CHAPTER 4

PARAMETER IDENTIFICATION

This chapter discusses the identification of parameters necessary for the classification process and various methods to obtain these parameters. The results of these methods are discussed in Chapter 5.

4.1 INTRODUCTION

Segmentation and separation processes are useful in separating the 46 individual chromosomes in the metaspread image. Separated chromosomes are arranged in descending order of numbers and are known as Karyotype. Karyotype helps in detecting the abnormalities in chromosomes. As stated already, this requires identification of certain parameters. The parameters involved are the centromere position, length of the chromosomes and centromere index (CI).

Various methods are proposed for centromere identification. They include intensity integrated Laplacian based method (Akila Subasinghe Arachchige et al 2010), centromere identification based on concave points (Mohammad Reza Mohammadi 2012), shape and intensity profile based algorithm (Xingwei Wang et al 2009), rule-based computer scheme (Xingwei Wang 2008), projection vector (Mehdi Moradi 2003), characteristic band for length and centromere index (Mehdi Moradi 2006), multistaged algorithm (Xiaoli Yanga et al 2013) which uses discrete curve evolution, gradient vector
flow using active contours, functional approximation of curve segments and support vector machine.

4.2 CENTROMERE POSITION

Centromere of a chromosome appears as a constriction in the body of the chromosome and has a significant role in the separation of chromosomes into daughter cells during mitotic cell division. Cells resulting from the replication and division of a single parent cell are called as daughter cells. Each chromosome has only one centromere. Centromere position is one of the important parameters for classification. Three different algorithms are applied in this work for identifying the centromere. They are:

- Medial axis transformation
- Projection vector
- Concave Function

4.2.1 Medial Axis Transformation (MAT)

The output of the MAT algorithm are skeleton images where a one dimensional view of the images is obtained. MAT proceeds in a way similar to a fire line which is propagating from the contour of a connected object towards the inside of the object. All the points which are lying in a position where at least two wave fronts of the fire line meet during the propagation give the skeleton of the image. This forms the medial axis of the object. MAT is mainly used in pattern classification applications. The steps in MAT algorithm are shown in Figure 4.1.
Shape of the object is the input and is initially converted to binary image. The input image is an individual chromosome image. The distance transform is applied on the binary image. The location of the centromere is obtained by drawing a profile perpendicular to the transformed image.

### 4.2.1.1 Image binarization

Pixels in a region can be grouped based on their intensity. Segmentation of such regions is done using thresholding which separates the regions into light and dark.

Thresholding creates binary images from gray scale images. Thresholding helps to replace all the pixels in the foreground with a value 1 (white) and all other pixels with a value 0 (black). Let \( g(x, y) \) be the threshold version of \( f(x, y) \) at some threshold \( T \). Mathematically, the threshold version is represented by Equation (4.1)

\[
g(x, y) = \begin{cases} 
1 & \text{if } f(x, y) \geq T \\
0 & \text{otherwise}
\end{cases} 
\]  

(4.1)

### 4.2.1.2 Distance transform on binary image

The perspective of applying distance transform on an image is region based shape analysis. Distance transform labels each object pixel \( p \) by
a distance close to the point \( q \) in the background. It is used for computing i) width measurement ii) differential estimator and iii) medial axis or object skeleton. The distance transform gives a metric or a measure of separation of the points in the image.

**Discrete Metrics**

The metrics are defined based on integer values as i) distance transform based on sequences of chamber masks ii) Displacement based distance transform iii) distance transform based on the Euclidean distance.

For the given pixel \( p,q \) and \( r \) any metric must satisfy the following properties

- \( d(p,q) \geq 0 \) and \( d(p,q) = 0 \text{ iff } p = q \)
- \( d(q, p) = d(p, q) \)
- \( d(p,r) \leq d(p,q) + d(q,r) \)

There is a tradeoff between the approximation of the Euclidean distance and the mask size.

The distance transform is based on the sequence of chamber masks and requires updation of weights with new optimization process whenever there is a change in the shape of the pixel, mask size and image dimension. Displacement based distance transform requires the coordinate values to be stored for achieving the error free output. Distance transform based on Euclidean distance is preferred as it gives error free outputs without any constraints.
There are three types of distance metrics or measures in digital geometry.

**Euclidean distance:** A straight-line distance between two pixels is known as the Euclidean distance and is represented by the Equation (4.2)

\[ d_{euclidean} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \]  

(4.2)

**City block distance:** Measures the path between the pixels based on 4-connected neighborhood. Pixels are 1 unit apart touching the edge and pixels diagonally are 2 units apart and represented by the Equation (4.3)

\[ d_{city} = |x_2 - x_1| + |y_2 - y_1| \]  

(4.3)

**Chessboard distance:** Measures the path between the pixels based on 8-connected neighborhood. Pixels touching on edges or corners are 1 unit apart and represented by the Equation (4.4)

\[ d_{chess} = \max(|x_2 - x_1|, |y_2 - y_1|) \]  

(4.4)

The representation of transformed metrics for the image metrics is given below:

<table>
<thead>
<tr>
<th>Distance Metrics/Measure</th>
<th>Illustration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean distance</td>
<td><img src="image.png" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>Distance Transform</td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.41</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.41</td>
</tr>
</tbody>
</table>


The distance metrics profile for a chromosome image is shown in Figure 4.2. From the profile it is understood that Euclidean distance gives better output compared to the city block and chessboard. Euclidean distance transform is better because it gives a smoother function with a global minimum point. Obtaining a global minimum point is an important issue in many of the applications. Euclidean distance transform gives optimal and error free output.

<table>
<thead>
<tr>
<th>City block distance</th>
<th>Chessboard distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
</tr>
</tbody>
</table>

**Figure 4.2 Distance Metrics Profile for a Chromosome Image**

### 4.2.1.3 Density and intensity profiles

Two different types of profiles are obtained namely density profile and intensity profile. The density profile determines the average gray scale...
value of every perpendicular line across the medial axis of the chromosome. It is computed by the Equation (4.5)

\[ D(x) = \left[ \frac{\sum_{i=1}^{n} g_i(x)}{n} \right] \tag{4.5} \]

where \( g_i(x) \) is the gray value of each pixel in a perpendicular line, and \( n \) is the number of all pixels in each perpendicular line. Intensity profile records the weighted width of every perpendicular line across the medial axis of a chromosome \((x)\). It is defined by the Equation (4.6)

\[ s(x) = \sum_{i=1}^{n} \left[ g_i(x) \times d_i(x)^2 \right] / \sum_{i=1}^{n} d_i(x)^2 \tag{4.6} \]

which corresponds to the sum of the product of the gray scale values \( g_i(x) \) and their corresponding Euclidean distance measures \( d_i(x) \) away from the medial axis of the perpendicular line, divided by the sum of the distance.

The density and intensity profiles help in obtaining the centromere position. The distance measure of this portion of the chromosome is less compared to other portions and is useful in getting the global minimum point. In the profile shown in Figure 4.2, the point where the global minimum exists is considered as the centromere position.

### 4.2.2 Projection Vector

A set of projections is used to obtain the information of the 2-D binary image. Pixels along a column, row, or diagonal of the binary image are summed up to get the projection. There are three types of projections i) horizontal projection ii) vertical projection and iii) diagonal projection. Here, horizontal and vertical projections are considered. The horizontal projection
$H(i)$ is defined as the vector sum of rows and vertical projection $V(j)$ is the vector sum of columns of an image matrix and is represented by the Equation (4.7)

$$H(i) = \sum_{j=0}^{n-1} B(i, j) \quad \text{and} \quad V(j) = \sum_{i=0}^{n-1} B(i, j)$$ (4.7)

Horizontal and vertical projections help in obtaining the image size, position and orientation. The algorithm for performing the projection vector is given below

- Obtain the histogram of the image
- Perform image binarization.
- Obtain the projection vectors
- Identify the centromere position.

### 4.2.2.1 Histogram of the image

The input image is initially converted to gray scale image. A histogram plot is obtained for the gray scale image. The histogram for any given image using probability density function is given by the Equation (4.8)

$$P(X_k) = \frac{n_k}{n} \quad k=0,1, \ldots, L-1$$ (4.8)

where $n_k$ is the number of occurrences of a gray value; $X_k$ is the intensity value of the image; ‘$X$’ and ‘$n$’ are the total number of samples in input image; $P(X_k)$ is the number of pixels that have specific intensity $X_k$.

The histogram is obtained by splitting the data range into equal sized classes. For each class the number of points are counted which falls into
the same class. The plot is between the frequency (nothing but the count of each class) of the classes and the variable which is to be tested. The image histogram is plotted with the intensity of the pixel along the x-axis and frequency of pixel along the y-axis.

The histogram plot is useful in the classification of chromosome images. The information on the counts of the classes are available in the histogram. The histogram corresponding to the image $f(x,y)$ has a dark object in a light background. The object and the background have clustered gray levels. To extract the object from background, a threshold value is sufficient. This separates the two modes.

Then the points in $(x,y)$ for which $f(x,y) > T$ are known as object points and the remaining points are the background points. Filtering is done on the histogram to remove the noise.

The filtered histogram is obtained using the Equation (4.9)

$$G(x) = \exp(x^2 / 2\sigma^2)$$

(4.9)

where $x$ is the coordinate relative to the center of the kernel and $\sigma$ is the standard deviation. After obtaining the binary image, projection vector is applied.

4.2.2.2 Projection vectors

The projection vectors are helpful in obtaining the morphological information of any image. The horizontal, vertical and diagonal projection vectors are shown in Figure 4.3. From the figure, it is understood that projection helps in getting the total number of white and black boxes along the row, column and diagonal.
This projection vector calculation is useful in identifying the centromere position in the chromosome image. The centromere is always identified as the narrowest part of a chromosome in the longitudinal direction. The centromere has the minimum width when the boundary is obtained for a chromosome. It is the global minimum point in the chromosome. Projection vector calculation also helps in obtaining this global minimum point.

![Horizontal Projection](image1.png) ![Vertical Projection](image2.png) ![Diagonal Projection](image3.png)

**Figure 4.3 Projection Vectors (Horizontal, Vertical and Diagonal)**
Source-http://ee.lamar.edu/gleb/dip/index.htm

### 4.2.3 Concave Function

Concave regions are identified based on angle and curvature calculations. The concave region is the region of maximum negative curvature as shown in Figure 4.4.
The steps in the algorithm include:

- Take Karyotyped image as input image
- Perform image binarization
- Obtain contour image
- Identify the concave region
- Obtain the weighted shortest path
- Obtain the centromere position

The karyotyped chromosome image is taken as input. The initial step is image binarization and is discussed in section 3.2.1. Contour detection is followed by image binarization. The contour detection is discussed in section 3.2.2. After finding the contour the concave points are identified along the contour.

4.2.3.1 Identification of concave region

The detection of concave region is necessary for identifying the centromere position. As stated earlier, the narrowest portion of the connected component is the centromere. Therefore the concave region is intuitively the
most important portion in identifying the centromere. The concave region identification is obtained using the cost function between the two points which is mathematically represented by the Equation (4.10)

$$E_s(A, B) = \frac{\text{dist}(A, B)}{\min\{\text{length}(A, B), \text{length}(B, A)\}}$$  \hspace{1cm} (4.10)

where \( A \) and \( B \) represent the two different points on the contour. \( \text{dist}(A, B) \) indicates the Euclidean distance between points \( A \) and \( B \). \( \text{length}(A, B) \) represents the clockwise length from \( A \) to \( B \) on the contour image. \( \min\{\text{length}(A, B), \text{length}(B, A)\} \) represents the smaller distance between the two lengths. The process is done on the closed contour. The minimized cost function for identifying the concave region is represented by the Equation (4.11)

$$\left(A^*, B^*\right) = \arg\min_{A, B} E_s(A, B)$$  \hspace{1cm} (4.11)

\( A^* \) and \( B^* \) are the pair of points which represent the concave region shown in Figure 4.5b.

### 4.2.3.2 Weighted Shortest Path

Weighted shortest path helps in connecting the pair of points obtained in the concave region. The strongest edge is identified to connect the identified pair of points in the image \( I \). This is done by applying the energy function given in the Equation (4.12).

$$E(I) = \left(\left|\frac{\partial}{\partial x} I\right| + \left|\frac{\partial}{\partial y} I\right|\right)$$  \hspace{1cm} (4.12)
Let $l$ be the path between $A^*$ and $B^*$ on image $I$ with 4 or 8 pixel connectivity. The cost function of the path $l$ between $A^*$ and $B^*$ is given by the Equation (4.13)

$$E(l) = \sum_{i=1}^{L} \frac{1}{E(I(l_i))}$$  \hfill (4.13)

where $l_i$ denotes the position of the pixel of the location $i$ of the concave region and $L$ is the length of $l$. Optimal length $l^*$ is represented by the Equation (4.14)

$$l^* = \arg \min_{l} E(l) = \arg \min_{l} \sum_{i=1}^{L} \frac{1}{E(I(l_i))}$$  \hfill (4.14)

Minimizing $E(l)$ helps in finding the weighted shortest path between points $A^*$ and $B^*$ on image $I$ as in Figure 4.5c. Figure 4.5d shows the centromere position.

Figure 4.5  Weighted Shortest Path a) Contour Image b) Concave Points c) Centromere Position d) Combined Representation of (a,b,c)
The results obtained are verified using distance measurement of a chromosome image. The distance measurement includes image binarization followed by contour identification which are discussed in 3.2.1 and 3.2.2. The process of distance measurement is discussed in the next section.

4.2.3.3 Distance measurement

The distance measurement requires the location of the edge point of the contour of any image. The contour of the telomere portions are not considered as the distances measured are very small and are not well defined at both the end of chromosomes. Each horizontal image coordinate of the contour is considered. The distance measure of the edge points for each set of horizontal edge pixels are calculated using the simple pairwise Euclidean distance formulation given by the Equation (4.15)

\[ d_{\text{Euclidean}} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \]  

(4.15)

where \((x_i, y_i)\) and \((x_j, y_j)\) are the edge coordinate points. Figure shows the contour image, connected horizontal edge points and the concave region with the minimum distance measure value.

Figure 4.6 Distance Measurement a) Contour Image b) Horizontal Lines c) Enhanced Horizontal Lines d) Combined Representation of (a & b) e) Concave Region
The distance measurement helps in identifying the Euclidean distance between the two coordinate points which gives the minimum value in the concave region. From this it is understood that identifying the concave region helps in obtaining the centromere position of the chromosomes. This is again verified using the distance measure calculation as shown in Figure 4.6 (a-e).

4.3 LENGTH OF CHROMOSOMES

Length of the chromosome is another parameter used for classification. The steps are given below.

- Perform image binarization
- Perform image thinning
- Calculate the length of the chromosome by counting number of white pixel
- Image binarization is discussed in section 3.2.1. Image thinning is performed on obtained binary image.

4.3.1 Image Thinning

Thinning is one of the morphological operations which removes the selected foreground pixels of the binary image. The output of thinning operation is a reduced binary image with one pixel thickness which gives the center line. Thinning reduces the image components to their essential information for further analysis.

For preserving the connectivity, the algorithm should satisfy the following properties:
• Connected image region should reduce to connected line component

• Thinned line should have minimum number of pixels

• Approximated endline location should be maintained

• Thinning results should give an estimate of the medial axis value

• Extra spurs or short branches should be minimized

Thinning is achieved by removing the outer boundary pixel value iteratively. Thinning is represented using the image with structuring element.

The sequence of structuring elements is given as \( \{B\} = \{B^1, B^2, B^3, \ldots, B^n\} \) where \( B^i \) is a rotated version of \( B^{i-1} \). Thus, thinning by a sequence of structuring elements is represented in the Equation (4.16)

\[
A \otimes \{B\} = \left( \ldots (A \otimes B^1) \otimes B^2 \right) \ldots \otimes B^n
\]

(4.16)

Thinning helps to shrink the binary image. It obtains the image skeleton represented by the Equation (4.17)

\[
\text{Thin}(I, Se) = I - \text{HMT}(I, Se)
\]

(4.17)

4.3.1.1 Hit or miss transform

The origin of the structuring elements is transferred to all the pixel points in the image and is compared with the original pixel value of the image. If the foreground and the background pixels in the structuring element match the foreground and background image pixels, then the pixel underneath
the structuring element is considered as foreground otherwise the pixel is
considered as background. The hit or miss transform is mathematically
represented by the Equation (4.18)

\[ HMT_b(X) = \left\{ x \mid (B_1)_x \subseteq X, (B_2)_x \subseteq X^c \right\} \] (4.18)

where \( B \) denotes the set object points \( X \) and their background. \( B = (B1, B2) \)

\( B1: \) The set formed from elements of \( B \) associated with an object

\( B2: \) The set formed from elements of \( B \) associated with the corresponding background

4.3.1.2 Image pruning

The image thinning is followed by image pruning. Pruning is performed based on mathematical morphology. Pruning is a complement to skeleton and thinning operations. It removes the small branches and unwanted parasitic components present in the image. Parasitic components are the branches present in the thinned line which should be removed. Any branch with three or less components are removed. Input set \( A \) with a sequence of structuring elements to find the end points of thinning is used in obtaining the desired output. Consider \( X_1 = A \ominus \{B\} \), where \( \{B\} \) is the sequence of the structuring element. The sequence has two possible structures each of which is rotated 90 degree for obtaining 8 elements in total as shown in Figure 4.7.

Three runs of thinning helps in obtaining the end points in \( X_1 \) which forms a set \( X_2 = \bigcup_{k=1}^{8} (X_1 \otimes B^k) \) where \( B^k \) is the same structuring element. This is followed by the dilation of end points using the input set \( A \) as a delimiter. This helps in obtaining \( X_3 = (X_2 \oplus H) \cap A \) where \( H \) is a \( 3 \times 3 \)
structuring element. This dilation prevents the non-zero pixel appearance outside the region of interest. The union of $X_1$ and $X_3$ gives a better result and is given as $X_4 = X_1 \cup X_3$

![Image](image.png)

**Figure 4.7 Image Pruning with Structuring Element**

The image representing all the sets are shown in Figure 4.7. Thus image pruning helps in removing the extra side branches of the thinned image. Length of the chromosome is obtained by counting the one width white pixels in the pruned image. The length of the chromosomes are tabulated and discussed in chapter 5. If chromosomes are bent, then image straightening operation is performed.

### 4.4 IMAGE STRAIGHTENING

The curved chromosomes are straightened using ImageJ software. The image straightening is done to calculate the length of the chromosomes.
ImageJ software is used for various image processing applications which include image preprocessing, image enhancement, image segmentation etc. Any type of image format can be given as input for ImageJ software. Figure 4.8 shows the dialog box of ImageJ software.

![Figure 4.8 Dialog box of ImageJ software](image)

The steps followed for image straightening using imageJ software are shown in Figure 4.9.

![Figure 4.9 Overall Block Diagram of Image Straightening](image)

The image is read using the file option. It is then converted to a gray scale image. There is a plugin option which helps to select the region of interest. One of the options in the region of interest is ‘straight to line’. This option straightens the bent image. Figure 4.10 shows the straightened image of the chromosome. The length of the chromosomes is then calculated.
4.5 CENTROMERE INDEX (CI)

The CI is another important parameter for classification. Centromere divides the chromosome into two portions. The upper portion of the centromere is the P-arm and lower portion of the centromere is the Q-arm as shown in Figure 4.11.

\[
CI = \frac{\text{Short Arm}}{\text{Total Arm}} = \frac{p}{p+q}
\] (4.19)

The tabulated CI values are discussed in chapter 5.
4.6 SUMMARY

The parameters for classification of chromosomes includes centromere, length of the chromosome and CI. Centromere is obtained using three different approaches. (i) MAT is very suitable only for straight chromosomes and the performance goes down when bent chromosomes are given as input. Getting the minimum global point is difficult. (ii) projection vector algorithm is suitable for images not less than 90 degree of bend. For highly curved images the algorithm is not suited and requires further modification. (iii) Concave Function helps in identifying the centromere of any curved or bend chromosome. From the three approaches concave function yields better accuracy in centromere detection. The length of the chromosome is obtained by counting the number of white pixels along the medial axis of the image using thinning operation. Image straightening is performed for straightening the curved and bend chromosomes. CI is another parameter which also helps in the classification of chromosomes. All the parameters help in obtaining the karyotyped image. The karyotyped image helps in finding the numerical and structural abnormalities in chromosomes. The abnormality detection at an early stage helps in reducing the birth defect issues.