K-Means Clustering of States in Dual-Beam Optical Trap Assays

Chao Liu

December 14, 2012

Abstract

Dual-beam optical trap is a common biophysics tool to study proteins on the single-molecule level. It has been used to characterize actin-based myosin function which forms the basis for muscle contractions. Data consisting of time-position traces obtained from the assay is currently manually analyzed to discriminate the two different states of the actin-myosin system [1]. In order to improve accuracy and efficiency, the unsupervised *k*-means clustering algorithm is applied to this classification problem using several features derived from the time-position traces. Different weighted combinations of features are evaluated to give the best performance. The algorithm correctly clusters simulated data with precision 0.97 and recall 0.96 and also performs very well on real data.

1 Introduction

At the molecular level, muscles contract when the motor protein myosin binds to actin and exerts a pull. The biophysical study of actin-myosin interaction on the single-molecule level can be performed using the dual-beam optical trap assay [1]. As shown in Figure 1(a), two micron-sized dielectric beads in solution are trapped under the electric potential of two highly focused laser beams and are connected to each other by a taut actin filament. They undergo Brownian motion while being confined in space by the traps' potential. When a myosin immobilized to the surface binds to the actin filament and executes a stroke along the filament, the attached beads experience an additional force and are displaced from their original position. Over time while the actin-myosin system switches between bound and unbound states, detectors track and record the position of each bead. Various factors can be then extracted from this data, including the length of time myosin is bound to actin and the step size taken by the myosin motor. These factors are important both in understanding normal myosin behavior and in characterizing changes seen in diseases such as cardiomyopathies caused by mutations in cardiac myosin genes [1, 2].

The critical data analysis step for dual-beam optical trap assays is thus the discrimination of bound and unbound states in the recorded traces. Difficulties arise from low signal-to-noise ratio because displacement due to Brownian motion is on the same order as that due to myosin binding and stroking (~10nm) so that short binding events may be misidentified as noise and vice versa. Moreover, states are currently assigned manually [1], a very time-consuming process which also produces large variability in the interpretation of the data by different human analyzers.

The purpose of this project is then to improve the accuracy and efficiency of classifying myosin-actin bound and unbound states in dual-beam optical trap assays by automating the interpretation of data using *k*-means clustering. This unsupervised learning algorithm exploits multiple features extracted from the recorded trace data, including mean displacements for each bead, standard deviation of displacements, signal to noise, and bead-bead correlation. Performance is first assessed on simulated data along with feature weights selection, then on real data taken in assays of actin filament and cardiac myosin.

2 Data

Although a large amount of data from dual-beam optical trap assays of the actin-myosin system already exists, there is much variability in the manual assignment of states due to various factors in the data not completely understood, so that the "correct labels" are not entirely known. Therefore it is informative to test a learning algorithm on cleaner simulated data first.

Two tethered Brownian motions representing displacements of the two trapped beads in a dual-beam optical trap assay are simulated as prescribed in [3], with several modifications. First, the steps of the two random walks are sampled from a bivariate Gaussian distribution with nonzero covariance in order to incorporate the correlation in displacements of two connected beads [1]. The variances of the two displacements are assumed equal but can be readily adjusted. Second, in the bound state, the values used for both the covariance and variances are lowered as observed in real data [1]. In addition, a fixed displacement representing the stroke taken by myosin is added to each bead's Brownian motion during the bound state. This simple addition does not entirely capture the physical details but should suffice for the purpose of simulation to test the learning algorithm. Finally, the durations of bound and unbound states are sampled from two exponential distributions with means expected of the waiting times of myosin unbinding and binding, respectively. The much longer average duration of the unbound state is mainly due to the diffusion limited rate of myosin binding. Key parameters used are summarized in Table 1. Several hundred bound and unbound events are generated. A short segment of simulated data is shown in Figure 1(b).

One set of real data containing roughly 60 binding and unbinding events taken from a dual-beam optical trap assay of actin filament and cardiac myosin is provided by Professor Jim Spudich's lab in Stanford's Biochemistry department. A short segment is shown in Figure 1(c). Note that here the displacement induced by the bound state is negative and, more importantly, that it is difficult to consistently manually assign states due to the much greater complexity present in real data but not in simulated data.

3 Methodology

The generated displacement-time data of the two beads is first evenly binned in time so that five features can be calculated for each bin: mean displacement of bead 1, mean displacement of bead 2, standard deviation of bead 1 displacement, standard deviation of bead 2 displacement, and correlation between bead 1 and 2. Two additional features, signal-to-noise ratio (SNR) of bead 1 and SNR of bead 2, are then calculated for each bin as the mean displacement divided by the standard deviation. A choice of thirty to fifty raw displacement data points per bin easily results in several thousand bins, or examples, depending on the amount of data used. Thus the training set consists of several thousand examples, each given as a seven-element feature vector.

The *k*-means clustering algorithm is then applied to the examples, with $k = 2$, to group them into two clusters representing the bound and unbound state. The two cluster centroids are initialized by setting their values equal to the features of two randomly chosen examples. Each iteration of the algorithm then (i) assigns examples to the closer cluster centroid and (ii) moves each cluster centroid to the mean of the points assigned to it. Convergence is achieved when the cluster centroids no longer change significantly.

In applying *k*-means clustering to the dual-beam optical trap problem, a few adjustments are made. Before calculating the distance between an example and a cluster centroid in step (i) of each iteration, each feature, whose magnitude may be on a very different scale from that of other features, is first normalized by its standard deviation across all examples. This prevents uncontrolled bias of cluster assignment based on large-valued features. After this rescaling, each feature may be assigned a weight based on empirical knowledge of how informative it is to cluster discrimination. The same normalization and weighting of a feature are applied to all examples.

After *k*-means clustering has grouped the data into two clusters, the labels "unbound" and "bound" are assigned simply based on prior knowledge of each state's characteristics.

4 Results

Results of *k*-means clustering of simulated data are given in Figure 2. Figure 2(a) shows the clusters found using only the features mean displacement 1 and 2, giving precision 0.91 and recall 0.99. False positives occur at the boundary of the two states, caused partly by the closeness of clusters and by the highly skewed cluster sizes, showing that using only the mean displacements, essentially thresholding, to perform clustering is not sufficient. Table 2 presents the result of using various different weighted combinations of the seven features. This reveals that the means and SNR's are very relevant features yielding high accuracy as compared to standard deviations and correlation. However, the best performance with both high precision and recall is achieved by a weighted combination of all seven features, with means and SNR's having the largest weights. Figure 2(b) displays the clusters found using this optimal weighted combination, producing precision 0.97 and recall 0.96.

Results of *k*-means clustering of real data are given in Figure 3. Figure 3(a) shows the clusters found using the optimal weighted combination of all seven features as determined in testing simulated data. Despite the complexity of real data, the learning algorithm has successfully discriminated two states by exploiting the features. Precision and recall are not calculated because the correct labels are not known unless a human subjectively assigns them. Nevertheless, a qualitative inspection of the assignments by the learning algorithm, given by a plot of a short segment of raw displacement-time trace in Figure 3(b), suggests that the correct classifications have indeed been made.

5 Conclusion and Future Work

Current manual assignment of bound and unbound states in data taken from dual-beam optical trap assays of the actin-myosin system is inconsistent and time consuming. Therefore an automated data processing is in need. The *k*-means clustering algorithm has been successfully implemented and tested on simulated data and produces high accuracy by exploiting a weighted combination of features available. It is observed to also perform surprisingly well on real data despite the significant increase in the complexity of real data due to various noisy factors ignored in simulation. Based on these results, successful application of this algorithm to real data in future work is promising. To this end, the advantage of unsupervised clustering over supervised learning is its adaptability. Different data sets taken from trap assays can give different characteristics such as standard deviation, displacement due to different stroke direction and size, and other variables. While a supervised learning algorithm restricts these values owing to its learning of specific feature values associated with known labels, the *k*-means unsupervised clustering readily group examples in any given data set.

Acknowledgements

I would like to thank Jongmin Sung from the Spudich group for the helpful discussions and for providing the real data.

References

[1] Sung J, Sivaramakrishnan S, Dunn AR, Spudich JA. (2010). Single-molecule dual-beam optical trap analysis of protein structure and function. *Methods Enzymol.* **475**, 321-75.

[2] Palmiter KA, Tyska MJ, Haeberle JR, Alpert NR, Fananapazir L, and Warshaw DM. (2000). R403Q and L908V mutant beta-cardiac myosin from patients with familial hypertrophic cardiomyopathy exhibit enhanced mechanical performance at the single molecule level. *J. Muscle Res. Cell Motil.* **21**, 609–620.

[3] Beausang JF, *et al.* (2007). Elementary simulation of tethered Brownian motion. *Am. J. Phys.* **75**, 520-23.

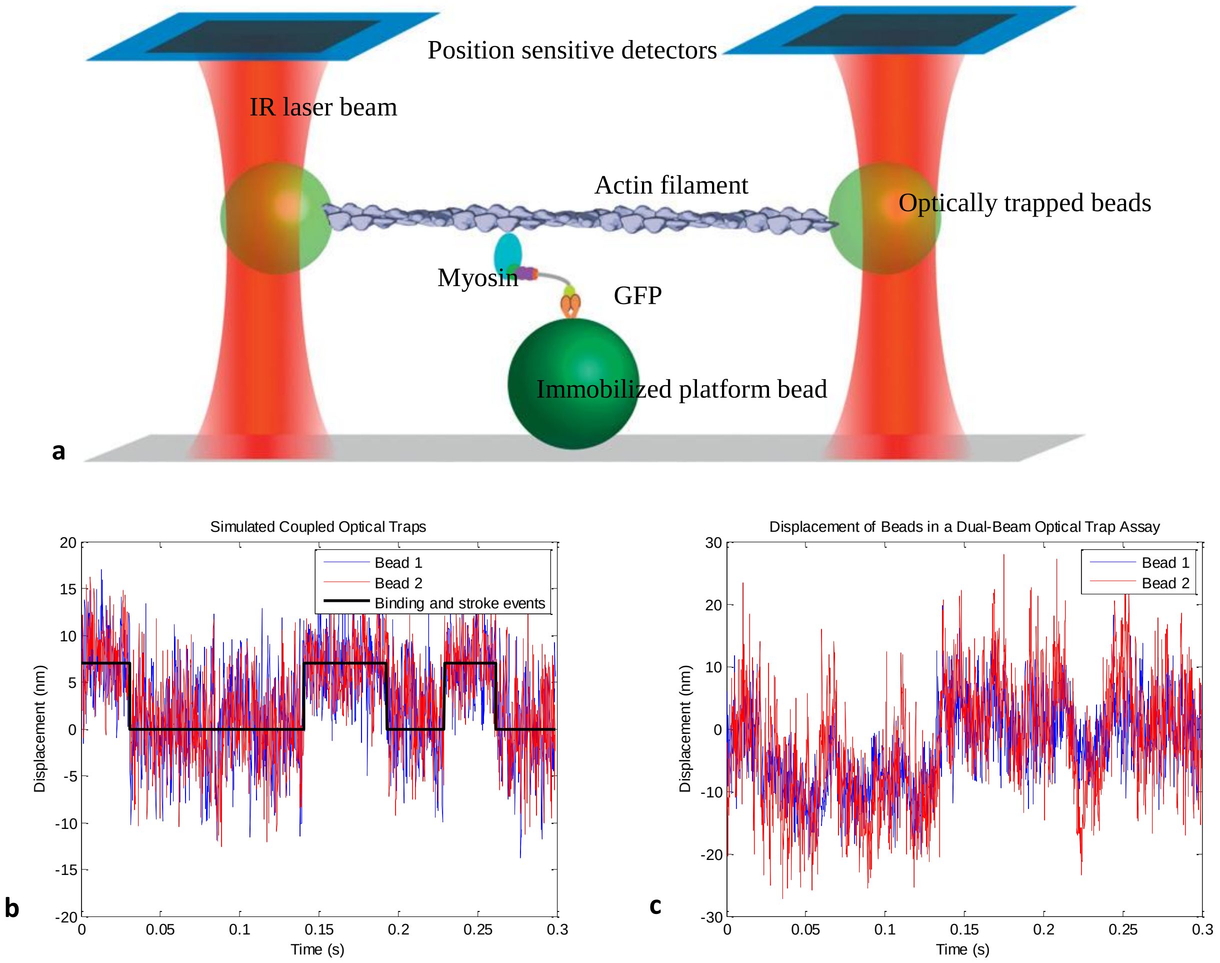

Figure 1. Dual-beam optical trap assay for study of myosin. a) Experimental setup. Drawing not to scale. Figure taken from [1]. b) A short segment of simulated displacement over time of the two trapped beads (red and blue). Binding and stroke events are simulated as positive displacements and shown in black. c) A short segment of real data taken from dual-beam optical trap assay of actin and cardiac myosin, provided by the Spudich lab. Bead 1 and 2 displacements shown in blue and red. Binding and stroke events in this data correspond to negative displacements. Labels of “bound” and “unbound” states are not clear due to the greater complexity of real data.

Stroke size	7 nm	Variance_unbound	1
Mean unbound time	0.1 s	Covariance_unbound	0.4
Mean bound time	18 ms	Variance_bound	0.5
Brownian bias	0.3	Covariance_bound	0.1

Table 1. Key parameters used in simulation of two coupled tethered Brownian motion

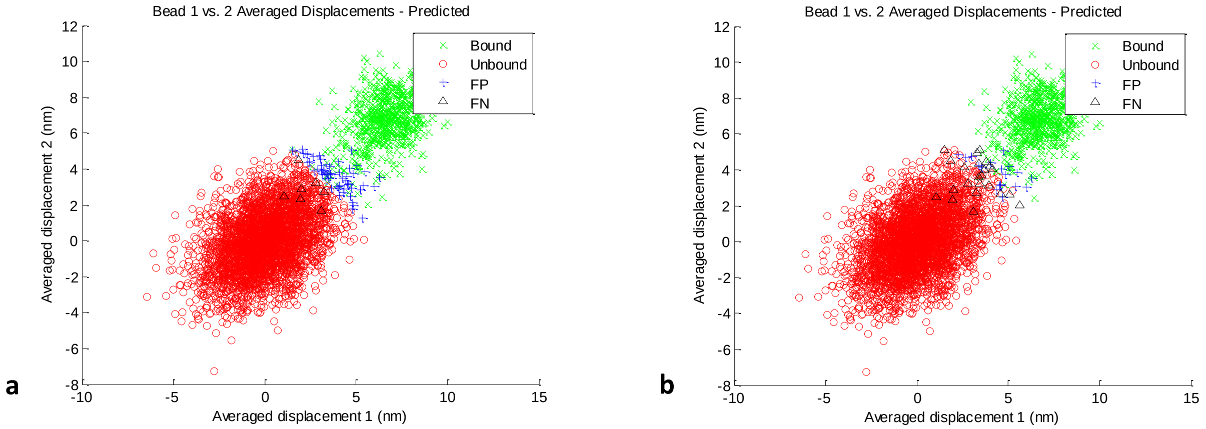

Figure 2. Result of *k*-means clustering on simulated data, displayed by plotting the feature mean displacements. Bound in green, unbound in red, FP (false positive) in blue, FN (false negative) in black. a) Using only mean displacements of bead 1 and 2. Precision = 0.91. Recall = 0.99. b) Using all seven features, weighted, where $w_{\text{mean}} = 1$, $w_{\text{sd}} = 0.3$, $w_{\text{snr}} = 0.7$, $w_{\text{corr}} = 0.2$. Precision = 0.97. Recall = 0.96.

w_{mean}	w_{sd}	w_{snr}	w_{corr}	Precision	Recall
1	0	0	0	0.91	0.99
0	1	0	0	0.25	0.89
0	0	1	0	0.99	0.92
0	0	0	1	0.23	0.58
1	1	1	1	0.97	0.91
1	0.3	0.7	0.2	0.97	0.96

Table 2. Results of *k*-means clustering on simulated data using weighted features

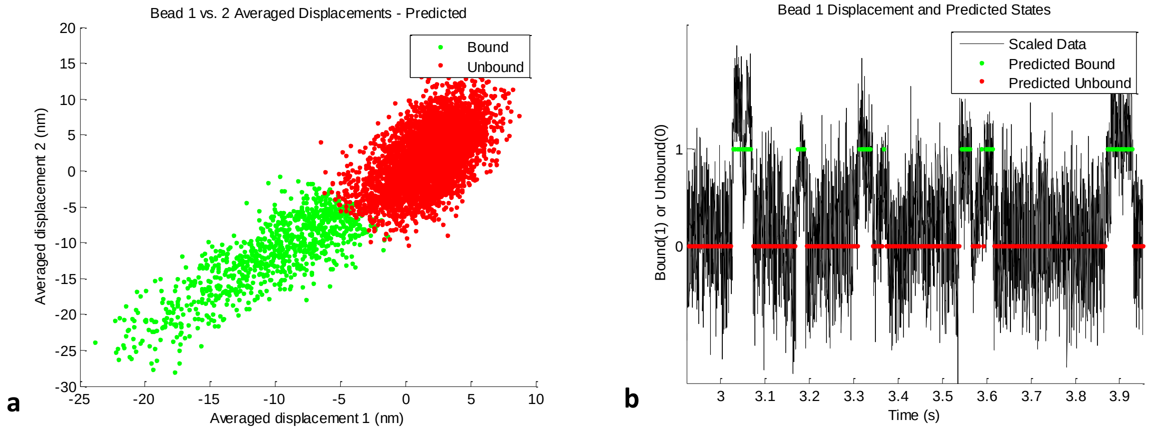

Figure 3. Result of *k*-means clustering on real data. a) Display of clusters by plotting the feature mean displacements. Predicted bound in green, unbound in red. All seven features used and weighted, where $w_{\text{mean}} = 1$, $w_{\text{sd}} = 0.3$, $w_{\text{snr}} = 0.7$, $w_{\text{corr}} = 0.2$. b) A short segment of raw trace of bead 1 displacement with assignment of bound state in green and unbound state in red. Displacement has been scaled for display.