

Creating Classification Features for Biological Images

Abstract

Computational Biology covers a broad spectrum of diverse applications within the field of molecular biology. Much of the computational work done so far has focused at the molecular level because of the strong computational characteristics of molecular biology. This tight focus has left relatively unexplored other aspects of biology such as cell biology that might benefit from computational techniques. This thesis examines the use of computational techniques in the field of cell biology. Towards this end, we develop a system based on image processing and machine learning to characterize the cellular events occurring in the cell division process meiosis and to classify images of cells exhibiting these events. The cellular events in question are the eight phases of Meiosis and the post-meiotic events that decide the phenotype of the resulting cells. The results of this thesis suggest the existence of significant application potential in computational techniques to the field of cell biology.

Acknowledgements

First, I would like to thank my advisor, Dr. Rich Maclin, for his persistent efforts in helping me finish this thesis. I would like to express my gratitude to Dr. Doug Dunham and Dr. Joe Gallian for their willingness to serve on my committee. I would also like to thank Dr. Qin Liu and her students, Kevin Wolfe and Jennifer Baumgardt, for providing images for this work and helping me with the biological aspects of this thesis. Finally, I would like to thank the CS faculty and my fellow students for their support during these two years of graduate study and making my stay in Duluth an enjoyable and memorable experience.

Contents

Introduction	1
Background	4
Meiosis	4
Phases of Meiosis	7
Wild-type and Mutant cell types	18
Image Analysis	21
Image Preparation	21
Image Histogramming and Segmentation	21
Image Normalization	22
Region Isolation	23
Region Identification	25
Machine Learning	27
Inductive Learning	28
Characterizing Cells during Meiosis	29
Features Examined	29
Feature: Number of Internal Regions	31
Feature: Main Region Occupancy by Internal Regions..	31
Feature: Area	31
Feature: Perimeter	32
Feature: Radius	32
Feature: Compactness	32
Feature: Smoothness	32
Feature: Texture	32
Feature: Concavity	34
Feature: Concave points	34
Initial Results	34
Result: Number of Internal Regions	35
Result: Main Region Occupancy by Internal Regions..	35
Result: Area	38
Result: Perimeter	38
Result: Radius	41
Result: Compactness	41
Result: Smoothness	41
Result: Texture	45
Result: Concavity	45
Result: Concave points	45
A Cell Classifier	49
Test Results	52

Future Work	58
Event Classification	58
Image Processing	59
Machine Learning	60
Conclusions	61
References	63

List of Figures

2.1	Schematic drawing of the reproduction process in multicellular organisms	6
2.2	Appearance of homologous pair of chromosomes at the beginning Of Meiosis	6
2.3	Picture of a cell in prophase I	8
2.4	Picture of a cell in prometaphase	9
2.5	Picture of a cell in metaphase I	11
2.6	Picture of a cell in anaphase I	12
2.7	Picture of a cell in telophase I	13
2.8	Picture of a cell in prophase II	15
2.9	Picture of a cell in metaphase II	16
2.10	Picture of a cell in anaphase II	17
2.11	Picture of a cell in telophase II	19
2.12	Images of cells exhibiting (a) wild-type, (b) ms2 mutation and (c) po mutation	20
2.13	Histogram of a cell in telophase I	20
2.14	Image of cells before and after applying normalization and their corresponding histograms	24
2.15	A cell isolated from a group of cells using region isolation tool	26
2.16	Regions identified in a cell image using region extraction tool	26
3.1	Images of cells exhibiting (a) wild-type, (b) ms2 mutation and (c) po mutation	30
3.2	Images of cells exhibiting (a) prophase I, (b) prometaphase, (c) telophase I, and (d) telophase II	30
3.3	Radial region lines used for computing region smoothness	33
3.4	Region chords used for computing region concavity	33
3.5	Plots of number of internal regions to main regions	36
3.6	Plots of main region occupancy by internal regions	37
3.7	Plots of area of internal regions	39
3.8	Plots of perimeter of internal regions	40
3.9	Plots of radius of internal regions	42
3.10	Plots of compactness of internal regions	43
3.11	Plots of smoothness of internal regions	44
3.12	Plots of texture of internal regions	46
3.13	Plots of concavity of internal regions	47
3.14	Plots of number of concave points in internal regions	48
3.15	A classifier for wild-type, ms6 and po mutation cell images	50
3.16	A classifier for prophase I, prometaphase, telophase I and telophase II cell images	51

3.17	Plot of the number of internal regions to main regions in wild-type, ms2 and po mutation test cell images	53
3.18	Plot of main region occupancy by internal regions in wild-type, ms2 and po mutation cell images	53
3.19	Plot of texture of internal regions in wild-type, ms2 and po mutation test cell images	54
3.20	Plot of the number of internal regions in prophase I, prometaphase, telophase I and telophase II test cell images	56
3.21	Plot of main region space occupancy by internal regions in prophase I, prometaphase, telophase I and telophase II test cell images	56

List of Tables

3.1	Test results of classifier for wild-type cell, ms6 and po mutation Cells	52
3.2	Test results of classifier for prophase I, prometaphase, telophase I and telophase II cells	55

1 Introduction

Computational Biology covers a broad spectrum of diverse fields ranging from techniques for determining molecular crystal structure based on X-ray crystallography data (Bruenger 1991; Nilges et al. 1991), to methods for simulating molecular interaction at various levels (Socci et.al., 1996; Warshel et.al., 1991), to the maintenance of Biological databases such as the Human GENOME project (Watson 1990) or the Ribosomal Database Project (Maidak et al. 1996), and the recognition of molecular features such as protein secondary structure (Holley et.al., 1989; Qian et.al., 1988; Rost et.al., 1993). Though these approaches vary significantly in the computational approaches used, they do share a strong focus on the molecular level of Biology. This focus is not surprising in that there is a strong computational aspect to much of the reasoning done at the molecular level and certain types of problems could not be approached without modern computational techniques. This tight focus has left relatively unexplored other aspects of biology such as cell biology that might benefit from computational techniques.

This thesis makes an attempt to examine the use of computational techniques in the field of cell biology. One area of cell biology that provides such an opportunity is the systematic examination of the process of cell division known as Meiosis and related mutations that occur in plant and animal cells during sexual reproduction. The process is characterized by a series of cellular events that take a single cell through a sequence of structural changes to form four new cells. The events in question are the eight phases of Meiosis process, namely prophase I, metaphase I, anaphase I, telophase I, prophase II, metaphase II, anaphase II and telophase II; and meiotic events such as *po* and *ms6* mutations, which decide the phenotype of the resulting cells. Cells undergoing Meiosis experience these events through a series of morphological changes unique to the events. These morphological changes can be quantitatively captured using computational techniques and be used to develop models characterizing these events. This thesis develops a system based on computer vision and machine learning techniques to characterize the cellular events governing the Meiosis process. The subject cells used for this work are from the maize plant, obtained from the Department of Biology at the University of

Minnesota, Duluth. The approach taken is to analyze digital images of cells undergoing Meiosis to obtain measurable, quantifiable features to be used to generate cellular maps characterizing the events governing the process.

Techniques for digital image analysis have a long history (Ballard et.al., 1982). They have played a part in a number of approaches to feature extraction for cells (Dawe et al. 1994; Wied et al. 1989; Wittekind et.al., 1987), but often these approaches have focused either on performing image transformation to make the image more clear for a human analyzer or have been used to make simple measurements with a human user performing analysis of resulting data.

More recently, researchers have begun to use image analysis and machine learning techniques to assist in the recognition of features associated with cells (Turner et al. 1993; Wohlberg et.al, 1993a; 1995). In these approaches, digital images of cells are analyzed using computer vision techniques and descriptive features are extracted that characterize aspects of the cells. Machine learning techniques are then used to determine a map to characterize the differences between a set of examples of cells that are exhibiting a certain property and cells that do not exhibit the property. One advantage of such a quantitative map is that human viewers often introduce biases in their analysis of images or may miss properties of images that require transformation of the image. A computer map allows for a recognition method that avoids such biases and is the focus of this work. Humans are often subjective in their observations. A cell biologist may misclassify a cell image based on some preconceived notions on cell types. Furthermore, he may tend to overlook certain critical and subtle aspects of the image that would play an important role in deciding the type of the cell image. A computer map would allow cell biologists to reinforce their observations through its results and in certain cases make decisions on their behalf when the level of observation required goes beyond human comprehension.

For this research, digital image analysis is used to produce quantitative models of cells in different states of the cell division process. For example, several digital images of cells in different states, (e.g., prophase I and metaphase I) are analyzed to produce cellular maps that

characterize precisely the differences that indicate which cells are in prophase I and which are in metaphase I. To produce such an appropriate cellular map of the different cell types, the creation of these maps is treated as an inductive learning problem. The goal of inductive learning is to determine a map that allows to differentiate between examples of objects that are part of a class (e.g., cells exhibiting prophase I properties) from objects in other classes (e.g., cells of type metaphase I). To do this, the inductive learner¹ is presented with examples to determine combinations of the features that allow it to distinguish between the different classes of the examples. The resulting map is used to classify new examples that were not part of the set of training examples. To produce appropriate maps, this research focuses on two aspects of the problem: (1) creating appropriate features to describe the different cells that are useful in characterizing the differences between cells; and (2) selecting from amongst the set of possible features the ones that best characterize the cells.

The crux of this thesis is then to characterize the meiotic and post-meiotic cellular events occurring in reproduction cells, and towards this end develop a system based on Computer Vision and Machine Learning techniques to classify cell images that exhibit these events.

The following chapter presents background material relevant to this work. This includes a discussion of Meiosis, along with the concepts of image analysis that form the basis of this research. Chapter 3 presents the features examined and extracted from cell images, the methodology used to extract the features, the final set of features used to generate maps of cellular events, and the results obtained by application of the cellular maps to cell images not part of the training data set. The last two chapters discuss future research and conclusions that arise from this work.

¹ A system that learns from a set of labeled examples (Quinlan, 1986).

2 Background

This chapter presents background material relevant to this thesis. The first section discusses the process of cell division Meiosis, the area of cell biology on which this research is focused. The second section presents concepts from image analysis used in this work.

2.1 Meiosis

Living organisms do not survive forever (Albert et.al., 1983). In order for the species to survive, they need to reproduce. Reproduction in an organism, as in evolution, begins at the cellular level. Cells exhibit two forms of reproduction; the first form of reproduction, known as asexual reproduction, contributes to the growth of an individual; the second form of reproduction, known as sexual reproduction, can help bring a new organism into existence. The asexual form of reproduction involves a process of cell division known as Mitosis. This process involves a single division of a cell into two cells that are genetically identical to the parent cell. It is experienced by both germ and somatic cells that make up the body of an organism; the former specialized in sexual reproduction and the later in other cellular functions. Mitosis causes cells to proliferate in the body and maintain the growth of an organism. It replaces worn-out cells with healthy cells to maintain the vitality of adult tissues. A single fertilized egg grows into an multicellular adult by repeatedly undergoing Mitosis. The other form of reproduction, known as sexual reproduction, involves the cell division process of Meiosis. Meiosis occurs only in germ cells and not in somatic cells. The Meiosis process involves the cell division of a germ cell into four gametes. Gametes are cells specialized in sexual fusion. Each gamete contains half the genetic complement of the germ cell. The type of gamete formed (sperm or egg) depends on the sex of the organism.

The reproductive cycle in an organism starts with the germ cell undergoing meiotic division to form four gamete cells. A germ cell contains two sets of chromosomes, one from each parent, and hence has a diploid ($2n$, where n =number of distinct chromosomes) amount of DNA. The

gametes formed get half the genetic complement of the germ cell with a single set of chromosomes, which give them a haploid (n) amount of DNA. Their chromosomes carry a mix of genes from the parents. The reproduction cycle culminates with the fusion of gametes, the sperm cells with the egg cell, to form a zygote, the first cell of a new individual. This process is known as fertilization. The zygote then replicates itself through Mitosis to form a multicellular organism. The schematic diagram in Figure 2.1 depicts this reproduction process in a multicellular organism.

When a cell is ready to undergo a cell division, the DNA found in its nucleus manifests itself in the form of chromosomes. As mentioned earlier, germ cells contain a diploid amount of DNA and hence have two sets of chromosomes, each coming from a different parent. These chromosomes occur in pairs, where one chromosome in the pair comes from the male parent and the other from the female parent. These pairs are called homologous chromosome pairs and the two chromosomes involved in it are called homologs. Chromosomes are highly coiled molecules of DNA containing single strands of nucleic acid known as chromatids. At the beginning of Meiosis, this single strand of chromatid in a chromosome duplicates itself to form a sister chromatid. The two chromatids are held together at a spot called the centromere where they join. The homologous pairs of chromosomes and the centromere play an important role in the Meiosis process. Figure 2.2 shows the appearance of homologous pair of chromosomes before Meiosis.

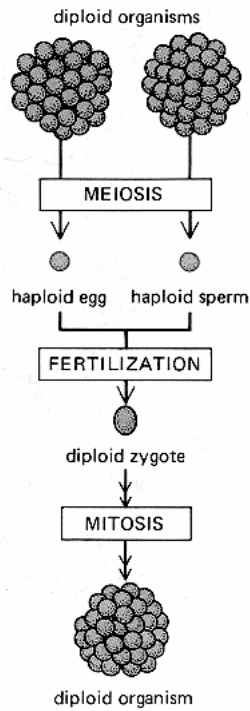


Figure 2.1: Schematic drawing (Albert et.al., 1983) showing the reproduction process in multicellular organisms.

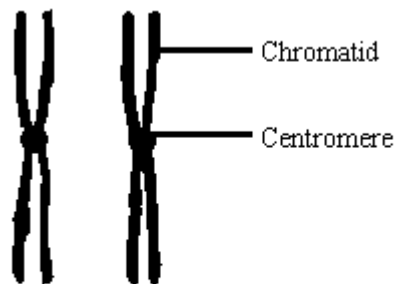


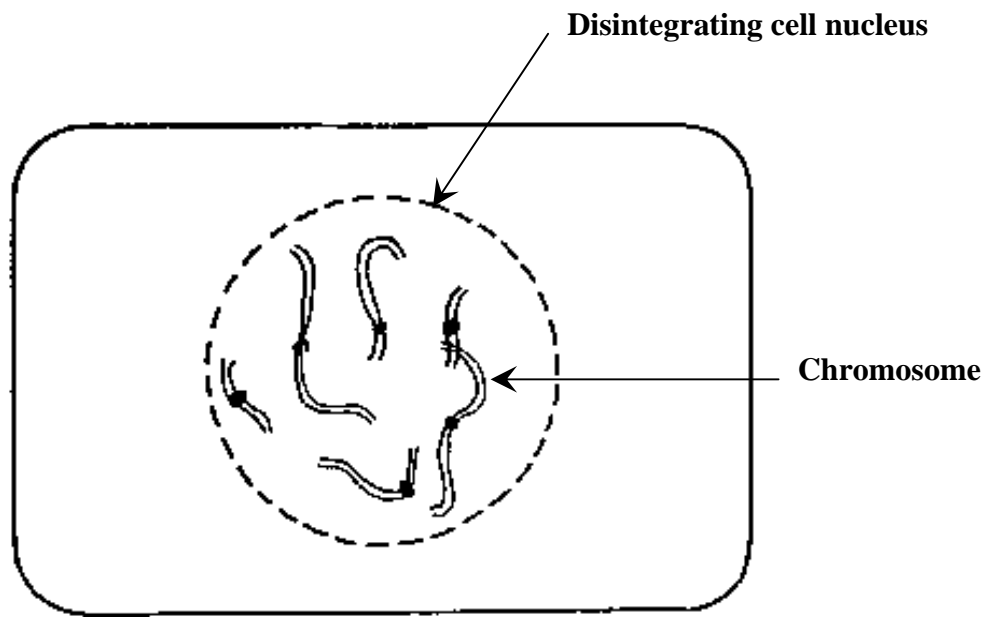
Figure 2.2: Appearance of homologous pair of chromosomes at the beginning of Meiosis.

2.1.1 Phases of Meiosis

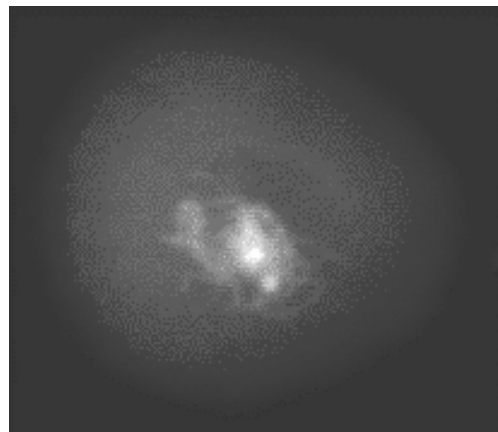
The process of Meiosis is spread over eight different phases. Over these eight phases, the cell undergoes two cell divisions. The first cell division occurs through the first four phases of the process and ends with the formation of two diploid cells. The second division occurs during the last four phases where the diploid cells undergo division to form four haploid cells. Following are the eight phases of the process of Meiosis:

Prophase I

Prophase I is the longest phase of Meiosis and takes about ninety percent of its total time. Elaborate morphological changes occur to the chromosomes of the cell during this phase. The beginning of the phase is marked by the disintegration of the nuclear envelope, which encloses the DNA. Chromosomes, which are otherwise invisible, start to shorten and thicken in size and become discernible. As time elapses, the nuclear envelope disappears and the chromosomes spread out through the cell. Homologous chromosomes seek out their counterparts and start pairing. When the pairing is complete, the paired homologs get connected between non-sister chromatids at points called chiasma. These are the points where the transfer of genetic information takes place. Towards the end of the phase, spindle fibres begin to form, connecting the homologous chromosome pair to the opposite poles of the cell. This part of prophase is called prometaphase and marks the end of this phase. Figure 2.3 (a) shows a stylized picture of a prophase I cell with its shortening chromosomes and a disintegrating nucleus. The image in 2.3 (b) shows a Maize cell exhibiting the phase. Figure 2.4 (a) shows the stylized picture of a cell in prometaphase. The homologous chromosomes are paired together and spindle fibres connecting them to the cell poles are visible. The image in 2.4 (b) shows the corresponding Maize cell in prometaphase. The chromosome fragments in the cell represent the homologous chromosome pairs. The spindle fibres are not visible in the cell image.



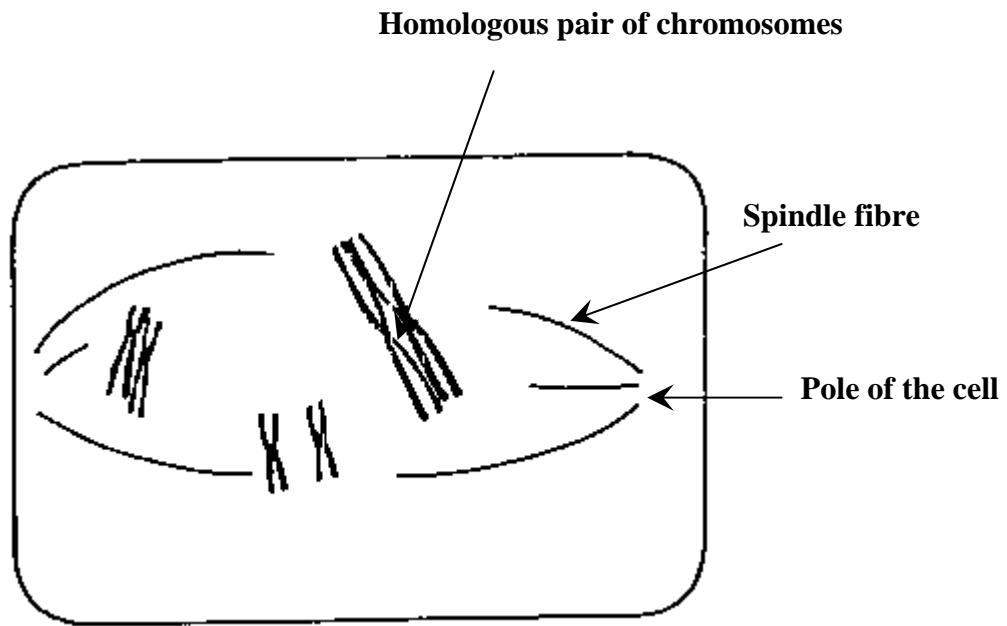
(a)



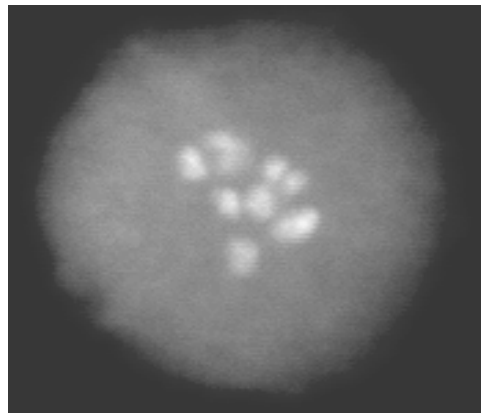
(b)

Prophase I

Figure 2.3: (a) A stylized picture of a cell exhibiting prophase I and (b) An image of a maize cell in prophase I.



(a)



(b)

Prometaphase

Figure 2.4: (a) A stylized picture of a cell exhibiting prometaphase and (b) An image of a maize cell in prometaphase .

Metaphase I

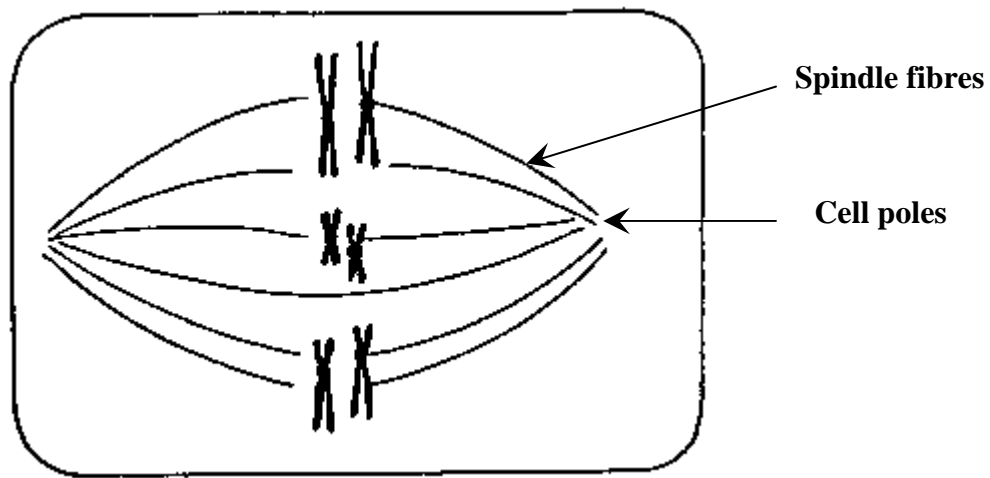
In this phase, the homologous chromosome pairs line up across the equatorial plane of the cell with the spindle fibres connecting them to the opposite poles of the cell. Figure 2.5 (a) shows a stylized picture of a cell exhibiting metaphase I. From the picture, it can be seen that the homologous chromosome pairs are lined up on the equatorial plane and the spindle fibres connect them to the opposite poles. Figure 2.5 (b) image shows a Maize cell in metaphase I. The cell exhibits homologous pairs of chromosomes aligned on the equatorial plane.

Anaphase I

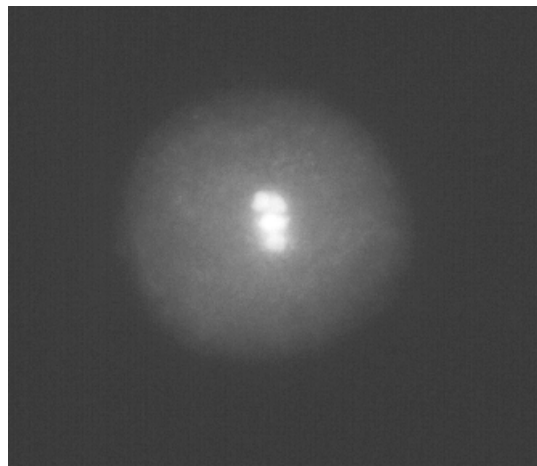
Homologous chromosome pairs, held together by chiasma, are separated from each other towards the poles by shrinking spindle fibres. The pairs break up at chiasma points and genetic information is transferred between the homologs. The genetic makeup of the homologs now comprises of a combination of genetic information from the parents. Figure 2.6 (a) shows a stylized picture of an anaphase I cell. The picture shows the shrinking spindle fibres separating homologous chromosome pairs away from each other. Figure 2.6 (b) shows the corresponding anaphase I Maize cell. Each chromosome from the pair is seen moving away from its counterpart.

Telophase I

The movement of chromosomes to the poles of the cell is complete. The spindle fibres begin to disappear and a cell plate, dividing the cell across the equatorial plane, starts to form. Figure 2.7 (a) shows the stylized picture of a cell in telophase II. The chromosomes are now at the poles of the cell. The cell plate can be seen at the top and bottom sides of the cell. The image in (b) shows the corresponding Maize cell in telophase I. The two chromosomes at the poles of the cell can be seen. The cell plate is barely visible.



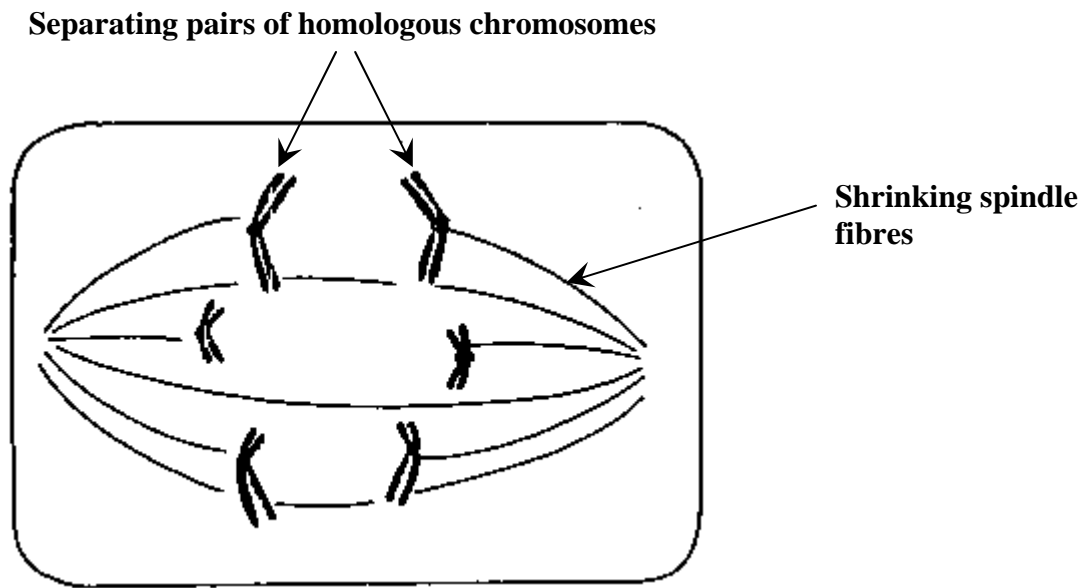
(a)



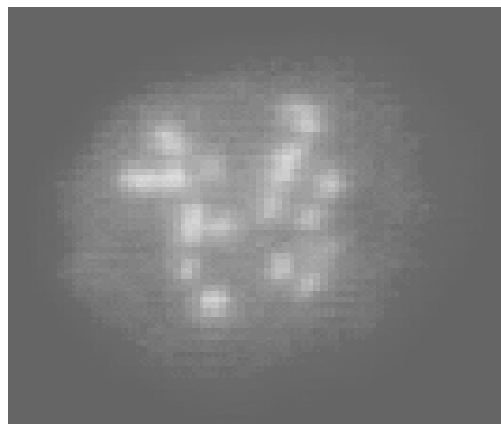
(b)

Metaphase I

Figure 2.5: (a) A stylized picture of a cell exhibiting metaphase I and (b) An image of a maize cell in metaphase I.



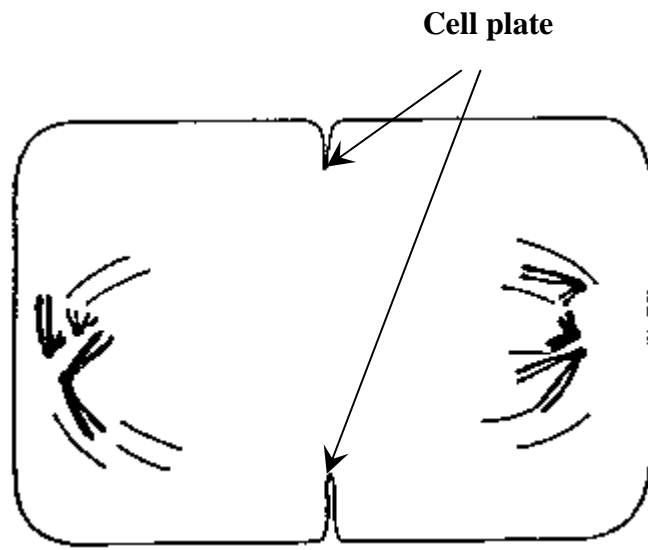
(a)



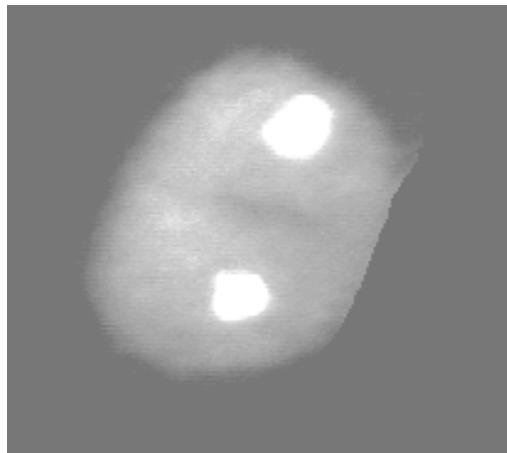
(b)

Anaphase I

Figure 2.6: (a) A stylized picture of a cell exhibiting anaphase I and (b) An image of a maize cell in anaphase I.



(a)



(b)

Telophase I

Figure 2.7: (a) A stylized picture of a cell exhibiting telophase I and (b) An image of a maize cell in telophase I.

This marks the end of the first cell division, resulting in the formation of two new cells. Each cell contains a single set of chromosomes containing genetic information from the two parents. The chromosomes each have two chromatids which makes the cells diploid. The cells continue with the second cell division to go from diploidy to haploidy.

Prophase II

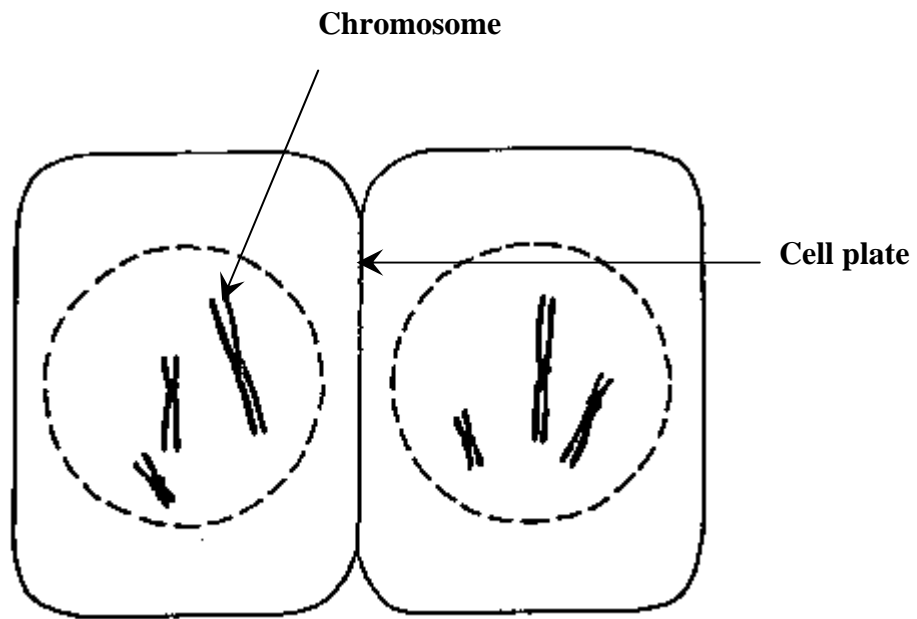
This phase marks the beginning of the second cell division. The chromosomes in each of the two cells formed of the first division become visible again. Spindle fibres are reformed that connect the centromere in the chromosome, holding the sister chromatids together, to the opposite poles of the cell. Figure 2.8 (a) shows the stylized picture of a prophase II cell. Chromosomes shorten and thicken. Figure 2.8 (b) shows the corresponding Maize cell in prophase II.

Metaphase II

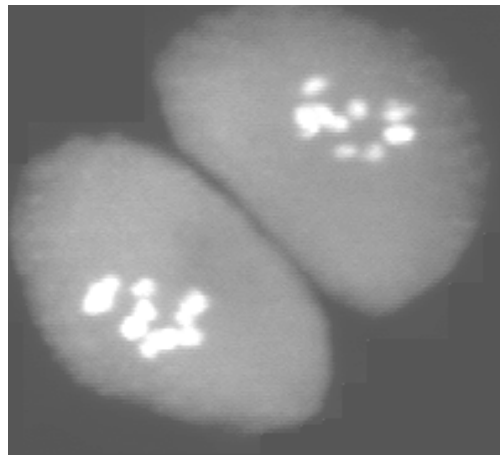
Chromosomes become aligned on the equatorial plane of the cells. The stylized image of a Metaphase II cell in Figure 2.9 (a) shows the chromosomes in the two cells aligned again on the equatorial plane. The Metaphase II cell of Maize in Figure 2.9 (b) is in a later state of metaphase II where the chromosomes start moving away from each other.

Anaphase II

Spindle fibres shrink, dividing the centromeres and separating the chromatids as chromosomes, towards the opposite poles of the cell. Figure 2.10(a) shows the stylized image of anaphase II cell. The centromeres of the chromosomes are pulled towards the cell poles by shrinking fibres. Figure 2.10 (b) shows the image of a Maize cell in anaphase II.



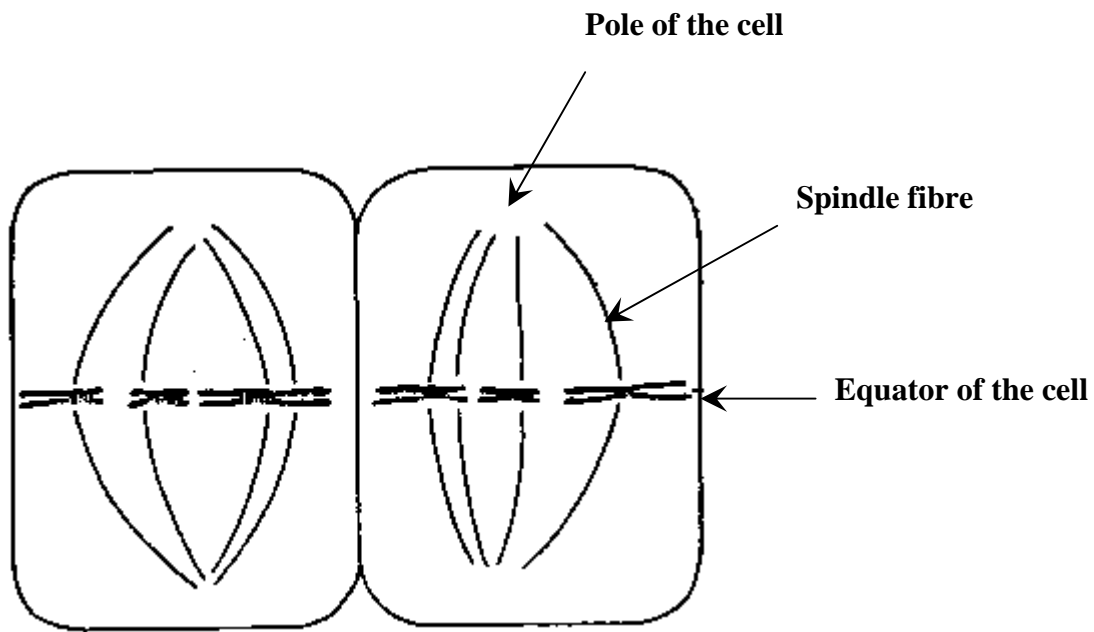
(a)



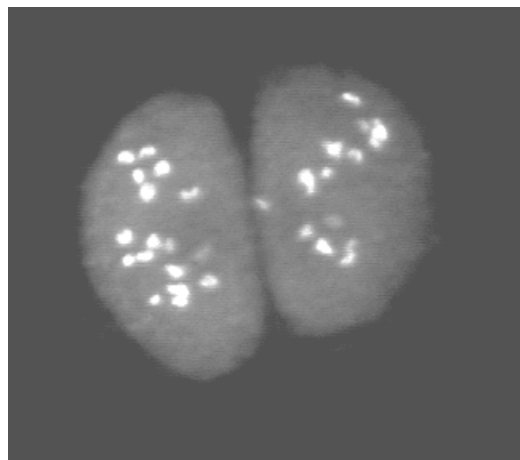
(b)

Prophase II

Figure 2.8: (a) A stylized picture of a cell exhibiting prophase II and (b) An image of a maize cell in prophase II.



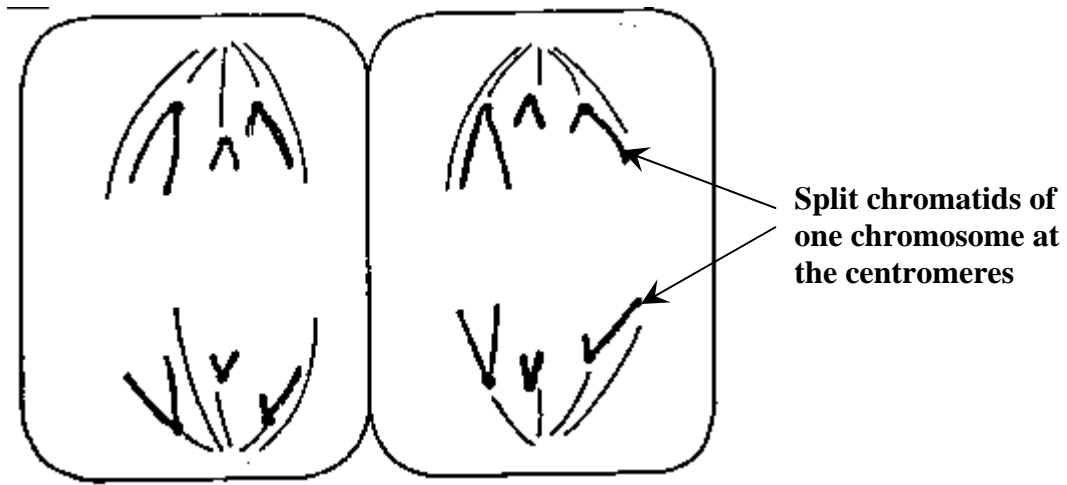
(a)



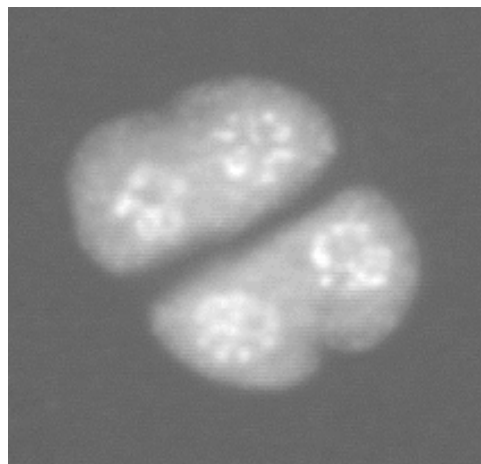
(b)

Metaphase II

Figure 2.9: (a) A stylized picture of a cell exhibiting metaphase II and (b) An image of a maize cell in metaphase II.



(a)



(b)

Anaphase II

Figure 2.10: (a) A stylized picture of a cell exhibiting anaphase II and (b) An image of a maize cell in anaphase II.

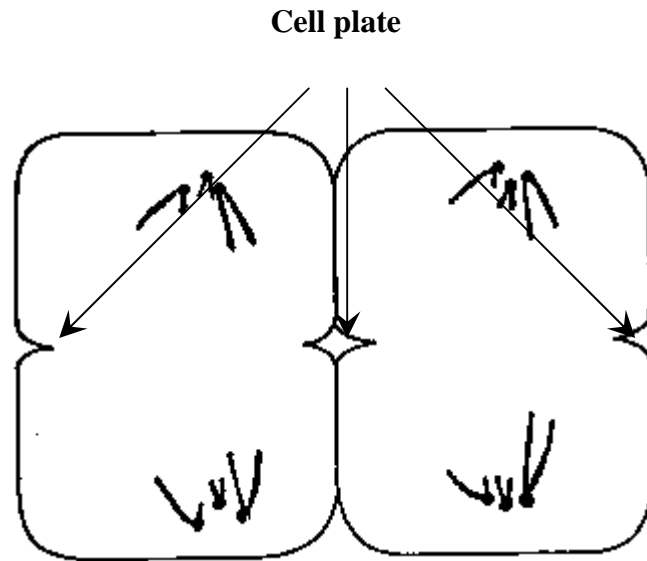
Telophase II

Movement of chromosomes to the poles is complete and spindles disappear. Cell plates are formed across the equatorial plane, dividing the cells into two and forming four haploid cells. These cells contain a single set of chromosomes having a mix of paternal and maternal genetic information. The formation of these cells marks the end of Meiosis. Figure 2.11 (a) shows the stylized picture of telophase II cell. The chromosomes are at the poles of the cell and the second cell plate dividing the two cells into four begins to form. Figure 2.11 (b) shows the corresponding telophase II maize cell. The cell plate in one of the cells has divided it into two cells. The cell plate in the other cell is in an early stage.

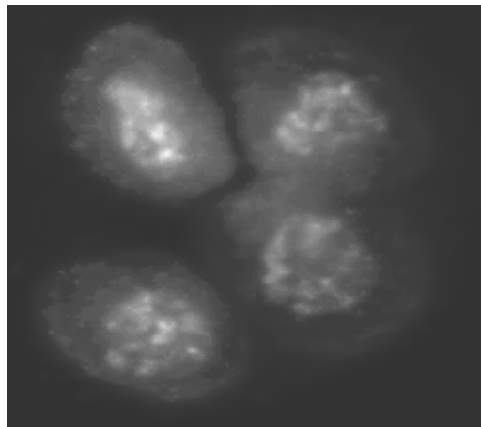
2.1.2 Wild-type and Mutant cell types

Meiotic cells sometimes deviate from the normal sequence of events dictating the cell division process and end up producing mutant cells. These mutations are of different type and are classified according to morphological deviations responsible for the mutations. The mutations that are of interest to this research are polymitotic (po) and ms6 mutations. These mutations are caused by a meiotic cell experiencing multiple cell divisions, in addition to the two prescribed by the process.

Cells that are produced through normal meiotic cell divisions are called wild-type cells. Figure 2.12 shows a (a) wild-type cell image as well as those exhibiting the (b) po and (c) ms6 mutations.



(a)



(b)

Telophase II

Figure 2.11: (a) A stylized picture of a cell exhibiting telophase II and (b) An image of a maize cell in telophase II.

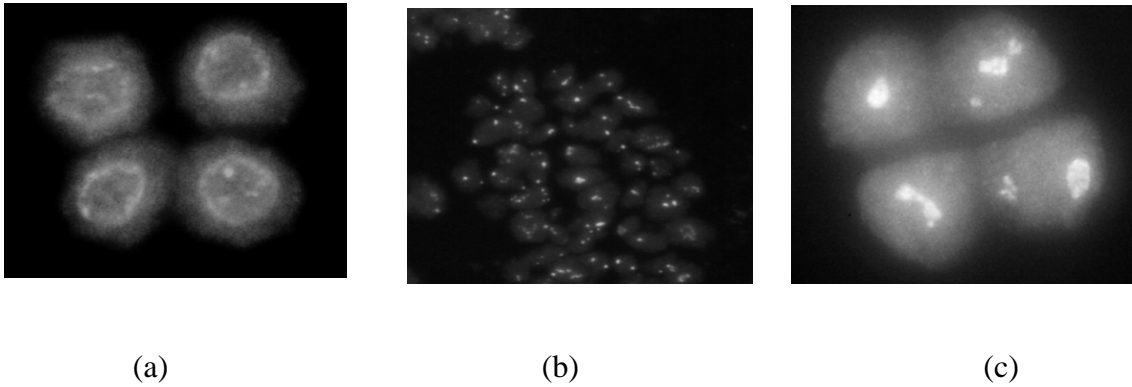


Figure 2.12: Example of cells during a particular stage of Meiosis exhibiting the (a) wild-type, (b) *ms6* mutation, and (c) *po* mutation.

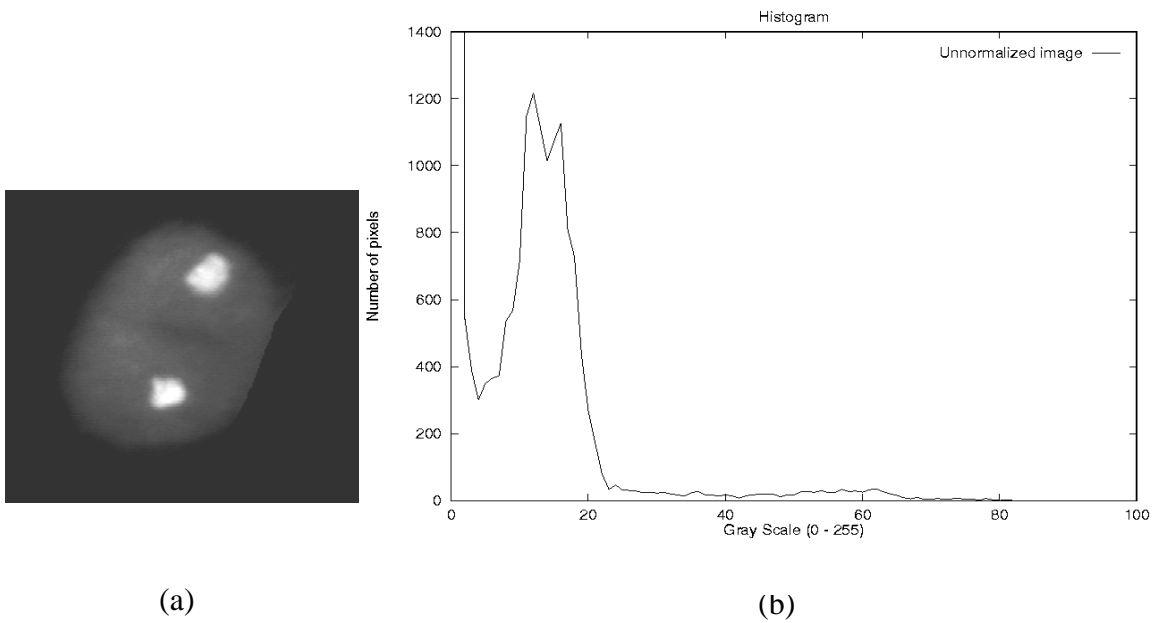


Figure 2.13: (a) An example of a cell in telophase I (b) with its corresponding intensity histogram.

2.2 Image Analysis

Image analysis forms a foundation of this work. Techniques in digital image analysis have often found their use in dealing with cell images. Most of these applications are focused on performing image transformations to make an image clearer to a human analyzer (Dawe et al. 1994; Wied, Bartels, & et al. 1989; Wittekind & Schulte 1987). Seldom have they been used to extract features from cell images except in a few applications (Wohlberg et. al., 1995). In this work, image analysis is used specifically to extract features descriptive of cells undergoing Meiosis. Following is a discussion on the image analysis techniques and the corresponding analysis tools developed for this research.

2.2.1 Image Preparation

The first step towards performing image analysis is to obtain images of the subject cells. This is done using a camera/microscope assembly, with a color camera fitted on top of a microscope and the entire unit connected to a computer. The camera takes pictures of cells on the observation slide below and stores them on the attached computer in some image file format, TIFF format in this case. The pictures taken are in color and are converted to grayscale for the sake of analysis.

2.2.2 Image Histogramming and Segmentation

A grayscale image is a collection of data points known as pixels, exhibiting different levels of intensity, ranging from 0 to 255. In order to analyze an image properly, it is very important to understand the distribution of the image pixels over the grayscale intensities. This is done through a process known as histogramming. A histogram of a grayscale image is a plot of the frequency of occurrence of image pixels at different gray levels (Ballard et. al., 1982). At any particular level in the grayscale, the plot gives the number of image pixels exhibiting that grayscale intensity. Figure 2.13 shows a grayscale image of a meiotic cell and its corresponding histogram plot.

A histogram plot helps in identifying the gray-level intensities in the image that are of most relevance to the analysis at hand. For example, the plot in Figure 2.13 shows that pixels in range 7 to 81 are part of the cell while pixels in the range 60 to 81 form the chromosomes in the cell and makeup for 7 percent of the cell space. Using this form of analysis, an image can be segmented into different parts, identifying portions that are of importance for further analysis and the ones that can be discarded.

2.2.3 Image Normalization

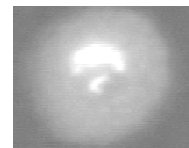
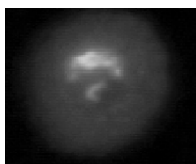
Images of cells obtained by the image preparation process described before exhibit different levels of brightness. These differences are in part due to variations in the staining process used in the slide preparation procedure, and the settings of the equipment used to capture the images. They prevent the application of a uniform approach of analysis to all images. In order to analyze the images at a constant brightness level, these differences have to be removed. This is done by a process known as normalization. Normalization moves the mean grayscale intensity of an image to a specific value and rearranges the image pixel intensities to differ from the mean value by a constant standard deviation. All images to be analyzed are normalized to the same mean grayscale value and standard deviation. This removes the brightness discrepancies between images and renders them useful for uniform analysis. The normalization process is performed by a tool that takes as input the image to be normalized and the values for the new mean gray intensity and the standard deviation for the normalized image. One drawback of this process is that the smoothness of the histogram curve is not maintained after the image is normalized. The reason for this behavior is that, normalization tries to rearrange the image pixels to exhibit a new intensity distribution pattern based on a new standard deviation. If the new standard deviation value is higher than the one exhibited by the image, the pixels are moved away from each other in their intensities causing the histogram plot to display a jagged contour. If the value is lower, pixels with different intensities are moved to occupy the same intensities thus shrinking the plot. This effect causes the normalized image to lose some of its gradual intensity variations, which might

affect the results of image analysis techniques like region extraction applied to the image further.

Figure 2.14 shows two images (a) and (b) with different brightness levels. As seen from the accompanying histograms in (e) and (f), Image 1 has a mean grayscale intensity of 32.7 and standard deviation of 11.32, and Image 2 a mean intensity of 57.6 and standard deviation 13.27. The two images are normalized to a mean grayscale value of 70 and a standard deviation of 15.0. The images after normalization and the corresponding histograms are also shown (c), (d), (e) and (f) respectively. The drawbacks of normalization are evident from the histogram plots of the two normalized images. The normalized histogram plots exhibit jagged contours since the images are normalized to a standard deviation values greater than the original values.

2.2.4 Region Isolation

Cells normally occur in groups and tend to overlap each other on the slide. Images capture cells as they occur and sometimes have more than one cell represented in them. This behavior makes analysis difficult, since only one cell can be analyzed at a time. Hence the cell to be analyzed needs to be isolated from the rest of the group before any further analysis can be carried out. This is achieved by a technique known as region isolation. In this technique, the user first interactively selects from the image a region containing the cell of interest. The selection is in the form of a closed curve along the cell edges. A region isolation tool then separates the region of selection from the rest of the image and creates a new image containing only the selected part. The isolation tool uses a variation of the flood-fill algorithm where it starts for a point inside the selected region and spreads outwards till it finds the user selected boundary enclosing the region.



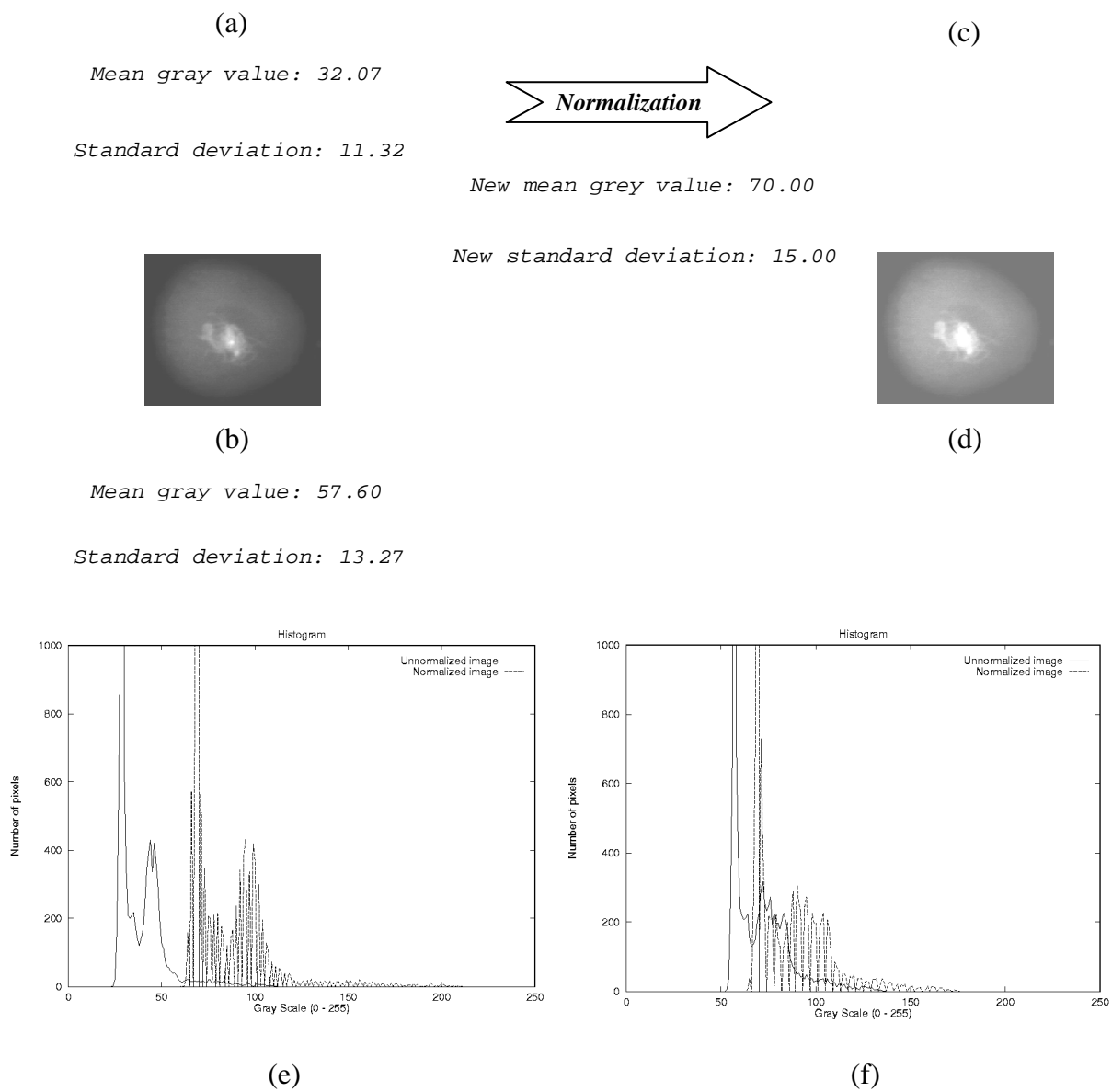


Figure 2.14: Images of two cells (a) and (b) in prophase I, exhibiting different levels of brightness. Images of these cells after normalization, (c) and (d), and their histogram plots before and after the normalization process, (e) and (f).

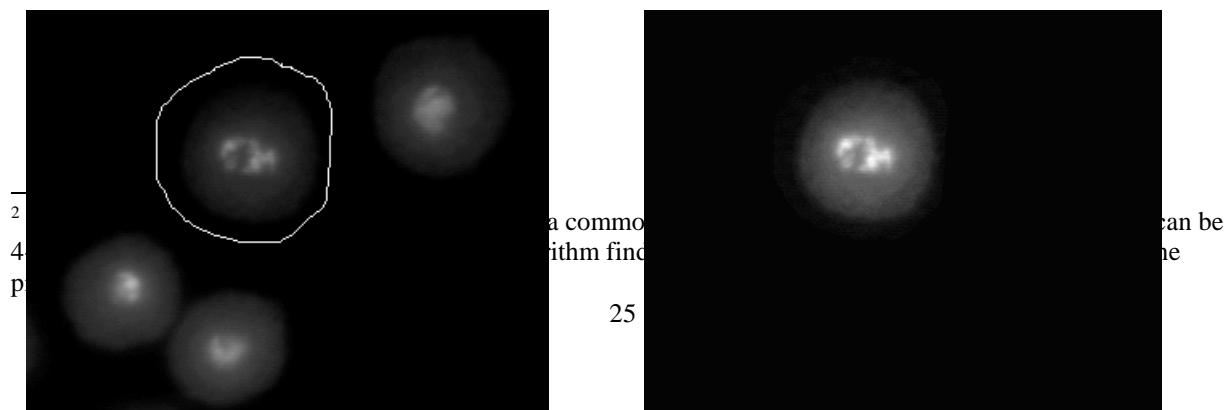
It retains the part lying inside the boundary and fills the rest of the image with a user supplied background intensity value. The drawback of this method is that it relies heavily on the user to

mark the region of interest around the cell. Techniques such as snakes (Witkin et. al., 1988) exist that automate much of the selection process and produce accurate regions of interest. Figure 2.15(a), shows an image containing a group of cells, with the cell of interest selected by the user. Figure 2.15(b), shows the image produced by the region isolation tool containing only the cell selected by the user.

2.2.5 Region Identification

The final step in the analysis of the cell images is the identification of the objects within the cells, the measurable properties of which can be used to distinguish cells of one type from the other. The objects that are of utmost importance for this sort of analysis are the cell nuclei and the chromosome fragments. These cellular entities undergo severe changes in shape, size and number during Meiosis. These changes can be measured quantitatively, and be used to characterize the events occurring during Meiosis.

The first step towards this is to analyze the histograms and intensity distribution plots of different cell type images to identify the grayscale ranges the defining objects fall in. Once the ranges are identified, the objects can be extracted from a cell image using a region extraction technique. A region extraction tool identifies regions of image pixels that share a common property. The common property in this case is the grayscale range occupied by the cell defining objects. The tool uses a blob growing² algorithm to extract regions of pixels from the image. Two types of regions are extracted: main regions that form groups of connected pixels exhibiting a common property; and internal regions that lie inside the main regions but do not exhibit the property. For every region extracted, the

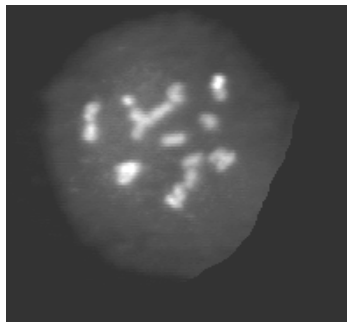


a common property can be identified with fine

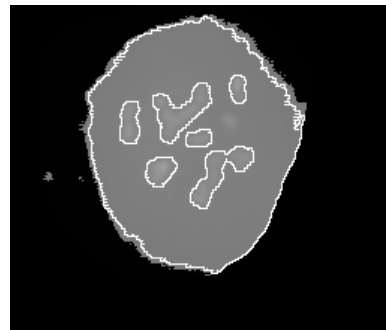
(a)

(b)

Figure 2.15: (a) An example of an image showing a group of cells with the cell of interest selected by a closed curve. (b) The resulting image generated by the region isolation technique containing only the selected cell.



(a)



(b)

Figure 2.16: (a) An image of a cell in prometaphase with its chromosome fragments visible. (b) The resulting image generated by the region extraction tool. The regions extracted are highlighted, with the outer region being the main region and internal regions approximating the chromosome fragments in the cell.

tool calculates a set of features that give an approximation of the shape and size of the region. These features are discussed in Chapter 3. Figure 2.16(a), shows an image of a cell in prophase

1. Figure 2.16(b), shows the regions identified by the extraction tool in the range 123 to 131 with their boundaries marked.

To summarize the entire analysis process and to give the order, in which the different techniques are applied, we reiterate the events again. The cells to be analyzed are first captured on digital images. Individual cells in an image are isolated from the rest of the group using region isolation tools. The images are normalized to a specific grayscale and standard deviation value to eliminate brightness discrepancies. Image histograms are studied to identify the grayscale ranges occupied by the cell defining objects like the chromosomes and the cell nucleus. These objects are then extracted from the images and their features measured to get an approximation of their shape, size and number.

Most of the image analysis process explained above is automated except for the image preparation and cell isolation parts that need human intervention. The image normalization and region extraction tool are part of a single system that takes as input the image of the isolated cell and analyzes it to generate descriptive features of regions extracted from the cell. These features are then fed to a cell classifier that classifies the cell according to its feature values. The details of the classifier are discussed in chapter 3.

2.3 Machine Learning

The cell classifier is based on the concepts of machine learning. The form of learning used in its construction is that of inductive learning.

2.3.1 Inductive Learning

Inductive learning is one of the basic mechanisms of machine learning (Quinlan, 1986). In this method, the learner is presented with examples of different types so that it generates a concept description for each type involved. These concept descriptions are then used to predict the types of examples that are not part of the training set. To illustrate this idea, consider the problem: Suppose a teacher wants the student to understand the concept of prime numbers. Instead of giving them the definition of a prime number, the teacher shows a list of numbers: 2, 3, 5, 7, 11, 13 etc. After seeing a large enough set, the students finally are able to conclude that a prime number is an integer greater than 1 that can only be divided by one and itself. This form of learning is known as inductive learning. In our case, the classifier is the learner. It is presented with example images of different cell types till a model characterizing each cell type is generated. The classifier then distinguishes between the cell images based on this model.

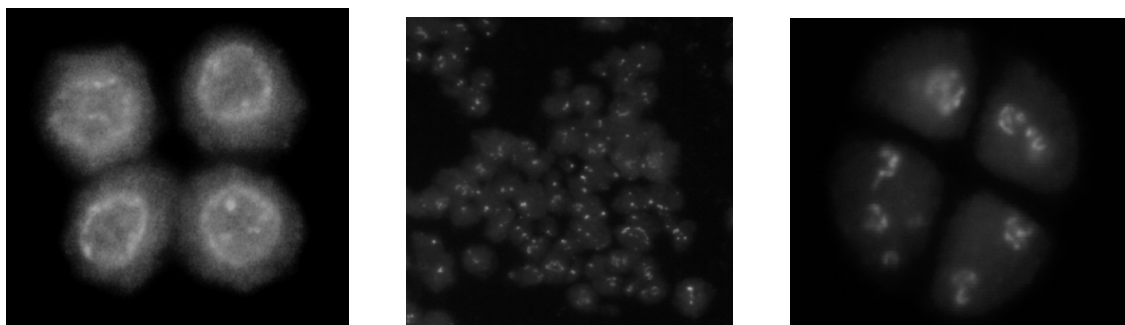
3 Characterizing Cells during Meiosis

This section discusses the methods used in building the classifier for the cell images. In doing so, it first presents the cellular features that were examined in the cell images and the results of these features observed in initial images. Based on these results, the features that were found useful are discussed and the resulting classifiers based on these features are presented. Finally, the results obtained from applying the classifiers to new images are discussed.

3.1 Features Examined

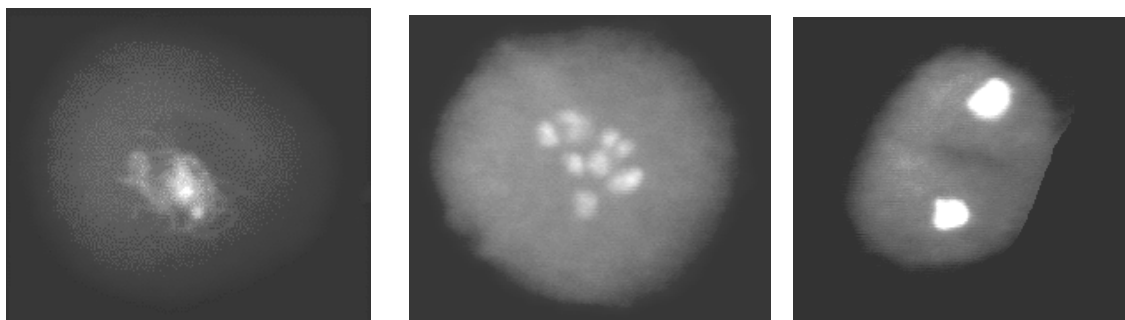
A classifier's ability to separate out objects of different classes is solely dependent on the distinguishing characteristics of its features. In order to produce an accurate classifier, it is necessary to identify good features in cell images. To give an example of this, consider the cell images shown in Figure 3.1 and 3.2. The cells shown in Figure 3.1 are of wild type, ms6 mutation and po mutation respectively. Those in Figure 3.2 are of four phases of Meiosis: prophase 1, prometaphase, telophase I and telophase II respectively. From a human observer's point of view, the cells exhibiting the mutations show two features characteristic of these mutations: (1) condensation of chromatin (po and ms6) and (2) large number of chromosome fragments (ms6). In the case of cells exhibiting meiotic phases, prophase I, telophase I and telophase II cells have one, two and four chromosome bodies respectively while prometaphase cells exhibit large number of chromosome fragments. The task then is to identify these features in a measurable way and use them for constructing the classifier.

The image analysis techniques discussed in Chapter 2 provide a framework for measuring these features. The histogramming tool identifies the grayscale range occupied by the cell nuclei and chromosome in a cell image, while the region extraction tool extracts these cellular objects by looking for pixel regions that fall in the specified grayscale range. The extracted objects are further analyzed to get an approximation of their shape, size and

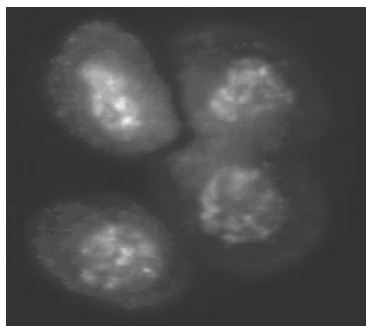


(a) (b) (c)

Figure 3.1: Images of cells exhibiting (a) wild-type, (b) ms6 mutation and (c) po mutation.



(a) (b) (c)



(d)

Figure 3.2: Images of cells exhibiting (a) prophase I, (b) prometaphase mutation, (c) telophase I, and (d) telophase II phases of Meiosis.

number. This is done by measuring the features of the extracted regions discussed below. These features are derived from earlier work on a system built to distinguish malignant cells from benign cells in breast tumor diagnosis (Wolberg et.al., 1993b). In that system, the features were mainly used to approximate the shape and size of the tumor cells. In our system, these features are used to measure the shape and size of the defining cellular entities (chromosomes and cell nucleus) of a cell.

The features are measured in both main and internal regions unless specified in their descriptions.

3.1.1 Feature: Number of Internal Regions

This feature counts the number of regions enclosed within the main region. As discussed in chapter 2, the region extraction tool finds two types of regions. The first type is the main region, which falls in the specified grayscale range. The second type is the internal region, which does not occupy the grayscale range but falls inside the main region.

3.1.2 Feature: Main Region Occupancy by Internal Regions

This feature measures the average percentage of main region area occupied by the internal regions. It gives an approximation of the average size of the internal regions with respect to the main regions.

3.1.3 Feature: Area

The area of a region is the pixel population of the region. In case of main regions, the region area is the number of pixels that fall within the region but lie outside the any internal regions contained in the main region. For internal regions, the feature is just their pixel population.

3.1.4 Feature: Perimeter

The number of pixels that fall on the periphery of the region. The perimeter pixels are those that form the boundaries of the region but are not a part of the region. The region boundary thickness is that of one pixel.

3.1.5 Feature: Radius

The radius of the region is measured by averaging the length of the radial line segments defined by the center of the region and the perimeter points.

3.1.6 Feature: Compactness

Perimeter and area are combined (Ballard et.al., 1982) to give a measure of the compactness of the region using the formula $\text{perimeter}^2/\text{area}$. This dimensionless number is minimum for a circular disk and increases with irregularity of the region.

3.1.7 Feature: Smoothness

The smoothness of a region contour is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it. Figure 3.3 shows the radial lines in the region used to compute smoothness.

3.1.8 Feature: Texture

The texture of the region is measured by finding the variance of the grayscale intensities in the component pixels.

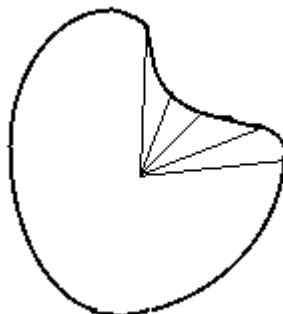


Figure 3.3: Radial lines of the region used for smoothness computation

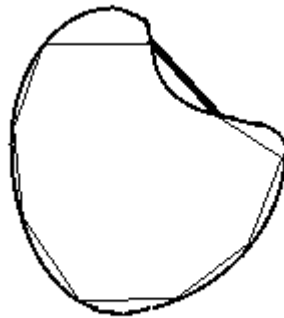


Figure 3.4: Region chords to compute concavity

3.1.9 Feature: Concavity

In a further attempt to capture the shape information, the number and severity of concavities or indentations in a region are measured. Chords are drawn between non-adjacent perimeter points and the extent to which the actual boundary of the region lies on the inside of the polygon formed by the chords is measured. Figure 3.4 shows the chords in the region used to calculate concavity.

3.1.10 Feature: Concave points

This feature is similar to Concavity but measures only the number rather than the magnitude of contour concavity.

3.2 Initial Results

The features described above were observed in a initial image data set comprised of fifteen images each of wild-type, ms6 mutation and po mutation cells and cells in prophase I, prometaphase, telophase I and telophase II stages of Meiosis. All images were normalized to remove brightness discrepancies, to a grayscale value of 127 and a standard deviation of 5.00. Histogram analysis revealed the cells to occupy a grayscale range of 123 to 255. Parts of the cells, excluding the chromosomes and cell nuclei, fell in the range of 123 to 131, while the chromosomes and the nuclei themselves occupied a range of 132 to 255. Regions of image pixels lying within the range 123 to 131 with their corresponding interior regions in the range 132 to 255 were extracted using the region extraction tool. The internal regions approximated the pixels occupied by the chromosomes and the cell nuclei in the images. Following are the results of the features observed on the internal regions extracted.

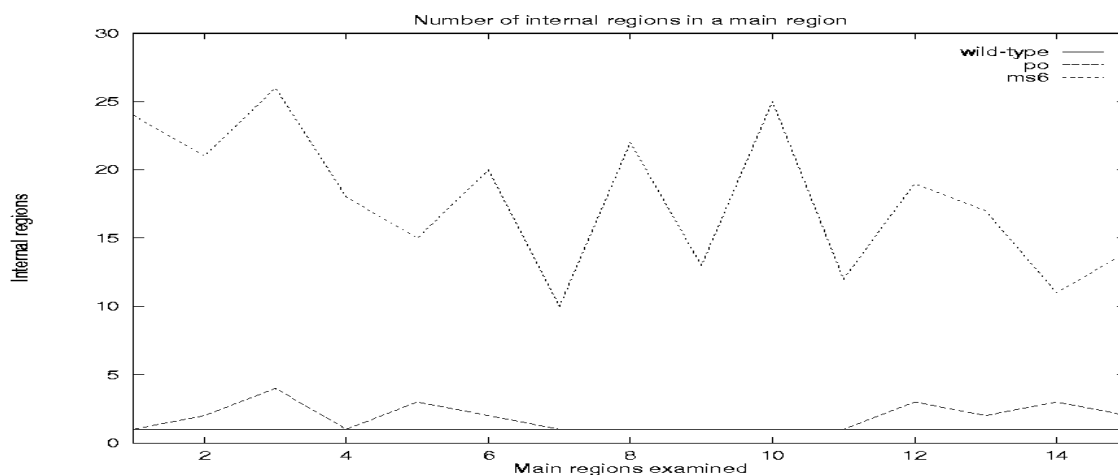
3.2.1 Result: Number of Internal Regions

The plot in Figure 3.5(a), shows that the number of internal regions in a wild-type cell is always one. This is due to the fact that wild-type cells have fully developed cell nuclei that do not exhibit fragmentation. Cells with po mutations have one or more internal regions. The number is one when the nucleus is condensed and greater when it is fragmented. Ms6 mutations exhibit many chromosome fragments and hence have a large number of internal regions.

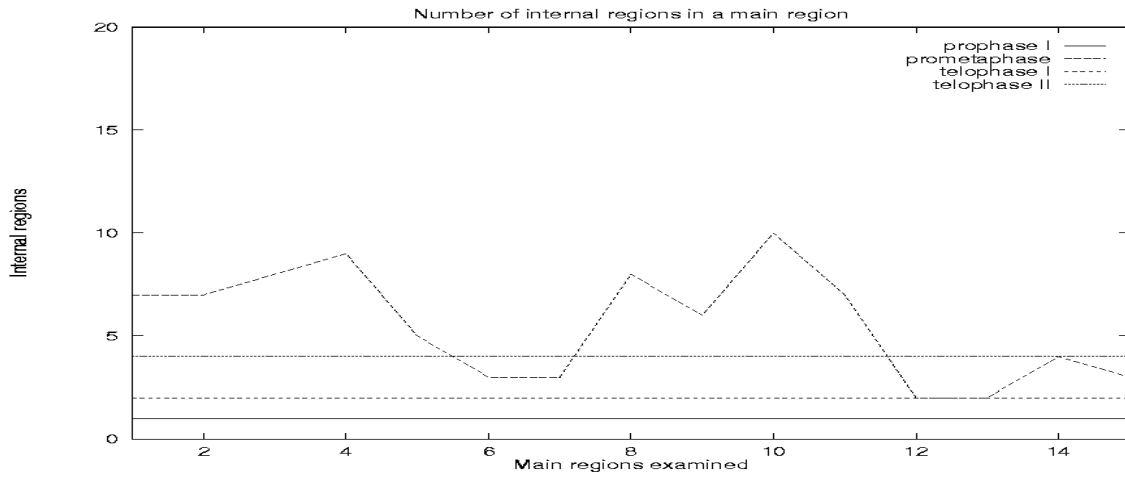
In case of the four phases of Meiosis observed for results, this feature brings out the defining aspects in each phase. The plot in Figure 3.5(b) shows the number of internal regions in prophase I cells is constantly one. Prophase I is characterized by the manifestation and condensation of chromosomes in the cell nucleus which are otherwise imperceptible. The one internal region found within the cell is the condensed chromatin. The cells in prophase I then move to prometaphase which is characterized by disintegration of the condensed chromatin into chromosome fragments. These chromosome fragments spread throughout the cell space and make up for the internal regions found in the cell. The number of internal regions varies from two to ten in the prometaphase cells observed. Telophase I is marked by the culmination of the first cell division of the Meiosis process. The cells have two condensed chromosome bodies that are held at the opposite poles of the cell. These two chromosome bodies are the internal regions found in the cell. Finally, telophase II marks the end of the second cell division with four well formed chromosome bodies stationed at the four ends of the cell. These cell bodies from the four internal regions in telophase II cells.

3.2.2 Result: Main Region Occupancy by Internal Regions

The plot in Figure 3.6(a), shows wild-type cells to have maximum main region occupancy values of the three cell types. This is because the nuclei in wild-type cells are well formed and make up for much of the cell space. In the case of po mutations, the

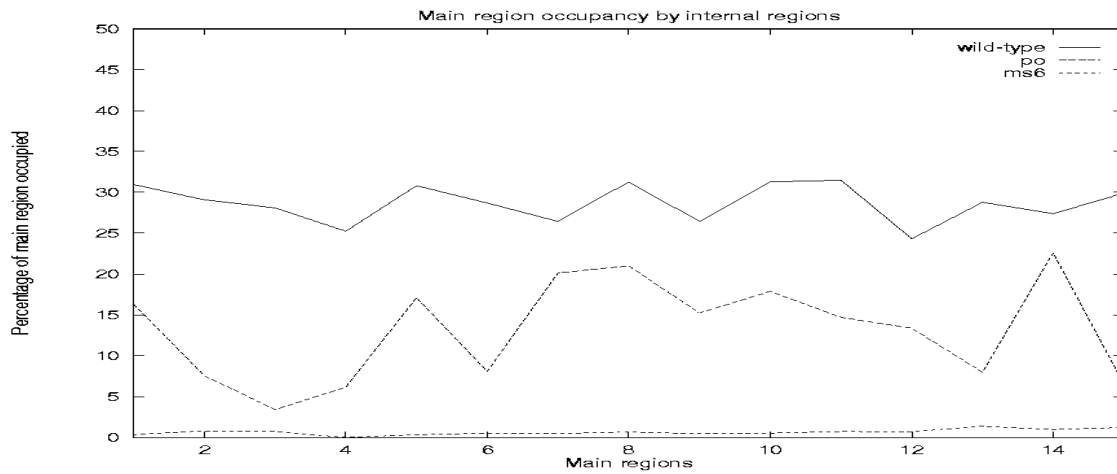


(a)

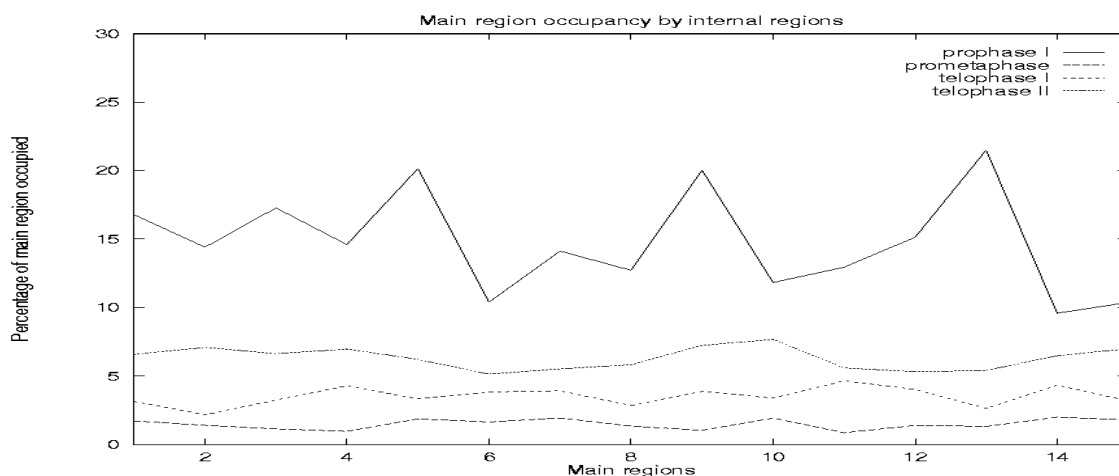


(b)

Figure 3.5: Plot of number of internal regions vs main region in (a) wild-type, ms6 and po mutation cells; (b) prophase I, prometaphase, telophase I and telophase II cells.



(a)



(b)

Figure 3.6: Plot of main region space occupancy by internal regions in (a) wild-type, ms6 and po mutation cells and (b) prophase I, prometaphase, telophase I and telophase II cells.

values are lower, because the chromosomes are condensed and often fragmented. The ms6 mutations have the minimum occupancy values of the three cell types due to the presence of large numbers of chromosome fragments, which is typical of ms6 mutations.

The average main region occupancy in case of the four meiotic phases studied is shown in Figure 3.6(b). Prophase I cells have maximum occupancy values. The one internal region found in a prophase I cell occupies much of the cell space. The main regions are smaller in comparison to the ones found in cells in other phases. This can be attributed to smaller size of cells in prophase I. Prometaphase cells have a large number of small chromosome fragments which give them lower occupancy values. Telophase I cells have two internal regions which

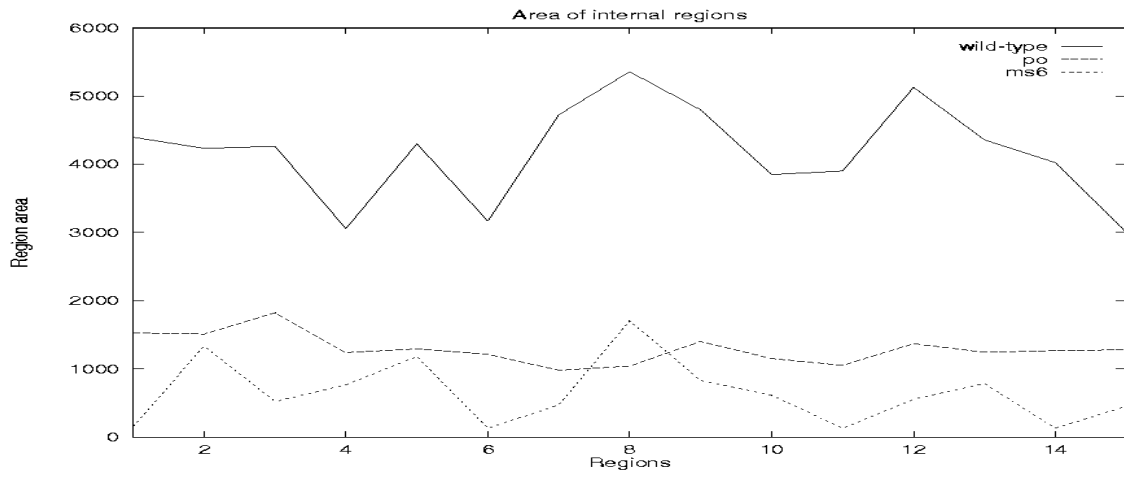
are much larger than those found in prometaphase. This gives them occupancy values higher than those in prometaphase. Telophase II cells have four internal regions with sizes on par with those found in prophase I. However, the main region sizes are greater, since they include all four cells formed in the phase. This leads to lower occupancy values as compared to those in prophase I.

3.2.3 Result: Area

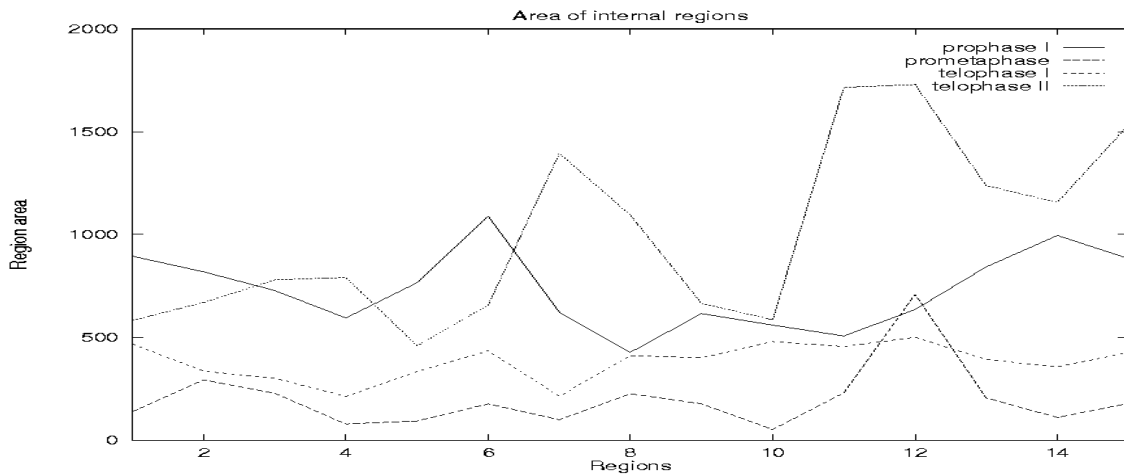
The results for this feature are consistent to the observations made in the preceding section. Figure 3.7(a) shows internal regions of wild-type cells have larger region areas than *ms6* and *po* mutation cell types. In the case of cells in the different phases of Meiosis, prophase I cells have internal regions with areas larger than prometaphase and telophase II, while telophase II internal regions have areas similar to prophase I cells. A few telophase II cells observed have areas way out of range which can be attributed to the differences in the magnification levels of the system used to capture the images.

3.2.4 Result: Perimeter

The results in Figure 3.8(a) and (b) can be attributed to the area difference of the internal regions in the three cell types.

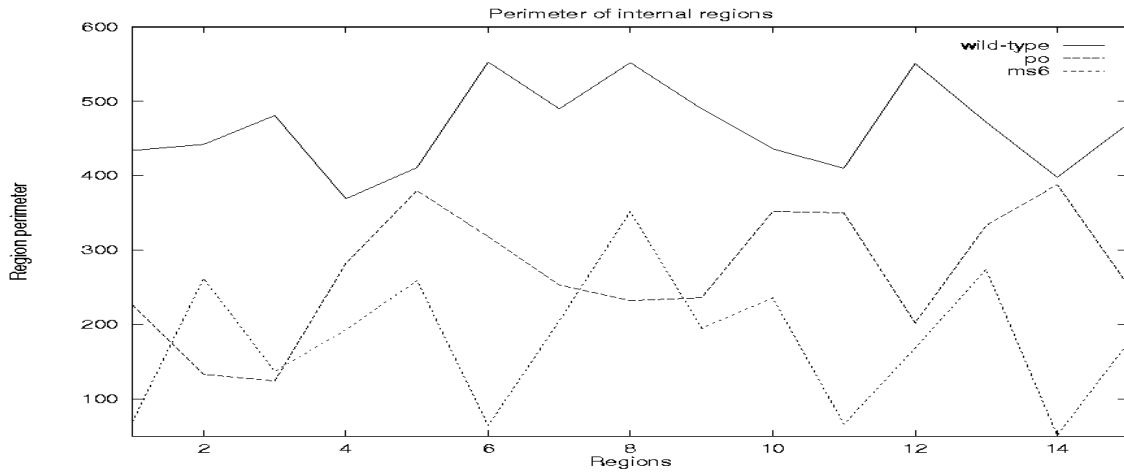


(a)

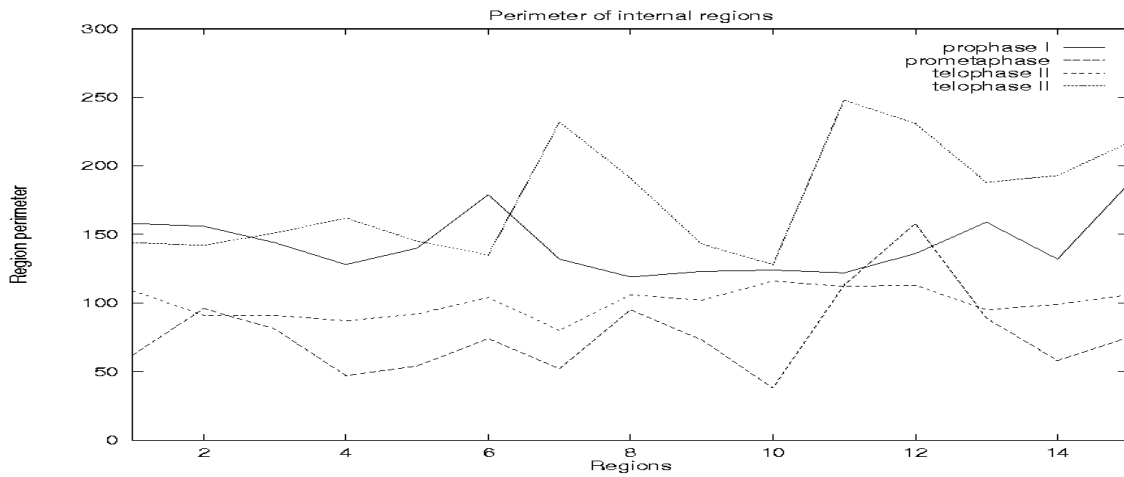


(b)

Figure 3.7: Plot of area of internal regions in (a) wild-type, ms6 and po mutation cells; and (b) prophase I, prometaphase, telophase I and telophase II cells.



(a)



(b)

Figure 3.8: Plot of perimeter of internal regions in (a) wild-type, ms6 and po mutation cells; and (b) prophase I, prometaphase, telophase I and telophase II cells.

3.2.5 Result: Radius

The values for region radius closely follow the values of area as is seen from the plot in Figure 3.9(a). The radius of internal regions in wild-type cells are relatively higher than those found in *ms6* and *po* mutant cells.

In the case of cells in the meiotic phases, radius values of internal regions in prophase I and telophase II cells are similar because of the similarity in the shapes and sizes of the nuclei found in these cell types. The values are greater than those found in prometaphase and telophase I cells. Telophase I cells have values higher than prometaphase cells because the two chromosome bodies found in telophase I cells are larger than the chromosome fragments in prometaphase.

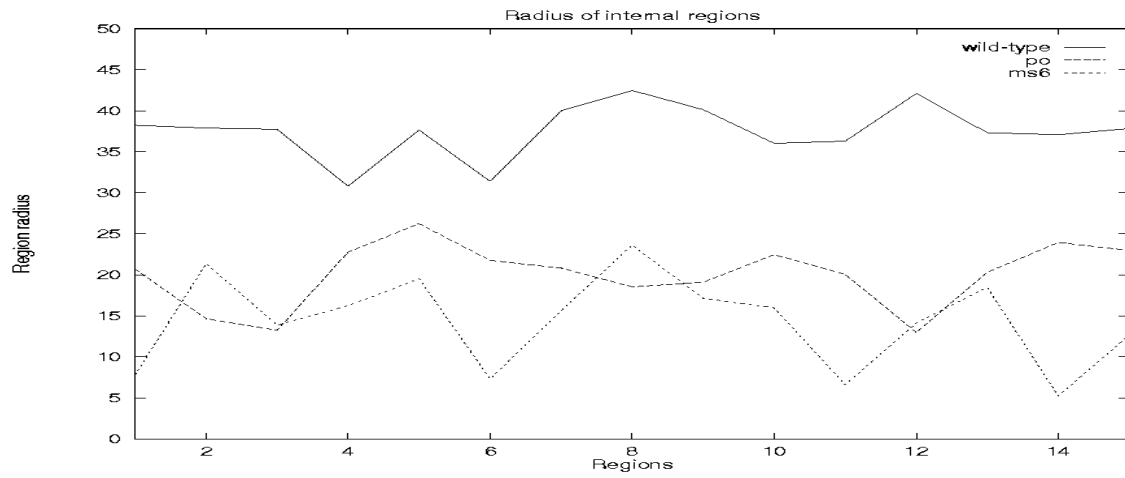
As with region area, region radius values are subject to discrepancies introduced by the images taken at different magnification levels.

3.2.6 Result: Compactness

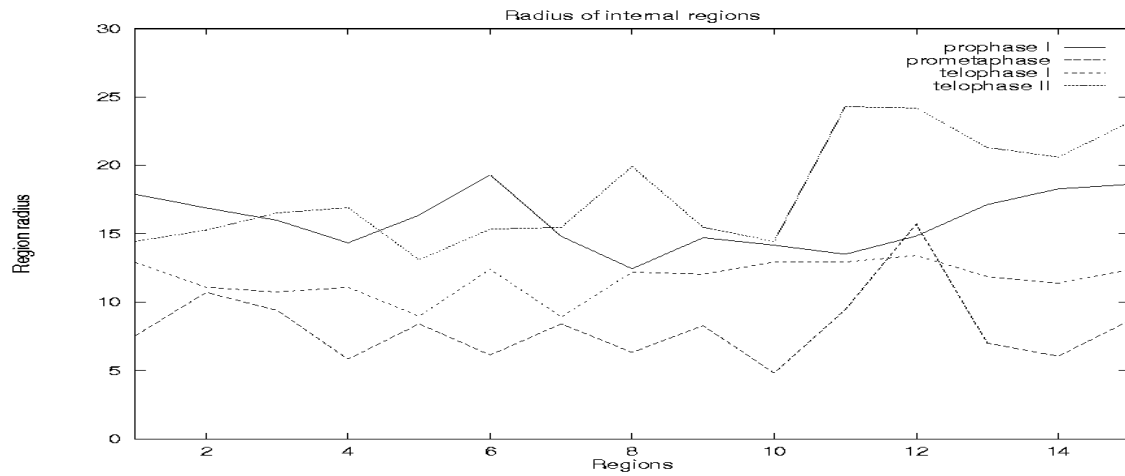
From Figure 3.10(a) and (b), it is evident that the compactness values of the internal regions do not exhibit any properties characteristic of the different cell types. This can be attributed to the irregularities in the shapes of the cell entities involved and the fact that the feature can only identify circular shaped regions.

3.2.7 Result: Smoothness

Region smoothness values from Figure 3.11(a) and (b) do not bring out any distinguishing feature in the cell types observed. The values are erratic which can be attributed to the roughness in the region boundaries and the unevenness in their shapes.

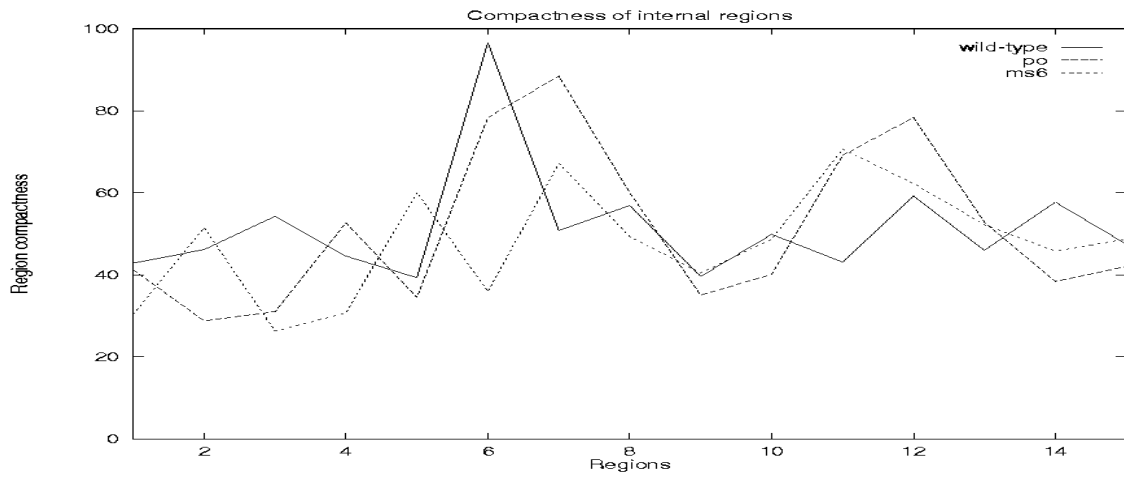


(a)

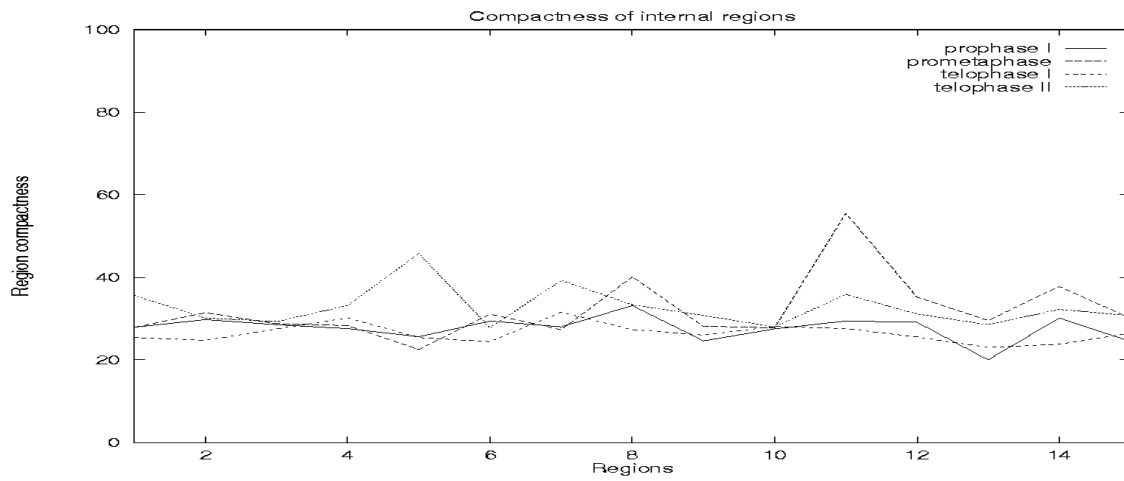


(b)

Figure 3.9: Plot of radius of internal regions in (a) wild-type, ms6 and po mutation cells; and (b) prophase I, prometaphase, telophase I and telophase II cells.

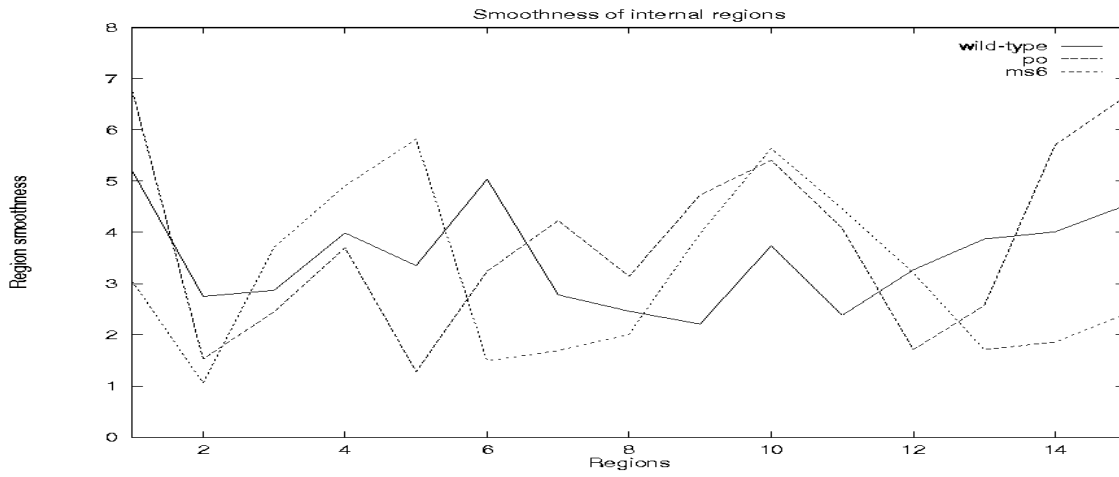


(a)

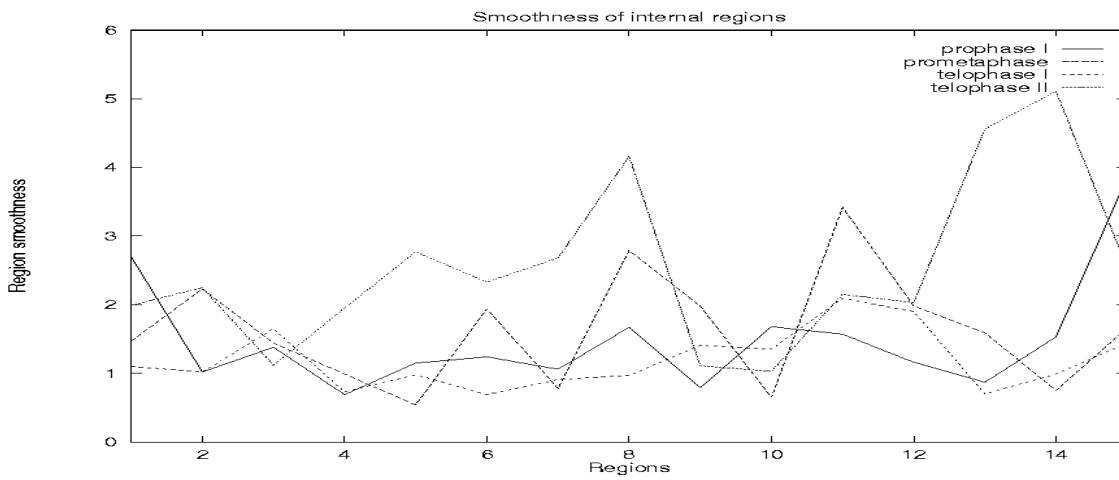


(b)

Figure 3.10: Plot of compactness of internal regions in (a) wild-type, ms6 and po mutation cells; and (b) prophase, prometaphase, telophase I and telophase II cells.



(a)



(b)

Figure 3.11: Plot of smoothness of internal regions in (a) wild-type, ms6 and po mutation cells; and (b) in prophase I, prometaphase, telophase I and telophase II cells.

3.2.8 Result: Texture

The plot in Figure 3.12(a) shows the texture values for the wild-type and mutant cells. From the plot, it can be seen that interior regions in wild-type cells have lower texture values than the two mutant cell types. This is because the pixels forming the cell nucleus in wild-type cells have low intensity variance than those forming the chromosome in po and ms6 mutant cells. This feature can be used as a distinguishing characteristic of wild-type cells from po and ms6 mutant cells.

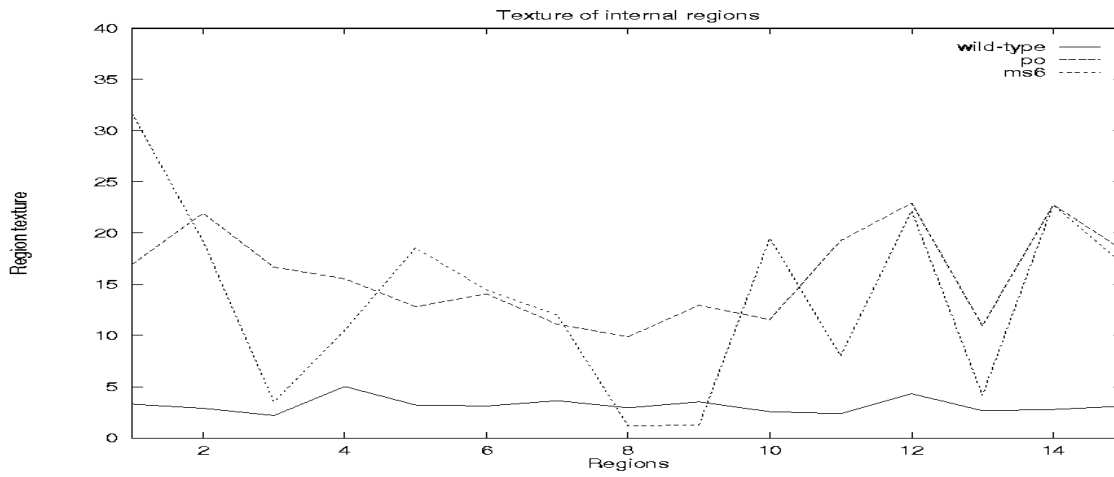
In case of meiotic cells in different phases, the values are not very distinctive of the phases as is evident from the plot in Figure 3.12(b).

3.2.9 Result: Concavity

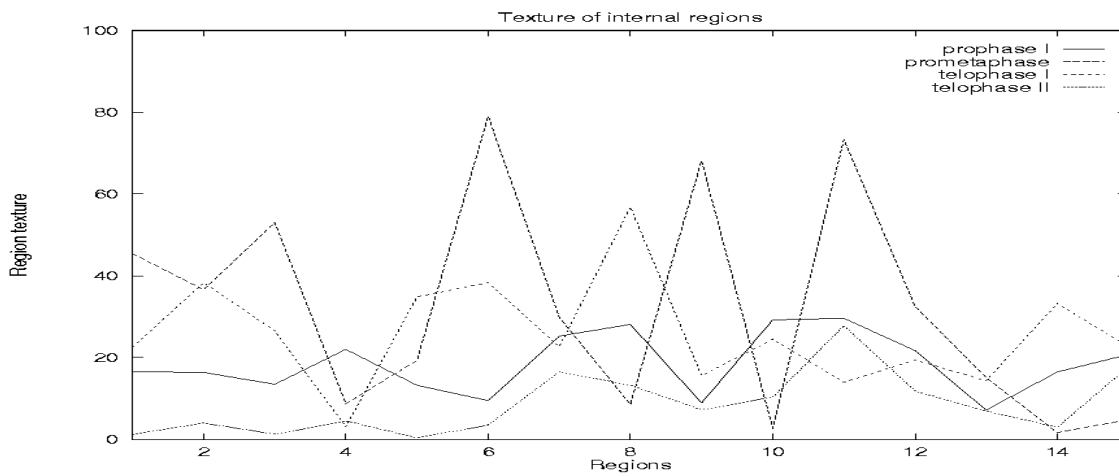
The concavity values from Figure 3.13 (a) and (b) do not seem to be very useful in terms of classifying the cells into the different cell types. The boundaries of the regions extracted show extreme roughness to produce any meaningful values for this feature. This can be attributed to the problem associated with the normalization process as discussed in Chapter 2. Normalization causes some of the pixels that lie on the boundaries of the region and would otherwise be part of the region to fall outside it giving the region a coarse exterior.

3.2.10 Result: Concave points

This feature, Figure 3.14 (a) and (b), measures the number of the contour concavities in a region. As with the concavity feature, this feature is vulnerable to the problems introduced by the coarseness of the boundaries of the extracted regions.

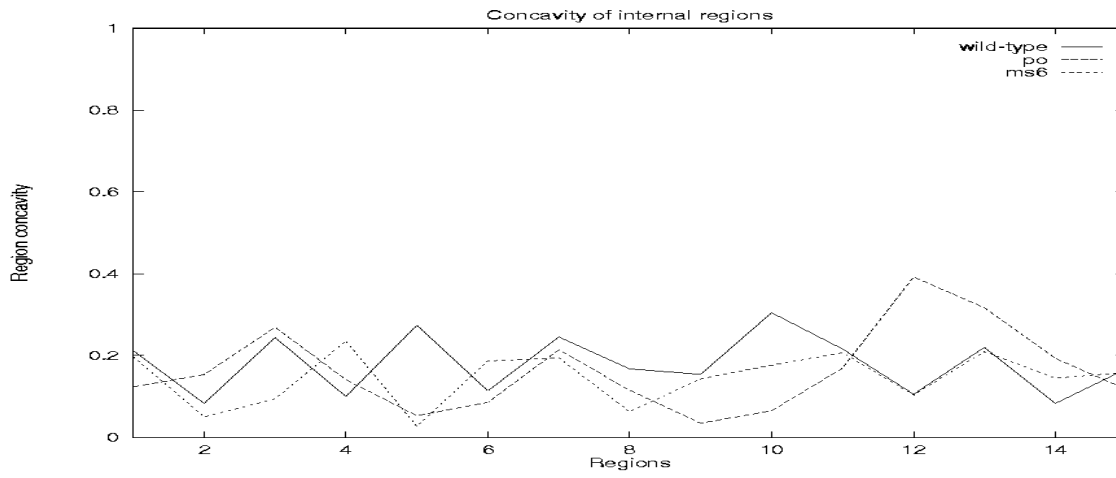


(a)

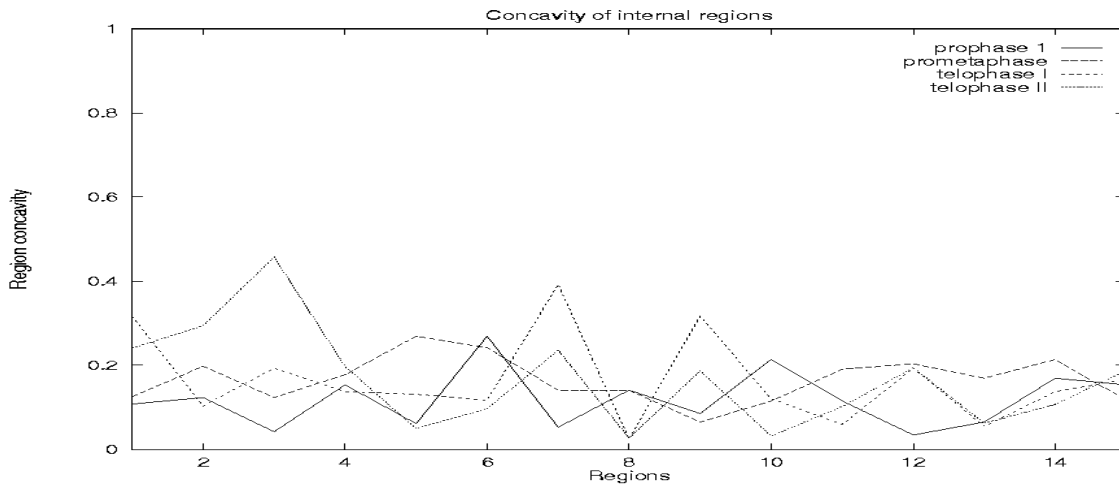


(b)

Figure 3.12: Plot of texture of internal regions in (a) wild-type, ms6 and po mutation cells; and (b) in prophase I, prometaphase, telophase I and telophase II cells.

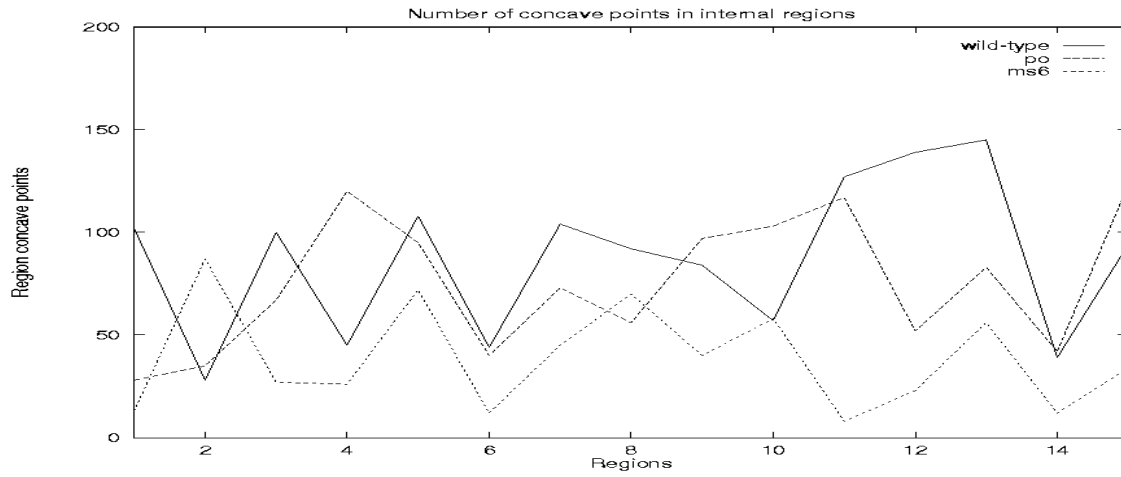


(a)

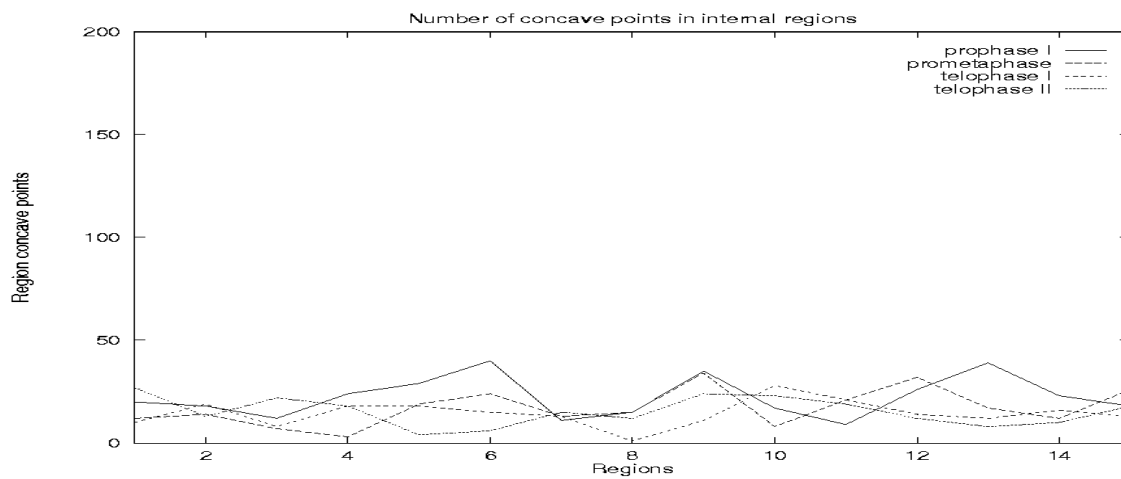


(b)

Figure 3.13: Plot of concavity of internal regions in (a) wild-type, ms6 and po mutation cells; and (b) in prophase I, prometaphase, telophase I and telophase II cells.



(a)



(b)

Figure 3.14: Plot of the number of concave points in (a) wild-type, ms6 and po mutation cells; and (b) in the internal regions in prophase I, prometaphase, telophase I and telophase II cells.

3.3 A Cell Classifier

From the above results, it can be seen that not all features examined are useful to the classifier construction process. Features like region compactness, smoothness, concavity and concave points do not seem to have any distinguishing characteristics for the different cell types and can be discarded. Of the remaining features examined, region area, radius and perimeter are seemingly useful, but are dependent on the cell size in the image. Images of cells are often taken at different magnification levels, which lead to proportional differences in the size of cells captured. These differences render features that depend on region size and pixel population useless for classification purposes and hence have to be discarded. The proportional differences in cell sizes can be removed by an image normalization technique, where all images are normalized to display their contents at one particular size and at the same level of detail. Features like area, perimeter and radius, which depend on image and cell size, can then be used as classification features. Unfortunately, the present working system does not incorporate this form of image normalization but can be extended to support it. Therefore, from the remaining features, the only ones that are independent of size discrepancies and bring out good distinguishing characteristics in cells are the region texture, number of internal regions and main region occupancy. Figures 3.15 and 3.16 show two classifiers based on the features found useful in the above analysis. The one in Figure 3.15, distinguishes between wild-type cells, po mutations and ms6 mutations while the other in Figure 3.16 distinguishes between cells in prophase I, prometaphase, telophase I and telophase II. It would be desirable to combine the two classifiers into a single classifier that distinguishes all the cell types involved. However, the current feature set is not sufficient to realize this goal.

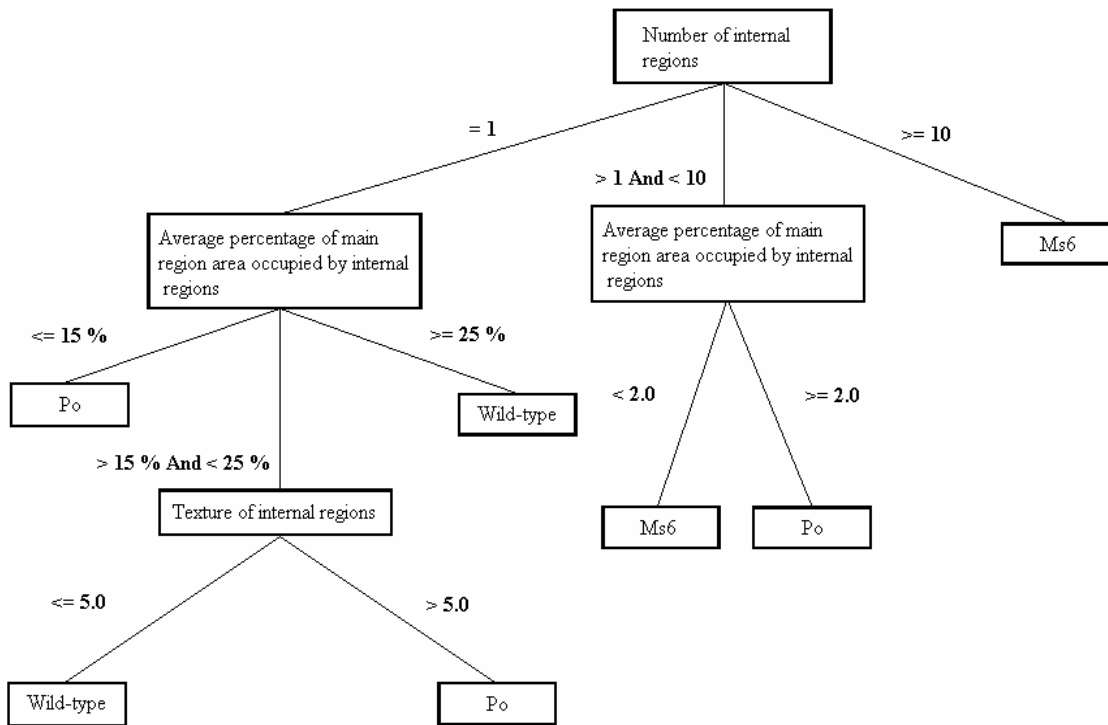


Figure 3.15: Classifier for wild-type, ms6 and po mutation cells.

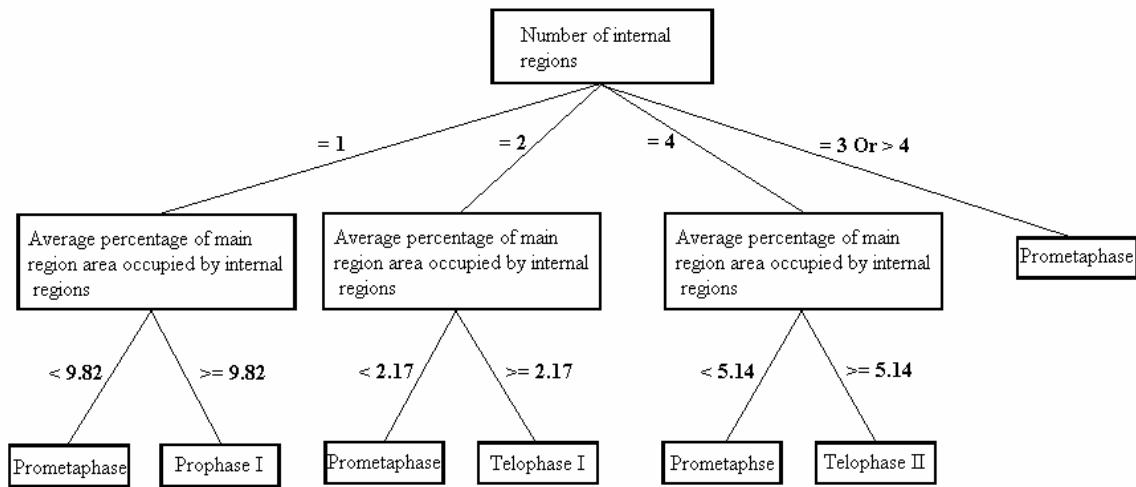


Figure 3.16: Classifier for cells in prophase I, prometaphase, telophase I and telophase II.

3.4 Test Results

The classifiers were tested on a test data set comprising of ten cell images each of wild-type, ms6 mutation and po mutation cells; and prophase I, prometaphase, telophase I and telophase II cells. The images were not a part of the data set used to build the classifier. Tables 1 and 2 show the results for the two classifiers respectively.

The classifier for wild-type and mutant cells could classify all the cell images tested correctly into their respective types as seen in Table 3.1.

Cell type	Proper classifications	Misclassifications
Wild-type (n = 10)	10	0
Po (n = 10)	10	0
Ms6 (n = 10)	10	0

Table 3.1: Test results of classifier for wild-type cells, ms6 and po mutation cells

Figures 3.17, 3.18 and 3.19 show the values seen in the cell images for the three features on which the classifier is built.

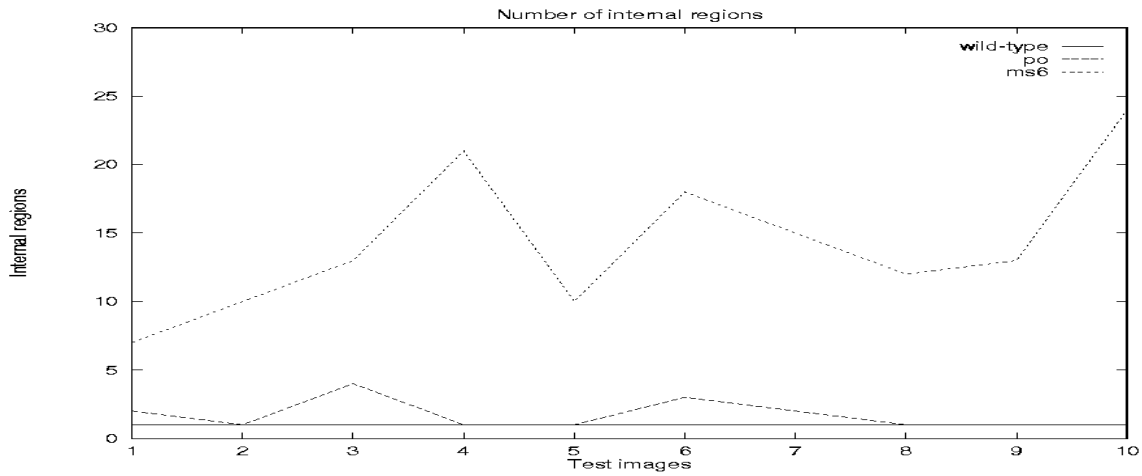


Figure 3.17: Plot of the number of internal regions in the test cell images of wild-type, po mutation, ms6 mutation cells.

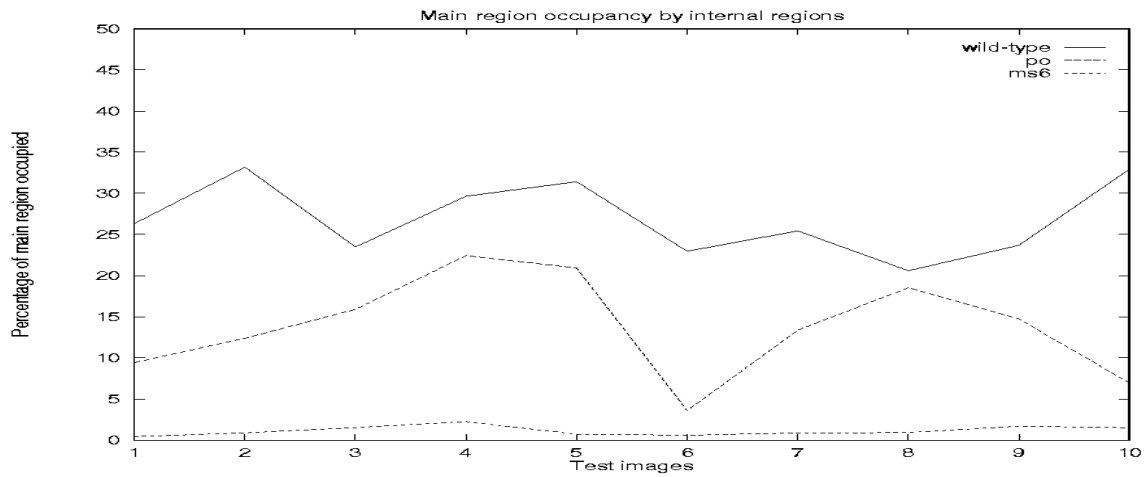


Figure 3.18: Plot of main region space occupancy by internal regions in test cell images of type wild-type, po and ms6 mutations.

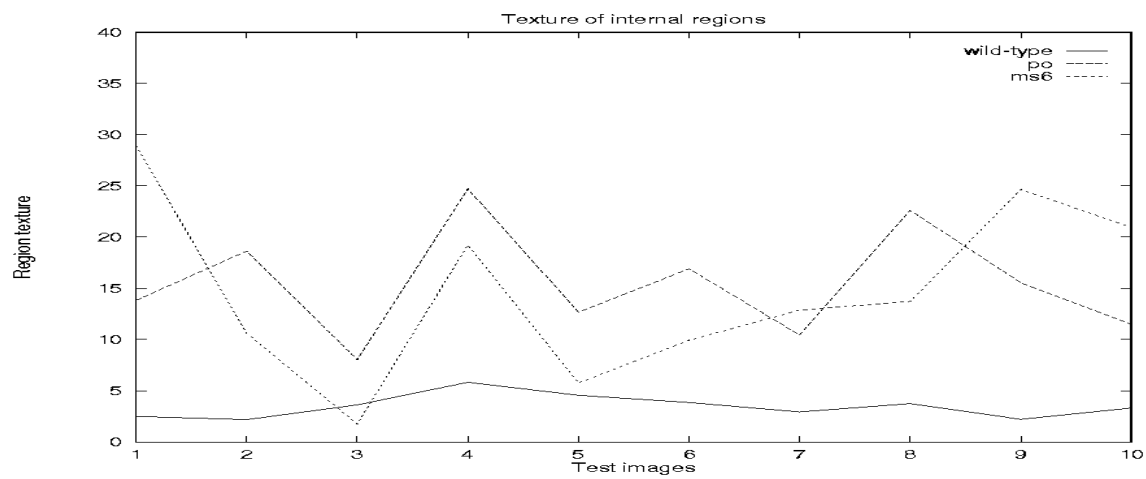


Figure 3.19: Plot of texture of internal regions in test cell images of type wild-type, po and ms6 mutations.

Table 3.2 shows the result of the classifier for prophase I, prometaphase, telophase I, and telophase II. The classifier was able to correctly distinguish all cell images of prophase I, telophase I and telophase II, except for two cell images of type prometaphase. Figures 3.20 and 3.21 shows the value plots of these images for the two classifier features.

Cell type	Proper Classifications	Misclassifications
Prophase I (n = 10)	10	0
Prometaphase (n = 10)	8	2
Telophase I (n = 10)	10	0
Telophase II (n = 10)	10	0

Table 3.2: Test results of classifier for prophase I, prometaphase, telophase I and telophase II cells.

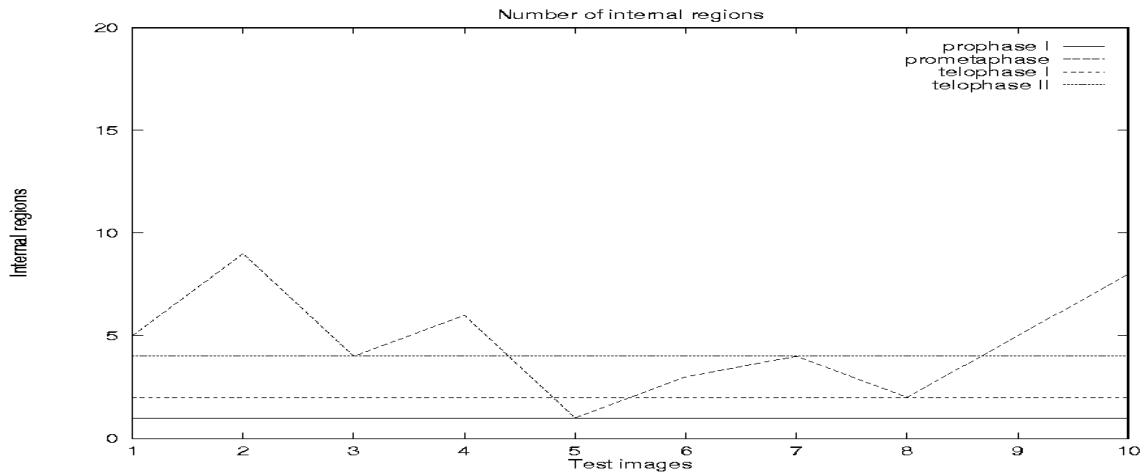


Figure 3.20: Plot of the number of internal regions in test cell images of prophase I, prometaphase, telophase I and telophase II cells.

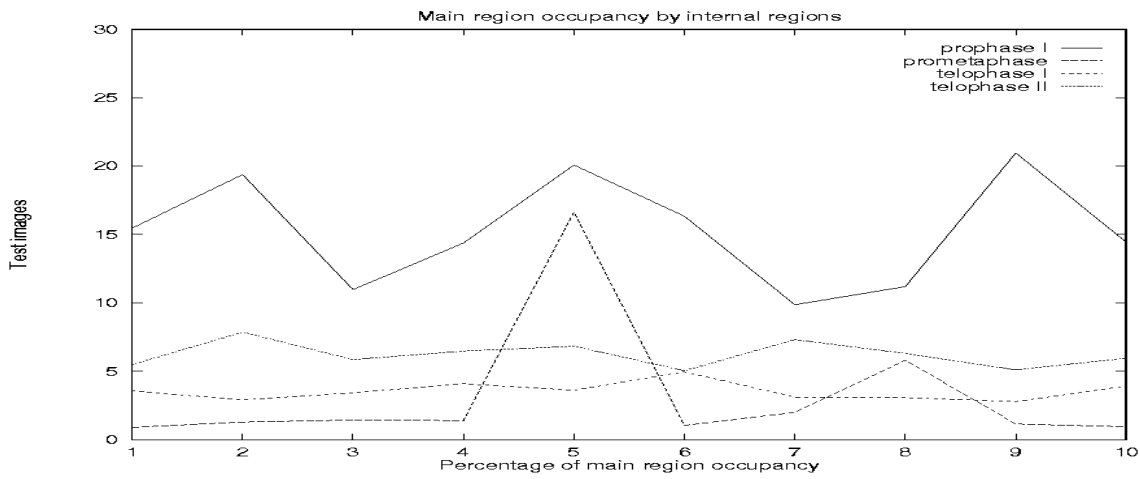


Figure 3.21: Plot of main region space occupancy by internal regions in test cell images of prophase I, prometaphase, telophase I and telophase II cells.

The two prometaphase images were classified as prophase I and telophase II respectively. The first image exhibited only one internal region with an average main region occupancy of 16.63 which is way into the range occupied by prophase I cells. The reason for this being that the chromosome fragments in the cell were clustered tightly together, which caused them to be extracted as a single region. The cell exhibited early stages of prometaphase where condensed chromosomes break into pairs and start spreading into the cell space away the cell center. The second image classified as telophase I, exhibited two internal regions which is characteristic of telophase I cells. The average main region occupancy was 5.80, which falls into the range occupied by telophase I cells. One cell feature that is unique to telophase I cells is the two chromosome regions are stationed at opposite ends of the cell. This feature is not seen in prometaphase cells with two chromosome fragments. Thus if the distances between the two regions can be measured, relative to their positions in the cell, this sort of misclassification can be avoided.

4 Future Work

This research focused on developing a system based on image processing and machine learning techniques to classify cells exhibiting different meiotic and post meiotic events. The system, though partially realized, is far from complete. The eventual goal of the system is to be able to completely characterize all the phases and events occurring during Meiosis and to automate the process of detection of the events in the cell images. Future work undertaken on the system will move in three different directions: event classification, image processing and machine learning.

4.1 Event Classification

Of the many cellular events occurring during Meiosis, the system is currently able to characterize only a handful, namely wild-type cells, po mutation and ms6 mutation of the post meiotic events; and prophase, prometaphase, telophase I and telophase II of the meiotic events. This can be attributed to lack of images exhibiting the different cellular events. The Meiosis process comprises of five more events, namely metaphase I, anaphase I, prophase II, metaphase II and anaphase II, which the system does not currently recognize and a plethora of meiotic mutations, some of which are yet to be discovered by cell biologists. So, as long as there are events to be discovered and classified, work on system will continue to persist. As mentioned before, the eventual goal for the system is to be able to automatically detect all the phases of Meiosis in cell images. In the case of cell mutations, the system may be able recognize them before they occur. A cell undergoes mutations as it deviates from the normal sequence of events governing the cell division process. The system should be able to detect these deviations early in a cell image and forecast the mutative path the cell would take.

4.2 Image Processing

In the area of image processing, future work will focus mainly on improving the system to incorporate better image analysis techniques and identifying new features in images to get a better approximation of the characteristics in different cellular events.

The first step towards improving the system is to replace the existing region isolation technique with one that is less dependent on user defined boundaries and more suggestive of the actual cell boundary. One such method is that of a snake (Witkin et.al.,1988), where a user defined curve around the cell converges to the actual boundary of the cell. A snake is a deformable spline, which seeks to minimize an energy function defined over the arclength of a closed curve. The energy function is defined in such a way that the minimum value occurs when the curve accurately corresponds to the boundary of the cell. The second improvement is to add another layer of image normalization to the system where images of cells taken at different magnification levels are normalized to display their contents at the same level of detail. This would allow features that are dependent on the cell size to be useful for classification purposes.

The feature set on which the system is currently based is certainly not comprehensive and can be extended to include additional features that reveal finer characteristics of the different cellular events. One example of this is the inability of the system to distinguish prometaphase cells, exhibiting only two chromosome fragments from telophase I cells. Telophase I cells have a distinctive characteristic of two condensed chromosomes placed at the opposite poles of the cell. If the distance between the chromosomes can be used as a feature, this form of misclassification can be avoided, since chromosome fragments in prometaphase cells are separated from each other by a shorter distance than the chromosome bodies found in telophase I.

4.3 Machine Learning

The machine learning part used in the current system can be said to be trivial to non-existent. The reason for this is the feature set used to build the classifier mainly comprised of a small number of features that captured good distinguishing characteristics in the cellular events. This fact obviated the need to extensively train the system to characterize the cellular events. However, in future, as the system is extended to cover more features and handle new cellular events, this observation will no longer be true, as the system would have to be trained to find the right combination of features that uniquely identify each cell type. Machine learning techniques will then play an important role in the system construction.

5 Conclusions

This thesis presents a system based on image processing and machine learning techniques to characterize cellular events occurring during the process of cell division Meiosis and to classify images of cells exhibiting these events. The events in question are the eight phases of Meiosis process: prophase I, metaphase I, anaphase I, telophase I, prophase II, metaphase II, anaphase II and telophase II; and post meiotic events such as *ms6* and *po* mutations that decide the phenotype of the resulting cells.

The system is based on extraction of features from cell images and construction of a classifier that distinguishes cell images of one type from another. The features are selected such that they are descriptive of the cells and give a good approximation of the shape, size and number of the cellular features such as chromosomes and nucleus found in cells. They are extracted using image analysis techniques including image histogram analysis, image normalization, region isolation and region extraction techniques. The construction of the classifier is treated as an inductive learning problem. Example images of different cell types are presented to the learner to determine a combination of features that distinguish examples of one cell type from another. The resulting classifier is then tested on cell images that were not part of the data set used to train the learner.

Two classifiers were developed as a result of this research. The first one characterized cells of wild-type, *po* and *ms6* mutations. The second classifier characterized cells of prophase I, prometaphase, telophase I and telophase II phases of Meiosis. The features constituting the classifiers revealed properties in cell images, a cell biologist would use for classification purposes. For example, the *number of internal regions* feature brought out the chromosome fragmentation characteristic in *po* and *ms6* mutations, which distinguishes them from the wild-type cells. Similarly, the *average main region occupancy by internal regions* feature separated the prometaphase cells from cells of other phases. The classifiers were tested on a data set consisting of ten images each of the different cell types involved. The results for the first classifier were quite impressive in the way of being able to correctly classify all the images of

the data set. The second classifier misclassified a couple of cell images of type prometaphase to prophase I and telophase I. These misclassifications can be attributed to the need of identifying more features to distinguish the cell types.

The system needs improvements and extensions in the form of employing better image analysis techniques and covering additional cellular events in its classification network. The system in its present state handles only a handful of cellular events occurring during Meiosis. The meiotic cell division process comprises of four additional phases and a large number of mutations. The eventual goal of this research is for the system to be able to automatically detect all the phases of the process and recognize cell mutations before they occur. In terms of this final goal, the thesis lays the foundations to such a comprehensive system.

References

- [Albert et.al.,1983] Albert, B.; Bray, D.; Lewis, J; Raff, M; Roberts, K; and Watson, J. D. 1983. *Molecular Biology of the Cell*. Madison Avenue, NY: Garland Publishing.
- [Ballard et.al.,1982] Ballard, D., and Brown, C. 1982. *Computer Vision*. Englewood Cliffs, NJ: Prentice Hall.
- [Bruenger, 1991] Bruenger, A. T. 1991. Simulated annealing in crystallography. *Annu. Rev. of Phys. Chem.* 42:197-223.
- [Dawe et. al., 1994] Dawe, R. K.; Sedat, J. W.; Agard, D. A.; and Cande, W. Z. 1994. Meiotic chromosome pairing in maize is associated with a novel chromatin organization. *Cell* 76:901-902.
- [Holley et.al., 1989] Holley, L., and Karplus, M. 1989. Protein structure prediction with a neural network. *Proceedings of the National Academy of Sciences (USA)* 86:152-156.
- [Maidak et.al., 1996] Maidak, B.; Olsen, G. J.; Larsen, N.; Overbeek, R.; McCaughey, M. J.; and Woese, C. R. 1996. The ribosomal database project (RDP). *Nucleic Acids Research* 24:82-85
- [Nilges et.al, 1991] Nilges, M.; Habezettl, J.; Bruenger, A. T.; and Holak, T. A. 1991. Relaxation matrix refinement of the solution structure of squash trypsin inhibitor. *Journal of Molecular Biology* 219:499-510.
- [Qian et.al., 1988] Qian, N., and Sejnowski, T. 1988. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology* 202:865-884.
- [Quinlan, 1986] Quinlan, J. 1986. Induction of decision trees. *Machine Learning* 1:81-106.
- [Rost et.al., 1993] Rost, B. and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology* 232:584-599.
- [Rumelhart et.al., 1986] Rumelhart, D.; Hinton, G.; and Williams, R. 1986. Learning internal representations by error propagation. In Rumelhart, D., and McClelland, J., eds., *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press. 318-363.
- [Socci et.al., 1996] Socci, N. D.; Onuchic, J. N.; and Wolynes, P. G. 1996. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.*

- [Turner et.al., 1993] Turner, M.; Austin, J.; Allinson, N.; and Thompson, P. 1993. Chromosome location and feature extraction using neural networks. *Image and Vision Computing* 11:235-239.
- [Warshal et.al., 1991] Warshel, A., and Aqvist, J. 1991. Electrostatic energy and macromolecular function. *Annu. Rev. Biophys. Chem* 20:267-298.
- [Watson, 1990] Watson, J. 1990. The Human Genome Project: Past, present, and future. *Science* 248:44-48.
- [Wied et.al., 1989] Wied, G.; Bartels, P; and et al., M. B. 1989. Image analysis in quantitative cytopathology and histopathology. *Human Pathology* 20:549-571.
- [Witkin et.al., 1988] Witkin, A., Kass, M., and Terzopoulos, D. 1988. Snakes: Active contour models. *International Journal of Computer Vision*. 1(4):321-331.
- [Wittekind et.al., 1987] Wittekind, C., and Schulte, E. 1987. Computerized morphometric image analysis of cytologic nuclear parameters in breast cancer. *Anal. Quant. Cytol. And Hist.* 9:480-484.
- [Wohlberg et.al., 1993a] Wohlberg, W. H.; Street, W. N.; Mangasarian, O. L. 1993. Breast cytology diagnosis via digital image analysis. *Analy. Quant. Cytol. And Hist.* 15:396-404.
- [Wohlberg et.al., 1993b] Wohlberg, W. H.; Street, W. N.; Mangasarian, O. L. 1993. Nuclear feature extraction for breast tumor diagnosis. *International Symposium on Electronic Imaging: Science and Technology*. 1905:861-870.
- [Wohlberg et.al., 1995] Wohlberg, W. H.; Street, W. N.; Mangasarian, O. L. 1995. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analy. Quant. Cytol. And Hist.* 17:77-87.