JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 0, 2021
© Mary Ann Liebert, Inc.
Pp. 1–14
DOI: 10.1089/cmb.2020.0226

Optimized Fluorescence-Based Detection in Single Molecule Synthesis Process

HSIN-HAO CHEN and CHUNG-CHIN LU

ABSTRACT

Single molecule sequencing is imperative to overall genetic analysis in areas such as genomics, transcriptomics, clinical test, drug development, and cancer screening. In addition, fluorescence-based sequencing is primarily applied in single molecule sequencing besides other methods, precisely in the fields of DNA sequencing. Modern-day fluorescence labeling methods exploit a charge-coupled device camera to capture snapshots of a number of pixels on the single molecule sequencing. The method discussed in this article involves fluorescence labeling detection with a single pixel, outrivals in high accuracy and low resource requirement under low signal-to-noise ratio conditions, as well as benefits from higher throughput comparing with others. Through discussion in this article, we explore the single molecule synthesis process modeling using negative binomial distributions. Furthermore, incorporating the method of maximum likelihood and Viterbi algorithm in this modeling enhances the signal detection accuracy. The fluorescence-based model benefits in simulating actual experiment processes and assisting in understanding relations between the fluorescence emission and the signal receiving events. Last but not least, the model offers potential candidates on fluorescence dye selection that yields more accurate experiment results.

Keywords: fluorescence dye selection, fluorescence labeling, genome sequencing, maximum likelihood, single molecule synthesis process.

1. INTRODUCTION

The field of biological science, thanks to collective advancement in technology, has recently been in exponential growth specifically in nucleotide sequencing, resulting in various analytical strategies and computing methods (Deamer et al., 2016, Shendure et al., 2017). Benefiting from newly available nucleic acid detection methods, genome sequencing detection has been greatly improved both on its performance and accuracy. Among them, the fluorescence signaling method has become one of the most efficient tools in its league (Schmitt et al., 2012). Its advantages include but not limited to an increased level of sensitivity, multiplexing capabilities, and simultaneous detection on fluorescence properties (Epstein et al., 2002).

Many agent supplementing methods are employed to distinguish organic molecules such as proteins and nucleic acids and to augment detection sensitivity in the controlled assay. Fluorescence labeling method stands out in its ranks, whereas it offers detection with light signal exposure when target and agent are

Department of Electrical Engineering, National Tsing Hua University, Hsinchu City, Taiwan.

stimulated during sequencing. In addition, each fluorescence agent is signified with its own unique fluorescence property or light wavelength signature. As a result, the detection devices are able to conduct simultaneous observations on more than one molecule, saving time resources and cutting down assay duration (Seo et al., 2005; Hermanson, 2013; Dean and Palmer, 2014).

A target molecule undergoing a fluorescent event is excited and emits light at a different wavelength than it was previously exposed to. The platform embedded with the charge-coupled device (CCD) camera receives the images of the process of the emission and detects the sequence from the data (Dean and Palmer, 2014; Valm et al., 2016). However, it requires an optimized signal extraction to increase signal-to-noise ratio (SNR), which is the main restricting factor for data collection and integrity as from the conventional fluorescence detection methods (Horne, 1986; Pollard et al., 2018). As we apply the methods to the detection of the single molecule synthesis, the disadvantages include higher error rate, higher cost per base, and lower throughput (Ardui et al., 2018). Although the consensus sequencing method is utilized to compensate for the high error rate (Hiatt et al., 2010), it requires higher computing power, a more complicated template, and more time resulting from necessary repetitive iterations.

The period of the synthesis and the emission intensity of the fluorescence dominate the performance and the accuracy of the experiment. To improve the efficiency of the decoding method and the cost per base in sequencing, we construct a statistical model to simulate polymerase synthesis and the fluorescent emission process and record the data with a single three-junction photodiode (Jansen-van Vuuren et al., 2016; Hsieh et al., 2019). The algorithm in this study utilizes the method of maximum likelihood (ML) and Viterbi algorithm to detect the signals and decode the sequences in lower SNR conditions (Viterbi, 1967). The identifiability of fluorescent events varies depending on wavelength given out from fluorescent dyes. Under certain scenarios, the selection of fluorescent dyes is crucial to detection accuracy, yet their efficiency cannot be measured until the sequencing is completed, resulting in time and resource consumption. Models hereunder help select fluorescent substances and show how different parameters affect performance. A part of this study has been published previously in Chen and Lu (2019).

2. MATERIALS AND MODELS

2.1. *The single molecule synthesis process*

The light emission produced by polymerase synthesis reaction plays an important role in a fluorescence-based single molecule sequencing as in Figure 1 (Eid et al., 2009). We take the long-read sequencing platform as an instance to exhibit the single molecule synthesis process (SMSP) since the DNA polymerase synthesizes the nucleotides in a sequence. SMSP is modeled as a discrete-time stochastic process $\{Z(t), t \geq 1\}$ on a state space $\{A, T, G, C, d\}$, where state A (T, G, or C) indicates that a dATP (dTTP, dGTP, or dCTP) is being incorporated by the DNA polymerase and state d indicates that no dNTP is being incorporated.

Let S_n and $T_n - 1$ be the start time and the stop time of the nth incorporation of dNTP for $n \geq 1$. Then

$$0 = T_0 \leq S_1 < T_1 \leq S_2 < T_2 \leq \cdots \leq S_n < T_n \leq \cdots$$

Let $X_n \equiv Z(S_n)$ be the nth nucleotide to be incorporated. The length of the nth interpulse duration is $W_n = S_n - T_{n-1}$ and the length of the nth incorporation period or pulse width (hereunder indicated as pulse width) is $Y_n = T_n - S_n$. The process $\{(W_n, Y_n), n \geq 1\}$ is assumed to be an alternating process such that W_n and Y_n are statistically independent. Owing to the lack of structural information of a short piece of DNA, it can be assumed that $\{X_n, n \geq 1\}$ is an independent and identically distributed sequence of random variables with uniform distribution over $\{A, T, G, C\}$ and is independent of the alternating process $\{(W_n, Y_n), n \geq 1\}$. Assume r is the smallest time period during which the fluorescence emits and l is the smallest time period during which the polymerase translocates to the next active site. According to the previous study, the distribution of the pulse width and the interpulse duration do not distribute uniformly, and we will model the distributions of W_n and Y_n by negative binomial distributions with parameters (l, q) and (r, p), respectively:

$$P(W_n = k) = C_{l-1}^{k-1} q^l (1-q)^{k-l}, k = l, l+1, l+2, \ldots. \tag{1}$$

$$P(Y_n = k) = C_{r-1}^{k-1} p^r (1-p)^{k-r}, k = r, r+1, r+2, \ldots. \tag{2}$$

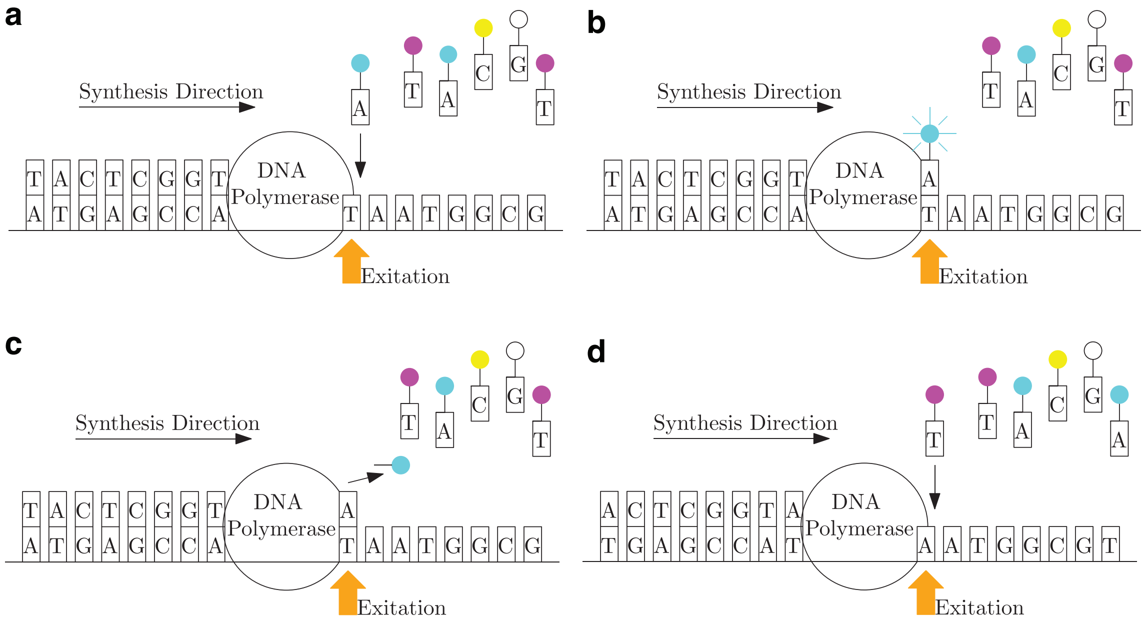

FIG. 1. The polymerase synthesis process. **(a)** Fluorescence-linked nucleotide is bound to the polymerase active site. **(b)** Fluorescence dye begins to illuminate when excited. **(c)** Fluorescence dye is disconnected from nucleotide, thus ceases to illuminate. **(d)** DNA polymerase translocates to the next active site.

With the aforementioned assumptions, the SMSP $\{(W_n, Y_n, X_n), n \geq 1\}$ becomes a discrete-time Markov chain $\{L(t), t \geq 1\}$ with the state space

$$\mathbf{S} = \{d_1, d_2, \ldots, d_l, A_1, A_2, \ldots, A_r, T_1, T_2, \ldots, T_r, G_1, G_2, \ldots, G_r, C_1, C_2, \ldots, C_r\}$$

of size $4r + l$. Let $S_i = \{A_i, T_i, G_i, C_i\}$. The state transition diagram of the Markov chain $\{L(t), t \geq 1\}$ is in Figure 2 and the state transition probabilities are

$$\pi_{d_i, s} = \begin{cases} 1 - q, & \text{if } s = d_i, \\ q, & \text{if } s = d_{i+1}, \text{ for } 1 \leq i \leq l - 1, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

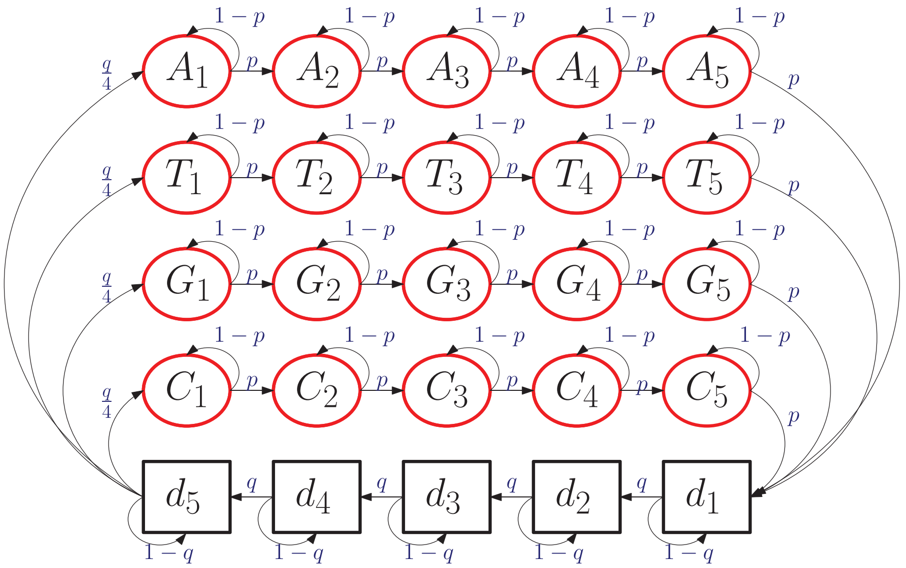
A_1 A_2 A_3 A_4 A_5
T_1 T_2 T_3 T_4 T_5
G_1 G_2 G_3 G_4 G_5
C_1 C_2 C_3 C_4 C_5
d_5 d_4 d_3 d_2 d_1
$1-p$
p
$\frac{q}{4}$
q
$1-q$

FIG. 2. The state transition diagram of the Markov chain $\{L(t), t \geq 1\}$. In the example shown, $l = 5$ and $r = 5$.

$$\pi_{d_l,\,s}=\begin{cases}1-q, & \text{if } s=d_l,\\ \frac{q}{4}, & \text{if } s\in S_1,\\ 0, & \text{otherwise,}\end{cases} \tag{4}$$

$$\pi_{A_i,\,s}=\begin{cases}1-p, & \text{if } s=A_i,\\ p, & \text{if } s=A_{i+1}, \text{for } 1\le i\le k-1,\\ 0, & \text{otherwise,}\end{cases} \tag{5}$$

$$\pi_{A_k,\,s}=\begin{cases}1-p, & \text{if } s=A_k,\\ p, & \text{if } s=d_1,\\ 0, & \text{otherwise.}\end{cases} \tag{6}$$

With the assumption in Equation (2), the transition probabilities, $\pi_{A_i,\,s}$, $\pi_{T_i,\,s}$, $\pi_{G_i,\,s}$, and $\pi_{C_i,\,s}$ are identical.

Next, we give an illustration of the resemblance of the characteristics of our SMSP model in Equations (1) and (2) to that of the previous study (Eid et al., 2009). We assume that the smallest pulse width and the smallest interpulse duration in the SMSP are both $\sim$100 ms with the time unit in the integrated circuit module being 25 ms. That means $l=5$ and $r=5$ in Equations (1) and (2) and the corresponding transition diagram is shown in Figure 2. Moreover, the largest pulse width and the largest interpulse duration are restricted to 500 ms and 5 s, respectively. As we limit the probability $p(Y_n\ge 21(500ms))\le 10^{-3}$ and $p(W_n\ge 201(5s))\le 10^{-2}$, which means the length of the pulse width and the interpulse duration are restricted, the probability mass functions are adjusted as in Figure 3a to $p=0.57$ and $q=0.05$. The distributions of the simulation results are similar to the previous study (Eid et al., 2009) in terms of pulse characteristics and trace statistics as shown in Figure 3.

2.2. *The emission process*

The fluorescence light emits as a nucleotide is incorporated by the DNA polymerase. We apply a method utilizing a three-junction photodiode to capture the image during the light emission (Hsieh et al., 2019). While no nucleotide is being incorporated, only ambient light is detected, and of course, ambient light also exists in an incorporation period. Consequently, the photodetector will output the fluorescence plus ambient light intensity signal during an incorporation period and the ambient light intensity signal only outside of the incorporation period.

Let $\{\boldsymbol{E}(t), t\ge 1\}$ be the output intensity signal of the photodetector associated with the SMSP $\{Z(t), t\ge 1\}$, called the emission process. Assume that the three-dimensional emission vector $\boldsymbol{E}(t)$ depends only on the state $Z(t)$ at time t. Note that $\{Z(t), t\ge 1\}$ represents the SMSP $\{(W_n, Y_n, X_n), n\ge 1\}$ and

$$Z(t)=\begin{cases} d & \text{, if } \sum_{i=1}^{n-1}(W_i+Y_i)<t\le\sum_{i=1}^{n-1}(W_i+Y_i)+W_n,\\ X_n & \text{, if } \sum_{i=1}^{n-1}(W_i+Y_i)+W_n<t\le\sum_{i=1}^{n}(W_i+Y_i).\end{cases}$$

2.2.1. The ambient light intensity signal. Under the condition that the ambient light source originated from the dNTPs or during the interpulse duration is in the steady state, the ambient light intensity signal can be modeled by a constant signal vector $\boldsymbol{a}_{Z(t)}$, which is undetermined and will be estimated for the photodetector in a pixel.

2.2.2. The fluorescence light intensity signal. The fluorescence light intensity signal during an incorporation period depends on the dye molecule bound with the nucleotide under synthesis as well as the distance between the dye molecule and the photodiode. Also, variance in the biochemical reactions affects the light intensity. The biochemical reactions take place in a closed space where fluorescence light can be fully captured by the photodetector. Such a closed space is called a synthesis well.

Assume that the DNA polymerase has a nominal fixed position in the synthesis well during an incorporation period so that the captured dNTP plus dye molecule by the enzyme has a nominal constant distance from the photodiode. Thus, if $Z(t)=x$ for $t\in[S_n, T_n-1]$, where $x\in\{A, T, G, C\}$, then the fluorescence light intensity signal will be $\Upsilon(t)\boldsymbol{s}_x, t\ge 1$, where $\boldsymbol{s}_x$ is the detected signal vector of the three-junction photodiode to the fluorescence light emitted from the dye molecule bound with a dxTP in a nominal distance from the photodiode.

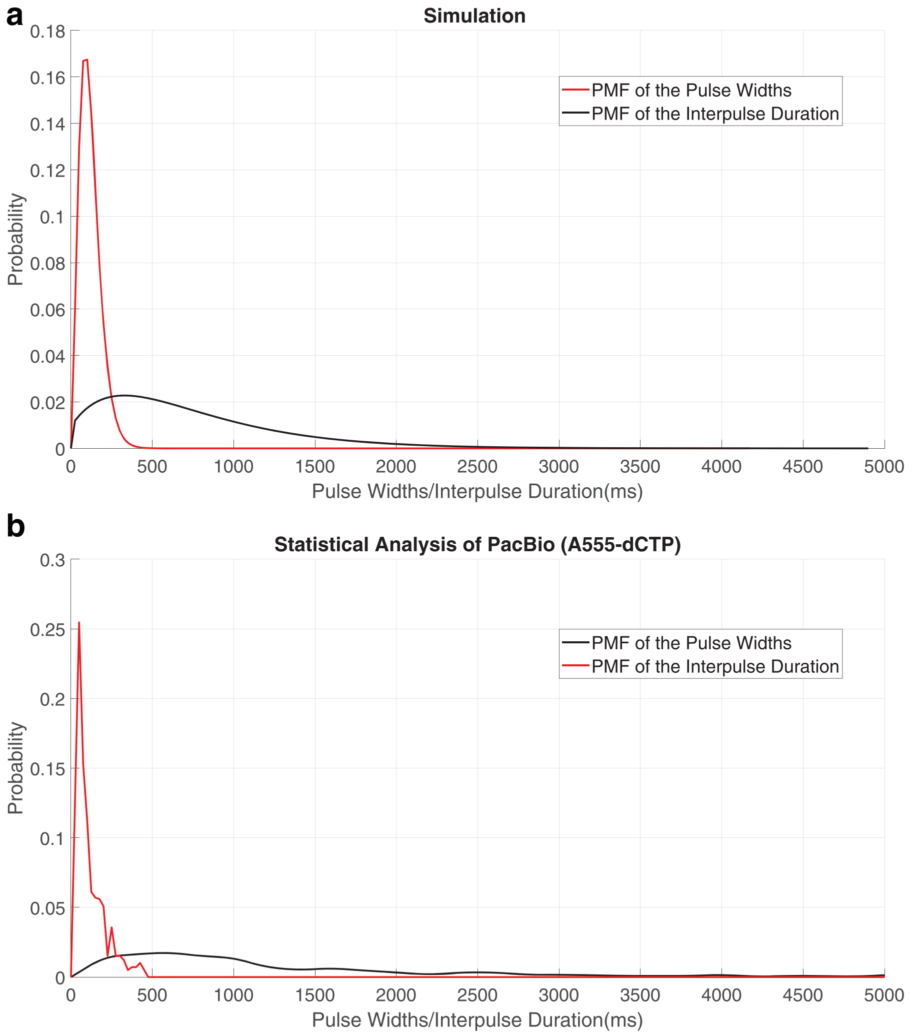

FIG. 3. **(a)** The red line represents the PMF of the pulse widths in Equation (2) with $p = 0.57$, and the black line represents the PMF of the interpulse duration in Equation (1) with $q = 0.05$. **(b)** The PMFs of the experiment in Eid et al. (2009). PMF, probability mass function.

$\Upsilon(t)$ is the fading coefficient due to the variation of the true distance from the dye molecule to the photodiode in the time t of the incorporation period relative to the nominal distance and $\Upsilon(t) = 0$ as $Z(t) = d$. Since constraints exist in the position of the DNA polymerase and the variation of the biochemical reactions, the emission intensity $\Upsilon(t)$ has the minimum and maximum values $(I_{A,\min}, I_{A,\max})$, $(I_{T,\min}, I_{T,\max})$, $(I_{G,\min}, I_{G,\max})$, and $(I_{C,\min}, I_{C,\max})$ when synthesizing ATP, TTP, GTP, and CTP, respectively. Each characteristic signal vector $\boldsymbol{s}_x$, $x \in \{A, T, G, C\}$, is an undetermined parameter vector and can be estimated. $\{\Upsilon(t), t \geq 1\}$ is the fluorescence intensity process in photon/ms.

2.2.3. The emission during an interpulse duration. The emission signal during the nth interpulse duration $[T_{n-1}, S_n - 1]$ is

$$\boldsymbol{E}(t) = \boldsymbol{a}_d, \;\; t \in [T_{n-1}, S_n - 1].$$

2.2.4. The emission during an incorporation period. The emission signal during the nth incorporation period $[S_n, T_n - 1]$ is

$$\boldsymbol{E}(t) = \Upsilon(t)\boldsymbol{s}_{Z(t)} + \boldsymbol{a}_{Z(t)}, \;\; t \in [S_n, T_n - 1].$$

2.3. The received process

The photodetector, as well as the readout circuit, will introduce noise to the emission process. From our noise measurement of the photodetector, the cross-correlations between three junctions are close to zero and, therefore, negligible to calculation. We model this noise as a white Gaussian vector process $\{\boldsymbol{N}(t), t \geq 1\}$ with covariance matrix

$$\Lambda = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix},$$

where σ_i^2 is the average noise power of the ith junction channel and will be estimated for each pixel. The joint probability density function of the noise vector $\boldsymbol{N}(t)$ at time t is

$$f_N(\boldsymbol{n}) = \frac{1}{\sqrt{(2\pi)^3|\Lambda|}} e^{-\frac{1}{2}\boldsymbol{n}'\Lambda^{-1}\boldsymbol{n}} = \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{n_i^2}{2\sigma_i^2}},$$

where $|\Lambda|$ is the determinant of the covariance matrix Λ. The noise process $\{\boldsymbol{N}(t), t \geq 1\}$ is assumed to be independent of the SMSP $\{Z(t), t \geq 1\}$ and the fading process $\{\Upsilon(t), t \geq 1\}$.

We assume that the noise process $\{\boldsymbol{N}(t), t \geq 1\}$ is additive to the emission process $\{\boldsymbol{E}(t), t \geq 1\}$ so that the received signal process $\{\boldsymbol{R}(t), t \geq 1\}$ is

$$\boldsymbol{R}(t) = \boldsymbol{E}(t) + \boldsymbol{N}(t), \quad t \geq 1 \tag{7}$$

$$= \begin{cases} \boldsymbol{a}_d + \boldsymbol{N}(t), & \text{if } Z(t) = d, \\ \Upsilon(t)\boldsymbol{s}_A + \boldsymbol{a}_A + \boldsymbol{N}(t), & \text{if } Z(t) = A, \\ \Upsilon(t)\boldsymbol{s}_T + \boldsymbol{a}_T + \boldsymbol{N}(t), & \text{if } Z(t) = T, \\ \Upsilon(t)\boldsymbol{s}_G + \boldsymbol{a}_G + \boldsymbol{N}(t), & \text{if } Z(t) = G, \\ \Upsilon(t)\boldsymbol{s}_C + \boldsymbol{a}_C + \boldsymbol{N}(t), & \text{if } Z(t) = C. \end{cases} \tag{8}$$

3. METHODS AND ALGORITHMS

In this section, the focus is positioned on the decoding algorithms. As a first phase, the pixel parameters are prepared by the default values or the values estimated from the experiment data. Next phase, the entire sequence is detected by the decoding algorithm and the pixel parameters.

1. **Initialization Phase:** The pixel parameters $\boldsymbol{s}_A$, $\boldsymbol{s}_T$, $\boldsymbol{s}_G$, $\boldsymbol{s}_C$, $\boldsymbol{a}_A$, $\boldsymbol{a}_T$, $\boldsymbol{a}_G$, $\boldsymbol{a}_C$, $\boldsymbol{a}_d$, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ are given by the default values or estimated from the experiment data.
2. **Decoding Phase:** $\{I(t), t \geq 1\}$ and $\{\alpha(t), t \geq 1\}$ of the alternating process $\{L(t), t \geq 1\}$ and the fluorescence intensity process $\{\Upsilon(t), t \geq 1\}$ are obtained by the known training nucleotide sequence $\{x_n, n \geq 1\}$ and the given estimated pixel parameters $\boldsymbol{s}_A$, $\boldsymbol{s}_T$, $\boldsymbol{s}_G$, $\boldsymbol{s}_C$, $\boldsymbol{a}_A$, $\boldsymbol{a}_T$, $\boldsymbol{a}_G$, $\boldsymbol{a}_C$, $\boldsymbol{a}_d$, and $\boldsymbol{\sigma}$.

3.1. Estimation of the pixel parameters

The pixel parameters $\boldsymbol{s}_A$, $\boldsymbol{s}_T$, $\boldsymbol{s}_G$, $\boldsymbol{s}_C$, $\boldsymbol{a}_A$, $\boldsymbol{a}_T$, $\boldsymbol{a}_G$, $\boldsymbol{a}_C$, $\boldsymbol{a}_d$, and $\boldsymbol{\sigma}$ can be estimated from the experimental measurement data. Since the four fluorescence dyes have the distinct wavelengths, the LEDs with the same wavelengths can be used to simulate the fluorescences and estimate the pixel parameters. While the LED light intensity increases as time progresses, the received values from three-junction photodiode are measured, in units of voltage, through photoelectric conversion (PC). Next, the pixel parameters, $\boldsymbol{s}_x$ and $\boldsymbol{a}_x \forall x \in \{A, T, G, C\}$, are estimated through a simple linear regression in Equation (8). $\boldsymbol{\sigma}$ is calculated by the measurement in the interpulse duration. Moreover, the R-squared (R^2) is calculated to determine if the simple linear regression is well fitted.

3.2. Decoding phase

Assume that the estimated pixel parameters $\boldsymbol{s}_A$, $\boldsymbol{s}_T$, $\boldsymbol{s}_G$, $\boldsymbol{s}_C$, $\boldsymbol{a}_A$, $\boldsymbol{a}_T$, $\boldsymbol{a}_G$, $\boldsymbol{a}_C$, $\boldsymbol{a}_d$, and $\boldsymbol{\sigma}$ are given. The likelihood function $f_{\boldsymbol{R}|Z,\Upsilon}(\boldsymbol{r}|z, \alpha)$ of the received signal $\{\boldsymbol{R}(t), t \geq 1\}$ to be $\{\boldsymbol{r}(t), t \geq 1\}$, given the SMSP $\{Z(t), t \geq 1\}$ to be $\{z(t), t \geq 1\}$ and the emission intensity process $\{\Upsilon(t), t \geq 1\}$ to be $\{\alpha(t), t \geq 1\}$, is

$$f_{\boldsymbol{R}|Z,\Upsilon}(\boldsymbol{r}|z,\alpha)=\prod_{t\geq 1} f_{\boldsymbol{R}(t)|Z(t),\Upsilon(t)}(\boldsymbol{r}(t)|z(t),\alpha(t))$$

since the noise process $\{\boldsymbol{N}(t), t\geq 1\}$ is a white Gaussian process, where

$$\begin{aligned} & f_{\boldsymbol{R}(t)|Z(t),\Upsilon(t)}(\boldsymbol{r}(t)|z(t),\alpha(t)) \\ &= \begin{cases} \prod_{j=1}^{3} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{r_j(t)-a_{d,j}}{\sigma_j}\right)^2}, & \text{if } z(t)=d, \\ \prod_{j=1}^{3} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{1}{2}\left(\frac{r_j(t)-a_{x,j}-\alpha(t)s_{x,j}}{\sigma_j}\right)^2}, & \text{if } z(t)=x\in\{A,T,G,C\} \end{cases} \end{aligned}$$

Note that $\alpha(t)=0$ as $z(t)=d$.

Now given the estimated pixel parameters $\boldsymbol{s}_A$, $\boldsymbol{s}_T$, $\boldsymbol{s}_G$, $\boldsymbol{s}_C$, $\boldsymbol{a}_A$, $\boldsymbol{a}_T$, $\boldsymbol{a}_G$, $\boldsymbol{a}_C$, $\boldsymbol{a}_d$, and $\boldsymbol{\sigma}$, the decoded versions $\{\hat{z}(t), t\geq 1\}$ and $\{\hat{\alpha}(t), t\geq 1\}$ of the SMSP $\{Z(t), t\geq 1\}$ and the emission intensity process $\{\Upsilon(t), t\geq 1\}$ can be obtained by the method of ML,

$$(\hat{z},\hat{\alpha})=\arg\max_{(z,\alpha)}\sum_{t\geq 1}\ln f_{\boldsymbol{R}(t)|Z(t),\Upsilon(t)}(\boldsymbol{r}(t)|z(t),\alpha(t)),$$

where $\ln f_{\boldsymbol{R}(t)|Z(t),\Upsilon(t)}(\boldsymbol{r}(t)|z(t),\alpha(t))$

$$=\begin{cases} \sum_{j=1}^{3}\ln\frac{1}{\sqrt{2\pi}\sigma_j}-\frac{1}{2}\sum_{j=1}^{3}\left(\frac{r_j(t)-a_{d,j}}{\sigma_j}\right)^2, & \text{if } z(t)=d, \\ \sum_{j=1}^{3}\ln\frac{1}{\sqrt{2\pi}\sigma_j}-\frac{1}{2}\sum_{j=1}^{3}\left(\frac{r_j(t)-a_{x,j}-\alpha(t)s_{x,j}}{\sigma_j}\right)^2, & \text{if } z(t)=x\in\{A,T,G,C\}. \end{cases}$$

Since the term $\sum_{j=1}^{3}\ln\frac{1}{\sqrt{2\pi}\sigma_j}$ is irrelevant to the maximization process, we will define a metric $m(z(t),\alpha(t))$ as follows:

$$m(z(t),\alpha(t))=\begin{cases} \sum_{j=1}^{3}\left(\frac{r_j(t)-a_{d,j}}{\sigma_j}\right)^2, & \text{if } z(t)=d, \\ \sum_{j=1}^{3}\left(\frac{r_j(t)-a_{x,j}-\alpha(t)s_{x,j}}{\sigma_j}\right)^2, & \text{if } z(t)=x\in\{A,T,G,C\}. \end{cases}$$

Then the ML decoded version of $\{Z(t), t\geq 1\}$ and $\{\Upsilon(t), t\geq 1\}$ is

$$\begin{aligned} (\hat{z},\hat{\alpha}) &= \arg\min_{(z,\alpha)}\sum_{t\geq 1} m(z(t),\alpha(t)) \\ &= \arg\min_{z}\sum_{t\geq 1}\min_{\alpha(t)} m(z(t),\alpha(t)). \end{aligned}$$

Subsequently, given a hypothetical SMSP $\{z(t),\ t\geq 1\}$, the minimization of the sum $\sum_{t\geq 1} m(z(t),\alpha(t))$ of metrics $m(z(t),\alpha(t))$ over the emission intensity process $\{\alpha(t), t\geq 1\}$ can be done by the minimization of the metrics $m(z(t),\alpha(t))$ over the intensity $\alpha(t)$ at each time t with $z(t)$ given.

When $z(t)\in\{A,T,G,C\}$, $\alpha^*(t|z(t))=\arg\min_{\alpha(t)} m(z(t),\alpha(t))$. Then we have

$$\alpha^*(t|z(t))=Q\left(I_{x,\,\min},\alpha^{\#}(t|z(t)),I_{x,\,\max}\right), \forall z(t)=x\in\{A,T,G,C\},$$

where for $a<c$,

$$Q(a,b,c)=\begin{cases} a, & \text{if } b<a, \\ b, & \text{if } a\leq b\leq c, \\ c, & \text{if } b>c, \end{cases}$$

and

$$\alpha^{\#}(t|z(t))=\frac{\sum_{j=1}^{3}\left(\frac{r_j(t)-a_{x,j}}{\sigma_j}\right)\left(s_{x,j}/\sigma_j\right)}{\sum_{j=1}^{3}\left(s_{x,j}/\sigma_j\right)^2}, \text{if } z(t)=x\in\{A,T,G,C\}$$

by doing the minimization of the metric $m(z(t), \alpha(t))$ over all $\alpha(t) \in \mathbb{R}$.

$$m^*(z(t)) = \begin{cases} \sum_{j=1}^{3} \left(\frac{r_j(t) - a_{d,j}}{\sigma_j}\right)^2, & \text{if } z(t) = d, \\ \sum_{j=1}^{3} \left(\left(\frac{r_j(t) - a_{x,j}}{\sigma_j}\right) - \alpha^*(t|z(t))\left(\frac{s_{x,j}}{\sigma_j}\right)\right)^2, & \text{if } z(t) = x \in \{A, T, G, C\}. \end{cases} \quad (9)$$

Now the ML decoded versions $\{z(t), t \geq 1\}$ and $\{\alpha(t), t \geq 1\}$ of the SMSP $\{Z(t), t \geq 1\}$ and the fluorescence intensity process $\{\Upsilon(t), t \geq 1\}$ are

$$(\hat{z}, \hat{\alpha}) = \arg \min_{(z,\, \alpha^{\#}(t|z(t)))} \sum_{t \geq 1} m^*(z(t)). \quad (10)$$

We apply the Viterbi algorithm (Viterbi, 1967) with the state transition diagram (Fig. 4a) and the related trellis diagram in Figure 4b to accomplish the minimization problem in Equation (10).

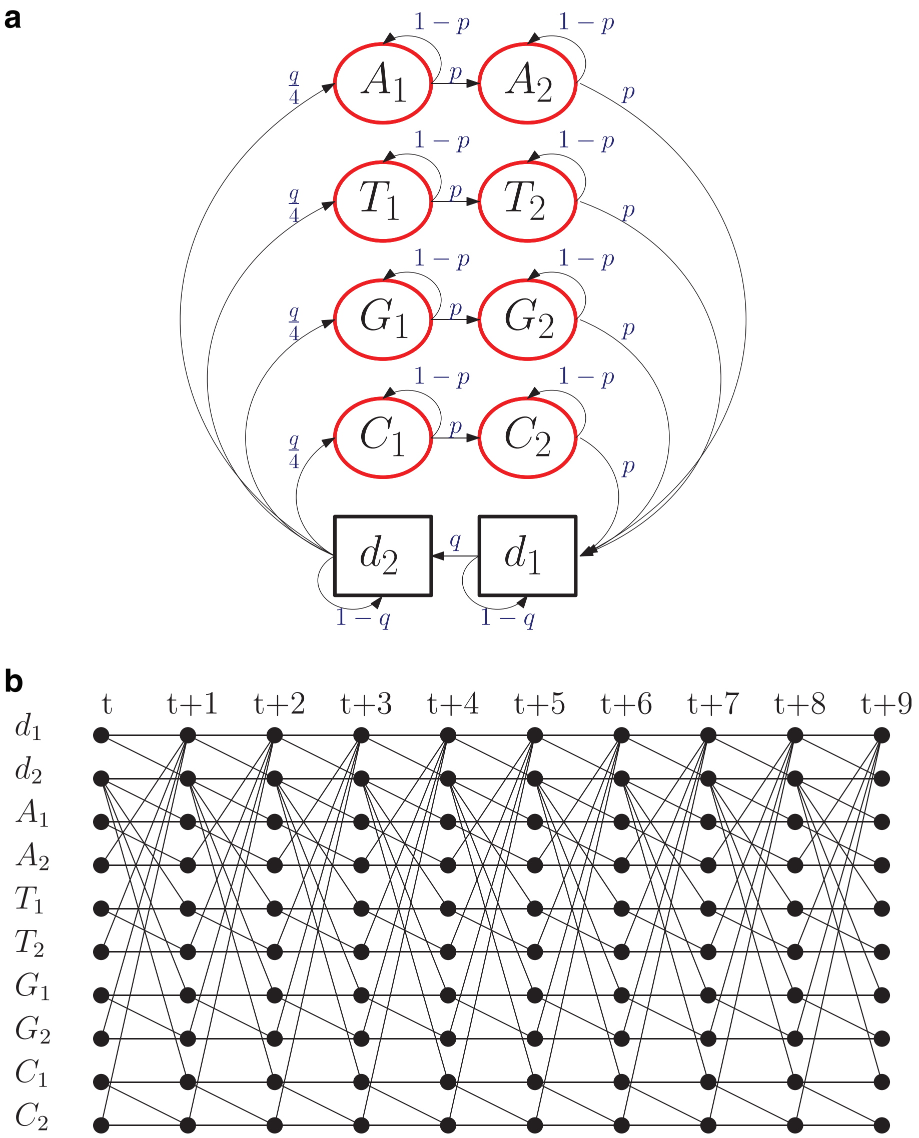

FIG. 4. **(a)** The state transition diagram of the Markov chain $\{L(t), t \geq 1\}$. In the example shown, $l=2$ and $r=2$. **(b)** An example of the trellis diagram corresponds with the state transition diagram in **(a)**.

3.3. Error performance under perfect information

Dye selection is one of the most important factors in the fluorescence labeling method. Different combinations of dye selection affect the performance of fluorescence signal detection, specifically in the process of sequencing. With the model established in this study, we compute the upper bound of the error detection probability of the distinct fluorescence combination, which will help to select the optimal group of dyes. The main characteristics of the fluorescence dyes governing signal detection are described through the model.

In the ideal scenario, assume the ambient light intensity source is in the steady state so that the ambient light intensity signal can be modeled by a constant signal vector $\boldsymbol{a}_d$. Also, assume the fading coefficient $\Upsilon(t) = \Upsilon_n,\ \forall t \in [s_n, t_n - 1]$.

With perfect information, that is,

$$\widehat{\boldsymbol{a}_x} = \boldsymbol{a}_d,\ \widehat{\boldsymbol{s}_x} = \boldsymbol{s}_x,\ \widehat{\sigma_i^2} = \sigma_i^2,\ \widehat{\alpha_n} = \alpha_n,\ \text{ and } \widehat{s_n} = s_n,\ \widehat{t_n} = t_n \forall x \in \{A, T, G, C\},$$

the error probability is

$$\begin{aligned} P(X_n \neq \widehat{X_n}) &= \sum_{x \in \{A, T, G, C\}} P(X_n = x) P(\widehat{X_n} \neq x | X_n = x) \\ &= \frac{1}{4} \sum_{x \in \{A, T, G, C\}} P(\widehat{X_n} \neq x | X_n = x) \\ &= \frac{1}{4} \sum_{x \in \{A, T, G, C\}} P(\cup_{y \in \{A, T, G, C\},\, y \neq x} F_{yx} | X_n = x) \\ &\leq \frac{1}{4} \sum_{x \in \{A, T, G, C\}} \sum_{y \in \{A, T, G, C\},\, y \neq x} P(F_{yx} | X_n = x), \end{aligned}$$

where F_{yx} is the event that

$$\sum_{k=s_n}^{t_n - 1} (\boldsymbol{r}(k) - \alpha_n \boldsymbol{s}_y - \boldsymbol{a}_d)^t \Lambda^{-1} (\boldsymbol{r}(k) - \alpha_n \boldsymbol{s}_y - \boldsymbol{a}_d) \leq \sum_{k=s_n}^{t_n - 1} (\boldsymbol{r}(k) - \alpha_n \boldsymbol{s}_x - \boldsymbol{a}_d)^t \Lambda^{-1} (\boldsymbol{r}(k) - \alpha_n \boldsymbol{s}_x - \boldsymbol{a}_d).$$

Given $X_n = x$, we have

$$\boldsymbol{r}(k) = \alpha_n \boldsymbol{s}_x + \boldsymbol{a}_d + \boldsymbol{n}(k),\ \ k \in [s_n, t_n - 1],$$

so that

$$\boldsymbol{r}(k) - \alpha_n \boldsymbol{s}_y - \boldsymbol{a}_d = \alpha_n (\boldsymbol{s}_x - \boldsymbol{s}_y) + \boldsymbol{n}(k),\ \ k \in [s_n, t_n - 1],$$

and

$$\boldsymbol{r}(k) - \alpha_n \boldsymbol{s}_x - \boldsymbol{a}_d = \boldsymbol{n}(k),\ \ k \in [s_n, t_n - 1].$$

Thus, the event F_{yx} is

$$\sum_{k=s_n}^{t_n - 1} (\alpha_n (\boldsymbol{s}_x - \boldsymbol{s}_y) + \boldsymbol{n}(k))^t \Lambda^{-1} (\alpha_n (\boldsymbol{s}_x - \boldsymbol{s}_y) + \boldsymbol{n}(k)) \leq \sum_{k=s_n}^{t_n - 1} \boldsymbol{n}(k)^t \Lambda^{-1} \boldsymbol{n}(k)$$

$$\Updownarrow$$

$$\frac{\alpha_n}{2} (\boldsymbol{s}_y - \boldsymbol{s}_x)^t \Lambda^{-1} (\boldsymbol{s}_y - \boldsymbol{s}_x) \ \leq\ (\boldsymbol{s}_y - \boldsymbol{s}_x)^t \Lambda^{-1} \frac{\sum_{k=s_n}^{t_n - 1} \boldsymbol{n}(k)}{t_n - s_n}$$

$$\Updownarrow$$

$$\frac{\alpha_n}{2} \sum_{i=1}^{3} \left(\frac{s_{y,i} - s_{x,i}}{\sigma_i} \right)^2 \leq \sum_{i=1}^{3} \left(\frac{s_{y,i} - s_{x,i}}{\sigma_i} \right) \left(\frac{1}{t_n - s_n} \sum_{k=s_n}^{t_n - 1} \frac{n_i(k)}{\sigma_i} \right).$$

Since the averaged scaled noises $\frac{1}{t_n - s_n} \sum_{k=s_n}^{t_n - 1} \frac{N_i(k)}{\sigma_i}$, $i = 1, 2, 3,$ has normal distribution $N\left(0; \frac{1}{t_n - s_n}\right)$ and are independent, the combined noise

$$N=\sum_{i=1}^{3}\left(\frac{s_{y,\,i}-s_{x,\,i}}{\sigma_i}\right)\left(\frac{1}{t_n-s_n}\sum_{k=s_n}^{t_n-1}\frac{N_i(k)}{\sigma_i}\right)$$

has normal distribution $N\left(0;\frac{1}{t_n-s_n}\sum_{i=1}^{3}\left(\frac{s_{y,\,i}-s_{x,\,i}}{\sigma_i}\right)^2\right)$ with zero mean and variance

$$\sigma^2=\frac{1}{t_n-s_n}\sum_{i=1}^{3}\left(\frac{s_{y,\,i}-s_{x,\,i}}{\sigma_i}\right)^2.$$

Thus, the conditional probability $P(F_{yx}|X_n=x)$ is

$$\begin{aligned}P(F_{yx}|X_n=x)&=P\left(N\geq\frac{\alpha_n}{2}\sum_{i=1}^{3}\left(\frac{s_{y,\,i}-s_{x,\,i}}{\sigma_i}\right)^2\right)\\&=P\left(N\geq\frac{\alpha_n}{2}(t_n-s_n)\sigma^2\right)\\&=P\left(\frac{N}{\sigma}\geq\frac{\alpha_n}{2}(t_n-s_n)\sigma\right)\\&=Q\left(\frac{\alpha_n}{2}(t_n-s_n)\sigma\right),\end{aligned}$$

where the Q function is defined as

$$Q(x)=\int_x^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt,\ \ \forall\, t>0.$$

The Q function has tight lower and upper bounds

$$\frac{1}{\sqrt{2\pi}x}\left(1-\frac{1}{x^2}\right)e^{-\frac{x^2}{2}}<Q(x)<\frac{1}{\sqrt{2\pi}x}e^{-\frac{x^2}{2}},\ \ \forall\, x>0$$

and a very good upper bound

$$Q(x)<\frac{1}{2}e^{-\frac{x^2}{2}},\ \ \forall\, x>0.$$

We now have an upper bound for $P(F_{yx}|X_n=x)$

$$P(F_{yx}|X_n=x)<\frac{1}{2}\exp\left\{-\frac{\alpha_n^2(t_n-s_n)^2\sigma^2}{8}\right\}=\frac{1}{2}\exp\left\{-\frac{\alpha_n^2(t_n-s_n)}{8}\sum_{i=1}^{3}\left(\frac{s_{y,\,i}-s_{x,\,i}}{\sigma_i}\right)^2\right\}$$

so that the error probability is upper bounded by

$$\begin{aligned}P(X_n\neq\widehat{X_n})&\leq\frac{1}{4}\sum_{x,\,y\in\{A,\,T,\,G,\,C\},\,y\neq x}P(F_{yx}|X_n=x)\\&<\frac{3}{2}\exp\left\{-\frac{\alpha_n^2(t_n-s_n)}{8}\min_{x\neq y}\sum_{i=1}^{3}\left(\frac{s_{y,\,i}-s_{x,\,i}}{\sigma_i}\right)^2\right\}\end{aligned}\tag{11}$$

Considering the expression, $-\frac{\alpha_n^2(t_n-s_n)}{8}\min_{x\neq y}\sum_{i=1}^{3}\left(\frac{s_{y,\,i}-s_{x,\,i}}{\sigma_i}\right)^2$, in Equation (11), there are four factors affecting the results of the fluorescence signal detection. α_n affects the expression with the power of two, while the nth time period of the incorporation, t_n-s_n, is with the power of one. The minimum distance among the wavelength of the distinct dyes denoted by $\min_{x\neq y}\sum_{i=1}^{3}(s_{y,\,i}-s_{x,\,i})^2$ is with the power of one, and the noise power is also with the power of one. Within the same environment and system, the time duration of the incorporation and specifically the minimum distance among the fluorescence dyes would influence the performance while the scientists would like to select the combination of the dyes.

TABLE 1. THE PHOTOELECTRIC CONVERSION VECTORS, $\boldsymbol{s}_x$ AND $\boldsymbol{a}_x$, ARE CALCULATED FROM THE MEASUREMENT WITH DISTINCT WAVELENGTHS OF LED LIGHT EMISSION

Wavelength (nm)	*RGB PC vector $\boldsymbol{s}_x$ (ms·mv/photon)*	*RGB PC vector $\boldsymbol{a}_x$ (mv)*
530	(0.153, 0.063, 0.025)	(1.390, 1.457, 1.105)
590	(0.098, 0.063, 0.059)	(1.915, 1.456, 0.924)
625	(0.078, 0.057, 0.076)	(1.856, 1.416, 0.891)
656	(0.069, 0.054, 0.087)	(1.595, 1.193, 0.527)

PC, photoelectric conversion; RGB.

4. SIMULATIONS

The measured data are to utilize four single-wavelength LEDs (530, 590, 625, and 656 nm) to simulate fluorescence effect on nucleotides (dATP, dTTP, dGTP, and dCTP), where 530 nm LED represents the emission for the phospholinked dATP and so forth. Assume that the pulse width and the interpulse duration are in the average of 100 and 200 ms and the parameters p and q are estimated from Equations (1) and (2), respectively. Since the shorter pulse widths and interpulse duration as compared with the previous study, the light-emitting pattern will be more restrictive, which means the internal and external environments of the experiments are allowed to be more flexible. Data collection involves the sampling procedures detailed hereunder. To compare the simulated data with the previous study (Eid et al., 2009), we established data with varying degrees of light intensity to achieve the different SNR to be used as a criterion for comparison. The intensity of the LED light emission follows the random walks in the nine bounded domains: [20,40], [30,60], [40,80], [50,100], [60,120], [70,140], [80,160], [90,180], and [100,200] (photon/ms). We generate 100 samples for each pattern and 1000 random nucleotides for each sample. For each nucleotide, the emission process is generated by the model of the random walk, and the SMSP is generated according to the transition diagram in Figure 2.

The received data derive from the LED emission process as the photodiode receives the LED light that involves voltages transformed from light emission and obtained from the photoelectric effect.

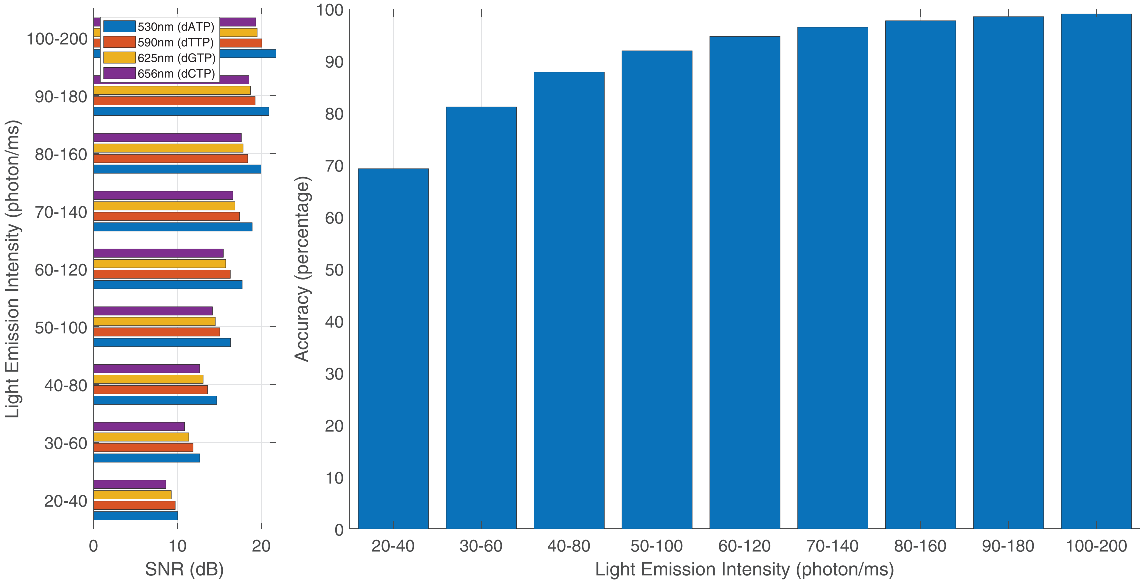

FIG. 5. The left diagram illustrates the related SNR calculated by different levels of the intensity of the four distinct fluorescences. The main diagram exhibits the accuracy of our decoding method with the different fluctuating light emission intensity. SNR, signal-to-noise ratio.

Table 2. The Comparison of the Performance between the PacBio Platform and Our Detection Architecture (Maximum Likelihood-Viterbi Algorithm Within a Three-Junction Photodiode)

Platform: PacBio		*SNR: 22–30 dB*		
	A555-dATP	*A568-dTTP*	*A647-dGTP*	*A660-dCTP*
Pulse width (ms)	133 ± 22	91 ± 13	117 ± 14	96 ± 10
Interpulse duration (ms)	770 ± 250	670 ± 220	960 ± 210	790 ± 230
	Correct	Mismatches	Insertions	Deletions
Performance (%)	82.9	4.4	5.1	7.6
ML-Viterbi algorithm		*SNR: 14.16–16.32 dB*		
	530-dATP	*590-dTTP*	*625-dGTP*	*656-dCTP*
Pulse width (ms)	100 ± 5.77	100 ± 5.77	100 ± 5.77	100 ± 5.77
Interpulse duration (ms)	200 ± 18.26	200 ± 18.26	200 ± 18.26	200 ± 18.26
	Correct	Mismatches	Insertions	Deletions
Performance (%)	91.94	8.06	0	0
ML-Viterbi algorithm		*SNR: 17.6–19.5 dB*		
	530-dATP	*590-dTTP*	*625-dGTP*	*656-dCTP*
Pulse width (ms)	100 ± 5.77	100 ± 5.77	100 ± 5.77	100 ± 5.77
Interpulse duration (ms)	200 ± 18.26	200 ± 18.26	200 ± 18.26	200 ± 18.26
	Correct	Mismatches	Insertions	Deletions
Performance (%)	97.72	2.28	0	0

dATP; dCTP; dGTP; dTTP; ML, maximum likelihood; SNR, signal-to-noise ratio.

We measure the received RGB values through PC while the LED light intensity increases as time progresses, and the relation is shown in Table 1. The PC values are distinct to our previous study (Chen and Lu, 2019) since under different environments such as distinct dyes and temperatures we will get different PC values.

The corresponding average SNRs are calculated from these different levels of the light intensity as in the left diagram of Figure 5.

5. RESULTS

The estimation of nucleotide sequences is derived from our detection method, called ML-Viterbi algorithm in this study. We apply the Smith–Waterman algorithm to align between the expected sequences and the measured sequences (Gotoh, 1982; Slater and Birney, 2005). In-depth data analysis is detailed in Figure 5 and Table 2. Deducting from the modeling in aforementioned, the sequencing accuracy results beyond 97% even with the fluctuating light emission intensity output ranging from 80 to 160 photon/ms and the SNR is <19 dB. Moreover, deletions and insertions, except mismatches, of sequencing will not occur under the scenario of SNR beyond 15 dB. In comparison, described from previous study in Eid et al. (2009), namely the PacBio platform, the ratio of signal to noise is $\sim 24 \pm 10$, ranging from 22 to 30 dB. The accuracy is approximately <85% without support from consensus sequencing method with higher SNR. The differences present the significance that the ML-Viterbi algorithm is outstanding at signal detection even under an unfavorable low SNR environment.

6. DISCUSSION AND CONCLUSIONS

Under the assumption of identical objective conditions of two detection methods, single photodiode detection prevails current image processing technology (multiple pixel inputs form CCD camera) at its processing speed, due to the single pixel processing. To be specific, the method in this study applying to the fluorescence labeling technique with a single photodiode delivers better performance, especially in the DNA sequencing, with its great accuracy (lower error rate), possible higher throughput, and potentially lower cost. Incorporating with the consensus sequencing method, the accuracy rate of data can be improved optimistically to nearly free of error (Hiatt et al., 2010). In addition, modeling on fluorescent dye selection and optimized combination reduce the time and resources required in comparison with the actual experiment. However, the throughput level is highly dependent on how advanced computing capability has to offer on the simultaneous sequencing of single photodiode groups. Furthermore, detection accuracy may be further enhanced by increasing the difference between the fluorescence characteristics to label four kinds of dNTPs. In summary, the algorithm elaborated in this article is definitely an excellent aid to the development of fluorescence-based sequencing.

ACKNOWLEDGMENT

We are grateful to Personal Genomics, Inc. for supporting the integrated circuit with the three-junction photodiode and the experiments of the LED system.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

No funding was received in support of this work.

REFERENCES

Ardui, S., Ameur, A., Vermeesch, J.R., et al. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* 46, 2159–2168.

Chen, H.H., and Lu, C.C. 2019. Optimized multiple fluorescence based detection in single molecule synthesis process under high noise level environment, 65–76. *In* Măndoiu, I., Murali, T., Narasimhan, G., Rajasekaran, S., Skums, P., and Zelikovsky, A., eds. *International Conference on Computational Advances in Bio and Medical Sciences*. Springer, Cham.

Deamer, D., Akeson, M., and Branton, D. 2016. Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518.

Dean, K.M., and Palmer, A.E. 2014. Advances in fluorescence labeling strategies for dynamic cellular imaging. *Nat. Chem. Biol.* 10, 512.

Eid, J., Fehr, A., Gray, J., et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138.

Epstein, J.R., Biran, I., and Walt, D.R. 2002. Fluorescence-based nucleic acid detection and microarrays. *Anal. Chim. Acta* 469, 3–36.

Gotoh, O. 1982. An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162, 705–708.

Hermanson, G.T. 2013. *Bioconjugate Techniques*, 395–465. Academic Press, Boston, Massachusetts.

Hiatt, J.B., Patwardhan, R.P., Turner, E.H., et al. 2010. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* 7, 119.

Horne, K. 1986. An optimal extraction algorithm for CCD spectroscopy. *Publ. Astron. Soc. Pac.* 98, 609.

Hsieh, H.Y., Peng, Y.H., Lin, S.F., et al. 2019. Triple-junction optoelectronic sensor with nanophotonic layer integration for single-molecule level decoding. *ACS Nano* 13, 4486–4495.

Jansen-van Vuuren, R.D., Armin, A., Pandey, A.K., et al. 2016. Organic photodiodes: the future of full color detection and image sensing. *Adv. Mater.* 28.24, 4766–4802.

Pollard, M.O., Gurdasani, D., Mentzer, A.J., et al. 2018. Long reads: their purpose and place. *Hum. Mol. Genet.* 27, R234–R241.

Schmitt, M.W., Kennedy, S.R., Salk, J.J., et al. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 109, 14508–14513.

Seo, T.S., Bai, X., Kim, D.H., et al. 2005. Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5926–5931.

Shendure, J., Balasubramanian, S., Church, G.M., et al. 2017. DNA sequencing at 40: past, present and future. *Nature* 550, 345–353.

Slater, G.S.C., and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 31.

Valm, A.M., Oldenbourg, R., and Borisy, G.G. 2016. Multiplexed spectral imaging of 120 different fluorescent labels. *PLoS One* 11, e0158495.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* 13, 260–269.

Address correspondence to:
Prof. Chung-Chin Lu
Department of Electrical Engineering
National Tsing Hua University
Hsinchu City 30013
Taiwan

E-mail: cclu@ee.nthu.edu.tw