Université de Montréal

A robust algorithm for segmenting fluorescence images and its application to single-molecule counting

par
Jacques Boisvert

Département de biochimie et médecine moléculaire
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en bioinformatique

septembre, 2014

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

A robust algorithm for segmenting fluorescence images and its application to single-molecule counting

présenté par:

Jacques Boisvert

a été évalué par un jury composé des personnes suivantes:

Sylvie Hamel,	président-rapporteur
Paul Maddox,	directeur de recherche
Sébastien Lemieux,	codirecteur
Jean Meunier,	membre du jury

Mémoire accepté le: . . . . . .

RÉSUMÉ

La microscopie par fluorescence de cellules vivantes produit de grandes quantités de données. Ces données sont composées d'une grande diversité au niveau de la forme des objets d'intérêts et possèdent un ratio signaux/bruit très bas. Pour concevoir un pipeline d'algorithmes efficaces en traitement d'image de microscopie par fluorescence, il est important d'avoir une segmentation robuste et fiable étant donné que celle-ci constitue l'étape initiale du traitement d'image. Dans ce mémoire, je présente MinSeg, un algorithme de segmentation d'image de microscopie par fluorescence qui fait peu d'assomptions sur l'image et utilise des propriétés statistiques pour distinguer le signal par rapport au bruit. MinSeg ne fait pas d'assomption sur la taille ou la forme des objets contenus dans l'image. Par ce fait, il est donc applicable sur une grande variété d'images. Je présente aussi une suite d'algorithmes pour la quantification de petits complexes dans des expériences de microscopie par fluorescence de molécules simples utilisant l'algorithme de segmentation MinSeg. Cette suite d'algorithmes a été utilisée pour la quantification d'une protéine nommée CENP-A qui est une variante de l'histone H3. Par cette technique, nous avons trouvé que CENP-A est principalement présente sous forme de dimère.

Mots clés: Segmentation, traitement d'image, microscopie par fluorescence, centromère

ABSTRACT

Live-cell fluorescence microscopy produces high amounts of data with a high variability in shapes at low signal-to-noise ratio. An efficient design of image analysis pipelines requires a reliable and robust initial segmentation step that needs little parameter fine-tuning. Here, I present a segmentation algorithm called MinSeg for fluorescence image data that relies on minimal assumptions about the image, and uses statistical considerations to distinguish signal from background. More importantly, the algorithm does not make assumptions about feature size or shape, and is thus universally applicable. I also present a pipeline for the quantification of small complexes with single-molecule fluorescence microscopy using this segmentation algorithm as the first step of the workflow. This pipeline was used for the quantification of a small histone H3 variant protein called CENP-A. We found that the CENP-A nucleosomes are dimers.

Keywords: Segmentation, image processing, fluorescence microscopy, centromere

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

CENP-A	Centromere Protein A
EMCCD	Electron Multiplying Charge Coupled Device
FDR	False Discovery Rate
FP	False Positive
FPR	False Positive Ratio
GFP	Green Fluorescent Protein
HD	H-Dome
IUWT	Isotropic Undecimated Wavelet Transform
LoG	Laplacian of Gaussian
LSM	Laser Scanning Microscope
MinSeg	Segmentation of fluorescent images with minimal assumptions
MSVST	Multi-Scale Variance Stabilizing Transform
NA	Numerical Aperture
QUBE	QUantification of photoBleaching Events
SCMOS	Scientific Complementary metal-oxide-semiconductor
SNR	Signal-to-Noise Ratio
STED	Stimulated emission depletion
STORM	Stochastic Optical Reconstruction Microscopy
TIRFM	Total Internal Reflection Fluorescent Microscopy
TP	True Positive
TPR	True Positive Ratio
YFP	Yellow Fluorescent Protein

To Jonas, my mentor, and Paul, my advisor.

ACKNOWLEDGMENTS

No words can truly express the gratitude I have for these people but I shall try nevertheless. I would like to thank Paul Maddox for the chance he took by making me part of his lab family, his counsel and his undeniable support. I would like also to thanks Jonas Dorn for his dedication,mentoring and friendship. This project would never have happened without the intervention of Sébastien Lemieux, who is a wonderful teacher. This work wouldn't have been possible without Abbas Padeganeh who did all the experimental work. I want to extend my undying gratitude to Joel Ryan for being there during work hour and after work hour, thanks buddy. I would also like to extend my appreciation to everyone in the Paul Maddox/Amy Maddox lab. For all the laughs, the fun, the memories. Incredible labmates made the experience more enjoyable than it had a right to be. The two algorithms used for benchmarking were given by Ihor Smal and I would like to acknowledged his help. Finally, my sincere gratitude to my beloved Lili for staying by my side and helping me every way she could.

CHAPTER 1

INTRODUCTION

In cell biology, the stoichiometry of molecules is an important factor in determining the role of a protein. A protein complex can be nonfunctional without all its subunits or it can interact with other different proteins depending on its current stoichiometry. To properly understand how a protein functions, it is important to know its exact composition. The methods currently used for analyzing the constitution of complexes aren't appropriate for highly dynamic processes, nor for complexes that cannot easily be extracted for biochemical analysis, either due to their size or due to their stability and that is why the composition of certain protein complexes has been challenging to resolve, for example the protein CENP-A in nucleosomes. This centromere protein's stoichiometry has been a debate in the last years [1–4]. The main goal of this work is to develop a new method using state of the art programs to analyze the complexes of important dynamic proteins like CENP-A.

Fluorescent microscopy is a powerful tool for the observation of *in vivo* protein interaction, localization and dynamics. The ability to be able to add a fluorophore like the green fluorescent protein (GFP) to a protein and be able to use a microscope to observe the GFP tagged protein through time was a major breakthrough. Total internal reflection fluorescent microscopy (TIRFM) is a type of fluorescent microscopy that is used to do single-molecule quantification. We propose to use the single-molecule power of TIRF with state of the art algorithms to characterize the stoichiometry of small proteins like CENP-A.

The first step in many processing pipelines for image analysis of fluorescent microscopy images is the segmentation. Segmentation is the separation of an image into regions of similar characteristics. The segmentation can create multiple clusters but in our case we are interested in binary segmentation. This segmentation creates two classes

of regions, foreground and background. The segmentation is the first step of an image processing pipeline because it reduces the complexity of the image. It is easier to extract information like area, shape, contour or intensities of an object of interest when you know what pixels belong to the object. The main challenge in fluorescent microscopy segmentation is the low signal-to-noise ratio. The signal is low compared to the noise because there is always a compromise between the frequency of acquisition, the total observation time and the intensity of the light. The total amount of light to which a sample can be exposed is limited by either its viability or the stability of the fluorophore.

In the second chapter, we propose a new segmentation algorithm for fluorescent microscopy called **MinSeg**. This work was made by both my supervisor Jonas Dorn and me. Briefly, the algorithm has 4 main steps. First, a background subtraction is applied to allow for global thresholding. Secondly, a noise estimation method is used to quantify the noise present in the image. Thirdly, the image is threshold based on the quantification of the noise into a binary image. Finally, the image is filtered by a function that takes into account the local pixels.

In the third chapter, we propose an algorithmic pipeline to do the QUantification of photoBleaching Events (QUBE) in TIRFM with minimal human interaction, while compensating for multiple sources of error. This pipeline has four main stages. First, a segmentation is done. This was done with the MinSeg algorithm presented in chapter 2. Secondly, a Gaussian mixture fitting is done to improve the estimation of the intensities and the localization of the features. Thirdly, a tracking of every feature through time is done. Finally an automatic analysis of the intensity profile is computed and corrections are applied. All the experimental work for this work was done by Abbas Padeganeh [1].

Be advise, blue words throughout the text can be found in the glossary.

1. Padeganeh A., Ryan J., Boisvert J., Ladouceur AM., Dorn JF., Maddox PS., Octameric CENP-A nucleosomes are present at human centromeres throughout the cell cycle., Curr Biol,9(23),764-9,2013

CHAPTER 2

MINSEG - AN ALGORITHM FOR SEGMENTATION OF FLUORESCENT IMAGES WITH MINIMAL ASSUMPTION

2.1 Introduction

In computer vision, the segmentation of an image, i.e. the separation of an image into regions such as foreground and background is the foundation of most image processing procedures. The goal of image segmentation is to reduce the information of an image from the ensemble of all pixel intensity values to the few landmarks or shapes that will be relevant for subsequent processing steps. For example, image segmentation is used in astronomy to find candidate signals that might be planets or stars [6]. The technique is also used for automatic recognition of handwriting, video surveillance, fingerprint analysis [7], iris analysis and facial recognition [8].

In our case, we are interested in the segmentation of fluorescence microscopy videos. Fluorescence microscopy is a powerful tool that allows direct visualization of the behavior of biological systems. However, in imaging of live species, such as cells, tissues, or organisms, there is always a compromise between signal-to-noise ratio (SNR), frequency of acquisition and the total observation time. The total amount of light to which the sample can be exposed is often limited by either viability, i.e. too much light will destroy the sample, or by photobleaching, a photon-induced chemical reaction that destroys their ability to emit photons [9–11]. Especially when the total amount of light tolerated by the sample has to be spread over multiple exposures for live imaging, the resulting fluorescence images are of low signal-to-noise ratio, which limits the choice of image segmentation algorithms.

Histogram-based methods such as Otsu's thresholding algorithm [12] have been highly successful in many image segmentation problems since they are easy to imple-

ment and allow segmentation of arbitrary shapes without time-consuming parameter adjustments. They work by analyzing the histogram of the pixel intensities to identify an intensity threshold that separates features from background. Histogram-based techniques assume that the foreground and background form two discernible distributions in the histogram, but that isn't always the case in live-cell imaging since the signal is often weak relative to the noise. To segment low signal-to-noise ratio images, additional assumptions are needed, such as the signal's shape. Due to the diffraction limit of the optics of microscopes, as well as the physics of surface tension, an assumption that can frequently be made is that the signal resembles a spot-like shape. This assumption has led to the development of many successful algorithms [13, 14], but due to their assumption about feature shape, they are limited to spot-like features, and usually require pre-specification of feature size, which may limit their robustness. In addition, while incorporation of prior knowledge of the problem can lead to powerful algorithms, such approaches tend to require the optimization of multiple tuning parameters whose relation with the final threshold becomes less and less intuitive as the method grows in complexity. As an alternative to explicit model-based algorithms, machine learning methods have been developed for fluorescence microscopy [15, 16]. However, especially in fundamental research, where the distribution of expected phenotypes is often not well understood a priori, and particularly with research involving animals, it may not be possible to create a sufficiently complete or large training set.

We sought to develop a generic segmentation algorithm for fluorescence microscopy which works for low SNR images, makes minimal assumptions about the images in general, no assumptions about feature shape in particular, and which will thus allow the segmentation of many different types of features with minimal parameter adjustment. Here, we propose an algorithm we call **MinSeg** for segmentation of fluorescent images with minimal assumptions (see algorithm 1).

Algorithm 1: MinSeg

input : Grayscale images

output: Binary Images

foreach *image X* **do**

  if *necessary* **then**

    background subtraction(X);

  end

  noise = Estimate Noise(X);

  foreach *pixel p of image X* **do**

    if *Intensity of* $p \geq$ *noise* **then**

      Value of $p == 1$;

    else

      Value of $p == 0$;

    end

  end

  foreach *pixel p of image X* **do**

    if *# of pixels around p with a value of 1* $\geq$ *threshold h* **then**

      Value of $p == 1$;

    else

      Value of $p == 0$;

    end

  end

end

The algorithm makes 3 assumptions:

1 - The signal is brighter than the surrounding background i.e.

$$I(x,y) + B(x,y) > B(x+\delta_x, y+\delta_y) \tag{2.1}$$

where $I(x,y)$ is the intensity at coordinate x,y, B(x,y) is the underlying fluorescent back-

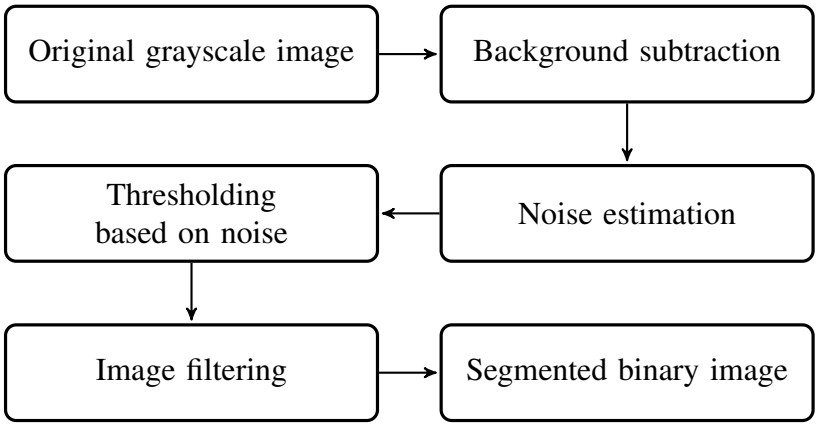

Figure 2.1: Flow chart of MinSeg algorithm

ground and δ_x , δ_y denotes a small change in coordinate, i.e. the surrounding fluorescent background.

2 - Signal is spatially correlated whereas noise is not spatially correlated.

3 - The spatial scale of the signal is different from the spatial scale of the background.

The first assumption is generally true for fluorescent microscopy where the signal is brighter than the fluorescent background because the protein of interest that was tagged with a fluorophore should have higher emission of photons, observed by a higher intensity in the image, than background autofluorescence, non-specific staining and even other fluorophores as long as those do not have the same excitation wavelength. Furthermore, fluorophores used in tagging have been engineered to increased their quantum yield. The second assumption requires the signals to be oversampled, which can be an issue with large-pixel EMCCD cameras or high NA/low magnification lenses because the signal won't spread over enough pixels. On the other hand, this makes the algorithm robust against hot pixels. If assumption three is violated, the algorithm won't be able to distinguish between background and signal and a model-based segmentation algorithm should be used in those cases. The algorithm requires almost no parameter adjustment, nor assumptions about feature shape, which makes it robust, versatile and easy to use, but also makes it unable to separate overlapping features. However, its reliable and statistically motivated distinction between features and background provides highly reliable

segmentation for any subsequent mixture-model algorithm.

MinSeg has four main steps (fig 2.1). First, a background subtraction algorithm is applied to allow global thresholding (fig 2.2B), which requires assumption 3. Secondly, a noise estimation method is used to quantify the noise present in the image. Thirdly, based on the estimation of the noise the image is thresholded into a binary image (fig 2.2C). Finally, the image is filtered with a kernel that replaces the value of the central pixel with 1 if the number of pixels is above a second threshold resulting in a binary image (fig 2.2D).

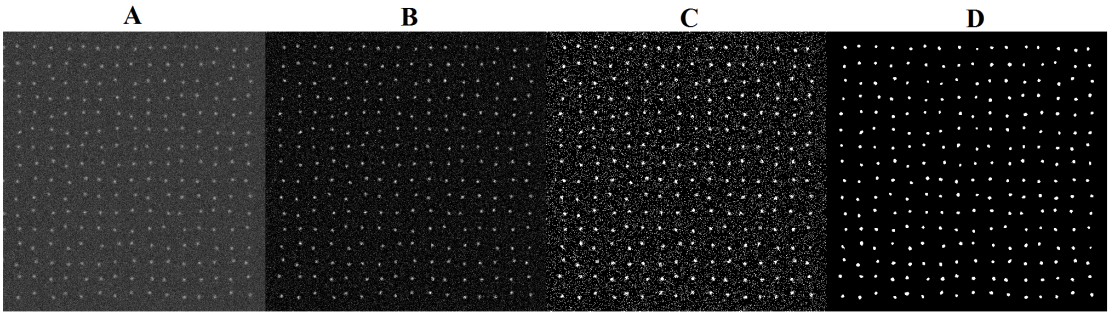

Figure 2.2: The 4 images show the processing pipeline of the MinSeg algorithm. Image A is a synthetic image with a SNR of 3 and containing 256 features. Image B is the same image after background subtraction. Image C is the image obtained after hard thresholding based on the noise. Image D is the image obtained after a second thresholding based on the vicinity of each pixel. The last image is the resulting image of MinSeg and is a binary image.

2.2 Background Subtraction

In fluorescence microscopy, the objects of interest are labeled with a fluorescent dye, which means that ideally, the only signal captured by the microscope is from the feature of interest, while the rest of the image remains black. In practice, this is not the case. The background comes from fluorescence that can originate from autofluorescence, non-specific staining or diffuse fluorescent molecules and is called fluorescent background. For total internal reflection microscopy, a non-uniform background can also come from

a small tilt in the angle of the cover-slip and the focus plane. This would make a side of the cover-slip closer to the laser, creating a background with a gradient effect. For practical purposes, this is also put in the same category and when we talk about fluorescent background this is taken into account. For our algorithm to work, it is important to subtract the fluorescent background because we do a global thresholding, and because our noise model assumes that the noise mean is centered at zero. A fluorescent background would influence the mean and break our assumption. The global threshold wouldn't take into account this change and if a background subtraction is omitted, we would under-threshold and have too many false positives.

In image processing, most methods developed for background subtraction were made to find moving objects [17]. In our case, a frame without any features isn't necessarily available and thus a method that can work on a single frame is needed. A median filter is often used to reduce salt and pepper noise. The filter operation will work well as long as the salt and pepper noise is 2 times smaller than the actual size of the kernel. In our case, the "salt" and "pepper" is actually the features and everything else is the background.

In fluorescence microscopy, the size of a feature is dictated by the diffraction limit because proteins are usually smaller than the wavelength of the light. By setting the median filter kernel size to at least 2 times the feature size, an image without the signal should be obtained and a background subtraction possible.

A median filtering is defined as the following (see algorithm 2). X is used to denote the initial image and Y is the filtered image. For every pixel p, let $X(x,y)$ be the intensity at coordinate x,y, where x represents the row and y the column. The results of the filtering with a window size of r will be the median of the intensity value located inside the r by r window centered at pixel x,y. For simplicity, r is odd. If r would be even, the center of the kernel and the median wouldn't be as well defined. Here is the pseudo-code of a naïve implementation of a median filter:

Algorithm 2: Naïve 2D Median Filter

input : Image X and filter size of size r
output: Filtered Image Y

foreach *pixel p of X at coordinate x,y* **do**
    for $i \leftarrow 0$ **to** r **do**
        for $j \leftarrow 0$ **to** r **do**
            // Create array with all the value inside the mask
            Array$[i+j]$ = X$[x-r/2+i, y-r/2+j]$;
        end
    end
    Quicksort(Array);
    // Middle value of the array is now the median
    $Y(x,y)$ = Array$[(r^2)/2]$;
end

This implementation has a complexity of $O(r^2)$ per pixel, and becomes really slow for large kernels in practical application [18]. Recently, a histogram based median filtering method was proposed with a complexity of $O(1)$ per pixel [19] (see algorithm 3). It was based on a previous histogram based method published [20]). This method works by keeping in memory a histogram for each column and updating it when needed, lowering the cost of calculating the median at each pixel. Each column histogram contains $2r+1$ pixels originally centered on the same row as the kernel. The first step is to update the column histogram to the right of the kernel by subtracting and adding one pixel to move it down one row. The second step is to update the kernel histogram by subtracting the leftmost column histogram and adding the column histogram that was just updated. Adding, subtracting and computing the median of a histogram isn't a function of the radius of the kernel but is in fact a function of the bit depth and as such is constant in terms of radius size. The side effect is an increase in memory and there is a constant in computational calculation introduced equal to the bit depth. This constant can become problematic for high bit depth image (fig 2.3A and fig 2.3B). The memory increases to $O(n * bitdepth)$ instead of O(r), where n is the number of columns in the image and r is the size of the mask. This approach is simple, efficient and most of the time the increase

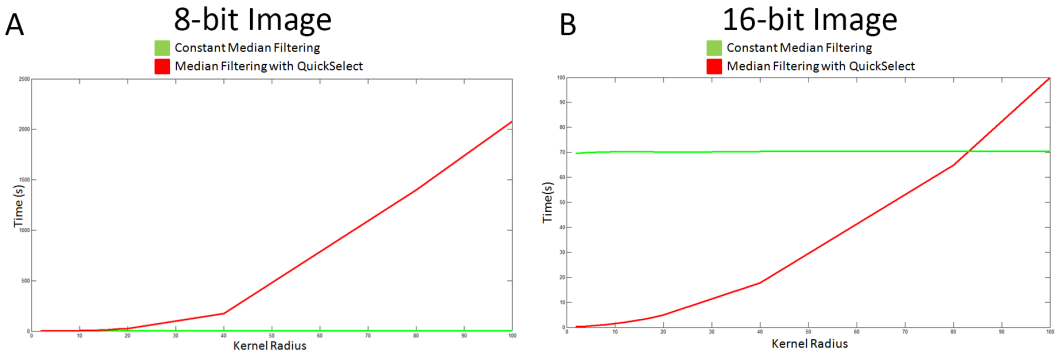

Figure 2.3: Panel A is the computation time for an 8 bit image as a function of the kernel radius for the naive median filter algorithm and the constant median filtering algorithm. Panel B is the computation time of an 16 bit image as a function of the kernel radius. The sharp change in time present in both panel graph is due to discreet steps.

in memory usage is not an issue because the dynamic range of microscopy images is rarely used to its full capacity and because of this it is possible to downscale images with a 16 bit depth to 8 bit depth image without any loss of information. The memory increases could be troublesome if the processing is made on large images that have a bit depth of 16 or higher can't be reduced to a lower bit depth without a loss in information. While a histogram-based median filter was available for Matlab , I wrote an implementation one for Java [21].

Algorithm 3: Median filter in constant time

input : Image X of size m,n and kernel size r
output: Filtered Image Y

Initialize kernel histogram H, column histogram $h_1..h_n$;
foreach *pixel p of X at coordinate x,y* **do**
    Remove $X_{x-r-1,y+r} from h_{x+y}$;
    Add $Xx+r,y+rtohy+r$;
    `// Add and subtract the relevent column histogram`
    H = $H + h_{y+r} h_{y-r-1}$;
    $Y_{x,y}$ = median(H);
end

As with other spatial filter techniques, border effects are inevitable. When the ker-

nel is centered on pixels near the border of an image, it is necessary to decide which value will be used outside of the border for calculation of the median. Four solutions have been implemented, null padding, symmetric padding, anti-symmetric padding and circular padding. Null padding assumes the value outside the border to be zeros. Symmetric padding reflects the trend of the values near the border. Anti-symmetric continues the trend of the values near the border. Finally, circular padding assumes that the image is circular on itself. For example, the pixels outside the left border are assumed to be the pixels situated at the right border. The image loops into itself. This is useful to replicate the behavior of filtering an image in the Fourier space.

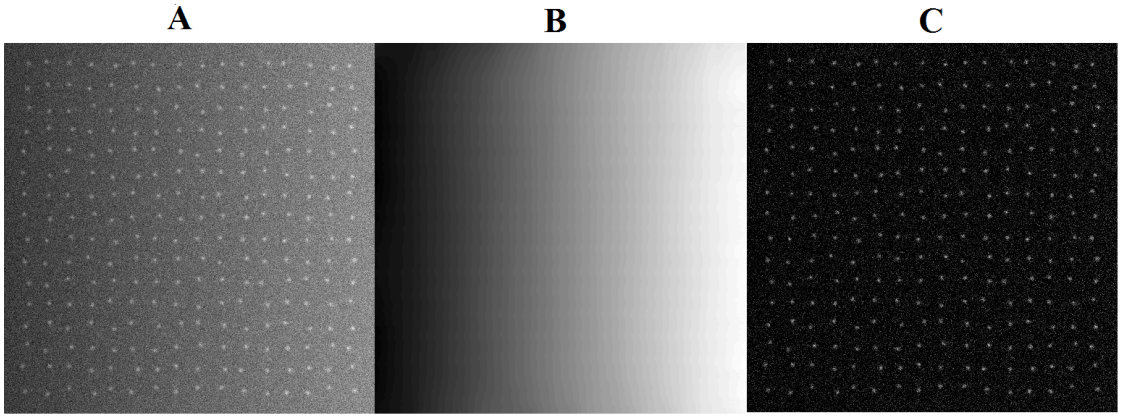

Figure 2.4: Panel A contains an image with a gradient fluorescent background and an SNR of 3. Panel B is the output of the median in constant time filter. Panel C is the subtraction of panel A by panel B.

One of the main drawbacks of the median filtering method for background subtraction is its limitation towards multiple levels of fluorescent background with size similar to the features or with highly inhomogeneous fluorescent backgrounds. This can be bypassed by multiple median filtering or with a pre-segmentation limiting the spatial domain of the filtering. For example, if we are trying to isolate fluorescent proteins situated inside a small nucleus of a cell, it can be tricky to get only the nucleus fluorescent background without losing any signal. This is especially true if the fluorescent background is also inhomogeneous inside the nucleus. A first background subtraction would be done to

eliminate everything beside the nucleus and then another background subtraction with a smaller kernel size would be done. This would give better results because the value outside of the nucleus would be set to null and thus not influence the fluorescent background inside the nucleus. A better estimation of the fluorescent background inside the nucleus is achieved. For fluorescence microscopy, median filtering offered a simple and robust algorithm for background subtraction (example of background subtraction on synthetic images fig 2.4). The main problem is the time involved in the computation since finding a median is computationally costly in time even with a complexity of O(1) per pixel in regard to the the kernel size. This is because the O(1) complexity while really low is still multiplied by the number of pixels and microscopy images can be very large containing millions of pixels and the image processing is usually done on image stacks containing hundreds of images. The drawbacks are acceptable in our case, since the other steps of the algorithm are not computationally costly making the whole segmentation take roughly 8 seconds for each image.

2.3 Noise estimation

In segmentation of fluorescent images, most algorithms do not do noise estimation because they are only interested in removing the noise. Estimating the noise is then not necessary. On the other hand, a good estimation of the noise can be useful for the selection of the appropriate parameters or for the suitable denoising method to be used [14]. Furthermore, in our case, it is possible to use the characterization of the noise to help the segmentation of the image without removing it. To be able to estimate the noise of an image properly, it is important to understand the origin of the noise. The digital images obtained from fluorescence microscope have multiple sources of noise, e.g. photon noise, dark noise, detector reading noise and read noise.

Dark noise, detector reading noise and read noise are inherent to the detector. Dark noise comes from the small electric current cursing through the detector even when no photons are entering. Detector reading noise comes from the amplification chain and the

converter of the detector device. Read noise comes from the quantification of the voltage to discrete levels of intensity. Photon noise comes from the random nature of photons emission. This type of noise could be theoretically reduced by increasing the light intensity or the exposure time but this is impractical in most applications. All the types of noise are usually considered to be spatially uncorrelated and independent. Photon noise and dark noise follow a Poisson distribution, detector reading noise follows a Gaussian distribution and read noise follows a uniform distribution. In most cases, the image processing community assumes that the noise is dominated by additive Gaussian type noise. Constant Poisson noise can be fit relatively well with a Gaussian fit, therefore assuming that the noise is Gaussian works pretty well in practice. On the other hand, if you have strong Poisson type noise on inhomogeneous fluorescent background, a local threshold would be necessary and estimating the noise on the whole image will overestimate the noise and lead to fewer features found. Thus, it is necessary to use a different approach to tailor a solution to the noise present in the image to get the best results.

Algorithms for noise estimation are mostly classified into two categories, block based [22, 23] or filter based [24–26]. Some algorithms use different approaches like wavelet based, but most of them fall into one of those two categories or use a combination of the two [27, 28]. Block based techniques create blocks by tessellating the image. Afterward, the noise is evaluated by calculating the variance of a set of homogeneous blocks. The challenge is to define those homogeneous blocks. Filter based techniques use a filter to extract the noise or get rid of the noise. If a noise-free image is obtained by the filter technique, the noise variance will be evaluated by calculating the difference between the noise-free image and the original image. The challenge with filter techniques is that the assumption that the difference between the filtered image and the original image is composed of only noise is often not true or not accurate.

For our algorithm, we have opted to use a derivative approach that falls in the category of filter techniques since a derivative can be implemented with different types of filters, like the Sobel operator. We chose this method because of its ease of use and

implementation.We estimate the noise from the variance, σ^2, of the third derivative.

$$\sigma_n^2 = \frac{var(\nabla_x^3 I(x,y))}{20^2}, \text{ where } \nabla_x^3 = I(x-1,y) - 3I(x,y) + 3I(x+1,y) - I(x+2,y) \quad (2.2)$$

Using the third derivative has the advantage of removing any contribution from the signal that can be approximated by up to quadratic functions. The differencing also has the property of increasing the noise variance by a factor of 20 for each dimension. This can be shown using Gaussian error propagation:

$$\sigma_{f(x_1,x_2..)}^2 = \sum_i \left(\frac{\delta f}{\delta x_i}\right)^2 \sigma_i^2 \quad (2.3)$$

$$\sigma_{\nabla_x^3(I)}^2 = (1 + 3^2 + 3^2 + 1)\sigma_n^2 \quad (2.4)$$

Thus, the variance of the high-order difference is dominated by the pixel-to-pixel noise variance, especially if the original image was noisy.

The variance calculated on the third derivative image is divided by 20 to the power of the number of dimensions. Of course, this approach assumes that the intensity $I(x,y)$ at position x,y is composed of the real intensity f with the addition of noise n drawn from a Gaussian distribution G of mean 0.

$$I(x,y) = f(x,y) + n(x,y) \text{ , } where\ n(x,y) \sim G(0,\sigma) \quad (2.5)$$

If the image is dominated by Poisson noise, like an image taken from a single-laser scanning confocal microscope where the excitation noise, photon noise, and detector noise all follow a Poisson distribution, this technique will not work adequately and another method is needed.

2.4 Distinction of signal from background and filtering

With the estimation of the noise, we threshold the image into a binary image,

$$BI(x,y) = 1 \; if \; I(x,y) > c * \sigma_n \; , \; 0 \;\; otherwise \tag{2.6}$$

Where BI is the binary image, $I(x,y)$ is the intensity at position x, y in the input image, c is a noise multiplier constant and σ_n the standard deviation of the estimated noise. This will lead to a uniform random distribution of pixels with value 1 where there is no signal because the noise should be spatially uncorrelated. On the other hand, if a signal is present, there will be clusters of 1s near each other. The probability of a pixel to be above the threshold is:

$$pSinglePixel = 1 - CDF_n(c, u = 0, \sigma = 1) \tag{2.7}$$

CDF_N is the cumulative distribution function of the normal distribution. For $c = 1.3$, pSinglePixel evaluates to 10%. That means that, we expect at most 10% pixels with value 1 in any neighborhood of I(x,y) where there is no signal. Conversely, we expect significantly more than 10% pixels with value 1 in a neighborhood where there is a signal, because signal should be correlated in space. Thus, the MinSeg considers a pixel as part of the signal if a sufficiently high number of its neighbors have their fluorescence intensity above the threshold defined(fig 2.5).

$$SI(x,y) = (BI(x,y) * H(x,y) > d \tag{2.8}$$

Where H is a binary convolution mask, BI the binary image obtained in the preceding step and SI is the resulting binary segmented image. Probabilistic Segmentation thus requires the determination of two thresholds, c and d, as well as the binary mask H. The threshold d is directly related to the number of pixels falsely considered as signal (FP).

$$FP = L_x L_y * (1 - CDF_b(p = pSinglePixel, k = d, n = \Sigma H)) \tag{2.9}$$

Where CDF_b is the cumulative distribution function to the binomial distribution and L_x/L_y are the width and height of the image, respectively. Therefore, by using FP as a parameter instead of d, we directly have an insight in the number of false positive pixels expected in the output segmentation image.

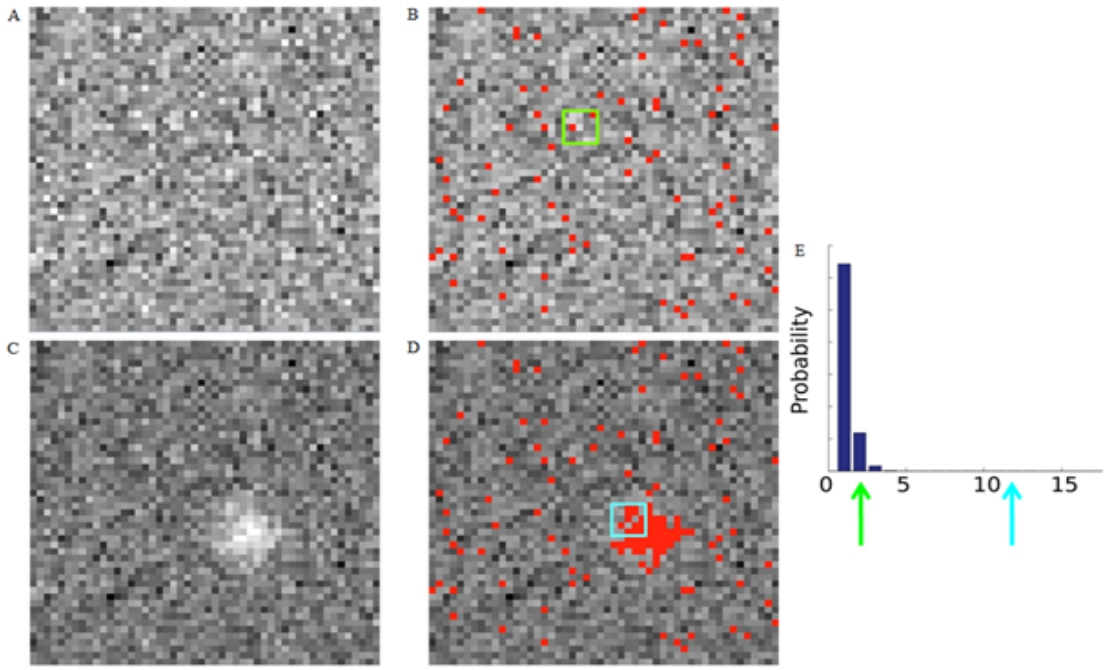

Figure 2.5: Image A is an image with only noise. Image B is the same image threshold with the noise estimation. Image C is an image with a feature and noise. Image D is the same image threshold with the noise estimation. E shows the probability distribution for a 5 by 5 mask. It represents the probability to find x lit pixel in a 5 by 5 mask. As can be seen, when a feature is present, the number of lit pixels is much higher than you would expect if only noise was present. The first arrow, present in green, shows the probability that the number of lit pixel in the mask present in panel B originates only from noise. The second arrow, present in blue, shows the probability that the number of lit pixel in the mask present in panel D originates only from noise

In practice, there is little reason to change these parameters from their default values. The H mask should be small, since choosing a mask that is larger than the features that are to be detected results in a morphological dilation of the features, while choosing a mask that is smaller than the features results in no adverse effects. However, a 3 by 3

mask allows for little dynamic range in choosing d and the expected number of false positives. Consequently, we use a 5 by 5 mask for 2D images, and a 5 by 5 by 3 mask for 3D images, with all pixels/voxels set to 1. The value of 1.3 for c allows us to reliably detect signals with SNR down to 2, and allows for a good dynamic range for choosing d and the respected number of false positives. Given c and H, only either d or FP can be chosen freely. FP is the more intuitive of the two parameters, thus we calculate d as a function of FP. FP is defined as the expected number of false positive pixels in an image, but with the dilation effect from the convolution, FP corresponds, in practice, to the number of false positive features in the image. By default, we set FP to 0.5, meaning we observe a false positive feature once every two images on average. It is important to note, that if the noise is underestimated, we will under-threshold and have more false positives than expected and if the noise is overestimated, the threshold will be too severe and we will miss some features. In practice, over-estimation is better than under-estimation because it is less detrimental to miss features than to pick false ones.

In summary, the algorithm substracts the fluorescent background with a median filter, estimates the noise with the third derivative, thresholds the image based on the noise estimation and finally uses a filter to threshold a pixel based on the number of pixels in its surrounding above the noise. This makes for a simple, efficient and easy to use algorithm.

2.5 Benchmark

2.5.1 Synthetic Images

We first tested our algorithm on synthetic images. We chose to compare MinSeg to spot detection algorithms, since spots are the most challenging features for the algorithm due to their small size, and many excellent algorithms have been developed for spot segmentation. We decided to use the synthetic data set of two studies comparing segmentation algorithms [13, 14].

From the first study, we decided to single out the two best unsupervised algorithms, H-dome [29] (HD) and Multi-Scale Variance-Stabilizing Transform [30] (MSVST) and tackle the same synthetic images.

H-Dome

H-Dome is a grayscale morphology operation scheme. The method assumes that the intensity distribution of the image is formed by N objects, background structures with intensity distribution I(i,j) and noise n(i,j) that can be multiplicative or additive. The algorithm has 3 mains steps: filtering, h-dome transformation and signal thresholding. The filtering is done with a Laplacian of a Gaussian filter(LoG). The filter is given by [31]

$$\nabla^2 G(x,y) = \frac{x^2+y^2-2\sigma_L^2}{\sigma_L^4} e^{\frac{x^2+y^2}{2\sigma_L^2}} \tag{2.10}$$

σ_L can be chosen based on the size of the feature in the image and the author suggests 2.5 pixels. The LoG operator will enhance the signal where features are present and diminish the background where there are no features. Applying the LoG filter gives the filtered LoG image as an output labeled by the author, I_σ. After the filtering, a grayscale reconstruction is applied. The LoG filtered image I_σ is filtered with a mask $I_\sigma - h$ where h > 0 and is a constant. This decomposes the image into a reconstructed image B_σ and a h-dome image H_σ.

$$I_\sigma(i,j) = H_\sigma(i,j) + B_\sigma(i,j) \tag{2.11}$$

B_σ represents the non-uniform background structures and H_σ the smaller noise structure and the features. For the final step, H_σ is used as a probability map. All the pixels in H_σ are raised to the power of S to compensate for the LoG filter that smoothens the image and to create a peak function that is similar to the probability density function of the feature distribution. S can be related to the minimum and maximum feature size and σ_L. The function

$$H_\sigma^s = (I_\sigma(i,j) - B_\sigma(i,j))^s \tag{2.12}$$

is the sampling function, denoted by $q(i,j|I)$. This function describes which areas are

more likely to contain features. Then, the authors sample N position/samples from $q(i,j|I)$ using a Monte-Carlo method, $x_l \; q(i,j|I)$, where $l = (1..N)$ and $x = (i,j)$. A mean-shift algorithm [32] is used to cluster the x_l sample, resulting in M clusters. For each cluster, the mean position $x_c = (i_c, j_c)$ and the variance R_c are calculated using only N_c samples belonging to that cluster.

$$x_c = \mathbb{E}[x_c^L] = N_c^{-1} \sum_{l=1}^{N_c} x_c^l \tag{2.13}$$

$$R_c = \mathbb{E}[(x_c^l - x_c)(x_c^l - x_c)^T] \tag{2.14}$$

Two criteria are used to distinguish between features and other structures.
1 - The number of samples N_c should be larger than the number of samples coming from a uniform intensity distribution in the region occupied by the cluster.
2 - The determinant of the covariance matrix R_c must be less than $\frac{\sigma_m^4}{s^2}$, where σ_m represents the maximum size of the features.

This comes from the fact that $(\sigma_m^2 ax + \sigma_L^2)s^{-1}$ is the upper bound of the intensity distribution. In conclusion, the method has 3 main parameters, σ_l, σ_{max} and h. σ_l and σ_{max} are related to the feature size. The parameter h is related to the signal-to-noise ratio. The authors claim that the method is insensitive to the power s and N, the sample size. The H-Dome algorithm used for this project was provided directly from the authors of [14].

Multiscale Variance-Stabilizing Transform

The Multiscale Variance-Stabilizing Transform uses the isotropic undecimated wavelet transform (IUWT) for the decomposition of the image. IUWT decomposes the image in K wavelet planes. The image I is convolved row by row with the 1D kernel [1/16,1/4,3/8,1/4/,1/16]. The kernel can be expanded if necessary by adding $2_{k-1} - 1$ zeros between the kernel coefficients. I_{k-1} is convolved with the kernel giving I_k as an

output. The wavelet plane k is computed with I_k and I_{k-1} by:

$$W_k(i,j) = I_{k-1}(i,j) - I_k(i,j), 0 < k \leq K \tag{2.15}$$

The variance-stabilizing transform is applied on I_k before the decomposition. Afterward, the wavelet coefficients are separated with a multiple statistical hypothesis test based on the Benjamini and Hochberg [33] procedure, which controls the false discovery rate (FDR). All the insignificant coefficients are zeroed and a reconstruction is done:

$$I(i,j) = I_K(i,j) + \sum_{k=1}^{K} W_K(i,j) \tag{2.16}$$

The obtained image thresholds all the pixels with negative values to zeros then the connected pixels with non zero values are considered features. The method has two main parameters, K and γ. K is linked to the depth of the decomposition and γ is the upper bound given to the FDR during the multiple statistical hypothesis procedure.

The first study used 3 different types of spot-like images, A, B and C (fig 2.6). Two types of features were modeled for the 3 images: round and elongated shapes. The round shapes are modeled using a 2D Gaussian intensity profile with a $\sigma_{max} = \sigma_{min} = 100nm$ and the elongated shapes with a $\sigma_{max} = 250, \sigma_{min} = 100nm$. All image types are 512 by 512 pixels with a pixel size of 50nm in x and y. 256 features are present in all image types and are placed randomly within a certain image region. Type A images were created by adding a uniform background of 10 representing uniform "fluorescent" background and Poisson noise independently to every pixel. Type B images were created by applying a gradient "fluorescent" background with a value of 10 on the left and linearly going to 50 towards the complete right. Since Poisson noise is intensity dependent, a correction was applied to ensure a constant SNR throughout the image. Type C images are constructed with large "fluorescent" background structures that could represent subcellular structures or acquisition artifacts. While the study benchmarks the algorithm on a range of 2-4 SNR, we limited our benchmarking to the hardest SNR, 2.

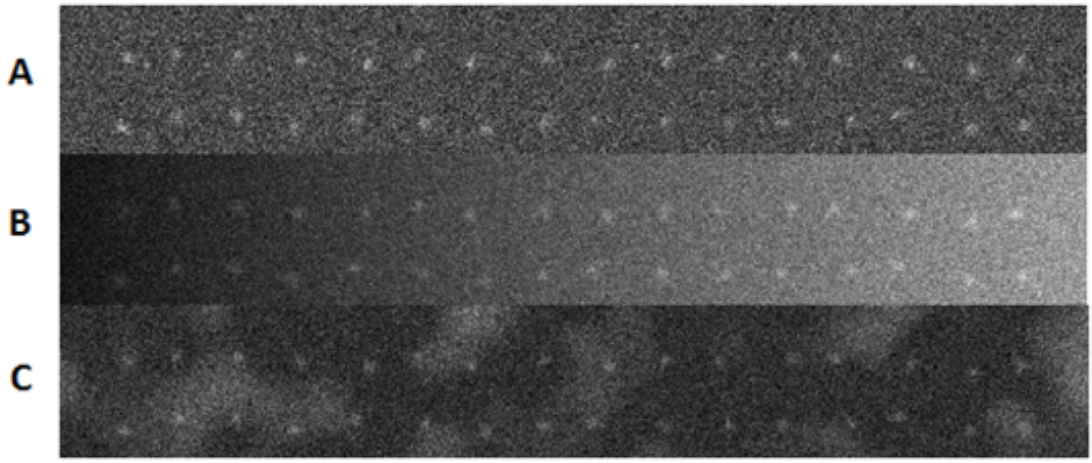

Figure 2.6: Image A,B,C shown are subsets of the original images used. Image A is an example of synthetic Type A images with a SNR of 4, the background is uniform. Image B is an example of the synthetic Type B images with a SNR of 4, the background is a gradient. Image C is an example of synthetic Type C images with an SNR of 4, the background is inhomogeneous.

To compare the algorithms, two measurements were used: the true-positive ratio (TPR) and the false-positive ratio (FPR).

$$TPR = N_{TP}/(N_{TP}+N_{FP}) = N_{TP}/N^0 \tag{2.17}$$

$$FPR = N_{FP}/N^0 \tag{2.18}$$

N_{TP} is the number of true positives. A true positive is a feature that was successfully detected. N_{FP} is the number of false positives. A false positive is a feature that was detected by the algorithm when that object is not present in the ground truth image. N^0 is the total number of features present in the ground truth image, 256 in our case. The TPR represents the sensitivity of the algorithm, the higher the TPR the more sensitive the algorithm is. The FPR represents the accuracy. A high FPR implies that the algorithm is picking up a lot of objects that are not present in the image. In other words, a low FPR indicates accuracy and a high TPR indicates sensitivity. It is worse to find false objects than it is to miss true objects. Thus it is acceptable to lower the TPR if the FPR is also

reduced. For the benchmark, all algorithms were set to a FPR of 0.01% because it is possible to tweak the parameters to find a lot of features, present or not.

	RoundType			Elongated Type		
Image Type	MinSg	MSVST	HD	MinSg	MSVST	HD
Type A	0.99	0.99	0.99	0.99	0.99	0.99
Type B	0.93	0.99	0.97	0.99	0.99	0.99
Type C	0.83	0.93	0.90	0.94	0.96	0.97

Table 2.I: True-positive ratio performance for MinSg, MSVST and HD on data set of SNR 2. The FPR for all the algorithms were estimated at the level 0.01%.

For all 3 types of images, MSVST performs slightly better (Table 1) but all three algorithms show a high sensitivity. Noteworthy is the results of our algorithm on Type C images that are a bit lower than HD and MSVST. This can be explained by the high textured "fluorescent" background present in that image type. Our background subtraction model may be too simple to properly subtract a double layer "fluorecent" background without losing some features. Using a more sophisticated approach to model the "fluorescent" background would probably boost the performance but the cost would be to lose one of the main appeals of our approach, the ease of use. MinSeg has the advantage over MSVST and HD of having really few parameters that needs to be tweaked to get good results. HD has the disadvantage of having 3 main parameters that are somewhat linked to the appearance of the features and its intensities. A group set with heterogeneous features would greatly impact HD where our method does not make any assumption about the feature size or shape.

2.5.2 Osteosarcoma well plate images

To compare our algorithm with others of its kind in a high-throughput setting, we decided to utilize the same sample found in [13] comparing multiple segmentation algorithms belonging to the same class type as ours. Like the synthetic benchmark, we focus our attention on the comparison between our results and the results of the H-Dome algorithm. Unlike the previous benchmark, MSVST was omitted because it was not part

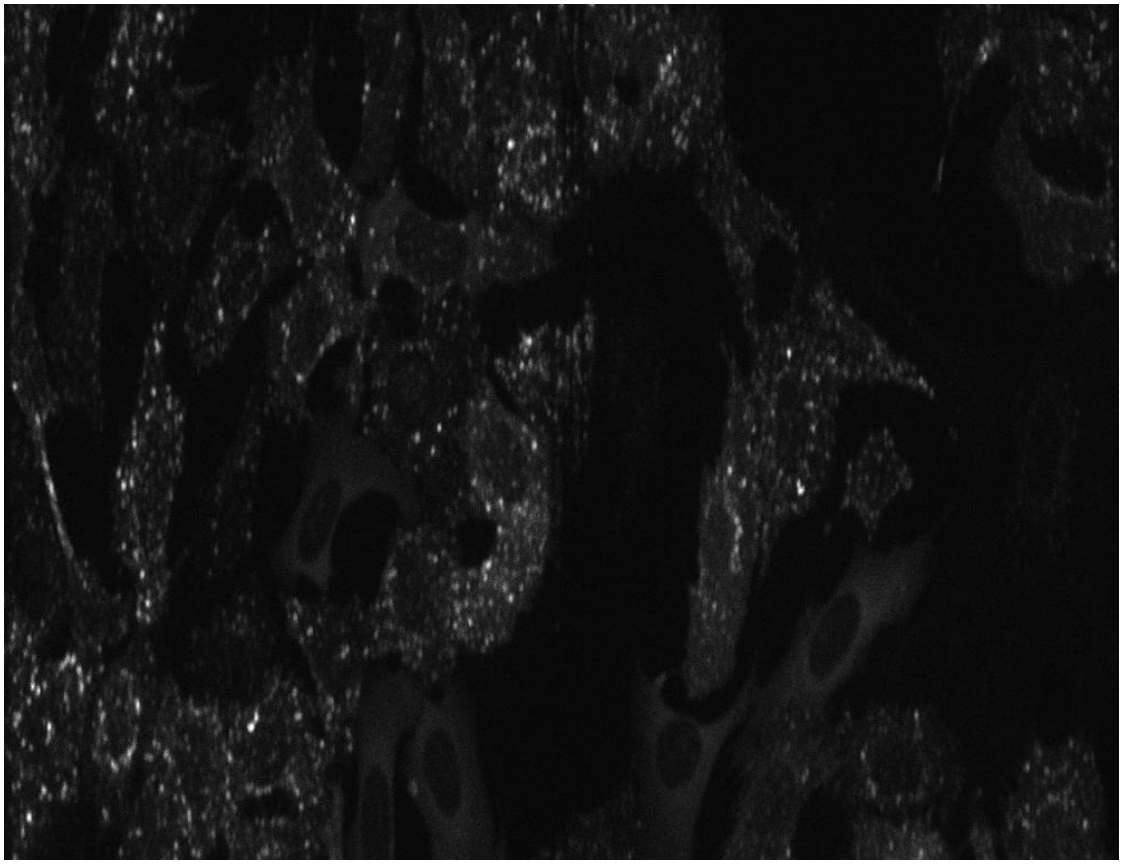

Figure 2.7: Example of an osteosarcoma well plate image with a dosage of category XII used for the high-throughput comparison between MinSeg and HD

of the benchmark found in [13]

The sample (fig 2.7) contains cells that in response to a drug will show a varying degree of vesicle-like structures. The algorithm has to estimate the average number of vesicles and then divide that number by the number of cells, returning an average number of vesicles per cells. The results are then regrouped by dosage level (fig 2.8). This should result in a sigmoidal pattern, where low dosage(I-VI) exhibit almost no vesicles and high dosage(VII - XII) exhibit a higher number of vesicles, plateauing at Dosage XI. As shown, MinSeg produces the anticipated sigmoidal graph, and could be used to detect the difference in a population exposed to different levels. In contrast, H-Dome shows a limited increase in vesicle number only. This could make it hard to identify

the onset of the high plateau. In an image with only noise, H-Dome will still find some features and on a high-throughput experiment, this can be problematic and undesirable.

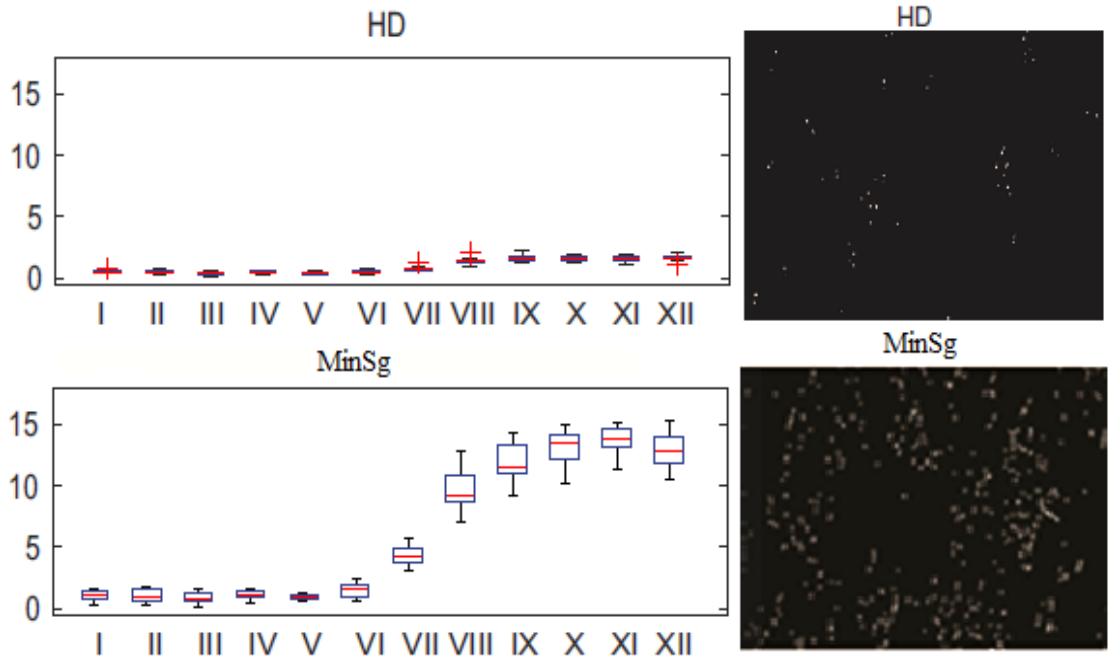

Figure 2.8: Top Left Panel, box plot showing the number of detected objects per cell for each dosage for the H-Dome algorithm. Top Right Panel, example of segmentation results for the H-Dome algorithm. Bottom left panel, box plot showing the number of detected objects per cell for each dosage for MinSeg. Bottom right panel, example of segmentation results for the MinSeg algorithm from the osteosarcoma well plate images set for dosage XII.

2.6 Experimental Results

It is important to prove that a method can actually be applied in a real world application. In this part, we show one experimental example of application of the MinSeg algorithm. This part was done by my supervisor Jonas Dorn. Another application example performed by myself is presented in Chapter 3.

We have successfully applied MinSeg to different types of fluorescent images, ranging from segmenting 3D Drosophila border cells [34] to single molecules [35]. Here,

we present an application that particularly highlights the strengths of the algorithm: The detection of P-bodies inside tissue culture cells (fig 2.9). P-bodies, or processing bodies, are agglomerates of both protein and RNA, and are thought to be involved in the regulation of translation, possibly by sequestering cytoplasmic RNA. To determine the role of the protein 4E-T in the regulation of P-body assembly, 4E-T distribution was observed in HeLa or U2OS cells by immunostaining with Alexa 488-conjugated antibodies, and imaged on a Zeiss LSM 510 single laser scanning confocal microscope with a 63x/1.4NA lens. Cells were either subjected to RNA interference or to drugs inhibiting key signaling cascades, before they were challenged by chemicals that either forced assembly or disassembly of P-bodies. We used MinSeg with default parameters first to detect the cell outlines without the need for an additional fluorophore - the diffuse signal of non-specific labeling or of cytoplasmic autofluorescence reliably provides enough signal for MinSeg to distinguish cells from fluorescent background. We further used MinSeg, still with default parameters, to detect DAPI-stained nuclei, which were used as seeds for a marker-based watershed algorithm [36] to separate touching cells, which produced satisfactory results and which did not require parameter adjustment when cell types were switched from U2OS to HeLa cells. Finally, we used MinSeg with background subtraction and Poisson noise estimation to segment P-bodies. There is no uniformly accepted criterion in the field to determine what is considered a P-body, and what background is. Frequently, only the very brightest spots are counted in manual analysis, which leads to skewed results [37]. MinSeg, through its statistical criteria for differentiating signal from background, allowed us to create an objective definition for "fluorescent agglomerates" of the labeled proteins, and therefore allowed characterization of the full distribution of sizes, revealing a novel mechanism of P-body assembly control.

2.7 Conclusion

MinSeg is an algorithm for segmenting fluorescence images that is highly convenient for several reasons: It robustly distinguishes signal from background in a wide variety of images where histogram-based approaches fail, it defines signal in a statistical fashion,

giving intuitive meaning to the threshold separating signal from noise and it requires, in practice, the tuning of only one single parameter, the background filter size, which leads to successful segmentation of arbitrary features with little effort. It is important to point out that the method necessitates a minimum resolution. If the sampling is too low, it will be impossible for the method to discern feature and noise because it assumes that features occupy a larger spatial resolution than noise. For the speed of the algorithm, the median filter is the bottle neck. For large images that have a huge inhomogeneous fluorescent background that takes a huge kernel, the constant median filtering has help a lot to lower the computational time but this step alone still takes a couple of seconds and if it is impossible to use the algorithm because of the size and bit depth of the image then the median filtering can take up to a couple of minutes. In terms of memory limitation, the median in constant time used by MinSeg for background subtraction can take a lot of memory if the image passed in input is an image with a large bit depth. If this occurs to be a problem or if the image cannot be reduced to a 8 bit image, an alternate slower median algorithm was also implemented.

A segmentation algorithm similar to ours was recently published [38]. They are using a feature-preserving non local mean filter [39] followed by a particle detection. It would have been interesting to compare it to our algorithm and this is certainly something we would like to accomplish in the near future. Furthermore, extending the algorithm to the exponential noise pattern of SCMOS camera is also something we would like to accomplish making the algorithm even more versatile.

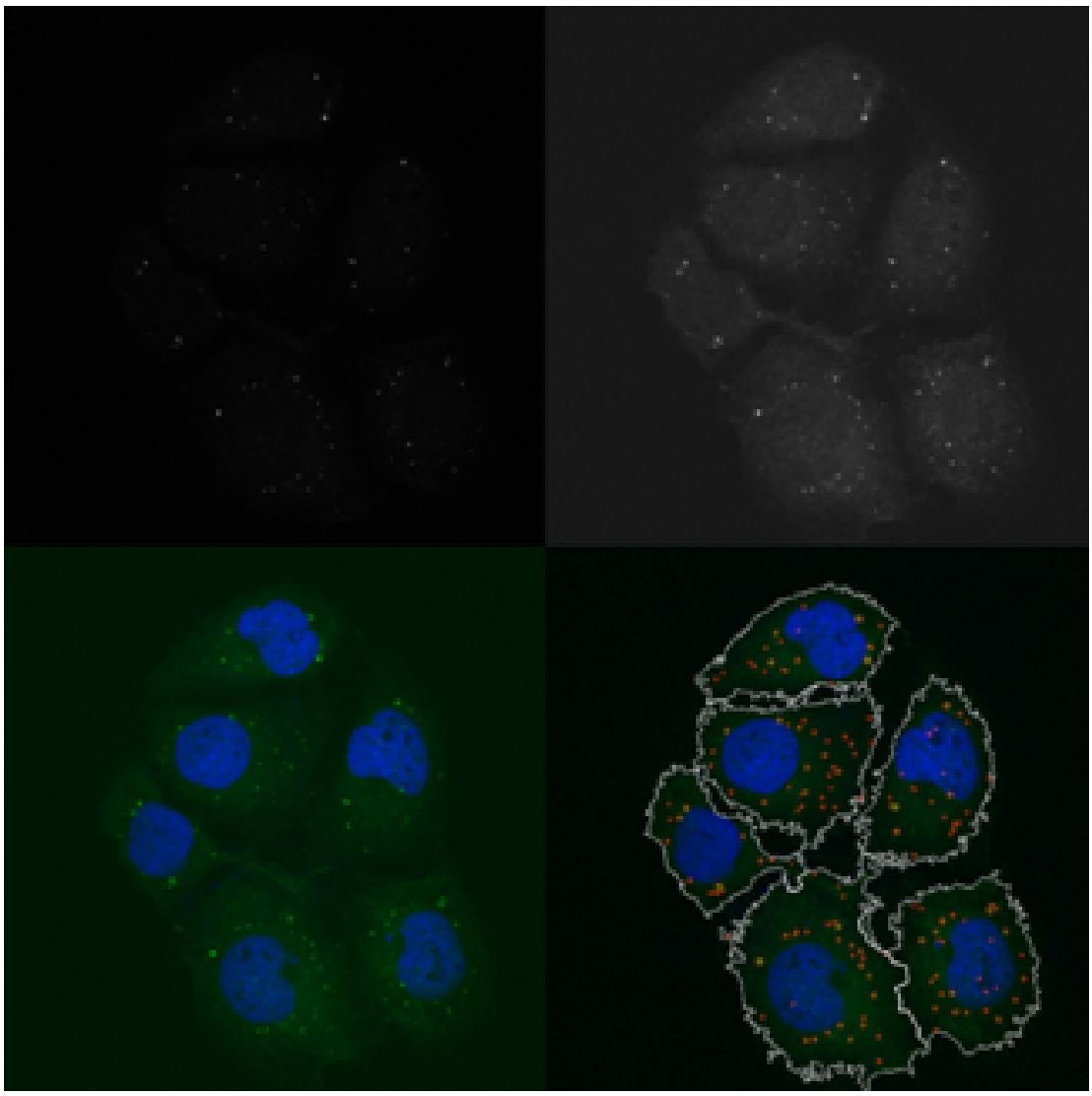

Figure 2.9: Real-world segmentation example. Top panels: GFP signal with normal and gamma corrected scaling to emphasize low intensities, respectively. Bottom left panel: Top right image overlaid with DAPI signal. Bottom right panel: Segmented image with cell outline (white) and vesicle outlines (red).

CHAPTER 3

QUANTITATIVE BLEACHING ESTIMATION(QUBE)

3.1 Introduction

The stoichiometry of molecules in cell biology is an important factor in determining the role of a protein. A protein complex can be nonfunctional without a proper number of constituents or it can interact with other different proteins depending on its current stoichiometry. For example, the hemoglobin of mammals is composed of four globular protein subunits; two alpha subunits and two beta subunits [40]. Without those, the hemoglobin wouldn't function properly because the conformation given by the four subunits is necessary for the binding of oxygen. In protein phosphorylation, the role of phosphorylated proteins is impacted by their stoichiometry [41, 42]. To understand how a protein functions, it is thus necessary to know its exact composition. The methods currently used for analyzing the constitution of complexes aren't appropriate for highly dynamic processes, nor for complexes that cannot easily be extracted for biochemical analysis, either due to their size or due to their stability and that is why the composition of certain protein complexes has been challenging to resolve, for example the protein CENP-A in nucleosomes. This centromere protein's stoichiometry has been a debate in the last years [1–4]. The goal is to develop a new method using state of the art programs to analyze the complexes of important dynamic proteins like CENP-A.

3.2 Centromere protein-A(CENP-A)

Proper chromosome segregation is one of the key functions of cell division. During mitosis, the microtubules attach to chromosomes by a protein complex called the kinetochore. Multiple kinetochores allow a chromosome to be linked to the two spindle poles and could cause defects in the segregation of the chromosomes and lead to aneuploidy [43]. This is why kinetochores need to be singletons. The kinetochores are located at the centromere. The centromere is the part of the chromosome where the kinetochore

assemble. A chromosome is said to be metacentric if the centromere is roughly located in the center. If the lengths are unequal the chromosome is said to be submetacentric. If the centromere is located at the tail of a chromosome it is said to be telocentric and if the entire length of the chromosome acts as a centromere it is said to be holocentric. One main characteristic of the centromere is the presence of centromere protein-A (CENP-A), a histone H3 variant. At the centromere, H3 is replaced by CENP-A and this is believed to mark the centromere epigenetically [44]. However, the exact composition of the centromeric nucleosomes has been subject to extensive debate [1–4]. Centromeric nucleosomes are potential anti-cancer targets. Since they have no known function outside cell division, their inhibition may have no effect on non-dividing cells. To develop such an inhibitor, it is necessary to understand the assembly of centromeric nucleosomes and their composition. Consequently we looked into ways to study the stoichiometry of biological complexes.

3.3 Analysis of complexes

Considering the importance of knowing the exact composition of a complex for understanding its function, many studies have aimed to count the number of subunits of complexes with a fundamental role for the cell, such as proliferation, division or regulation. Multiple biochemistry approaches have been developed to quantify the stoichiometry of proteins.

Western blot

Western blots can be used to detect the stoichiometry of phosphorylation sites. This is achieved by separating the purified sub-unit of a complex via Western blot and then using antibodies to quantify the stoichiometry by looking at the relative strength of the antibodies. This gives the relative composition of each subunits. This approach has some drawbacks; it is time-consuming, it may be hard to isolate and purify the protein complex. [45].

Mass spectrometry

Another powerful tool for analyzing protein complexes is mass spectrometry. This technique analyzes the mass to charge ratio of particles. In biological assays, proteins are studied. The purified protein complex is fragmented into smaller peptides. The fragments are then ionized to generate charged peptides. The mass spectrometer calculates the mass-to-charge ratio. By using a protein database, the proteins inside the protein complex can be characterized and their relative abundance known. Mass spectrometry analysis is fast and efficient. One of the drawbacks of the technique is the extensive purification involved prior to the mass spectrometry and analysis of dynamic change is lost unless the complex is stable in multiple forms. Also it has a limited capability to detect low abundance peptides [46].

Crystallography

By analyzing the X-ray diffraction of a pattern generated by hitting a crystal with an X-ray, it is possible to reconstitute the molecular conformation of biological macromolecules. Crystallography gives high resolution of the conformation and constitution of proteins but the creation of the crystal is tedious, hard and costly. Also, the crystal obtained is static and as such, crystallography is not the proper tool to analyze proteins that change dynamically in conformation/composition.

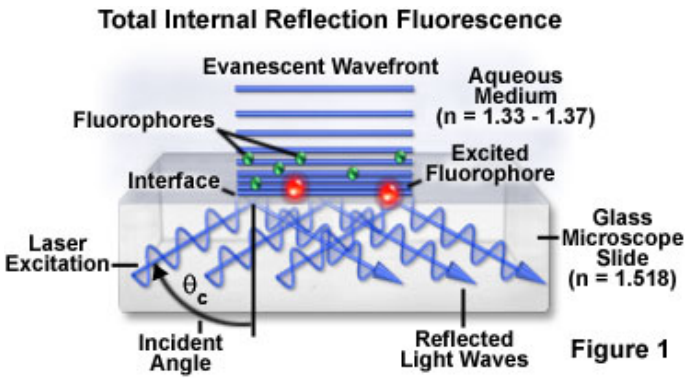

Figure 3.1: Schematic for total internal reflection microscopy reproduced from [47].

Total internal Reflection Microscopy

In cell biology, an abundance of proteins interface with the cell membrane or near the cell membrane. A new microscopic technique was developed to be able to visualize those interactions mitigating the interference of the fluorescent background found in other techniques like confocal microscopy. This technique is based on the total internal reflection fluorescence microscopy (TIRFM) [48]. This method generates a thin evanescent wave of fluorescence that is approximately 100 nm thick (fig 3.1). This wave is produced by an excitating laser hitting a surface with a critical angle. That angle should be high enough that the laser is totally internally reflected instead of refracted. This will create a polarized electromagnetic field with the same frequency as the incident light. This field decays exponentially with the distance from the surface. Thus, the farther the fluorophore is from the surface, the lower the chance of excitation. This gives good signal near the cover slip, reducing the illumination coming from the fluorescent background. TIRFM has been used in multiple different studies:
- Measurement of the binding rates of proteins with the cell membrane receptor of an artificial matrix [49–52].
- Tracking of secreted granules in cells during the secretion process [53] [54].
- Single-molecule dynamics [55][56].
Several studies have combined total internal reflection microscopy with fluorescence recovery rates [49–51] or with resonance energy transfer to study the diffusion of certain proteins inside the cell membrane [57–59].

3.4 Counting with TIRFM

Total internal reflection microscopy is capable of imaging single molecules at a high contrast because of its inherent propriety of illuminating only the fluorophores closest to the cover slip. How is it possible to validate that single-molecules are being observed rather than agglomerates or multiple molecules? To prove this statement, i.e. the assay was effectively made on single-molecules and not on agglomerates, researchers look at the number of photobleaching steps quantified in one optically refracted spot. The maximum resolution of an optical system is limited to the size of its objective and inversely

proportional to the wavelength. This is known as Abbe's diffraction limit:

$$d = \frac{\lambda}{2n \sin \theta} \tag{3.1}$$

A light with a wavelength of λ travelling in an environment with refractive index n and intersecting a spot with angle θ will create a spot of diameter d. In microscopy, this limit is around 200 nm meaning that any structure that is smaller than 200 nm will still have a diameter of the size described by Abbe's limit. Thus, is it possible for multiple small molecules to be located in one optically refracted spot.

Photobleaching is the irreversible loss of the photon emission property of the fluorophore following prolonged light excitation. During the transition from a single state to a triplet state, the fluorophores may interact with other molecules leading to a dark state. In a dark state, the fluorophore has lost the ability to emit photons and thus becomes invisible to detectors of fluorescence. There is also a possibility that the molecule undergoes a process that is reversible and this will create an effect called blinking. Blinking occurs when a fluorophore goes to a dark state and comes back to its prior state. The actual reaction responsible for the blinking of fluorophores hasn't been conclusively identified. Multiple theories have been proposed to explain the blinking of fluorophores like the protonation of the chromophore, the cis-trans isomerization of the chromophore or even the formation of a zwitterion [60]. In general, fluorophores are optimized such that they blink as little as possible.

Photobleaching is a behavior that is usually considered problematic in fluorescent microscopy because it destroys the probe and thus renders proteins undetectable. Nevertheless, it is possible to exploit the property of photobleaching to obtain insight in protein dynamics or in the stoichiometry of small protein complexes. If a complex is supposed to contain one copy of a fluorescent protein, there should only be one sharp drop of intensity when the fluorophore of that protein finally bleaches, while the complex is observed through time. With the same reasoning, complexes of fluorescent proteins

should exhibit as many (or fewer, see below) photobleaching steps as there are fluorophores. Using photobleaching, it is therefore possible to quantify the stoichiometry of a complex or to validate that the molecule observed is really a singleton by counting the number of photobleaching steps [61–68]. However, if the protein complex contains 15 or more fluorophores, it will become very difficult to count the exact copy number, since the intensity profile of the complex fits an exponential decay [69].

In most studies, the research follows the following steps:Samples are fixed to the cover slip and images are acquired with a total internal reflection microscope. After acquisition, homemade software will automatically find the spots in the images and return the intensity profile of those spots as output. An additional step is sometimes undertaken if there is diffusion in the sample. For example, for identifying the number of binding sites to the receptor Nicotinic Acetylcholine, a tracking software was necessary because of the diffusion of the spots [61]. Afterwards, the photobleaching steps are counted manually. Sometimes, to facilitate the interpretation, a filter is used to reduce the noise in the intensity profile. For the analysis of the stoichiometry, a binomial is fit to the observed distribution of bleaching steps to assess the stoichiometry that would explain the observed population of bleaching steps [67]. Moreover, some studies count the subunit populations by the intensity distribution of the spots [70][71]. The problematic with the use of intensities is that the intensity throughout an image as well as the population of spots may vary. The distance to the focal point changes the intensity of the fluorophores and since not all spots are at the same distance, they do not all have the same maximum intensity. The intensity is also subject to the alignment of the fluorophores with the polarized light since TIRFM light is polarized. Those combined effects will create a range of average intensities that is prone to give false results unless corrected but correcting for angle and distance for each spot would be a complicated process. Before drawing conclusions from the total number of steps, several corrections are therefore necessary.

3.5 Complex analysis with TIRFM

All the studies seen thus far have used manual analysis to count the number of bleaching steps. This has many disadvantages. The user may have a bias toward a certain type of spot; bright ones, or isolated ones. This will in turn lead a bias toward the analysis of a subpopulation of the whole data. In any case, this is an unknown factor introduced in the study. Furthermore, to do complex analysis, it is important to have a large number of data collected to extract statistical information and be able to apply stochastic corrections. Studies with manual analysis have a low number of observed spots and thus cannot correct certain biases introduced by the method, such as pre-bleaching, multiple bleaching or labeling ratio.

The segmentation of the image in regions of interest is generally automated and so is the tracking of those spots throughout movies but to achieve a high enough population for complex analysis, the detection of bleaching steps needs to be automated as well. In recent years, researchers have started to develop tools to correct that gap. McGuire H. et al. proposed a method of step detection [72]. The first goal of the algorithm would allow distinction between a bleaching event and a blinking event or pure noise fluctuation. To be able to do exactly this, the algorithm uses an iterative approach. At the start of each iteration, the algorithm calculates a sigma defined by:

$$\sigma = \frac{SNR * N_{ff}}{\varphi} \tag{3.2}$$

where SNR is the signal to noise ratio, N_{ff} is the noise in the intensity fluctuation and φ is an empirical value obtained with a given SNR. If the intensity drops below this sigma, a step is detected. At the end of the iteration, the noise is evaluated again so that the sigma changes between iterations. Also, the algorithm takes the average of short segments of intensity. The size of the segments will increase with each iteration. The algorithm stops when no reduction of steps was obtained. To decrease the number of false positives, three more steps are evaluated. Those represent the minimal times per-

mitted between photobleaching steps, the maximal step amplitude and the minimal step amplitude. To discard other artifacts, the authors included additional criteria to establish a track as viable. A goodness-of-fit x^2 with a threshold of 1.5 is applied on step duration and step amplitude to ascertain the validity of a track. If it fails, it is rejected. While the method does take the user bias out of the equation, the authors failed to use the high number of spots analyzed in any relevant way and analyzed the population by applying a binomial fit on the data with no correction for pre-bleaching or labeling ratio.

Another interesting technique that is used in super-resolution microscopy was developed by Munck and al, called Photobleaching microscopy with non-linear processing (PiMP) [73]. The method is interesting for us as well because even though the technique was developed for confocal and widefield microscopy, the same idea could be applied to TIRF microscopy and photobleaching assays. When a fluorophore bleaches, the image shows peaks and troughs because the bleaching is non-uniform. This random process changes the visible spot population of fluorophores. The main idea is to take the absolute difference of two consecutive images. The rationale is that the difference is caused by photobleaching (fig 3.2). The illumination of the first images is corrected by α_n, the overall brightness ratio, and this gives the equation for the differential image:

$$D_n(x,y) = |\alpha_n I_n - I_{n+1}| \tag{3.3}$$

By looking at the differential image we can get information from multiple fluorophores contained in a diffracted limited spot. In optics, the resolution limit of an optic microscope is defined by Abbe's law : $D_{min} = \frac{\lambda}{2NA}$. This means that anything smaller than the diffraction limit will show as a single spot defined by this limit. Multiple fluorophores, if they are smaller than the diffraction limit, can be in that vicinity but only one spot can be observed. With PiMP, it is possible to go beyond Abbe's limit because the differential image created by the technique contains less information. Only the spot that bleached between two consecutive frames can be seen and thus if two fluorophores are in the same vicinity they will not show in the same image. This holds unless the two flu-

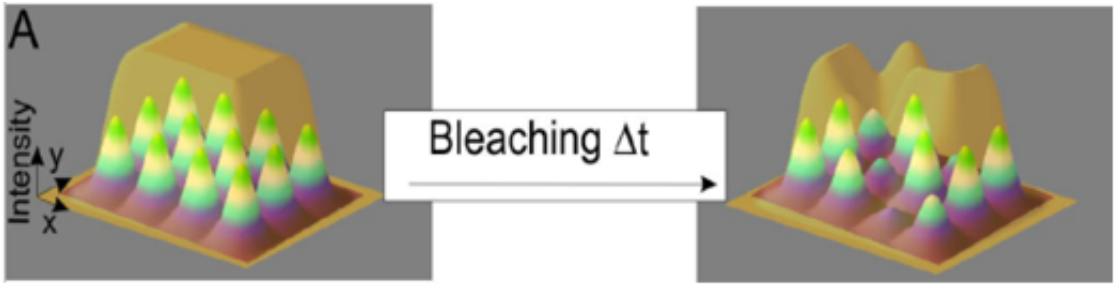

Figure 3.2: Illustration of multiple fluorophores located at the same diffracted limited spot of equal intensity. During imaging, the fluorophores of these complexes bleach and due to statistical uneven bleaching peaks and troughs appear. Reproduced from Figure 1 in [73]

orophores bleach at the same time, but the probability of two fluorophores bleaching at the same time is low as long as exposure time is short and will not contribute statistically to the overall final population calculation. The authors also applied correction to compensate for bias, like labeling density and varying bleach rates. Moreover, the technique was tested on a confocal microscopy assay where they imaged Drosophilia neuromuscular junction in larval tissue. They were able to resolve the ring-like structure of the immunolabeled C-terminal Bruchpilot antigen that was imaged by super-resolution microscopy such as STORM and STED. Interestingly, the method could also be used to count the stoichiometry of complexes with the same technique on a TIRFM. By looking at differential images from total internal reflection microscopy, it would be possible to count the number of fluorophores contained in a complex by counting the number of spots that appeared at the same location in space but not in time. Since a spot in the differential images is a bleaching event, this would amount to the same as counting the number of bleaches without the problem of identifying steps in intensity profiles. For the technique to work properly, a lot of caveats need to be addressed. Proper fluorescent background and intensity correction need to be applied and a complex filter is necessary. Moreover, if the sample isn't fixed, or if there is stage drift, drifting correction is also needed. The parameter also needs to be fine tuned for every experiment but this would be an interesting approach for TIRFM microscopy single-molecule counting.

We proposed an algorithmic pipeline to do the QUantification of photoBleaching Events (QuBE) in TIRFM with minimal human interaction, while compensating for multiple sources of error.

3.6 QuBE

QuBE is a pipeline reuniting four algorithms (fig 3.3) to achieve quantification of photobleaching steps. The four steps consist of the segmentation, a Gaussian mixture fitting, tracking and photobleaching step analysis.

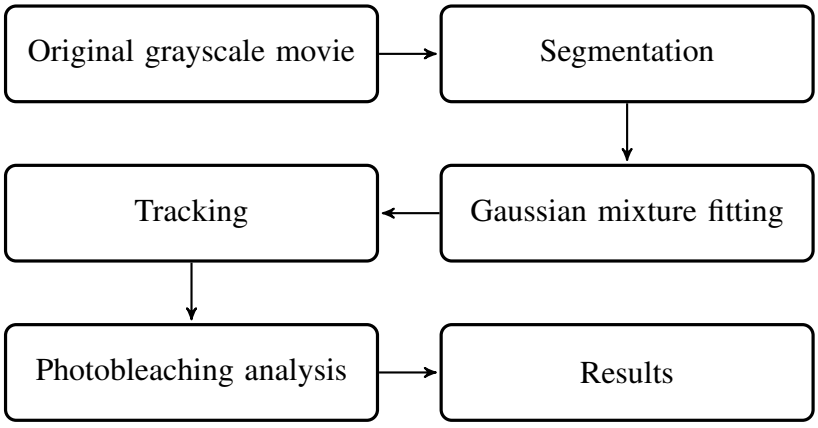

Figure 3.3: Flow chart of QuBE

3.6.1 Segmentation

The first step to be able to automatically quantify fluorescence microscopy videos is to segment the images. The role of the segmentation is to divide the image into a binary image dividing the image into two categories: features and background. To accomplish this, we use the MinSeg algorithm, described in chapter 2. MinSeg is a good choice because it is robust, doesn't require extensive parameter optimization and I had experience with the algorithm. Other powerful segmentation algorithms would also have been a good choice, like MSVST or H-Dome described in chapter 2. For our experimental case, we used MinSeg on the TIRFM images of CENP-A GFP-tagged single molecule complexes. To make the segmentation more robust, a mean of 3 images was used for

each time point. At each time point t, the image is averaged with $t-1$ and $t+1$ to diminish the noise. This is acceptable because of the absence of drift in our experiment but isn't part of the segmentation pipeline. Our features are bound to antibodies attached to a cover slip so any molecule that moves shouldn't be quantified but this is specific to the experiment and not to the method in general. The main parameters used were a false positive expectation of 10^{-8} and the median filter size for the background subtraction was a 13 by 13 window because after testing the parameters, those were found to be the best. We used a lower false positive expectation than the default parameter because of the abundance of spots in the image. With already a high number of features, I think it is better to lose possible features and lower the false positive ratio.

Algorithm 4: Gaussian mixture fitting

input : Grayscale image X and coordinates of features
output: Coordinates and intensities

```
Do a hierarchical clustering based on distance;
foreach cluster c do
    Fit position and intensities of spots in cluster c;
    Test distances of all spots in cluster c;
    Test intensities of all spots in cluster c;
    if A spot was discarded then
        Restart fitting with N − 1 spots;
    else
        Fit N + 1 spots;
        Use F-test to decide between N + 1|N spots;
    end
end
```

3.6.2 Gaussian Mixture fitting

The centroids and intensities of the spots found during segmentation are important for subsequent steps in the pipeline. Thus, the estimation of those values is of utmost importance. For 2D spots, it has been shown that a 2D Gaussian gives a good approximation of the shape of a diffracted limited spot [74]. A Gaussian mixture fitting [75] was introduced to the pipeline to achieve a better estimation of the intensity and localization

of the spots (see algorithm 4). First, we cluster the spots into different groups depending on their overall distance to each other. Spots will belong in the same group if they are lower than 6σ from each other, where σ is the width of the Gaussian representing one spot. Then, for each cluster, every spot in the cluster has its intensity linearly fit because fitting the position and intensity as a non-linear fit takes a lot of computation time. This is followed by a non-linear least square fit for the fluorescent background and position of the type:

$$\min_x |f(x)|^2 = \min_x (f_1(x)^2 + f_2(x)^2 + \cdots f_n(x)^2) \tag{3.4}$$

In our case, the objective function to be minimized $f(x)$ is the residual between the model and the actual image. The Jacobian is given by the gradient of the Gaussian function. The gradient is obtained by taking the first derivative of function:

$$aG(x,y) + bg = I \tag{3.5}$$

where a is the amplitude, $G(x,y)$ is the Gaussian function of the spot at coordinate x, y, bg is the background contribution and I the intensity of the spot. This gives equation:

$$\partial I = ae^{\frac{(i-x)^2}{2\sigma^2}} * ae^{\frac{-(i-x)}{\sigma^2}} \tag{3.6}$$

Where I is the Intensity in the actual image, a is the amplitude previously fitted linearly and $G(x,y)$ is the value of a Gaussian at position x,y of the type: $e^{\frac{-(x-c)}{2\sigma^2}}$

After the fit, two statistical tests are used to further decrease the possible false positives on the spots in the same cluster. The first test is made on the distance between the spots. If the distance between the spots is not statistically significant, the spot with the lowest intensity is discarded and the procedure is started over with one less spot. The second test is made on the significance of the amplitude versus the fluorescent background. If the spot is not significantly higher than the fluorescent background, it is removed and the spots are fitted again. Both tests are Fisher's exact tests and are done with a significance level of α=95%. When all the spots have passed the F-tests, a fit with N+1 spot is made in an attempt to rescue some spots. In practice, it is possible for a spot to be divided by

the segmentation into two centroids, each individual spot intensity could be low enough that they both be discarded because they share the intensity of the real spot, so both centroids would be discarded but by trying to find a spot afterwards it is possible to find the proper centroid. If the fit is significantly better, done with an F-test, the new spot is kept. This improves the sensitivity of the segmentation.

3.6.3 Tracking

Spots were tracked using the u-track algorithm [76]. U-track is a global-nearest neighbor algorithm that makes multiple passes over the data to track features across gaps (e.g. due to blinking). We used the following parameter settings: no merge/split, minimum track length of 5 frames, time window of 10 frames and search radius of 1 to 3 pixel.

3.6.4 Intensity Profile Analysis

Once the tracking for each of the complexes is done, we have a history for each individual feature throughout a movie. With this trajectory, it is possible to extract the intensity of the feature for each timepoint and create what we call an intensity profile. When fluorophores bleach, sudden drops in intensities should be observed. By counting those sudden drops, quantification of the number of fluorophores incorporated into the complexes is possible. To increase the accuracy of the method, a median filter is used to reduce the noise on the intensity profile (fig 3.5). Median filters have the property to reduce noise while preserving edges (i.e. the sharp drops). It is also straightforward to use and robust. After the filtering, the difference of the profile is computed.

$$I_t = \frac{\delta f(t)}{\delta t} = f(t+1) - f(t) \tag{3.7}$$

Where I_t is the intensity after the difference, and $f(t)$ is the intensity at timepoint t in the profile. We calculated the normal probability distribution on the difference I_x because the distribution is mostly symmetric and was found to be a robust estimation (fig 3.6). A t-test is used to test if I_t comes from a normal distribution with a given mean and standard

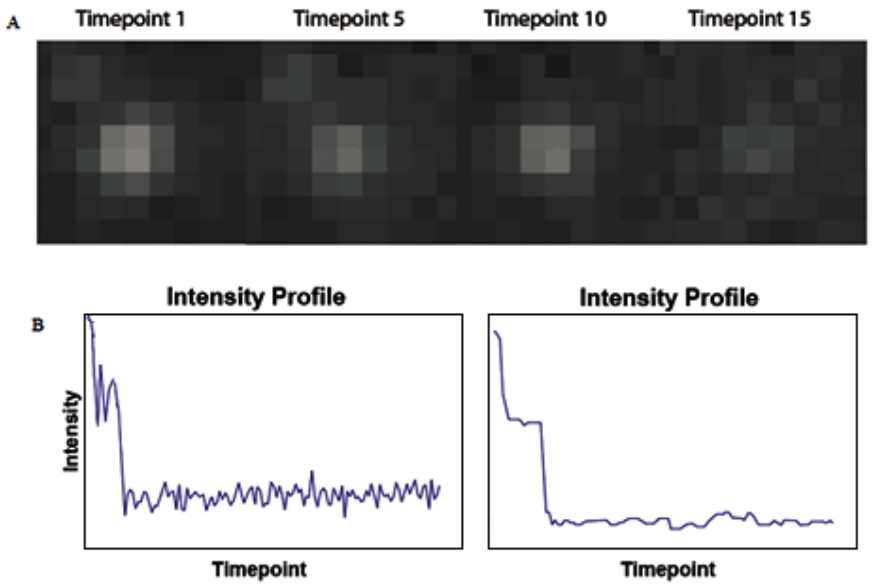

Figure 3.4: Panel A is a spot of interest isolated. Panel B is the intensity profile (left) of the spot shown in panel A and the same profile smoothed with a median filter(right).

deviation. If not, it is considered a possible photobleaching edge. The directionality of the edge is also calculated by looking at the sign of the difference. If the edge is going upward instead of downward, which is what a photobleaching event should produce, the profile is discarded. A sharp upward edge is often the result of a fluorophore blinking. The intensity profile is cut into $n+1$ segments, where n is the number of possible photobleaching steps. A last t-step is made between each segment to ensure that their means are statistically different from each other, if they are not statistically different, they are joined and the juncture is removed. The number of significant intensity drops is the number of photobleaching events. With this number, we can estimate the number of fluorophores present in a complex but we need to take into account the following corrections.

3.6.5 Correction

The following correction scenario is based on the experiment we have done but the rationale could be used for complexes with a higher expected number of molecules. We

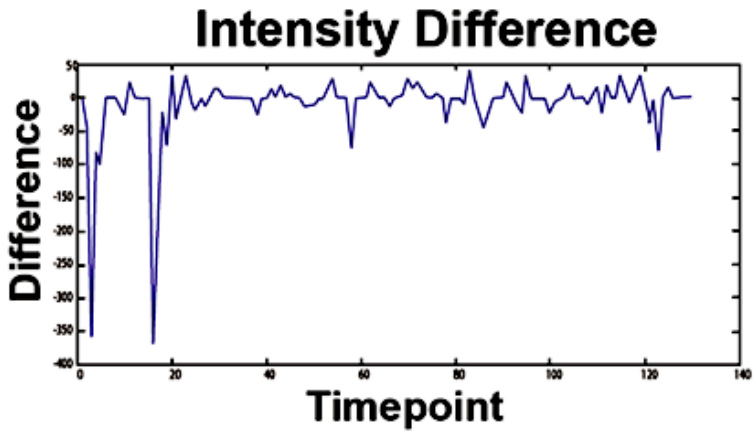

Figure 3.5: Plot of the intensity difference. Noteworthy are the two inverse peaks that correspond with bleaching at times 3 and 18.

wanted to know whether our protein of interest, CENP-A, was present as one or two copies in the centromeric nucleosome, or whether there was a mixture of stoichiometries. If the protein was present as one copy, we would expect a single photobleaching event, if the protein was present as two copies, we would expect single or double photobleaching steps. However, a mixture of single or double events is what we'd expect from a mixture as well. Since we did indeed observe both single and double photobleaching events, the question became whether the observed distribution was indicative of a mixture of stoichiometries or not.

We want to find X, the ratio between the number of complexes containing two copies of CENP-A Δ, and the sum of Δ and the number of complexes with one copy Σ:

$$X = \frac{\Delta}{\Delta + \Sigma} \tag{3.8}$$

If there are only complexes with two copies of CENP-A, X is 1. The raw data cannot be directly interpreted because of three major biases: labeling ratio, pre-bleaching and misclassification.

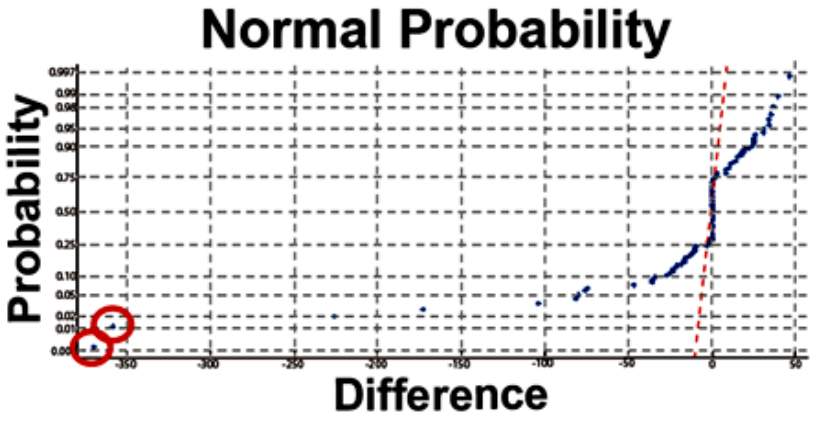

Figure 3.6: Example of a normal probability plot corresponding to the difference of the intensity. The red circle puts emphasis on the two low data points that have a low probability of belonging to the same distribution and corresponding to the two high peaks in (fig 3.5).

Labeling ratio

The first bias is the labeling ratio. A labeling ratio of 100% means that every single copy of CENP-A carries a fluorophore. Since the labeling ratio is not 100%, some complexes will incorporate copies of a CENP-A without a fluorophore. This means that a single photobleaching event that suggests a complex with one copy of CENP-A may in fact come from a complex with two copies of CENP-A, but only one of them carries a fluorophore. So assuming a labeling ratio $L \in [0,1]$:

$$\frac{D_o}{D_o + S_o} = \frac{D_\Delta}{D_\Delta + S_\Delta + S_\Sigma} = \frac{L^2 X}{L^2 X + 2L(1-L)X + L(1-X)} \tag{3.9}$$

Where, D_o is the number of double events observed; S_o is the number of single events observed; D_Δ is the number of complexes with two labeled fluorescence proteins, S_Δ is the number of complexes with two molecules but with only one labeled; S_Σ is the number of complexes with only one molecule labeled. Resolving the equation for X, which is

the ratio we are looking for, we obtain:

$$X = \frac{R_o}{L - R_o + LR_o} \ , \ R_o = \frac{D_o}{D_o + S_o} \tag{3.10}$$

D_o and S_o are the observable ratio and are thus the ratio in the raw data. The labeling ratio can be assessed with a western blot by comparing the expression level of CENP-A-tagged and CENP-A. By doing this, we have estimated the labeling ratio to be around 80% (fig 3.7).

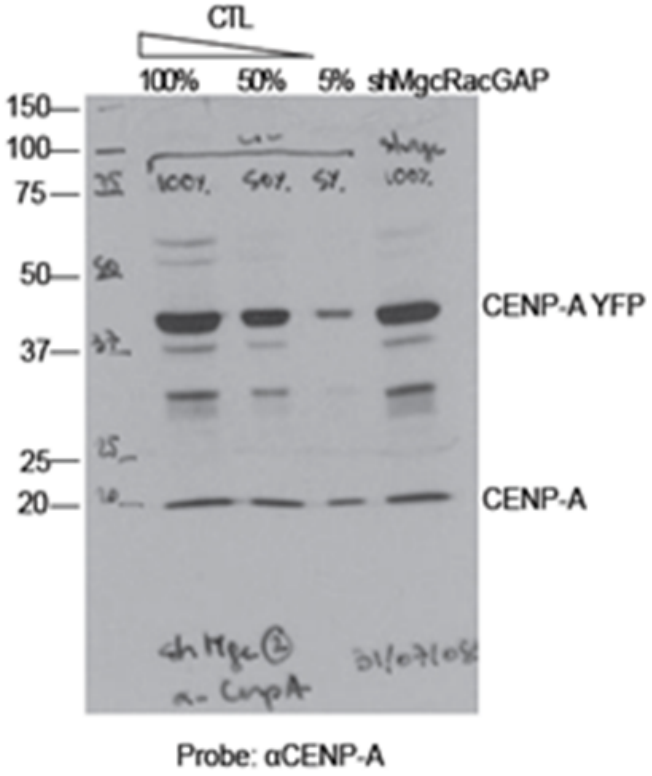

Figure 3.7: Western blot used to calculate the labeling ratio. The labeling is calculated by taking the ratio between the CENP-A and CENP-A-YFP band.

Pre-bleaching

The second bias is what we call pre-bleaching. Pre-bleaching is the degradation of a fluorophore prior to the acquisition of the images. Both the ambient light, and the

medium in which the sample resides in can degrade the capacity of a fluorophore to emit fluorescence and is thus another bias towards the observation of single molecules. Thus, a complex containing two molecules with fluorophores can be observed as a complex with only one molecule with a fluorophore since the other fluorophore could have bleached before the observation. We devised an experiment to estimate the percentage of fluorophores that will be pre-bleached. We used GFP-GST since the ratio of double molecules is supposed to be 1 in this experiment and the labeling ratio is also known and is equal to 1 since every GFP-GST construct should contain two GFP. Given the probability of pre-photobleaching P_b for each fluorophore, we can calculate the estimated fraction of fluorophores that have undergone pre-bleaching:

$$R_o = \frac{D(1-P_b)^2}{D(1-P_b)^2 + 2D(1-P_b)P_b + S(1-P_b)} \tag{3.11}$$

Where D and S are the real number of double and single molecules, but since we observe D_o and S_o, we can change the formula to reflect this and get:

$$D = \frac{D_o}{(1-P_b)^2} \ , \ S = \frac{S_o}{(1-P_b)} - \frac{2DP_b}{(1-P_b)^2} \tag{3.12}$$

Thus

$$R_o = \frac{D_o}{D_o - 2D_o(1-P_b) + S_o(1-P_b)} \tag{3.13}$$

For the experiment of GFP-GST, the discrepancy between single and double is due only to pre-bleaching and knowing that the real number of single molecules should be inexistent, P_b was isolated and estimated to be around 13%.

Misclassification

Another correction that needs to be done is linked to misclassification. The time-lapse between two images isn't short enough that it is impossible for two fluorophores to bleach at the same time in one complex. A single bleaching detected by the method could be in theory a double bleaching happening too fast to be classified properly. To correct that method bias, we look at the time occurrence between each bleaching. What

we would expect to see from that kind of graph would be an exponential decay because fluorophore bleaching is a Poisson event and the difference between two Poisson distributions is an exponential distribution [75]. By fitting the histogram of time between bleaching events with an exponential decay curve, we can estimate the number of events that would happen during a single exposure. Since double events that happen at the same time should look like a single event, we change the number from single to double, effectively correcting for misclassification. The reason we don't directly use the intensity to discern double from single events is because it is not correct to infer that a profile with two times the intensity has two times the number of fluorophores. The intensity values are influenced by the orientation of the fluorophore or its localization in the medium. The light is polarized thus a fluorophore in sync with the polarity of the light would be at its maximum output while a fluorophore almost perpendicular to the polarity would almost have no contribution. In other words, the yield of a given fluorophore is a function of its orientation and an intensity two times greater could mean that there is two times more fluorophores or that the fluorophores are two times more in sync with the light if the relation between orientation and yield is linear. A fluorophore closer to the cover slip will have a stronger emission than a fluorophore farther away, therefore that also influences the contribution of a fluorophore. The fluorophores of a complex are in all probability at different distances from the coverslip and thus do not contribute equally to the intensities values. For all those reasons, I think it is impractical to use intensities directly.

3.7 Results

The aforementioned method was used on the CENP-A-YFP assay and found that $99.0\% \pm 3.4\%$ of CENP-A was present in form of dimers (fig 3.8). As a form of control, the method was also used on a H2B-GFP assay that is known to be a dimer and similar results were obtained, $96.1\% \pm 7.1\%$. Furthermore, manual analysis was also performed to see if both would give similar results (not shown). As expected similar results were obtained by the manual analysis and QuBE. This data suggests that centromeric nucle-

osomes contain two copies of CENP-A. To further test the method, an experiment with cytosolic enhanced green fluorescent protein (eGFP) HeLa cell lysates was performed to see if single events were possible to be captured or if the method was biased toward analyzing only double events. Even though a good part of the population is effectively single, the double population probably reflects the nature of GFP to self-dimerize. While

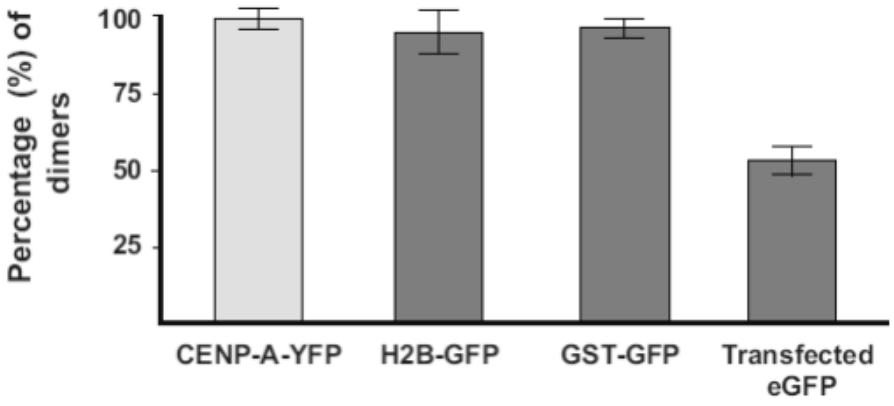

Figure 3.8: Percentage of dimers detected using QuBE with purified nucleosomes of CENP-A-YFG, H2B-GFP, GST-GFP and cellular lysate from eGFP transfected HeLa cells. This is corrected data for labeling ratio, pre-bleaching and misclassification.

we have done our best to correct for possible biases we still have some concerns. While the GFP-GST experiment is a good way to estimate pre-bleaching, it may not be realistic to apply the pre-bleaching of GFP-GST to the CENP-A-YFP experiment for diverse reasons. The regents used are not the same, the time taken to prepare the sample isn't exactly the same and the percentage can probably vary from user to user since people don't exactly do the same manipulations at the same speed. The GFP-GST experiment did give us an insight in the magnitude of the population that undergo pre-bleaching and thus helped us alleviate the problem and also shows that a sensible population of fluorophores does bleach before the acquisition starts. It is also noteworthy to point out that all those corrections are possible because of the high number of analyzed particles. It would be pointless to use the same correction with a manual analysis because those corrections are only applicable if the data set is large enough. While QuBE works well for step analysis of small protein complexes, the higher the possible number of steps, the

more corrections are needed and the harder it is to detect the steps. The steps become less statistically significant the more fluorophores are present in an optically refracted spot because the overall contribution of each is lessened. In other words, if you have two fluorophores, and one fluorophore bleaches, you lose 50% of your intensity, but if you have ten fluorophores present and one bleaches, you lose 10% of your intensity making the drop less significant. This makes the approach of using a t-test to single out steps less robust. The algorithm was implemented in Matlab. In terms of speed, QuBE analyzes a video in roughly 1 hour. The Gaussian mixture fitting is the slowest algorithm in the pipeline. The fitting of multiple variables can be extremely computationally costly.

3.8 Conclusion

We have shown that QuBE can be applied to automatically analyze single-molecule videos in experimental controls and employed successfully to the characterization of CENP-A. One possible improvement would be in adapting the algorithm for multithreading, taking the advantage of multiprocessor computers. While every step in the pipeline needs to be done sequentially, the steps themselves could be faster by exploiting parallel programming. The segmentation of the videos could be switched to multithreads by segmenting one image per thread. The same rationale can be applied for the Gaussian mixture fitting. This would make the analysis even faster. Furthermore, we would like to do an implementation in a more open source language, like C/C++ making it available for image processing package such as openCV. We are planning to publish Qube in the near future.

CHAPTER 4

CONCLUSION

In chapter 2, I presented a new robust, versatile and easy to use segmentation algorithm for fluorescent microscopy images, called MinSeg. In practice, it is easy to use because it takes the tuning of a single parameter, making it a powerful tool for the segmentation of arbitrary features. Like previously stated, the main drawback of the method is the minimum sampling needed for the algorithm to work. If the sampling of the feature is not large enough, the method cannot distinguish the feature from the noise. Something to keep in mind is that the minimum sampling is linked to the signal-to-noise ratio. The higher the signal-to-noise ratio the lower the sampling can be before the algorithm breaks. Knowingly, if the sampling is too low, another powerful segmentation algorithm like MSVST should be used instead. On the computational time, MinSeg is satisfactory for most practical applications, its bottleneck being the background subtraction that was optimized. If the speed ever becomes a problem I would suggest to do a background subtraction beforehand or use a different algorithm altogether. One main improvement that we would like accomplish is to adapt the algorithm to a greater range of noise patterns like the exponential noise pattern found in SCMOS cameras. This would extend the robustness and versatility of the algorithm. Another line of inquiry would be to try new background subtraction methods or new noise estimation methods and compared them with the results we have with the method we are currently using. Another thing I would have liked to do is create an online resource that permits you to upload a figure, input the parameter, do the segmentation server-side and see the segmentation. Most microscope setups are computerized but the actual image processing is made on another computer. This has the disadvantage that you need to take a picture with the microscope, transfer it to the analysis computer and then try a segmentation to see if your experimental setup is in line with the restriction of the algorithm. Having that online resource, you could take the picture, go online, try the segmentation and directly see if the segmentation works instead of acquiring the data and hoping that your acquisition setup is good enough for

the algorithm.

In chapter 3, I presented QuBE, a novel automated pipeline using state of the art algorithms for the quantification of photobleaching events. QuBE was employed successfully to characterize CENP-A and while the overall method is restricted to protein complexes with a low copy-number of subunits, the quantity of data analyzed by the method gives the possibility to correct for some experimental biases by using statistic that couldn't be applied without a large data set. The method corrects for three main biases, the label ratio, the pre-bleaching and the misclassification. Furthermore, the fact that the method is fully automated frees the experiment from possible user bias. While the speed of the method wasn't problematic in our experimental setup, if it proved to be a problem in the future, I would suggest dropping the Gaussian mixture fitting because while it helps make the overall method more robust, the rationale behind the whole method still works without it. The Gaussian mixture fitting is the slowest algorithm in the pipeline and is thus a good target to optimize. Another solution to the computational speed would be to re-implement the code to use parallel processing. This could improved the speed of the segmentation step, the Gaussian mixture fitting step and the overall analysis. While all the algorithms must be done sequentially, it is possible to parallelize each step individually. I would have also liked to code the pipeline in a more open source language or image processing package like openCV. While matlab has the advantage of having a large image processing package that comes with a lot of build-in functions, this also limits the availability of the program and its usefulness.

An answer is only as good as the question. In the same line of thought, a tool is only as useful as the one using it. Overcomplication brings obfuscation. That is why developing tools that are easier of approach and utilization while maintaining their robustness and strength is important. MinSeG achieves robust segmentation with ease of use and was shown to be usable in practical biological inquiries. On the other hand, QuBE by using a combination of multiple algorithms is not as easy to use but it is at least modular. This means that every algorithm chosen to accomplish a step in the pipeline can

be replaced easily and thus tailored by the user for their needs or knowledge. The work on the quantification of photobleaching steps in single-molecule experiments has also brought forth attention to the different biases present intrinsically to this kind of experiment. The importance of the corrections in the quantification of photobleaching steps in single-molecule experiments can't be stress enough. I think that the community that undertakes this kind of quantification needs to be sensitized to the different biases and needs to use proper tools to correct for those biases. While QuBE is not without fault it is a step in the right direction.

BIBLIOGRAPHY

[1] R. Camahort M. Shivaraju M. Mattingly B. Li S. Nakanishi D. Zhu A. Shilatifard J. L. Workman and J. L. Gerton. Cse4 is part of an octameric nucleosome in budding yeast. *Mol Cell*, 35(6):794–805, 2009.

[2] E. K. Dimitriadis C. Weber R. K. Gill S. Diekmann and Y. Dalal. Tetrameric organization of vertebrate centromeric nucleosomes. *Proc Natl Acad Sci*, 107(47): 20317–22, 2012.

[3] I. J. Kingston J. S. Yung and M. R. Singleton. Biophysical characterization of the centromere-specific nucleosome from budding yeast. *J Biol Chem*, 286(5):4021–6, 2011.

[4] J. S. Verdaasdonk and K. Bloom. Centromeres: unique chromatin structures that drive chromosome segregation. *Nat Rev Mol Cell Biol*, 12(5):320–32, 2011.

[5] Boisvert J Ladouceur AM Dorn JF Maddox PS. Padeganeh A, Ryan J. Octameric cenp-a nucleosomes are present at human centromeres throughout the cell cycle. *Curr Biol*, 9(23):764–9, 2013.

[6] A. Q. Amara Sascha P. Pynpoint: an image processing package for finding exoplanets. *Monthly notices of the royal astronomcal society*, 427(2):948–955, 2012.

[7] Y. W. L Hong A Jain. Fingerprint image enhancement: algorithm and performance evaluation. *Pattern Analysis and Machine Intelligence IEEE Transactions*, 20(8): 777–789, 1998.

[8] A. M. L. Adeshina Siong-Hoe; Loo Chu-Kiong. Real-time facial expression recognitions: A review. *Conference on innovative technologies in intelligent systems and industrial applications*, pages 375–378, 2009.

[9] D. J. Stephens and V. J. Allan. Light microscopy techniques for live cell imaging. *Science*, 300(5616):82–86, 2003.

[10] J. F. Dorn G. Danuser and G. Yang. Computational processing and analysis of dynamic fluorescence image data. *Methods in cell biology*, 85:497–+, 2008.

[11] Q. Wu F. A. Merchant and K. R. Castleman. *Microscope image processing*. Elsevier/Academic Press, 2008.

[12] N. Otsu. A threshold selection method from gray-level histograms. *EEE Trans. Sys.*, 9(1):62–66, 1979.

[13] P. Ruusuvuori T. Aijo S. Chowdhury C. Garmendia-Torres J. Selinummi M. Birbaumer A. M. Dudley L. Pelkmans and O. Yli-Harja. Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images. *BMC Bioinformatics*, 11:248, 2010.

[14] I. Smal M. Loog W. Niessen and E. Meijering. Quantitative comparison of spot detection methods in fluorescence microscopy. *IEEE Trans Med Imaging*, 20(2): 282–301, 2010.

[15] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:511–518, 2001.

[16] G. J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Interscience Publication, 2004.

[17] M. Piccardi. Background subtraction techniques: a review. *IEEE Internation Conference on Systems Man. and Cybernetics*, 1-7:3099–3104, 2004.

[18] A. V. Aho J. E. Hopcroft and J. D. Ullman. *The design and analysis of computer algorithms*. Addison-Wesley Pub, 1974.

[19] S. Perreault and P. Hebert. Median filtering in constant time,âĂİ ieee transactions on image processing. *IEEE Transactions on Image Processing*, 16(9):2389–2394, 2007.

[20] M. O. Ahmad and D. Sundararajan. A fast algorithm for two-dimensional median filtering. *IEEE Transactions on Circuits and Systems*, 34(11):1364–1374, 1987.

[21] J. Boisvert. Github code, 2013. URL https://github.com/maurtheus/probabilistic-segmentation.

[22] A. A. a. E. Dubois. Fast and reliable structure-oriented video noise estimation. *IEEE Trans. Circuits Syst. Video Technol*, 15(1), 2005.

[23] G. V.-S.-F. S. Aja-Fernandez M. Martin-Fernandez and C. Alberola-Lopez. Automatic noise estimation in images using local statistics. additive and multiplicative cases. *Image Vis. Comput.*, 27(6), 2009.

[24] J.-S. Lee. Refined filtering of image noise using local statistics. *Comput. Vision Graphics Image Process*, 15:380–389, 1981.

[25] J. Immerkaer. Fast noise variance estimation. *Comput. Vision Image Underst.*, 64 (2):300–302, 1996.

[26] C. B. a. M. V. R. New method for noise estimation in images. *Proc. IEEE-Eurasip Int. Workshop on Nonlinear Signal and Image Processing*, 1:290–293, 2005.

[27] S.-M. Yang and S.-C. Tai. Fast and reliable image-noise estimation using a hybrid approach. *Journal of Electronic Imaging*, 19(3), 2010.

[28] R.-H. P. D.-H. Shin S. Yang and J.-H. Jung. Block-based noise estimation using adaptive gaussian filtering. *IEEE Trans. Consum. Electron*, 51(1):218–226, 2005.

[29] L. Vincent. Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms. *IEEE Transactions on Image Processing*, 2(2):176–201, 1993.

[30] B. Zhang M. J. Fadili J. L. Starck and J. C. Olivo-Marin. Multiscale variance-stabilizing transform for mixed-poisson-gaussian processes and its applications in bioimaging. *IEEE International Conference on Image Processing*, 1-7:3029–3032, 2007.

[31] R. C. Gonzalez and R. E. Woods. *Digital image processing*. Pearson Prentice Hall, 2008.

[32] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603âĂŞ619, 2002.

[33] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-*, 57(1):289–300, 1995.

[34] C. Laflamme G. Assaker D. Ramel J. F. Dorn D. She P. S. Maddox and G. Emery. Evi5 promotes collective cell migration through its rab-gap activity. *J Cell Biol*, 198(1):57–67, 2-12.

[35] A. Padeganeh J. Ryan J. Boisvert A. M. Ladouceur J. F. Dorn and P. S. Maddox. Octameric cenp-a nucleosomes are present at human centromeres throughout the cell cycle. *Curr Biol*, 23(9):764–9, 2013.

[36] J. Cousty G. Bertrand L. Najman and M. Couprie. Watershed cuts: minimum spanning forests and the drop of water principle. *IEEE Trans Pattern Anal Mach Intell*, 31(8):1362–74, 2009.

[37] M. Cargnello J. Tcherkezian J. F. Dorn E. L. Huttlin P. S. Maddox S. P. Gygi and P. P. Roux. Phosphorylation of the eukaryotic translation initiation factor 4e-transporter (4e-t) by c-jun n-terminal kinase promotes stress-dependent p-body assembly. *Mol Cell Biol*, 32(22):4572–84, 2013.

[38] L. Yang Z. Qiu A. H. Greenaway and W. P. Lu. A new framework for particle detection in low-snr fluorescence live-cell images and its application for improved particle tracking. *IEEE Transactions on Biomedical Engineering*, 59(7): 2040–2050, 2012.

[39] L. Yang R. Parton G. Ball Z. Qiu A. H. Greenaway I. Davis and W. P. Lu. An adaptive non-local means filter for denoising live-cell images and improving particle detection. *Journal of Structural Biology*, 172(3):233–243, 2010.

[40] M. F. Perutz. Haemoglobin: structure, function and synthesis. *Br Med Bull*, 32(3): 193–4, 1976.

[41] Wu R Haas W Dephoure N Huttlin EL Zhai B Sowa ME and Gygi SP. A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat Methods*, 8(8):677–83, 2011.

[42] Tan CS and Bader GD. Phosphorylation sites of higher stoichiometry are more conserved. *Nat Methods*, 9(4):317, 2011.

[43] D. W. Cleveland Y. Mao and K. F. Sullivan. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell*, 112(4):407–21, 2003.

[44] B. E. Black M. A. Brock S. Bedard V. L. Woods Jr. and D. W. Cleveland. An epigenetic mark generated by the incorporation of cenp-a into centromeric nucleosomes. *Proc Natl Acad Sci*, 104(12):5008–13, 2007.

[45] Schleich K Warnken U Fricker N OztÃijrk S Richter P Kammerer K SchnÃűlzer M Krammer PH and Lavrik IN. Stoichiometry of the cd95 death-inducing signaling complex: experimental and modeling evidence for a death effector domain chain model. *Mol Cell*, 47(2):306–19, 2012.

[46] Prakash A Piening B Whiteaker J Zhang H Shaffer SA Martin D Hohmann L Cooke K Olson JM Hansen S Flory MR Lee H Watts J Goodlett DR Aebersold R Paulovich A and Schwikowski B. Assessing bias in experiment design for large scale mass spectrometry-based quantitative proteomics. *Mol Cell Proteomics*, 389 (10):1017âĂŞ31, 2007.

[47] Nikon. Tirf schematic, 2013. URL http://www.microscopyu.com/articles/fluorescence/tirf/tirfintro.html.

[48] D. Axelrod. Total internal reflection fluorescence microscopy in cell biology. *Methods Enzymol*, 361(1):1–33, 2003.

[49] N. L. Thompson T. P. Burghardt and D. Axelrod. Measuring surface dynamics of biomolecules by total internal reflection fluorescence with photobleaching recovery or correlation spectroscopy. *Biophys J*, 33(3):435–54, 1981.

[50] J. Ries E. P. Petrov and P. Schwille. Total internal reflection fluorescence correlation spectroscopy: effects of lateral diffusion and surface-generated fluorescence. *Biophys J, vol*, 95(1):390–9, 2008.

[51] R. Gilmanshin C. E. Creutz and L. K. Tamm. Annexin iv reduces the rate of lateral lipid diffusion and changes the fluid phase structure of the lipid bilayer when it binds to negatively charged membranes in the presence of calcium. *Biochemistry*, 33(27):8225–32, 1994.

[52] T. Funatsu Y. Harada M. Tokunaga K. Saito and T. Yanagida. Imaging of single fluorescent molecules and individual atp turnovers by single myosin molecules in aqueous solution. *Nature*, 374(6522):555–9, 1995.

[53] T. Lang I. Wacker I. Wunderlich A. Rohrbach G. Giese T. Soldati and W. Almers. Role of actin cortex in the subplasmalemmal transport of secretory granules in pc-12 cells. *Biophys J*, 78(6):2863–77, 2000.

[54] J. L. Johnson J. Monfregola G. Napolitano W. B. Kiosses and S. D. Catz. Vesicular trafficking through cortical actin during exocytosis is regulated by the rab27a effector jfc1/slp1 and the rhoa-gtpase-activating protein gem-interacting protein. *Mol Biol Cell*, 23(10):1902–16, 2012.

[55] N. T. Umbreit D. R. Gestaut J. F. Tien B. S. Vollmar T. Gonen C. L. Asbury and T. N. Davis. The ndc80 kinetochore complex directly modulates microtubule dynamics. *Proc Natl Acad Sci*, 109(40):16113–8, 2012.

[56] H. Yamamura C. Ikeda Y. Suzuki S. Ohya and Y. Imaizumi. Molecular assembly and dynamics of fluorescent protein-tagged single kca1.1 channel in expression system and vascular smooth muscle cells. *Am J Physiol Cell Physiol*, 302(8): C1257–68, 2012.

[57] C. E. Fowler P. Aryal K. F. Suen and P. A. Slesinger. Evidence for association of gaba(b) receptors with kir3 channels and regulators of g protein signalling (rgs4) proteins. *J Physiol*, 580(1):51–65, 2007.

[58] I. Riven E. Kalmanzon L. Segev and E. Reuveny. Conformational rearrangements associated with the gating of the g protein-coupled potassium channel revealed by fret microscopy. *Neuron*, 38(2):225–35, 2003.

[59] D. S. Martin R. Fathi T. J. Mitchison and J. Gelles. Fret measurements of kinesin neck orientation reveal a structural basis for processivity and asymmetry. *Proc Natl Acad Sci U S A*, 107(12):5453–8, 2010.

[60] H. E. Seward and C. R. Bagshaw. The photochemistry of fluorescent proteins: implications for their biological applications. *Chem Soc Rev*, 38(10):2842–51, 2009.

[61] P. D. Simonson H. A. Deberg P. Ge J. K. Alexander O. Jeyifous W. N. Green and P. R. Selvin. Counting bungarotoxin binding sites of nicotinic acetylcholine receptors in mammalian cells with high signal/noise ratios. *Biophys J*, 99(10): L81–3, 2010.

[62] N. Groulx H. McGuire R. Laprade J. L. Schwartz and R. Blunck. Single molecule fluorescence study of the bacillus thuringiensis toxin cry1aa reveals tetramerization. *J Biol Chem*, 286(49):42274–82, 2011.

[63] K. Kitamura M. Tokunaga A. H. Iwane and T. Yanagida. A single myosin head moves along an actin filament with regular steps of 5.3 nanometres. *Nature*, 397 (6715):129–34, 1999.

[64] M. C. Leake J. H. Chandler G. H. Wadhams F. Bai R. M. Berry and J. P. Armitage. Stoichiometry and turnover in single, functioning membrane protein complexes. *Nature*, 443(7109):355–8, 2006.

[65] M. H. Ulbrich and E. Y. Isacoff. Subunit counting in membrane-bound proteins. *Nat Methods*, 4(4):319–21, 2007.

[66] M. Tokunaga K. Kitamura K. Saito A. H. Iwane and T. Yanagida. Single molecule imaging of fluorophores and enzymatic reactions achieved by objective-type total internal reflection fluorescence microscopy. *Biochem Biophys Res Commun*, 235 (1):47–53, 1997.

[67] K. Nakajo M. H. Ulbrich Y. Kubo and E. Y. Isacoff. Stoichiometry of the kcnq1 - kcne1 ion channel complex. *Proc Natl Acad Sci*, 107(44):18862–7, 2010.

[68] W. Ji P. Xu Z. Li J. Lu L. Liu Y. Zhan Y. Chen B. Hille T. Xu and L. Chen. Functional stoichiometry of the unitary calcium-release-activated calcium channel. *Proc Natl Acad Sci*, 105(36):13668–73, 2008.

[69] S. K. Das M. Darshi S. Cheley M. I. Wallace and H. Bayley. Membrane protein stoichiometry determined from the step-wise photobleaching of dye-labelled subunits. *Chembiochem*, 8(9):994–9, 2007.

[70] J. L. Swift A. G. Godin K. Dore L. Freland N. Bouchard C. Nimmo M. Sergeev Y. De Koninck P. W. Wiseman and J. M. Beaulieu. Quantification of receptor tyrosine kinase transactivation through direct dimerization and surface density measurements in single cells. *Proc Natl Acad Sci*, 108(17):7016–21, 2011.

[71] A. G. Godin S. Costantino L. E. Lorenzo J. L. Swift M. Sergeev A. Ribeiro da Silva Y. De Koninck and P. W. Wiseman. Revealing protein oligomerization and densities in situ using spatial intensity distribution analysis. *Proc Natl Acad Sci*, 108(17):7010–5, 2011.

[72] H. McGuire M. R. Aurousseau D. Bowie and R. Blunck. Automating single subunit counting of membrane proteins in mammalian cells. *J Biol Chem*, 287(43): 35912–21, 2012.

[73] S. Munck K. Miskiewicz R. Sannerud S. A. Menchon L. Jose R. Heintzmann P. Verstreken and W. Annaert. Sub-diffraction imaging on standard microscopes through photobleaching microscopy with non-linear processing. *J Cell Sci*, 125 (9):2257–66, 2012.

[74] A. Santos and I. T. Young. Model-based resolution: applying the theory in quantitative microscopy. *Applied Optics*, 39(17):2948–2958, 2000.

[75] J. G. Skellam. The frequency distribution of the difference between 2 poisson variates belonging to different populations. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 109(3):296, 1946.

[76] K. Jaqaman D. Loerke M. Mettlen H. Kuwata S. Grinstein S. L. Schmid and G. Danuser. Robust single-particle tracking in live-cell time-lapse sequences. *Nature Methods*, 5(8):695–702, 2008.

GLOSSARY

binary image

A binary image is an image that contains two possibles values, often 0 and 1. Binary image are frequently used as logical mask. p2, p7, p15, p37

bit depth

The bit depth is the number of bits used to quantify the intensity of a pixel. For example, an image with a bit depth of 8 bits, can have values ranging from 0 to 255. p9, p10, p26

DAPI

4',6-diamidino-2-phenylindole is a fluorescent stain that binds to A-T rich region in DNA. DAPI is used as a blue fluorescent stain for DNA. p25, p27

diffraction limit

The resolution of a microscope is fundamentally limited by diffraction. The limit is bound by Abbe's law : $d = \frac{\lambda}{2nsin\theta}$, where λ is the wavelength of the light, $nsin\theta$ is also called numerical aperture and d is the minimum diameter. For a microscope using green light, this resolution is roughly 250 nm. This means that anything smaller than 250 nm will show like a spot with a diameter of 250 nm. This is an important concept in subcellular microscopy because most proteins are smaller than the diffraction limit . p4, p8

fluorescent background

Background fluorescence can originate from autofluorescence, non-specific staining or diffuse fluorescent molecules. In this text, background artifacts, background fluorescent and inhomogeneous illumination called background fluorescent plural. p5–p8, p11–p13, p17, p25, p26, p30, p31, p36, p39

hot pixel

A bright dot defect is when a pixel is always "on" and has the maximum value. Those pixel can easily be found by taking an empty image that should only contain

noise and locate the pixels with the maximum value. This defect is known in the field as hot pixel. p6

kernel

A kernel, mask or filter window is a small matrix. A convolution is made with the image and an image to produce a desirable effect like edge detection, blurring, sharpening or smoothing. plural. p7–p10, p12, p19, p26

labeling ratio

Since not every tagged-protein will incorporate a GFP, it is possible for a complex to have less GFP present in the complex than the number of copies of the tag-protein. Labeling correction tries to correct for this bias. p34, p35, p42

mask

See kernel. p9, p15–p18

P-body

Processing bodies are distinct foci within the cytoplasm. It has been shown that P-bodies are involved in mRNA decay. p24, p25

pre-bleaching

Pre-bleaching is the degradation of a fluorophore prior to the acquisition of the images. p34, p35, p42

quantum yield

number of photon emitted versus the number of photon absorbed. A high quantum yield denotes a higher energy transfer from absorption to emission. p6

salt and pepper noise

An image with salt and pepper noise will have dark pixels and hot pixels. Degradation in an image can lead to this type of noise or bit errors in transmission. Dead pixels and hot pixels in a camera can produce this type of noise but will be non-random and always have the same signature. This type of noise is usually reduced by using a median filter. p8

signal-to-noise ratio

Signal-to-noise ratio is a measure to evaluate the level of the signal versus the background noise. Multiple definition for the SNR exists, when I refer to SNR I refer to the following definition : $SNR = \frac{\mu}{sigma}$, where μ is the mean of the signal and σ is the standard deviation of the noise. p2–p4, p7, p11, p17, p19–p22, p34, p50

Sobel

Sobel operator is an image gradient filter. For a 2D image, the operator uses two filters, one in the x dimension and one in the y dimension. The operator has the form :

$$Sobel\ operator = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} .$$ p14

tessellation

Tessellation of an image is the fragmentation of the image into blocks that do not overlap and create no gaps. p13

zwitterion

A zwitterion is a dipolar ion that is neutral. It has both a negative and a positive charge. The zwitterion can be present in both a cationic and an anionic form depending on the environment. p32