31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Single Molecule Diffusion Coefficient Estimation by Image Analysis of Simulated CCD Images to Aid High-Throughput Screening

Pengfei Song, *Student Member, IEEE*, Lloyd M. Davis, and Gregory R. Bashford, *Senior Member, IEEE*

Abstract—**Extension of one-dimensional signal analysis to two-dimensional image analysis could accelerate conventional methods of high-throughput screening in the discovery of new pharmaceutical agents. This work describes a first step taken towards this goal – the evaluation of image-analysis based estimation strategies of the diffusion coefficient of a single molecule transported within a microfabricated flowcell. A computer simulation of single-molecule imaging by a charge-coupled device (CCD) camera is used to determine if it is possible to distinguish three different types of molecules with different diffusion coefficients. The Gaussian fitting algorithm finds the variance of the transverse trajectory, which increases linearly with the diffusion coefficient; the path analysis algorithm determines the diffusion coefficient from cumulative summation of the squared displacement along the imaged path; the detector area analysis algorithm determines the number of resolvable positions or pixels in the imaged trajectory. Of the three methods, the path analysis strategy appears to provide the most reliable measure of diffusion coefficient with relative error of 13.6% and 6.4% between single molecules with diffusion coefficients of 2.85e-7 and 1.425e-7 cm^2/s. The detector area analysis method can statistically distinguish between single molecules with diffusion coefficients of 5.7e-7 and 1.425e-7 cm^2/s at the $p_{0.05}$ level.**

I. Introduction

High-throughput screening (HTS) methods are currently the focus of extensive research and development in the creation of bioanalytical tools for the rapid discovery of new pharmaceutical agents. Fluorescence correlation spectroscopy (FCS) is a commonly used technique in HTS. The development rate of new pharmaceutical compounds in recent years has greatly accelerated due to the creation of novel protein adaptation methods like combinatorial biosynthesis [1] and directed evolution [2]. However, one of the major roadblocks against efficient drug discovery is the backlog of the large number of potential compounds needing to be screened for their therapeutic potential. Therefore, an obvious need exists for developing new and improved HTS techniques to mitigate this backlog.

Our long-term goal is to accelerate conventional methods of rapid bioanalysis and enable novel, rapid bioanalysis methods by capitalizing upon image analysis methods designed for single-molecule imaging and tracking. This goal is underpinned by the central hypothesis that a substantial increase in throughput and information content for these assay methods will be realized by extending one-dimensional signal analysis of fluorescence counts versus time to two-dimensional image analysis. In this paper, we offer one example of potential bioassay improvement, the analysis of CCD images obtained by a 2D FCS analogue. The FCS analogue was chosen because it is a means of measuring the diffusion coefficient of a molecule and chemical binding kinetics, both of which are valuable information in HTS. For example, in HTS one may aim to quantify the fraction of molecules with different diffusion coefficients, such as small fluorescently labeled ligands and ligands bound to a large protein. The identification of a single molecule from a measure of its diffusion by FCS is subject to large errors [3], but a 2D image provides greater information content than the 1D signal versus time and hence has the potential for enabling more rapid bioanalysis.

Means for a faster rate of travel of molecules through the detection zone are also examined by using electro-osmosis to create bulk solution flow rather than relying only on diffusion. The potential increase in information content and analysis throughput that this technique is expected to provide is appealing and bodes well for substantially decreasing the experimental cost and the time of current HTS methods based on FCS.

In this paper, we compare and contrast three different image analysis methods for estimating the single-molecule diffusion coefficient based on computer modeling. We have created a single-molecule imaging simulation to evaluate the feasibility of single-molecule diffusion estimation within a CCD image. Then we implement the three image analysis algorithms onto a Region of Interest (ROI) manually selected from the images created with this model representing the CCD frame. Three different diffusion coefficients were set up for each of the three image analysis algorithms, and our objective was to compare and contrast each analysis algorithm's capability of separating different image streaks with different diffusion coefficients.

II. Background

A. Random Walk

The path traced by a freely diffusing single molecule in a liquid could be modeled as a random walk. The path length,

P. Song is with the Department of Biological Systems Engineering, University of Nebraska-Lincoln, 23 L. W. Chase Hall, Lincoln, NE 68583 USA (e-mail: psong@huskers.unl.edu).

L. M. Davis is with the Center for Laser Applications, University of Tennessee Space Institute, Tullahoma, Tennessee 37388 USA.

G. R. Bashford is with the Department of Biological Systems Engineering, University of Nebraska-Lincoln, 230 L. W. Chase Hall, Lincoln, NE 68583 USA (e-mail: gbashford2@unl.edu).

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

therefore, is distributed according to a normal distribution. When setting up the starting point of the single-molecule diffusion as the coordinate origin, the diffusion distance has the relationship with diffusion time as below [4]

$$E[r^2] = 2Dt \qquad (1)$$

where $E[]$ denotes the expectation value, r is the distance between the current diffusion position and the origin, D is the diffusion coefficient, and t is the diffusion time. If we consider a series of times

$$t = nT, n = 0,1,2,...,N \qquad (2)$$

where T is the step time, we can cumulatively sum up the squared distance during each step time as below

$$L^2 = \sum_{n=0}^{N} r_n^2 \qquad (3)$$

where $E[r_n^2] = 2DT$. Thus, if we plot L^2 versus the number of steps n, theoretically a straight line with slope $2DT$ would be obtained, by which the diffusion coefficient could be estimated.

B. Previous Work

Single-molecule detection (SMD) has been of interest recently in both medical assays and DNA sequencing [5]. Bunfield and Davis [6] proposed a Monte Carlo simulation of single-molecule detection, which is useful for improving one's quantitative understanding of the trade-offs and limitations that photophysical and instrumental parameters play in the choice of experimental setup, and for optimizing the choice of parameters for a particular SMD application. Xu and Yeung [7] took advantage of the analog-to-digital conversion time of a CCD camera to generate a smeared image of single-molecule emission, by which they realized the direct measurement of molecular diffusion coefficients and unimolecular photodecomposition rates for single fluorophores in free solution. However, attempts to image and track single moving molecules in free solutions have met with limited success due to diffusion of molecules out of the depth of field [8] and within the image plane [7], both of which limit tracking time. These problems originally limited the use of single molecule tracking to large, slowly-diffusion substances such as proteins and viruses [9], giving a method termed single-particle tracking (SPT) [10]. One approach to overcome this limitation is to restrict the volume in which the molecule may diffuse, thus increasing the observation time.

An important difference between SPT and our algorithms is that, in SPT, it is assumed the movement of a particle of interest occurs on a time scale much slower than the image integration time [11]. Our simulated image of a molecule that moves over certain distance of the view is within one frame exposure period, while in SPT, multiple frames are used to locate and track single molecules. Thus, if estimation of single-molecule diffusion parameters could indeed be obtained from a single frame exposure, the potential for accelerating HTS is significantly increased.

Single-molecule imaging and tracking are not trivial and require judicious selection of equipment and experimental parameters. A commonly-used strategy, fluorescence microscopy requires an excitation laser with 10-20 mW output power at the fluorophore wavelength(s). Prior to investing time and equipment in an experiment, it is helpful to first assess the likelihood of successfully imaging a particular type of biomolecule, based on the required experimental conditions and equipment specifications, by using a simulation. Our simulation models the main features of fluorescence imaging, including the photophysical properties of fluorescent molecules, laser-molecule interactions, effects of electrical and pressure fields on these molecules in solution, the use of channels to constrain diffusion (with boundary conditions that assume reflection from channel walls), CCD camera pixilation and read-out noise, and fluorescence collection optics.

III. Materials and Methods

A. Simulation and ROI selection

The simulation (written in MATLAB) includes variables representing molecule diffusion, CCD specifications, flowcell channel dimensions and boundaries, molecule photophysical properties, laser parameters, and emission optics. A typical image (100×100 pixels, microscope NA = 1.2) created with this model representing a single CCD frame is shown in Fig. 1. This model simulates molecules (30-base ssDNA tagged with Rhodamine 6G) freely diffusing within a 20 μm-wide, 100nm-deep flowcell microchannel. A uniform bulk flow is present (500 μm/s, from top to bottom in the image). In the frame, two single molecules are present. One (represented by the track on the right) has moved three-fourths of the way from the top of the frame (about 15 μm) within a 40-ms frame acquisition time and then experienced photodestruction and stopped emitting photons, while the second molecule (on the left) moves outside of the field of exposure during the CCD frame exposure time period. The main function of this simulation is to iterate through a loop, which represents a discrete period of time. Within each iteration or time step, a fluorescent molecule is allowed to move due to uniform flow, electrophoretic forces, and diffusion, while emitting photons according to the interactions of this molecule with an excitation beam of light. These emitted photons fall on the CCD pixels with a probability given by the point spread function (PSF) of the imaging optics. Some photons are not detected due to optical

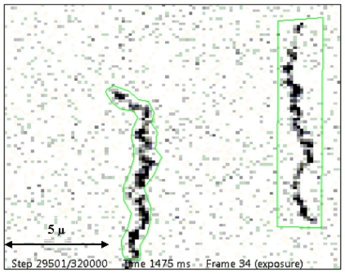

Fig. 1. Model of single-molecule diffusion within a CCD frame, and the ROI selection (green lines around the "streaks"). Each "streak" represents a single fluorescent molecule moving during a frame integration time. Two ROI selection methods were used, one is following the outline of the streak (the left one), and one is drawing a quadrangle to cover the whole steak inside (the right one).

transmission losses and the detector quantum efficiency. Finally, noise counts modeled with Poisson statistics are added to the CCD pixels, to represent read-out noise and background scattering.

Two ROI selection methods were used as shown in Fig. 1. The "following the streak outline" method was designed for the image analysis Algorithm B (see section III.B), which could effectively decrease the interference of noise in computing the centroids of the single-molecule streaks. The "drawing-quadrangle" method was designed for both of Algorithm A and Algorithm C, which could help the analysis algorithm focus on the molecule-activity-related areas and obtain all the information of the streaks.

B. Image analysis algorithms

We designed three image analysis algorithms to estimate the single-molecular diffusion coefficient based on the CCD images.

1) Algorithm A (Gaussian fitting)

This algorithm is based on the Gaussian fitting method, and it was originally inspired by the observation that a smaller molecule with greater diffusion coefficient generally generates a wider transverse streak than that of a larger molecule with smaller diffusion coefficient, as shown in Fig. 2. Hence, this method is to add the pixel values along the column direction inside the ROI and form an 1 × N vector containing the summation information for N columns. Then we use Gaussian fitting to fit this vector and obtain the variance of the fitted curve, which would be used to compare the level of diffusion among different single molecules.

2) Algorithm B (Path analysis)

The diffusion paths length is related to the diffusion coefficient by using (1)–(3). However, it is difficult to compute and pick out the exact paths that the single molecule has passed through only based on the simulation CCD images, so we use this algorithm to reestablish the approximate path of the single molecule by following the centroid of each row inside the ROI. Then we calculate and cumulatively sum up the squared diffusion distance during each step time, plot the accumulated squared distance against the number of steps, and estimate the diffusion coefficient by calculating the slope of the plotted curve. Moreover, the movement caused by bias flow in y-direction was removed from each diffusion distance within each step to obtain the actual single-molecule diffusion.

3) Algorithm C (Detector area analysis)

The third algorithm is based on the hypothesis that the greater the diffusion coefficient, the more detector area (in this case, the detector area corresponds to the number of pixels in the CCD frame) that the single molecule will pass through within a certain period of diffusion. Therefore, we first filter the frame image with a threshold value which has been carefully chosen by visual observation to preserve enough image information and eliminate the noise, and obtain a binary image of the CCD frame; then we sum up the number of pixels with the pixel value of 1 and perform the normalization. Finally we statistically analyze and compare these summation results among the three different diffusion coefficients.

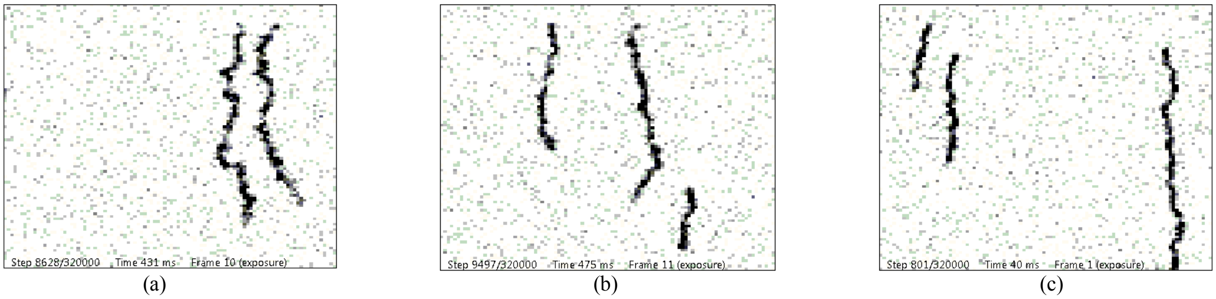
Step 8628/320000 Time 431 ms Frame 10 (exposure)
Step 9497/320000 Time 475 ms Frame 11 (exposure)
Step 801/320000 Time 40 ms Frame 1 (exposure)
(a)
(b)
(c)

Fig. 2. Three CCD frames of the diffusion single molecules with different diffusion coefficients. (a) is the simulation CCD frame of 30-base ssDNA tagged with Rhodamine 6G, with the diffusion coefficient of 5.7 × 10^{-7} cm^2/s; (b) is the simulation of a single molecule with half of the diffusion coefficient as in (a), which is 2.85 × 10^{-7} cm^2/s; (c) is the simulation of a single molecule with one-fourth of the diffusion coefficient as in (a), which is 1.425 × 10^{-7} cm^2/s.

IV. RESULTS

A. Experiment 1: Tests based on Algorithm A

Fifty ROIs (fifty different streaks) were selected for each of the three different kinds of single molecules with different diffusion coefficients, which are the same as the coefficient values in Fig. 2. The distribution of the Gaussian fitting variances is shown as normalized histograms in Fig. 3 (a). The t-test result is shown in Table II.

B. Experiment 2: Tests based on Algorithm B

In experiment 2, the same number of ROIs was utilized to the same kinds of single molecules as in experiment 1. For each kind of molecule, we estimate the ensemble average

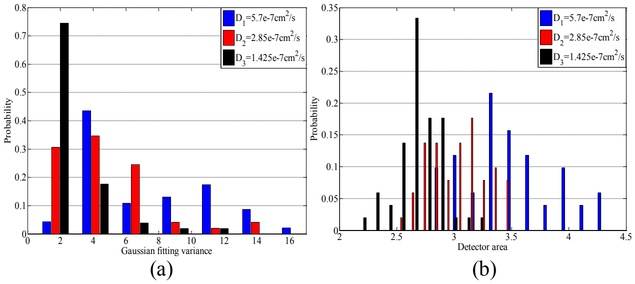

Fig. 3. (a) is the distribution of the Gaussian fitting variances (in units of pixels squared) of three different single molecules. (b) is the distribution of the detector area (in units of pixels per step time, step time = 4e-4 s) of the three different molecules.

$E[L^2]$ of the accumulated squared diffusion distance of fifty ROIs. The results are shown in Fig. 4. Also, the estimation of the diffusion coefficient by calculating the slope of the plotted curve was shown in Table I.

C. Experiment3: Tests based on Algorithm C, threshold=3 pixel value

As the same in experiment 1, fifty ROIs were selected for three different diffusion coefficients. The distribution of the detector area is shown as normalized histograms in Fig. 3 (b). The t-test result is shown in Table II.

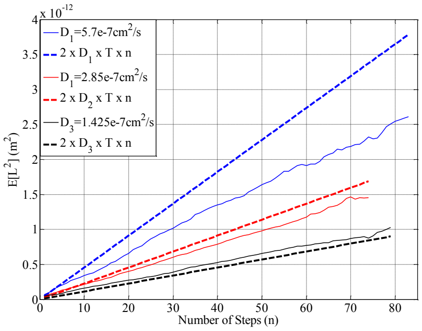

Fig. 4. The solid lines are the actual expectation of the accumulated squared diffusion distance against the number of steps, while the dashed lines are the theoretical curve based on (1). The step time is 4.0×10^{-4} s.

TABLE I
Estimation of the Diffusion Coefficient of Experiment 2

D_1[a]=5.7 × 10^{-7}		D_2=2.85 × 10^{-7}		D_3=1.425 × 10^{-7}	
Slope	3.858e-7	Slope	2.462e-7	Slope	1.516e-7
R. Err. [b]	32.3%	R.Err.	13.6%	R. Err.	6.4%

[a] D_1, D_2 and D_3 are all in the unit of cm^2/s

[b] R. Err. is the abbreviation of relative error

TABLE II
Summary Table of T-test on Experiment 1&3

Exp.1 (Gaussian fitting variances)		D_2		D_3	
		$p_{0.05}$[b]	t	$p_{0.05}$	t
	D_1[a]	1.671	0.521	1.664	0.977
	D_2			1.671	0.601
Exp.3 (Detector area)		D_2		D_3	
		$p_{0.05}$	t	$p_{0.05}$	t
	D_1	1.664	0.942	1.667	1.744
	D_2			1.660	1.014

[a] D_1, D_2 and D_3 are of the same values as in Table I

[b] $p_{0.05}$ is based on an independent two-sample t-test. The null hypothesis is that the mean of one sample is equal to the mean of the other sample, and the significance level is at 5%

V. Discussion

The objective of this paper is to compare and contrast the capabilities of three different image analysis algorithms to separate three kinds of single molecules with different diffusion coefficients. Both experiment 1 and experiment 3 verified the theory and hypothesis of Algorithm A and Algorithm C respectively, by showing that the mean value of the Gaussian fitting variances and the detector area vary with different diffusion coefficients. In experiment 1, the means of the Gaussian fitting variances vary approximately linearly with the assigned diffusion coefficient, but all the t-tests among the different diffusion coefficients failed to reject the null hypothesis at the 5% significance level, which means that none of the three distributions in Fig. 3 (a) is separable with each other. In experiment 3, no linear relationship could be found among the means; however, in the t-test between D_1 and D_3, the t value of 1.744 is greater than the $p_{0.05}$ which is 1.667, which means that the null hypothesis was rejected at the 5% significance level, and the distribution of D_1 and D_3 in Fig. 3 (b) are separable with each other. Thus, an unknown diffusion coefficient generated either by distribution of D_1 or D_3 could be differentiated by Algorithm C. Experiment 2 indicated that the most promising results for direct estimation of the diffusion coefficient are from the slope of the curve. Especially, for single molecules with D_2 and D_3, the relative error between the calculation of the slope and the theoretical value of diffusion coefficient is 13.6% and 6.4% respectively, which suggests a potential method of quantitative estimation of an unknown diffusion coefficient from the CCD images.

One cause of the large standard deviation and less satisfying t-test of the above experiments is the multiple and complicated shapes of image streaks caused by the freely diffusion single molecules. Currently, all types of streaks without specific selections were used in all the experiments in order to obtain as much information as possible, which would also introduce considerable interferences into the system. Another cause, especially for experiment 2, is the uniform choice of the microscope magnification. In Fig. 2, the transversal width of the streak decreases with the reduction of the diffusion coefficient; however, once the lateral path variance stays within a pixel, further decrease in diffusion is harder to detect. Therefore, an improved ROI selection method with better signal-to-noise and directional extraction of ROI, and an adaptive magnification method with better display of the lateral path variance of the streaks will be considered in future work.

References

[1] H. G. Menzella and C. D. Reeves, “Combinatorial biosynthesis for drug development,” *Curr. Opin. Microbiol.*, vol. 10 (3), pp. 238–245, June 2007.

[2] Y. Yoshikuni, T. E. Ferrin, and J. D. Keasling, “Designed divergent evolution of enzyme function,” *Nature*, vol. 440, pp. 1078–1082, Apr. 2006.

[3] J. Enderlein and M. Kollner, “Comparison between time-correlated single photon counting and fluorescence correlation spectroscopy in single molecule identification,” *Bioimaging*, vol. 6, pp. 3-13, Jan. 1998.

[4] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1992, pp. 95–99.

[5] R. A. Keller *et al.*, “Analytical applications of single-molecule detection,” *Anal. Chem.*, vol. 74, pp. 316A–324A, June 2006.

[6] D. H. Bunfield and L. M. Davis, “Monte Carlo simulation of a single-molecule detection experiment,” *Applied Optics*, vol. 37, pp. 12–20, Apr. 1998.

[7] X. H. Xu and E. S. Yeung, “Direct measurement of single-molecule diffusion and photodecomposition in free solution,” *Science*, vol. 275, pp. 1106–1109, Feb. 1997.

[8] T. Kues, A. Dickmanns, R. Luhrmann, R. Peters, and U. Kubitscheck, “High intranuclear mobility and dynamic clustering of the splicing factor U1 snRNP observed by single particle tracking,” *Proc. Natl. Acad. Sci. U. S.*, vol. 98, pp. 12021–12026, Oct. 2001.

[9] U. Kubitscheck, O. Kuckmann, T. Kues and R. Peters, “Imaging and tracking of single GFP molecules in solution,” *Biophysical Journal*, vol. 78, pp. 2170–2179, Apr. 2000.

[10] M. J. Saxton, “Single-particle tracking: the distribution of diffusion coefficients,” *Biophysical Journal*, vol. 72, pp. 1744–1753, Apr. 1997.

[11] N. Destainville and L. Salome, “Quantification and correction of systematic errors due to detector time-averaging in single-molecule tracking experiments,” *Biophysical Journal*, vol. 90, pp. L17–L19, Jan. 2006.