decoding complexity. As the DNA string is a sequence of alphabet size 4, a quaternary code is a natural candidate for error correction. Although the LDPC code supports non-binary alphabets, it suffers from the slow speed in decoding compared to the binary case. Hence, we focus on binary LDPC codes for DNA storage.

In this paper we present an asymmetric nanopore channel model. We propose to use two binary LDPC codes for one quaternary DNA sequence, and show decoding algorithms that exchange information between the two LDPC codes. The information exchange step is inspired by Turbo codes, and addresses the asymmetry in the channel errors. Simulation results show that our proposed binary codes have bit error rate comparable to quaternary codes, but has a 10x speed up.

Ⅱ. PRELIMINARIES

We first introduce the asymmetric channel model of nanopore sequencing. In an (n, k) DNA code, 2k bits of information are encoded into a DNA codeword \( x = (x_1, x_2, \ldots, x_n), x_i \in \{A, T, G, C\}, 1 \leq i \leq n \). Nanopore sequencing can read the DNA codeword by monitoring small changes in the ionic current flowing through the pore and dwell time in the nanopore [8], the distribution of which is shown approximately in Fig. 1. Hard estimation decisions are made from the current and time measurements with substitution errors. In other words, the channel is discrete and memoryless. The received word is denoted by \( y = (y_1, y_2, \ldots, y_n) \), \( y_i \in \{A, T, G, C\}, 1 \leq i \leq n \).

![Fig. 1. Measurement distribution in nanopore sequencing.](image)

From Fig. 1, we can see that the nanopore channel is asymmetric among the possible erroneous transitions: transitions between T and C are more often, however less so between G and A. The error probabilities are shown as in Fig. 2, for some \( p_3 < p_1 < p_2 \). This is a slight generalization of the model in [12], and we assume \( p_1, p_2, p_3 \) satisfy the specific relation that

\[
p_1 = \lambda, p_2 = 3\lambda, p_3 \approx 0, \tag{1}
\]
for some parameter $0 < \lambda < 1/3$. We note that for other choices of error probabilities such that $p_3 \ll p_1 < p_2$, our proposed coding scheme works similarly.

In order to approach channel capacity, we need to optimize the mutual information of the channel input and output, among the possible distributions for the transmitted symbols. We now demonstrate that the uniform distribution is near optimal. For this purpose, we consider the error probabilities as in (1). Denote by $Cap$ the channel capacity, and $X,Y$ the random variables of the channel input and output. Due to the symmetry of $A,G$ and $C,T$, we set $q_1 = q_3, q_2 = q_4$ as the probabilities for the transmitted symbols $A,G,C,T$, respectively. The capacity

$$Cap = \max_{q_1,q_2} I(X;Y)$$

$$= \max_{q_1,q_2} H(Y) - H(Y|X)$$

$$= -2q_1(1 - p_1) + 2q_3 p_1 \log(q_1 (1 - p_1) + 2p_1 q_3)$$

$$- 2(q_1 p_1 + q_2 (1 - p_1)) \log(2q_1 p_1 + 2q_2 (1 - p_1))$$

$$+ 2q_1 ((1 - p_1 - p_3) \log(1 - p_1 - p_3) + 2p_1 \log(p_1) + 2p_3 \log(p_3))$$

$$+ 2q_3 ((1 - p_1 - p_2) \log(1 - p_1 - p_2) + p_1 \log(p_1) + p_2 \log(p_2)).$$

is maximized when the percentage of $A$ is from 25 to 31, for $0 < \lambda \leq 0.04, p_3 = 0$. We can see this probability is close to $1/4$. Accordingly, all other 3 symbols also have a probability close to $1/4$. The mutual information under uniform distribution is very close to the capacity for the best distribution. Specifically, using Equation (2), when $\lambda = 0.035$, the mutual information of uniform distribution is 1.32 bits and the capacity is 1.34 bits. Thus one can use the uniform distribution for the transmitted symbols, hence a linear block code with uniform information symbols is suitable for such channels.

We map the four types of nucleotides to two bits. When a quaternary code is used, these two bits correspond to the vector representation of elements in the finite field $GF(4)$; when a binary code is used, they are viewed as a 2-bit vector in $\{0,1\}^2$. In this paper, the mapping is chosen to be $A \mapsto (0,0), G \mapsto (1,1), T \mapsto (1,0), C \mapsto (0,1)$. According to the channel model in Fig. 2, another (non-equivalent) mapping can be $A \mapsto (0,1), G \mapsto (1,1), T \mapsto (0,0), C \mapsto (1,0)$. When the symbols are viewed as $2$ bits, the choice of the mapping affects the raw bit error rate (BER). Under uniform distribution, the former mapping gives the raw BER

$$BER = \frac{1}{4}(p_1 + 2p_2 + p_1) + \frac{1}{4}(p_1 + 2p_2 + p_1)$$

$$+ \frac{1}{4}(p_1 + 2p_3 + p_1) + \frac{1}{4}(p_1 + 2p_3 + p_1)$$

$$= 2p_1 + p_2 + p_3. \quad (8)$$

For the latter mapping, we have a raw BER

$$BER = 4p_1 + \frac{1}{2}p_2 + \frac{1}{2}p_3. \quad (9)$$

Since $p_3$ is small, we ignore the last term in both BERs. In this paper, we only consider the former mapping, and assume that the condition $p_2 < 4p_1$ is satisfied.

Next, we briefly review the LDPC decoding sum-product algorithm (SPA) (see, e.g., [13]). For simplicity, we describe the algorithm for the binary case. Consider a binary LDPC code defined by an $(n-k) \times n$ parity-check matrix $H = (h_{i,j})$. The code can be represented by a Tanner graph, with $n$ variable nodes (VNs) and $n-k$ check nodes (CNs). An edge exists between the $j$-th variable $VN_j$ and the $i$-th check $CN_i$ if and only if $h_{i,j} = 1$. Let $N(i)$ denote the set of neighboring nodes of the node $i$. When the Tanner graph has no cycles, SPA can be proved to be the bit-wise maximum a posteriori (MAP) decoder [14].

Let $x^i = (x_1^i, x_2^i, \ldots, x_n^i), y^i = (y_1^i, y_2^i, \ldots, y_n^i)$ be the transmitted and received binary words, respectively. In SPA, the log-likelihood ratio (LLR), $\log \frac{p(x^i=0|y^i)}{p(x^i=1|y^i)}$, is passed between neighboring variable nodes and check nodes in iterations. The LLR passed from node $i$ to node $j$ in the algorithm, called extrinsic information, is the aggregated information of the incoming messages to node $i$ from its neighbors except $j$, i.e., $N(i) - \{j\}$.

Let $iter$ be the index of the current iteration. Define the information passed from the variable node $VN_j$ to the check node $CN_i$ by $L_{j\rightarrow i}^{iter}$, where $1 \leq j \leq n$ and $i \in N(i)$. Define the information passed from the check node $VN_i$ to the variable node $CN_j$ by $L_{i\rightarrow j}^{iter}$, where $1 \leq i \leq n-k$ and $j \in N(i)$. In the $0$-th iteration, $L_{0\rightarrow i}^{iter}$ is initialized by $L_{0\rightarrow i}^{iter} = Ch_j$, where for the given received $y_j$, $Ch_j$ is the channel information defined by

$$Ch_j = \log \frac{p(y_1^j | x^i_1 = 0)}{p(y_1^j | x^i_1 = 1)}. \quad (10)$$

The message from a CN to a VN is given by

$$L_{i\rightarrow j}^{iter} = 2 \tanh^{-1} \left( \prod_{j' \in N(i)-(j)} \tanh(\frac{1}{2} L_{j'\rightarrow i}^{iter-1}) \right). \quad (11)$$

The message from a VN to a CN is updated by

$$L_{j\rightarrow i}^{iter} = Ch_j + \sum_{i' \in N(j)-(i)} L_{i'\rightarrow j}^{iter-1}. \quad (12)$$

Fig. 2. Error probabilities in the nanopore channel, $p_3 \ll p_1 < p_2$. This figure illustrates the error probabilities in the nanopore channel, $p_3 \ll p_1 < p_2$. The figure shows the error probabilities for different combinations of error probabilities for the transmitted symbols $A,G,C,T$. The error probabilities are given by $p_1, p_2, p_3$, and $p_4$, with $p_3 \ll p_1 < p_2$. The figure highlights the differences in error probabilities between the symmetric and asymmetric cases. In the symmetric case, $p_1 = p_2 = p_3 = p_4$, while in the asymmetric case, $p_1 < p_2$. The figure is useful for understanding the performance of the channel under different error probability scenarios.
Then the overall soft information is obtained for each $VN_j$:

$$L_{j}^{soft,iter} = Ch_j + \sum_{i \in N(j)} L_{i}^{iter-1}$$  \hspace{1cm} (15)$$

and the corresponding symbols is decoded as $\check{x}_j^{1,iter} = 1$ if $L_{j}^{soft,iter} < 0$, and decoded as 0 otherwise. The algorithm stops if $\check{x}_j^{1,iter} \cdot H^T = 0$ or the maximum number of iterations is reached.

III. BINARY LDPC CODES FOR DNA STORAGE

In this section, we present the binary LDPC codes for DNA storage. We propose our decoder that is inspired by Turbo decoder.

The construction of the binary codes is straightforward: we take $k$ bits of information, and encode them into a binary LDPC codeword of length $n$. Repeat for a second binary LDPC code. The two LDPC codes are denoted by $C^1, C^2$, and have the same rate $k/n$. The two codewords are denoted by $x^1 = (x^1_1, x^1_2, \ldots , x^1_n)$, for $i = 1, 2$, respectively. Then we store $x_j = (x^1_j, x^2_j)$ as one DNA quaternary symbol, for all $1 \leq j \leq n$. After the nanopore channel, we receive the corrupted $j$-th DNA symbol, $y_j = (y^1_j, y^2_j)$, $1 \leq j \leq n$, and then form the two binary received words $y^1 = (y^1_1, y^1_2, \ldots , y^1_n)$, $i = 1, 2$.

Below, we use notations similar to the sum-product algorithm described in Section II, except that we add a subscript 1 or 2 to denote the corresponding notation for the first or the second LDPC code. For example, denote by $VN^1_j, VN^2_j$ the $j$-th variable node in the first and the second LDPC codes, respectively; denote by $L_{j,i}^{1,iter}, L_{j,i}^{2,iter}$ the information passed from note $j$ to node $i$ in the two codes.

The baseline decoder. We propose a baseline decoder that runs two SPAs on the two codewords independently, with some modifications to the channel information $Ch^1_j, Ch^2_j$. Conditioned on the received quaternary $y$, the bit-wise MAP rule of the first LDPC is:

$$\check{x}^1_i = \arg \max_{x^1_i \in \{0,1\}} p(x^1_i | y) \hspace{1cm} (16)$$

$$= \arg \max_{x^1_i \in \{0,1\}} \sum_{~x^1_i} p(x^1_i | y) \hspace{1cm} (17)$$

$$= \arg \max_{x^1_i \in \{0,1\}} \sum_{~x^1_i} p(y|x^1_i)P(x^1_i) \hspace{1cm} (18)$$

$$= \arg \max_{x^1_i \in \{0,1\}} \sum_{~x^1_i} \prod_{j=1}^{n} \frac{p(y_j | x^1_j)}{\mathbb{1}_{x^1_i \in C^1}} \hspace{1cm} (19)$$

where $\sim x^1_i$ means all possible binary vectors $x^1 \in \{0,1\}^n$ such that $x^1_i$ is fixed, and $\mathbb{1}_{x^1_i \in C^1}$ is the indicator function that $x^1_i$ is a codeword of the first LDPC code $C^1$. Here (18) follows from Bayes’ rule and $y$ being fixed, and (19) follows from the memoryless channel and the uniformity of the transmitted codewords. In a conventional binary LDPC, the bit-wise MAP rule has the term $\prod_{j} p(y_j | x^1_j)$ instead of $\prod_{j} p(y_j | x^1_j)$ in (19), corresponding to the channel information. Effectively, the new channel has the binary input $x^1_i$ and the quaternary output $y_i = (y^1_i, y^2_i)$, and (12) is modified to

$$Ch^1_j = \log \frac{p(y_j | x^1_j = 0)}{p(y_j | x^1_j = 1)} \hspace{1cm} (20)$$

When $y_j = C$,

$$Ch^1_j = \log \left( \frac{p(y_j = C | x^1_j = C) + p(y_j = C | x^1_j = A)}{p(y_j = C | x^1_j = G) + p(y_j = C | x^1_j = T)} \right) \hspace{1cm} (21)$$

$$= \log \frac{1 - p_1 - p_2}{p_1 + p_2} \hspace{1cm} (22)$$

Similarly, when $y_j = A$,

$$Ch^1_j = \log \left( \frac{p(y_j = A | x^1_j = A) + p(y_j = A | x^1_j = C)}{p(y_j = A | x^1_j = G) + p(y_j = A | x^1_j = T)} \right) \hspace{1cm} (23)$$

$$= \log \frac{1 - p_1 - p_3}{p_1 + p_3} \hspace{1cm} (24)$$

When $y_j = T$,

$$Ch^1_j = \log \left( \frac{p(y_j = T | x^1_j = C) + p(y_j = T | x^1_j = A)}{p(y_j = T | x^1_j = G) + p(y_j = T | x^1_j = T)} \right) \hspace{1cm} (25)$$

$$= \log \frac{p_1 + p_2}{1 - p_1 - p_2} \hspace{1cm} (26)$$

When $y_j = G$,

$$Ch^1_j = \log \left( \frac{p(y_j = G | x^1_j = C) + p(y_j = G | x^1_j = A)}{p(y_j = G | x^1_j = G) + p(y_j = G | x^1_j = T)} \right) \hspace{1cm} (27)$$

$$= \log \frac{p_1 + p_3}{1 - p_1 - p_3} \hspace{1cm} (28)$$

The same analysis can be applied to the second LDPC. Much like conventional LDPC, for cycle-free Tanner graphs, the baseline decoder is a bit-wise MAP decoder, conditioned on $y$, for each individual LDPC.

However, this baseline decoder can be improved upon. Observe that in Fig. 2, if the received symbol is $y^1_j = C = (0, 1)$, and the first bit is decoded to be 1, then it is much more likely that the transmitted symbol is $x^1_j = T = (1, 0)$ than the case of $x^1_j = G = (1, 1)$. One can verify that given $y^1_j = C$ or $T$, the probability of having complementary bits ($x^1_j = 1 - x^2_j$) is much larger than identical bits ($x^1_j = x^2_j$). We note that Turbo codes (e.g., [13]) have a similar property. In particular, the Turbo codes can be viewed as two separate convolutional codes, sharing identical but permuted information bits. The two identical information bits pass extrinsic information between each other hence connecting the two convolutional codes. Inspired by Turbo codes, we propose to pass auxiliary information between the two bits when receiving $C$ or $T$.

In our proposed decoders, the two LDPC codes run their sum-product algorithms as in Section II concurrently, with the modifications as below. If a received symbol $y^1_j$ is $C \text{ or } T$, then an auxiliary term is passed from $VN^1_j$ to $VN^2_j$ in each iteration. We note that different from Turbo codes, the auxiliary term does not appear as an addend in the overall soft information of (15) due to two reasons: (i) the two bits $y^1_j, y^2_j$ are not deterministically complementary of each other, (ii) the function in (13) is nonlinear. Therefore, we cannot eliminate
the auxiliary term entirely when information is passed back from $VN_j^2 \to VN_j^1$, and the algorithm for the auxiliary information exchange is not straightforward. Regardless, we call the auxiliary term *extrinsic information* to highlight the resemblance to Turbo codes. Next, we propose three decoding algorithms.

**Algorithm 1.** Our first decoding algorithm is shown in Algorithm 1. In this algorithm, we pass the extrinsic information from $VN_j^1 \to VN_j^2$ if the received quaternary symbol $y_j = C$ or $T$, and the extrinsic information of the $j$-th variable node is defined as

$$Le_j^{1 \to 2} = \sum_{i' \in N(j)} L^{1, \text{iter}-1}_{i' \to j}.$$  \hspace{1cm} (29)

In every iteration, the extrinsic information is passed from the first LDPC to the second LDPC. To generate the message from a variable node $VN_j^2$ to a check node $CN_j^2$ in the second LDPC code, we combine the above extrinsic information, the channel information of (20), and the messages from $VN_j^2$, $i' \in N^2(j)$. Specifically, equation (14) is modified to

$$L_{j \to i'}^{2, \text{iter}} = Ch_j^2 + \sum_{i' \in N^2(j) \setminus \{j\}} L_{i' \to j}^{2, \text{iter}-1} - \alpha Le_j^{1 \to 2},$$  \hspace{1cm} (30)

for some constant parameter $0 < \alpha < 1$. Moreover, to generate the overall soft information of $VN_j^2$, (15) is modified to

$$L_{j}^{2, \text{soft,iter}} = Ch_j^2 + \sum_{i' \in N^2(j)} L_{i' \to j}^{2, \text{iter}-1} - \alpha Le_j^{1 \to 2}.$$  \hspace{1cm} (31)

Equivalent to (30) (31), define the new channel information as

$$\overline{Ch}_j^{2, \text{iter}} = Ch_j^2 - \alpha Le_j^{1 \to 2},$$  \hspace{1cm} (32)

and simply replace $Ch_j^2$ with $\overline{Ch}_j^{2, \text{iter}}$ in the sum product algorithm of Section II. The extrinsic information from $VN_j^2$ to $VN_j^1$ is defined similarly, and the algorithm of the first LDPC uses the new channel information similar to (32).

In expression (29) of the extrinsic information, the coefficient $-\alpha$ captures the fact that $x_j^1$ and $x_j^2$ are complementary with a large probability. Note that $Le_j^{1 \to 2}$ is the overall soft information at $VN_j^2$ as in (15) except the channel information term. The channel information is excluded because it is unchanged during the iterations, and would be amplified over time since $\alpha < 1$.

**Algorithm 2.** In the second algorithm, we derive an alternative channel information expression. In the following, we assume $y_j = C$, but the derivation for the case $y_j = T$ is similar and hence omitted.

We define $Le_j^{1 \to 2}$ the same as (29), and we treat it as the LLR of $x_j^1$ given $y_j = C$:

$$\log \left( \frac{p(x_j^1 = 0 | y_j = C)}{p(x_j^1 = 1 | y_j = C)} \right) = Le_j^{1 \to 2}.$$  \hspace{1cm} (33)

Since $x_j^1 \in \{1, 0\}$, the corresponding probability can be computed from LLR:

$$p(x_j^1 | y_j = C) = \frac{e^{-Le_j^{1 \to 2}}}{1 + e^{-Le_j^{1 \to 2}}} e^{x_j^1 Le_j^{1 \to 2}}.$$  \hspace{1cm} (34)

### Algorithm 1: Binary LDPC decoder for DNA storage

**Initialize:**

1. For $1 \leq j \leq n$, $i \in N(j)$ initialize $L_{j \to i}^{1, 0} = Ch_j^1$ and $L_{i \to j}^{2, 0} = Ch_j^2$ by (20)

2. for iter = 1 : $I_{\text{max}}$ do

3. Stop if $x_{1, \text{iter}}(H^1)^T = 0, x_{2, \text{iter}}(H^2)^T = 0$ or iter = $I_{\text{max}}$

**Message passing between two codes**

4. for $1 \leq j \leq n$ do

5. if $y_j = C = (0, 1)$ or $y_j = T = (1, 0)$ then

6. $Ch_{1, \text{iter}}^j = Ch_j^1 - \alpha \sum_{i' \in N(j)} L_{i' \to j}^{2, \text{iter}-1}$

7. $Ch_{2, \text{iter}}^j = Ch_j^2 - \alpha \sum_{i' \in N(j)} L_{i' \to j}^{1, \text{iter}-1}$

8. else

9. $Ch_{1, \text{iter}}^j = Ch_j^1$

10. $Ch_{2, \text{iter}}^j = Ch_j^2$

11. end if

12. end for

**Check node update**

13. for $1 \leq i \leq n - k$, $j \in N(i)$ do

14. $L_{i \to j}^{1, \text{iter}} = 2 \log(\prod_{i' \in N(i) \setminus \{j\}} \tanh(\frac{1}{2} L_{i' \to i}^{1, \text{iter}}))$

15. $L_{i \to j}^{2, \text{iter}} = 2 \log(\prod_{i' \in N(i) \setminus \{j\}} \tanh(\frac{1}{2} L_{i' \to i}^{2, \text{iter}}))$

16. end for

**Variable node update**

17. for $1 \leq j \leq n$, $i \in N(j)$ do

18. $L_{j \to i}^{1, \text{iter}} = Ch_{i, \text{iter}}^1 + \sum_{i' \in N(j) \setminus \{i\}} L_{i' \to j}^{1, \text{iter}}$

19. $L_{j \to i}^{2, \text{iter}} = Ch_{i, \text{iter}}^2 + \sum_{i' \in N(j) \setminus \{i\}} L_{i' \to j}^{2, \text{iter}}$

20. end for

21. for $i = 1 : n$ do

22. $L_{i}^{1, \text{soft,iter}} = Ch_{i, \text{iter}}^1 + \sum_{i' \in N(j)} L_{i' \to i}^{1, \text{iter}}$

23. $\hat{x}_i^1 = \begin{cases} 1, & \text{if } L_{i}^{1, \text{soft,iter}} < 0, \\ 0, & \text{else.} \end{cases}$

24. Compute $\hat{x}_i^2$ similarly

25. end for

26. end for

Assume $y_j = C$, from (20), a new channel information from the extrinsic information $Le_j^{1 \to 2}$ in iteration iter is set to be

$$Ch_j^{1 \to 2, \text{iter}} = \log \left( \frac{p(y_j = C | x_j^2 = 0)}{p(y_j = C | x_j^2 = 1)} \right) + \log \left( \frac{p(x_j^2 = 0 | y_j = C)}{p(x_j^2 = 1 | y_j = C)} \right),$$  \hspace{1cm} (35)

which follows from Bayes’ rule and the uniformity of $x_j^2$. Moreover,

$$p(x_j^2 = 0 | y_j = C) = \sum_{\gamma \in \{0, 1\}} p(x_j^2 = 0, x_j^1 = \gamma, y_j = C) p(x_j^1 = \gamma | y_j = C) \hspace{1cm} (36)$$

When $\gamma = 0$,

$$p(x_j^2 = 0 | x_j^1 = 0, y_j = C) \hspace{1cm} (37)$$

$$= \sum_{\gamma \in \{0, 1\}} p(x_j^2 = 0 | x_j^1 = \gamma, y_j = C) p(x_j^1 = \gamma | y_j = C) \hspace{1cm} (38)$$

$$= \sum_{\gamma \in \{0, 1\}} p(x_j^2 = 0 | x_j^1 = \gamma, y_j = C) p(x_j^1 = \gamma | y_j = C) \hspace{1cm} (39)$$

$$p(x_j^2 = 0 | x_j^1 = 0, y_j = C) \hspace{1cm} (40)$$
where (44) follows from the uniformity of the transmitted symbol $x_j$ and the channel transition in Figure 2. Similarly, we can calculate for the case $\gamma = 1$,

$$p(x_j^2 = 0|x_j^1 = 1, y_j = C) = \frac{p_2}{p_1 + p_2},$$

and thus obtain

$$p(x_j^2 = 0|y_j = C) = \frac{p_1}{1 - p_1 - p_2} p(x_j^1 = 0|y_j = C)$$

Similarly, we have

$$p(x_j^2 = 1|y_j = C) = \frac{1 - 2p_1 - p_2}{1 - p_1 - p_2} p(x_j^1 = 0|y_j = C)$$

In the above calculation, the probability $p(x_j^1|y_j = C)$ should be computed based on the extrinsic information from the first LDPC as in (34). Similarly, we have

$$p(x_j^1 = 1|y_j = C) = \frac{p_1}{p_1 + p_2} p(x_j^2 = 0|y_j = C)$$

We use the new channel information $C_{H_j^{1,2}} = C_{H_j^{2,1}} + \rho_{j} C_{H_j^{1,2}}$ by equations (20) (36) (48) (49) (34) in our SPA. As a special case, if $p_1 = 0$, then we have $C_{H_j^{2,1}} = C_{H_j^{1,2}} + \rho_{j}$.

We can also consider when the received word is C and the channel information of the first LDPC should be computed based on:

$$p(x_j^1 = 0|y_j = C) = \frac{p_1}{p_1 + p_2} p(x_j^2 = 0|y_j = C)$$

When $y_j = T$,

$$p(x_j^2 = 0|y_j = T) = \frac{p_1}{p_1 + p_2} p(x_j^1 = 0|y_j = T)$$

$$p(x_j^2 = 1|y_j = T) = \frac{1 - 2p_1 - p_2}{1 - p_1 - p_2} p(x_j^1 = 0|y_j = T)$$

Finally, we pass the check message from $V_1$, $V_2$, and $V_3$ to the check node $e_j$.

Algorithm 3. In the third decoding algorithm, we add an imaginary variable node $u_j$ and an imaginary check node $e_j$ in the Tanner graph, for every $j$ such that $y_j = C$ or $T$. The imaginary nodes serve the purpose of extrinsic information exchange. As Fig. 3 shows, after iteration $iter - 1$, we pass the variable message from $V_{N_1}^{j}$ and $u_j$ to the check node $e_j$. Then, we pass the check message from $e_j$ to $V_{N_1}^{j}$.

$$p(x_j^2 = 0|y_j = T) = \frac{p_1}{1 - p_1 - p_2} p(x_j^1 = 0|y_j = T)$$

$$p(x_j^2 = 1|y_j = T) = \frac{1 - 2p_1 - p_2}{p_1 + p_2} p(x_j^1 = 1|y_j = T).$$

where $u_j$ is set to have a fixed LLR of

$$L_{u_j} = \log \frac{p(u_j = 1|y_j = C)}{p(u_j = 1|y_j = C)}$$

Here (67) is due to (61) and the uniformity of $x_j$.

IV. SIMULATION

We construct LDPC with the degree distribution pair $p(X) = \sum_{d=1}^{d_n} \rho_d X^{d-1}$ and $\lambda(X) = \sum_{d=1}^{d_n} \lambda_d X^{d-1}$, where $\rho_d$ denotes the fraction of all edges connected to degree-$d$ check nodes, $d_c$ is the maximum check node degree, and $\lambda_d$ and $d'_d$ are defined similarly for variable nodes. A good distribution pair is found using density evolution [13], according to which a random parity-check matrix is picked. In this paper, the distribution is $p(X) = 0.814X^6 + 0.186X^7$, $\lambda(X) = 0.931X^2 + 0.045X^3 + 0.0234X^4$ which is in fact very close to the commonly used (3.6) distribution. Namely, every variable node has degree 6, and every check node has degree 3. For the quaternary code, we first choose the same degree distribution as the binary code.
Then in the parity check matrix the ones are randomly changed to one of the three non-zero elements of GF(4).

**TABLE I**

<table>
<thead>
<tr>
<th></th>
<th>quaternary</th>
<th>method 1</th>
<th>method 2</th>
<th>method 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>one iteration</td>
<td>0.8589s</td>
<td>0.038s</td>
<td>0.06s</td>
<td>0.05s</td>
</tr>
<tr>
<td>one codeword</td>
<td>8.53s</td>
<td>1.08s</td>
<td>1.05s</td>
<td>1.05s</td>
</tr>
</tbody>
</table>

The bit error rate of quaternary codes, the 3 proposed decoding algorithms, and the baseline binary decoder is shown in Fig. 4. The code has $n = 1000, k = 500$. For Algorithm 1, we pick the best parameter $\alpha = 0.1$. We can see that when $\lambda$ is small, the error rate of Algorithms 2 and 3 are better than the baseline binary decoder and the quaternary codes, Algorithm 1 has a comparable error rate to the quaternary codes. When $\lambda$ is large, quaternary codes work better. Algorithm 3 performs best because it has the most interactions between the two binary LDPC codes.

Moreover, the simulation time of the algorithms are shown in Table I. Our binary codes are 14 times faster on each iteration and 8 times faster on each codeword than quaternary codes. Given the demand of high-volume data storage, the speed-up can improve system performance significantly.

**V. CONCLUSION**

In conclusion, we present binary LDPC codes for DNA storage to combat asymmetric sequencing errors. Our decoding algorithms utilize the extrinsic information exchange to accommodate the dependence of the errors of the two bits in a nucleotide symbol. We demonstrate the desirable performance of our codes in terms of bit error rate and decoding speed. One possible future work is to investigate error correction codes with soft information from sequencers.

**REFERENCES**