TÉCNICO LISBOA

Cell cycle staging from DAPI and fluorescence microscopy

Ivan Sahumbaiev

Thesis to obtain the Master of Science Degree in

Bioengineering and Nanosystems

Supervisor(s): Professor Joao Miguel Raposo Sanches

Examination Committee

Chairperson: Professor Luis Joaquim Pina da Fonseca
Supervisor: Professor Joao Miguel Raposo Sanches
Member of the Committee: Doctor Manya Afonso

November 2015

Step by step and the thing is done.

Charles Atlas

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisors Prof. Joao Sanches and Doctor Raquel Seruca for their continuous support during my master thesis. The scientific guidance was crucial during the development of this work.

I would like to thank also Anabela Ferro for help her with the biological questions underlying this work and her availability to clarify all doubts I had and with scientific discussions.

My sincere thanks also go to Anton Popov for helping me and supporting me through my academic life.

Last but not the least, I would like to thank my Mom, Dad and my brother Alex. My parents always encouraged me to ask questions, to be curious about how things work. Thanks for encouraging me to be an independent thinker, and having confidence in my abilities to pursue new things and the ones that inspire me.

Abstract

Cell cycle denotes a set of biological processes and stages that occur sequentially along the dynamic evolution (life) of typical eukaryotic cells.

In eukaryotes, the cell cycle is divided in two major parts: growing(interphase) phase and division (mitosis) phase. The interphase can be divided in three sub-phases: gap 1 (G1 phase), in which a cell starts to increase in size; synthesis (S phase), in which the Deoxyribonucleic acid (DNA) replication and protein synthesis initiates; and gap 2 (G2 phase), during which cell growth continues and preparation for cell division occurs [1].

Getting cell cycle information is of great interest for biological and pharmacological research in order to understand the underlying biochemical processes associated with some pathological conditions and its therapeutical assessment. However, studies of the cell cycle have traditionally relied on the analysis of populations of cells, and they often require specific markers or the use of genetically modified systems, making it difficult to determine the cell cycle stage of individual, unsynchronized cells [2]. The most common method to determine cell phases is based on flow cytometry, which destroys the natural organization of cellular due to its fluidic requirements. So, for rare and unique biological samples, flow cytometry is not an option.

In this project, was developed a new approach for determination of cell cycle phases based on fluorescence microscopy and 4′,6′-diamidino-2-phenylinodole (DAPI) nuclear dye. DAPI dye is fluorescent stain that binds strongly to DNA and can be excited with ultraviolet light (maximum emission is 461 nm). Such new approach was chosen because DAPI binds stoichiometrically to DNA, allowing, in an unsupervised manner, the correlation and the quantification of features as the area and total intensity of the DAPI-stained nuclei of acquired fluorescent images, which are intrisically related to changes that occur in the nucleus throughout cell cycle progression. Moreover, developed method allows the preservation of the natural architecture of the samples analyzed.

Keywords

Cell Cycle, DAPI staining, Fluorescence Microscopy

Contents

List of Figures

List of Tables

Abbreviations

DNA Deoxyribonucleic acid

Cdk Cyclin−dependent kinase

DAPI 4′,6′-diamidino-2-phenylinodole

FUCCI Fluorescent ubiquitination-based cell-cycle indicator

FM Fluorescence Microscopy

ML Machine Learning

EM Expectation Maximization algorithm

SVD Singular Value Decomposition

SLAE System of Linear Algebraic Equations

GMM Gaussian Mixture Models

STD Standard deviation

USTD Weighted uncorrected standard deviation

DMEM Dulbecco's modified eagle medium

PBS Phosphate buffered saline

1

Introduction

Contents

1.1 Motivation

Progression through the cell cycle is one of the most fundamental features of cells. The coordination between genome duplication and faithful chromosome segregation to daughter cells is an integral part of growth and reproduction, and it essential to ensure genome stability and maintenance [8]. Deregulation of cell cycle control promotes genome instability and has been implicated in developmental abnormalities and numerous diseases, particularly cancer [9, 10].

Cell cycle status and progression has traditionally measured using population-based methods such as flow cytometry, which is generally not compatible with high-resolution cell biological techniques and does not allow tracking of individual cells over time [2]. Recent approaches have resulted in the development of methods to accurately determine and track the cell cycle phase of individual cells and to combine this information with other cellular features assessed by imaging, such as localization of a protein or morphological changes of organelles and cells. Most of these methodologies involve selective labeling of replicating cells [11, 12] staining with specific cell cycle markers [13] or expression of cell cycle phase-specific reporters [14]. Although these methods have proven useful for the study of key aspects of cell cycle regulation and coordination with other cellular functions such as Deoxyribonucleic acid (DNA) repair, senescence or apoptosis [15], they can only probe specific cell cycle stages, and thus combinatorial use of multiple methods is required to probe a given process comprehensively throughout the entire cell cycle.

1.2 Objective and Original Contributions

This project aims to evaluate the cell cycle progression through 4′,6′-diamidino-2-phenylinodole (DAPI) staining with fluorescence microscopy images, considering the analysis of inter-cellular and intracellular features. The approach is based on the accurate, quantification of the total intensity of nuclei and area of cells stained with the DAPI [16], through image processing.

The herein developed pipeline is based on nuclei segmentation, which is a relatively low throughput process. To overcome this drawback, an automatic algorithm was developed, that enables the correct segmentation of all nuclei present in each DAPI stained image. The segmentation pipeline consists of several steps, which are: denoising, which allows to convert heterogeneous objects to homogeneous; then color and contrast adjustment; and last step is Otsu thresholding.

1.3 Thesis Outline

In the first part of master thesis, theoretical background of the eukaryotic cell and it function is given. In the first Section, the cell cycle phases and mechanisms of transition between phases for eukaryotic cell are described. The second part of the Section is dedicated to the basics of fluorescence microscopy and a brief explain on the general methods used, to obtain fluorescence images is presented. Additionally, fluorescence dye DAPI is represented, as a substance, which is used for observation of the fluorescence phenomenon in biology. Section two is dedicated to image processing

toolkit, especially, cell segmentation and feature extraction from images.

The second part of the master thesis is devoted to data mining basics for analysis of fluorescence images. Additionally, the main algorithms for the cluster analysis problem are listed.

The third part of the master thesis is presented proposed method for determination of the cell cycle phase, which based on two features: area and total intensity of the nuclei. Additionally, the results of developed fully automatic segmentation algorithm is shown. Furthermore, the competitive analysis of designed algorithm is presented.

2

Biological background

Contents

2.1 The cell

Cell is the fundamental structural and functional unit of living organisms. Every eukaryotic cell has a plasma membrane, a cytoplasm and well defined nucleus.

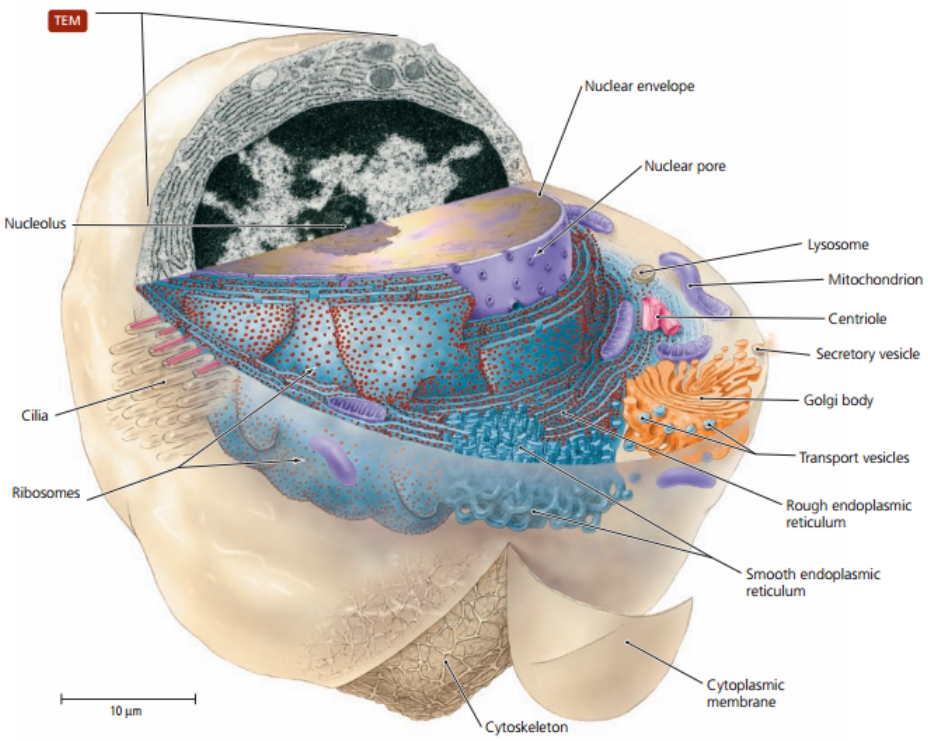


Figure 2.1: Structure of the eukaryotic cell. (Image source [3])

The *plasma membrane*, which surrounds the cell and keeps it intact, regulates what enters and exits from cell. The plasma membrane is a semipermeable phospholipid bilayer because it is active to certain molecules. Proteins present in the plasma membrane play important roles in allowing substances to enter the cell. The *nucleus* is a large structure that can often be seen with a light microscope. The nucleus contains the genetic material in chromosomes and is the control center of the cell: it controls the metabolic functioning and structural characteristics of the cell. The nucleolus is a region inside the nucleus. The *cytoplasm* is the portion of the cell between the nucleus and the plasma membrane. Cytoplasm is a gelatinous, semifluid medium that contains water and various types of molecules suspended or dissolved in the medium. The presence of proteins accounts for the semifluid nature of cytoplasm. Cells also have a *cytoskeleton*, a network of interconnected filaments and microtubules in the cytoplasm. The elements of the cytoskeleton maintain cell shape and allow the cell and its contents to move.

The Plasma Membrane. The plasma membrane separates the interior of the cell, termed the cytoplasm, from the outside. Plasma membrane integrity is necessary to the life of the cell. The phospholipid molecules, in plasma membrane, has a polar head and nonpolar tails. Because the polar heads are charged, they are hydrophilic (water-loving) and face outward, where they are likely to encounter a

watery environment. The nonpolar tails are hydrophobic (water-fearing) and face inward, where there is no water. Plasma membranes also contain a substantial number of cholesterol molecules. These molecules stabilize the phospholipid bilayer.

The Nucleus. The nucleus is a prominent structure in human cells. The nucleus is of primary importance because it stores the genetic information that determines the characteristics of the body‘s cells and their metabolic functioning. The unique chemical composition of each person‘s DNA forms the basis for DNA fingerprinting. Every cell contains a copy of genetic information, but each cell type has certain genes turned on and others turned off depending on the cell’s function. The protein content of a cell determine its structure and the function it can perform. The nucleus is separated from the cytoplasm by a double membrane known as the nuclear envelope, which is continuous with the endoplasmic reticulum. The nuclear envelope has nuclear pores of sufficient size to permit the passage of proteins into the nucleus and ribosomal subunits out of the nucleus. Additionally, the double membrane of the nuclear envelope surrounds and contains cellular DNA, protecting the vital genetic information.

The Cytoskeleton. Several types of filamentous protein structures form a cytoskeleton that helps maintain the cell‘s shape and either anchors the organelles or assists their movement as appropriate. The cytoskeleton includes microtubules, intermediate, filaments, and actin filaments. Microtubules can assemble and disassemble in a centrosome-dependent manner. The centrosome, lies near the nucleus, and is the cell region that contains the centrioles. Microtubules radiate from the centrosome, helping to maintain the shape of the cell and acting as tracks along which organelles move. It is well known that during cell division, microtubules form spindle fibers, which assist the movement of chromosomes. Intermediate filaments differ in structure and function. Because they are tough and resist stress, intermediate filaments often form cell-to-cell junctions. For example, intermediate filaments join skin cells in the outermost skin layer, the epidermis. Actin filaments are long, extremely thin fibers that usually occur in bundles or other groupings.

2.2 The Cell Cycle

The cell cycle is a complex biological process in which a set of cellular stages occurs in a sequential manner progressing to the cell division. The cell cycle is a period of time in which a cell is formed from its dividing parent cells until its own division into two cells occurs. [17]

In eukaryotes (cells with nucleus), the cell cycle is divided in two major phases: a growing(inter) phase and mitotic phase (M phase) as shown in Figure 2.2. The interphase consists of three discrete sub phases: gap 1 ($G1$ phase), in which a cell is increases size; synthesis (S phase), in which the DNA replication and protein synthesis takes place; and gap 2 ($G2$ phase), during which growth continues and preparation for cell division occurs. The last phase of the cell cycle is mitosis (M), during which the division of the nucleus happens. Cells not progressing through the cell cycle are called quiescent cells and remain in the so-called gap 0 ($G0$ phase). [17–19]

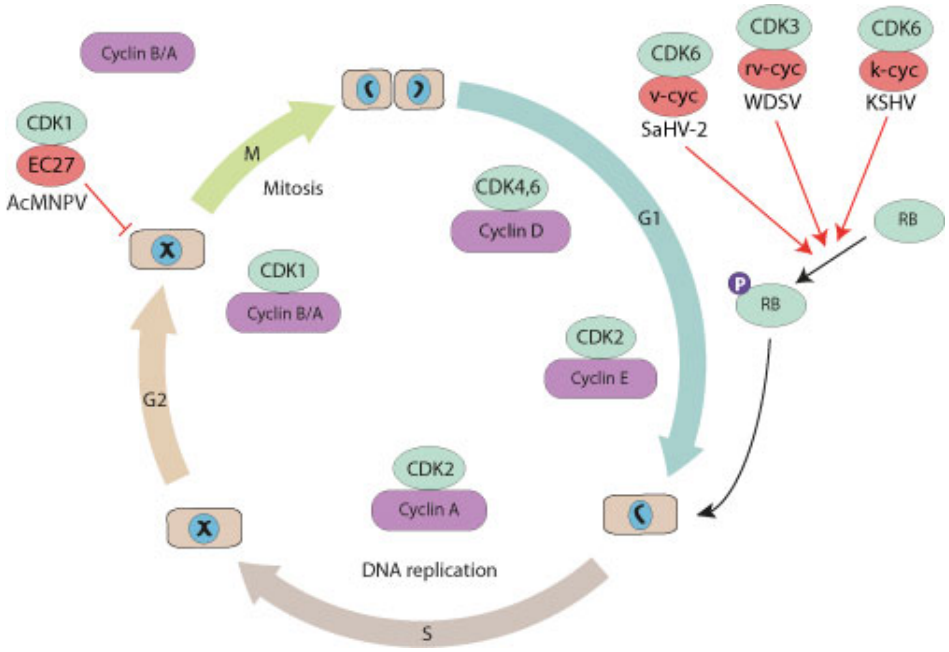


Figure 2.2: Phases of the eukaryotic cell cycle. (Image source [4])

The evolution along the cell cycle is mainly regulated by the Cyclin−dependent kinases (Cdks) and their respective activating cyclins. To avoid uncontrolled cell growth, the so-called checkpoints — a complex network of additional regulatory mechanisms, which are substantial to ensuring the correct order of the cell cycle events and the complementation of the one cell cycle before starting the next one — have to be passed. While the concentration of the Cdks remains approximately constant throughout the cell cycle, the cyclins are expressed and degraded periodically.

G0 phase. Cells in G0 phase do not have growth factors and mitogens in their immediate environment, do not express cyclins, and show high concentrations of cell cycle inhibitors, such as p27. These cells are named quiescent or post-mitotic cells and do not undergo cell division. [18–21]

G1 phase. The $G1$ phase directly follows cell division and is frequently also called post-mitotic presynthesis phase. The cell starts to grow, the content of the cell (cytoplasm) with the functional machinery (organelles) is formed [22]. During transition from $G1$ to S, cyclins D and E, and Cdk 2,4,6 are predominate. The D-type cyclins, the first cyclins expressed in the cell cycle, form complexes with Cdk4 and Cdk6 in the cytoplasm. These complexes after phosphorytlation of the Cdk subunit mediated by CAK are transported into the nucleus[19], where they initially phosplorylate Rb and some of its related pocket proteins, resulting in the release of HDAC from the RB/E2F complexes[23–25], and thus, in expression of the cyclin E [25–28]. Cyclin E forms complexes with Cdk2 in the cytoplasm. In the nucleus the cyclin E-Cdk2 complexes pass their bound p27 to activate cyclin D-Cdk4/6 complexes, leading to the activation of the cyclin E-Cdk2 complexes by additional phosphorylation by the cyclin D, to the destruction of the cyclin D-Cdk4/6 complexes [19]. Cdk4/6 and the phosphorylated cyclin D (pCyclin D) are transported back to the cytoplasm, where pCyclin D is degraded via ubiquitin-proteasome pathway, mediated by SCF [17, 29]. Additionally, it is involved in inhibition of p27 gene

expression via phosphorylation of transcription growth factor β (TGF-β). it also phosphorylates p27 [30, 31] to initiate its degradation by the ubiquitin-proteasome pathway mediated by SCF and, thus, enables the cell to pass the $G1 - S$ transition [18–21].

S phase. During S phase of the eukaryotic cell cycle, chromosomal DNA is replicated precisely once as a prelude to its segregation to the daughter cells at mitosis. After translocation of p27 from the nucleus the cell cycle irreversibly passes the $G1 - S$ checkpoint, also known restriction point (see Figure 2.3).

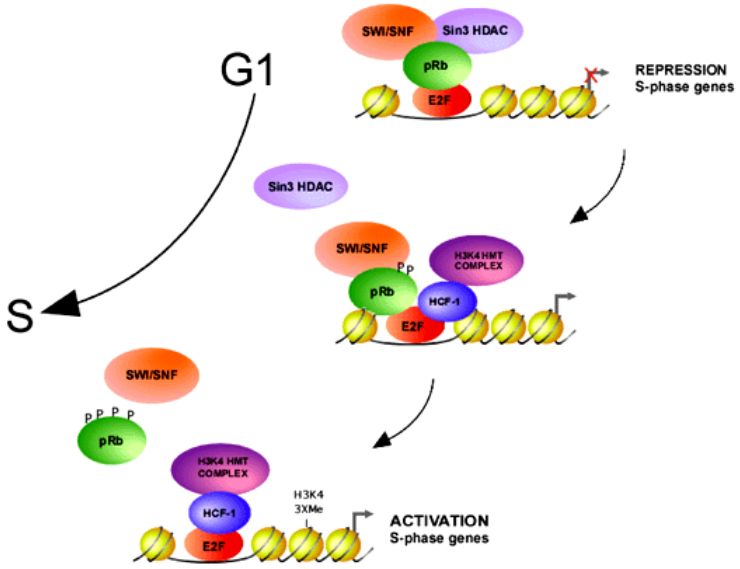


Figure 2.3: Model for activation of S-phase promoters by E2F. (Image source[5])

E2f is involved in the transcriptional activation of a variety of cell cycle-regulated genes [25, 27, 28, 32–34]. The transcription of cyclin A and — by passing some regulatory steps, i.e. with a lag — also of the B-type cyclins is activated by the trimeric NF-Y and for the B-type cyclins [35]. E2F also seems to be involved in regulating the transcription of cyclins A and B, because the cyclin A promotor as well as cyclin B promotor are containing both positively- and negatively-acting E2F binding sites [34–36]. Degradation of the cyclin E-Cdk2 complex releases Cdk2, which is bound immediately by cyclin A to initiate DNA replication, to phosphorylate and, thus, inactivate the DP-E2F heterodimer, and to secure S phase progression [37, 38]. If both subunits of the E2F heterodimer become phosphorylated E2F is degraded via the ubiquitin-proteasome pathway [39–41], but it seems that an additional phosphorylation of E2F through cyclin A-pCdk1 complexes is important to indicate this degradation [40]. Phosphorylated cyclin E (pCyclin E) is transported to the cytoplasm where it is degraded by the ubiquitin-proteasome pathway mediated by SCF [18–21].

G2 phase. The $G2$ phase begins when DNA replication has completed and in the cytoplasm cyclin B and Cdk1 form complexes, in which the Cdk subunits is promptly phosphorylated twice [19, 37,

42, 43]. The resulting inactive complexes accumulate in the cytoplasm until their activation via dephosphorylation of the inhibitory phosphorylation site and their transport to the nucleus where they regulate $G2/M$ transition [44, 45]. In case of DNA damage the cyclin B-Cdk1 complexes remain in or are transported back to the cytoplasm caused by inhibitory phosphorylation of Cdk1 mediated by Weel ($G2/M$ checkpoint)[37, 42, 43]. In the meantime, cyclin A-Cdk2 complexes decompose and cyclin A binds to Cdk1 to form an inactive, on the Cdk subunit twofold phosphorylated complex [46]. This complex is activated via dephosphorylation of the inhibitory phosphorylation site of Cdk1 mediated by Weel [42]. The active cyclin A-pCdk1 complexes activate phosphorylation of a variety of cytoskeletal proteins and together with cyclin B-Cdk1 complexes they secure the $G2/M$ phase transition [18–20].

Mitosis

The cyclin B-Cdk1 complexes, also known as MPF, are transported to the nucleus and activated, thus ensuring the progression through M phase and the correct completion of one cell cycle before starting the next one. Mitosis is a continuous process, and it conventionally divided into five stages: prophase, prometaphase,metaphase,anaphase and telophase.

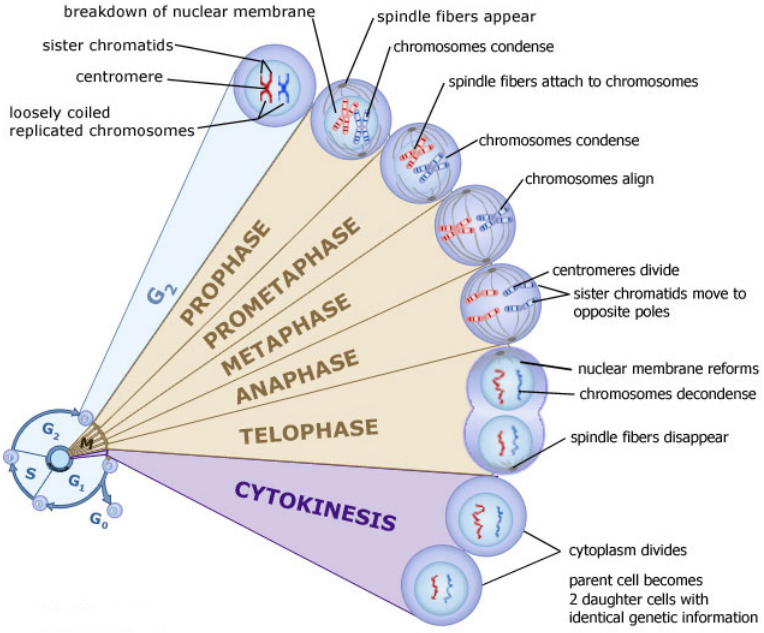


Figure 2.4: Mitotic phases. (Image source [6])

Prophase. Prophase occupies more than half of mitosis time. The nuclear membrane breaks down to form a number of small vesicles and the nucleolus disintegrates. A structure known as the centrosome duplicates itself to form two daughter centrosomes that migrate to opposite polars of the cell. The centrosomes organize the production of microtubules that form the spindle fibers that constitute the mitotic spindle. The chromosomes condense into compact structures. The replicated chromosome consists of two identical chromatids (or sister chromatids) held together by centromere. [38]

Prometaphase. The chromosomes, led by their centromeres, migrate to the equatorial plane in the midline of cell - at right-angles to the axis formed by the centrosomes. This region of the mitotic spindle is known as the *metaphase plate*. The spindle fibers bind to a structure associated with the centromere of each chromosome called a kinetochore. Individual spindle fibers bind to a kinetochore structure on each side of the centromere. The chromosomes continue to condense [47].

Metaphase. The chromosomes align themselves along the metaphase plate of the spindle apparatus[48].

Anaphase. The shortest stage of mitosis. The centromeres divide, and the sister chromatids of each chromosome are pulled apart and move to the opposite ends of the cell, pulled by spindle fibers attached to the kinetochore regions. The separated sister chromatids are now referred to as daughter chromosomes. (It is the alignment and separation in metaphase and anaphase that is important in ensuring that each daughter cell receives a copy of every chromosome.)

Telophase. The nuclear membrane reforms around the chromosomes grouped at either pole of the cell,the chromosomes uncoil and become diffuse, and the spindle fibers disappear.

Cytokinesis. The final cellular division form two new cells by the constriction of the cytoplasm and these new cells enter interphase.

2.3 Overview of techniques for determination the cell cycle phases

The study of DNA replication machinery is currently leading to the detection of novel biomarkers for cancer detection, outbursting into cell cycle directed therapies. The analysis of cell cycle behavior is a promising challenge to biologists and several standard methods, described below, have been used to determine the cell cycle phases.[49]

2.3.1 Fluorescence phenomenon

When energy levels of atoms bound in molecules are excited the result is optical radiation, this process is known as *luminescence*. Photoluminescence is a luminescence which is caused by ultraviolet, visible or infrared radiation. *Fluorescence* is photoluminescence which occurs when a material absorbs photons at some certain wavelength or group of wavelength from ultraviolet and visible spectrum and then emits photons of different band of wavelength.

Fluorescence phenomenon is illustrated in Figure 2.5 by the Jablonski diagram which describes photo physical processes in molecular system [50]. If the incident photon has sufficient energy matching the electronic band gap between the ground state (S_0) and the first excited state (S_1) of the fluorophore the photon can be absorbed, see Figure 2.5. If the photon has slightly more energy than between the electronic states it can still be absorbed and the excess energy brings the fluorophore in an even higher state of vibrational and/or rotational energy within S_1. Thus, a fluorophore can be excited by a range of wavelengths as long as the minimum energy is higher or equal to the $S_0 \rightarrow S_1$

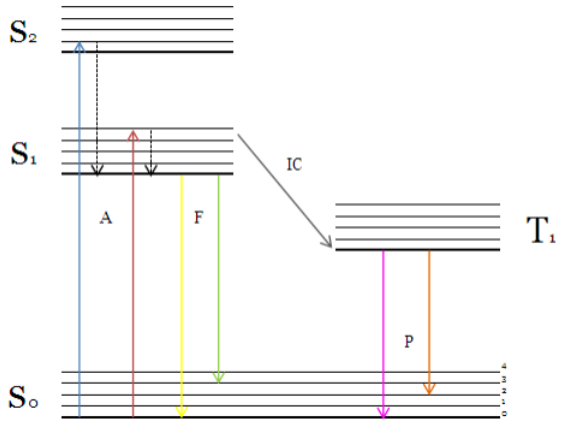


Figure 2.5: Jablonski diagram: A - Absorption, F - Fluorescence, IC - Intersystem crossing, P - Phosphorescence, S_0 - Ground state, S_1 - Higher energy states, T_1 - Triplet state

transition. The fluorophore is then relaxed very quickly to the lowest state of S_1 by internal conversion, phonons or some other process. From the ground state of S1 the fluorophore then relaxes to one of the many states of the electronic ground state S_0, and transmits a range of wavelengths typically redshifted from the excitation wavelength. This is ***fluorescence*** [51].

2.3.1.A Fluorescence microscopy

Nowadays, magnification and estimation size and structure of biological objects plays key role in understanding of working mechanisms. Fluorescence Microscopy (FM) gives a big contrast advantage as it allows labeling of specific structures in the sample The interesting parts of the sample are the only ones giving a signal and the background is essentially black. The possibility of several colors also exists as fluorophores of different transmission wavelengths can be used. The fluorophores are attached to molecules that attach to different parts of the sample, so called markers [51].

2.3.2 Flow Cytometry

Flow cytometry is a widely used powerful tool for studying many aspects of cell biology.It specifically allows for multi-parametric analysis of the physical and chemical characteristics of cells at a high rate(over a thousand cell/second). The idea of the method is to create a flow of the cells, but converting regular cell culture leads to destroying the natural architecture of the nuclei. Since, the flow has been created, the cells are passing through, they are excited by a laser, allowing measurements of cell size and internal complexity, as well as detection of fluorescent antibodies or stains on the cell (see Fig. 2.6)[7]. Furthermore, this technique rapidly quantifies small differences between cell populations using statistically-significant number of events. [52, 53]

Flow cytometry is one of the most commonly used methods to evaluate DNA content and has been used for the past four decades as the gold standard tool for cell cycle analysis. This method relies on the labeling of cells with DNA fluorophores such as DAPI, in order to accurately assess the DNA content of a cell.[2]

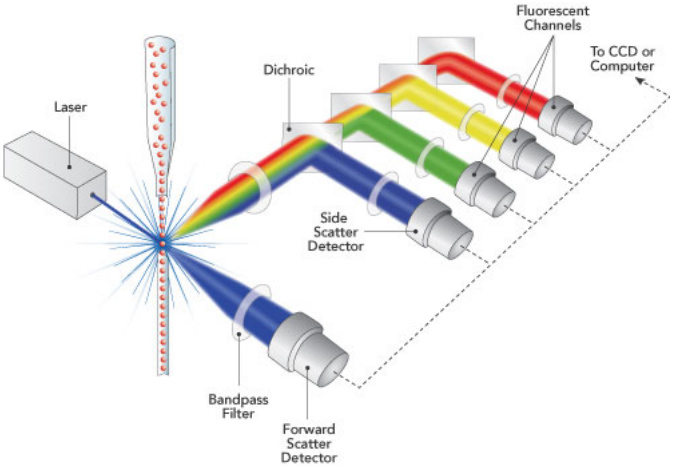


Figure 2.6: Flow cytometry overview. (Image source [7])

Despite being a commonly used technique, it presents some limitations, such as requiring specific instrumentation, not allowing tracking of individual cells over time and limited ability to pair the cell cycle status of individual cells with subcellular morphological features, which require detection by high-resolution analysis. Unlike fluorescence microscopy, flow cytometry uses cellular suspension cultures. [2, 54].

2.3.3 Labelling cells

The color of the dye is related with the wavelength of the specific radiated energy after excitation. Dyes injected for these purposes should have the following properties: *a*) they should be visible, either immediately during or after chemical reaction; *b*) they should remain in the injected cell, either because they are too large to move across the cell membrane and through gap junctions or because they are strongly bound by the cytoplasm; *c*) they should not be toxic, although this requirement can be relaxed if the tissue is to be processed immediately after the cell has been injected; *d*) they should be stable and not break down to give products with different properties; *e*) they should withstand histological processing. Three classes of compound are used for this purpose:

1. Inherently fluorescent molecules and those tagged with a fluorescent probe. Lucifer Yellow (MW 457) and carboxyfluorescein (MW 376) are the most popular fluorescent compounds for determining overall cellular architecture. Both pass through gap junctions and carboxyfluorescein cannot be fixed. Lucifer Yellow withstands fixation well but as with all other dyes some fluorescence intensity is lost. Passage through gap junctions can be prevented by conjugation of the fluorophore to dextrans. Dextrans (MWs 3000-70000) can be coupled to fluorescein, rhodamine isothiocyanate or Texas Red[55].

Advantages	Disadvantages
1. Can be pressure injected or iontophoresed. 2. Can be seen in living cells with appropriate fluorescent illumination. 3. Are not toxic provided the amount injected is kept fairly low. 4. Do not break down. 5. Will withstand routine fixation and embedding techniques, provided the fixative or mounting agents do not generate auto-fluorescence.	1.Limit of detection determined by threshold of fluorescence. Detection levels can be improved by electronic image intensification. 2. Fluorescence fades under continuous illumination. This can be reduced by using anti-fade mounting agents. 3. Fluorescein fades particularly fast, but is more fluorescent than rhodamine or Texas Red. 4. Sometimes become incorporated into cellular organelles with time, making fluorescence particulate. 5. Margin between visible not toxic, and visible but toxic is narrow.

Table 2.1: The advantages and disadvantages of two fluorescent dyes: Lucifer Yellow and Carboxyfluorescein

2. The *carbocyanine dyes*. Octadecyl (C_{18})-indocarbocyanine (DiL) and oxycarbocyanine (DiO) are highly fluorescent lipophilic compounds. They dissolve in, and diffuse throughout, the lipids of the plasma membrane. They are not toxic and they have been reported to remain in the cell membrane for up to one year [56]. The diffusion rate for these compounds is slow, however, carbocyanines with unsaturated alkyl chain segments (FAST-DiI and FAST-DiO) exhibit accelerated diffusion rates. The polyunsaturated "DiASP" compounds (N-4(4-dilinoleylaminostyryl)-N-methylpyridinium iodide and related molecules) are also reported to diffuse more rapidly. Because the carbocyanines are insoluble in water they must either be pressure injected into cells in solution in DMSO or alcohol or applied to the cell membrane in which they rapidly dissolve. DiI and DiO can be visualized by fluorescence microscopy. DiI has similar excitation properties to rhodamine, excited by green it fluoresces red. DiO is similar to fluorescein in that it is excited by blue light and produces green fluorescence. DiAsp has a broad excitation spectrum and fluoresces orange.

Advantages	Disadvantages
1. They are not toxic and can remain in the cell membrane without harm over several years.	1. Not water soluble. 2. They tend to fade quickly particularly in laser scanning confocal microscopy. 3. Long diffusion times.

Table 2.2: The advantages and disadvantages of two fluorescent dyes: Octadecyl-indocarbocyanine and oxycarbocyanine

3. *Biocytin*. Intracellular marker [57] comprising a highly soluble conjugate of biotin and lysine that has a high binding affinity for avidin. The injected biocytin is visualised by attaching a label to avidin, e.g. a fluorescent label such as FITC or rhodamine, or a chromogenic enzyme such as HRP.

Advantages	Disadvantages
1. Highly soluble in aqueous solutions. 2. Can be pressure injected or iontophoresed. 3. Low toxicity. 4. Does not break down. 5. Good fluorescent, visible light, or electron microscopic visibility after avidin reaction.	1. Can only be seen after avidin reaction. 2. Reaction penetration limited to about 100 μm even with detergents or surfactants so tissue may have to be sectioned. 3. Some ultrastructural degradation from penetration agents. 4. Can pass between coupled cells. 5. Occurs naturally in trace amounts.

Table 2.3: The advantages and disadvantages of intercellular marker Biocetin

2.3.4 DAPI stain

DAPI or 4′,6′-diamidino-2-phenylinodole is a commonly used fluorescent stain that binds strongly to A-T rich regions of the DNA and weakly binds to the RNA. The chemical formula is shown in Fig. 2.7.

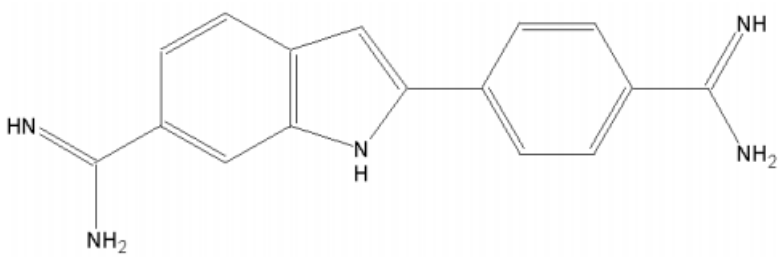


Figure 2.7: DAPI- 4′,6′-diamidino-2- phenylindole — chemical formula

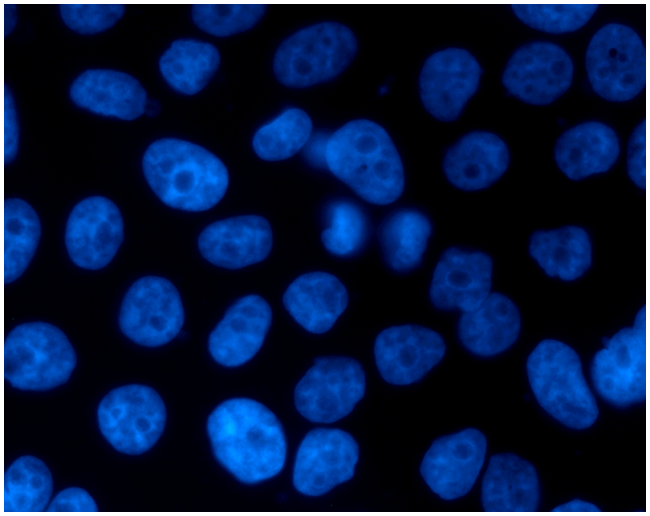


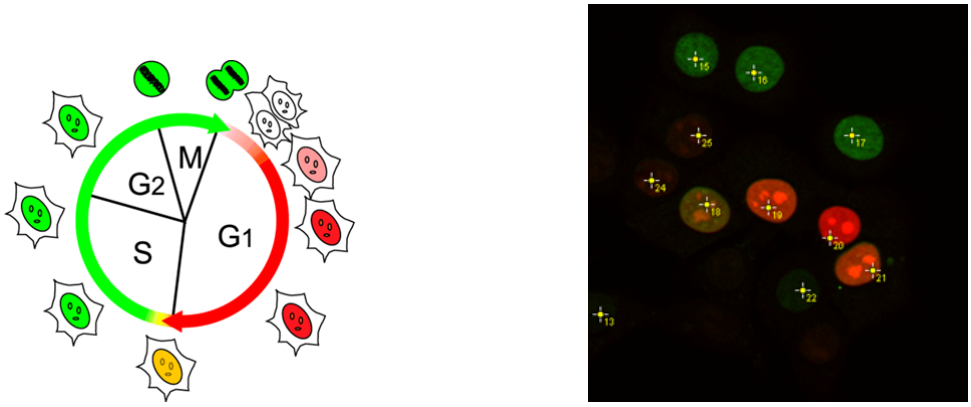
Figure 2.8: Fluorescence microscopy image with DAPI stain

The injection of the dye is making manually by creating several pores in the cell membrane. The dye can be excited with ultraviolet light and stained cells can be captured with fluorescent microscopy. Absorption maximum of the DAPI is at 356 nm and its emission maximum is at 462 nm. Due due

wave length — 462 nm, DAPI-stain fluoresces mainly in the blue light range. The image of the cell culture stained with DAPI is shown in Fig.2.8

2.3.5 Fluorescent ubiquitination-based cell-cycle indicator (FUCCI)

This method is based on fluorescent ubiquitination-based cell cycle indicator and exploits cell-cycle-dependent proteolysis of the ubiquitinated oscillators, Cdt1 and Geminin, to specifically mark the $G1/S$ transition in living cells. By fusing the red- and green-emitting fluorescent proteins mKO2 and Azami Green (mAG) to portions of Cdt1 and Geminin, respectively [58]. Specifically, the nuclei of cells in $G1$ phase (and G0) appear red, while those of cells in $S/G2/M$ appear green (see Figure 2.9). During the transition from $G1$ to S phase, cell nuclei turn yellow, clearly marking cells that have started the DNA replication. This system, thus allows the easy visual readout of the cell cycle progression.



(a) **(b)**

Figure 2.9: FUCCI labeling process. During *G1* phase, the nuclei of FUCCI-expressing cells appear red; during *S/G2/M* appear green.

The color changes exhibited by FUCCI are based upon the reciprocal activities of the ubiquitin E3 ligase complexes APC^{Cdh1} and SCF^{Skp2} [59]. While APC^{Cdh1} functions primarily during $G1$ phase, SCF^{Skp2} is most active during S, *G2*, and M phases. Consequently, the APC^{Cdh1} and SCF^{Skp2} substrates Germinin and Cdt1 are specifically degraded during $G1$ and $S/G2/M$, respectively [60].

The fluorescence chimeras utilized by FUCCI are distributed exclusively in the nucleus, this method is amenable to imaging studies using fluorescent biosensors designed to track signaling dynamics in live cells [61]. FUCCI, is a powerful tool to visualize cell-cycle progression in living cells, lays a foundation for studying the cell cycle in a variety of cellular contexts. In particular, its ability to mark the $G1/S$ transition with high contrast.

The core idea of current work is to use DAPI-stained and FM images to develop a bioimaging tool that enable the cell cycle assesment of single or populations of cells, because treatment the cell culture with **FUCCI!** (**FUCCI!**) is expansive. Also DAPI has some unique properties, specifically binds stoichiometrically to DNA (see chapter 2.3.4) the intensity of a blue color in fluorescent images correspond to the amount of DNA in the nuclei. Visually it is complicated to estimate how many cells

are in G_1 or S or G_2/M (image processing algorithms will be described in next chapters). Biologically it can be done with the well-established FUCCI system.

2.4 Biological material

2.4.1 Cell culture

Fluorescence microscopy (FM) images obtained from cells in culture were the basis of this work, though only DAPI (blue) information was used in the course of this work to extract the quantitive features that intrinsically mirrors the changes that occur in the nucleus throughout cell cycle progression.

A total of 47 fluorescent images, comprising 998 DAPI-stained nuclei were acuired from NMuMG-Fucci2 *in vitro* cultures. Fucci2-expressing cells were obtained from Riken Institute, Japan. Cells were grown in complete Dulbecco's modified eagle medium (DMEM) supplemented with 10$\%$ FBS(fetal bovine serum), 1$\%$ penicillin/streptomycin and 10μg/mL insulin and seeded onto glass cover slips in 6-well plates until they were 70-80$\%$ confluence [62]. Cells were then washed with 1 mL of Phosphate buffered saline (PBS) (NaCl 137 mM, KCl 2.7 mM, Na2HPO4 10 mM, KH2PO4 7.4 mM) and then fixed with 1 mL of 4$\%$ formaldehyde in PBS (freshly made) for 15 min at room temperature in the dark.Cells were then quenched for 10 min at room temperature with 1 mL 10 mM of NH_4Cl and subsequently permeabilized for 10 min with 0.2$\%$ Triton X-100. Nuclei of fixed cells were stained with DAPI 1 mg/mL for 2 min, in the dark at room temperature and the coverslips were mounted on slides using Vectashield plain mounting medium. Prepared slides were kept at 4 $^\circ$C and protected from light prior to imaging.

2.4.2 Fluorescence imaging

Images were captured Zeiss Apotome Axiovert 200M ImagerZ1 fluorescence microscope and with the 40X/1.3 oil DICI(UV)VIS-IR objective(carl Zeiss, Thornwood, NY). Aiming at extracting complete information on DAPI-stained nuclei, multiple images in different planes along the z-axis (60 stacks) were acquired and then merged together by projecting into a single image. The acquisition parameters were maintained constant in all experiments. Captured images were processed with Zeiss Axion Vision and ImageJ software.[63]

3

METHODOLOGY

Contents

3.1 Images preprocessing pipeline

In this work, automatic bioinformatic application was designed and implemented in MatLab®, to manage the images sets and process them with the algorithms described throughout this chapter.

The pipeline for the preprocessing stage of a single FM image is depicted in Fig. 3.1. The blue channel was only used, corresponding to the DAPI plane of the FM images. The first and second steps of this pipeline are the common ground of the several analyses pursued. Firstly, a denoising algorithm followed by a contrast and intensity adjustment were applied to the DAPI plane of the FM images. Then, the nuclei segmentation strategy pursued consisted on the consecutive implementation of Otsu's method, morphological operators. Ultimately, each nucleus becomes uniquely labeled, which it turn enables the acquisition of morphological features of each one.

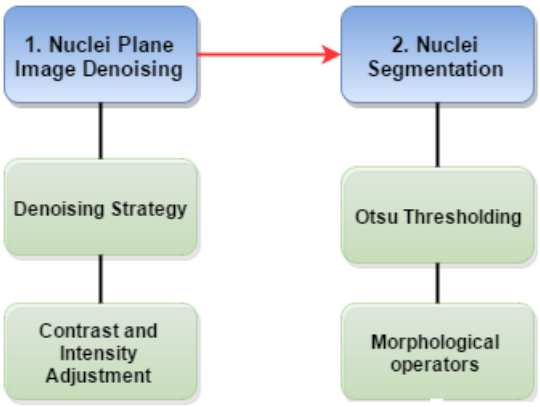


Figure 3.1: FM images preprocessing pipeline

A detailed description of each stage of the preprocessing pipeline is displayed in the following sections of this chapter.

3.1.1 Nuclei plane image denoising

Denoising it is first step in image processing procedure, because all instrumentation systems always originate some amount of noise. Mathematically developed methods can partly overcome this hindrances, emphasizing to a greater extent the underlying relevant data [64].

Fluorescence microscopy related data are corrupted by noise that follows the Poisson distribution, given the discrete nature of the acquisition [65] . The probability function can be defined as:

$$Pr(X = k) = \frac{\lambda^k \exp^{-k}}{k!},$$

where $\lambda > 0$ is distribution parameter.

The denoising strategy is following a Bayesian algorithm which was developed to remove Poisson intensity noise, where the optimization task follows a maximum-a-posterior (MAP) criterion [65]:

$$\widehat{Z} = argmin_z E(Z, Y),$$

where the energy function is given by $E(Z,Y) = E_y(Z,Y) + E_z(Z)$.

$E_y(Z,Y)$ is the data fidelity term, where the Poisson distribution models the observation noise and $E_z(Z)$is the prior term regularizing the solution. The regularizer term is necessary to introduce some apriori information about the solution, since the estimation problem is ill-posed [64].

Assuming that observations are independent and noise compliance with a Poisson distribution, the data can be described by the anti-logarithm of the product of all observation probabilities as follows:

$$E_y(Z,Y) = -\log\left[\prod_{i,j=0}^{N-1,M-1} p(y_{i,j}|z_{i,j})\right] = \sum_{i,j=0}^{N-1,M-1} |z_{i,j} - y_{i,j}\log(z_{i,j}) + C|,$$

where C is a constant.

The prior distribution function used is denominated a log total variation (TV) potential function since it uses logarithms of ratios of neighboring pixel intensities, allowing interpretation of differences between neighbors according to orders of magnitude.

$$TVlog \sqrt[\infty]{log^2 \frac{z}{\varsigma}},$$

where z and ς are neighboring pixels.

The TV-log had faults in homogeneous regions, leading to efficient high frequency noise removal in these regions, whilst it has a smaller penalization in sharp transitions, which is useful when considering biological images with intrinsically abrupt transitions that should be preserved, such as:

$$E_z(Z) = \alpha \sum_{i,j=0}^{N-1,M-1} \sqrt{\log^2 \frac{z_{i,j}}{z_{i-1,j}} + \log^2 \frac{z_{i,j}}{z_{i,j-1}}},$$

where α is a positive prior parameter.

Since the resulting energy function is nonconvex and complex, a change of variable is performed such that $w = g(z) = \log(z)$ [65]

3.1.2 Segmentation

Segmentation is a process that allows the division of images into regions that distinguish objects of interest. This process is characterized for having a typically low accuracy and consistency output when applied to most images. It is often referred to as the first, most important and most difficult step in image processing, determining the success of the final analysis. Several segmentation techniques, namely watershed and thresholding techniques, clustering and edge detection methods and hybrid techniques, among others, have been used in order to partition images with the best outcome possible.[66]

In this work, the segmentation strategy was applied to the denoised and contrast and intensity adjusted DAPI plane of the FM images. This strategy consisted on the consecutive implementation of Otsu thresholding, morphological operators to each image. An overall illustration of the segmentation strategy is shown in Fig.

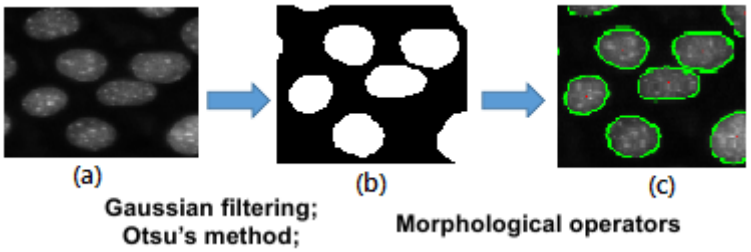


Figure 3.2: Overview of the cell nuclei segmentation and labeling procedure. (a) — input DAPI plane image; (b) — unique mask for each cell based on Gaussian filtering and Otsu's thresholding; (c) — obtained boundaries in the original image.

3.1.2.A Otsu thresholding

Otsu's method is a common technique applied for image thresholding segmentation. This method assumes that a gray level image contains two classes of pixels (foreground and background pixels) following a bi-modal histogram. The threshold value corresponds to a value within the deep and sharp valley between the two classes, represented as peaks in the histogram. Gray level images are then converted into binary, where black pixels correspond to the background and white pixels correspond to foreground objects. Moreover, a uniform background is required, in order to avoid diffuse object-background transitions and consequent difficulty to make an optimal threshold level. [67–69]

Formulation. Let a given picture be represented by a discrete number of gray levels [1,2,...,L], and the number of pixels at each level be denoted by n_i. The total number of pixels in the image is $N = \sum_{i=1}^{L} n_i$. By normalizing the histogram data and describing it as a probability distribution, we can write $p_i = \frac{n_i}{N}$, such that $p_i \geq 0$ and $\sum_{i=1}^{L} p_i = 1$.

Let us assume a random intensity threshold level k dichotomizing pixels into 2 classes C_0 and C_1. C_0 encompasses levels $[k+1, ..., L]$. The probabilities of the class occurrence are respectively $w_0 = Prob(C_0) = \sum_{i=1}^{k} p_i = w(k)$ and $w_1 = Prob(C_1) = \sum_{i=k+1}^{L} p_i = 1 - w(k)$.

The class mean levels are respectively $\mu_0 = \sum_{i=1}^{k} iProb(i|C_0)) = \sum_{i=1}^{k} \frac{p_i}{w_0} = \frac{\mu(k)}{w(k)}$ and $\mu_1 = \sum_{i=1}^{k} iProb(i|C_1)) = \sum_{i=1}^{k} \frac{p_i}{w_1} = \frac{\mu_T - \mu(k)}{1-w(k)}$ with $\mu(k) = \sum_{i=1}^{k} p_i$; $\mu_T = \mu(L)$. Finally, the class variances are given by $\sigma_0^2 = \sum_{i=1}^{k} (i-\mu_0)^2 Prob(i|C_0) = \sum_{i=1}^{k} (1-\mu_0)^2 \frac{p_i}{w_0}$ and $\sigma_1^2 = \sum_{i=k+1}^{L} (i-\mu_1)^2 Prob(i|C_1) = \sum_{i=k+1}^{L} (1-\mu_1)^2 \frac{p_i}{w_1}$.

The discriminating criterion measure used, evaluating the "goodness"/separability of the threshold level *k* is $\eta = \frac{\sigma_B^2}{\sigma_T^2}$, where $\sigma_B^2 = w_0(\mu_0 - \mu_T)^2 + w_1(\mu_0 - \mu_T)^2 = w_0 w_1 (\mu_1 - \mu_0)^2$ represents the between class variance and $sigma_T^2 = \sum_{i=1}^{L} (i-\mu_T)^2 p_i$ represents the total variance.

The optimization problem is defined by finding the threshold k^* that maximizes the object function η.

$$\eta(k^*) = \max_{1 \leq L \leq \eta(k)}, 0 \leq w(k) \leq 1$$

Or, since $sigma_T^2$ is constant with respect to k,

$$\sigma_B^2 = \max_{1 \leq L \leq \sigma_B^2(k)}, 0 \leq w(k) \leq 1,$$

where $\sigma_B^2(k)$ can be written as:

$$\sigma_B^2(k) = \frac{[\mu_T w(k) - \mu(k)]^2}{w(k)\,[1 - w(k)]}.$$

3.1.2.B Morphological operators

Thresholding techniques, such as Otsu's method, enable the definition of the border between objects and background by a threshold, however these methods present difficulties in the separation close adjacent objects. For that reason, morphological operators were employed. [70]

Firstly, the structures that were lighter than their surroundings and connected to the image border were suppressed. Afterwards, the filling of holes present in each object and the morphological opening of the images were performed. These both steps allowed the filling of background pixels that were present within objects and the open up of spaces between just-touching objects, by removing pixel noise from the binary image, respectively. [67]

3.1.3 Feature Extraction

In order to determine the cell cycle phase of each cell present in the acquired FM images, the area and total in tensity of DAPI in each nucleus were used. Both features are intrinsically related to changes that occur in the nucleus during cell cycle progression, namely, the growth of the nucleus in G1 and G2 phases, and the DNA replication during the S phase.

The area of each nucleus is defined as the actual number of pixels in the region (N_{PR}), such as

$$Area = \Sigma n_{PR}$$

The measurement of the total intensity of nuclear DAPI staining yields the relative amount of DNA in each nucleus and was defined as

$$Totalintensity = \int_A Intensity \cdot dA = \sum_{i=1}^{N} Intensity_i,$$

where N is the total number of pixels within nucleus.

Theoretically, by the area versus total intensity features plane, shown in Fig.3.3, one should expect a first region corresponding to cells in G1 phase, in which a significant rise of areas and total intensity value would be detected, characteristic of the cell growth with 2N DNA content during this phase. Then, a region of almost constant areas and increased total intensities (from 2N to 4N) should be indicate the S phase, in which DNA replication takes place. Finally, a region with approximately the doubled total intensity of the first region and spread areas should appear, indicating the cell growth in G2 with fixed DNA content, 4N.

3.1.4 Data Standardization

Data standardization is common pre-processing technique applied in data mining. A direct application of geometric measures to attributes, such as distance in case of the k-means clustering

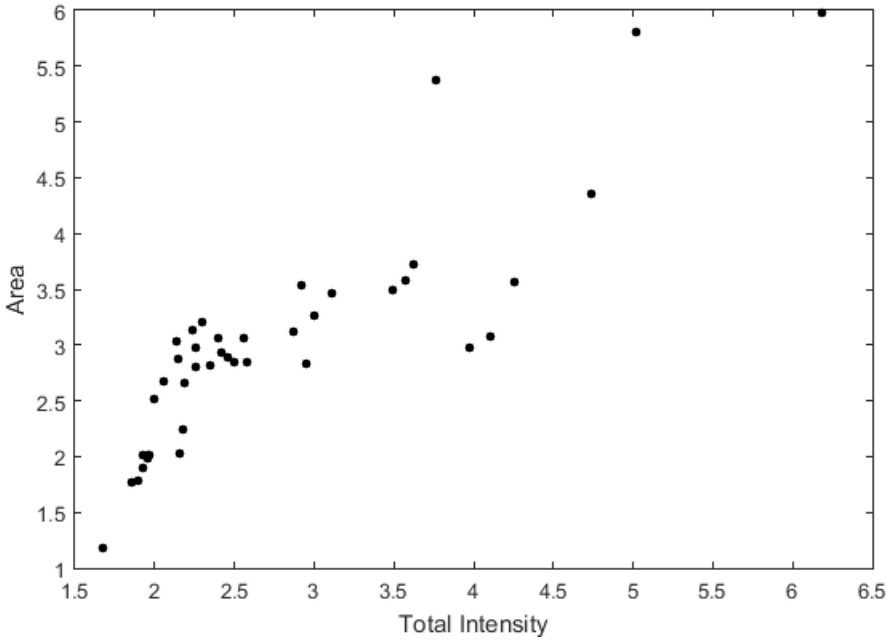


Figure 3.3: Area vs. Total intensity obtained from set of biological material.Each point corresponds to a different nucleus.

algorithm, highly depend on their range. Attributes with large range will give rise to bigger contributions to the metrics, in comparison with attributes with small ranges. Additionally, the attributes should be dimensionless because the numerical values of the ranges of dimensional attributes depend on the units of measurement and, therefore, the choice of the units of measurements may greatly affect on the results of clustering.[71]

In this work two distinct types of normalization were separately applied to the data, namely, z-score normalization and Weighted uncorrected standard deviation (USTD) standardization defined respectively as:

$$x_{ij}^{*} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

$$x_{ij}^{*} = \frac{x_{ij}}{\sigma_j}$$

in which x_{ij}^{*} represents the normalized attribute value, x_{ij} represents the raw data and μ_j and σ_j represent the mean and Standard deviation (STD) for the values of the j^{th} attribute, area and total intensity in this work.

The z-score normalization returns data with mean 0 and standard deviation 1 and contrary, the USTD standardization gives rise to data with standard deviation 1 and transformed mean equal to $\frac{\mu_j}{\sigma_j}$.[72, 73]

For numerical data sets with attributes following a Gaussian distribution, the most commonly employed standardization is the z-score normalization. However, both techniques have been reported as the best ones available, together with the min-max normalization, besides being very identical.[72, 73]

In this work, the use of min-max normalization was not pursued, since this type of normalization does not give rise to equal contributions of variables to the similarity measures and therefore does not achieve equalization of the means of the attributes. Therefore, the comparison between non-

normalized, z-score normalized and USTD standardized data was performed, in order to determine the highest throughput approach.[71]

3.2 Learning the cell phase

Machine learning is a scientific discipline that explores the construction and study the algorithms that can be learn from data [74]. Such algorithms operate by building a model from example inputs and using that to make predictions or decisions [75].

Supervised learning is a widely used framework in Machine Learning (ML). It is the task of deriving a function from training data. This includes the problem of classification of inputs. A typical example in is deciding, whether diagnosed tumor is considered ”benign” or ”malignant”, based on previous observation, which have been marked as such by the doctor.

The learning algorithm is confronted with a training set (e.g. the tumors marked by the doctor), which consists of input/output pairs. This set serves as a sampling of the whole problem space, which the algorithm covers (e.g. all the tumors). The first step is to use pairs of the training set to teach the algorithm some correct mappings of the problem space. This is where the ”supervision” comes from, the learner is told what the right outcome for the function is for an example and is therefore supervised. Afterwards the algorithm can be confronted with unseen inputs/samples for which it determines the outputs/target, solely based on its experience with the training set. The algorithm is said to generalize over the problem space.

Besides the classification of inputs, another typical ML task is regression. Regression comes very close to interpolation, which is the task of finding a function that is exactly correct in the points given by the training set. Regression adds the assumption that the samples of the training set are affected by noise. Thus, the task is to find an approximation for the function, from which the training set was sampled. It turns out that the tasks of classification and regression basically solve the same problem and the solution of one can also be used for the other using an appropriate transformation.

Most of the widely known ML techniques belong to this category, such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Bayesian Statistics, Kernel Estimators and Decision Trees.

Unsupervised Learning. Unlike in supervised learning this category of algorithms is not confronted with input/output pairs, but only with input pairs. The algorithm then finds similarities among the input samples and/or gives them a structure.

A typical example application would be the recommendation system of an online shop. It analyzes which products should be advertised to a customer based on information about products he bought before and products bought by other customers purchasing the same articles beforehand.

It is closely related to density estimation in the field of statistics. Typical approaches are clustering such as with the k-Nearest Neighbor Algorithm, Hidden Markov Models (HMMs) or the self-organization of ANNs.

Reinforcement Learning (RL) is located between supervised and unsupervised learning Like in unsupervised learning, there are no correct outputs, which known apriori would belong to some inputs. The algorithm receives from an environment a numeric value about how well the generated output meets the desired outcome. Through many iterations the algorithm tries many possibilities until it makes a correct prediction about which outputs or decisions will lead to a good overall feedback. Algorithms following this framework are also said to learn with a critic.

A common subproblem is to make correct predictions of the feedback in yet unseen situations. This can be done by generalizing over situations. This subproblem of generalizing over situations is a supervised learning problem. A typical application would be to teach a computer how to play a game. The feedback it receives can be its score or even just the information, whether it won or lost. By repeated playing the actor learns which (sequence) of situations and actions leads to a better score. An actor can learn playing while doing so (online), or be trained beforehand, e.g. by playing against itself.

Solution techniques in this category consist of Dynamic Programming, Monte Carlo methods and Temporal Difference algorithms.

Evolutionary Learning. This category covers methods, which simulate processes in nature that are perceived as ”learning” from experience. Evolution is one such process, that does improve the chance of survival of a species.

This kind of learning is simulated by Genetic Algorithms (GAs). The basic idea is to have a string or array representation for the solution of a problem which serves as the DNA of a solution. Each solution has a certain fitness, which is a real valued number. Its evaluation is similar to the objective function in mathematical optimization problems. In the first step GAs randomly generate a population of solutions of which the fittest are selected for the second step of reproduction, i.e. a new generation of solutions. This second step takes place using the definition of two operators, namely crossover and mutation. Crossover is the generation of new solutions by combining the DNA of two old ones. Mutation is the alteration of the new solutions. In a third step, the fittest new solutions replace the least fit solutions of the old population. This is then considered the next generation. By repeating steps two and three for a number of iterations, the solution space is effectively sampled in many more or less random points. The fittest one is then used. Through the mutation operator, local solutions are sought to be improved, while the crossover operator is based on the assumption that combinations of existing solutions again result in good solutions.

Ant Colony Optimization (ACO) is a second learning approach inspired by biology. ACO models the behavior of ants searching for food. While first wandering around randomly, an ant which has found food leaves a pheromone trail on its way back to the colony. Other ants are likely to follow a path once they found such a trail. While the pheromones on a path evaporate over time, the pheromones on a shorter path will be reinforced more, because the ants return earlier. The path is more often used as a consequence.

Evolutionary and biologically inspired methods in the field of ML are sometimes used as search

heuristics and stochastic optimization techniques. Other techniques, which serve exactly this purpose, but which are not so strongly correlated with machine learning are Tabu Search and Simulated Annealing.

As we can see from the differing categories and applications of ML, the border of what is not clear. Genetic Algorithms as learning methods it could also be seen as a heuristic for mathematical optimization [76, 77]. On the other hand, even the framework of RL, which fits very well an intuitive notion of the term "Machine Learning" can be used as a heuristic for mathematical optimization [78, 79].

Viewing at the four basic categories of ML, supervised, unsupervised, reinforcement, and evolutionary learning, only unsupervised learning is worthwhile for a further investigation on how to employ it on the problem of the cell cycle phase determination directly.

3.3 Cluster Analysis

The main idea of clustering is [80]:

"Given a number of objects or individuals, each of which is described by a set of numerical measures, devise a classification scheme for grouping the objects into a number of classes such that objects within classes are similar in some respect and unlike those from other classes. The number of classes and the characteristics of each class are to be determined."

This definition emphasizes the most important properties of clustering:

- Objects within the formed clusters should be $homogeneous$, while the clusters should be $heterogeneous$. It is necessary, as clusters should help us distinguish between object.
- Number and attributes of clusters should be found out by the algorithm, not given as the input data. So it should not only assign objects to groups, but also determine their structure.

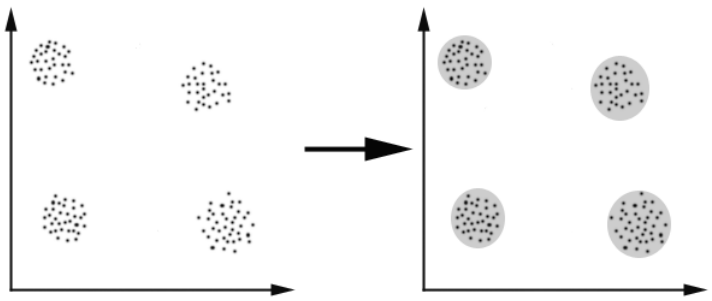


Figure 3.4: Clustering example

Example of clustering of numerical data, consisting of a set of objects described by two variables (so they may be interpreted as points in two-dimensional space) is given on Figure 3.4.

Mathematical Formulation Given a set of n objects $S = \{O_1, O_2, ..., O_n\}$, let $C = \{C_1, C_2..., C_k\}$ be a partition or S, i.e., a set of subsets of S such that $\cup_{i=1}^{k} C_i = S$ and $C_i \cap C_j = \emptyset$ for $\leq i \neq j \leq k$.

Each subset C_i (where $1 \leq i \leq k$) is called a *cluster*, and C is called a *clustering result*.

A dataset containing objects to be clustered is usually represented in one of two formats: the *data matrix* and the *similarity* (or *dissimilarity*) *matrix*. In a data matrix, the rows usually represent objects, and the columns usually represent features or attributes of the objects. Suppose, there are n objects and p attributes. We assume the rows represent total intensity of nuclei and the columns represent area of the nuclei, such that entry (i,e) in the data matrix D represents the total intensity of the nuclei i under area e, where $1 \leq i \leq n$ and $1 \leq e \leq p$. The ith row in the data matrix D (where $1 \leq i \leq n$), D_i, represents the expression vector of total intensity of the nuclei i across all p areas. The similarity (or dissimilarity) matrix contains the pairwise similarity (or dissimilarity) of the area and total intensity of the nuclei. Specifically, entry (i, j) in the similarity (or dissimilarity) matrix Sim represents the similarity (or dissimilarity) of total intensity of the nuclei i and area of the nuclei j.

Similarity metrics The measure used to compute similarity (or dissimilarity) between a pair of objects is called a *similarity metric*. Many different similarity metrics have been used in clustering. The two most popular similarity metrics are correlation coefficient and Euclidean distance. Correlation coefficient is a similarity measure (a high correlation coefficient implies high similarity) while Euclidean distance is a dissimilarity measure (a high Euclidean distance implies low similarity).

3.3.1 Clustering Algorithms

Clustering algorithms could be divided into two groups that follow two fundamentally different strategies [81].

1. *Hierarchical* or agglomerative algorithms start with each point in its own cluster. Clusters are combined based on their ”closeness”, using one of many possible definitions of ”close”. Combination stops when further combination leads to clusters that are undesirable for one of several reasons.

2. The second group of algorithms involve point assignment. Points are considered in some order, and each one is assigned that cluster in which it best fits. This process is normally preceded by a short stage in which initial clusters are estimated. Variations allows occasional combining or splitting of clusters, or may allow points to be unassigned if they outlayers (points which are too far from any of the current cluster).

3.3.2 Hierarchical clustering

This algorithm can only be used for relatively small datasets, but even so, there are some efficiencies by careful implementation. For this clustering algorithm Euclidean space it is desirable and allows to represent a cluster by its centroid or average of the point in the cluster.

WHILE it is not time to stop, pick the best two clusters to merge;

    Combine those two clusters into one cluster

END

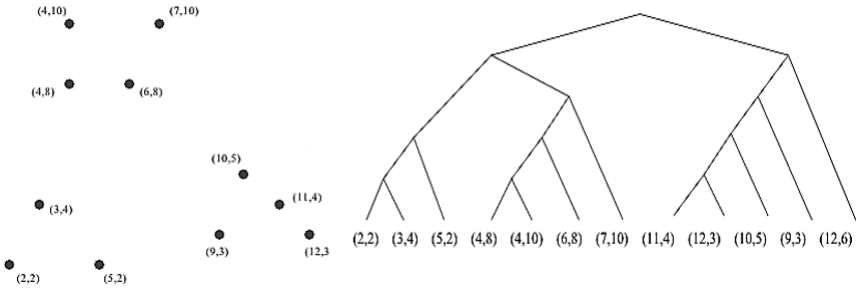


Figure 3.5: Tree showing the example of complete grouping of the points

3.3.3 Point-assignment algorithms

In contrast to hierarchical clustering this group of algorithms have not restrictions for input data set. In case of DAPI images it is very useful, because dataset could be different depending on image and it size could vary. Also this algorithms commonly used for Gaussian distribution separation.

K-means is one of classical methods in the cluster analysis [82, 83]. Suppose that given dataset $\{x_1, x_2, ..., x_n\}$ consisting of N observations of a random D-dimensional Euclidean variable x. Clusters could be a group of data point whose inter-point distance are small compared with the distances to points outside of the cluster. In this case we can use D-dimensional vectors μ_k, where $k = 1..N$, in which μ_k is a prototype associated with the k^{th} cluster. The centers of the clusters are represented by μ_k. The main idea is to partition the data points into some number K (given value) of clusters. For each data point x_n there is corresponding set of binary indicator variables r_{nk}, where $k = 1...K$ describing which of the K clusters the data point x_n. An objective function, sometimes call a distortion measure, given by:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{3.1}$$

Which represents the sum of the squares of the distances of each data point its assigned vector μ_k. The goal of K-means algorithm is to find values for the $\{r_{nk}\}$ and the $\{\mu_k\}$ so as minimize J. That could be improved by iteration process:

1. Initial values for the μ_k are chosen;
2. Minimize J with respect to the r_{nk}, keeping μ_k fixed;
3. Minimize J with respect to the μ_k, keeping the r_{nk} fixed.

Because J in 3.1 is a linear function of r_{nk}, this iteration process optimization can be performed easily to give a closed form solution. The terms involving different n are independent that is why we can optimize n separately by choosing r_{nk} to be one for whichever value of k gives the minimum value of $||x_n - \mu_k||$, this can be expressed as:

$$\begin{cases} 1, & \text{if } k = argmin_j ||x_n - \mu_k|| \\ 0, & \text{otherwise} \end{cases}$$

In third step optimization of the μ_k with the r_{nk} held fixed is produced. The objective function J is a quadratic function of μ_k and can easily be solved:

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

The denominator in this expression is equal to the number of points assigned to cluster k, and so this result has a simple interpretation, namely μ_k equal to the mean of all of the data points x_n assigned to the cluster k. The two steps (re-assigning data points to clusters and re-computing the cluster mean) are repeated in turn until there is no further change in the assignment[75].

3.4 Model-Based Clustering Approach

Clustering algorithms based on probability models offer a principled alternative to heuristic-based algorithms. In particular, the model-based approach assumes that the data is generated by a finite mixture of underlying probability distributions such as multivariate normal distributions. The Gaussian mixture models has been shown to be a powerful tool for many applications [84–86]. With the underlying probability model, the problems of determining the number of clusters and of choosing an appropriate clustering method become statistical model choice problem [87, 88].

3.4.1 The model-based framework

The mixture model assumes that each component (group) of the data is generated by underlying probability distribution. Suppose the data **y** consists of independent multivariate observation $y_1, y_2, ...y_n$. Let *G* be the number of components in the data. The likelihood for the mixture model is

$$L_{MIX}(\theta_1, ..., \theta_G|y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(y_i|\theta_k), \tag{3.2}$$

where f_k and θ_k are the density and parameters of the *k*-th component in the mixture, and τ_k is the probability that an observation belongs to the *k*th component ($\tau_k \geq 0$ and $\sum_{k=1}^{G} \tau_k = 1$).

In the Gassuian mixture model, each component *k* is modeled by the multivariate normal distribution with parameters μ_k (mean vector) and Σ_k (covariance matrix):

$$f_k(y_i|\mu_k, \Sigma_k) = \frac{exp\{-\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\}}{\sqrt{det(2\pi\Sigma_k)}}. \tag{3.3}$$

Geometric features (shape, volume, orientation) of each component *k* are determined by the covariance matrix Σ_k. In [84] proposed a general framework for exploiting the representation of the covariance matrix in terms of its eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k^T, \tag{3.4}$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalue of Σ_k, and λ_k is a scalar. The matrix D_k determines the orientation of the component, A_k determines its shape, and λ_k determines its volume.

Allowing some but not all of the parameters in equation 3.4 to vary results in a set of models within this general framework that is sufficiently flexible to accommodate data with widely varying characteristics. Several common models are outlined below. Constraining $D_k A_k D_k^T$ to be the identity matrix *I* corresponds to Gaussian mixtures in which each component is spherically symmetric. The equal volume spherical model (denoted *EI*), which is parameterized by $\Sigma_k = \lambda I$, represents the most constrained model under this framework, with the smallest number of parameters. The unequal volume spherical model (*VI*), $\Sigma_k = \lambda_k I$, allows the spherical components to have different volumes, determined by a different λ_k for each component *k*. The unconstrained model(*VVV*) allows all D_k, A_k and λ_k to vary between components. The unconstrained model has the advantage that it is the most general mode, but has the disadvantage that the maximum number of parameters need to be estimated, requiring relatively more data points in each component. There are a range of elliptical models with other constraints and fewer parameters. For example, with the parameterization $\Sigma_k = \lambda D A D^T$, each component is elliptical, but all have equal volume, shape and orientation (denoted *IEEE*) [85, 88]. Also considered the model in which $\Sigma_k = \lambda_k B_k$, where B_k is a diagonal matrix with $|B_k| = 1$. Geometrically, the diagonal model corresponds to axis-aligned elliptical components.

Σ_k	Distribution	Volume	Shape	Orientation	Reference
λI	Spherical	equal	equal	NA	[84, 89–91]
$\lambda_k I$	Spherical	variable	equal	NA	[84, 91]
$\lambda D A D^T$	Ellipsoidal	equal	equal	equal	[84, 91–93]
$\lambda_k D_k A_k D_k^T$	Ellipsoidal	variable	variable	variable	[84, 91, 93]
$\lambda D_k A D_k^T$	Ellipsoidal	equal	equal	variable	[84, 90, 91]
$\lambda_k D_k A D_k^T$	Ellipsoidal	equal	equal	variable	[84, 91]

Table 3.1: Parameterizations of the covariance matrix Σ_k in the Gaussian model and their geometric interpretation.

The diagonal model implementation, the desired number of clusters *G* is specified, and then the model parameters (τ_k, μ_k and Σ_k appropriately constrained, for $1 \leq k \leq G$) are estimated by the Expectation Maximization algorithm (EM) algorithm (see Section 3.4.2). In the EM algorithm, the Expectation steps and Maximization steps alternate. In the E-step, the probability of each observation belonging to each cluster is estimated conditionally on the current parameter estimates. In the M-step, the model parameters are estimated given the current group membership probabilities. When the EM algorithm converges, each observation is assigned to the group with the maximum conditional probability. In the clustering context, the EM algorithm for mixture models is usually initialized with a model-based hierarchical clustering or with k-means (see Section 3.3.3).

3.4.2 Expectation Maximization Algorithm

Iterative relocation methods for clustering via mixture models are possible through EM and related techniques [86]. The EM algorithm [94, 95] is a general approach to maximum likelihood in the presence of incomplete data. In EM for clustering, the "complete" data are considered to be $y_i =$

(x_i, z_i), where $z_i = (z_{i1}, ..., z_{iG})$ with

$$z_{ik} = \begin{cases} 1 & \text{if } x_i \text{ belongs to group k} \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

constitutes the "missing" data. The relevant assumptions are that the density of an observation x_i given z_i is given by $\prod_{k=1}^{G} f_k(x_i|\theta_k)^{z_{ik}}$ and that each x_i is independent and identically distributed according to a multinomial distribution of one draw on *G* categories with probabilities $\tau_1, ..., \tau_G$. The resulting complete-data loglikelihood is

$$l(\theta_k, \tau_k, z_{ik}|x) = \sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik}[log\tau_k f_k(x_i|\theta_k)]. \tag{3.6}$$

The quality $\widehat{z}_{ik} = E[z_{ik}|x_i, \theta_1, ..., \theta_G]$ for the model (3.6) is the conditional expectation of z_{ik} given the observation x_i and parameter value. The value z^*_{ik} of $\widehat{z}_{ik}$ at a maximum of 3.2 is the conditional probability that observation *i* belongs to group *k*; the classification of an observation x_i is taken to be $\{j|z^*_{ij} = max_k z^*_{ik}\}$.

The EM algorithm iterates between an E-step in which values of $\widehat{z}_{ik}$ are computed from the data with the current parameter estimates, and an M-step in which the complete-data loglikelihood (3.6), with each z_{ik} replaced by its current conditional expectation $\widehat{z}_{ik}$, is maximized with respect to the parameters (see Figure 3.6). Under certain conditions [95–97], the method can be shown to converge to a local maximum of the mixture likelihood (3.2). Although the conditions under which convergence has been proven do not always hold in practice, the method is widely used in the mixture modeling context with good results. Moreover, for each observation *i*, $(1 - max_k z^*_{ik})$ is a measure of uncertainty in the associated classification [98].

Initialize $\widehat{z}_{ik}$ (this can be from a discrete classification 3.5)
repeat
M-step: maximize 3.6 given $\widehat{z}_{ik}$ (f_k as in 3.3)

$$n_k \leftarrow \sum_{i=1}^{n} \widehat{z}_{ik}$$

$$\widehat{\tau}_k \leftarrow \frac{n_k}{k}$$

$$\widehat{\mu}_k \leftarrow \frac{\sum_{i=1}^{n} \widehat{z}_{ik} x_i}{n_k}$$

$\widehat{\Sigma}_k$: depends on the model (see [85])

E-step: compute $\widehat{z}_{ik}$ given the parameter estimates from the M-step

$$\widehat{z}_{ik} \leftarrow \frac{\widehat{r}_k f_k(x_i|\widehat{\mu}_k, \Sigma_k)}{\sum_{j=1}^{G} \widehat{r}_j f_j(x_i|\widehat{\mu}_j, \Sigma_j)}, \quad \text{where } f_k \text{ has the form 3.3}$$

until convergence criteria are satisfied

Figure 3.6: EM algorithm for clustering via Gaussian mixture models.

EM algorithm is sensitive to input data, because in each step the inverse matrix should be calculated, for that reason singular value decomposition should be used. Moreover, the parameters correspond to cluster shape, volume and orientation are dependent from obtained covariance matrix. To determine adequate values, eigenvectors should be obtained, which could be done through powerful tool singular value decomposition.

3.4.3 Singular Value Decomposition

The Singular Value Decomposition (SVD) is a widely used technique to decompose a matrix into several component matrices, exposing many of the useful and interesting properties of the original matrix. The decomposition of a matrix is often called a factorization. Ideally, the matrix is decomposed into a set of factors (often orthogonal or independent) that are optimal based on some criterion. The decomposition of a matrix is also useful when the matrix is not of full rank. That is, the rows or columns of the matrix are linearly dependent. Theoretically, one can use Gaussian elimination to reduce the matrix to row echelon form and then count the number of nonzero rows to determine the rank. However, this approach is not practical when working in finite precision arithmetic. A similar case presents itself when using LU decomposition where L is in lower triangular form with 1's on the diagonal and U is in upper triangular form. Ideally, a rank-deficient matrix may be decomposed into a smaller number of factors than the original matrix and still preserve all of the information in the matrix. The SVD, in general, represents an expansion of the original data in a coordinate system where the covariance matrix is diagonal.

The most general case is to provide a decomposition for a rectangular matrix. It is possible to decompose a matrix that is not square nor symmetric by first considering a matrix A that is of dimension $m \times n$ where $m \geq n$. This assumption is made for convenience only; all the results will also hold if $m \leq n$. As it turns out, the vectors in the expansion of A are eigenvectors of the square matrices AA^T and A^TA. The former is a outer product and results in a matrix that is spanned by the row space A. The latter is a inner product in a matrix that is spanned by the column space of A.

The *singular values* are the nonzero square roots of the eigenvalues from AA^T and A^TA. The eigenvectors of AA^T are called the "left" singular vectors (U) while the eigenvectors of A^TA are the "right" singular vectors (V). By retaining the nonzero eigenvalues $k = min(m, n)$ a SVD can be constructed. That is

$$A = U\Lambda V^T, \tag{3.7}$$

where U is an $m \times n$ orthogonal matrix ($U^TU = I$), V is an $n \times n$ orthogonal matrix ($V^TV = I$), and Λ is an $m \times n$ matrix whose off-diagonal entries are all $0'$s and whose diagonal elements satisfy

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0.$$

It can be shown that the rank of A equals the number of nonzero singular values and that the magnitudes of the nonzero singular values provide a measure of how close A is to a matrix of lower rank. That is, if A is nearly rank deficient (singular), then the singular values will be small. In general, the SVD represents an expansion of the original data A in a coordinate system where the covariance matrix Σ_A is diagonal. This is called SVD because the factorization finds values or eigenvalues or characteristic roots (all the same) that make the the following characteristic equation true or singular. That is

$$|A - \lambda I| = 0.$$

Using the determinant this way helps solve the linear system of equations thus generating an *n*-th

degree polynomial in the variable λ This polynomial, that yields *n*-roots, is called the characteristic polynomial.

A set of linear algebraic equations can be written as

$$Ax = B, \tag{3.8}$$

where A is a matrix of coefficients $(m \times n)$, and B $(m \times 1)$ is some form of a system output vector. The vector x is what we usually solve for. Considering (3.7) and from property of orthogonal matrices, that inverse to it is a Hermitian conjugate:

$$\Lambda Y = C, \tag{3.9}$$

where $Y = V^T x$ and $C = U^T B$. From equation (3.9) all components of *Y* can be found according:

$$y_i = \frac{c_i}{\lambda_i}, \;\; i = 1, .., n,$$

where *n* is System of Linear Algebraic Equations (SLAE) dimension. Then, equation (3.8) has only one solution if $\lambda_i \neq 0$; has infinite number of solutions if $c_i \neq 0$; and can not be solved if $\lambda = 0$. For any case can be put conditions, that residual rate $\|R = Ax - B\|$ should be minimal. That conditions can be described as:

$$y_i = \begin{cases} y_i = \frac{c_i}{\lambda_i} & \lambda_i > \varepsilon \\ 0 & \lambda_i \leq \varepsilon \end{cases}, \tag{3.10}$$

where ε is accuracy. In other words the solution of the (3.8) according (3.10) can be found as:

$$x = VY$$

4

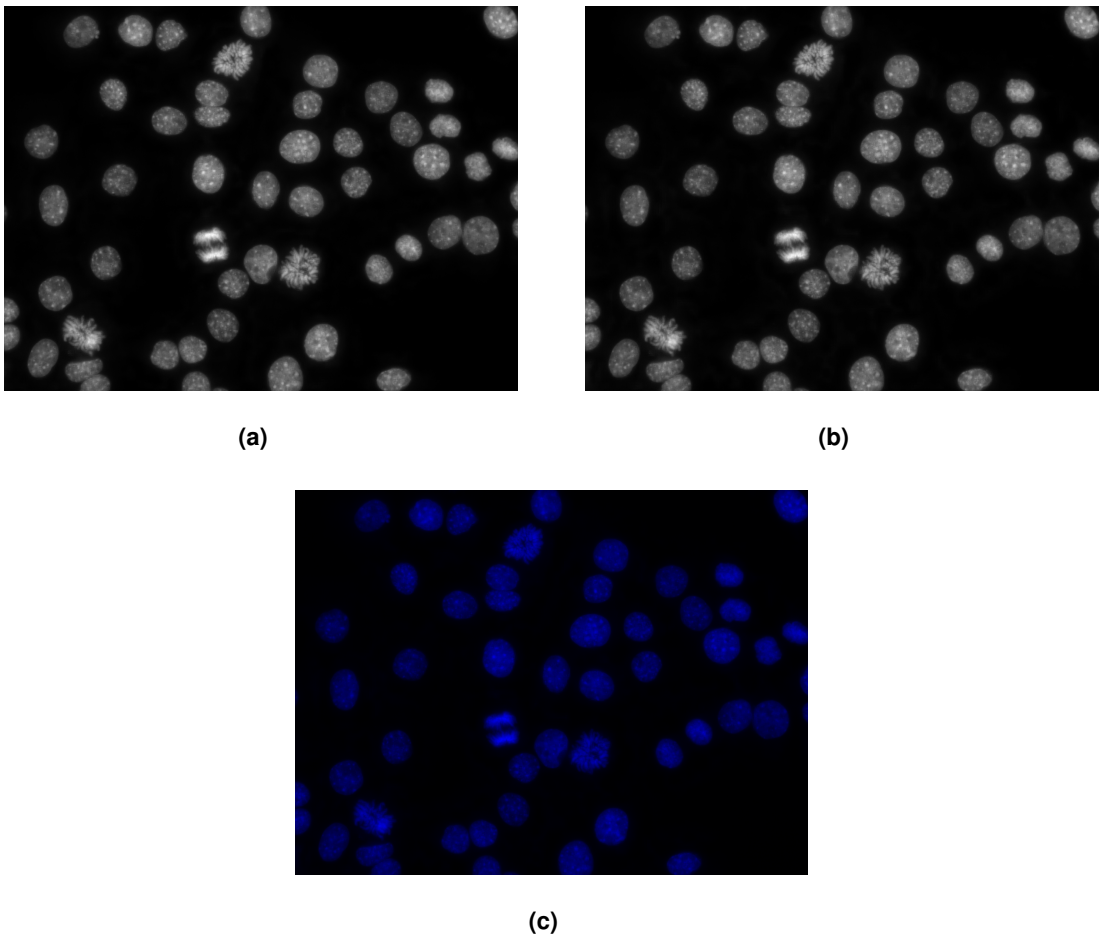
Results

Contents

In this Section all obtained results are shown. First, some of acquired and treated fluorescence images are shown. Then, the clustering analysis is performed upon the extracted features of the analyzed FM is presented. Validation of the bioimaging analysis was performed with FUCCI system.

4.1 Image Processing

Grayscale images acquired with FM are in high resolution and can be well treated with thresholded methods (see Section 3.1). In current work two kind of images were used: with 40 and 60 stacks, which after were computed in the a single DAPI-plane image. Experimentally, was determined that the greater number of stacks allows to consider all structure of the cell, it means that it is possible more larger describe processes inside the cells. In Figure 4.1 is showed fluorescence images in which cells were stained with DAPI.

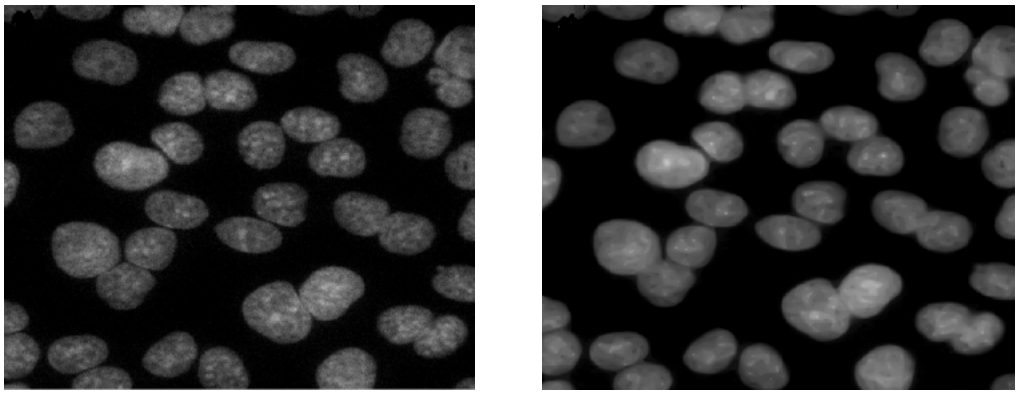


(a) **(b)**

(c)

Figure 4.1: Set of FM images. (a) - original DAPI-plane image obtained from 40 stacks; (b) - original DAPI-plane image obtained from 60 stacks; (c) - DAPI-plane image obtained from 60 stacks, considering prior knowledge about DAPI dye.

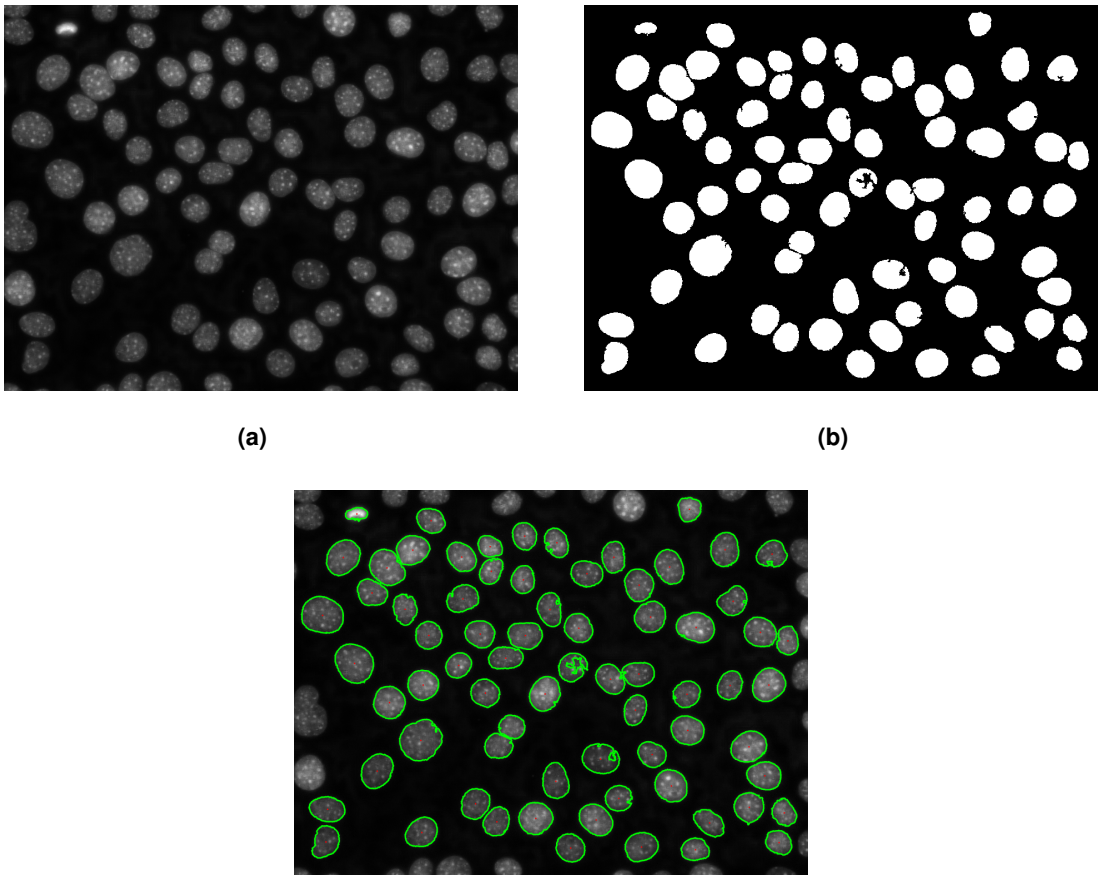
The output of FM is digital, it means it contains a certain level of noise. Removing noise in image is necessary, because such pre-processing converts all objects from heterogeneous to homogeneous ones. In Figure 4.2 is shown denoising results. Based on denoised images main features are estimated according to the segmentation algorithms, which are described in Section 3.1. Each cell has a



(a) (b)

Figure 4.2: Illustration of the denoising procedure for the DAPI plane image. (a) - Original image zoomed;(b) - Denoised DAPI plane image zoomed.

certain mask by its unique structure. In Figure 4.3 is shown the masks of each cell and boundaries of its structure.



(a) (b)

(c)

Figure 4.3: Illustration of the segmentation procedure for the DAPI plane image. (a) - Original image;(b) - Mask of each cell from original image;(c) - Obtained boundaries of each cell, based on mask.

It should be noticed, that cells on edges of the image were not taken account because of incom-

plete information extractable from the DAPI-stained nuclei.

4.2 Cluster analysis

After acquiring all data from the fluorescent images, these were subjected to clustering algorithms analysis. This work is based on the assumption that processes inside cells, occurring during the cell cycle, can be described as stochastic processes. Consequentially, the Gaussian Mixture Models (GMM) strategy was chosen, specifically, EM algorithm which calculates the probability that a given nuclei falls to a certain cluster. Additionally, based on the biological knowledge it was determined the number of clusters, and initial conditions for clustering algorithms. The number of clusters established herein correspond to the number of active phases: *G1,S* and *G2*. Initial conditions are: exist dependence between phase and intensity of the cell during the cell cycle ,and separation between clusters should be described by nonlinear function. All requirements were implemented in k-means algorithm, which was first iteration of EM algorithm. Figure 4.4 shows the clustering results.

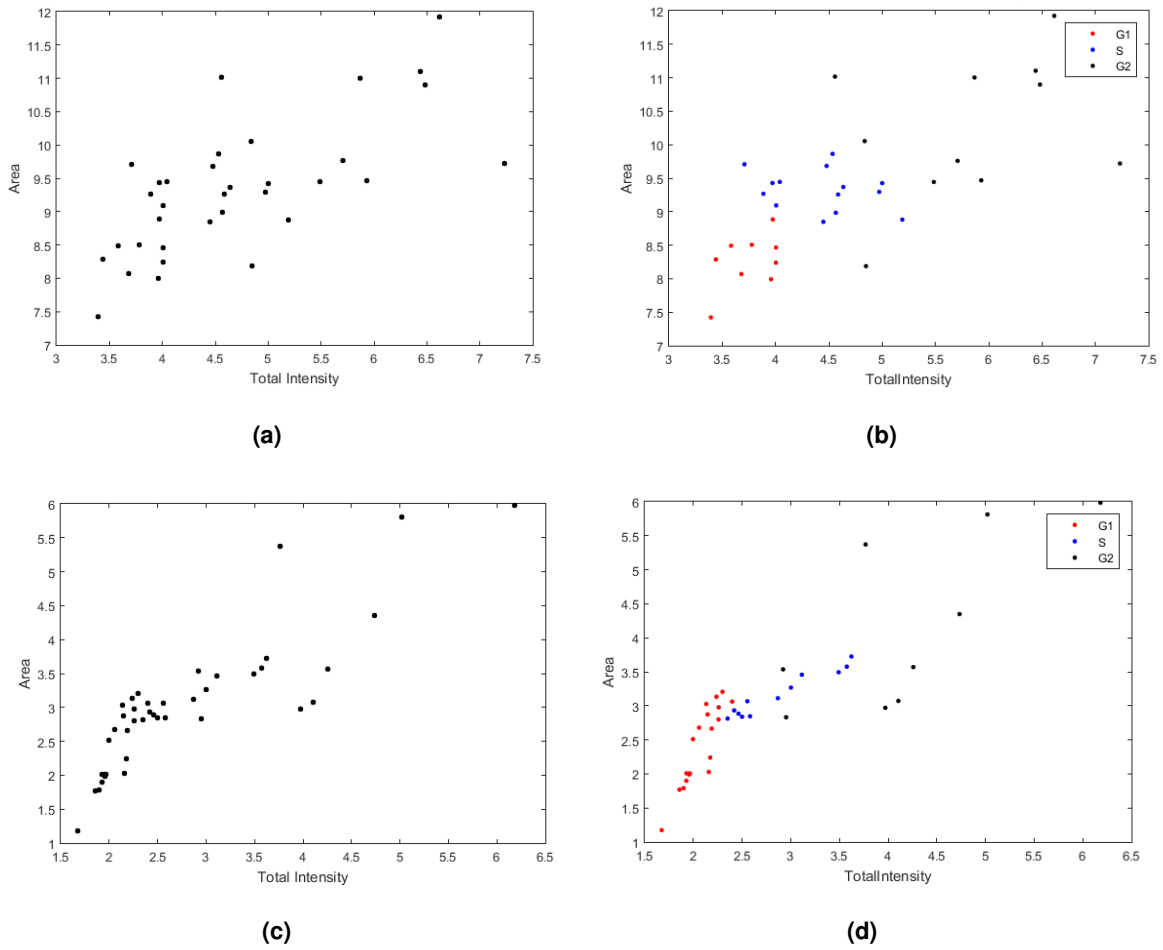


Figure 4.4: Illustration of the clustering procedure for some of the DAPI plane images from Fig.4.3. (a),(c) - Input data, obtained from image; (b),(d) - Clustering results.

It should be noticed that the data was normalized (each data point was divided by the standard deviation of all dataset) before clustering algorithms were applied. The normalization process is

necessary as it helps to scale data, thus leading to better clustering results.

Since distribution between clusters were determined, the algorithm should be compared with already existing methods. For verification of the clustering results it was used the FUCCI system, which is a powerful tool used in cell biology that allows the colored readout of the cell cycle progression. In Figure 4.5 are shown representative images acquired with FUCCI and the corresponding DAPI-plane images.

A total of 47 fluorescence images comprising 998 DAPI-stained nuclei were analyzed and compared with both methods. The analysis of bioimaging tool performance revealed that the cell cycle distributions using FUCCI (G1 - 53%, G1/S - 13% and S/G2 - 33%) and the new analysis tool (G1 - 62%, S - 29% and G2 - 9%) are similar. The difference observed might be due to the variability inherent to the cell cultures used during analysis. The segmentation algorithm revealed a high efficiency and throughput ability. The accuracy was calculated by using the sensetivity and specificity parameters by using the mathematical formulas described below:

$$Sensitivity = \frac{TP}{TP + FN} Specificity = \frac{TN}{TN + FP} Accuracy = \frac{TP + TN}{TN + FP + TN + FN},$$

where TN — the number of true negative; TP — the number of true positive; FN — number of false negative; FP — number of false positive.

The achieved sensitivity, specificity and mean accuracy for the analazed datase was 96%, 93% and 94.5%, respectively. Moreover, the average time of analysis for one FM image is approximately ten seconds, with 97% rate of nuclei classification. This new bioimaging approach, based on the area and total intensity retrieved from DAPI stained nuclei is a highly accurate tool for cell cycle analysis. The developed methodology can be further explored and be used in future works.

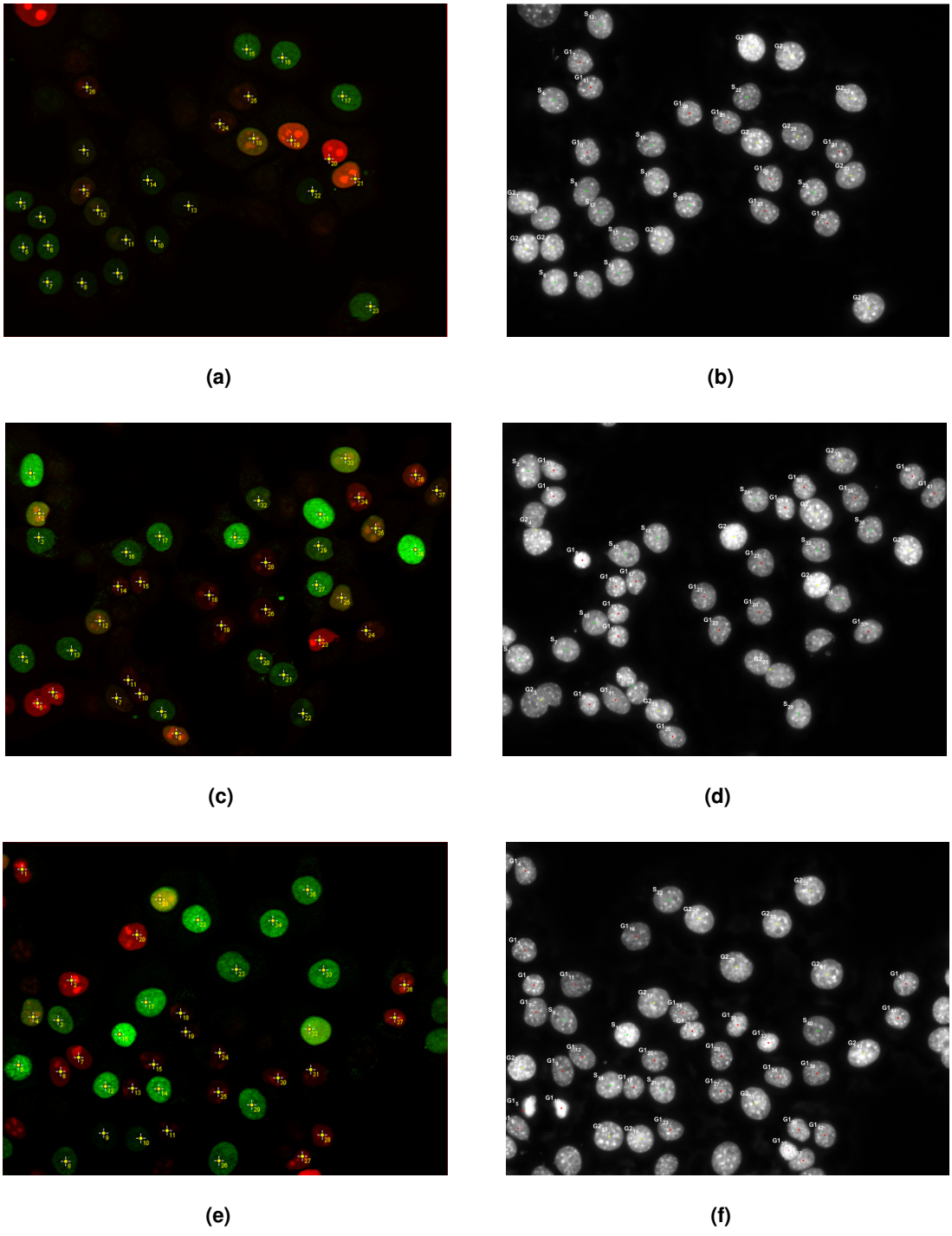


Figure 4.5: Several images which were used for verification process. (a),(c), (e) - Images obtained with FUCCI, where red color correspond *G1* phase, yellow to *G1 - S* transitioning cells, and green correspond *S/G2/M* phase; (b), (d), (f) - Classified DAPI-plane image according to the clustering results.

5

Discussion

In this chapter all obtained results of cell cycle staging are discussed. Based on FM images acquired *in vivo* cell cultures, was proved, that total intensity and area of the DAPI-stained nuclei, strongly correlate with the cell cycle phase. To determine such dependence two-level approach was designed. In the first level, an image-processing pipeline was applied to extract specific features from each DAPI-stained nuclei — area and total intensity. In the second level data mining analysis was implemented, specifically, a new clustering algorithm was developed, which comprises a mix of unsupervised and supervised machine learning algorithms.

The average accuracy achieved for developed methodology is 94.5%. The obtained accuracy is based on verification process between results of the designed algorithm and FUCCI. Currently, other imaging-based methods for assessing the cell cycle status of individual cells include metabolic labeling procedures that probe cells transversing S-phase [12], staining methods that use specific cell cycle markers such as cyclics, proliferating cell nuclear antigen (PCNA) or Cdt1 [13]. Others are very laborious as they involve the generation of cellular systems that stably express various cell cycle phase specific reporters [14, 62], as FUCCI that was used as validator. Contrarily to the designed algorithm, these methods are able to probe only specific stages of the cell cycle. For example, immunolabeling after incorporation of modified DNA precursors such as the nucleotide analog BrdU, chlorodexyuridine (CldU), allows the precise detection of S-phase cells only. Similarly, S-phase cells can also be identified by high expression of DNA polymerase of PCNA [99], whereas mitotic cells can be detected by immunostaining with the mitotic marker phosphor-histone H3 [100]. Cell cycle reporters can be integrated to generate stable cell lines or transgenic animals expressing cell cycle-specific fluorescence markers. The widely used FUCCI system, based on the expression of the cell cycle oscillators Cdt1 and germinin tagged with different fluorescence proteins, marks cells in G1 or S/G2/M phases, respectively [58], whereas visualization of cell division can be achieved by the marked changes in cell morphology. However, these are not universal system as they are specific for the cellular systems chosen, or need a combinatorial use of the available tools as none of these immunostaining or reporter-based methods is sufficient to determine the cell cycle stages of all individual cells in a population. This approach is technically complex and requires the use of multiple spectral imaging channels, reducing the capability for concomitant visualization of other cellular features, such as GFP-labeled cellular structure.

Contrarily, the proposed strategy of this work uses a single and inexpensive compound, the commonly used DNA binding dye DAPI, to quantify the cellular DNA content by imaging. In order to obtain relevant information from fluorescence images, an automatic segmentation procedure was designed. Segmentation is a complex procedure, which consists of several steps, particularly, denoising and thresholding. In this work, denoising procedure is based on Poisson statistics, which adequate describes image, corrupted by noise. Denoising step is crusial as it converts nuclei into homogeneous objects. Additionally, having homogeneous objects significantly increases the thresholding throughput. For this work, the thresholding was always performed by separating the background color from others, particularly, all other colors except background, related to objects (see chapter 3.1.2.A). The final step of image processing pipeline is feature extraction (see chapter 3.1). In this work, two fea-

tures were used, specifically, area and total intensity of the nuclei stained by DAPI. Both features are intrinsically related to changes that occur in the nucleus during cell cycle progression, namely, the growth of the nucleus in G1 and G2 phases, and the DNA replication during the S phase. Additionally, color versus total intensity, on the feature plane has three regions: the first corresponds to significant rise of areas and total intensities; the second region, which correspond to DNA replication, shows almost constant areas values, but increased total intensities values ranging from 2N to 4N; and the third region, with approximately two-folded total intensity of the first region, and high values of areas. After extracting the features derived from DAPI-stained nucleus, the clustering algorithm has been developed. The main requirement for clustering algorithm was to implement a blind strategy, meaning that the true distribution of cells in the cell cycle phases was unknown. Nowadays, is possible to perform such analysis through unsupervised machine learning algorithms, specifically, cluster analysis. In the interest of data representation, the points assignment clustering approach was chosen and each of nucleus corresponds to one point on the feature plane. To specify clustering approach, the biological background of cell cycle was strongly used. Particularly, the events occurring during the eukaryotic cell cycle are a stochastic processes, consequently, all changes could be described with normal (Gaussian) distribution. Considering three active phases: G1, S and G2, the number of Gaussians is equal to three, hence, the strategy of Gaussian mixture models was chosen. In terms of cluster analysis, each Gaussian distribution correspond to one cluster in the feature plane. Such approach could be implemented by modeling covariance matrix of each cluster. Since, a cluster is a group of points it has a several parameters, namely volume, shape and orientation. On that basis, several models for covariance matrix can be defined. In this work, the most general model (VVV) was used; for an extensive description of the parameters used, specifically, the covariance matrix of each cluster, refer to chapter 3.4.1. In the VVV model, the covariance matrix has full rank, hence obtained shape is an ellipsoid. The volume of the cluster strongly depends of the points number and is determined by eigenvectors. For the implementation of the defined model, the EM was chosen. The central idea of the latter, is the estimation of the probability of a belonging point to a certain cluster. In this work, the EM algorithm was applied as an unsupervised classification method, but due to high dependence of the input data, the initial conditions should be defined. To solve the sensitivity problem, an extra classical unsupervised strategy was adopted. An example of this type of analytical strategy is the k-means algorithm, where points are assigned to different clusters depending to their distance to that given cluster. Several analytical assumptions were made considering the biological background. Particularly, if the amount of DNA is increasing (doubled) during S-phase, the mean total intensity of the third cluster, that corresponds to G2-phase, would be two-fold the total intensity first cluster, i.e., the G1-phase can be described as following: $\mu_{G2} = 2\mu_{G1}$. Furthermore, there is no clear transition between cell cycle phases, hence, decision boundary function should be non linear. Mathematically, it is represented as the squared Euclidean distance, defined as: $d = \sum_{i=1}^{K} \sum_{j=1}^{N} ||x_j - c_i||^2$, where d — distance between data point and the centroid.

Finally, a fully automatic, unsupervised classification algorithm was designed. This method has been validated by observing the expected enrichment of cell populations in specific cell cycle stages

and by direct comparison with the classification obtained with FUCCI system. The validation process was made manually, and the developed method achieved an average accuracy of 94%, which denotes the good ability of the new algorithm to correctly classify the cell cycle phases of cells, based on DAPI staining.

However, algorithm relies fundamentally on the accurate segmentation of nuclei for DNA content quantification. Therefore, deficient nuclei segmentation hinders the classification method. The discrepancies observed between the new bioimaging tool and FUCCI can be explained with the biological background. The majority of misclassified nuclei are in the transitioning phases between G1 and S phase. Here, cells are slightly bigger than ”pure” G1 cells, but the increase in size and the DNA duplication are not striking enough that allows a clear S-phase classification; other misclassified points correspond to abnormal cells that randomly appear *in vivo* cultures. Moreover, the classification of early stages of mitosis (G2/M population) needs improvement, as these cells are counter intuitively classified as G1. So, if the assessment of the mitotic population is needed, the analysis of DNA content based on DAPI-stained nuclei must be further combined with other morphological features or mitosis-specific markers is needed.

6

Conclusions

Studying and analyzing the cell cycle is an urgent task, because of importance understand main processes during the cell cycle, hence, there is possibility to control them. Flow cytometry is commonly used method to determine cell cycle phases of biological samples, but uit requires specific and expensive equipment.

The aim of this master thesis was to develop a new approach for the determination of the cell cycle phases, based only on fluorescence microscopy and on the nuclear stain, DAPI dye. Such new approach was chosen because DAPI binds stoichiometrically to DNA, thus the intensity of a blue color in fluorescent images correspond to amount of the DAPI in the nuclei. The fact that DAPI is an inexpensive and a universally used nuclear stain turns this new bioimaging tool extensively accessible for the cell cycle analysis

All fluorescence images were subjected to denoise and segmentation algorithms meaning that means that a pattern was found for each cell analyzed, as well as the estimation of the *area* and *total intensity* for each object. Since true distribution cells under phases was unknown, the machine learning approach was chosen, specifically - cluster analysis. Moreover, the initial conditions were obtained from the known biological background of the cell cycle, and expectation-maximization algorithm was applied. All the results obtained were compared with FUCCI system. Additionally, the defined algorithm, under initial conditions was compared with the common clustering analysis algorithm — k-means. The chosen clustering-based strategy showed the best results, with an accuracy of 94.5%. There are still some classification discrepancies that need to be tackle, either by improving algorithm itself or by using combined methods of validation(e.g., FUCCI system along with specific markers for the appropriate cell cycle phase).

Finally, developed tool is based solely on the quantitative analysis of DAPI-stained fluorescent images and that allows the cell cycle staging of individual cells, as well as the generation of population-based cell cycle profiles, while preserving cell's natural architecture.

Bibliography

[1] J. W. Rice, D. A. Warner, C. D. Kelly, M. P. Clough, and J. T. Colbert, "The theory of evolution is not an explanation for the origin of life," *Evolution: Education and Outreach*, vol. 3, no. 2, pp. 141–142, 2010.

[2] V. Roukos, G. Pegoraro, T. C. Voss, and T. Misteli, "Cell cycle staging of individual cells by fluorescence microscopy," *Nature protocols*, vol. 10, no. 2, pp. 334–348, 2015.

[3] M. . with Cody at University of Dallas, "Cell structure," https://www.studyblue.com/notes/note/n/ch-4-cell-structure--function/deck/1127855.

[4] ViralZone, "Modulation of host cell cycle by viral cyclin-like protein," http://http://viralzone.expasy.org/all_by_protein/879.html.

[5] T. C. for DNA Fingerprinting and Diagnostics, "Cell cycle regulation," http://www.cdfd.org.in/labpages/cellcycle.html.

[6] U. of Leicester, "The cell cycle, mitosis and meiosis," http://www2.le.ac.uk/departments/genetics/vgec/schoolscolleges/topics/cellcycle-mitosis-meiosis.

[7] F. of Medicine and C. Density, University of Alberta, "Flow cytometry," http://http://flowcytometry.med.ualberta.ca/.

[8] M. Malumbres and M. Barbacid, "Cell cycle, cdks and cancer: a changing paradigm," *Nature Reviews Cancer*, vol. 9, no. 3, pp. 153–166, 2009.

[9] P. Nurse, "A long twentieth century of the cell cycle and beyond," *Cell*, vol. 100, no. 1, pp. 71–78, 2000.

[10] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

[11] W. Sterzel, P. Bedford, and G. Eisenbrand, "Automated determination of dna using the fluorochrome hoechst 33258," *Analytical biochemistry*, vol. 147, no. 2, pp. 462–467, 1985.

[12] D. Jackson and P. R. Cook, "Analyzing dna replication i: Labeling animals, tissues, and cells with bromodeoxyuridine (brdu)," *Cold Spring Harbor Protocols*, vol. 2008, no. 8, pp. pdb–prot5031, 2008.

[13] H. Leonhardt, H.-P. Rahn, P. Weinzierl, A. Sporbert, T. Cremer, D. Zink, and M. C. Cardoso, "Dynamics of dna replication factories in living cells," *The Journal of cell biology*, vol. 149, no. 2, pp. 271–280, 2000.

[14] M. Hesse, A. Raulf, G.-A. Pilz, C. Haberlandt, A. M. Klein, R. Jabs, H. Zaehres, C. J. Fügemann, K. Zimmermann, J. Trebicka *et al.*, "Direct visualization of cell division using high-resolution imaging of m-phase of the cell cycle," *Nature communications*, vol. 3, p. 1076, 2012.

[15] G. I. Evan and K. H. Vousden, "Proliferation, cell cycle and apoptosis in cancer," *Nature*, vol. 411, no. 6835, pp. 342–348, 2001.

[16] E. Trotta and M. Paci, "Solution structure of dapi selectively bound in the minor groove of a dna t· t mismatch-containing site: Nmr and molecular dynamics studies," *Nucleic acids research*, vol. 26, no. 20, pp. 4706–4713, 1998.

[17] D. Robertis, "Cell and molecular biology," 1987.

[18] M. Tessema, U. Lehmann, and H. Kreipe, "Cell cycle and no end," *Virchows Archiv*, vol. 444, no. 4, pp. 313–323, 2004.

[19] A. Obaya and J. Sedivy, "Regulation of cyclin-cdk activity in mammalian cells," *Cellular and Molecular Life Sciences CMLS*, vol. 59, no. 1, pp. 126–142, 2002.

[20] D. Johnson and C. Walker, "Cyclins and cell cycle checkpoints," *Annual review of pharmacology and toxicology*, vol. 39, no. 1, pp. 295–312, 1999.

[21] K. Vermeulen, D. R. Van Bockstaele, and Z. N. Berneman, "The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer," *Cell proliferation*, vol. 36, no. 3, pp. 131–149, 2003.

[22] C. Ziegler and C. Behl, "Cell aging: molecular mechanisms and implications for disease," 2014.

[23] A. S. Lundberg and R. A. Weinberg, "Functional inactivation of the retinoblastoma protein requires sequential modification by at least two distinct cyclin-cdk complexes," *Molecular and cellular biology*, vol. 18, no. 2, pp. 753–761, 1998.

[24] R. A. Weinberg, "The retinoblastoma protein and cell cycle control," *Cell*, vol. 81, no. 3, pp. 323–330, 1995.

[25] K. Ohtani, J. Degregori, and J. R. Nevins, "Regulation of the cyclin e gene by transcription factor e2f1," *Proceedings of the National Academy of Sciences*, vol. 92, no. 26, pp. 12 146–12 150, 1995.

[26] R. Müller, "Transcriptional regulation during the mammalian cell cycle," *Trends in Genetics*, vol. 11, no. 5, pp. 173–178, 1995.

[27] N. Dyson, "The regulation of e2f by prb-family proteins," *Genes & development*, vol. 12, no. 15, pp. 2245–2262, 1998.

[28] J. W. Harbour and D. C. Dean, “The rb/e2f pathway: expanding roles and emerging paradigms,” *Genes & development*, vol. 14, no. 19, pp. 2393–2409, 2000.

[29] P. R. Yew, “Ubiquitin-mediated proteolysis of vertebrate g1-and s-phase regulators,” *Journal of cellular physiology*, vol. 187, no. 1, pp. 1–10, 2001.

[30] J. Vlach, S. Hennecke, and B. Amati, “Phosphorylation-dependent degradation of the cyclin-dependent kinase inhibitor p27kip1,” *The EMBO journal*, vol. 16, no. 17, pp. 5334–5344, 1997.

[31] H. Morisaki, A. Fujimoto, A. Ando, Y. Nagata, K. Ikeda, and M. Nakanishi, “Cell cycle-dependent phosphorylation of p27 cyclin-dependent kinase (cdk) inhibitor by cyclin e/cdk2,” *Biochemical and biophysical research communications*, vol. 240, no. 2, pp. 386–390, 1997.

[32] K. Helin, “Regulation of cell proliferation by the e2f transcription factors,” *Current opinion in genetics & development*, vol. 8, no. 1, pp. 28–35, 1998.

[33] S. J. Weintraub, C. A. Prater, and D. C. Dean, “Retinoblastoma protein switches the e2f site from positive to negative element,” 1992.

[34] A. Schulze, K. Zerfass, D. Spitkovsky, S. Middendorp, J. Berges, K. Helin, P. Jansen-Dürr, and B. Henglein, “Cell cycle regulation of the cyclin a gene promoter is mediated by a variant e2f site,” *Proceedings of the National Academy of Sciences*, vol. 92, no. 24, pp. 11 264–11 268, 1995.

[35] T. K. Fung and R. Y. Poon, “A roller coaster ride with the mitotic cyclins,” in *Seminars in cell & developmental biology*, vol. 16, no. 3. Elsevier, 2005, pp. 335–342.

[36] W. Zhu, P. H. Giangrande, and J. R. Nevins, “E2fs link the control of g1/s and g2/m transcription,” *The EMBO journal*, vol. 23, no. 23, pp. 4615–4626, 2004.

[37] P. R. CLARKE and E. KARSENTI, “Regulation of p34 ccto2 protein kinase: new insights into protein phosphorylation and the cell cycle,” *Journal of cell science*, vol. 100, pp. 409–414, 1991.

[38] B. Lewin, “Driving the cell cycle: M phase kinase, its partners, and substrates,” *Cell*, vol. 61, no. 5, pp. 743–752, 1990.

[39] K. W. Kohn, “Functional capabilities of molecular network components controlling the mammalian g1/s cell cycle phase transition,” *Oncogene*, vol. 16, no. 8, pp. 1065–1075, 1998.

[40] M. Xu, K.-A. Sheppard, C.-Y. Peng, A. S. Yee, and H. Piwnica-Worms, “Cyclin a/cdk2 binds directly to e2f-1 and inhibits the dna-binding activity of e2f-1/dp-1 by phosphorylation.” *Molecular and cellular biology*, vol. 14, no. 12, pp. 8420–8431, 1994.

[41] B. D. Dynlacht, K. Moberg, J. A. Lees, E. Harlow, and L. Zhu, “Specific regulation of e2f family members by cyclin-dependent kinases.” *Molecular and cellular biology*, vol. 17, no. 7, pp. 3867–3875, 1997.

[42] J. J. Tyson, “Modeling the cell division cycle: cdc2 and cyclin interactions.” *Proceedings of the National Academy of Sciences*, vol. 88, no. 16, pp. 7328–7332, 1991.

[43] M. J. Solomon, T. Lee, and M. W. Kirschner, “Role of phosphorylation in p34cdc2 activation: identification of an activating kinase.” *Molecular biology of the cell*, vol. 3, no. 1, pp. 13–27, 1992.

[44] P. Nurse, “Universal control mechanism regulating onset of m-phase.” *Nature*, vol. 344, no. 6266, pp. 503–508, 1990.

[45] J. Pines and T. Hunter, “Human cyclins a and b1 are differentially located in the cell and undergo cell cycle-dependent nuclear transport.” *The Journal of Cell Biology*, vol. 115, no. 1, pp. 1–17, 1991.

[46] W. Zachariae and K. Nasmyth, “Whose end is destruction: cell division and the anaphase-promoting complex,” *Genes & Development*, vol. 13, no. 16, pp. 2039–2058, 1999.

[47] H. F. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell *et al.*, *Molecular cell biology*. Citeseer, 2000, vol. 4.

[48] L. H. Hartwell and T. A. Weinert, “Checkpoints: controls that ensure the order of cell cycle events,” *Science*, vol. 246, no. 4930, pp. 629–634, 1989.

[49] G. H. Williams and K. Stoeber, “The cell cycle and cancer,” *The Journal of pathology*, vol. 226, no. 2, pp. 352–364, 2012.

[50] W. M. Yen, H. Yamamoto *et al.*, *Phosphor handbook*. CRC press, 2006.

[51] J. W. Lichtman and J.-A. Conchello, “Fluorescence microscopy,” *Nature methods*, vol. 2, no. 12, pp. 910–919, 2005.

[52] M. G. Macey, *Flow Cytometry*. Springer, 2007.

[53] A. L. Givan, *Flow cytometry: first principles*. John Wiley & Sons, 2013.

[54] W. Godfrey, D. Hill, J. Kilgore, G. Buller, J. Bradford, D. Gray, I. Clements, K. Oakleaf, J. Salisbury, M. Ignatius *et al.*, “Complementarity of flow cytometry and fluorescence microscopy,” *Microscopy and Microanalysis*, vol. 11, no. S02, pp. 246–247, 2005.

[55] R. L. Gimlich and J. Braun, “Improved fluorescent compounds for tracing cell lineage,” *Developmental biology*, vol. 109, no. 2, pp. 509–514, 1985.

[56] D. P. Kuffler, “Long-term survival and sprouting in culture by motoneurons isolated from the spinal cord of adult frogs,” *Journal of comparative neurology*, vol. 302, no. 4, pp. 729–738, 1990.

[57] K. Horikawa and W. Armstrong, “A versatile means of intracellular labeling: injection of biocytin and its detection with avidin conjugates,” *Journal of neuroscience methods*, vol. 25, no. 1, pp. 1–11, 1988.

[58] A. Sakaue-Sawano, H. Kurokawa, T. Morimura, A. Hanyu, H. Hama, H. Osawa, S. Kashiwagi, K. Fukami, T. Miyata, H. Miyoshi *et al.*, "Visualizing spatiotemporal dynamics of multicellular cell-cycle progression," *Cell*, vol. 132, no. 3, pp. 487–498, 2008.

[59] H. C. Vodermaier, "Apc/c and scf: controlling each other and the cell cycle," *Current Biology*, vol. 14, no. 18, pp. R787–R796, 2004.

[60] H. Nishitani, Z. Lygerou, and T. Nishimoto, "Proteolysis of dna replication licensing factor cdt1 in s-phase is performed independently of geminin through its n-terminal region," *Journal of Biological Chemistry*, vol. 279, no. 29, pp. 30 807–30 816, 2004.

[61] J. Zhang, R. E. Campbell, A. Y. Ting, and R. Y. Tsien, "Creating new fluorescent probes for cell biology," *Nature Reviews Molecular Cell Biology*, vol. 3, no. 12, pp. 906–918, 2002.

[62] A. Sakaue-Sawano, H. Kurokawa, T. Morimura, A. Hanyu, H. Hama, H. Osawa, S. Kashiwagi, K. Fukami, T. Miyata, H. Miyoshi *et al.*, "Visualizing spatiotemporal dynamics of multicellular cell-cycle progression," *Cell*, vol. 132, no. 3, pp. 487–498, 2008.

[63] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "Nih image to imagej: 25 years of image analysis," *Nature methods*, vol. 9, no. 7, pp. 671–675, 2012.

[64] M. Bertero and P. Boccacci, *Introduction to inverse problems in imaging*. CRC press, 1998.

[65] I. C. Rodrigues and J. M. R. Sanches, "Convex total variation denoising of poisson fluorescence confocal images with anisotropic filtering," *Image Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 146–160, 2011.

[66] R. C. Gonzalez and R. E. Woods, "Digital image processing 3rd edition," 2007.

[67] C. Wählby, I.-M. SINTORN, F. Erlandsson, G. Borgefors, and E. Bengtsson, "Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections," *Journal of Microscopy*, vol. 215, no. 1, pp. 67–76, 2004.

[68] A. Lin, L. Wu, B. Zheng, and H. Zan, "The combination of local fuzzy-entropy-based transition region extraction with otsu thresholding method for image segmentation," in *Image and Signal Processing, 2009. CISP'09. 2nd International Congress on*. IEEE, 2009, pp. 1–4.

[69] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[70] A. Pinidiyaarachchi and C. Wählby, "Seeded watersheds for combined segmentation and tracking of cells," in *Image Analysis and Processing–ICIAP 2005*. Springer, 2005, pp. 336–343.

[71] M. M. Suarez-Alvarez, D.-T. Pham, M. Y. Prostov, and Y. I. Prostov, "Statistical approach to normalization of feature vectors and clustering of mixed datasets," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. The Royal Society, 2012, p. rspa20110704.

[72] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. Siam, 2007, vol. 20.

[73] G. W. Milligan and M. C. Cooper, "A study of standardization of variables in cluster analysis," *Journal of classification*, vol. 5, no. 2, pp. 181–204, 1988.

[74] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning*, vol. 30, no. 2-3, pp. 271–274, 1998.

[75] C. Bishop, "Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn," 2007.

[76] S. Marsland, *Machine learning: an algorithmic perspective*. CRC press, 2014.

[77] T. M. Mitchell, "Machine learning and data mining," *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[78] V. V. Miagkikh and W. F. Punch III, "An approach to solving combinatorial optimization problems using a population of reinforcement learning agents," in *Proceedings of the genetic and evolutionary computation conference*, vol. 2, 1999, pp. 1358–1365.

[79] V. Miagkikh, "A survey of reinforcement learning and agent-based approaches to combinatorial optimization," 2012.

[80] B. S. Everitt, "Unresolved problems in cluster analysis," *Biometrics*, pp. 169–181, 1979.

[81] B. Babcock, M. Datar, R. Motwani, and L. O'Callaghan, "Maintaining variance and k-medians over data stream windows," in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2003, pp. 234–243.

[82] J. J. Rocchio, "Document retrieval system-optimization and," 1966.

[83] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[84] J. D. Banfield and A. E. Raftery, "Model-based gaussian and non-gaussian clustering," *Biometrics*, pp. 803–821, 1993.

[85] G. Celeux and G. Govaert, "A classification em algorithm for clustering and two stochastic versions," *Computational statistics & Data analysis*, vol. 14, no. 3, pp. 315–332, 1992.

[86] G. J. McLachlan and K. E. Basford, "Mixture models: Inference and applications to clustering," *Applied Statistics*, 1988.

[87] A. Dasgupta and A. E. Raftery, "Detecting features in spatial point processes with clutter via model-based clustering," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 294–302, 1998.

[88] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," *The computer journal*, vol. 41, no. 8, pp. 578–588, 1998.

[89] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[90] F. Murtagh and A. E. Raftery, "Fitting straight lines to point patterns," *Pattern recognition*, vol. 17, no. 5, pp. 479–483, 1984.

[91] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models," *Pattern recognition*, vol. 28, no. 5, pp. 781–793, 1995.

[92] H. P. Friedman and J. Rubin, "On some invariant criteria for grouping data," *Journal of the American Statistical Association*, vol. 62, no. 320, pp. 1159–1178, 1967.

[93] A. J. Scott and M. J. Symons, "Clustering methods based on likelihood ratio criteria," *Biometrics*, pp. 387–397, 1971.

[94] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[95] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[96] R. A. Boyles, "On the convergence of the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 47–50, 1983.

[97] L. Xu and M. I. Jordan, "On convergence properties of the em algorithm for gaussian mixtures," *Neural computation*, vol. 8, no. 1, pp. 129–151, 1996.

[98] H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert, "Inference in model-based cluster analysis," *Statistics and Computing*, vol. 7, no. 1, pp. 1–10, 1997.

[99] D. A. Brott, J. D. Alvey, M. R. Bleavins, F. A. de la Iglesia, and N. D. Lalwani, "Cell cycle dependent distribution of proliferating cell nuclear antigen/cyclin and cdc2-kinase in mouse t-lymphoma cells," *Journal of cellular biochemistry*, vol. 52, no. 3, pp. 362–372, 1993.

[100] B. Pérez-Cadahía, B. Drobic, and J. R. Davie, "H3 phosphorylation: dual role in mitosis and interphase this paper is one of a selection of papers published in this special issue entitled 30th annual international asilomar chromatin and chromosomes conference and has undergone the journal's usual peer review process." *Biochemistry and Cell Biology*, vol. 87, no. 5, pp. 695–709, 2009.