AUTOMATIC METHOD FOR CAVEOLAR STRUCTURE DETECTION AND INTENSITY DISTRIBUTION ANALYSIS FROM MICROSCOPY IMAGES

Harri Pölönen[1]*, Maurice Jansen*[2]*, Jussi Tohka*[1]*, Elina Ikonen*[2] *and Ulla Ruotsalainen*[1]

[1]Institute Signal Processing, Tampere University of Technology
P.O. Box 553, FI-33101 Tampere, Finland
harri.polonen@tut.fi, www.tut.fi
[2]Institute of Biomedicine, University of Helsinki, Finland

ABSTRACT

Fluorescent fusion proteins of caveolin oligomerize to form plasma membrane pits, called caveolae. Amount of caveolin protein in a pit can be estimated by fluorescence intensity of the pit in microscopy image. In this study an automatic method is introduced for pit recognition, intensity measurement and intensity distribution parameter estimation. Dots are recognised and separated from non-caveolar structures. Intensities are measured with a new automatic method, which is capable of estimating intensities from all the recognised pits. Intensity distribution is cleaned up from outliers and modelled with a mixture model of normal distributions. Optimal parameter set of mixture model is searched automatically with a genetic algorithm.

1. INTRODUCTION

Caveolae are plasma membrane pits that form upon oligomerization of caveolin proteins [1]. Each caveola is considered to contain a set number of caveolin molecules (Pelkmans and Zerial [2] have estimated this to be 144 ± 39). These complexes can be recognised in the microscopy image (see Figure 1) when caveolin is tagged with a fluorescent fusion protein. The number of caveolin proteins in each caveolae can be estimated by measuring the intensity of caveolar fluorescence from the image. Caveolae have a tendency to form clusters of two or more pits and therefore intensities form quantal groups according to the number of caveolae grouped together.

Microscopy images from cells are blurry and noisy containing pixel-to-pixel variation, which must be smoothed out. After image processing (e.g. deconvolution) caveolae appear in images as bright and fairly symmetric circular dots of variable sizes. Detecting caveolaes and separating them from background and the determination of caveolae intensity automatically is a challenging task. Some of the caveolae are also located so close to each other that they disturb each others intensity estimation.

A method used for measuring intensity of a caveolar structure is described in [2]. Briefly, five rings are formed around the center pixel of a dot with each ring having one pixel greater radius than previous one. First ring contains just the center pixel and the outmost ring has a diameter of nine pixels. Average intensity within each ring is calculated and average value of the outmost ring representing background is subtracted from all values. A one-dimensional normal distribution is then fitted to these values symmetrically set around center pixel value to represent radial sweep, and the area under the curve is used as an estimate of the dot intensity.

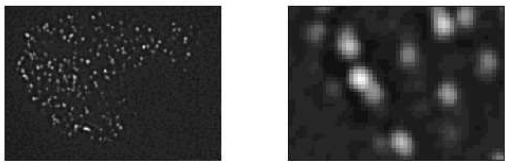

Figure 1: A cell with caveolae and a zoomed detail (pixel size 87×87 nm).

This previous method lacks of capability to handle closely located dots, which has lead into excluding all the closely located dots from the analysis. Thereby the amount of observations in reduced, which decreases the statistical significance of results and deductions. While the number of observable caveolae in an image of a cell is just some hundreds, it is crucial to estimate as many of them as possible. More over, systematic exclusion of certain type of the data (e.g. by the distance to the nearest dot) creates unreliability to statistical deductions.

The intensity distribution estimation poses an optimization problem, which regular curve fitting can't solve efficiently. Fitting a curve is highly dependent on the initial quess and requires therefore manual adjusting and setting. Identifying and separating the quantal clusters and estimating respective curve parameters needs an efficient and exact method, which would increase the quality and reliability of parameters of interest (e.g. proportion of observations in each quantal cluster, widths of clusters). In this study a genetic algorithm is applied for this purpose. The number of quantal cluster present in the observations is estimated automatically.

2. THEORY AND METHODS

The amount of caveolin in a pit can't be observed directly while the pixel size (here 87×87 nm) in a TIRF microscopy image is about the same as the size of the caveolar pit (about 50-150 nm) itself. The amount of fluorescent caveolin in a caveolar structure can be seen as multinormal-like distributed dots in the image. For simplicity, a caveolar structure (single or a group of several) is referred as a 'dot' from now on, and more sophisticated notion is used when needed.

The dots are recognised simply by finding all the pixels which are brighter than their eight-pixel neighborhood. All the highest valued pixels of each caveolae get recognised with a group of background pixels. Background detections are removed automatically as in [3] using Otsu's [4] method

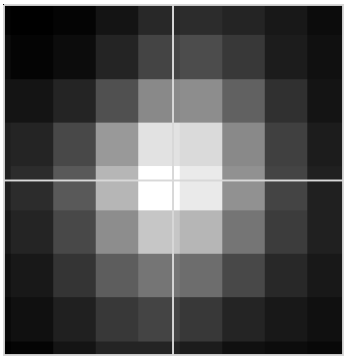

Figure 2: A caveolar structure from a cell divided into quarters according to estimated center location.

to separate potential caveolae areas and background.

If the dots were completely separated in the image i.e. far from each other, the intensity would be easy to measure by summming the pixel values within the dot. However, this method doesn't work with closely located dots but here the following similar procedure is performed. The exact location (center point) of the dot is estimated by fitting a truncated bivariate normal distribution to the nine (three times three) centermost pixels of the dot. Majority of dots in true images are separated enough, that the centermost nine pixels stay mostly undistrubed. The true center location can be estimated as the mean of the best fitting bivariate normal distribution. Finding the best fit can be done for example with a grid search of possible parameters (mean is restricted inside center pixel, variance below a fixed reasonable constant) or with some more advanced method.

According to the estimated true center location, the image is divided into four square-shaped quarters of size four times four pixels and the pixel values falling inside each of the four squares are calculated. (See Figure 2.) Values from pixels partially inside squares are summed only partially according to their area inside the square. Quarter values are multiplied by four to get an estimate for the whole dot intensity. Most of the dots in images are small enough that they fit into this eight times eight pixel area. If there are other dots nearby, the highest quarter value probably contains the highest amount of disturbance and the lowest value of these quarters would presumably to be the least disturbed. However, tests have shown that the second and third highest valued quarters are the most reliable estimators for the whole dot intensity. This is probably because the highest and lowest quarters are also most sensible to error produced by erroneous estimate for the center location of the dot. Therefore, either second or third largest quarter value should be used as intensity estimate.

The obtained intensity distribution needs to be cleaned up of too high values. These merely distract mixture model parameter estimation and they are removed in the following way. Observed intensities are sorted to ascending order and simple Mahalanobis distance (normalized Euclidean distance)

$$D(x_i) = \sqrt{(x_i-\mu)\sigma^{-2}(x_i-\mu)} = \frac{x_i-\mu}{\sigma}$$

is calculated from each observation x_i to the mean intensity μ with variance σ^2. Treshold is set to the first observation x_n

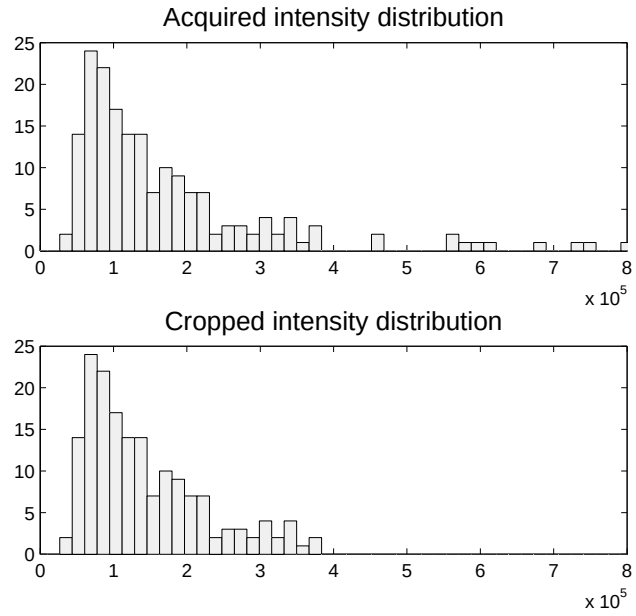

Figure 3: A caveolar structure from a cell divided into quarters according to estimated center location.

with property $D(x_n) > 1 + D(x_{n-1})$ which is removed along with all the consequent observations. In other words, after the increase in distance is greater than one in sorted data, observations are considered as outliers. This method works well, while the observed intensities are mostly in one large stack with just some outliers (see Figure 3 for an example).

Intensity distribution is assumed to be a finite mixture model

$$f(x) = \sum_{i=1}^{n} a_i \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \qquad (1)$$

consisting of n one-dimensional normal distributions with component means μ_i, variances σ_i^2 and mixture parameter a_i (see [5] for finite mixture models). Each normal distribution represents a quantal cluster of caveolae. Fitting a mixture model to the intensity distribution establishes an optimization problem.

The number of components in mixture model, i.e. clusters in intensity distribution, needs to be determined. It is assumed that the components are located with a constant distance from each other because dots in the image are assumed to be either single caveolae or clusters of two or more caveolae. Therefore, mixture models with three to six components are fitted to the data, and the model whose components are located closest to constant intervals is chosen as the appropriate model. The distance of component intervals to constant intervals, with n component means $\mu_1, \ldots, \mu_n$, is measured as a sum of squared errors from the mean component interval

$$C = \sum_{i=1}^{n-1} (\bar{\mu} - (\mu_{i+1} - \mu_i))^2,$$

with

$$\bar{\mu} = \frac{1}{n-1}\sum_{i=1}^{n-1}(\mu_{i+1}-\mu_i) = \frac{\mu_n - \mu_1}{n-1}.$$

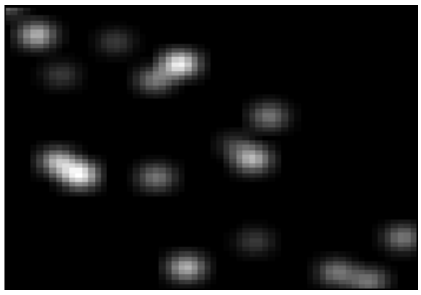

Figure 4: A detail of a simulated image.

With the data used here, three or four component model usually fits best.

The mixture model parameter estimation problem can be solved automatically with a genetic algorithm [6], developed initally for brain imaging applications. Genetic algorithm maximises the likelihood function with respect to the parameter vector, and finds global maximum avoiding local maxima (see reference for more detailed information).

3. DATA AND RESULTS

3.1 SIMULATION

Two simulations were performed to test the accuracy of intensity measurement in controlled conditions. While the true intensity values of simulated dots was known, accuracy of the measurements could be observed.

a) Simulation of constant valued dots without disturbance from neighboring dots was used to test the inaccurateness of a measurement method itself.

b) Simulation with an image with multiple dots was used to test the capability to handle closely located dots.

A simulated dot was sized nine times nine pixels and created from a bivariate normal distribution, whose mean was located randomly inside the center pixel. Dot pixel values are propability distribution function values multiplied with a constant to achieve the wanted total intensity value. In simulation a) 1000 dots were estimated one by one, and in simulation b) the image consisted of 500 randomly located dots in a 400×400 pixel image, allowing partial and complete overlapping of dots.

a) Results of the test with single dots without noise caused by neighboring dots can be seen in the Figure 5 and in the Table 1.

Table 1: Results from simulation with single undisturbed dots with true intensity value 10 000.

Method	Mean	St.dev	CV
Pelkmans&Zerial	9754.3	170.64	1.75%
New	9798.1	24.81	0.25%

The new method developed performs better with this simulation having lower variation, and both methods have quite low bias. Note also that the results of the method by Pelkmans and Zerial [2] are not distributed symmetrically around the mean value and therefore even a normally distributed data wouldn't result as normally distributed estimates, which can pose a problem when estimating intensity distribution parameters.

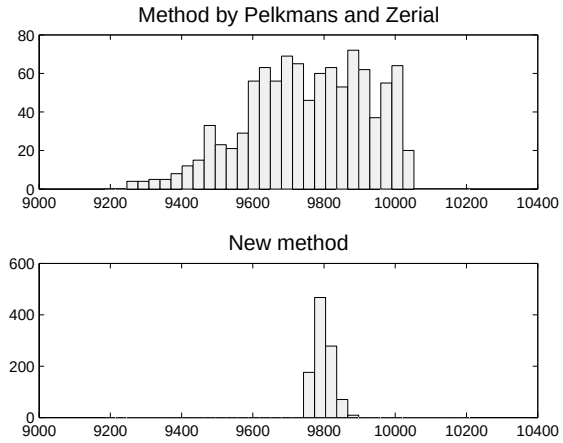

Figure 5: Results from simulation with 1000 single undisturbed dots with true intensity value 10 000.

Table 2: Results from simulation with multiple dots.

		True	Pelkmans	New
Mean	μ_1	10321	8356	10023
	μ_2	19856	16946	19441
	μ_3	30635	26115	30347
St.deviation	σ_1	2467	2912	2576
	σ_2	2815	2544	2915
	σ_3	3409	2839	4479
Mixture parameter	a_1	0.4329	0.4949	0.4424
	a_2	0.4000	0.3376	0.3684
	a_3	0.1671	0.1675	0.1892

b) In the second simulation the purpose is to test the ability to measure tightly stacked dots and clusters. Dot intensity values were created randomly from a mixture model of three normal distributions. Because dots are located randomly throughout the image and some of them are overlapping making estimation more difficult.

Intensity measurement results can be seen in Figure 6 and Table 2. Out of the 500 dots initially created 431 dots were detected, which implies some overlapping. Results show that the new method produces more accurate results. Especially the parameters of the first and second quantal cluster are close to the true ones.

3.2 TRUE MICROSCOPY DATA

A real cell image was acquired with TIRF (Total internal reflection fluorescence) microscopy and deconvolded using Huygens software by Scientific Volume Imaging to reduce pixel-to-pixel noise. Cell were HeLa cells stably expressing caveolin-1 with a C-terminal GFP tag. Dots are detected as earlier, and intensities are estimated with both methods. A proper mixture model is searched for both intensity estimates.

While the true distribution behind estimates is not exactly known, evaluation of the quality of results is more difficult.

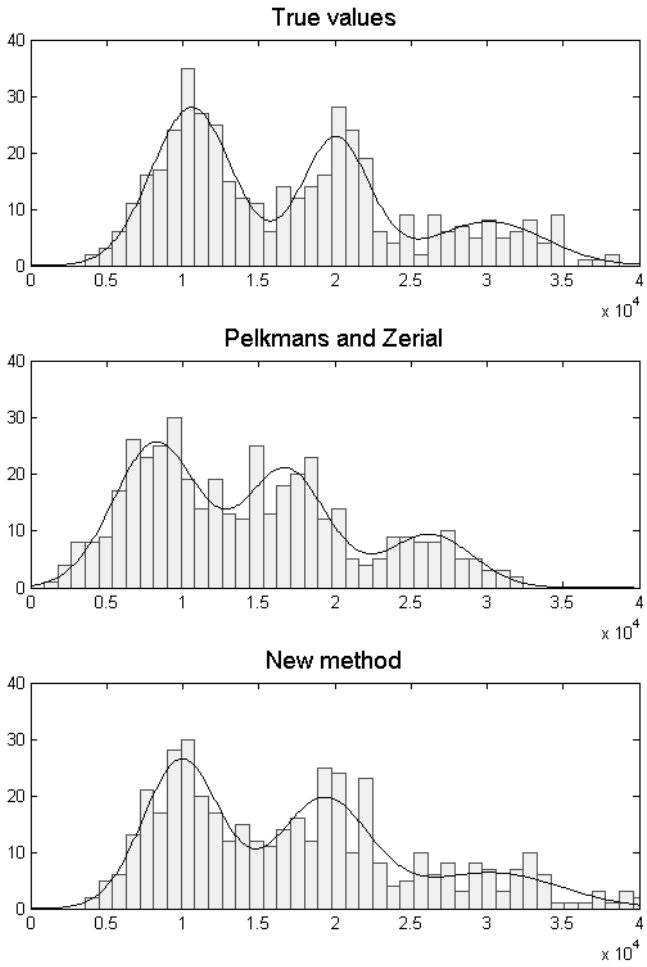

Figure 6: Results from simulation image with multiple dots.

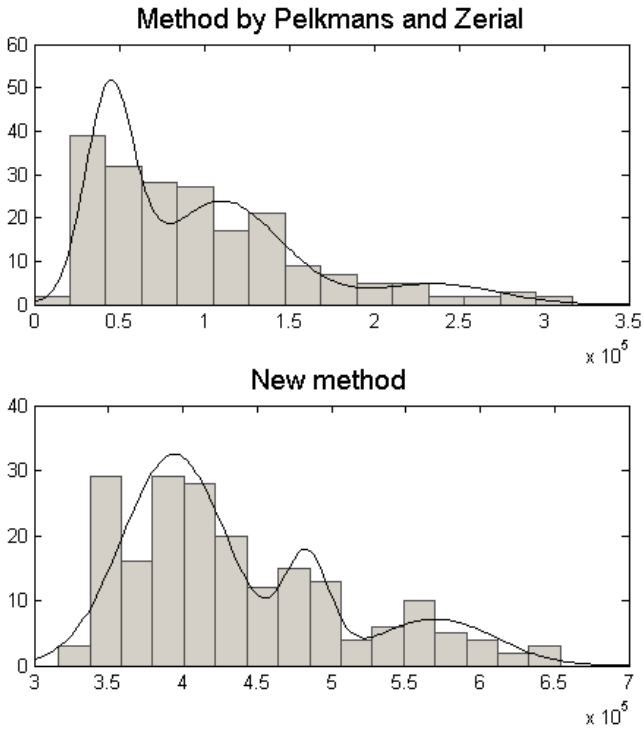

Figure 7: Intensity histograms from a real cell with fitted mixture model.

The intensity distributions produced by both methods with estimated mixture models can be seen in Figure 7. It can be said that the methods clearly produce different results, but it would require a thorough investigation to determinen which method has the more correct results. Int he results from the new method, the quantal clusters are located with constant intervals and they are visually pretty well separable. The genetic algorithm manages works well in both cases in determining the parameter of mixture model.

4. DISCUSSION

The new method developed performs well in simulations. Using Otsu's method in deciding which dots belong to background works well with this problem, as there is no background observations visible in intensity distribution histograms. The intensity measurement method manages to exclude the disturbance effect of neighbor dots, which allows the estimation of every detected dot and therefore increases the reliability of results and deductions. Genetic algorithm works well in finding caveolae intensity distribution parameters, with no initial guess or restrictions needed. Assumed constant intervals of quantal clusters are used as a template for finding the correct number of mixture components, making the intensity distribution analysis automatic. Genetic algorithm is can also be easily tuned for different purposes, if for example prior knowledge of quantal clusters is available or different hypotheses needs to be tested. Other algorithms, e.g. EM algorithm with good initial guess could work pretty well with this simple parameter estimation, but the genetic algorithm is fully automatic and more flexible.

The new method is fully automatic, starting from an image and producing the needed intensity distribution parame-

ters. Automaticness allows to process larger amounts of data, because no manual interference is needed. The method is also pretty fast, which might be important for example with time series image analysis. The larger amount of data that can be analysed and the higher number of dots per image that can be estimated, helps to increase the reliability.

Another solution for dot intensity estimation could be the following procedure. If image was considered as a plane or a surface and the intensity values as a third dimension orthogonal to the plane, then the dots would be as multinormal distributed hills on the surface. A perfect method, i.e. a method which would extract all the information from the image, would then be the one which estimates all individual multinormal distributions straight from the image. Then individual variances, volume below each hill and other parameters could be measured within the limits of some restrictions and each caveolae would get an individual set of parameters. This method however would require an algorithm capable of estimating overlapping multinormal distribution parameter reliably and efficiently, which is problematic with current algorithms. Each caveolae would have six parameters to estimate (two for mean, three for covariance, one for volume). This poses a very complex optimization problem, which has to be solved e.g. with an genetic algorith used here.

REFERENCES

[1] Parton, R. G. "Caveolae and caveolins," *Curr. Opin. Cell Biol.* vol. 8, 542-548. August 1996

[2] L. Pelkmans and M. Zerial, "Kinase-regulated quantal assemblies and kiss-and-run recycling of caveolae," *Nature*, vol. 436, pp. 128–133, July 2005.

[3] A. Niemisto, L. Hu, O. Yli-Harja, W. Zhang, and I. Shmulevich, "Quantification of in vitro cell invasion through image analysis "in *Proc. 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'04), San Francisco, California, USA*, September 1-5, 2004, pp. 1703-1706.

[4] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 6266, Jan. 1979.

[5] G.McLahlan and D.Peel, *Finite Mixture Models*. Wiley Series in Probability and Statistics, John Wiley and Sons, 2000.

[6] Tohka J., Krestyannikov E., Dinov I. D.,MacKenzie Graham A., Shattuck D. W., Ruotsalainen U., Toga A. W., "Genetic algorithms for finite mixture model based voxel classification in neuroimaging," *IEEE Transactions on Medical Imaging*, vol 26, no 5, Page(s): 696-711, 2007.