Journal of Biomedical Optics

SPIEDigitalLibrary.org/jbo

Toward automated denoising of single molecular Förster resonance energy transfer data

Hao-Chih Lee
Bo-Lin Lin
Wei-Hau Chang
I-Ping Tu

SPIE

Toward automated denoising of single molecular Förster resonance energy transfer data

Hao-Chih Lee,[a] Bo-Lin Lin,[b] Wei-Hau Chang,[b] and I-Ping Tu[a]
[a]Academia Sinica, Institute of Statistical Science, Taipei, Taiwan
[b]Academia Sinica, Institute of Chemistry, Taipei, Taiwan

Abstract. A wide-field two-channel fluorescence microscope is a powerful tool as it allows for the study of conformation dynamics of hundreds to thousands of immobilized single molecules by Förster resonance energy transfer (FRET) signals. To date, the data reduction from a movie to a final set containing meaningful single-molecule FRET (smFRET) traces involves human inspection and intervention at several critical steps, greatly hampering the efficiency at the post-imaging stage. To facilitate the data reduction from smFRET movies to smFRET traces and to address the noise-limited issues, we developed a statistical denoising system toward fully automated processing. This data reduction system has embedded several novel approaches. First, as to background subtraction, high-order singular value decomposition (HOSVD) method is employed to extract spatial and temporal features. Second, to register and map the two color channels, the spots representing bleeding through the donor channel to the acceptor channel are used. Finally, correlation analysis and likelihood ratio statistic for the change point detection (CPD) are developed to study the two channels simultaneously, resolve FRET states, and report the dwelling time of each state. The performance of our method has been checked using both simulation and real data. [DOI: 10.1117/1.JBO.17.1.011007]

Keywords: change point detection; denoising; dimension reduction; fluorescence resonance energy transfer; molecular imaging; total internal reflection.

Paper 11266SS received May 27, 2011; revised manuscript received Sep. 16, 2011; accepted for publication Sep. 19, 2011; published online Feb. 8, 2012.

1 Introduction

The Förster resonance energy transfer (FRET) is a radiationless process between two fluorophores, donor and acceptor, whose intensities ratio defines transfer effeciency (E) and reports the in-between distance $\{E = [1 + (R/R_0)^6]^{-1}\}$, where R_0 is the Förster distance between the donor and the acceptor. As a single biological molecule is labeled by a FRET pair and immobilized,[1] its conformation dynamics associated with the function can be recorded by using a two-channel fluorescence microscope to track FRET changes for an extended period of time.[2] In practice, this type of single-molecule FRET (smFRET) microscope can be realized through a confocal configuration,[3] by which one molecule is imaged at a time, or a wide-field configuration that allows for hundreds of molecules to be simultaneously monitored by a pixelated detector. To achieve the detection of single-molecule fluorescence in a wide-field configuration, total internal reflection (TIR) has been employed to generate an evanescent wave to excite a thin layer near the interface such that the fluorescence background from the bulk can be greatly reduced.[1,3–6]

The wide-field smFRET data are embedded in movies of N time-frame of CCD images. Each image is divided into two half-images: one of the donor (usually Cy3, TMR, Alexa, or Atto 555) and the other of the acceptor channel (Cy5, Alexa, or Atto 647). To maximize the temporal resolution and achieve the longest single-molecular time traces before photo-bleaching occurs, one should increase the CCD frame rate as fast as possible and keep the laser illumination power as low as possible, yielding the time trace data with poor signal-to-noise (S/N) ratio. A wide-field smFRET microscopy can be realized by a prism type TIR (PTIR)[1] or a high-numerical aperture (NA) objective type (OTIR) (Fig. 1).[4,6] We chose the OTIR as it is easy to build and offers an empty space on top of the sample slide that would allow for sample manipulation from above. Compared to PTIR, OTIR requires an additional dichroic mirror underneath the objective to direct the excitation beam into the objective lens and yet to prevent the reflected excitation beam from entering the pathway of the fluorescence collection optics. The photon collection is thus comprised in the OTIR configuration despite having a high NA. We have experienced higher fluorescence background that might come from the objective lens or its associated elements as reported.[7,8] All those conditions have together yielded very noisy data so that it becomes very challenging to studying hundreds of thousands of noisy time traces to select meaningful ones by human efforts. To facilitate the data reduction from smFRET movies to smFRET traces and address the noise-limited issues, we developed a denoising recipe that utilizes novel statistical approaches based on the spatial and/or temporal correlation at different steps in the work flow. Due to these new algorithms, the work flow can be automated.

As to the estimation and removal of the background, local subtraction[5,9,10] and profile fitting[11] are two commonly used methods, both of which utilize local information around the fluorophores. Interestingly, as we introduced a global method based on higher-order singular value decomposition (HOSVD) to extract spatial and temporal features in a movie for denoising

Address all correspondence to: I-Ping Tu, Institute of Statistical Science, No. 128, Sec. 2, Academia Rd., Nangang Dist., Taipei 11529, Taiwan. Tel: 886 2 27871952; E-mail: iping@stat.sinica.edu.tw.

0091-3286/2012/$25.00 © 2012 SPIE

the system errors, we found that it could estimate the non-uniform TIR background quite well. As to mapping and registering the fluorophore coordinates in the two channels, we used temporal correlation to locate those spots representing the donor signals bleeding through the acceptor channel and then used their coordinates to determine the best linear transformation matrix. We found the transformation matrix was not adequate because non-uniform deviations on the coordinate mapping were observed across the whole image. To identify the pairs, our algorithm performs a spatially exhaustive search in the neighboring pixels of the spot coordinates predicted by the matrix. Once the trace pair was found, we checked whether it was a FRET trace by studying if there were donor and acceptor anti-correlated patterns along the pair traces. For those FRET traces, we used change point detection (CPD) to detect the status change points and reported the dwelling time. To do so, we derived a likelihood ratio statistic given a change point position under a multivariate Gaussian model and used it to search all time points in the interval between adjacent change points to find the new change point candidates. Conventionally, the status change points are solved by hidden Markov model (HMM),[14] multivariate Gaussian HMM (MGHMM),[15] or time order clustering (TOC).[16] In contrast to those methods, CPD is a deterministic approach that does not require assigning initial values such as state means and variance. As such, our work flow employing the CPD algorithm could be executed without human intervention.

This paper is organized as follows: Section 2 introduces the analysis methods, and these methods are summarized in five algorithms. In Sec. 3 we present our real data analysis and simulation studies. This paper ends with a brief discussion.

2 Methods

The main goal of the FRET data analysis is to target the FRET traces and detect the status change points along the trace. Our strategy includes three steps, which are summarized in a flow chart in Fig. 2. The first step is to denoise the system error. We applied HOSVD to estimate the non-homogenous system error. Traditionally, the corresponding approach is to perform local subtraction of the intensities in the surrounding pixels of a spot. These two approaches are shown to match quite well on both simulated data and real data. One extra benefit from the HOSVD approach is that the estimated system error could give feedback to the experimenters so as to optimize their instrument setup.

The second step is to find the paired FRET traces. In order to make it automated, three sub-steps are executed by applying three algorithms. They involve 1. locating the fluorophores from the acceptor images; 2. mapping the coordinate system of the donor images and acceptor images; 3. finding all possible fluorophore paired traces.

The third step is to detect the change points along the FRET trace pairs. Instead of detecting the change points along the FRET trace resulting from the ratio of the acceptor to the combination of the acceptor and donor, we treated the donor and the acceptor traces as multivariate variables and detected the change points along them. To detect the change points, we adopted the multivariate Gaussian model and used the log-likelihood ratio as our statistics.[12,13] We also checked whether those traces are locally anti-correlated across time. The detection criterion is based on the p-value. Once the threshold (given a p-value) is determined, the computation is deterministic, involving no random initials, which includes the state mean and variance

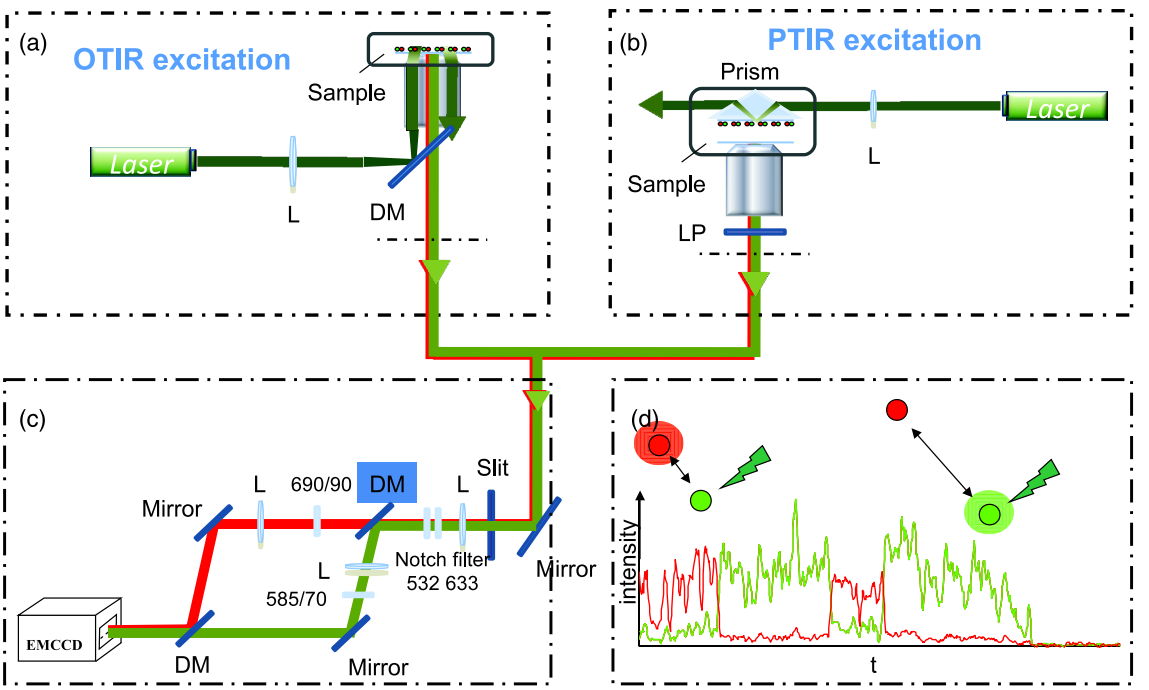

Fig. 1 smFRET microscopy configuration. (a) Objective TIR (OTIR): the laser is focused onto the back focal plane to generate an evanescent wave at cover-slip interface (L: lens; DM: dichroic mirror). (b) Prism TIR (PTIR): the laser beam is shined onto the prism to generate evanescent wave at the quartz slide-buffer interface (L: lens; LP: long-pass filter). (c) Emitted photons are separated into two channels by a dichroic system, the dual-view system (DM: dichroic mirror, EMCCD: electron multiplication CCD). (d) Illustration of a single molecular FRET time trace.

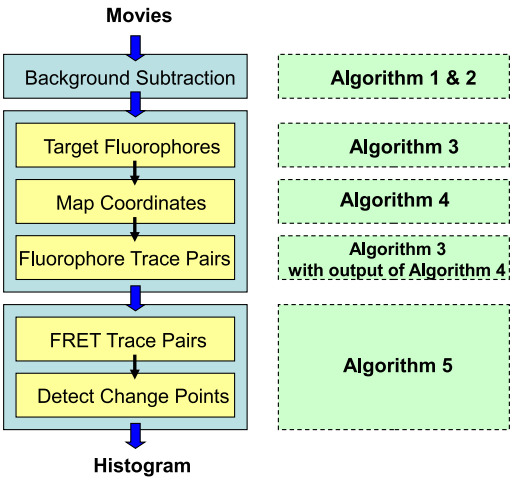

Fig. 2 The flow-chart of automated denoising (ADN).

values or pre-specified number of states as in commonly used methods such as HMM,[14] MGHMM,[15] and TOC.[16]

To test our algorithm, both simulation and real data were used. The sample used to generate real data shown in Figs. 3 to 6 was a 16 bp GC-rich double strand DNA with Cy3 and Cy5 attached at the 5' ends modified from a design previously described[17] to allow for specific binding of a protein, of which the details will be published in a separate paper. To compare the performance of CPD with MGHMM (Fig. 7), the movie data were provided by Ha's laboratory website (http://www.cplc.illinois.edu), and the sample is a DNA plus its binding protein.

2.1 *System Error Denoising*

One particular feature of the smFRET experiment is that the fluorescent molecules are immobilized. Thus, we proposed a local model: let (i_0, j_0) be the center of a fluorophore, then the signal can be modeled as

$$I^L_{i_0,j_0}(i,j,k) = F^{\mathrm{ps}}{}_{i_0,j_0}(i,j) \cdot g_{i_0,j_0}(k), \quad \times \text{for } 1 \le i \le I, 1 \le j \le J, 1 \le k \le K, \quad (1)$$

where $F^{\mathrm{ps}}{}_{i_0,j_0}(\cdot,\cdot)$ is the unimodal point spread function centered at (i_0, j_0) while $g_{i_0,j_0}(\cdot)$ describes the signal emitted by the fluorophore on the time trace. Since there are many fluorophores, we have the global model:

$$I^G(i,j,k) = B(i,j,k) + \sum_{l=1}^{M} I^L_{i_l,j_l}(i,j,k) + N^G(i,j,k), \quad (2)$$

where B models the system error and N^G describes the random noise. $I^G(\cdot,\cdot,\cdot)$ is multi-arrayed data with global structure $B(\cdot,\cdot,\cdot)$ and many local structures $I^L_{i_l,j_l}(\cdot,\cdot)$. We tried to decompose the structures by HOSVD:[18] $I^G(i,j,k) = \sum_{m,n,l} a_{m,n,l} f_{x,m}(i) f_{y,n}(j) f_{z,l}(k)$, where each $f_{\cdot,\cdot}$

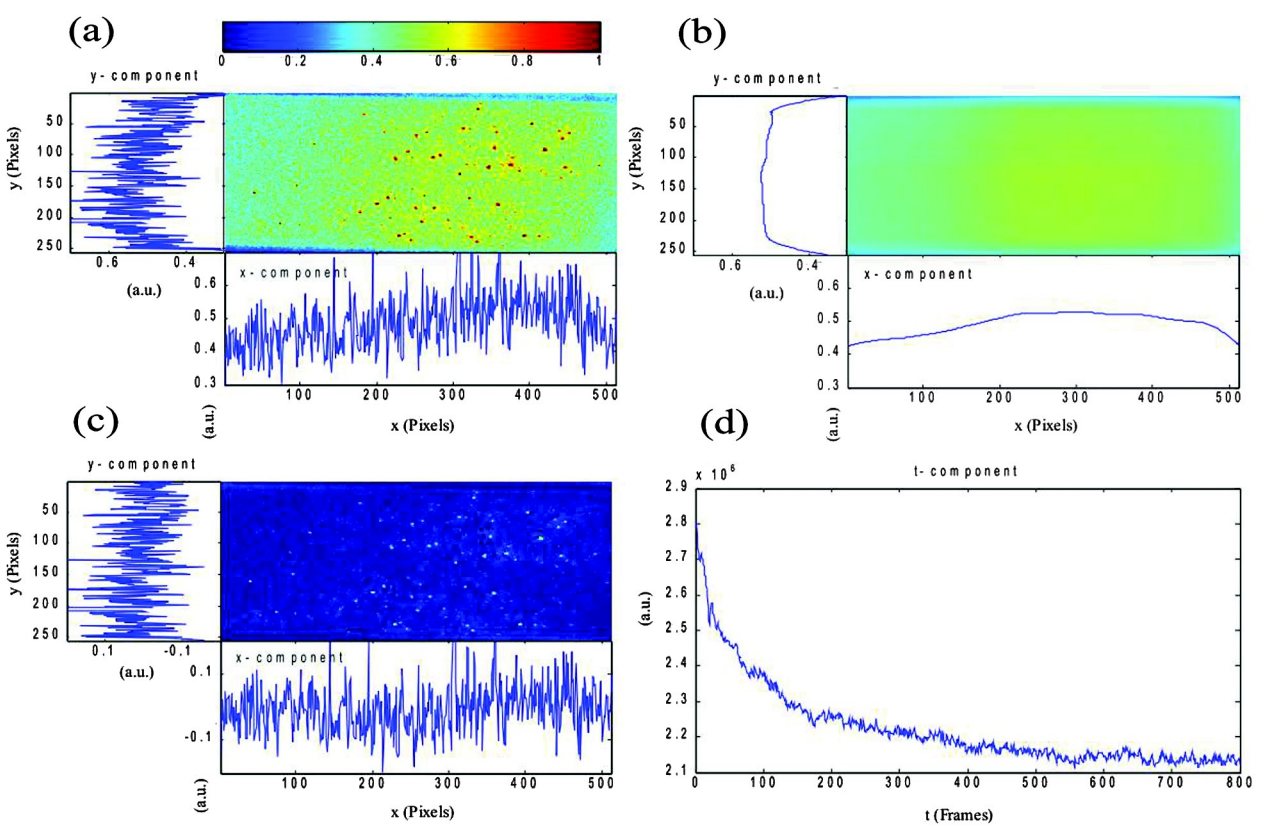

Fig. 3 One typical image frame and its estimated system error. (a) Raw image. (b) System error estimated by HOSVD. (c) Image after removing system error. (d) Time component. Original pixel values are rescaled into [0,1]. Plots at the left and under images (a), (b), and (c) show the 256th column and 128th row of each image. It is clear that the envelop trend of the 128th row in (a) is pretty much taken away after denoising as shown in (c). The width of side panels is limited to 0.4 to magnify the envelop trend. The time-component in (d) shows the temporal decay of the intensities.

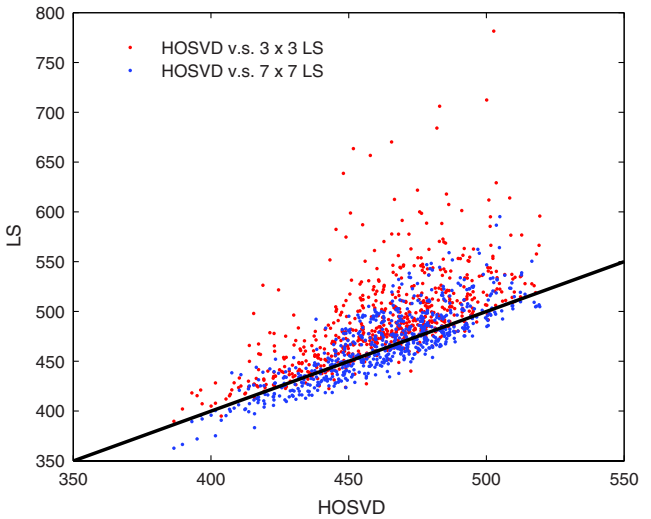

Fig. 4 Error estimated by HOSVD (*x*-axis) versus that by local subtraction (*y*-axis). Note that almost all the red circles fall on the upper side of the 45-deg line, indicating the estimated system error from 3×3 LS are usually higher than those from HOSVD. On the other hand, values of 7×7 LS agree better with those from HOSVD.

represents one linear mixture on one mode of the array and each product term of $f_{x,}f_{y,}f_{z,}$ represents one component. $a_{,,,}$ refers to the projected size of the data on the corresponding component. The solution to this decomposition problem can be obtained by the alternating least square algorithm,[19,20] described in Algorithm 1. It may be helpful to link this model with the commonly used SVD (singular value decomposition), which can be viewed as a two-order version of this model.

Most data analysis would model signal components with larger variance (eigenvalue) and noise with smaller variance; however, this is not so in our case. In real data analysis, we found the first component with the largest eigenvalue describes the system error $B(\cdot,\cdot,\cdot)$ very well. The data structure indicates that this result would not be a surprise. Each signal $I^L_{:,:}(\cdot,\cdot,\cdot)$, localized within 5 by 5 pixels, would unlikely produce large variance. On the other hand, the system error, representing global information, can contribute a large portion of the total variance. The phenomenon that the first component captures the global information has also been reported on an integrative analysis of DNA microarray set combining different studies.[21] To be specific, we used $a_{1,1,1}f_{x,1}(i)f_{y,1}(j)f_{z,1}(k)$ to estimate $B(i,j,k)$ and delete it from the data as a denoising process. We further developed Algorithm 2 to smooth this term in order to avoid the possible mixture of the signals. Fig. 3 provides a typical example.

We also designed simulation experiments to check the performance of this algorithm. The result is very supportive. We compared this algorithm with the concurrent local background subtraction method. This part is reported in section 3.1.

2.2 *Extract FRET Traces*

2.2.1 *Locate fluorophores*

The fluorophores we are interested in have two characters: they are immobilized, and they last for a period of time. Therefore, the corresponding pixel values are supposed to be comparatively high over a range of spatial index and over a period of time. The image with the system error denoised is transformed to a binary image, depending on whether the pixel value exceeds a given threshold. The threshold is set to be the overall mean plus only one standard deviation to avoid missing the candidate fluorophore cluster. Subsequently, the clusters with the value of one are searched throughout the spatial index and the time frame index as well. The detailed procedure is stated in Algorithm 3.

2.2.2 *Image registration (map the coordinate systems)*

Traditionally, the mapping of coordinates relies on imaging a calibration slide made of fluorescence beads of broad spectrum before collecting the data from the experimental slides as a daily practice.[5,6] The fluorescence spots in the donor channel and their leakage pairs in the acceptor channel are used for the registration of the two coordinate systems by solving a transformation matrix. Alternatively, we took an approach that used the leakage spots found in the experimental images to determine the mapping between the two channels so that the work on the fluorescence beads could be omitted. Leakage spots refer to those

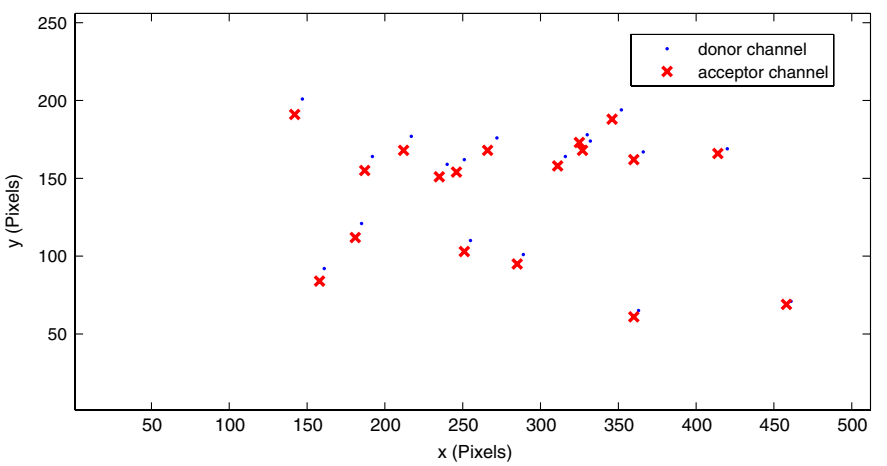

Fig. 5 Correlated leakage pairs throughout the half CCD image. As the two channels are superimposed, the apparent linear relationship between leakage pairs can be observed in the center of the field but not outside the central region.

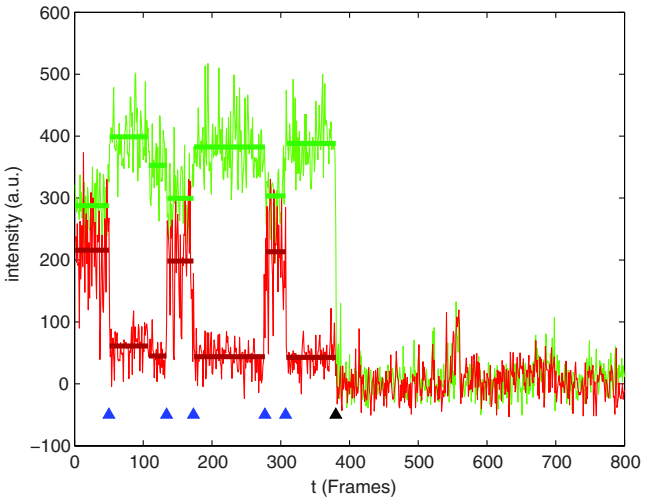

Fig. 6 FRET detection in experimental data. Blue triangles indicate the candidates of FRET events. Means of each interval $[c_i, c_{i+1}]$ are marked by a solid line.

donor-only fluorophores carrying abnormal intensities so that they bleed through into the acceptor channel. As a result, the values of those leakage pixels increase or decrease simultaneously in both channels. Having a spot located in the acceptor channel by the previous sub-step, we looked for its leakage pair in the donor channel by searching in the neighborhood of the corresponding spot coordinates. (Note that a simple translation relationship between the acceptor and the donor coordinate systems represents a good approximation to begin with when the two channels are aligned roughly parallel to each other.) When a set of leakage pairs was found, we applied least squared error method to determine the transformation matrix. This algorithm is briefly described in Algorithm 4.

2.2.3 *Search fluorophore paired traces*

Given the acceptor fluorophores and the coordinate transformation matrix, we applied Algorithm 3 again to search all possible donor fluorophores on the corresponding locations. In this part, the traditional method searches the candidate spots with good S/N at the corresponding locations in the donor channel by averaging the initial, say, 10 time frames. With this regard, we adopt a temporally exhaustive search along the time trajectory to include all possible fluorophores. The necessity comes from considering the following scenario in which the traditional approach will likely miss a donor spot while the greedy search will not: the FRET occurs for a period longer than 10 time frames in the beginning, such that no fluorophore would be detected in the donor channel based on the initial average.

2.3 *Change Point Detection*

2.3.1 *Check if FRET occurs*

Given a fluorophore trace pair, we check if there exists any real FRET event by anti-correlations. One means to do so is to calculate the correlation over the whole trace pair, which can be easily automated. However, this approach would sometimes generate an artifact when a time trace carries a dark state period and/or a long range of random noise. Instead, the human inspection approach usually focuses on the local anti-correlation around the candidate change point. To mimic human inspection, we applied a change point method to locate the possible change points and calculated the local correlations around them for

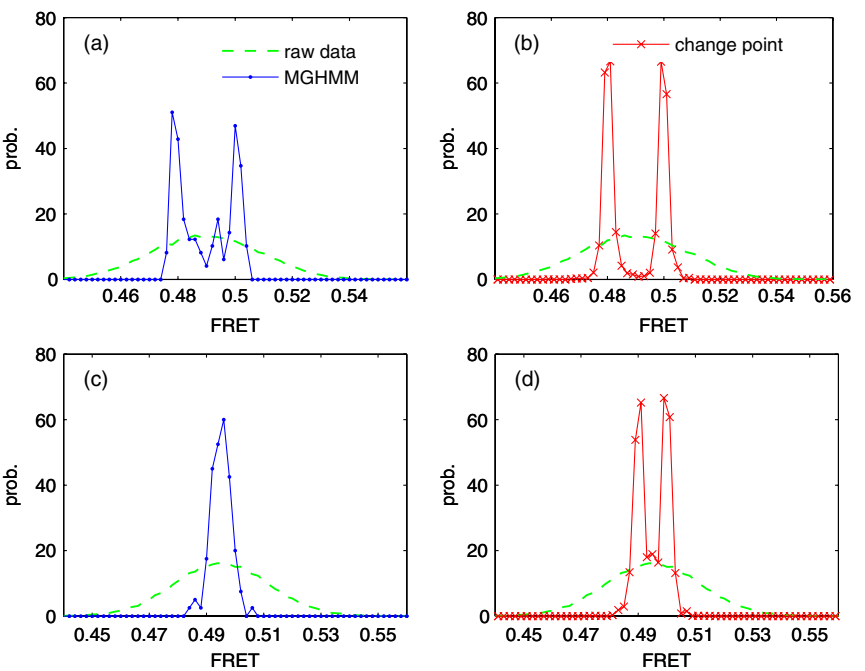

Fig. 7 FRET histogram of real data. There are 7073 and 8475 traces used for the DNA-only group and the DNA + protein group, respectively. (a) Histogram generated from unprocessed raw data. (b) Time traces are processed by two-state MGHMM, and the histogram is plotted using the mean values of each state. (c) Time traces are processed by our CPD method.

checking if FRET occurs. Our approach at this step can greatly shrink the size of the candidate set.

2.3.2 *Detecting change points*

Change point detection is a statistical method to detect different states in a sequential data set. A very brief introduction on CPD is presented in the Appendix. For the theoretical background, please refer to Siegmund[12] and Chen and Gupta.[13] Previously, Watkins and Yang[22] applied a likelihood ratio test to locate the intensity jumps as the change point based on individual photon arrival times in a single molecule trace extracted from one channel. In contrast, the CPD used here is based on multivariate Gaussian model that allows us to simultaneously analyze the paired intensity traces.[23] Fig. 6 is a typical example. The tuning parameter for this approach is the significant size (the threshold for p-value). Reasonable p-value is in the range between 0.1 and 0.01 : .1 (more flexible) or 0.05 (marginal) or 0.01 (for the sake of multiple testing). The computation is deterministic, such that it involves neither random initials nor a pre-specified number of states as those commonly used methods like HMM,[14] MGHMM,[15] or TOC.[16] which not only saves the computation time but also minimizes human intervention.

Algorithm 1: Alternating least square algorithm[19].
Given initial vectors f_x^0, f_y^0, f_t^0. Denote $f_{x,i}$ as i'th element of vector f_x.
input I(i, j, k) {# input movies of smFRET data}
for p = 0, 1, . . . P **do** {# repeat P iterations}
  for i = 1, . . . I **do**
    $f_{x,i}^{p+1} = \sum_{j,k} I(i,j,k) f_{y,j}^p f_{t,k}^p$
  for j = 1, . . . J **do**
    $f_{y,j}^{p+1} = \sum_{i,k} I(i,j,k) f_{x,i}^p f_{t,k}^p$
  for k = 1, . . . K **do**
    $f_{t,k}^{p+1} = \sum_{i,j} I(i,j,k) f_{x,i}^p f_{y,j}^p$
  normalize so that $\|f_x^{p+1}\|_2 = \|f_y^{p+1}\|_2 = \|f_t^{p+1}\|_2 = 1$
output f_x^p, f_y^p and f_t^p

Algorithm 2: Background estimation
input I(i, j, k), f_x, f_y.
f_x and f_y are outputs of Alogrithm 1.
$\tilde{f}_x$, $\tilde{f}_y$ are degree two local polynomial regressions of f_x, f_y with span 0.2.
$\tilde{f}_{t,k} = \sum_{i,j} I(i,j,k) \tilde{f}_{x,i} \tilde{f}_{y,j}$.
output $(B_{HOSVD})_{i,j,k} = \tilde{f}_{x,i} \tilde{f}_{y,j}$.

Algorithm 3: Localization of fluorophore
Let $D_l(i_0, j_0) = \{(i,j)\} | i - i_0 | \le l, |j - j_0| \le l\}$ be the square index centered at (l_0, j_0) with edge size $2l + 1$ and $|A|$ be the size of the set A.
input I(i, j, k). Denote the k'th frame of I by I_k.
$\mu_k = mean(I_k), \sigma_i = std(I_k)$.
for k = 1, . . . , K **do**
  get $E_k = [(i,j) | I(i,j,k) > \mu_k + \sigma_k]$. {# select high-intensity pixels}
for i_0, j_0 in E_k **do**
if $|D_1(i_0, j_0) \cap E_k| < 5$, eliminate (i_0, j_0) from E_k. {# keep aggregated high-intensity pixels}
$E = \cup_{r=1}^{K-W} \cap_{0 \le l < w} E_{r+l}$ for some chosen W {# keep sustained and aggregated high-intensity pixels}
for i_0, j_0 in E **do**
$A_{i_0,j_0}(k) = argmax[I(i,j,k) | (i,j) \in D_2(i_0, j_0)]$
$C(i_0, j_0) = mode_k(A_{i_0,j_0}(k))$
if $C(i_0, j_0) \ne (i_0, j_0)$, eliminate (i_0, j_0) from E.
output E. {# select a local maximum as a representative}

Algorithm 4: Coordinate transformation for two channels
Denote the time trace in donator and acceptor channels at (i, j) point by $I^D(i,j,k)$ and $I^A(i,j,k)$, respectively.
input I_k^A into Algorithm 3 with E as its output. {# get the coordinate indices for acceptor fluorophores}
Let E_A and E_D be two empty sets.
for i_0, j_0 in E **do**
  for $(r, s) \in D_4(i_0, j_0)$ **do**
    $\rho_{r,s}$ = correlation coefficient of $I^A(i_0, j_0, \cdot)$ and $I^D(r, s, \cdot)$.
    $(r^*, s^*) = argmax_{r,s} \rho_{r,s}$.
    if $\rho_{r^*,s^*} > 0.8$, {# search the possible leakage pairs around (i_0, j_0)}
    $(i,j) \to E_A$, $(r^*, s^*) \to E_D$ {# register the donor and acceptor coordinates for leakage pairs}
use least square method to find A which map the index $(r^*, s^*) \in E_D$ to its corresponding index $(i_0, j_0) \in E_A$.
output the transformation matrix A.

Algorithm 5: Change point detection.
Initially, let the set of change points C be $\{1, N\}$. Denote (X_D, X_A) as bi-variate traces of donor, and acceptor S_r is defined in the Appendix.
repeat
  Sort $C = \{c_1, c_2, \ldots, c_n\}$ such that $c_1 < c_2 < \ldots < c_n$.
for i = 1, . . . , n **do**
$r^* = argmax_{r \in (c_i, c_{i+1})} S_r$
if $S_{r^*} >$ threshold (c_{i+1}, c_i) and $min(r^* - c_i, c_{i+1} - r^*) > 5$
$r^* \to C$ {# find the possible change point in interval (c_i, c_{i+1}) and update the change point set C.
until no more r^* is significant in each interval
for i = 1, . . . , n **do**
  ρ_i = correlation coefficient of $(X_{D,j}, X_{A,j}), c_i - 20 \le j \le c_i + 20$.
  if there exists an i such that $\rho_i < -0.5$,}
  then label (X_D, X_A) as a FRET candidate. {# check an anti-correlated pattern around change points}
Collect mean values and dwelling time in each interval $[c_i, c_i + 1]$ for all FRET candidate to plot a histogram.

3 Real Data Analysis and Simulations

3.1 *Results of Background Subtraction*

To investigate the performance of HOSVD in estimating the system error, we simulated the data according to model 2. Poisson random variables $X_{i,j,k}$ are generated with conditional mean: $B(i,j) + \sum_{l=1}^{100} 1000 \cdot F^{ps}{}_{i_l,j_l}(i,j) g_{i_l,j_l}(k)$, $1 \le i \le 256$, $1 \le j \le 512$ and $1 \le k \le 40$. We generated the system error $B(i,j)$, simulating real data; thus, we fitted a bi-variate function $B(i,j)$ on a real image as

$$\begin{aligned} B(i,j) = {} & 349.8 + 185.6\hat{j} - 158.5\hat{i} - 1104.6\hat{j}^2 \\ & + 929.2\hat{i}\hat{j} + 976.4\hat{i}^2 + 2229.5\hat{j}^3 - 1213.7\hat{i}\hat{j}^2 \\ & - 785\hat{i}^2\hat{j} - 1490.7\hat{i}^3 - 1350.3\hat{j}^4 + 454.3\hat{i}\hat{j}^3 \\ & + 530.6\hat{i}^2\hat{j}^2 + 117.2\hat{i}^3\hat{j} + 675.5\hat{i}^4, \end{aligned}$$

with $(\hat{i}, \hat{j}) = (i/256, j/512)$. We also set $F^{ps}{}_{i_l,j_l}$ to be the standard Gaussian distribution centered at the uniformly distributed random indices (i_l, j_l). The minimum distance between (i_l, j_l) is set to be greater than 6. Finally, $g_{i_l,j_l}(k) \in \{0, 1\}$ is a step function with dwelling time distributed as an exponential random variable in each state. Precisely, 100 fluorophores intensity were simulated with mean 1000 and the system error with mean 400. The performance of our

Table 1 Errors in background estimation. The simulation setup is in Section 3.1. Note that the sever maximum error in 5×5 LS occurs when averaging over a region containing other fluorophores. Since fluorophores locate at least 6 pixels apart, the 3×3 LS is free from this concern.

	Maximum error	Root mean square error
Algorithm 1 only	36.0 ± 4.9	15.6 ± 1.1
Algorithms 1 and 2	13.6 ± 1.6	5.1 ± 0.5
3×3 LS	19.5 ± 1.6	5.7 ± 0.1
5×5 LS	57.4 ± 24.3	8.3 ± 3.3

algorithm is evaluated by maximum error and the mean squared error between B and B_{HOSVD} at (i_l,j_l). We compared our algorithm with the commonly used local subtraction[5] ($l\times l$ LS, $2l+1$ refers to the size of the background) at (i_0,j_0), which can be specified as

$$B_{\mathrm{LS}}(i_0,j_0,k|l) = \left(\sum_{i,j\in D_l(i_0,j_0)} I(i,j,k) - \sum_{i,j\in D_{l-1}(i_0,j_0)} I(i,j,k)\right)/8l.$$

$D_l(i_0,j_0)$ is defined in Algorithm 3. The result is shown in Table 1. With 100 replications of simulation, the maximum error obtained by coupling Algorithms 1 and 2 was evaluated to be 13.6 ± 1.6. As contrasted with the signal intensity of 1000, this indicated an error of about 1.3%.

We further compared the performance between HOSVD and LS on real data sets. In this comparison, the size of system error is unknown. We applied both methods to estimate system error underlying 134 manually selected time traces. Fig. 4 shows a linear relationship between estimated values of HOSVD and LS, indicating compatibility between the two methods. Particularly, the estimated system error by HOSVD is most consistent with those from 7×7 LS whereas 3×3 LS usually overestimates the system error. Choosing the size l in local subtraction method is an art: larger l may reduce the point spread function impact but would have the risk of entering another fluorophore's influence circle.

3.2 *Results of Searching FRET Trace Pairs*

Fig. 5 shows an image example of the leakage pairs to demonstrate its feasibility on mapping the coordinates. Fig. 6 gives an example of the searched FRET trace pair. Generally, all the single-molecule FRET analysis methods have the first three sub-steps in common: target fluorophores, perform image registration, and search for the trace pairs. Thus, we set up a task to check the performance of searching the FRET trace pairs from all the fluorophore trace pairs. There were 4421 trace pairs from an experiment data set consisting of 14 movies passing through our algorithm. We let an experimenter who is experienced and rigorous with regards to manual selection screen the FRET traces. The manual selection is based on a more stringent criteria listed as follows:

1. the traces exhibit single-step photo-bleaching
2. the average fluorescence intensity along the trace is constant
3. the traces have a normal signal strength
4. the acceptor channel is fluorescent

Thus, the experimenter selected only 79 FRET trace pairs out of the pool of 4421 (a typical example of such unambiguous FRET traces is shown in Fig. 6). On the other hand, by using the FRET screening step in our algorithm, more than 80% of the 4421 trace pairs were efficiently eliminated, yielding a subset of 583 candidate FRET trace pairs. Seventy-seven among the 79 manually chosen ones were preserved in the subset of 583. We checked those two traces that were selected manually but not by our algorithm to find they could be omitted as their anti-correlations, -0.4 and -0.2, were not significant. This result, summarized in Table 2, demonstrates that our algorithm can shrink the candidate set by a factor of approximately 8.

3.3 *Results of Change Point Detection*

To compare the performance of our CPD method with that of MGHMM, we simulated a three-state system. While CPD needs the threshold for p-value, MGHHM requires inputs of random initials and the number of states. To be specific, let (X_D,X_A) follow a bi-variate three-state Markov model with parameters as follows:

1. Initial distribution $\pi=(1/3,1/3,1/3)$.
2. Transition probability matrix
$$A=\begin{pmatrix}0.99 & 0.005 & 0.005\\ 0.005 & 0.99 & 0.005\\ 0.005 & 0.005 & 0.99\end{pmatrix}.$$
3. Observation distributions of (X_D,X_A) are two-dimensional Poisson distributions with (800,200), (500,500), and (520,480) as the mean values for the three states.

Table 2 Results of selecting FRET trace pairs. There are 14 movies in this data set. The upper row shows the ratio of manually selected targets to the candidate pairs for each movie. The lower row shows the ratio of targets to those pairs selected by our algorithms.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
targets / total traces	10/172	5/204	11/361	8/285	11/268	4/323	3/400	2/335	2/344	7/329	5/335	1/331	3/392	7/342
selected targets / selected traces	10/40	5/29	10/52	8/36	10/44	4/28	3/35	2/41	2/52	7/50	5/44	1/28	3/55	7/49

Their corresponding standard deviations are (28.3, 14.1), (22.4, 22.4) and (22.8, 21.9).

In other words, we simulated a situation in which two compact conformations exist plus an unfolded state with distinct FRET efficiencies 0.5, 0.48, and 0.2, respectively. Furthermore, this transition matrix A would allow the dwelling time to be distributed as a geometric random variable with mean 100 units. We would like to test whether or not these algorithms can distinguish these two conformers. Figs. 8(a) and 8(b) display some typical histograms based on 100 simulated time traces. Note that two populations of different FRET efficiency are distinguishable by both approaches. Surprisingly, as the difference was made smaller by letting the two FRET efficiencies be 0.5 and 0.49, our algorithm appeared to perform better with respect to the resolution [Figs. 8(c) and 8(d)]. The resolution of CPD is increased because it provides more efficient de noising by taking the average of the intensities over the dwelling time. The reason MGHMM fails in this case is because the likelihood does not gain as much as the paid penalty associated with introducing one more state.

We also studied the performance of our CPD algorithm by employing a real smFRET model data set provided by TJ Ha's group at UIUC (http://www.cplc.illinois.edu/). The data contains movies taken from two donor-acceptor labeled DNA samples. One is the DNA-only control, and the other is treated with a protein that can induce the change in FRET efficiency. A total of 15,548 raw time traces were processed by the UIUC's package to select the FRET traces.[5] Those traces were further studied by the MGHMM and the CPD method for comparisons. Fig. 7 displays the histogram results generated by the two methods. MGHMM and CPD both enhance the resolution in the histograms while CPD seems to give slightly sharper modes, consistent with the results obtained from the simulation data (Fig. 8). In summary, the comparison of the performance suggests that our CPD approach can be either better than or as good as the MGHMM in terms of removing the random noise. Most importantly, to perform MGHMM usually requires many tries to escape the trap of local minimum whereas there is no such need for the CPD approach.

4 Discussion

The development of our algorithms has involved implementation of several novel ideas. First, we utilized the knowledge on the smFRET system to model the data and applied HOSVD to clean the system error effectively. Second, we adapted to the leakage spots for registration of the donor-acceptor mapping coordinates. Finally, we employed the characteristic of coincidental change of the donor and acceptor traces to implement the multivariate change point analysis to detect the smFRET events. By using these approaches, automatic computing has replaced human manual efforts, largely if not completely, including the fluorescence bead slide setting,[5,9–11] human inspection on choosing FRET traces, and parameter assignment for the microstatus change detection. Nevertheless, the full automation in the current version of our package falters at the step of selecting meaningful FRET traces because we have implemented only an anti-correlation rule using a less stringent condition in order not to miss any interesting traces. Furthermore, by module segmentation, our algorithm package allows for combination with other software packages. The sharpness of the CPD methods depends on the dwelling time in each state. When the dwelling time is too short, we may not see a clear boundary between states. Thus, our package also includes an option for MGHMM analysis, which applies a model selection criterion to set the boundary. For those interested in the details of various analysis methods for smFRET time trajectories[14,24,25] and their comparisons, we recommend a comprehensive article by Bianco and Walter,[26] which appeared while we were preparing this article.

A wide-field TIR equipped with a CCD has made it possible for collecting single-molecule FRET data in a high throughput manner. Recently, the smFRET method has been applied to various biological processes to reveal dynamic behaviors. These processes include catalytic RNA,[10,27] polymerase-nucleic interactions,[28,29] ribosome translation,[30,31] spliceosome

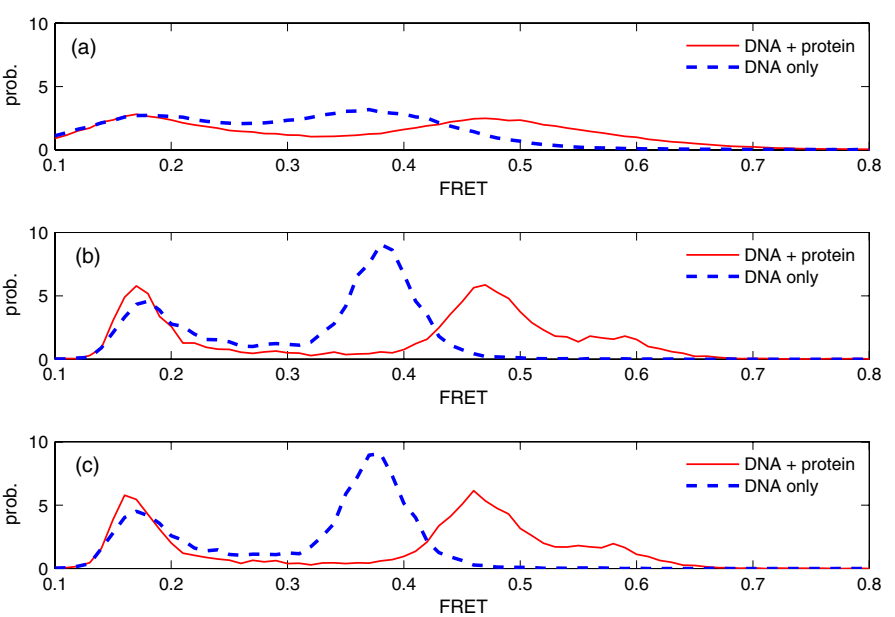

Fig. 8 FRET histogram of simulation data. Parameters of the simulation are in section 3.2, with the FRET efficiencies (0.5, 0.48) in (a), (b) and (0.5, 0.49) in (c), (d). Green, blue, and red lines show the histogram of raw data, MGHMM, and CPD respectively.

assembly,[32,33] vesicle fusion,[34] and the intrinsic protein disorder involved in the process.[35] In addition to revealing the dynamics of biological molecules, smFRET can be employed to study the structure of biological molecules. However, most researchers in this field find it difficult to relate FRET efficiency to a physical distance. Since the anisotropy of the fluorophores' dipoles can be characterized[5] and the quantum efficiency of the two channels can be calibrated,[36] it is now possible to convert smFRET to absolute distance. By obtaining a set of distances between different sites of a biological complex through a large number of smFRET measurements and triangle analysis, the partial structure of the complex might be determined.[17,37,38] As the TIR solution, the dual-view splitter, and the high-sensitivity CCD are available from the market nowadays, it would be easy for an experimenter to set up a wide-field smFRET and generate a large volume of data. Interestingly, very often the time trace from a wide-field smFRET experiment is compared to that from the single ion-channel technique.[39] However, unlike the single ion-channel recording, from which a time trace is reported in real-time, the time traces from wide-field smFRET experiment are obtained through post-imaging data processing. Therefore, to turn the wide-field smFRET microscope into a friendly tool with real-time data analysis potential, it is crucial to improve its data processing efficiency dramatically so that an experimenter can see the time traces immediately at the end of an experimental session to have a clue about what to do for the next. In order to facilitate the post-imaging processing for the smFRET movie data in an online manner, we have created a highly efficient analytical package that combines a series of algorithms that allow for full automation of the work flow of the smFRET data processing.

Appendix: Change Point Detection

This section serves as a very brief introduction of change point detection. At first, a hypothesis test is constructed as

$$H_0: x_i \sim N(\mu, C), \qquad 1 \le i \le N. \quad H_1: x_i \sim N(\mu_1, C),$$
$$1 \le i \le r \quad \text{and} \quad x_i \sim N(\mu_2, C), \qquad r+1 \le i \le N,$$

where $x_i, \mu, \mu_i \in R^p$, and $C \in R^{p\times p}$ is an unknown symmetric and positive definite matrix. Here, $p = 2$ in this case. By likelihood ratio test, the statistic S_r can be derived:

$$S_r = T_r^2/(N - 2 + T_r^2),$$

where r refers to a candidate of a change point and

$$T_r^2 = \frac{r(N-r)}{N}\sum_1^N (x_i - \bar{x}_N)^T S^{-1}(x_i - \bar{x}_N),$$

and

$$S = \frac{\sum_{i=1}^r (x_i - \bar{x}_r)(x_i - \bar{x}_r)^T + \sum_{i=r+1}^N (x_i - \bar{x}_{N-r})(x_i - \bar{x}_{N-r})^T}{N-2}.$$

Because r is unknown, $\max_r S_r$ is our test statistic for the hypothesis. The critical value to reject the null hypothesis can be calculated by a p-value approximation, once the significance size is determined. The approximation has been obtained by Srivastava and Worsley[23] as follows:

$$P(\max_r S_r > c) \le \sum_{r=1}^{N-1} P(S_r > c) - \sum_{r=1}^{N-2} P(S_r > c, S_{r+1} > c)$$
$$\approx 1 - G_{p,v}(c) - q_1 \sum_{r=1}^{N-2} t_r + q_2 \sum_{r=1}^{N-2} t_r^3,$$

where

$$t_r = \left[1 - \left(\frac{r}{N-r}\frac{N-r-1}{r+1}\right)^{1/2}\right],$$

$$q_1 = g_{p,v}(c)\{2c(1-c)/\pi\}^{1/2}\Gamma((N-2)/2)/\Gamma((N-1)/2),$$

$$q_2 = q_1[(p^2-1)/c + (v^2-1)/(1-c) - (p+v)(p+v-1)]/(12(p+v)),$$

$v = N - p - 1$, $G_{p,v}$ and $g_{p,v}$ are cdf and pdf of $Beta(p/2, v/2)$, respectively.

This part is summarized in Algorithm 5.

Acknowledgments

We thank Academia Sinica (AS-99-TP-AB5) and the National Science Council of Taiwan (NSC 98-2118-M-001-022) for the funding to this work. We are thankful for critical discussions with Chi-Fu Yen and Dr. Yen-Chen Lin for OTIR smFRET instrumentation and data analysis and to Dr. Chin-Yu Chen for helping the smFRET DNA experiments.

References

1. T. Ha et al., "Ligand-induced conformational changes of single RNA molecules," *PNAS* **96**(16), 9077–9082 (1999).
2. X. Zhuang et al., "A single-molecule study of RNA catalysis and folding," *Science* **288**(5473), 2048–2051 (2000).
3. P. R. Selvin and T. Ha, Eds., *Single Molecule Techniques: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, New York, p. 507 (2008).
4. M. Tokunaga et al., "Single molecule imaging of fluorophores and enzymatic reactions achieved by objective-type total internal reflection fluorescence microscopy," *Biochem. Biophys. Res. Commun.* **235**(1), 47–53 (1997).
5. R. Roy, S. Hohng, and T. Ha, "A practical guide to single-molecule FRET," *Nat. Methods* **5**, 507–516 (2008).
6. L. S. Churchman et al., "Single molecule high-resolution colocalization of Cy3 and Cy5 attached to macromolecules measures intramolecular distances through time," *PNAS* **102**(5), 1419–1423 (2005).
7. L. J. Friedman, J. Chung, and J. Gelles, "Viewing dynamic assembly of molecular complexes by multi-wavelength single-molecule fluorescence," *Biophys. J.* **91**(3), 1023–1031 (2006).
8. W. P. Ambrose, P. M. Goodwin, and J. P. Nolan, "Single-molecule detection with total internal reflection excitaiton: comparing signal to background and total signals in different geometries, *Cytometry* **36**(3), 224–231 (1999).
9. P. Schluesche et al., "NC2 mobilizes TBP on core promoter TATA boxes," *Nat. Struct. Mol. Biol.* **14**, 1196–1201 (2007).
10. S. V. Solomatin, M. Greenfeld, S. Chu, and D. Herschlag, "Multiple native states reveal persistent ruggedness of an RNA folding landscape," *Nature* **463**(7281), 681–684 (2010).
11. S. J. Holden et al., "Defining the limits of single-molecule FRET resolution in TIRF microscopy," *Biophys. J.* **99**(9), 3102–3111 (2010).

12. D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*, Springer-Verlag, New York (1985).
13. J. Chen and A. K. Gupta, "On change point detection and estimation," *Commun. Stat.—Simulat. Comput.* **30**(3), 665–697 (2001).
14. S. McKinney, C. Joo, and T. Ha, "Analysis of single-molecule FRET trajectories using hidden Markov modeling," *Biophys. J.* **91**(5), 1941–1951 (2006).
15. Y. Liu et al., "A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis," *J. Phys. Chem. B* **114**(16), 5386–5403 (2010).
16. R. H. Goldsmith and W. E. Moerner, "Watching conformational and photodynamics of single fluorescent proteins in solution," *Nat. Chem.* **2**, 179–186 (2010).
17. C. Y. Chen et al., "Mapping RNA exit channel on transcribing RNA polymerase II by FRET analysis," *PNAS* **106**(1), 127–132 (2009).
18. L. D. Lathauwer, B. D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.* **21**(4), 1253–1278 (2000).
19. T. Zhang and G. H. Golub, "Rank-one approximation to high order tensors, *SIAM J. Matrix Anal. Appl.* **23**(2), 535–550 (2001).
20. T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.* **51**(3), 455–500 (2009).
21. L. Omberg, G. H. Golub, and O. Alter, "A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies," *Proc. Natl. Acad. Sci. U. S. A.* **104**(47), 18371–18376 (2007).
22. P. Watkins and H. Yang, "Detection of intensity change points in time-resolved single-molecule measurements," *J. Phys. Chem. B* **109**(1), 617–628 (2005).
23. M. S. Srivastava and K. J. Worsley, "Likelihood ratio tests for a change in the multivariate normal mean," *J. Am. Stat. Assoc.* **81**(393), 199–204 (1986).
24. F. Qin and L. Li, "Model-based fitting of single-channel dwell time distributions," *Biophys. J.* **87**(3), 1657–1671 (2004).
25. J. E. Bronson et al., "Learning rates and states from biophysical time series: a Baysian approach to model selection and single-molecule FRET data," *Biophys. J.* **97**(12), 3196–3205 (2009).
26. M. Bianco and N. Walter, "Analysis of complex single-molecule FRET time trajectories, *Methods Enzymol.* **472**, 153–178 (2010).
27. S. E. McDowell, J. M. Jun, and N. G. Walter, "Long-range tertiary interactions in single hammerhead ribozymes bias motional sampling toward catalytically active conformations," *RNA* **16**(12), 2414–2426 (2010).
28. T. D. Chrsitian, L. J. Romano, and D. Rueda, "Single-molecule measurements of synthesis by DNA polymerase with base-pair resolution," *PNAS* **106**(50), 21109–21114 (2009).
29. S. Liu et al., "Initiation complex dynamics direct the transitions between distinct phases of early HIV reverse transcription," *Nat. Struct. Mol. Biol.* **17**, 1453–1460 (2010).
30. P. V. Cornish et al., "Following movement of the L1 stalk between three functional states in single ribosomes," *Proc. Natl. Acad. Sci. U. S. A.* **106**(8), 2571–2576 (2009).
31. J. B. Munro et al., "Correlated conformational events in EF-G and the ribosome regulate translocation," *Nat. Struct. Mol. Biol.* **17**, 1470–1477 (2010).
32. J. Abelson et al., "Conformational dynamics of single pre-mRNA molecules in spliceosome assembly," *Nat. Struct. Mol. Biol.* **17**, 504–512 (2010).
33. A. A. Hoskins et al., "Ordered and dynamic assembly of single spliceosomes," *Science* **331**(6022), 1289–1295 (2011).
34. H. K. Lee et al., "Dynamic ca2+-dependent stimulation of vesicle fusion by membrane-anchored synaptotagmin," *Science* **328**(5979), 760–763 (2011).
35. U. B. Choi et al., "Beyond the random coil: stochastic conformational switching in intrinsically disordered proteins," *Structure* **19**(4), 566–576 (2011).
36. J. J. McCann et al., "Optimizing methods to recover absolute FRET efficiency from immobilized single molecules," *Biophys. J.* **99**(3), 961–970 (2010).
37. J. Andrecka et al., "Single-molecule tracking of mRNA exiting from RNA polymerase II," *PNAS* **105**(1), 135–140 (2008).
38. A. T. Brunger et al., "Three-dimensional molecular modeling with single molecule FRET," *J. Struct. Biol.* **173**(3), 497–505 (2011).
39. E. Neher and B. Sakmann, "Single-channel currents recorded from membrane denerved frog muscle fibres," *Nature* **260**(5554), 799–802 (1976).