12-2005

Automated Analysis of Fluorescent Microscopic Images to Identify Protein-Protein Interactions

Sankar Venkataraman
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Sankar Venkataraman entitled "Automated Analysis of Fluorescent Microscopic Images to Identify Protein-Protein Interactions." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Electrical Engineering.

Hairong Qi, Mitchel John Doktycz, Major Professor

We have read this thesis and recommend its acceptance:

Mohammed Ferdjallah

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Sankar Venkataraman entitled “Automated analysis of fluorescent microscopic images to identify protein-protein interactions”. I have examined the final electronic copy of the thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Electrical Engineering.

Hairong Qi
Co-Major Professor

Mitchel John Doktycz
Co-Major Professor

We have read this thesis
and recommend its acceptance:

Mohammed Ferdjallah

Acceptance for the Council:

Anne Mayhew
Vice Chancellor and Dean of
Graduate Studies

(Original signatures are on file with official student records.)

AUTOMATED ANALYSIS OF FLUORESCENT MICROSCOPIC IMAGES TO IDENTIFY PROTEIN-PROTEIN INTERACTIONS

A Thesis
Presented for the
Master of Science Degree
The University of Tennessee, Knoxville

Sankar Venkataraman
December 2005

Dedication

This thesis is dedicated to my parents Venkataraman Kalyanaraman, Jayanthi Venkataraman, my sister Sri Vidya and *bawa* Bhasker.

Acknowledgements

I would like to thank *amma* and *appa* for providing me with a platform to study without any hindrances. I know the struggle they've been through to raise a brat like me and orienting me towards studies was quite a nightmare for them and I thank them for not giving up on me. I have to mention the role of my sister in making me what I am today. She is not exactly my inspiration, but was and still is a constant source of support and if not for her, I am sure I would have happily landed up making *dosas* at an *udipi* restaurant. But, here I am and I can never thank her enough. Thank you for everything V*izzia*.

I would like to thank Dr. Doktycz, my advisor at Oak Ridge National Laboratories, for providing me with this opportunity to work under him at a time when people were not convinced I could do good research. I owe a lot to Dr. Doktycz for believing in me. I can never forget the day when Dr. Qi, my advisor at The University of Tennessee, Knoxville, said that "things have to even out and since you've been through enough trouble, try not to worry as your good times will start soon". I owe it to her for making me believe in myself and for being such a wonderful advisor. I would like to thank Dr. Mohammed Ferdjallah for reviewing my thesis work and being on my committee. I also would like to thank Jenny for her valuable insights and for the images she acquired during my work. I could not have done any analysis without her help or her images.

There are a lot of people who do not have much to do with this thesis but have been a constant source of support and inspiration and I would like to thank them for that. I really appreciate the support of my lab mates *Balls*, who actually listened to a lot of my weird

thoughts (thanks for being there *macha*), Gaurav, who was not exactly an ideal lab mate but quite a queer one to have (in a good way bud), Hongtao and Lidan who were always there with a smile on their face. I should really thank *Teja bhai* and *Sagi bhai* for their guidance and without whom I'd have missed out on so much fun. I would like to thank all the people in my apartment complex for feeding me day or night (literally). Thank you guys, without you I'd starve, or worse, cook everyday. I would like to thank my apartment mate Adrija, for not kicking me out of the house and tolerating my nuisance.

I want to thank Rathu and Sirsa for being such good friends and a source of support during the last few years. I want to thank all my friends Sridhar, Sagar, Ranjith, Prashant, Bond, Chandu, Sudheer, Safia, Sisira, Ruma (I think I should stop the list; a thesis has some limits to how many pages it can have you know) and a hundred others for being there for me.

Now that the most important part is done with, I will begin the less interesting stuff.

Abstract

The identification and confirmation of protein interactions significantly challenges the field of systems biology and related bio-computational efforts. The identification of protein-protein interactions along with their spatial and temporal localization is useful for assigning functional information to proteins. Fluorescence microscopy is an ideal method for assessing protein localization and interactions as a number of techniques and reagents have been described. Historically, data sets obtained from fluorescence microscopy have been analyzed manually, a process that is both time consuming and tedious. The development of an automated system that can measure the location and dynamics of interacting proteins inside a live cell is of high priority. This paper describes an automated image analysis system used to identify an interaction between two proteins of interest. These proteins are fused to either Green Fluorescent Protein (GFP) or DivIVA, a bacterial cell division protein that localizes to the cell poles. Upon induction of the DivIVA fusion protein, the GFP-fusion protein is recruited to the cell poles if a positive interaction occurs.

There were many problems that came into the picture during the development for an automated system to identify these positive interactions. There were basic segmentation

and edge detection problems and the problems caused by inclusion bodies (will be discussed in the sections to follow). Different known procedures to obtain thresholds, and edges were evaluated and the apt ones for our analysis were implemented. A proper flow of advanced image processing and feature extraction algorithms was laid out. These steps were used to analyze the datasets of acquired images. Various methods applied are discussed in detail. The experiments conducted along with the results generated are discussed extensively. A statistical feature set used to quantify the image based information and to aid in the determination of a positive interaction is developed.

Various image processing and feature extraction algorithms used to analyze fluorescence microscopic images were also applied to Atomic force microscopic images with a few modifications. There was a basic problem of uneven background noise and this was removed using a common procedure that is used to remove uneven illumination in DIC images. These AFM images were analyzed and quantized using numerical descriptors defined during the analysis of fluorescent microscopic images.

Table of Contents

List of Figures

Chapter 1

Introduction

This thesis is the result of work in the field of quantitative automated image analysis of fluorescence and atomic force microscopic images. Focus was channeled towards development and implementation of robust algorithms for quantitative feature extraction enabling the automated image analysis of various cells and their structure.

1.1. Motivation

Knowledge of the cell its structure and its functionality forms the basic motivation of the field of biotechnology. A “proteome” is defined as the total set of proteins expressed in a given cell at a given time and ‘Proteomics’ refers to the science and the process of analyzing and cataloging all the proteins encoded by a genome. Functional and location proteomics with their high content information is revolutionizing current research in the post genomic era [Dav 04]. A Protein is characterized by its structure, sequence, expression level, activity and location. The location of a protein in particular is pretty

useful in understanding its function. An area of protein characterization that is still in its fledgling stages but likely to be extremely useful in the post-genomic era is that of protein sub-cellular localization that essentially describes the location within a particular cell type where one finds a given protein. The organelle where the protein is located gives a context for it to carry out its role. Each organelle provides a different biochemical environment that influences the associations that a protein may form and the reactions that it may carry out. Thus the knowledge of such data could be invaluable to us.

The identification of protein-protein interactions along with spatial and temporal localization data is vital for assigning functional information to proteins. There are greater than 30,000 genes of the human genome and they are speculated to give rise to about $1x10^6$ proteins through a series of post-translational modifications and gene splicing mechanisms [Pen 03]. Majority of them are expected to operate in concert with other proteins in complexes and networks to orchestrate the myriad of processes that impact cellular structure and function. Implications of these studies are based on the premise that the function of unknown proteins may be discovered if captured through their interaction with a known protein target of known function.

Undertaking a comprehensive study of protein localization and interactions typically involves the analysis of huge datasets of images. Historically, these datasets have been analyzed manually, a process that is found to be highly biased, time consuming and

inconsistent. Thus the explosive need for automated approaches to experimentally identify positive protein-protein interactions and localizations is exposed. The motivation is thus to develop software with generic algorithms to automatically quantify and analyze image based information.

1.2. Objective

The objective of work leading to this thesis was to automate the process of identifying positive interactions between two proteins of interest in fluorescence microscopic images. These proteins are fused to either Green fluorescent protein (GFP) or DivIVA, a bacterial cell division protein that localizes to the cell poles [Din 02] . Upon induction of the DivIVA fusion protein, the GFP-fusion protein is recruited to the cell poles if a positive interaction occurs. This included the use of existing algorithms in image analysis and when required, the development of a sequential combination of techniques to obtain satisfying results.

A significant part of the algorithm development was devoted to various preprocessing steps used to define cell boundaries and segment them from background. This appended with advanced feature extracting algorithms were used to assemble a complete chain of processing steps to obtain an automated system capable of identify positive protein interactions in fluorescent microscopic images. The same algorithms with minor

modifications were used to extract quantitative attributes of a specimen under study from atomic force microscopy images.

1.3. Contribution

The current thesis works contributions are focused in the field of automated image analyses for fluorescent microscopic images and atomic force microscopic images. Contributions pertaining to the specific research are as follows:

- Using the Differential interface contrast (DIC) image to define cell contours and corresponding localized image enhancement of the fluorescence image.
- Successful statistical feature extraction.
- Identify and quantify positive protein localization spots.
- Extending the same algorithm with a few modifications to images from other modalities (AFM).
- A logical way of avoiding the problem of inclusion bodies in automating the process for this specific case.

Combinations of various known techniques are employed to obtain image based statistical parameters attributing the specimen under study.

1.4. Thesis outline

The thesis is organized as follows: chapter 1 consists of the basic objective, motivation and contribution of this thesis. Chapter 2 gives a brief background for the work done with a focus on the basic biology required to understand the work and the different types of microscopy and some amount of detail on digital image analysis. It is followed by chapter 3 that describes the various method employed in the work and the materials used for the same (will elaborate after writing the chapter). Chapter 4 contains the results observed during various stages of the work and is followed by chapter 5 giving concluding remarks that includes the discussion of results and future work in the field.

Chapter 2

Background

2.1. The cell

The structure and composition of living organisms vary vastly, from a single celled bacterium to complex multi-cellular organisms with differentiated cell types and interconnected organ systems. There are myriad systems that act in concert with each other to produce and sustain a living organism. Because cells are the 'basic units of life', the study of cells, cytology, can be considered one of the most important areas of biological research. Though we have known about cells for over three centuries, we are still discovering new structures and molecules in them. The knowledge of these various organelles and their respective functions has been the cynosure of much research in the last few decades.

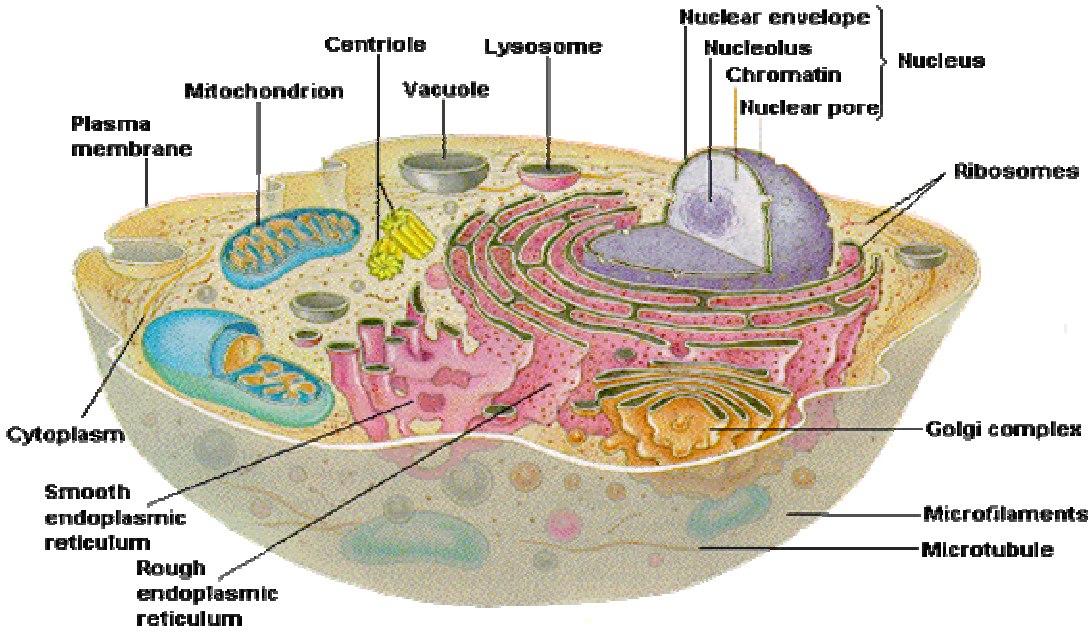


Figure 2.1. An animation of the cell [Hus 95]

Figure 2.1 illustrates the various parts of a cell. Information required by a living cell to exist resides inside the “nucleus” of every cell. These instructions tell the cell what role it is to play in the body. They are in the form of a molecule called the De-oxyribonucleic acid (DNA) that acts like a blueprint with a set of instructions. A DNA strand is made of letters which form words and which in turn form sentences. Such sentences are “genes”.

Genes are instruction manuals for the body as they contain the directions for building all the proteins that essentially make our body function. Study of genes, Genomics, has gained high prominence as it allows a means of constructing biological pathways by integrating information in order to identify key components within those pathways. Each

DNA fragment is one gene and each gene has a specific instruction to carry so as to produce a protein.

Proteins are responsible for every function of a cell and are very small and are usually difficult to see even with the best microscopes around. Specific machinery inside the cell reads a gene and creates a ribonucleic acid (RNA) every time there is a need to produce a protein. RNA moves from nucleus to cytoplasm and where the protein manufacturing machinery - Ribosome, reads the message and produces a protein as per the specifications sent out by the gene. Thus to make one protein, we need a number of other highly specialized proteins and thus the humungous number of proteins.

The identification of protein-protein interactions along with spatial and temporal localization data is vital for assigning functional information to proteins. There are greater than 30,000 genes of the human genome and they are speculated to give rise to about $1x10^6$ proteins through a series of post-translational modifications and gene splicing mechanisms [Pen 03]. Majority of them are expected to operate in concert with other proteins in complexes and networks to orchestrate the myriad of processes that impact cellular structure and function. Implications of these studies are based on the premise that the function of unknown proteins may be discovered if captured through their interaction with a known protein target of known function.

2.2. Microscopy

2.2.1. Fluorescence microscopy

The fluorescence emitted when a cell exposed to fluorescent dyes tagged to biologically active contents of it, like the proteins, is a very handy tool to observe and analyze various reactions and interactions. This is called *fluorescence microscopy*, and the study of protein interactions is a field where this procedure becomes very useful and is gaining a high priority [Dav 04] **.**

The first compound light microscope, invented by Zacharias Jansen in 1595, has gone through enormous evolution. The modern light microscope is a versatile instrument for microscopic analysis. The construction of the first epi-illuminated fluorescence microscope by Ploem, made the light microscope a useful instrument for fluorescence microscopy. Figure 2.2 illustrates a schematic of a wide-field fluorescence microscope.

When fluorescent molecules absorb a photon of a specific energy for an electron in a given orbital, the electron rises to a higher energy level which is a highly unstable state for an electron. Thus, it tries to come back to the ground state by releasing energy in the form of light and heat. This emitted light is fluorescence. In this procedure, cells are tagged with a fluorescent dye like the Green Fluorescent Protein (GFP) and then illuminated with filtered light. Light emitted from the dye is viewed using a filter. Thus,

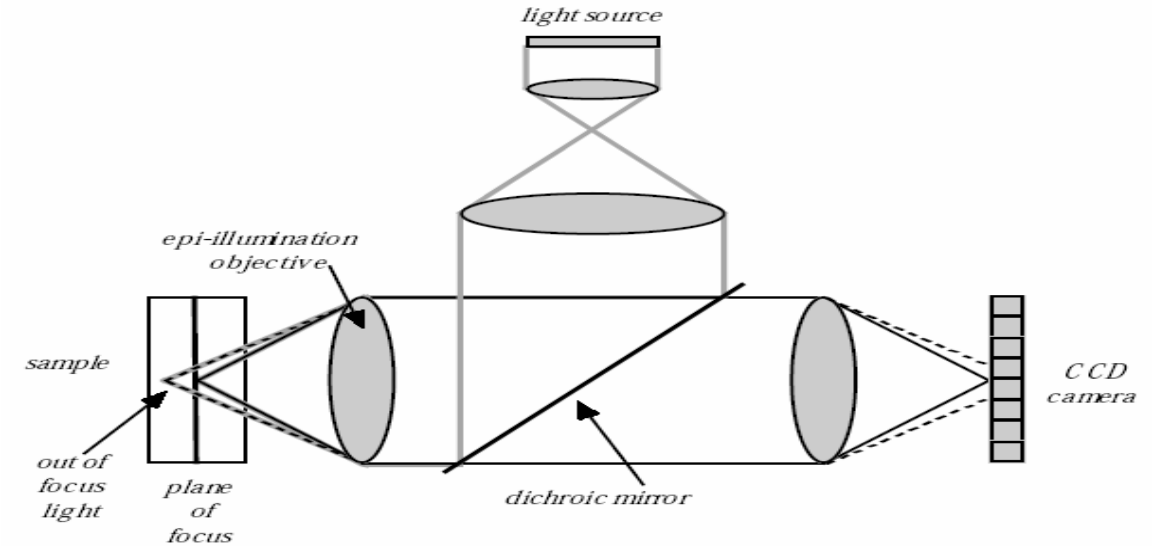


Figure 2.2. Schematic diagram of conventional wide-field fluorescence microscope [Kem 99]

specific parts of a cell can be viewed by tagging them with specific fluorescent dyes allowing for a quantitative evaluation of multiple parts simultaneously. In epi-illumination, the illumination of the sample and the detection of its emitted fluorescence light are done using the same objective lens. This strongly reduces penetration of illumination light in the detection light path, which makes the detection of the weak fluorescence light feasible. A strong characteristic of the epi-fluorescence microscope is its wide-field illumination, which enables the simultaneous imaging of the entire focal plane. Modern scientific grade fluorescence microscopes are excellent tools for acquiring microscopic images of two dimensional samples with a discriminating power of well below one micrometer. The wide-field illumination turns out to be a major drawback of

the microscope. Since the whole sample is illuminated simultaneously, it will not only excite fluophores in the focal plane but also in the out-of-focus regions of the sample as well. When a fluorescence sample is illuminated with light of the proper wavelength (in the absorption spectrum of the fluorescence molecules), it emits light of a longer wavelength. This emitted light can then be detected using, for example, a CCD camera. A camera will acquire a two-dimensional image of the emitted light intensity. The acquisition of both the in-focus and out-of-focus light, results in poor resolution of a conventional wide-field fluorescence microscope along the optical axis which, can be overcome by various de-convolution techniques [Nee 03].

2.2.2. Atomic force microscopy

AFM has found extensive use in many areas of cell and molecular biology as it is one of the most powerful tools for determining the surface topology of bio-molecules at a sub-nanometer resolution [Jur 96]. Unlike X-ray crystallography and electron microscopy (EM), the AFM allows bio-molecules to be imaged not only under physiological conditions, but also while biological processes are at work. AFM operates by measuring attractive or repulsive forces between a tip and the sample [Bin 86]. In its repulsive "contact" mode, the instrument lightly touches a tip at the end of a leaf spring or "cantilever" to the sample.

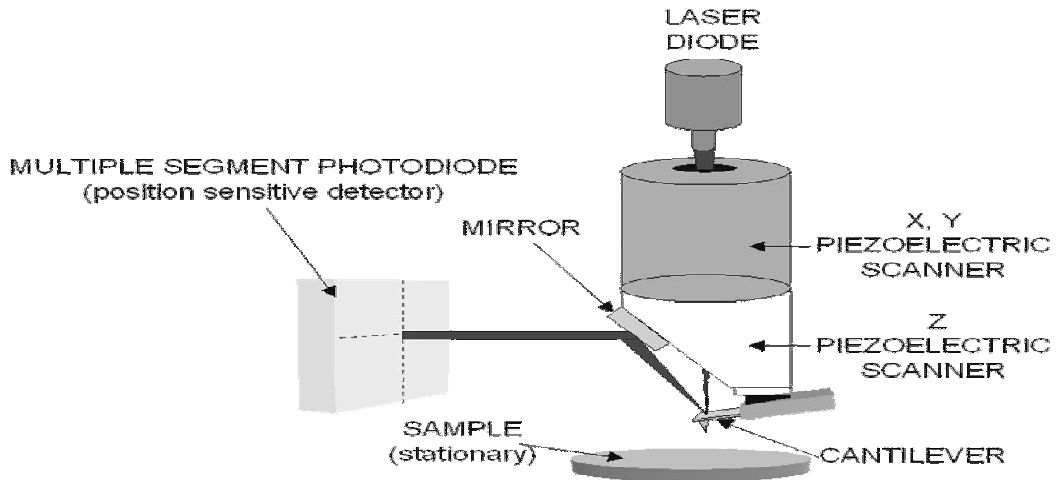


Figure 2.3. Atomic force microscope schematic [Mic 03]

As a raster-scan drags the tip over the sample, some sort of detection apparatus measures the vertical deflection of the cantilever, which indicates the local sample height. Thus, in contact mode the AFM measures hard-sphere repulsion forces between the tip and the sample. Figure 2.3 illustrates a schematic of an Atomic force microscope.

In non-contact mode, the AFM derives topographic images from measurements of attractive forces; the tip does not touch the sample [Alb 91]. An AFM can achieve a resolution of 10 pm, and unlike electron microscopes, can image samples in air and under liquids.

2.3. Literature review - automated sub-cellular localization

The technique used to develop applications such as sub-cellular protein localization of late is drawn from the fields of fluorescence microscopy, pattern recognition and machine learning. The goal is to develop methods that allow for the numerical description and subsequent classification of the patterns found in fluorescent light microscope images of cells. Such images are obtained by labeling one or more sub-cellular structures with fluorescent dyes and then collecting images of the resulting pattern of fluorescence using a microscope, which in turn leads to the problem of describing these patterns in a way that is acquiescent to further processing [Bol 97].

There are quite a few advantages of automated sub-cellular localization. The most important being that the quantitative description of images facilitates standardization that was not previously possible. An immediate comparison of a new pattern with many existing patterns from the database so constructed could be of immense potential. The same goes with studying protein-protein interactions. We would be in a better position to obtain an insight into the complex protein interaction and localization mechanisms with such a system at our disposal.

Initially, Zernike moments [Zer 34] and Haralick texture [Har 79] features were used to quantify images, as they were invariant to translation and rotation of the cells within the filed of view. They would also serve as a completely general set of descriptors and allow

adding additional image classes without redesigning the basic feature set. The classification system then used with the features was a classification tree, as implemented in S-Plus (Mathsoft Seattle, WA USA) tree() function. This implementation was based on Classification and Regression Trees (CART) and the second classifier was back propagation neural network and was implemented using PDP++. The image feature data were separated into distinct training and test sets so as to assess the performance of the two classifiers. The classification tree had an accuracy of 69% [Jur 96] [Bol 97] and the back propagation network was accurate to about 84% [Jur 96] [Bol 97]. Chinese Hamster Ovary (CHO) cells were used for experiments in [Jur 96] [Bol 99].

When the above procedure was applied to a larger number of patterns in HeLa cells, many of the patterns could not be distinguished. Then in [Bol 01], a set of new features were added to the existing list to address the challenge of distinguishing all major classes of localization patterns. Various image processing routines were carried out using various MATLAB functions but the single cells still are isolated by manually defined polygons. The new feature set that was added comprised of morphological and statistical features. Classification was carried out by using back-propagation neural networks (BPNN) using the NETLAB (http://www.ncrg.aston.ac.uk/) scripts for MATLAB. The mean and standard deviation of the training data were used to normalize the train and test sets. One of the most vital steps in pattern recognition is an appropriate choice of features to represent an image. In an attempt to optimize the same, a subset of features was selected

from the available ones by employing the stepwise discriminant analysis [Jen 77] using the STEPDISC function of SAS (SAS Institute, Cary, NC, USA). Neural networks were chosen owing to the failure of other approaches including linear discriminant analysis, decision trees, and k-nearest neighbor classifiers. The classifier so described in [Bol 01] was able to correctly recognize 83% of previously unseen cells and 98% accuracy on homogeneously prepared cells.

Later, in [Hua 02], a pattern analysis method to compare sets of fluorescence microscope images was developed essentially to evaluate the differences in protein sub-cellular distribution in an objective fashion. They presented a method for quantifying changes in sub-cellular protein distributions and applied a standard statistical test to determine the significance of those changes. In that work, the same set of features discussed previously were used to compare two sets of images (e.g. before and after treatment with a drug) with the task being to determine if these matrices were statistically different. A multivariate statistical approach called the Hotelling T^2-test was used for the above defined purpose. The system so developed in [Roq 02], was able to distinguish between two sets of images that were previously indistinguishable by visual analysis and also could identify situations in which two patters showing the same distribution.

A Protein Sub-cellular Image Database (PSLID) was described in[13] that collects and structures 2-D through 5-D fluorescence microscope images, annotations, and derived

features in a relational schema. Image interpretation was achieved using Sub-cellular Location Features that have previously shown capable of recognizing all major sub-cellular structures and of resolving patterns that cannot be distinguished by eye. The paper used previously devised numerical descriptors to compare and classify protein patterns. The previous work was incorporated into PSLID (http://murphylab.web.cmu.edu/) to provide a comprehensive application incorporating relational database machine learning and statistical inference. It was an example of applying data mining on top of a relational database and achieving query interpretation.

An improved set of numeric features for describing images that are fairly robust to image intensity binning and spatial resolution was described in [Mur 03]. The features were used to train neural networks and were validated by the fact that they can accurately recognize all major sub-cellular patterns with accuracy higher than those reported earlier. The features were subsequently used to create a Sub-cellular location trees that group similar proteins and provide a systematic framework for describing the same.

Results obtained for multi-cell images were described in [Hua 04], thus suggesting a classification technique for sub-cellular patterns in tissue images. Since texture features essentially represent repetitive local patterns in an image, they are invariant of the number of cells and should also work for partial cells.

2.4. Image analysis

The human visual system is quite complicated as it has a continuous transfer of information between the system itself and the brain. An image captured by the eye is sent to the brain for analyzing the same. The brain analyzes and interprets data from the image. The field of machine vision deals with the development of a system to replicate the above described one. The visual system can make excellent qualitative and quantitative judgments from a field of view. To develop a system with similar attributes along with actual numerical descriptors accentuating the quantitative judgments made by it is the basic driving force for the work described in this thesis. For example, a human visual system can count the number of cells in a given microscopic image and probably identify positive protein interactions with some amount of pre determined knowledge. How do we make a machine do the same for us?

A digital image represents an image as an array of numbers. It could be a two-dimensional array like an ordinary gray scale picture, or a three-dimensional array like a stack of images combined together to give volume information. There can be a four-dimensional image with a time series of images with the fourth dimension being the time itself. There are images that represent wavelengths like the color image that comprises of three different images in three channels: the red, green and the blue channel, corresponding to the fact that the human eye retina has three cones to sense color.

Though higher dimensions are not common, the use of three-dimensional images has become quite a common feature in the field of fluorescence microscopy.

Each element of a 2D array is called a ‘pixel’ (short form for picture element) and the corresponding terminology for a 3D image is a ‘voxel’ (volume picture element). The process of manipulating an image is called image processing where the input image is changed to suit the needs for a particular analysis. The basic difference between an image processing system and an image analyzing one is in their outputs as the former gives out an improved image and the latter has a set of numerical data attributing the input image.

There are various image analysis problems but often, most of them share common formulations and solutions although the specific algorithm used maybe quite different. The methodology employed remains mostly consistent, including in general six processing stages: sample preparation, image acquisition, image pre-processing, image segmentation, feature extraction, pattern classification, and evaluation.

Chapter 3

Materials and Methods

This chapter gives an overview of the various methods employed throughout the work leading to this thesis. The basic inevitable steps for any image analytic system include, sample preparation, image acquisition, image restoration / image pre-processing, segmentation, feature extraction and pattern recognition. Figure 3.1 shows different blocks of an automated image analytic system.

3.1. Sample preparation and image acquisition

Careful planning and selection of imaging modalities and staining methods can significantly reduce complexity of the image analysis procedure that follows. Procedures that could avoid redundant or disturbing background can be very vital.

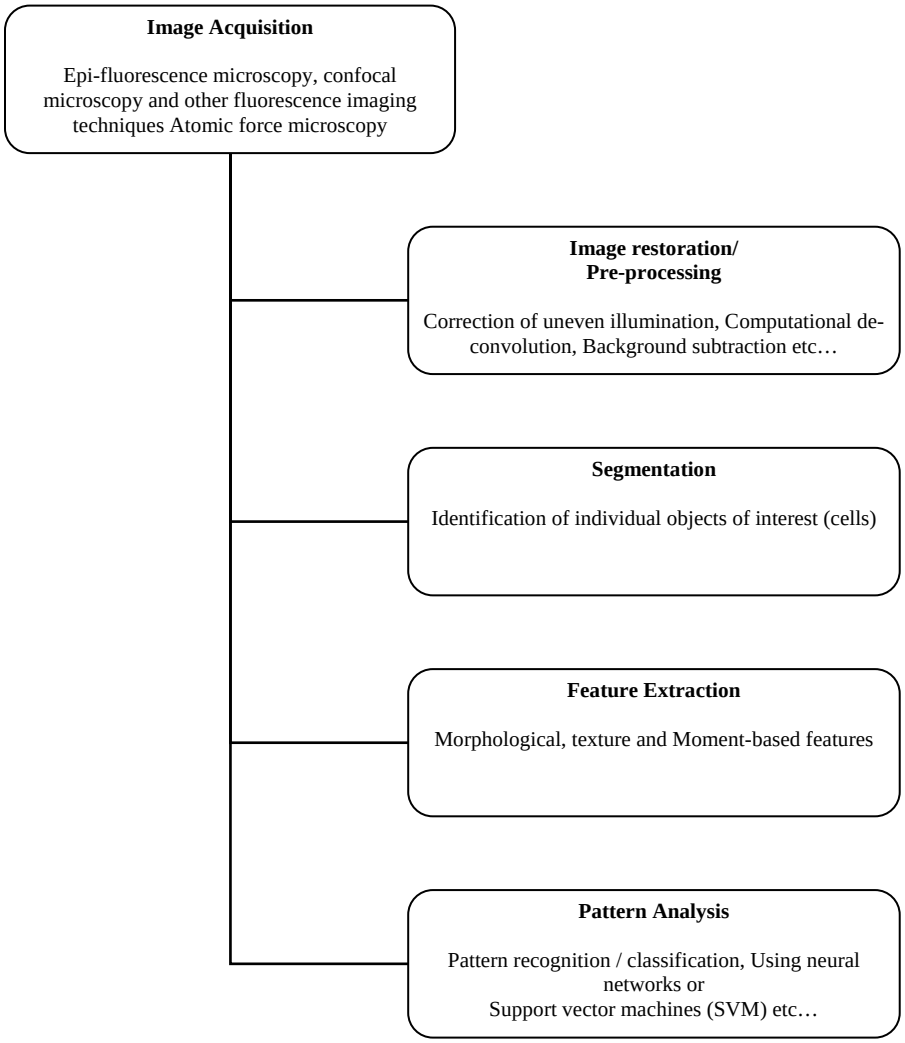


Figure 3.1. A flow chart for automated image analysis

Most of the work in this thesis was done on fluorescent images of Escherichia coli. Escherichia coli strain BL21-DE3 (Invitrogen, Carlsbad, CO) was co-transformed with two compatible vectors encoding pairs of potentially interacting proteins from R. palustris fused to either DivIVA or green fluorescent protein (GFP). The R. palustris gene products tested in this study includes GroES1 (RPA1141), GroES2 (RPA2165), GroEL1 (RPA1140), and GroEL2 (RPA2164) [Lar 04]. For this assay, expression of the DivIVA fusion protein is tightly regulated by an arabinose inducible promoter [Guz 95] and the GFP fusion protein is expressed constitutively from a T7 promoter. Co-transformed cells were grown for at least 6 hours at 30˚C or 37˚C in LB medium containing 50 μg/ml ampicillin and 15 μg/ml chloramphenical to maintain plasmid selection and then imaged using a Leica SP2 confocal laser scanning microscope to determine the localization pattern of the GFP-fusion protein. After assessment of the baseline pattern of GFP localization, arabinose was added to the medium to a final concentration of 0.2% to induce expression of the DivIVA-fusion protein. The cells were incubated for an additional hour at 30˚C or 37˚C. Following induction of the DivIVA-fusion protein, the cells were imaged again to determine if a change in the pattern of GFP-fusion protein localization occurred. If the GFP-fusion protein is recruited to the cell poles following expression of the DivIVA-fusion protein, the data is interpreted as showing a positive interaction between the two proteins of interest. Images were acquired using Leica Confocal Software (LCS).

To stain cell membranes, E. coli cells were grown in liquid LB medium as described above. Approximately 15 minutes prior to harvesting the cells, 200 ng/ml FM5-95 (Molecular Probes, Eugene, OR) was added directly to the culture to stain the membranes. The cells were then harvested by centrifugation, washed two times with phosphate buffered saline, and prepared for microscopy. A set of fluorescent and its corresponding differential interference contrast (DIC) images are then acquired.

3.2. Image preprocessing / restoration

This step is where we tend to employ various image processing algorithms to help us interpret the images acquired in a better fashion. These steps try and reduce imperfections caused during image acquisition procedures leaving a better looking image for further segmentation procedures whose complexity is pretty much proportional to the amount of distortions or noise present in the image. Figure 3.2 illustrates a flow chart with basic image processing steps.

Preprocessing steps typically include simple yet effective techniques such as smoothing and histogram based processing (e.g. Histogram equalization), or complex algorithms such as de-convolution to reduce the effect of smoothing. Background subtraction is usually a vital step of this procedure as it helps narrowing down the region of interest (ROI).

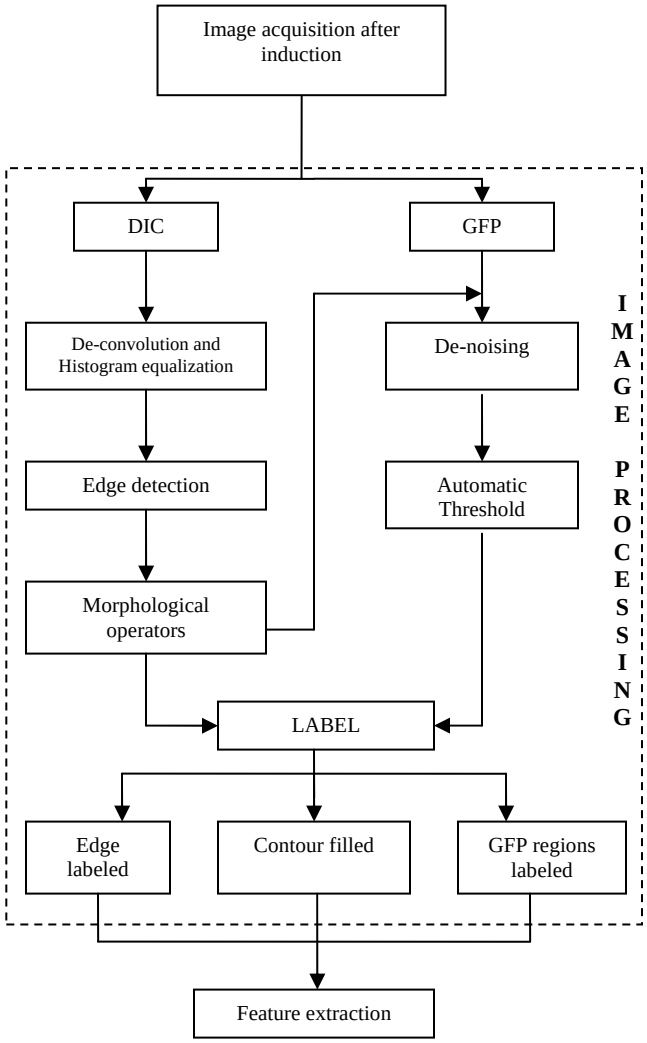


Figure 3.2. Flow chart describing the various image processing algorithms

The algorithm implemented in this thesis uses processed and consequently labeled DIC images to identify ROI in the fluorescent images. Since the DIC images are essentially used to mark ROI (area corresponding to cells) in fluorescent images, we needed to define cell boundaries in it. From observation, it was evident that the DIC images were slightly blurred and the gradients along cell boundaries needed some amount of enhancement in order to extract their boundaries using segmentation procedures. For this purpose, the DIC images underwent a de-convolution followed by a histogram equalization procedure.

De-convolution is a procedure to recover an image from its degraded observation by assuming *a priori* knowledge of the type of degradation and blind de-convolution obtains and estimate of the original image without assuming and prior knowledge of the method of degradation. This by itself is a huge area of research and has many complex algorithms dedicated to it [Nee 04]. Since this de-convolution step in our algorithm is used to just enhance the edge based information in the images, a *Lucy-Richardson* [Luc 79] [Ric 72] filter provided by MATLAB was used over more complex approaches. This de-convolution restores an image that was degraded by convolution with a point-spread function (PSF). The type of degradation occurred during image acquisition from the microscope is assumed to a particular value and the process of de-convolution is carried out. The algorithm is based on maximizing the likelihood of the de-convolved resultant

image being an instance of the original image under Poisson statistics. In our case, the PSF is assumed to be Gaussian with a mask size of 3-by-3 and a sigma of 0.5.

De-convolution was followed by a contrast adjusting technique, popularly known as histogram equalization. Though simple, it could lead to interesting results in many cases. It employs a monotonic, non-linear mapping that reassigns intensity values of pixels in the input image such that the output image contains a uniform distribution of intensities. This, results in a flat histogram with the dynamic range of grayscale intensities stretched over the entire spectrum of 0-255 for an 8-bit image. Thus the small intensity difference along the boundaries is enhanced and made obvious. The process of histogram equalization is illustrated in Figure 3.3.

3.3. Image segmentation

The segmentation problem has been present since the beginning of image analysis, where one tries to find object boundaries within an image. Segmentation in image processing is considered to be one of the most important and also one of the most difficult tasks. In this section we describe in detail, some of the methods considered during our work and discuss briefly a few other methods that are pertinent to Cytometry. Segmentation techniques can be classified into region-based and edge-based ones where edge-based methods try and connect boundaries or edges of objects to create enclosed regions and region-based methods find connected regions of foreground and split these up into

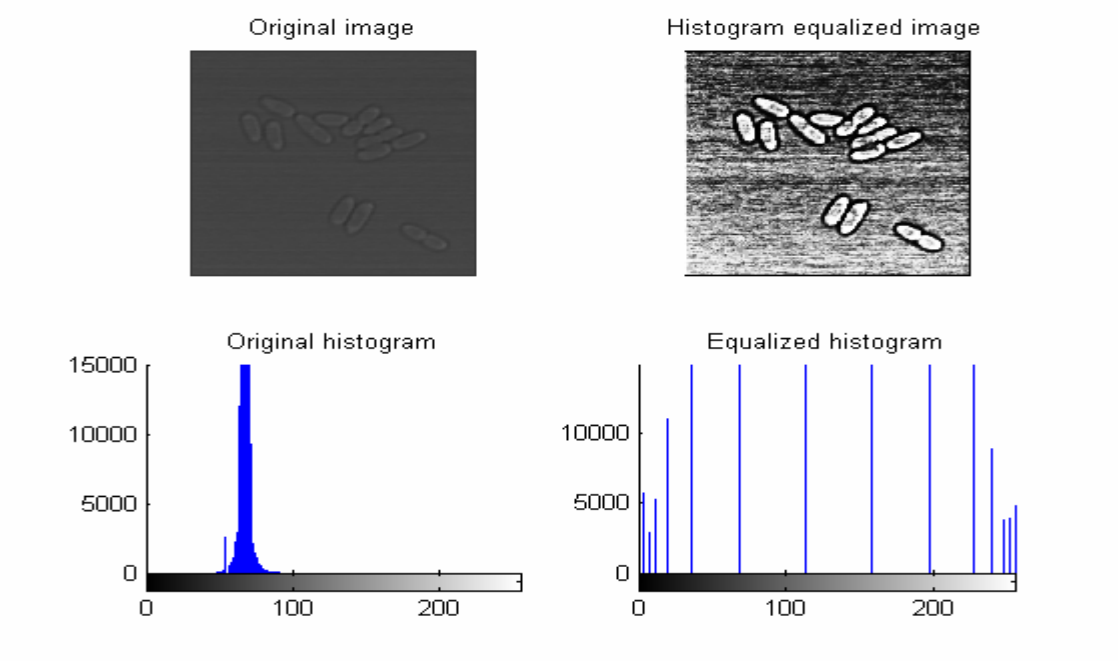


Figure 3.3. Histogram equalization

individual objects. Segmentation is basically the act of separating an image into foreground and background, where the edge-based methods trace a border between the two and the region based ones try to find a property that separates the two in a well-behaved manner. A concert of edge-based and region-based methods can be aptly applied in some cases to obtain good segmentation results. The imaging conditions and the staining methods employed could play a vital role in selecting a segmentation technique thus making the 'optimum method', highly image dependent.

Since the DIC images have a high contrast along the cell boundaries, edge-based methods are used to segment cells from the background and since the protein localization spots within fluorescent images display a significant difference from their background, region-based segmentation is applied. Images with fuzzy borders but significantly varying objects and background, divert our attention towards a region-based segmentation algorithm and thresholding is one such application.

3.3.1. Thresholding

A simple thresholding operation can most of the times achieve this separation of an image into foreground and background. Despite (or perhaps, due to) its simplicity, thresholding can be a very powerful method to separate foreground from background but the choice of the threshold itself remains a challenge. In the simplest case everything in the image that is brighter or darker than the threshold belongs to the object and the rest belongs to the background. This is called a global threshold; however, the use of local thresholds is typically more sensitive to noise. Any feature, for example, color, texture or shape, that essentially separates the field of view into foreground and background can be employed to perform the function of thresholding.

There are quite plenty of methods that can help us decide on a specific threshold, none of them being perfect in all circumstances. There are the popular histogram-based methods [Ros 83] [Sez 85], but other methods incorporating spatial information[Pal 89] and various other parameters [Kap 85] [Tsa 85] of the image are also quite common. The

most popular ways to threshold an image is to find a minimum in the histogram that corresponds to the most stable point. That is, moving the threshold up or down will affect a minimum number of pixels at that point. This is apt when there is just one global minima in the histogram of an image but problems galore in the presence of more than one minima or if there is no significant minimum (uni-modal). The problem of more than one minimum can be solved in a relatively easier fashion by iterative smoothing until only one minimum remains or to directly pick the deepest minimum in some sense. The second problem is much more difficult to solve with the use of different transformations applied to the data in such a way that the histogram is no longer uni-modal, but this method also increases the instability of the threshold point. There are other common methods as that of Otsu [Ots 79] or the equivalent iterative version proposed by Ridler and Calvard [Rid 78]. These methods split the histogram into two parts, so that the threshold is located in the middle between the means of the two classes. This works well if the distribution is made up of two classes of equal variance which usually does not happen that often. But in our case as the de-noised fluorescence image contains two distinct peaks (background (0) and GFP), the algorithm works pretty well and thus has been employed to obtain corresponding binary threshold images (Figure 3.4).

3.3.2. Edge based methods

There are various edge-based segmentation techniques that have been proposed and like the thresholding algorithms, none is optimum for all kinds of images. One such edge

a. b.

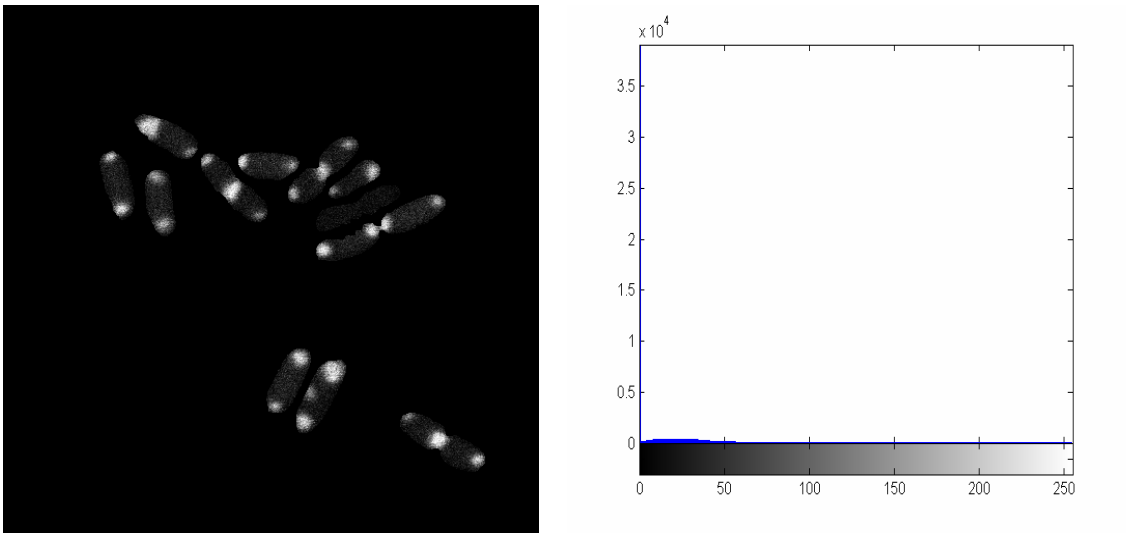


Figure 3.4. De-noised GFP image and its histogram

a. De-noised GFP image b. Histogram (observe the background is made zero) and fore ground is the remaining pixels

detectors is the *Canny* edge detector [Can 86]. It initially identifies candidate edge pixels through a set of edge-detection criteria; the image is convolved with two square masks (highpass filter), producing estimates of the horizontal and vertical components of the brightness gradient at every pixel. The intensity gradient at each pixel location can then be estimated by taking the linear combination of these directional values, providing an estimated magnitude and direction**.** For all pixels, “non-maximum suppression" based on the gradient magnitude is performed by exploring in the direction of steepest gradient. A pixel is kept as a possible edge point only if it has a larger gradient than its neighbors located in the direction closest to that of the gradient, and than its neighbors located in the opposite direction. The remaining local maxima belong to one-pixel-wide edge segments.

Thresholding based on gradient magnitude is then performed on these points. Any point above a high threshold is kept, as well as any segment connected to it which consists of points above a lower threshold, reducing the probability of subdividing a segment whose magnitude fluctuates near the high threshold. A *Canny* edge detector is used to identify the edges of individual cells in our algorithm.

Edge-based methods are built on differences or derivatives, in the image thus making them more sensitive to noise. A purely edge-based method faces the problem of connecting edges to form connected boundary of objects and thus is often combined with region based methods, to distinguish between the object and the background.

Another method of segmentation starts with connected edges and then try to find their correct position, as is done, e.g., when using snakes [Kas 88] or active shape models [Coo 95]. The time complexity of these segmentation algorithms is a primary factor in deciding to opt for the apt one and it is the main reason why we did not pursue with iteratively refined active shape models or active contours, which tend to be orders of magnitude slower than the more direct methods that we have applied.

3.3.3. Connected component labeling

Once the foreground has been separated from the background, the next step in segmentation is that of object identification or labeling the region of interest (ROI). But

before this, the image foreground needs to be processed so as to avoid stray pixels from getting labeled as individual objects and the contours have to be smoothened.

Morphological operators

Morphology is an image processing technique based on shapes of objects observed in an image. The value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors. The size and shape of the neighborhood is defined as a structuring element. This can take many different shapes, e.g. disk, diamond, rectangle, line, etc. This structuring element can be used to construct a morphological operation that is sensitive to its specific shape and size in the input image. There are a number of operations that are used. A few of them are *erosion*, *dilation*, *opening*, *closing*, etc.

Erosion: The output pixel value is determined as the minimum of all the pixels lying in the neighborhood (defined by the structuring element) of the input pixel in the input image. That is, in a binary image if any of the pixels in the defined neighborhood of the input pixel is 0, the output pixel value is set to 0. It is essentially computed by taking the minimum of a set of differences.

Dilation: The output pixel value is determined as the maximum of all the pixels lying in the neighborhood (defined by the structuring element) of the input pixel in the input

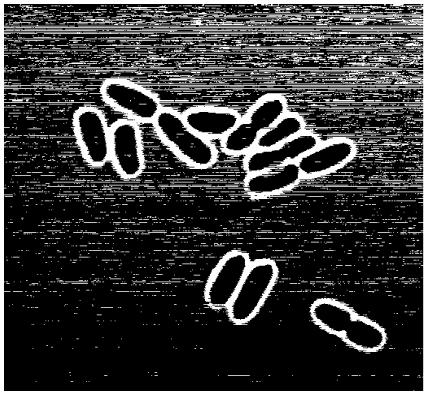
image. That is, in a binary image if any of the pixels in the defined neighborhood of the input pixel is 1, the output pixel value is set to 1. It is basically computed by taking the maximum of a set of sums.

Erosion and dilation are dual operators but in general are not the inverse operation of each other.

Opening and closing: As can be understood from Figure 3.5, erosion shrinks and image and dilation expands the same. The process of opening or closing and image are used to smooth object contours, but in different approaches. Opening is basically defined as the dilation of an eroded function (image in our case) and it tends to smooth the object contour by breaking down narrow isthmuses and eliminates thin protrusion. The closing operation is basically defined as the erosion of a dilated function and tends to fuse narrow breaks, long thin gulfs and typically fills gaps in the contour. All these operations are again based on the neighborhood defined by the structuring element.

Experiments with these morphological operators and the obtained results are illustrated in Figure 3.5. The operation of closing is used to complete the cell contours and these are filled using a binary fill option in MATLAB. This binary image is then labeled using a labeling function provided by MATLAB as *bwlabel* [Har 92] that tags independent groups of objects in the image with a unique label. This works on the principle of neighborhood similar to the region growing technique, as it tags the neighboring (4 or 8

a.

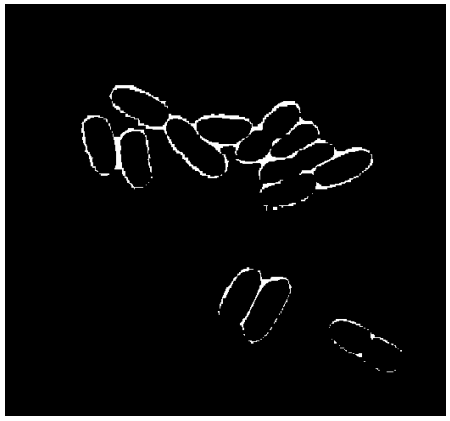


b.

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

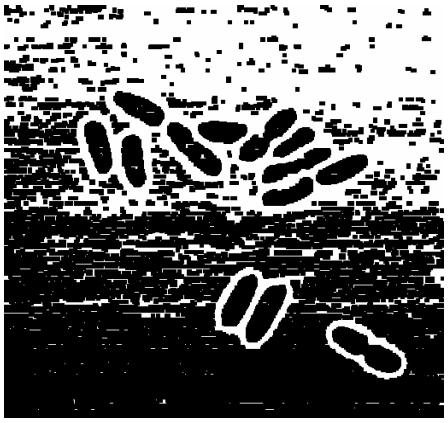
Origin

c.



d.

e.



f.

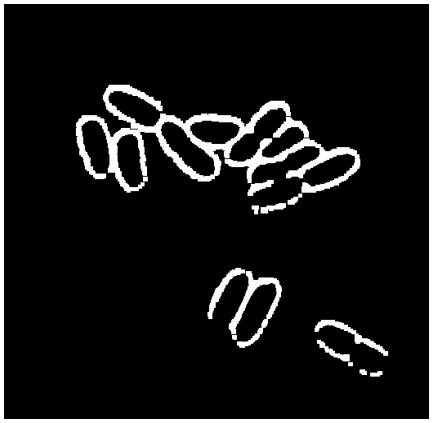


Figure 3.5. Morphological operations

a. Given input image b. Structuring element (*disk* radius = 3) c. Erode operation
d. Dilate operation e. Closing operation f. Opening operation

neighborhood) pixels with the same label. A fresh label is assigned to the next set of neighbors while scanning the entire image. The morphological operators are also used to improve the de-noised GFP image. The GFP image often contains background noise (Figure 3.6.a), which can be removed by considering just the area occupied by the cell (Figure 3.6.b). This is obtained from the corresponding labeled DIC image as explained above. Now, the fluorescent image is devoid of any background noise and an automated global threshold described above is applied on the de-noised fluorescence image to obtain its binary image. The same set of morphological operators of opening and closing were used to remove any speckle noise that might be present in the de-noised, binary GFP image. This binary fluorescent image was then subsequently labeled to identify localization spots as positive objects.

a. b.

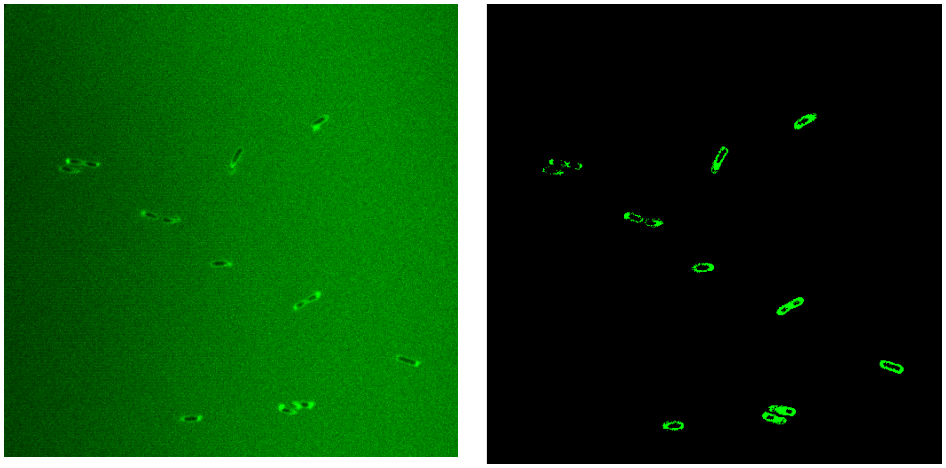


Figure 3.6. GFP image de-noising

a. GFP fluorescence image before De-noising b. GFP fluorescence image after De-noising

3.4. Feature extraction

A list of relevant features that can be obtained from fluorescence microscopic images of cells is described in [chap 2 ref *9*]. Features relevant to this study were selected and include:

- *Number of cells in an image* - This is calculated by counting the number of labels obtained from the DIC image using the function bwlabel provided in MATLAB.
- *Area occupied by individual cells* - This is calculated by counting the number of pixels under each filled contour label.
- *Perimeter of individual cells* - This is calculated by counting the number of pixels under each edge label.
- *Diameter of individual cells* - Diameter is calculated as the value of the greatest eccentricity, i.e. longest distance between any two points in an edge image as shown in Figure 3.7.
- *Roundness factor* - This feature quantifies the shape of an object (the cell) with respect to a circle. It is calculated as follows,

$$\left(perimeter^{2}/_{area}\right) - 4\pi \qquad \text{..... (1)}$$

- *Center of gravity (COG) of cells and center of fluorescence (COF) of GFP localization* - These features use a common equation stated as follows,

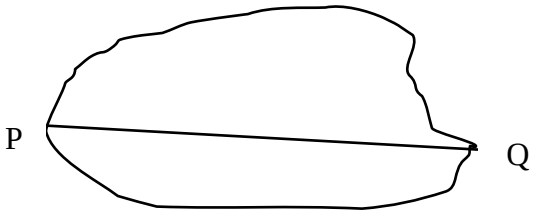


Figure 3.7. Diameter

$$x_i = \sum\sum xI_i(r,c) \Big/ A_i$$

$$y_i = \sum\sum yI_i(r,c) \Big/ A_i$$

…… (2)

where *i* is the cell count, *x* and *y* are the coordinates for COG of the object (in our case, a labeled cell), $I_i(r,c)$ is the image intensity at the location *(r,c)* and A_i is the area of the i^{th} cell. These are used to localize fluorescence within the cell.

- *Distance of GFP localization regions from the COG of the cell* - This feature is obtained by calculating the Euclidean distance between the COG and COF's within the cell which provides us with another qualitative measurement with respect to the orientation of the GFP-fusion protein localization sites within the cell.
- *Percentage area occupied by protein clusters within each cell* - This is calculated from the ratio of areas of GFP-fusion protein localization sites within a cell and that of the cell itself. It is given as follows,

$$\text{Percentage area occupancy} = \frac{\text{Area of GFP localization sites within a cell}}{\text{Area of the cell}}$$

- *Number of regions of GFP-fusion protein localization within each cell* - This is the feature that quantifies the success of the co-localization. It is extracted by treating each cell as an individual entity and labeling the sites of GFP-fusion protein localization within it. This gives the number of localization sites within each cell. Ideally, this number would be 2 for our test system. The possibility of other values is discussed in the next section.
- *Percentage of cells with desired localization regions within an image - This feature calculates the percentage of cells that display the desired GFP-fusion protein localization pattern using the number of cells, the number of localization sites within each cell, and their respective distances from the center of gravity of the cell to determine the result.*

3.5. Pattern recognition

3.5.1. Positive localization spots

This is an integral part of an automated image analysis system, where features extracted by using the above procedures are put to use to identify patterns. The pattern pertinent to the current experiment is observing two localization spots, one at each pole of the cell. This is the ideal case and should be identified as positive localization. A third localization spot at the centre of the cell is observed during the process of cleavage of the cell and this has to be considered as a positive interaction too. This simple pattern can be marred by

confusion when more than two or three localization spots are identified within the boundaries of a single object (cell or cells??) at different orientations.

The task is thus to identify the number of localization spots and their location within the cell. The case of two localization spots is easily identified as a positive interaction, but in the case of three localization spots, there is some amount of ambiguity. Three localization spots within a cell can illustrate that the cell is undergoing the process of cleavage and can be considered as a positive interaction or there could be more than one cell in the labeled object overlapping each other or there could be a possibility of the cell being out of focus during image acquisition procedure. Thus, additional information supporting the decision of a positive or a negative interaction has to be obtained. Upon observation, it was found that cells that undergo the process of cleavage tend to have a higher roundness factor as compared to the normal ones. This criterion was used to ascertain the presence of a positive interaction in the case of 3 identified localization spots.

Distance from the center of cell t the centre of GFP localization spots was used as a metric to identify location of these spots within the cell.

Number of regions of GFP-fusion protein localization within each cell is obtained by treating each cell as an individual entity and labeling the sites of GFP-fusion protein localization within it. This gives the number of localization sites within each cell. As discussed above, ideally this number would be 2 for our test system.

3.5.2. Inclusion bodies[1]

The presence of inclusion bodies in the sample is an experimental problem that can be inherent to the biological system under study. However, it acts as a hurdle to automating the process of image analyses. A unique, logical method of avoiding the problem of inclusion bodies due to GFP-fusion protein over expression in bacterial cells is implemented here. It is achieved by identifying the presence of inclusion bodies in the sample before induction of the DivIVA-fusion protein by acquiring a set of images (DIC and fluorescence) before induction. These images go through the same set of image pre-processing, segmentation and feature extraction procedures discussed above. The percentage area occupied by the localized fluorescence within each cell is calculated, which after experimentation and observation was found to be less than 60% for images with inclusion bodies. If inclusion bodies are present before induction of the DivIVA-fusion protein, the sample is not analyzed further. The flow of procedure followed to test for the presence of inclusion bodes is given in Figure 3.8.Sample test images with and without inclusion bodies are shown in Figure 3.9.

[1] Intracellular protein aggregates that are usually observed in bacteria upon protein over expression.

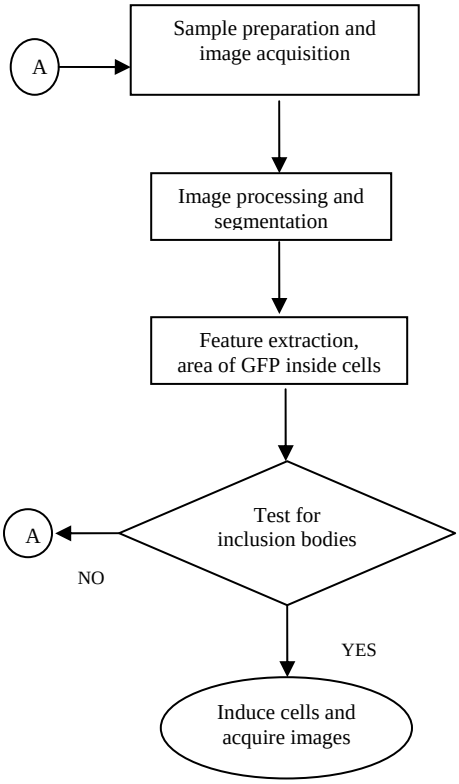


Figure 3.8. Chart describing the flow of procedure for testing the presence of inclusion bodies

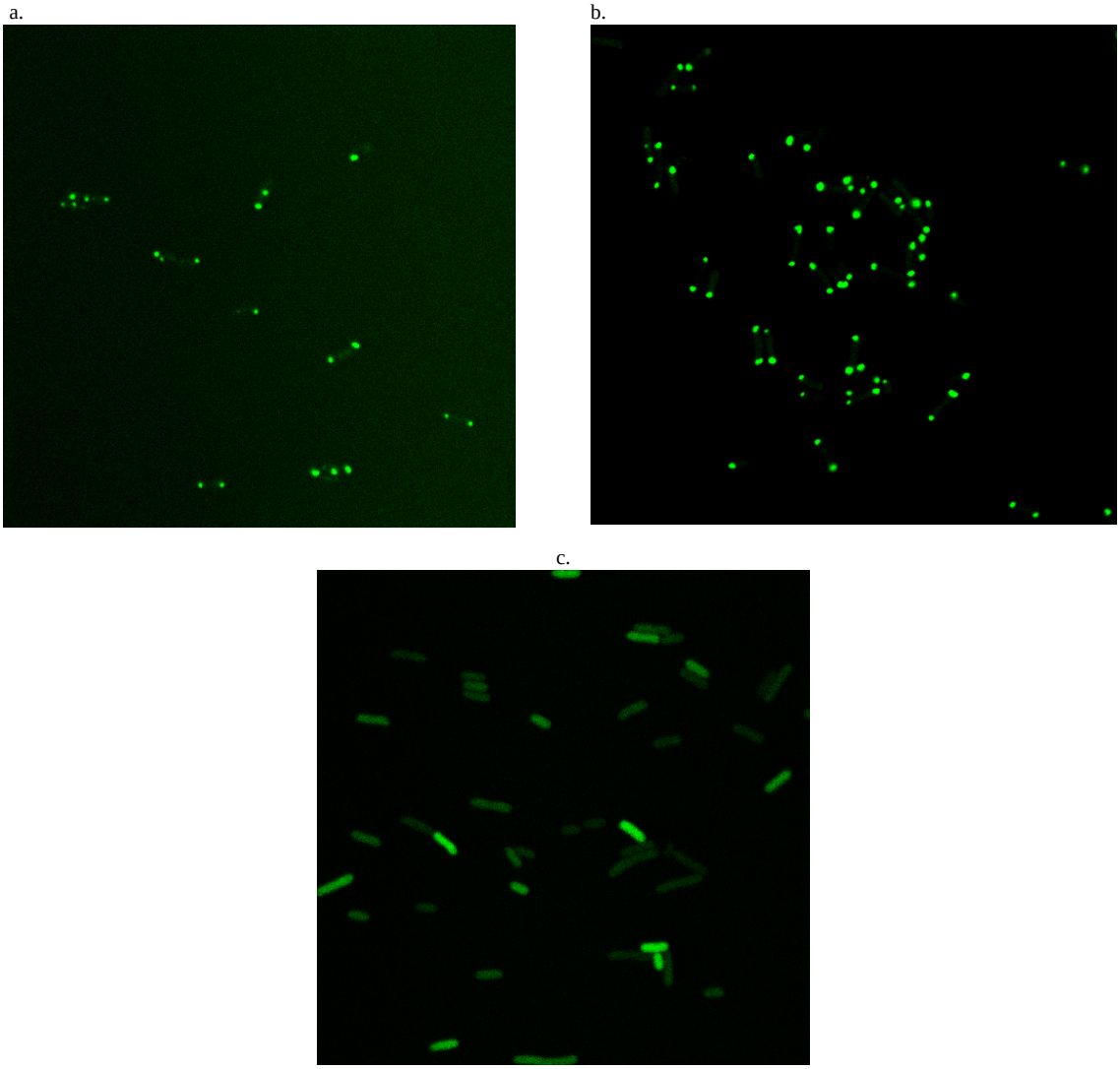


Figure 3.9. Visual similarity between image of a positive interaction and inclusion bodies

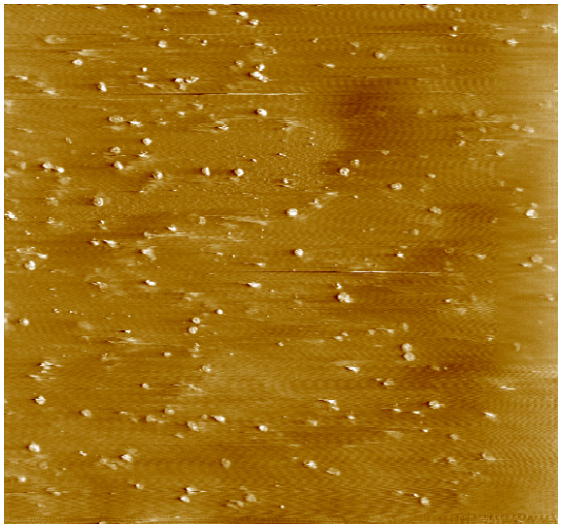
a. Cells with GFP-fusion protein localization at the poles corresponding to a positive protein-protein interaction b. Cells displaying inclusion bodies before induction c. Cells showing cytoplasmic GFP-fusion protein localization before induction

3.6. Background separation for AFM images

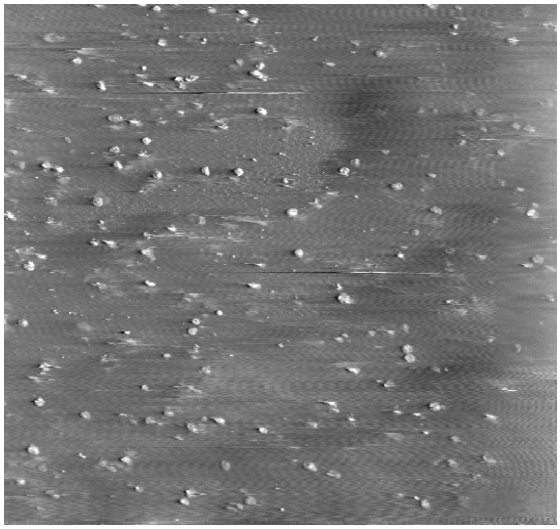
A typical AFM image (Figure 3.10.a) contains background features that can easily be mistaken for valid object features. A simple but effective technique that employs morphological operators is used to estimate the background (Figure 3.10.b) of the image and subtract the same.

A structuring element (*disc* with radius = 10) that is bigger (too large a structuring element gives a very poor estimate of the background and a very small element gives rise to a dark image after background subtraction) than the object of interest is defined. An *opening* operation as described in the previous sections is carried on the image using the defined structuring element. This gives a rough estimation of the background and this image is subtracted from the original to obtain an image devoid of background noise. This image is again *opened* using a small structuring element (disc shaped with radius = 1) to clear noise (Figure 3.10.d).

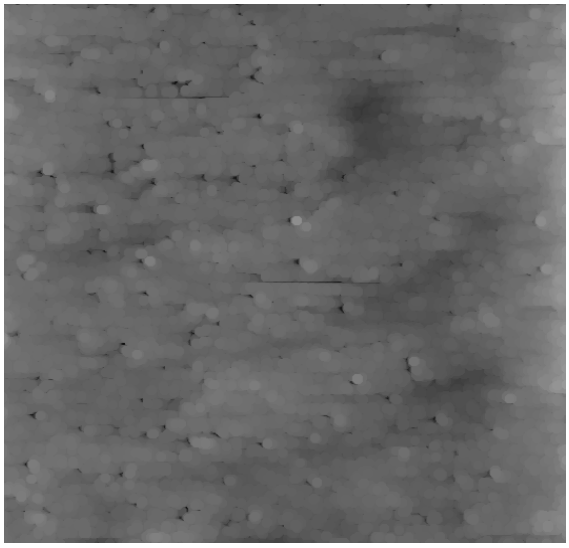
Once the background and foreground (object of interest) are separated, the objects of interest have to be properly marked and labeled. A simple thresholding function, as described in section 3.3.1 can be used to obtain a binary image with desired objects.



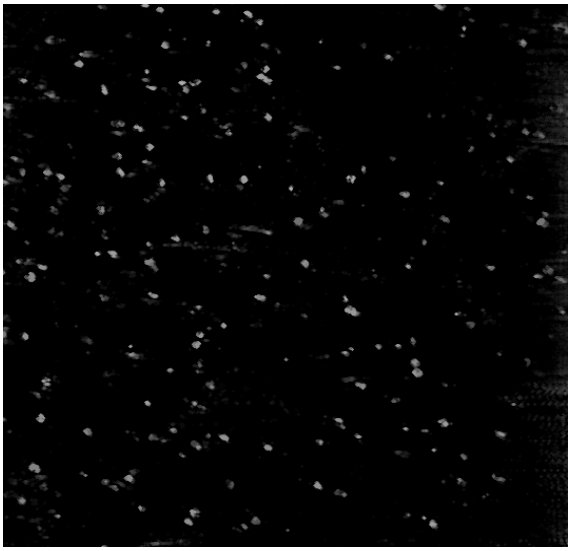
(a) Given AFM image of E.Coli spheroplasts (cells with their outer wall enzymatically digested away)



(b) Gray scale image of the given image



(c) Estimated background of the image



(b) Background subtracted and opened image

Figure 3.10. AFM image background estimation and removal

Different objects from this binary image are labeled with unique numbers using the connected component labeling operation described in section 3.3.3. This labeled image is used as an input for further feature extraction algorithms describe in the previous sections where simple morphological features are calculated from the labeled binary image.

Chapter 4

Discussion – Analysis of Results

The previous chapter dealt with the details of different strategies used in the work of this thesis. This chapter discusses the implementation of the strategies mentioned, and analyzes the obtained results. The various experiments considered and the involving discussion along with the criteria of selecting one method over another are laid out in this chapter.

4.1. Experimental image dataset

Once the sample is prepared, it is placed on a glass slide and it is mounted on to the optical Leica microscope fitted with a 64x objective lens. An appropriate field of view with the desired specimen is selected by browsing through the slide using the stage movement knobs provided. We can then zoom in or out using the software provided by the microscope and adjust the focus while viewing the field of view for clarity which is defined by the user's discretion. Once the focus is set, it is usually locked and the

brightness of the fluorescence image or the contrast of the DIC image can be adjusted to appropriate values as per the user's discretion. There are no fixed values as they vary over the field of view with respect to the GFP expression level or the focal plane etc. The Lieca Confocal Software (LCS) provides us with a fluorescent and its corresponding DIC image.

The experiments are carried on a set of images comprising of fluorescent and corresponding DIC image. In fluorescence microscopy, the DIC images are usually not vital but in our approach, the DIC images play a very important role in defining the boundaries of the cell. Thus, care has to be taken during the acquisition of these images to provide them with good contrast as the parameters used in edge detection depend on the contrast level of the DIC images. The essential information obtained from the DIC image is the cell shape and its boundary that can identify the contour of a cell and thereby making the analysis of its corresponding fluorescence image localized to those contours.

Since the person acquiring images may not be the same one analyzing the data, there is a need for defining an ideal DIC image to help reduce the complexity in the analyses domain. There are no fixed parameters to define an ideal DIC image but there are a few guidelines that can be followed so that the consequent processing part is made easy. Clumps of cells are avoided while choosing the field of view as there would be hundreds

a. b.

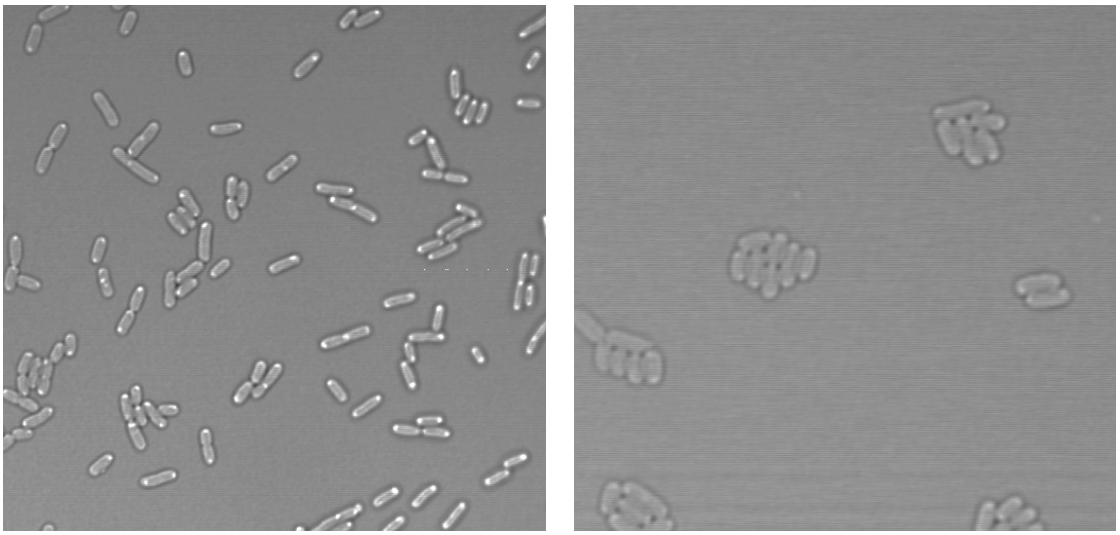


Figure 4.1. DIC images

a. DIC image with adjoining cells but evident edge information that can be easily extracted

b. DIC image with adjoining cells but marred edge information that is very difficult to extract

of cells in the sample under view and most importantly, the boundaries of cells in the DIC image should be clearly visible. Figure 4.1 illustrates an example for a usable and an unusable DIC image. Once a good contrast image is obtained, a high threshold can be applied to achieve useful edge information.

4.2. Performance metrics

The basic idea of our analysis is to identify positive localization patterns (one GFP localization spot at each of the poles) in each cell after inducing them. Ideally there have to be one at each pole of the cell but often, this is not the case. Quite a few cells in the

field of view could be out of focus and thus display just one spot or no spots at all. The number of cells with positive localization patterns was calculated using the developed automated system and the same was obtained from the observations of an expert.

From the test images that were analyzed, the percentage of positive interactions calculated by our automated system was in close agreement with the observations of an expert. The percentage of positive interactions is defined as the ratio of number of cells with positive localization patterns (localization spots, one at each pole) to the total number of cells in the image.

4.3. Experimental steps

The edge feature from the DIC images becomes difficult to extract when there are overlapping cells and cells that lie in close proximity to each other. The occlusion problem caused by overlapping cells is compromised by avoiding such images during the acquisition period. Cells in close proximity were distinguished from each other as the DIC images typically gave a thick boundary and thus the inner contour of the boundaries were used to isolate cells from each other. Once the boundaries are extracted, the fluorescence image is used to identify and label the sites of GFP localization. A common problem with fluorescent images is the presence of background signal.

Each image processing step described in the previous chapter has a specific application

and a reason behind its application in our analysis. The experiments and logic leading the choice of these methods are discussed in this section.

4.3.1. Preprocessing

The purpose of DIC images in our analysis as stated above is to identify cell contours in the spatial domain. Most of the images require some kind of preprocessing to enhance edge information. There are various methods for various kinds of images. De-convolution and histogram equalization of the DIC image were found to be useful for obtaining sharp edge features. De-convolution by itself is a major field of research and improved techniques that output very sharp fluorescent images from dull, blurred ones are available, but often are quite expensive. Since this step in our analysis is applied to merely enhance the ability of DIC images to output better edges, simple and effective de-convolution algorithms offered by MATLAB as built in functions were considered. From the list of available de-convolution functions in MATLAB, both blind de-convolution and Lucy-Richardson algorithms worked well on the test images and had similar outputs, whereas regularized and Wiener de-convolution gave rise to poor resultant edge information (Figure 4.2). Lucy-Richardson was selected for our analyses as it processed images much faster than blind de-convolution algorithm. This step is followed by histogram equalization which basically stretches the grayscale values of the DIC image over the entire dynamic range (0-255 for an 8 bit image).

a. b.

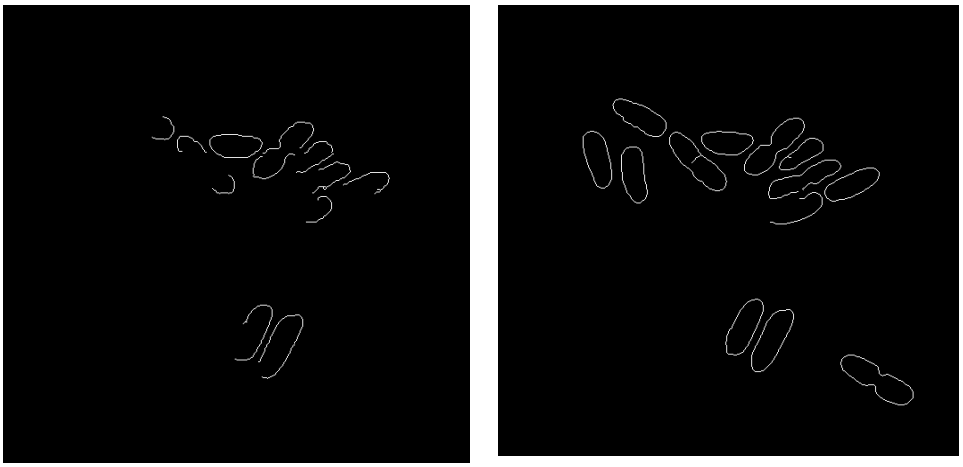


Figure 4.2. Experiment results with de-convolution

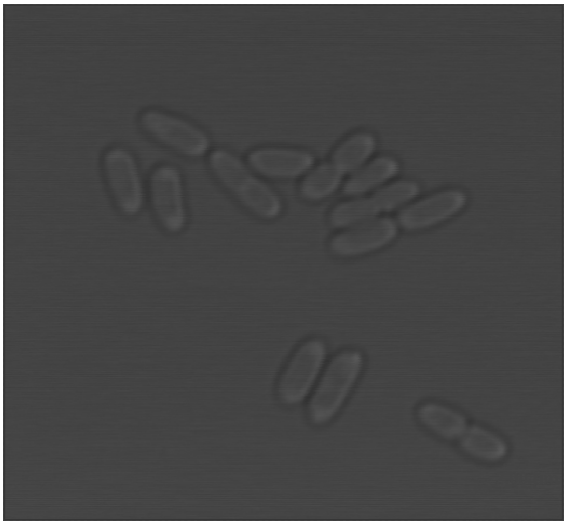
Edge information (using *Canny* filter) obtained after restoring the DIC image using

a. Regularized filter b. Wiener filter

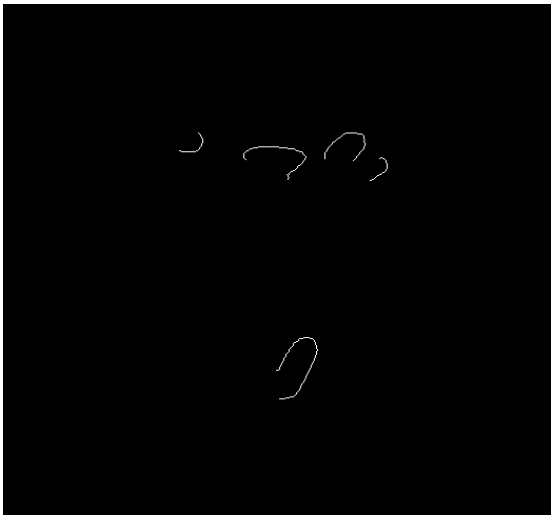
Figure 4.3 illustrates the importance of de-convolution and histogram equalization on test DIC images during our analysis.

4.3.2. Extracting edges

The cell boundaries can be visualized by several techniques. We evaluated the use of a membrane dye and the DIC image. While both methods worked, the use of DIC images

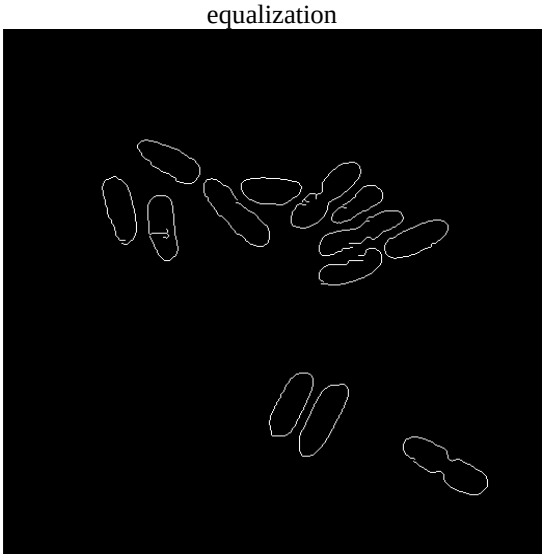


(a) Given DIC image



(b) Edge information without histogram equalization

(c) Edge information with histogram equalization and without De-convolution

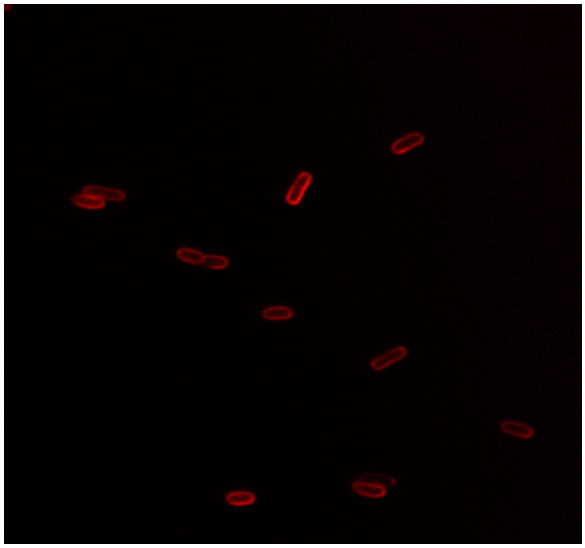


(b) Edge information with Histogram equalization and de-convolution

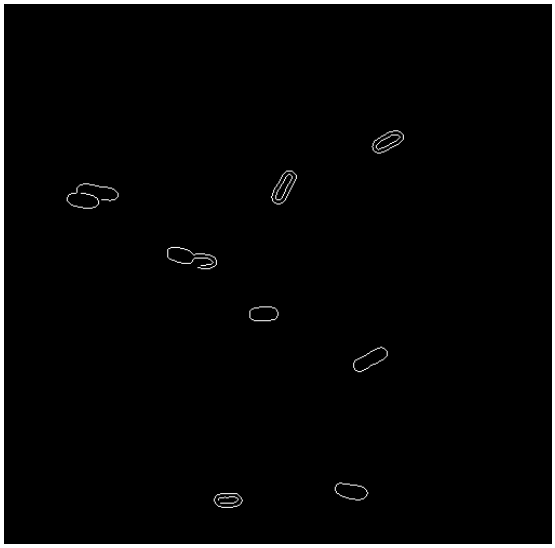
Figure 4.3. Edge information (obtained using *Canny* filter) results from a sample DIC image illustrating the importance of histogram equalization and de-convolution

proved to be more robust for defining the cell outline when examining images of clumped cells as shown in Figure 4.4. The use of images with stained membranes gave a fair indication of the cell boundaries for isolated cells. In addition, the use of DIC images had the experimental advantage of not requiring an additional incubation step and the difference in intensity gradients along the cell outline enabled definition and identification of regions for further analyses. In a DIC image, rapid increase of the difference in intensity gradient along the inner boundary of cells was used. Thus, as long as the inner boundaries of any two cells do not overlap, they can be evaluated successfully as two separate cells. Analysis of a DIC image was chosen over the use of images of membrane staining dyes to determine cell boundaries after evaluating different approaches including the use of active contours and various edge detectors. Effective determination of cell boundaries is mandatory for identifying the localization pattern of the GFP–fusion protein inside the cell. There are various types of edge detection techniques and their suitability often depends on the chosen application. A *Canny* edge detector, a *Sobel* edge detector and the procedure of active contours that is commonly used in the segmentation of images obtained from microscopy, were considered for extracting edge information.

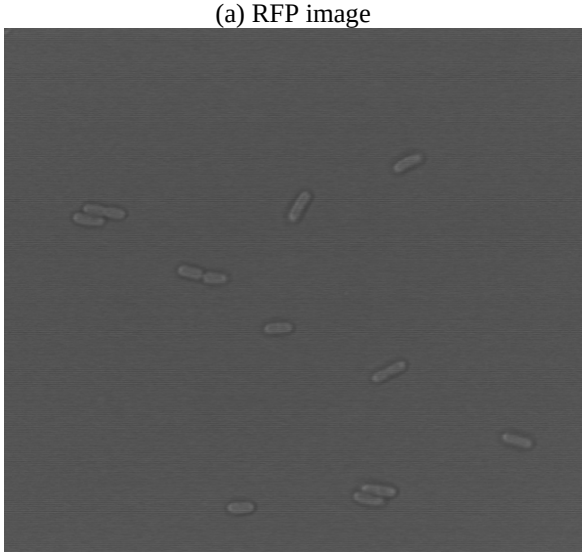
There were two major constraints with the active contour approach. The first one was that of initializing the *snake*. Since we intended the entire process to be automated, and most



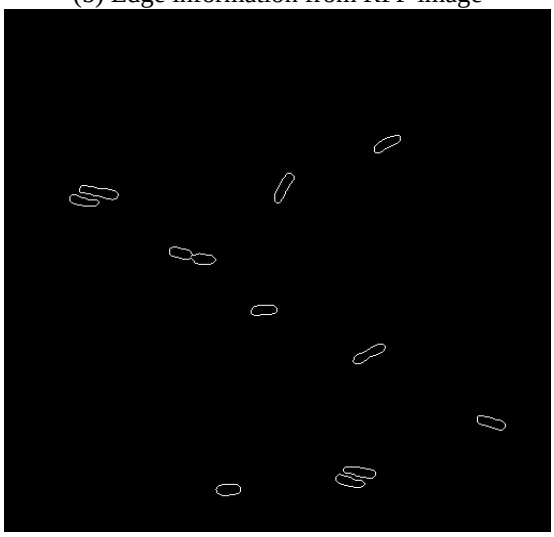
(a) RFP image



(b) Edge information from RFP image



(a) DIC image



(b) Edge information from DIC image

Figure 4.4. RFP membrane dye image Vs DIC image for an edge detector

of the active contour methods described manual initialization of the *snakes*, it was a major set back. Another constraint was that of speed. This method of obtaining cell contours produced good results for images with single cells but took nearly 25 seconds on an Intel 1.6 GHz processor with a 512 MB RAM. A more complex active contour algorithm was required to overcome the problem of isolating individual cells from an image containing a group of closely associated cells.

We chose to employ simple edge detectors like the *Sobel* and *Canny* filters after the preprocessing steps mentioned above as they were simple and faster to implement, on the order of 2-3 seconds on the same machine. Some amount of trial and error was done to determine the threshold level used in the detectors and care was taken to remove unwanted information (weak edges). Of the detectors evaluated, *Canny* filters were found to be the most versatile with respect to our dataset owing to their sensitivity as the *Sobel* filter failed to capture many vital edges and thus was not used in our analysis.

The DIC image shows a thick boundary to the cells, thus producing a ring-like binary image. Upon observation that the inner side of the ring led to more consistent boundary determination, the weak outer edges were discarded by keeping about 10% of the lowest intensity value by using a high threshold in the *Canny* detector.

4.3.3. Morphological operations

While performing morphological operations, particular attention was taken in choosing an appropriate structuring element. The choices of parameters for morphological operations that follow are made in accordance with the resolution of images. The shape and size of the structuring element is defined by the object features (cell contours) under study. The resolution of the image determines the spatial dimensions of objects in an image and this, in turn, determines the parameters relating to the structuring element used for morphological operations. For this reason, care must be taken while choosing the structuring element and its dimensions to avoid overlapping of closely spaced cells in the final image. Since these cells possessed smooth corners, a *disk* shaped structuring element was employed and a radius of 3 pixels was chosen, taking into consideration the spatial dimensions (in pixels) of the cell. A higher or lower dimension for the structuring element would tend to disrupt the information in an image by adding or removing vital information, depending on the morphological function employed.

4.3.4. Data structure

The resultant binary images contain relevant data at specific spatial locations while the rest of the image is featureless and can be ignored in further processing steps. This is achieved by tabulating spatial coordinates of the various features from the labeled image, thereby enhancing the speed of subsequent operations forming a three-dimensional data matrix as mentioned in the previous chapter. This saves computational time by ignoring

the null values in each image and thus is a vital component of the algorithm. An example of the data matrix is shown in Figure 4.5.

4.3.5. Fluorescence images

As described earlier, fluorescence images undergo a different set of preprocessing steps but follow a similar procedure to label images for creation of a data matrix. A unique aspect of our algorithm is its ability to discard any bleeding of the fluorescent signal. Potential noise derived from background fluorescence in areas near the cell features is limited with the use of a data matrix.

This saves computation time and acts as a simple yet efficient noise limiting technique. Inclusion bodies are metabolically inactive materials within the cytoplasm or nucleus of a

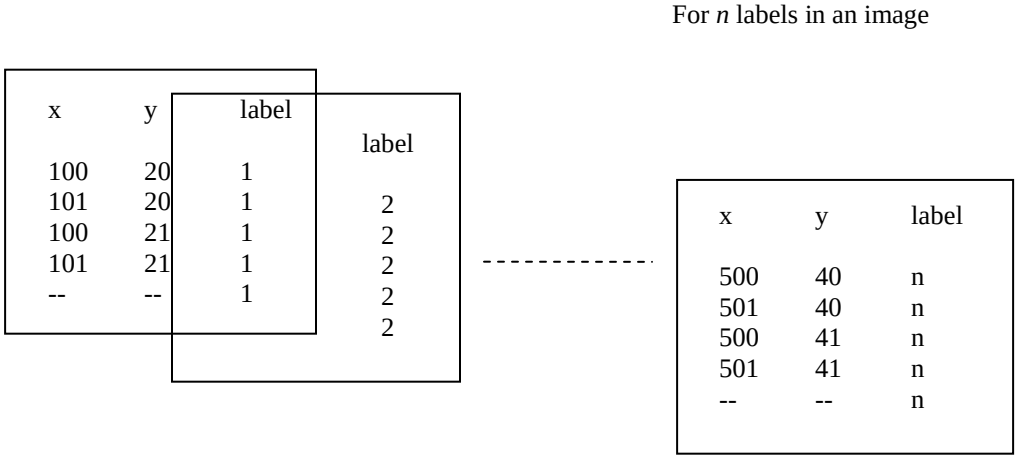


Figure 4.5. A sample data structure (3 dimensional 3 column) matrix for n labels in an image

cell. In this particular assay, over-expression of the GFP-fusion protein can lead to the formation of inclusion bodies. Unfortunately, these inclusion bodies have a tendency to accumulate at the poles of *E. coli* cells and look very similar to the sites of GFP localization associated with a positive protein-protein interaction. Distinguishing inclusion bodies from localization sites is important for reducing the number of false positive results associated with this assay. For this reason, experimental testing for inclusion bodies was conducted before computationally-based assessment of sub-cellular protein localization.

This problem is specific to this particular assay and may not be a consideration for other types of cells, labels, or protein localization experiments. Upon the observation of GFP localization at the poles in the images taken before induction of the DivIVA fusion protein, the sample is presumed to contain inclusion bodies and discarded, thereby nullifying its effect in the next stage of the study. At present, this is our best defense against misinterpretation of the data caused by inclusion bodies. A set of features that can be ascribed to inclusion bodies will be discussed in future work.

The results obtained after morphological and labeling operations are displayed in Figure 4.6.

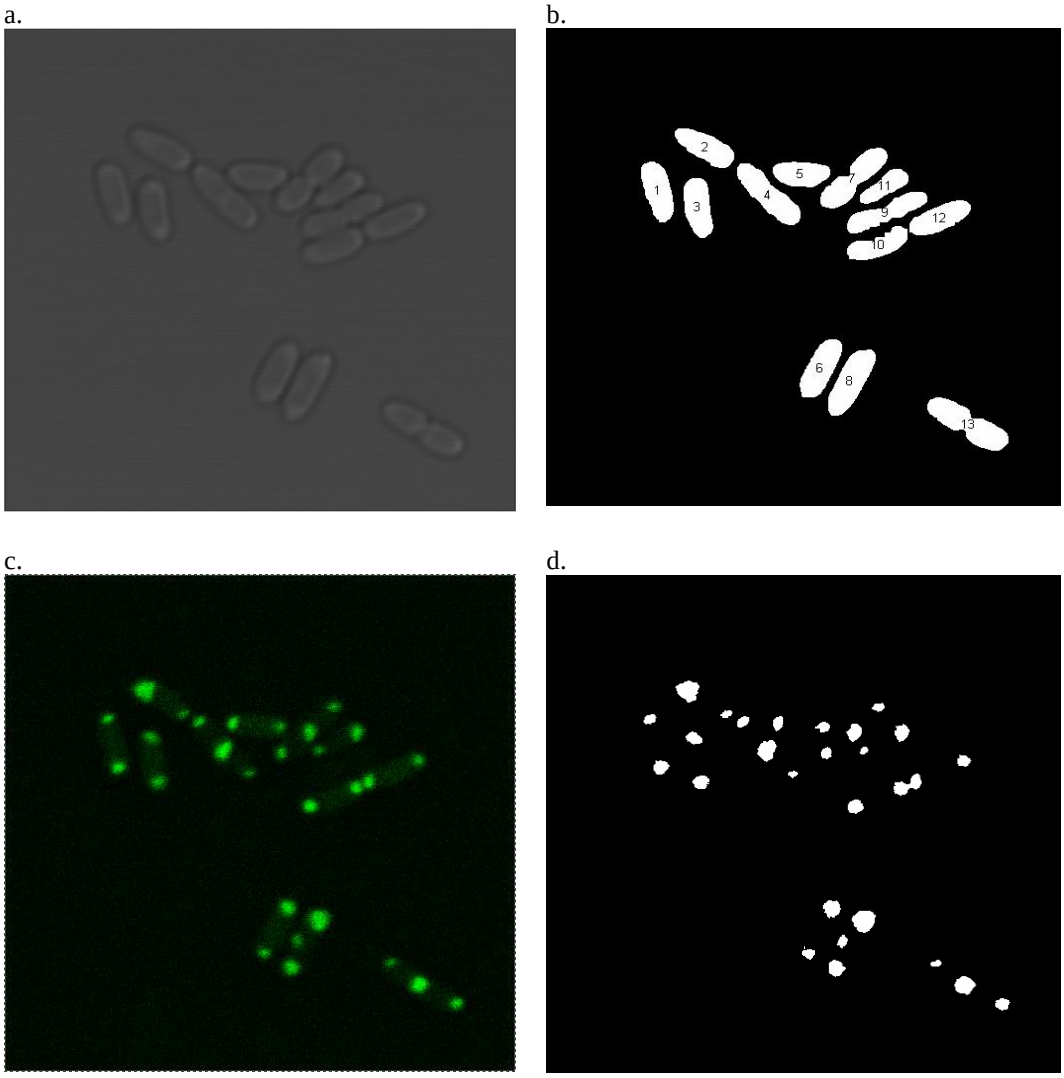


Figure 4.6. Image processing steps leading to a final pseudo colored image from sample DIC and fluorescent images of E. coli cells expressing a GFP-fusion protein

(a) Original DIC image; (b) Original fluorescence image; (c) Cell contours (from DIC image) obtained using after morphological operations (d) Identification of GFP-fusion protein localization sites using thresholding and morphological operations

4.4. Pattern recognition

In the case of this assay, we know the desired sites of protein localization and have designed the algorithm to determine whether the GFP localization occurs at the cell poles as expected for a positive protein-protein interaction. Once the number of GFP localization sites in each cell is identified, the distance between their respective COF's and COG is calculated and compared with the diameter of the cell. This procedure segments the cell into three parts (Figure 4.7) along the diameter, where the first and third segments are considered to be the cell poles. The number of segments is limited to three due to the small size of bacterial cells and the limits of optical resolution. The presence of localized GFP in the first and third segments is considered a positive result (protein-protein interaction) and other patterns are considered a negative result (no interaction).

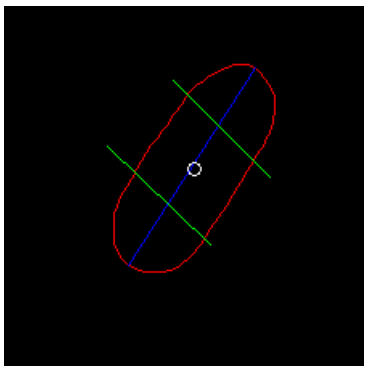


Figure 4.7. Pseudo colored image of a cell showing it divided into 3 parts along its diameter.

Blue – Diameter of the cell; Green – Hypothetical lines drawn at 1/3rd distance from either end of the diameter; Red – Cell contour from Canny edge detector; White – Center of gravity of the cell

A negative result having one or three sites of GFP localization may be interpreted several ways. For example, a cell showing localization in all three segments may be undergoing division as DivIVA is known to localize to the medial region of the cell during this stage of the cell cycle [16]. Alternatively, this result may occur if multiple cells are overlapping each other but have been identified as a single cell. Likewise, localization in only one segment may suggest that the cell under observation is slightly out of focus (a problem with the image acquisition procedure). This way, cells with GFP localization patterns not in the first and third segments are discarded as ambiguous results owing to improper segmentation of cells or overlapping of cells. Cells with uniform fluorescence in all segments are considered a negative result. The presence of inclusion bodies could generate visually positive but technically negative results, also leading to an ambiguity and are therefore evaluated before the induction of DivIVA fusion protein expression.

The *Roundness factor* described in the previous chapter acts as a tool to identify dividing cells and other cells displaying ambiguous results. Most cells in these classes display a different shape (dumbbell) than non-dividing cells. In such a situation, a dividing cell can be identified with a high roundness factor. The roundness factor is calculated using values obtained for area and perimeter of the cell as shown in equation 1. Thus, the various features extracted are put to use to enhance the quality of interpretation. The statistical features extracted from the data using various algorithms discussed in

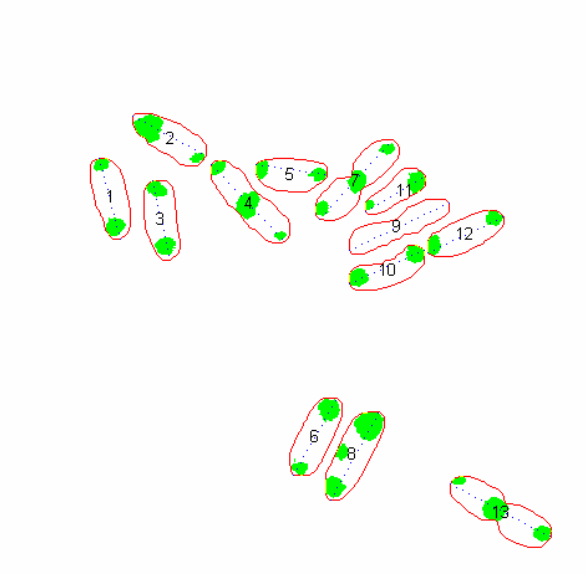


Figure 4.8. Final pseudo colored image

Final Pseudo colored image showing cell boundaries, cell diameter, sites of GFP-fusion protein localization, and cell labels (numbers) at the centers of gravity of each cell.

chapter 3 were used to characterize a pair of sample test images. These results are tabulated in Table 1. A final image (Figure 4.8) that represents individual features in different color channels is generated by pseudo coloring the target locations.

4.5. AFM image analyses

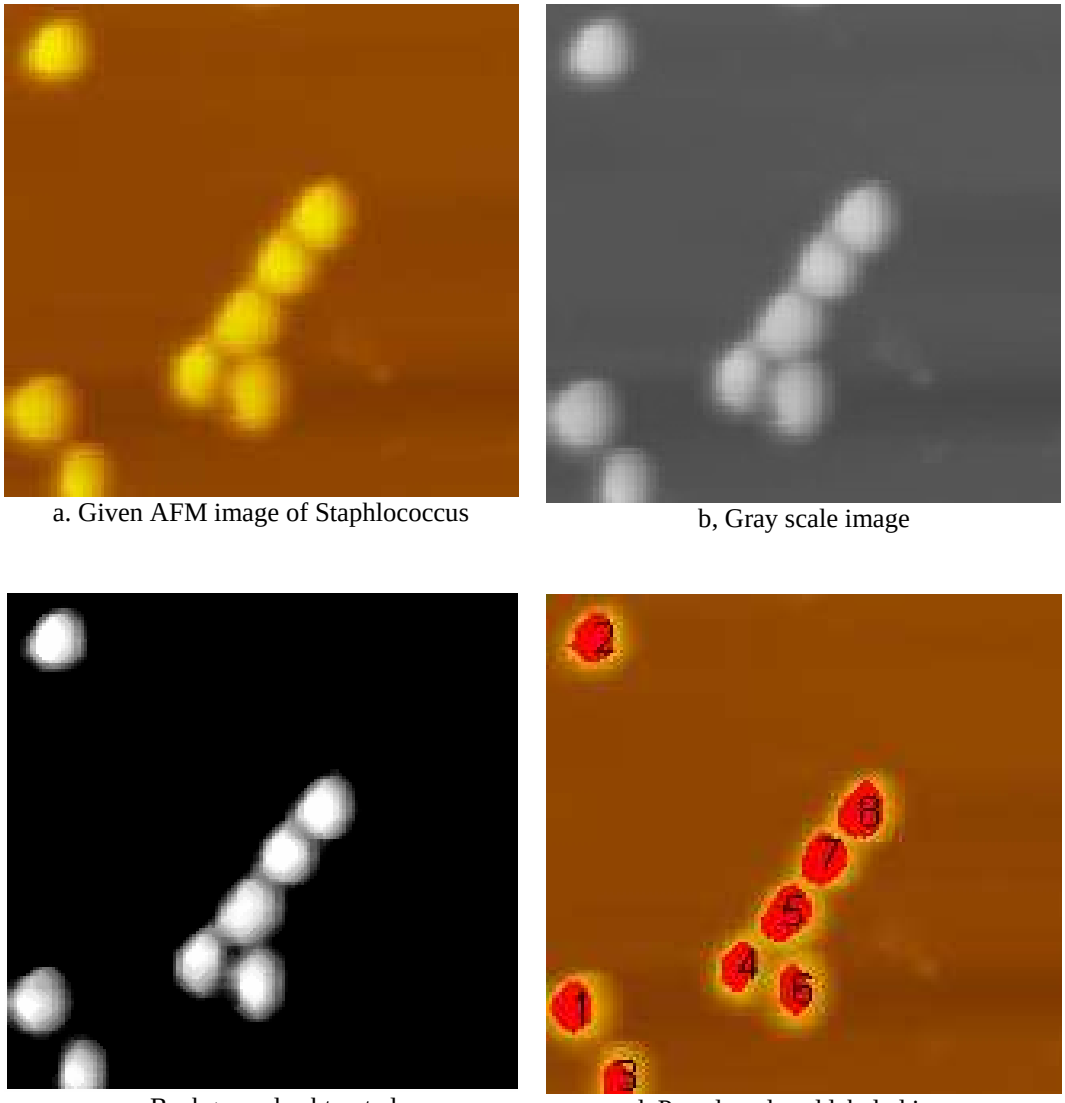
Once an AFM image of the sample under study is acquired, it goes through various image processing algorithms mentioned in the previous chapter before it can be converted to its

Table 1. Statistical Features of cells shown in Figure 4.8.

Cell No.	Area	Per	Diameter	Thinness Ratio	Number of GFP spots	COG	
1	1491	148	67	135.433	2	178	114
2	1461	141	67	128.433	2	131	162
3	1460	147	66	134.433	2	196	154
4	1876	167	86	154.433	3	183	225
5	1270	132	58	119.433	2	160	258
6	1595	150	71	137.433	3	371	278
7	1843	177	87	164.433	3	165	311
8	1872	166	81	153.433	3	385	308
9	1540	181	89	168.433	0	202	344
10	1256	143	66	130.433	2	237	336
11	922	113	56	100.433	2	173	343
12	1414	142	68	129.433	2	208	398
13	2163	192	92	179.433	3	432	427

Area, perimeter and diameter are given in pixels.
Percentage positive interactions in figure 4.8 – 84.6 (11 out of 13)

corresponding feature space. Segmentation of the region of interest is vital as it defines the area of desired objects within the image. There are a number of different features that can be extracted from a certain image but not all are necessarily pertinent. The selection of a set of pertinent features used to describe an image is by itself a huge task and obviously is application dependent. Sample AFM images of Sample AFM image (Figure 4.9.a) of E.Coli spheroplasts (cells with their outer wall enzymatically digested away) and Staphlococcus acquired at the Oak Ridge National Laboratories were used for our experiments.



a. Given AFM image of Staphlococcus

b, Gray scale image

c. Background subtracted

d. Pseudo colored labeled image

Figure 4.9. AFM image analysis

Table 2. Statistical Features of the spheroplasts segmented from image 4.9.a

Object No.	Area (micron2)	Perimeter (microns)	Diameter (nanometers)	Thinness Ratio	Center of gravity (x y)	
1	9.09126	2.43622	950.72	28.43363	136	10
2	9.56662	2.43622	1010.14	28.43363	16	17
3	5.46662	1.42608	713.04	11.43363	158	24
4	7.7246	2.19854	950.72	24.43363	122	62
5	11.8840	2.73332	1188.4	33.43363	105	76
6	7.4375	2.25796	950.72	25.43363	130	79
7	11.11154	2.61448	1069.56	31.43363	87	88
8	11.40864	2.6739	1128.98	32.43363	72	100

Simple morphological and shape defining features like the number of objects in the image, individual areas, diameter, perimeter and thinness ratio are calculated in terms of the number of pixels.

These numerical descriptors are converted physical quantities and converted to a physical quantity (Angstroms) form the raw data generated by the AFM (Table-2).

Chapter 5

Conclusion and Future work

This chapter summarizes basic contributions of the work discussed in this thesis. A discussion of future work based on these results is also presented.

5.1. Contributions and conclusions

Assigning functional information to the large number of proteins is a major concern in the field of proteomics. A direction towards that goal is to analyze protein-protein interactions, which often involves large sets of fluorescence microscopy images. Automated analyses of these large data sets can improve the speed, accuracy, and consistency of such analyses and a step towards the development of such a system was achieved by this work. A system for quantifying and identifying the location patterns of labeled proteins in live cells was successfully developed via algorithms implemented in MATLAB. Unique solutions to solve problems due to the ambiguity arising from cells undergoing division, adjoining cells, and the problems caused by background

fluorescence have been offered. This automated system achieved a percentage of about 84.6% in identifying the number of cells with positive interactions which was in close agreement with the one observed by an expert. A set of statistical descriptors were used to quantify the images to allow a provision for content base image retrieval system. AFM images were successfully analyzed, and quantified using various image processing and feature extraction algorithms developed in this work. The task here was to identify the objects of interest and extract statistical information from them and this was achieved successfully.

During this work, many problems were encountered and the solutions offered could be used to more or less similar image analysis projects. General solutions such as image thresholding methods, edge detection techniques and more specific solutions pertaining to connected component labeling and unique feature extraction techniques have been discussed and can be applied to various kinds of images. All methods developed in this work are applicable to real world images.

5.2. Future work

Work presented in this thesis is by no means a final solution to the problems behind its motivation. Results obtained from the work described in this thesis, are the foundation for further research focusing on the analysis of unknown protein interactions and setting up a database for localization patterns from many other protein-protein interactions.

The algorithm developed for this study can be easily extended to other applications involving multiple fluorescent labels or other imaging modalities with slight modifications. Such an algorithm can also be employed to reduce the size of image datasets by selecting those that possess desired features, such as positive interactions or specific location patterns.

In future, paths that mix the work done so far with active shape models [Coo 95] and those that dig deeper into the connections of digital morphology and statistical descriptors would be interesting to tread upon. Automated analyses for AFM images is a field with wide scope for research and the work presented in this thesis can be used as the first step towards the same. Developing algorithms to extract more vital statistical information from AFM images and to fuse such information with that extracted from other channels of the AFM is an interesting challenge. Combining these possibilities with the opportunity of applying the methods developed in the field of proteomics and the field of biomedicine in general seems like an ideal extension of the work done in this thesis.

A paper titled “Automated image analysis of fluorescent microscopic images to identify protein-protein interactions” by the author has been selected at the IEEE, *Engineering in medicine and biology society* and also has been short listed for the student paper competition. (will add the bioinformatics paper details once we send it out).

References

[Alb 91] Albrecht, T.R., Grütter, P., Horne, D., and Rugar, D., "Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity", *J. Appl. Phys.* 69(2), 668-673, 1991

[Bin 86] Binnig, G., Quate, C.F., and Gerber, Ch. (1986) Atomic force microscope. Phys. Rev. Lett. 56(9), 930-933

[Bol 97] M. V. Boland, M. K. Markey and R. F. Murphy, "Automated Classification of Cellular Protein Localization Patterns Obtained via Fluorescence Microscopy", *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 594-597 (1997)

[Bol 99] Boland,M.V. and Murphy,R.F. (1999) After sequencing: Quantitative analysis of protein localization. *IEEE Eng Med Biol Mag*, **18**, 115-119.

[Bol 01] Boland,M.V. and Murphy,R.F. (2001) A neural network classifier capable of recognizing the patterns of all major sub-cellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics*, **17**, 1213-1223.

[Can 86] Canny,J. (1986) A Computational Approach to Edge Detection. *IEEE Trans Pattern Anal Mach. Intell*.**8**, 679-698.

[Coo 95] T. F. Cootes, C. J. Taylor, D. H. Cooper et al. “Active Shape Models - their training and application.” *Computer Vision and Image Understanding* **61(1)**, pp. 38–59, January 1995.

[Dav 04] Davis,T.N. (2004) Protein localization in proteomics. *Curr. Opin Chem. Biol.*, **8**, 49-53.

[Din 02] Ding,Z., Zhao,Z., Jakubowski,S.J., Krishnamohan,A., Margolin,W. And Christie,P.J. (2002) A Novel Cytology-Based, Two-Hybrid Screen for Bacteria Applied to Protein-Protein Interaction Studies of a Type IV Secretion System. *J. Bacteriology*, **184**, 5572-5582.

[Edw 97] Edwards,D.H., and Errington,J. (1997) The *Bacillus subtilis* DivIVA protein targets to the division septum and controls the site specificity of cell division. *Mol. Microbiol*, **24**, 905-915.

[Gla 96] Glasbey,C.A. (1996) Problems in digital microscopy [abstract]. In *XVIIIth International Biometric Conference program, Amsterdam*, 183-200.

[Guz 95] Guzman,L., Belin,D., Carson, M.J., and Beckwith, J. (1995) Tight regulation, modulation, and high level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol*, **177**, 4121-4130.

[Har 79] R.M. Haralick, “Statistical and Structural Approaches to Texture”. *Proceedings of the IEEE,* VoL. 67, No. 5: 786 - 804, 1979.

[Har 92] Haralick,R.M., and Shapiro,L.G. (1992) *Computer and Robot Vision: Vol I*, Addison-Wesley, Reading, MA.

[Hua 02] K. Huang, J. Lin, J.A. Gajnak, and R.F. Murphy, "Image Content-based Retrieval and Automated Interpretation of Fluorescence Microscope Images via the Protein Subcellular Location Image Database", *Proc 2002 IEEE Intl Symp Biomed Imaging (ISBI 2002)*, pp. 325-328 (2002).

[Hua 04] K. Huang and R. F. Murphy, "Automated Classification of Subcellular Patterns in Multicell images without Segmentation into Single Cells", *Proc 2004 IEEE Intl Symp Biomed Imaging (ISBI 2004)*, pp. 1139-1142 (2004).

[Hus 95] Robert J. Huskey. Class notes BIOL 201, University of Virginia, http://www.biologie.uni-hamburg.de/b-online/library/bio201/bio201.html, 1995

[Jen 77] R.I. Jennrich, "Stepwise discriminant analysis," Statistical Methods for Digital Computers, K Enslein, A Ralston, and HS Wilf, Editors, pp. 77-95, John Wiley & Sons, New York, 1977

[Jur 96] Jurvelin, J.S., Müller, D.J., Wong, M., Studer, D., Engel, A. & Hunziker, E.B. (1996). Surface and sub-surface morphology of bovine humeral articular cartilage as assessed by atomic force- and transmission electron microscopy. *Journal of Structural Biology* **117**, 45-54.

[Kap 85] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, ‘‘A new method for gray-level picture thresholding using the entropy of the histogram,’’ *Graph. Models Image Process.* **29**, 273–285 ~1985.

[Kas 88] Michael Kass, Andrew Witkin, Demii Terzopoulos, “Snakes: Active contour models”, *International Journal Of Computer Vision,* 1988, 321 331.

[Kem 99] Van Kempen, Thesis: Image restoration in fluorescence microscopy, 1999.

[Lar 04] Larimer,F.W., Chain,P., Hauser,L., Malfatti,S., Do,L., Land,M.L., Pelletier,D.A., Beatty, J.T., Lang,A.S., Tabita,F.R., Gibson,J.L., Hanson,T.E., Bobst,C., Torres,J.L., Peres,C., Harrison,F.H., Gibson, J. and Harwood,C.S. (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. *Nat. Biotech*, **22**, 55-61 .

[Luc 74] Lucy,L.B. (1974) An iterative technique for the rectification of observed distributions. *Astron. J*., **79**, 745-754.

[Mic 03] Michigan tech. Module components. Atomic force microscopy, http://www.phy.mtu.edu/nue/afm.htm, 2003.

[Mur 02] R. F. Murphy, M. Velliste, and G. Porreca, “Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images”, *J. VLSI Sig. Proc. 35:* 311-321 (2003).

[Nee 03] Neelamani,R., Choi,H., and Baraniuk,R.G. (2004) ForWaRD: Fourier-Wavelet Regularized Deconvolution for Ill-Conditioned Systems. *IEEE Trans Signal Proc*, **52**, 418-433.

[Ots 79] Otsu,N. (1979) A threshold selection method from gray level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66.

[pal 89] N. R. Pal and S. K. Pal, ‘‘Entropic thresholding,’’ *Signal Process.* **16**, 97–108 ~1989

[Pen 03] Elizabeth Pennisi. "A Low Number Wins the GeneSweep Pool." *Science* **300**, 1484 (2003).

[Pri 02] Price,J.H., Goodacre,A., Hahn,K., Hodgson,L., Hunter,E.A., Krajewski,S., Murphy,R.F., Rabinovich,A., Reed,J.C. and Heynen,S. (2002) Advances in Molecular labeling, High Throughput Imaging and Machine Intelligence protend Powerful Functional Cellular Biochemistry Tools. *J. Cell Biochem Suppl*. **39**, 194-210.

[Reu 97] Reutter,B.W., Klein,G.J. and Huesman,R.H. (1997) Automated 3-D Segmentation of Respiratory-Gated PET Transmission Images. *IEEE Trans Nuclear Sci.*, **44**, 2473–2476.

[Ric 72] Richardson,W.H. (1972) Bayesian-Based Iterative Method of Image Restoration. *J. Opt. Soc. Am.*, **62**, 55.

[Rid 78] T. Ridler and S. Calvard. Picture thresholding using an iterative selection method.SMC, 8(8):630–632, 1978.

[Roq 02] E.J.S. Roques and R.F. Murphy, "Objective evaluation of differences in protein subcellular distribution", *Traffic 3:* 61-65 (2002).

[Ros 83] A. Rosenfeld and P. De La Torre, "Histogram concavity analysis as an aid in threshold selection," *IEEE Trans. Systems Man Cybernet.*, vol. 13, pp. 231-235, 1983.

[Sez 01] M. I. Sezan, ''A peak detection algorithm and its application to histogram-based image data reduction,'' *Graph. Models Image Process.* **29**, 47–59 ~1985

[Sha 01] Shattuck DW and Leahy R.M. (2001) Automated Graph-Based Analysis and Correction of Cortical Volume Topology. *IEEE Trans Med Imaging*,**20**, 1167-1177.

[Tsa 85] W. H. Tsai, ''Moment-preserving thresholding: A new approach,'' *Graph. Models Image Process.* **19**, 377–393 ~1985

[Xu 98] Xu,C. and Prince,J.L. (1998) Snakes, shapes, and gradient vector flow. *IEEE Trans. Image Processing*, **7**, 359–369.

[Zer 34] F. Zernike. Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode (Diffraction theory of the cut procedure and its improved form, the phase contrast method). *Physica*, **1**:pp. 689-704, 1934.

Vita

Sankar Venkataraman was born in the exotic city of Hyderabad in a much more exotic country, India on the first of October of 1980. He did his high school from Visakha Valley School situated in the coastal city of Vishakhapatnam. He went on to obtain his Bachelor of Engineering in the field of Electronics and instrumentation from Osmania University in 2003. During this period, he designed and implemented a micro-controller based electro-oculogram stimulator for the Nizam's Institute of medical Sciences (NIMS), Hyderabad. Following that, Sankar started his graduate program at The University of Tennessee, Knoxville in fall 2003, with a research focus in the field of medical imaging / image processing. He joined Dr. Mitchel J. Doktycz's group at the Oak Ridge National Laboratories (ORNL) as a research assistant in May 2004. He worked under the combined guidance of Dr. Doktycz and Dr. Hairong Qi from the University of Tennessee on the automation of fluorescence image analysis to identify positive protein-protein interactions resulting in the work presented in this thesis. His areas of interest are image processing, pattern recognition and computer vision.