M2 Project proposal: Statistical quantification of spatial differences in cellular structure

1 Introduction

A key feature of eukaryotic cells is the compartmentalization of cellular functions into a complex system of distinct organelles, each separated by membranes. Fluorescent markers attached to proteins of interest are a useful proxy for studying organelles and their behaviour. Using them in conjunction with modern fluorescent microscopy technologies has allowed for the visualization of a range of sub-cellular organelles. Schauer et al. (2010) demonstrated that kernel density estimation is a powerful technique for studying global cellular organization and for quantification of cell biology. Kernel density estimation is a non-parametric statistical technique which is useful for describing distributions of biological experimental data when no obvious parametric model is available.

The workflow we have established in our laboratory to obtain a density map of an endomembrane’s spatial distribution is summarised in Figure 1: (a) The fibronectin micropattern (red) is manufactured in a fixed shape (here in a crossbow) which constrains the cell to grow in a standardised shape; (b) the cell is plated onto the micropattern; (c) the cell is fixed and stained ready for immunofluorescence imaging; (c) the cell is imaged in 3D using a fluorescent microscope, with the green spheres representing the endomembranes; (d) the cell images are processed using deconvolution and segmentation techniques to produce 3D co-ordinates of the endomembranes; (e) tens of cells are imaged and analysed, then all co-ordinates are pooled and aligned using the micro-pattern; (f) density estimation is carried out on these aligned co-ordinates.

![Figure 1: (a) Fibronectin pattern. (b) Cell plating. (c) Immunofluorescence staining. (d) 3D image acquisition. (e) Deconvolution and segmentation. (f) Alignment of tens of cells. (g) Density estimation.](image)

2 Using kernel estimation to detect significant differences in density maps

Figure 1 contains the workflow to produce a density map of one endomembrane. If we repeat this procedure, then we have multiple density maps. A natural question is thus to compare these maps, e.g. to establish a valid control experiment protocol, repeated experiments should
give maps which are not significantly different, or to detect the differences induced by a drug treatment compared to the control sample.

Let \( \{X_1, X_2, \ldots, X_{n_1}\} \) and \( \{Y_1, Y_2, \ldots, Y_{n_2}\} \) be \( d \)-variate random samples from their respective common densities \( f_1 \) and \( f_2 \). The kernel density estimates of \( f_1 \) and \( f_2 \) are

\[
\hat{f}_1(x; H_1) = n_1^{-1} \sum_{i=1}^{n_1} K_{H_1}(x - X_i), \quad \hat{f}_2(x; H_2) = n_2^{-1} \sum_{j=1}^{n_2} K_{H_2}(x - Y_j)
\]

where \( K \) is the kernel function with \( K_{H_\ell}(x) = |H_\ell|^{-1/2}K((H_\ell)^{1/2}x) \), and \( H_\ell \) is a bandwidth matrix, for \( \ell = 1, 2 \). To test null hypothesis \( H_0 : f_1 \equiv f_2 \), we require a discrepancy measure between the two density functions \( f_1 \) and \( f_2 \). We follow Anderson, Hall & Titterington (1994) who use an \( L_2 \) criterion \( T(f_1, f_2) = \int_{\mathbb{R}^d} |f_1(x) - f_2(x)|^2 \, dx \). Theorem 1 from Duong (2011) shows that under suitable regularity conditions, a kernel estimator of this test statistic \( \hat{T} \) follows an asymptotic normal distribution, i.e. as \( n_1, n_2 \to \infty \), \( [(n_1^{-1} + n_2^{-1})\sigma_\ell^2]^{-1/2}(\hat{T} - \mu_T) \overset{d}{\to} N(0,1) \). The most important parameters to estimate are \( \mu_T \) (via the bandwidth matrices \( H_1 \) and \( H_2 \)) and \( \sigma_\ell^2 \). Fortunately this has been done already in Duong (2011). The advantage of this test over their competitors is that it does not require bootstrap re-sampling to compute the null distribution, and so can be used in a ‘black-box’ manner by non-statistician users.

Once we establish from this test that two samples are different overall, then we are interested further in the sources of this difference. We follow the framework set up by Duong et al. (2009) who characterise the regions of the sample space which contribute the most to the overall difference using the difference between the two density functions \( q = f_1 - f_2 \), as show in Figure 2 for two 1D density functions. We use 1D functions for illustrative purposes because they are the easiest to visualise, but we will be working with 2D and 3D data. In the right panel in this figure, the most significant differences between these two densities correspond to the local extrema in the density difference \( q = f_1 - f_2 \).

\[
\text{Figure 2: Density difference. Left: Normal mixture density } f_1 = 1/2N(0, 1/4) + 1/2N(3/2, 1/4)). \quad \text{Centre: Normal density } f_2 = N(-1/2, 1/4). \quad \text{Right: Density difference } q = f_1 - f_2. \quad \text{The most} \quad \text{significant differences between these two densities correspond visually to the local extrema in } q.
\]

2.1 Global thresholding

The approach taken in Duong et al. (2009) is to look for thresholds based on the height of of the density difference \( q \) to define level sets. From the left panel of the illustrative example in Figure 3, we obtain the level sets \([-0.12, 0.16] \cup [1.34, 1.66] \) (denoted by the solid horizontal green lines on the \( x \)-axis) by cutting at \( q_1 = 0.62 \), and \([-0.93, -0.58] \) (solid horizontal purple...
line on the $x$-axis) at $q_2 = -1.20$. The solid green lines indicate the range of values of $x$ where $f_1$ is mostly greater than $f_2$, and the solid purple lines vice versa. The crux of this approach is to find optimal threshold values. Duong et al. (2009) used a technique borrowed from data mining, known as the Patient Induction Rule Method or PRIM. This method suited to flow cytometry FACS data which is usually more than 5D, but kernel based approaches are more intuitive for 2D and 3D microscopy data. Our current approach in the laboratory is to find thresholds based on the probability regions of the individual density estimates. These are defined as the smallest region which contains $\alpha$ probability mass. This implicit definition allows us to define thresholds for any density function since we choose from probabilities which are bounded by 0 and 1, and so we do not need to know the range of density values. We have a good understanding how to interpret the regions for single density estimates: we need to develop an analogous rigorous interpretation of the size and placement of density difference level sets. For example, we know that $\mathbb{P}(X \in [-0.12, 0.16] \cup [1.34, 1.66]) = \int_{[-0.12,0.16] \cup [1.34,1.66]} f_1(x) \, dx = 0.5$, and $\mathbb{P}(Y \in [-0.93, -0.58]) = \int_{[-0.93,-0.58]} f_2(x) \, dx = 0.5$, but we are not sure how to interpret probabilistically the light green shaded area $\int_{[-0.12,0.16] \cup [1.34,1.66]} |f_1(x) - f_2(x)| \, dx$ and the pink shaded area $\int_{[-0.93,-0.58]} |f_1(x) - f_2(x)| \, dx$.

![Figure 3](image_url)

Figure 3: Density difference regions. The horizontal solid green line is where sample 1 ($f_1$) is mostly denser than sample 2 ($f_2$), vice versa for the horizontal solid purple lines. Left: Level based, global thresholds. The horizontal dashed lines are the thresholds $q = 0.62$ and $q = -1.20$, which result in two level sets (horizontal solid green and purple lines). These level sets also correspond to the smallest probability regions with probability mass 0.5 for $f_1$ and $f_2$. Right: Significant curvature based, local thresholds. The density values where the curvature is significantly less/greater than zero is coloured in solid green/purple curves on the $q$ function. These in turn project onto the $x$-axis to give significant curvature regions (horizontal solid green and purple lines).

### 2.2 Local thresholding

An alternative to these highest probability regions are thresholds based on higher order properties of density difference function $q$. Local minima in $q$ correspond to where the first derivative is zero, and the second derivative is positive; and local maxima where the first derivative is zero and the second derivative is negative. These local extrema are single points. For example, a local maximum can be extended to encompass its local neighbourhood to define a local modal region, using a technique known as feature significance, which uses hypothesis testing on the derivative of a kernel density estimate. So the modal region is defined as the points where
the curvature (or second derivative) is significantly negative \( \{ x : \text{reject } H_0(x), \text{ where } H_0(x) : \frac{\partial^2 f(x; h)}{\partial x^2} = 0 \} \), see Duong et al. (2008). This local approach contrasts with the global approach of highest probability regions. On the right panel of Figure 3, a significance level of 5% gives the ‘bumps’ based on a kernel density estimate \( \hat{f}_1(x; 0.12) \) denoted by the solid green curves and the ‘anti-bumps’ based on \( \hat{f}_2(x; 0.2) \) by the solid purple curves. These in turn give projections on the \( x \)-axis, namely \([-0.17, 0.17] \cup [1.29, 1.70] \) and \([-1.46, 0.50] \). These significant curvature regions are optimal for the individual densities \( f_1 \) and \( f_2 \) separately, so we do not how appropriate they are estimating the significant curvature regions of \( q = f_1 - f_2 \). We need thus to adapt feature significance for a single density function to the difference between density functions.

These two approaches give similar results in this illustrative example because this is an easy example to see that the two density functions are clearly different. However for 2D and 3D experimental data, the differences can be more subtle, and more so since we must first estimate the density maps from the data. We are looking to optimise either of these approaches to find statistically significant regions of differences between density maps which are biologically important.

3 Project members

This project is supervised by Tarn Duong (email: tduong@curie.fr, tarn.duong@gmail.com) in the Molecular Mechanisms of Intracellular Transport Laboratory, Institut Curie, Paris, as part of an interdisciplinary research team. The head of the laboratory is Bruno Goud (email: Bruno.Goud@curie.fr) who is also the principal investigator. The other members are Kristine Schauer (post-doctoral researcher, email: Kristine.Schauer@curie.fr) and Jean-Philippe Grossier (doctoral student, email: Jean-Philippe.Grossier@curie.fr) who are both biologists.

References


