NSERC USRA Report Summer 2015

Cluster Analysis of Super Resolution Fluorescence Images

Ki Woong Sung

1 Biological Backgound

Direct Stochastic Optical Reconstruction Microscopy (dSTORM) is a novel, high resolution technique to produce fluorescence images. It uses conventional organic fluorophores which are fluorescent molecules that emit light when excited in either a spontaneous or induced way. Common examples of fluorophores are labeled antibodies, used as markers to indicate locations of certain proteins. The fluorophores attached to certain proteins are activated which then "blink" intermittently. Due to the stochastic nature of the fluorophores, only a subset of them are activated at once. Then their positions are precisely determined by fitting Gaussian to each and localizing them, which means taking an image. After localization, they are deactivated to prepare activation of another subset of the fluorophores. Repetitive activation, localization, and deactivation – more than few thousand times – and reconstruction of all the images into one by overlapping them yield 10-fold better resolution compared to conventional fluorescence microscopy [3]. Also, with dSTORM, it is possible to generate 3D or multi-coloured images.

However, there still exists spatial uncertainty which may hinder precise interpretation of reconstructed images. For instance, two close blinks may be due to two separate proteins or two fluorophores on the same protein. To date, the spatial resolution is about 20 nm in the lateral dimensions and about 50 nm in the axial dimension. Thus careful interpretation of the reconstructed image is critical. For this project, B lymphocytes were studied, which make antibodies and are critical in the immune system. The dSTORM data comparing resting and activated B cells were given to be analyzed.

2 Cluster Analysis Methods

2.1 Hopkins Index

Hopkins index, developed in 1954, compares the data to random distribution [1], shown in Equation 1:

$$Hop = \frac{\sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} P_i + \sum_{i=1}^{n} I_i}, \tag{1}$$

where n is the number of data points, I_i is a distance from i^{th} data point to its nearest neighbour, and P_i is a distance from i^{th} random point to its nearest neighbour. The output is a single value ranging from 0 to 1. Hopkins index of close to 1 suggests clustered data of some sort, and 0.5 suggests totally random distribution of data. While this method is quick and simple, Hopkins index only concerns the nearest neighbour and should not be used alone. An extreme example is a group of points along a line or curve – the nearest neighbour distances of all the points are very small if each point is considered separately, but a line of points is not a cluster. Hence, it can only be used as a crude indication of the presence of one or more clusters.

2.2 Ripley's Functions

Ripley's K function seeks number of neighbour points within certain radius r [2], shown in Equation 2:

$$K(r) = \frac{A}{n^2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} I(d_{ij} < r), \tag{2}$$

where n is the number of data points, A is the area of the data space, d_{ij} is the distance between the i^{th} and j^{th} points, and $I(t)$ is the indicator function such that $I = 1$ if t is true and 0 otherwise. With totally random data, $K(r) = \pi r^2$, which means expectation of $K(r)$ increases as r^2. Nonetheless, one major problem with this equation is that points outside the boundary are ignored while they fall within distance r of a point in the data space . To correct for this error, a weight which depends on each pair of points can be multiplied to the indicator function part of the equation, yet this effect is negligible for this project because of large data space and large number of data points. Typically, if the data space is a rectangle, then r ranges from 0 to the half of the smaller side of the data space.

There are some derivatives of the Ripley's K function, which are more commonly used than the K function:

$$L(r) = \sqrt{\frac{K(r)}{\pi}}, \tag{3}$$

$$H(r) = L(r) - r, \tag{4}$$

so that expectation of $L(r)$ is r and that of $H(r)$ is 0. The H function indicates the clusters, if any, most clearly out of the three functions, and is used for this project. A "bump" in the H function suggests a cluster of radius roughly equal to r at the peak, and relative density proportional to the height of the peak.

2.3 Markov Clustering

Markov Clustering algorithm is another simple and fast technique based on simulation of random walk [4]. With a graph – consisting of nodes and edges – as its input, the algorithm identifies clusters by a bootstrapping procedure. It computes a Markov (stochastic) matrix from the graph, uses alternating two operators, expansion and inflation, rescales the matrix to render it stochastic again, and repeats the steps until equilibrium is reached. A stochastic matrix is a square matrix where all the entries are non-negative, and each column add up to 1. Expansion is defined as multiplying a matrix by itself, modelling the spreading out of random walks. And inflation is defined as taking every entry to a power of a inflation parameter I, where $I > 1$, modelling evaporation of walks between different clusters. The MCL will never reach equilibrium if $I = 1$. This parameter controls the granularity of the output clusters: Higher inflation parameter means detecting smaller clusters, and vice versa.

Unlike the two methods mentioned above, MCL algorithm is highly resistant to noise that is unavoidably present in data. Although it categorizes all points, including noise, as being in a cluster, noise that lies far from clusters can be removed after the algorithm by eliminating all clusters with smaller number of points than a threshold. Also, it clearly reveals where the clusters are and how many there are, whereas Hopkins and Ripley's give ambiguous results.

For this project, nodes refer to coordinates of data points, and edges refer to weights between the points. Since higher weights should be assigned to closer points, weights are calculated to follow exponential function so that they always decrease exponentially. Also, the MCL algorithm favours undirected, symmetric matrices and thus input matrices are calculated to be such.

3 Results and Analyses

For comparing resting and activated B cells, only the results were calculated from one cell each, since most cells revealed the same pattern. Based on Hopkins indexes for both types of cells, it can be confirmed that there are clusters present in those cells, as previously determined. In Figure 1 shows that activated B cells have greater mean and median with smaller IQR than resting B cells, suggesting that localizations in activated B cells are closer to each other and the nearest neighbour distances are not as heterogeneous as in resting B cells.

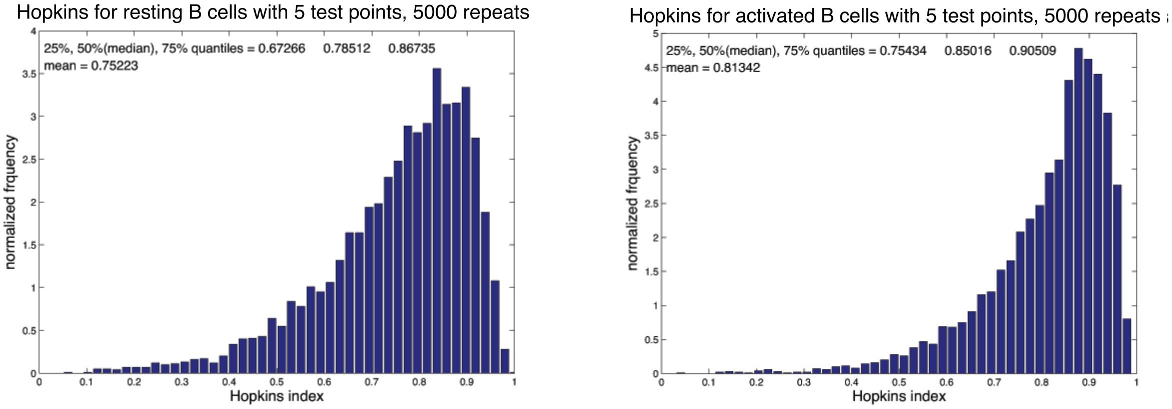

Figure 1: Hopkins index histograms for resting and activated B cells.

Ripley's H functions inform more clearly the presence of clusters of certain sizes. As seen in Figure 2, the peak of the curve is more to the right and up in activated B cells than in the other type. This suggests that in general clusters in activated B cells are larger by roughly a factor of 2, and denser.

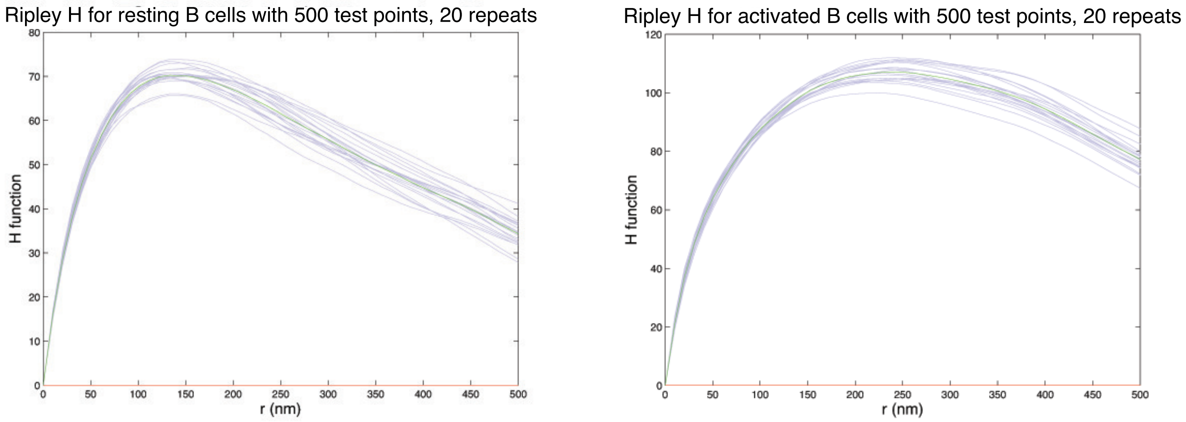

Figure 2: Ripley's H functions for resting and activated B cells.

Markov Clustering confirms the results from the previous 2 methods (Please refer to Figure 3). There are indeed clusters present in both cells, and the clusters in activated B cells are denser and larger. Based on the line graphs, cluster radii can be estimated as follows:

$$r = \sqrt{\frac{area}{\pi}}. \tag{5}$$

Calculating cluster radii for resting and activated B cells using the medians of the approximate cluster areas, $r = 140\ nm$ and $230\ nm$, respectively, which agree with the radii at the peaks of the Ripley's H graphs in Figure 2.

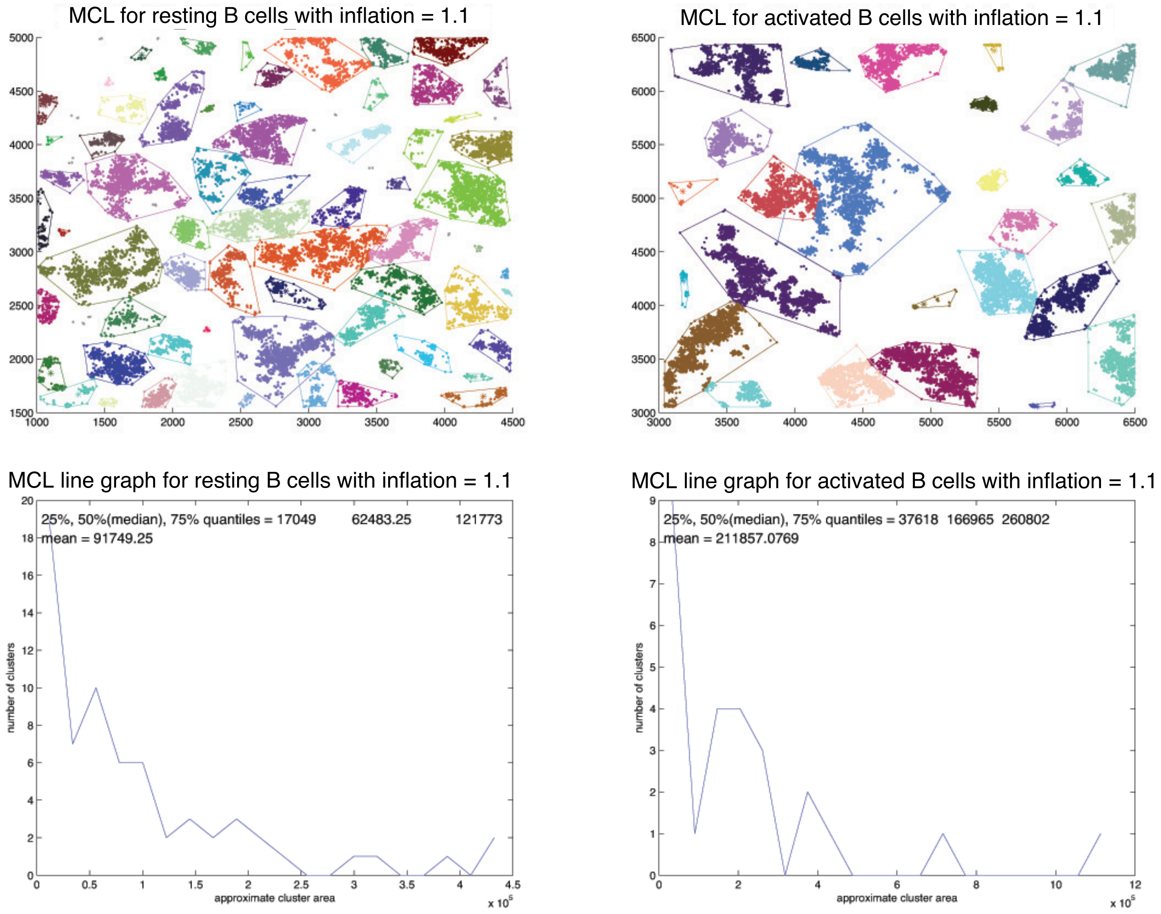

Figure 3: Markov Clustering for resting and activated B cells.

4 Acknowledgement

I would like to thank the NSERC USRA program for funding this project, Daniel Coombs for his supervision along with critical advice, Libin Abraham for general guidance, Edwin Moore for his explanation of biological background, and David Scriven, Josh Scurll, Alejandra Herrera-Reyes, and Reza Tafteh for helping me with technical details.

References

[1] B. Hopkins and J. G. Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2):213–227, 1954.

[2] M. A. Kiskowski, J. F. Hancock, and A. K. Kenworthy. On the use of ripleys k-function and its derivatives to analyze domain size. *Biophysical Journal*, 97:1095–1103, 2009.

[3] S. van de Linde, A. Loschberger, T. Klein, M. Heidbreder, S. Wolter, M. Heilemann, and M. Sauer. Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nature Protocols*, 6(7):991–1009, 2011.

[4] S. van Dongen. *Graph clustering by flow simulation*. PrintPartners Ipskamp, Enschede, 2000.