Estimation of Band Level Resolutions of Human Chromosome Images

W. Sethakulvichai\textsuperscript{a}, S. Manitpornsut\textsuperscript{b}, M. Wiboonrat\textsuperscript{c}, W. Lilakiatsakun\textsuperscript{d}, A. Assawamakin\textsuperscript{e}, S. Tongsima\textsuperscript{f}

IT Graduate School, Mahanakorn University of Technology, Bangkok, Thailand\textsuperscript{a, d}
Department of Computer and Multimedia Engineering, University of the Thai Chamber of Commerce, Bangkok, Thailand\textsuperscript{b}
College of Graduate Study in Management, Khon Kaen University, Thailand\textsuperscript{e}
National Center for Genetic Engineering and Biotechnology (BIOTEC), Klong Luang, Pathumthani, Thailand\textsuperscript{c, f}

Abstract—Current estimation methods for band level resolution of human chromosome images in cytogenetic laboratories are time consuming and required experienced specialists to manually perform. To alleviate this problem, in this paper, a computerized approach to estimate band level resolution is proposed. The intensity gradient profile and sign profile of chromosome images are utilized to count the number of bands. Then band level resolutions of chromosome images are classified into three categories: 400-, 500-, and 550-band levels by using k-nearest neighbor algorithm. The experimental results show the accuracy of the proposed algorithm. A discussion on how to improve the overall accuracy is also provided.

Keywords: image processing, band level resolution, feature extraction, medial axis determination, k-nearest neighbor.

I. INTRODUCTION

Automated systems are now widely integrated to almost everywhere surrounding our everyday life. Most of the systems are aiming to reduce a time consumption of repeated and tedious tasks. Chromosome analysis is one of those laborious and time-intensive activities, routinely conducted in cytogenetic laboratories [1] – [5]. The evolution of computer graphic in the 21st century has changed the way of human chromosome analysis. Integrating computerized system such as image processing into this area has been an open research topic since 1960s [2]. However, there are plenty of aspects for research in this area required for improvement, e.g. information from the chromosome images, processing speed, accuracy rate, reliability, easy to adaptation, and competitive cost.

The band level of resolution of each human chromosome image is one of the additional information that laboratories need to specify in karyotype report. Actually this information represents the total number of bands on chromosomes approximately half of the cell [6]. Most laboratories determine this number by identifying the specific landmarks on the specific chromosomes. From many literature reviews, automatic estimation system for band level resolution of human chromosome images has not been mentioned much in this research field. Furthermore, it is expected that knowing the band level resolution before the classification process will help managing the proper dataset for the chromosome classification system.

II. BACKGROUND

Cytogenetic laboratories are one of those specialized laboratories where many special tests for study microbiological behavior of all living things are usually conducted [7]. Karyotype analysis is a special activity, consisting of several steps, i.e. collecting cell samples, culturing cell samples, staining chromosomes, acquiring image of chromosomes, segmenting image into individual chromosomes, classifying chromosomes into their classes, rearranging and representing into karyogram and finally analyzing the karyogram [8].

The results of the karyotype analysis give valuable information in many aspects such as, for parental screening, sex determination, identification of abnormality in chromosomal diseases such as Down’s syndrome, Klinefelter syndromes, malignant tumor, cancer, leukemia, etc. [7].

G-banding is the technique that cytogenetic laboratories use to stain the specific content of chromosomes so that they express their distinctiveness pattern on their chromosome images [1]. This information is the main information that cytogenetic staff use to classify the chromosome into their correct classes. Although, there are many more others staining techniques available for karyotype analysis in this present, and some of them give more information of individual chromosomes, g-banding technique is the most widely used for the karyotype analysis due to its low cost and short preparation duration. In addition, information achieved after the analysis is adequate for medical practitioner to make decisions. Therefore, this work will be focused on the images of the g-banding chromosome only.
During the cell culturing process, cells go through their cells cycle processes as shown in Figure 1 [8]. Chromosomes in each cells, during mitosis process, has different stages of condensation [9]. Most of currently automatic karyotype analysis research is focusing on the images of metaphase chromosomes because the chromosome at this stage is in the most condensed form. However, in reality, cytogenetic need to conduct further examination on the pro-metaphase or even prophase chromosomes in many cases [10]–[11]. At the earlier stages of mitosis, the structures of chromosomes are less condensed and the overall length of all chromosomes in one cell is longer than in metaphase cell. Furthermore, the total numbers of bands along the length of all chromosomes are also different. The more number of bands the earlier the stage of mitosis.

The level of resolution of human chromosome image is an important number to indicate the stages of chromosomes being analyzed. It is normally determined by counting the number of bands seen in a haploid set (22 autosomes + X and Y) [6]. However, this is really difficult to perform in reality. Most laboratories currently evaluate the level of resolution by identifying the land marks on the chromosome images and scoring the karyotype image by using the scoring table as shown in Table I. At least three of the land marks criteria to be obtained to indicate the band resolution [10]. There is also a requirement to specify the number of band resolution in the karyotype report to confirm that the analysis is appropriate for the referral reason. Table II shows lowest standard acceptable for a given reason for referral in constitutional

**Table I** G-banding resolution evaluation score [10].

<table>
<thead>
<tr>
<th>Band resolution</th>
<th>Land marks Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>300 band</td>
<td>2 dark bands on 8p (8p12 &amp; 8p22) 3 dark bands on 10q (10q21, 10q23, 10q25) 20p12 visible 22q12 distinct</td>
</tr>
<tr>
<td>400 band</td>
<td>3 dark bands on mid–4q (q22-28) 3 dark bands mid-5q (5q14, 5q21, 5q23) 2 dark bands on 9p (9p21 &amp; 9p23) 13q33 distinct</td>
</tr>
<tr>
<td>500 band</td>
<td>7q33 &amp; 7q35 distinct 3 dark bands on 11p (11p12, 11p14, 11p15.4) 14q32.2 distinct 4 dark bands on 18q (18q12.1, 18q12.3, 18q21.2, 18q22)</td>
</tr>
<tr>
<td>550 band</td>
<td>5q31.2 distinct 8p21.2 visible 2 dark bands on 11pter (11p15.2 &amp; 11p15.4) 22q13.2 distinct</td>
</tr>
<tr>
<td>700 band</td>
<td>2p25.2 distinct 2q37.2 distinct 10q21.1 and 10q21.3 resolve 17q22-q24 resolves into 3 dark bands</td>
</tr>
</tbody>
</table>

**Table II** Minimum G-banding resolution for referral reason [10].

<table>
<thead>
<tr>
<th>Reason for referral</th>
<th>Minimum level of resolution of G-banding images quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmation of aneuploidy e.g. direct lymphocyte, direct CV or solid tissue culture preparation</td>
<td>&lt; 300 band</td>
</tr>
<tr>
<td>Exclusion of known large structural rearrangements e.g. lymphocyte, solid tissue, CVS direct preparation or amniotic fluid cell preparation</td>
<td>300 band 400 band</td>
</tr>
<tr>
<td>Identification and exclusion of small expected structural rearrangements e.g. lymphocyte, solid tissue, CVS culture or amniotic fluid preparation</td>
<td>400 band 400 band</td>
</tr>
<tr>
<td>Routine amniotic fluid and CV culture preparations</td>
<td>400 band 500 band</td>
</tr>
<tr>
<td>Abnormal ultrasound scan associated with AF, CV and solid tissue referrals</td>
<td>500 band 700 band</td>
</tr>
<tr>
<td>Blood referrals, not covered by exclusion criteria</td>
<td>550 band 700 band</td>
</tr>
<tr>
<td>For microdeletion syndroms (when no FISH probe is available)</td>
<td>700 band</td>
</tr>
</tbody>
</table>
analysis as recommended by the Association for Clinical Cytogenetics [10].

The standard ideograms of chromosomes class 1 shown in Figure 2 provide schematic representations of chromosomes corresponding to approximately 300, 400, 550, 700 and 850 bands [6]. It is clearly noticeable that different resolution yielding different banding or intensity profile.

Apart from this, different levels resolution of G-banding images acquired from different staining process creates a tedious problem to most laboratories. Sometime, a slightly different of period when capturing the metaphase moment could cause a different in level resolution of banding patterns.

III. PROPOSED METHODOLOGY

Figure 4 shows the different banding levels of chromosome number 1, acquired at the different moments during the cell division process. The longer chromosomes give many more bands than the shorter one [11].

Since the shape and the number of bands of the same class chromosomes are varied according to their relative mitosis stages, classification of chromosomes at the level resolution different from the resolution of trained database is prone to a huge number of error rates. Estimation the band levels resolution of chromosomes prior to the classification stage is a promising approach to enhance the accuracy of the classification when incorporate this module into the system to help selecting the proper band level database to the analyzing subject. In the next section, automatic band level resolution estimation algorithm is described in detail.

A. Image Enhancement

The first step is to eliminate noises, enhance image contrast and adjust the brightness of the images by using the following methods:

- Median filtering – to get rid of pepper noises.
- Contrast enhancement – to adjust a low-contrast gray-scale image.

Figure 4 Diagram shows the proposed automatic chromosomes band counting system.
• Intensity enhancement – to compensate non-uniform brightness and contrast behavior of the images by using histogram equalization.
• Segmentation – to convert the image to binary by using Otsu’s thresholding algorithm. Then, closing, opening and filling image operations are performed to mask out the individual chromosomes region and smooth the image boundary.

B. Medial axis determination

A simple thinning algorithm is applied here to determine an axis of an individual chromosome, as shown in Figure 6(a). However, the length of thinning line is shorter than that of the chromosome, thus requiring line extension. Bresenham’s line algorithm is applied to extend the tip points of the thinning lines to tip point of the chromosome, as illustrated in Figure 6(b). Finally, as displayed in Figure 6(c), the medial axis of individual chromosome image is formed.

C. Intensity profile extraction

Intensity profile is extracted from the individual chromosome image on the perpendicular lines of the medial axis. In Figure 7(a), the perpendicular lines $S_o$ are generated from the local slopes of the medial axis. However these perpendicular lines might overlap and therefore result in incorrect intensity profile. The orientations of the perpendicular lines $S_o$ are rearranged to eliminate the overlapping lines as shown in Figure 7(b).

Let $S_p$ is the $1 \times n$ vector of rearranged slopes. The $S_o$ is rearranged by using a moving mean value as:

$$S_o(i) = S_o(i) \quad \text{where } i = 1 \text{ and } n$$

$$S_o(i) = \text{Mean}(o_{\delta_1}, S_o(i - 1), S_o(i), o_{\delta_2}, S_o(i + 1)) \quad \text{where } i = 2, 3, 4, \ldots, n - 1$$

(1)

The $o_{\delta_1}$ and $o_{\delta_2}$ are the smoothing weights calculated from the difference between the current slope and previous slope and the current slope and the next slope respectively.

The sampling is done by collecting all the intensity value on each perpendicular line and calculating the average value along this line except three pixels that closed to the medial axis. This is because most of the chromosome images appeared to be “white” in the middle of the chromosome body along its medial axis as noticeable in Figure 7 (b). The original intensity profile as shown Figure 8 (a) is then smoothed by using a moving average filter. Let $I_o$ is the original intensity profile and $I_s$ is the smoothed intensity profile.

The span of the moving average of this work is set to 5 (experimentally selected). Transformation from the original to the smoothed profile can be expressed as;

$$I_s(i) = I_o(i) \quad \text{Where } i = 1$$

$$I_s(i) = \begin{cases} \text{Mean}(I_o(i - 1), I_o(i), I_o(i + 1)) & \text{Where } i = 2 \text{ to } n - 2 \\ \text{Mean}(I_o(i - 2), I_o(i - 1), I_o(i), I_o(i + 1), I_o(i + 2)) & \text{Where } i = n - 1 \end{cases}$$

(2)

D. Intensity gradient determination

The intensity gradient profile is the rate of change in intensity value along the medial axis line defined as:

$$G_{I_5} = \nabla I_s = \frac{\partial I_s}{\partial x}$$

(3)
where $G_{IS}$ is a vector returns the one-dimensional numerical gradient of $I_x$. $G_{IS}$ corresponds to $\frac{\partial I_x}{\partial x}$, the differences in x (chromosome medial axis) direction. Figure 8(c) shows a sample of intensity gradient profile derived from the intensity profile.

E. Band counting

The bands are counted by detecting the changes in direction of the gradient intensity profile. The band is divided into 2 types: white bands and dark bands. The rate of change in intensity between the edge of white and dark band should be high. The algorithm is set to count the changes in sign (+/-) of the gradient profile along the medial axis. The gradient profile is converted to the sign profile as

$$Y = \begin{cases} 
1, & \text{where the } G_{IS}(i) > 0 \\
0, & \text{where the } G_{IS}(i) = 0 \\
-1, & \text{where the } G_{IS}(i) < 0
\end{cases}$$

(4)

where, $Y$ is the sign profile calculated from a gradient profile. Figure 8(d) shows the sample of sign profile.

F. K-Nearest Neighbors

The k-NN is one of the supervised data mining techniques for classifying data into classes. The method is based on the majority of the k ($k = \text{number of neighbors}$) nearest distances between the features of the training dataset to the features of the testing data [13]. Four features which are number of bands, length, width and aspect ratio are used in the proposed system. The number of neighbors was set to vary from $k = 1$ to $15$. The features of test dataset are then compared to the nearest trained features.

IV. EXPERIMENTS AND RESULTS

The dataset of chromosome images are from the Center for Medical Genetics Research, Rajanukul institute, Thailand. In this research, 210 karyogram images are collected according to the availability of information of the band level resolution. All chromosomes were stained using G-banding technique. These chromosomes were cultured, stained, captured by cytogenetic specialists. Karyograms were also prepared by using the current analysis system in the laboratory. However, image intensity and scaling are very diverse. The band level resolution of each images were identified by experienced staff. Three band levels resolutions are investigated, i.e. 400, 500 and 550 band levels.

<table>
<thead>
<tr>
<th>DESCRIPTION</th>
<th>BAND RESOLUTION</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>Testing</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>

Seventy metaphase cell images for each band level resolution are employed in experiment, or 210 images in total. Each metaphase cell image consists of 46 individual chromosomes. These samples are divided into two sets, training and testing dataset as shown in table III. All data in each band level resolution are randomly selected and divided into 80% for training and 20% for testing, respectively. The experimentation is divided into 2 sections, band counting and band estimation.

A. Band counting

The test is conducted to count number of bands of individual chromosomes from the training dataset of 400, 500 and 550 band level resolutions, respectively. The output of each run contains the following features.

1) The computed bands:
This feature is computed by the technique as described in the previous section.

2) The Aspect ratio (AR):
This feature is defined by the ratio of length and the width of each chromosome.

Figure 8 Output profiles from the band estimation algorithm, (a) intensity and smoothed intensity profiles, (b) original input, (c) gradient intensity profile, and (d) sign profile.
\[ AR = \frac{\text{length}}{\text{width}} \]  

This feature indicates the morphological shape of the chromosome. The high value of \( AR \) means long and/or narrow chromosome, thus implying the high value of band level resolution. On the other hand, the low \( AR \) implies the low value of band level resolution.

The results are plotted as shown in figure 9, which is the plot of counted number of band vs. the aspect ratio of all chromosomes from each band level datasets. This graph shows the three clusters of these dataset in a 2D feature space spanned by the relation between numbers of bands counted from the proposed algorithm and theirs relative aspect ratios. The border areas of these clusters are distributed dispersedly but the center of populated areas of each cluster can be observed.

B. Estimation of Band Level Resolution

The band level resolution is estimated by implementing k-nearest neighbor (k-NN) classification model. The features are number of bands, length, width and aspect ratio, respectively. The number of neighbors was set to vary from \( k = 1 \) to 15. For simple image representation, Figure 10 displays the relationship of the target node and the neighbor nodes in the 2D feature space of the Count Number and the Aspect Ratio features.

Figure 11 shows the accuracy percentage of the classification results of each band level and the overall test set. The maximum accuracy percentage is 72\%, 93\%, and 79\% for classification of 400-, 500-, and 550-band level resolutions. The overall accuracy is 76\% at \( k = 3, 9 \) and 11, respectively.

V. DISCUSSION AND FUTURE WORK

This work has shown the method for extracting features from individual chromosomes, i.e. intensity profile, gradient intensity profile, sign profile and aspect ratio. By using k-NN technique, the classification of human chromosome band level resolution can be identified with the maximum accuracy of 76\%. The main issues that affect the accuracy of the system are the chromosome image quality and the accuracy of band level resolution from experts.

The actual chromosomes in the sample glass slides are in three dimensions. When they are captured by a camera, they are projected to two-dimensional images. Some of them are overlapped or crossed by others. The more of chromosomes are overlapped, the lower quality of the chromosome images is. The low quality of chromosome image affects the sign profile and therefore decreases the classification accuracy of the proposed system.

Another effect on the system accuracy is the classification results from experts. Experts do not count the actual number of bands in the chromosome image since actual counting takes too much time for human. Instead, they notice the landmark of some chromosomes to identify the band level resolution. This is an estimation approach. Unfortunately, the chromosome images have wide variation in terms of landmarks, length, width, and band intensity.

To account for the aforementioned issues, our future work is considering more features from the chromosome images, e.g.

\[ \text{Accuracy} = \frac{\text{correct classification}}{\text{total classification}} \times 100\% \]

\[ k = 1, 3, 5, 7, 9, 11, 13, 15 \]

\[ \text{Accuracy} = \frac{\text{correct classification}}{\text{total classification}} \times 100\% \]

\[ k = 1, 3, 5, 7, 9, 11, 13, 15 \]

The location of centromere (or centromere index), the normalized band intensities, scaling effect and compensation of band counting in the overlapped areas. Other classification techniques are also under investigation. In addition, the interactive cross validation with the experts will be conducted to reduce human errors.
ACKNOWLEDGMENT

Special thanks to Dr. Verayuth Praphanphoj, M.D., Center for Medical Genetics Research, Rajanukul Institute, Thailand and all his staff for supporting data and information.

REFERENCES


