Single-molecule Biophysics: Machine Learning for Automated Data Processing

Junhong Choi[1] and Soomin Cho[2]

***Abstract*— Single-molecule fluorescence microscopy has been greatly successful in understanding biophysics at molecular level. This technique has been advancing toward higher throughput, which creates a need for a data-analysis tool to distinguish molecule of interest from other fluorescence signals. Here, we have used supervised machine-learning approaches to filter biological events of our interest, and present three approaches applicable to different data set size.**

I. INTRODUCTION

Single-molecule Biophysics has led a stride in understanding biological mechanism at the most fundamental level. With an aid of recently developed single-molecule techniques, we can now measure forces exerted by conformation change in a single molecule, distances between target molecules, and hybrid approaches to measure both simultaneously [1]. Among various techniques, a single-molecule fluorescence microscopy (SMFM) method measures molecular dynamics for many molecules simultaneously. Each SMFM experiment usually yields a time-course data for many fluorescing molecule within a microscope field of view. Since many aspects of biology can be modeled through stochastic model, higher throughput of each experiment leads to a high accuracy in measurements for dynamic parameters.

SMFM has a versatile utility to be used in understanding many biological mechanisms. One of such area is a translation, a process carried out by a ribosome that decode messenger RNA (mRNA) to synthesize protein in all organisms [2]. During translation, ribosome needs to move through mRNA from one codon, three mRNA bases that are mapped to one amino acid of long protein chain, to the other through series of small steps. Ribosomes carry out these steps through corresponding conformation. By using a labeled ribosome to probe this conformational change through the intensity of fluorescence signal, Puglisi and coworkers discovered underlying mechanisms of canonical and non-canonical decoding by ribosome [2, 3].

Since dynamic for each steps of translation are probabilistic, a large number of sample is needed to provide accurate measurements. Recently, the Puglisi lab has utilized zero-mode waveguides technology to improve a throughput of assay from order of 1,000 to 100,000 [4]. As a throughput of assay increases, a need to filter unwanted signal and retain molecules of interest arises. With a small number of observed molecules, it is possible to visually pick signals that are corresponding to biological event based on control experiments. However, in the order of 10,000 to 100,000 molecules per experiment, manual data processing becomes quickly infeasible. Using machine-learning approach, an automated data processing can increase the throughput of data processing. Since a classification label for each molecule is easily provided by visual inspection, we used supervised learning algorithms to take advantage of an accessible information. After quick implementations, we found three different algorithms that can be used to maximally help filter out unwanted data depending on the size of experiment.

II. METHOD

A. *SMFM Experimental Setup for acquiring dataset*

For each experiment, one end of mRNA strand (5' end) was labeled with Biotin and immobilized on the microscope slide surface containing Neutravidin through Biotin-Neutravidin chemistry. A ribosome small subunit (30S) was labeled with Cy3B (green fluorescing molecular dye), and formed pre-initiation complex (PIC) before experiment. After the start of data acquisition through charge-coupled device (CCD) camera, a ribosome large subunit labeled with BHQ-2 (quencher molecule corresponding to Cy3B) along with necessary elongation factors were injected into a microscope slide. During the translation of mRNA, small and large subunit of ribosome undergoes series of conformation changes that can be monitored through intensity level of Cy3B. [2, 3, 4, 5]

B. *Preprocessing of Data*

To train and test, we preprocessed raw imaging data resulted from four experiments, and used two for training and two for testing. During experiment, fluorescently labeled ribosomes were immobilized on a surface of microscope slide, and illuminated by one laser. Movie was collected using CCD camera after optically filtering out excitation laser. Due to spreading of the fluorescence spots from each image, we identify each spot and sum up over 4-by-4 pixels (Fig. 1.). Then, we calculate the background of the image to remove, and scale a signal from each molecule from 1 to 100. In a final data structure, we get an m-by-n matrix with m to be a number of frames (usually 10 frame per seconds and 4,800 frames for 480 seconds movie), and n to be number of molecules recorded [5].

C. *Feature Selection*

To select appropriate features, we visually labeled data used for training and testing, and separated out wanted and unwanted signal for each data. Then, we compared

[1] Department of Applied Physics, Stanford University, Stanford, CA 94305, USA

[2] Department of Statistics, Stanford University, Stanford, CA 94305, USA

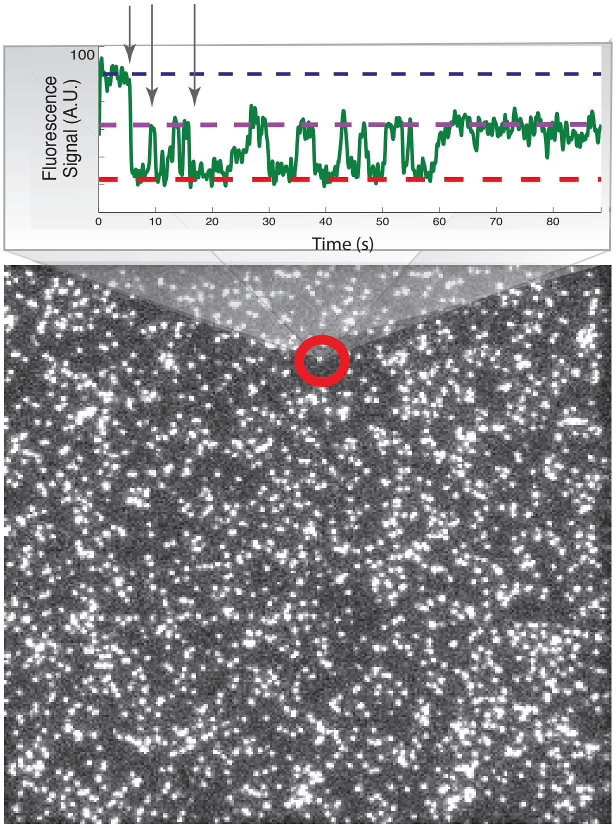


Fig. 1. Preprocessing of data: We obtain a movie from the microscope, and look for a fluorescence spot for an immobilized molecule. After preprocessing steps detailed in text, we extract a scaled time-course signal for each molecules. (Top) This panel shows a signal from wanted molecule, which has three intensity states. The first state (Blue horizontal line) signifies before the assembly of full translation complex. The second state (Red line) signifies waiting of transfer RNA (tRNA) to decode mRNA information, and the third state (Purple line) corresponds to the binding of tRNA and conformation change of ribosome to decode and synthesize protein with adding one amino acid per a codon at a time. Top arrows show where specific events happen. (Bottom) this is raw image acquired using our experimental setup.

distribution of each features between wanted and unwanted data to choose 10 features in total. For the first two, we simply calculated average and variance of signal. For the rest of eight features, we used differentiation of signal to identify rapid change in signal, which we achieved through convolution using a Prewitt filter well used for edge detection in one dimensional image (Fig. 2.) [6]. For the third feature, we looked for when a minimum value within the first 300 frames (30 seconds) occurs, which corresponds to the assembly of a whole translational decoding complex after the start of the experiment. For the fourth feature, we looked for when maximum value within 80 frames (8 seconds) after a fall in signal occurs. For the fifth feature, we looked for when minimum value within 80 frames after a surge in signal occurs, and for the sixth feature, we looked for when maximum value within 80 frames occur after the second fall. For the last four features, we used means and variances of gradient of the signal for whole 480 seconds and for 150 seconds. We have optimized each features crudely by comparing their performance.

We visualized features to see that used features induce two different distribution for wanted and unwanted signal, and therefore used features are indeed relevent (Fig. 3.).

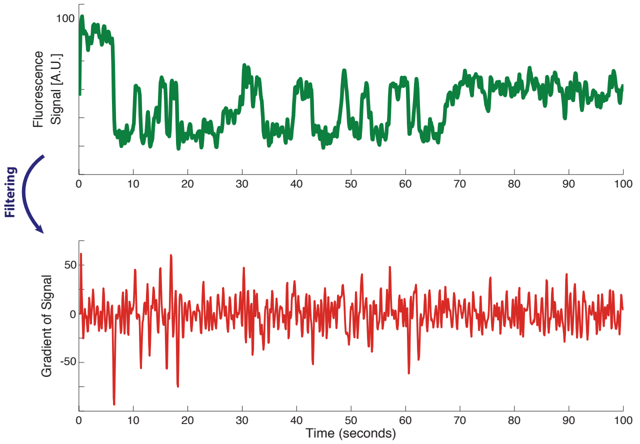


Fig. 2. Feature selection: We calculated gradient of each signal to detect edge of the image. (Top) Fluorescence signal of wanted form after preprocessing. (Bottom) Gradient of signal calculated convolving with Prewitt filter. Time was cut off at 100 seconds to increase visual resolution of underlying structure within signal.

D. *Supervised Learning*

Then, we have used supervised learning package built in MATLAB software (MathWorks) for Support Vector Machine (SVM) model and Naive Bayes (NB) model. To compare with a non-supervising method, we also implemented Exclusion (EX) method. For this method, we looked for a maximum and a minimum value for each feature from training set and retained test samples in which falls within these ranges. This method was also provided a baseline for maximizing true positive (retaining a good molecule from the set) accuracy, which has an increasing importance in a smaller data set. We also devised another hybrid method of a Majority Voting (MV), which compares results from three method mentioned (SVM, NB and EX), and follow a majority decision for all molecules.

III. RESULTS

Using 10 features derived from 5,000 molecules of 4,800 frame time-course data, we achieved at least 60 percent accuracy in labeling using any methods for the two criteria, true positive accuracy (TP, labeling of wanted molecule to be wanted) and true negative accuracy (TN, labeling of unwanted molecule to be unwanted), for 2,000 test molecules and 3,000 training molecules. Each method used has different trade-off between TP and TN accuracy. Using NB, we could filter out up to 90 percent of the unwanted test data by losing up to 30 percent of wanted data. Using SVM, we could retain around 90 percent of wanted data, comparable to EX method, while successfully filtering out 60 percent of unwanted data. We also changed regularization parameter

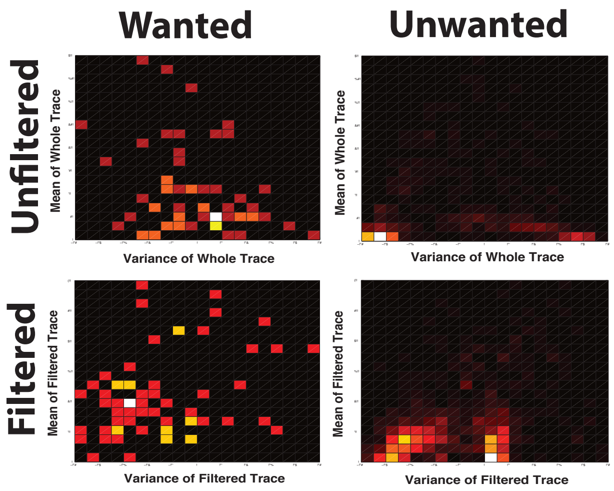


Fig. 3. Feature Validation: For each features, distribution of features of wanted and unwanted signals are compared to see relevence of used feature for classification. In this figure, we compared first two features on top two histograms, and last two features on bottom two histograms.

for SVM, which resulted in trade-off between retaining of wanted data and filtering of unwanted data. EX method was good at retaining data (90 percent), but not great in filtering (40 percent), as we have expected. Performance of MV method resulted compromise between SVM and NB, with 80 percent retaining and 70 percent filtering. We have also tried other hybrid schemes with different weighting from three methods, but their performances were similar to presented four methods here. In the end, we have three methods with distinctive performance that can be used depending on the size of data.

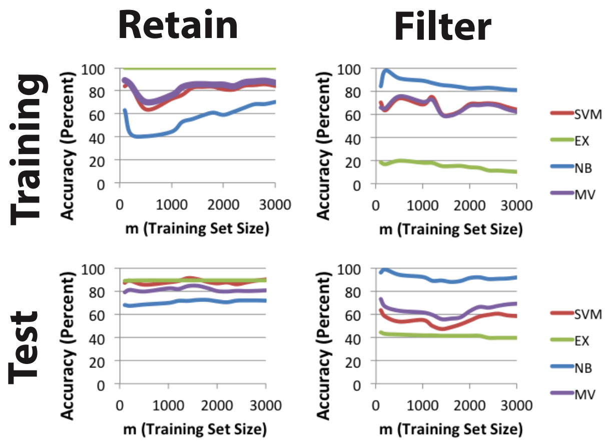
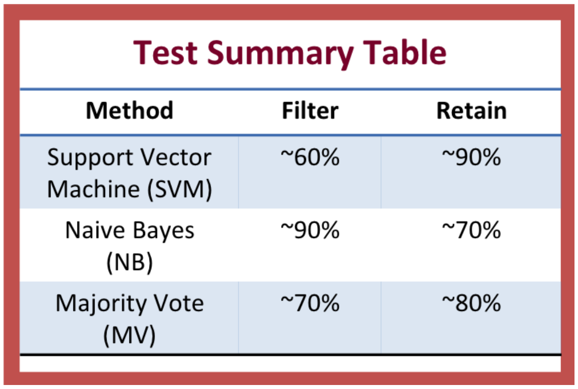


Fig. 4. Result from learning: We used 3,000 molecules for training and 2,000 molecules for testing. Training and Testing showed similar trend with respect to training data size and for each methods used. NB method is the best in classifying unwanted molecule, and SVM is the best in classifying wanted molecule. In general, there is trade between retaining and filtering accuracy, which is conceptually correct; more aggresive the filter is, it will likely to filter out wanted material as well. EX method seems to be similar in retaining ability with SVM (for testing set), but SVM is better at filtering. Excluding EX method, we present three methods (NB, SVM, MV) to be used in different circumstances depending on the size of data set and how aggressive filtering we want it to be.

IV. Discussion

Simple implementation of a supervised machine-learning approach performed considerably well, with one particular method resulting in nearly 90 percent filtering. Since each step of ribosome conformation changes is distributed exponentially, a boundary between wanted and unwanted data is not a clear cut. Due to this inseparable nature of data, a visual recognition method using neural networks would have required a huge set of training data. In our method, training data set size did not matter much, although with a larger training set resulted in the stabilization of the performance. Other machine learning approaches such as the k-mean clustering or PCA would have not performed as well, as the structure of signal is probabilistically distributed and preprocessing using scaling would have resulted in very different temporal structure as well as structure in intensity level for each ribosome conformation, leading to a continuous distribution of time-course signal. The supervised learning algorithm utilized labeling information resulted from visual inspection and features designed to match visual processing of data, which was the most reasonable and appropriate for this project.

Approaching the problem with multiple methods within a supervised learning and combining results, we have achieved three different policies that can be used depending on the size of the data set. If a data set size is in a scale of 10,000 or more molecules, the most aggressive method of NB would filter out most of the unwanted data (in general, for 10,000 molecules, around 8,000 are not wanted and 2,000 are wanted, and 500 wanted molecules give a good estimate of kinetic parameters within experiment), while giving enough wanted molecules to be analyzed further for accurate measurements. For the data set size of 5,000, the most conservative method, SVM, would retain most of the wanted data while filtering more than 50 percent of unwanted data. For the data set size in between 5,000 to 10,000, a hybrid approach of MV would result in a good balance between filtering and retaining.


Test Summary Table

Method	Filter	Retain
Support Vector Machine (SVM)	~60%	~90%
Naive Bayes (NB)	~90%	~70%
Majority Vote (MV)	~70%	~80%

Fig. 5. Summary table of different methods on testing set

V. Conclusions

In this project, we implemented supervised learning approach to automate data processing in a single-molecule experiment. Extracting relevent features from time-course data required imagination and logical steps in recognizing wanted shape in one-dimensional signal. Using ten of such features, We have presented three different models that can be used in data processing of certain data set sizes.

VI. FUTURE

In a future endeavor, increasing feature spaces with finding more relevant features might improve overall performance. Our next goal is to improve retaining to 95 percent, while filter out nearly 70 percent, which can be used in practical setting immediately independent of data set size.

ACKNOWLEDGMENT

We thank the member of the Puglisi laboratory for useful discussion, especially Jin Chen and professor Jody Puglisi. We also thank professor Ron Dror for helpful advise. Lastly, we thank NIH for funding experiments needed in this project.

REFERENCES

[1] T. J. Ha, ”Single-molecule methods leap ahead,” Nature Methods, vol. 11, pp. 1015-1018, 2014.

[2] C. E. Aitken, A. Petrov, and J. D. Puglisi, ”Single ribosome dynamics and the mechanism of translation,” Annu Rev Biophys, vol. 39, pp. 491-513, 2010.

[3] J. Chen, A. Petrov, M. Johansson, A. Tsai, S. E. O’Leary, and J. D. Puglisi, ”Dynamic pathways of -1 translational frameshifting,” Nature, vol. 512, pp. 328-32, Aug 21 2014.

[4] J. Chen, R. V. Dalal, A. N. Petrov, A. Tsai, S. E. O’Leary, K. Chapin, et al., ”High-throughput platform for real-time monitoring of biological processes by multicolor single-molecule fluorescence,” Proc Natl Acad Sci U S A, vol. 111, pp. 664-9, Jan 14 2014.

[5] J. Chen, A. Petrov, A. Tsai, S. E. O’Leary, and J. D. Puglisi, ”Co-ordinated conformational and compositional dynamics drive ribosome translocation,” Nat Struct Mol Biol, vol. 20, pp. 718-27, Jun 2013.

[6] J.M.S. Prewitt ”Object Enhancement and Extraction” in ”Picture processing and Psychopictorics”, Academic Press, 1970.