Journal of **Kim Il Sung** University (Natural Science)
Vol. 4 No. 1
Juche104(2015)

# Plant Chromosome Pairing using Neural Networks and Cluster Analysis

*Ho Un Hyang, Ri Sok Jun* and *Kim Kang*

**Abstract** Until now, all processes in manual karyotyping analysis were performed based on visual recognition by naked eyes. And various methods of chromosome image analysis have been developed via computer aided image analysis methods.

On the basis of previous studies about computer aided image analysis methods we have distinguished individual chromosomes by means of skeletal line extraction from rye chromosome image including B chromosomes and developed homologous chromosome paring and ploidy determination algorithm by neural network and cluster analysis which use length, arm ratio, total pixel number, the presence of satellite and secondary constriction as characteristic index. Our new homologous chromosome paring algorithm can lead automatic and efficient identification of chromosomes and correct chromosome paring and ploidy determination even in the case of presence of B chromosomes.

**Key words** chromosome classification, chromosome pairing, artificial neural network, B chromosome

## Introduction

Chromosomes are genetic information carriers and karyotyping analysis on the chromosome images constitutes an important stage in clinical and cytogenetic studies requires much labors, times and funds. Researchers have developed a number of karyotyping algorithms using a computer-based image processing in order to minimize much labors and ensure the objectivity, the exactness and the promptness.

Generally, karyotyping analysis contains chromosome-counting, determination of chromosomal characteristic indices, pairing of homologous chromosomes, determination of polyploidy, making karyogram and ideogram, and the output of karyotype formula [1, 4]. Each stage of karyotyping is important, but the most essential stage among them is chromosome classification to distinguish individual chromosomes and homologous chromosome pairing.

By recently developed automatic counting algorithms of chromosomes firstly skeletal lines can be defined [2, 10] and on the basis of the lines, the touching and overlapping chromosomes can be separated [4, 6－9], but in some instances such as the presence of T-shape overlapping [12] and satellite and B chromosome above methods are not suitable. Also, multicolor karyotyping analysis method [3] have been developed, but these technologies have inherent limitations, that is, in certain situations, may result in chromosomal misclassification and be expensive analysis cost.

B chromosomes are dispensable elements that do not recombine with the A chromosomes during meiosis which exist in only some plant line [11].

The study on the application of artificial neural networks in analyzing and classifying the human banded chromosomes has been previously performed [5, 6, 8, 13].

In this paper, we have done chromosome classification using skeletal line extraction on non-band stained plant chromosome images including B chromosome and chromosome paring and ploidy determination by neural network and cluster analysis.

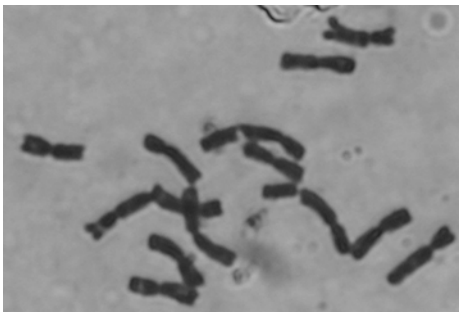

Fig. 1. Rye chromosome image including B chromosome

### 1. Materials

We have made chromosome specimen of root tip from rye (*Secale cereal L.*) JKN line (2n=14+0～6B) including B chromosome. Chromosome images were obtained by computer connected to light microscope "Olimpus".

Fig. 1 shows the original rye chromosome image.

### 2. Chromosome Classification based on Skeletal Line Extration

In this paper, based on digital image processing techniques the new system was developed to recognize individual chromosomes from a chromosome image and to count chromosomes automatically.

Chromosome images converts to gray scale images and was performed the conversion of binary images using the method of Otsu's threshold.

Automatic counting of chromosomes is composed of two broad stages namely preprocessing and counting.

#### 2.1. Preprocessing

The preprocessing stage involves the making of binary images, median filtering, thinning and cleaning.

Fig. 2. Rye chromosome binary image

At making of binary images stages the pixels below Otsu's threshold is regarded as the background and the image is separated into several regions by grouping of the pixels separated from background. Each region can be comprised of one or more than two chromosomes (Fig. 2).

At median filtering removes various noise and converts the small holes in the body of chromosomes to the region of chromosomes. And this stage also smoothes the chromosome contours so as not to grow small and unwanted branches.

At thinning stage the single-pixel-wide skeletons are obtained from the binary image resulted from the median filtering.

When $N$ is pixel number in chromosome region, $(x_i, y_i)$ ($i$=1∼$N$) is position of $i^{th}$ point, $M$ is pixel number of contour and $(u_j, v_j)$ ($j$=1 ∼ $M$) is position of contour point, $d_i = \min_j\left(\sqrt{(x_i-u_j)^2+(y_i-v_j)^2}\right)$ is distance to nearest contour point which is given to each point of chromosome region.

After defining of the points of contour of chromosomes the value of distances between the nearest contour points located from every point in a chromosome image are calculated and the values are given to those points. The skeletal points in a chromosome image are the points that the contour distance value $d_i$ is bigger than not only half of the minimum chromosome width but also the values to eight points around it.

And also calculated skeletal points are discontinuous because chromosome images are not smooth. Therefore, the interpolation lines to skeletal points are calculated using the least squares method and these lines regarded as the imaginary axes of the chromosomes thereby the skeletal lines are obtained by aligning and connecting the skeletal points in direction of the lines. The obtained skeletal lines are given in forms of a curved line by which are chromosomes are divided to two regions in the longitudinal direction. In the vertical direction of the skeletal lines the width and the average width of chromosomes calculated by investigating chromosome images. Chromosome width at a point means pixel numbers located on the vertical lines to the skeletal lines.

When $L$ is pixel number of skeletal point and $d_k^0$ is the contour distance value to skeletal point, the average width of chromosome can be calculated by following formula

$$\overline{w} = \frac{\sum_{k=1}^{L} d_k^0}{L}$$

The width of chromosomes is almost unchanged in an image (Fig. 3).

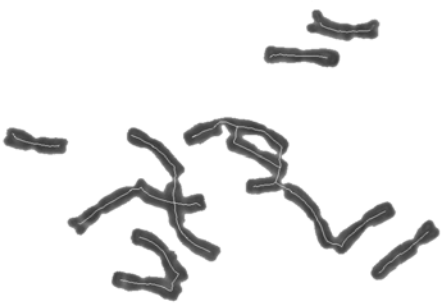

Fig. 3. Chromosome image with skeletal line after thinning stage

After calculating of the average width of chromosomes, the slight connections between chromosomes are recognized. At all points on the skeletal lines the width of chromosomes are calculated and if the values are bigger than half of the average width, the presence of slight connections are regarded. This threshold namely half value of the average width is the value setting a single chromosome with a small bending not to being separated two parts in a contour of it.

At the stage of cleaning based on the determination of the chromosomal skeletal lines the ranges separated from those lines regarded as the non-chromosomal ranges, and converted to the background. And also, the ranges corresponding to

the skeletal lines of which length is smaller than chromosomal average width should be removed (Fig. 4).

Fig. 4. Chromosome image separated into individual chromosomes

**2.2. Determination of chromosome numbers**

Skeletonized images resulting from preprocessing stage are taken as input to separate chromosomes and determine chromosome numbers. At this stage it is essential to separate several touching and overlapping chromosomes each other.

Firstly endpoints and crossover identify among the points along skeletal lines. End points are in the skeletal lines that has only one neighboring skeletal pixel and the points that has three neighboring skeletal pixels are crossovers, namely with overlapping of two chromosomes and touching of endpoint of other chromosome.

Start from the end point that appears first in raster-scan order and keep tracing the skeletal lines until another end point is reached. Appearing the end points those are marked and moving to the neighboring skeletal lines those pixels are marked. This process continued until endpoints and crossovers are reached. Appearing the crossover the interpolation line of the latest appeared five skeletal pixels are given and moved to the points in the nearest direction with a interpolation line to mark this point. This process is repeated continuously. At the ending of cycle in skeletal lines all the skeletal pixels appeared in cycling process are marked and in next counting process these are not regarded as skeletal pixels. And also the skeletal pixels appeared in cycling process are not marked and therefore are still remained as the skeletal pixels. Whether this point will be crossover in next counting or not is determined according to the numbers of unmarked pixels among the skeletal pixels around this point. If endpoint appears, at this stage the number of chromosome is increased by one and confirmed chromosomal range. The center of chromosomal range is skeletal lines connecting two endpoints. In the case of being chromosomes touched and overlapped the contour of chromosomal range near crossovers is determined by connecting the contour pixels of non-touching range (Fig. 5).

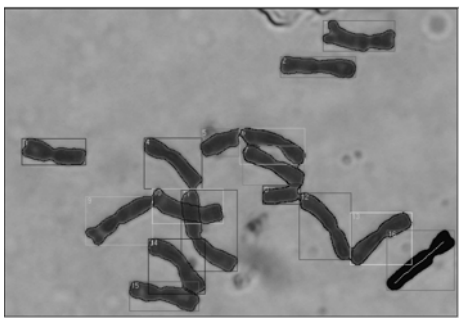

Fig. 5. Chromosome image after auto-counting

**3. Determination of Chromosomal Characteristic Indices**

After individual chromosomal ranges are isolated from chromosome images, each chromosomal characteristic index such as the length, the arm ratio, area and so on are calculated.

Firstly centromere and second-constriction sites in chromosomes are confirmed. Centromere and second-constriction sites are identified the points which the value of chromosome width determined at that point are smallest among the skeletal pixels. If the smallest points appear more than

three, only two points are selected in decreasing order of the width. Every chromosome has not second-constriction sites. Among the smallest points the point near by medial of chromosome (the midpoint of skeletal line) is identified centromere and other point is identified second-constriction site. If the smallest point is not detected because the centromere site of chromosome are overlapped other chromosomes, crossover will be guessed centromere site.

In next stage satellite is confirmed. As satellite is the range that a bit away from chromosome, it can be identified with other chromosomal range at chromosome counting stage. Because the width of connecting region between satellite and chromosome is very small to chromosomal average width. Among the skeletal lines the lines of which length are notably small are corresponded to satellite region.

The value of following formula is minimum in the chromosomes including satellite.

$$\sqrt{(a_x{}^2 x_b + a_y{}^2 x_0 + a_x a_y (y_b - y_0) - x_b)^2 + (a_x{}^2 x_0 + a_y{}^2 y_b + a_x a_y (x_b - x_0) - y_b)^2} + \\ + \min(\sqrt{(x_1 - x_b)^2 + (y_1 - y_b)^2},\ \sqrt{(x_2 - x_b)^2 + (y_2 - y_b)^2})$$

Where, $(x_0,\ y_0)$ —center point of chromosome, $\{a_x,\ a_y\}$ —direction unit vector of chromosome, $(x_1,\ y_1)$, $(x_2,\ y_2)$ —point of both end points in chromosome, $(x_b,\ y_b)$ —center point of satellite chromosome.

Next stage determines satellite chromosome.

The skeletal lines not being satellite are elongated backward and forward to confirm the nearest chromosome to satellite site and the satellite site is included in this chromosomal range.

Then the length of chromosomes is calculated. The endpoint of skeletal line is not real endpoint of chromosome and placed in telomere region of chromosome and is in its contour. As elongated skeletal lines backward and forward, the points that that lines are crossed with the contour of chromosome are regarded as endpoints of chromosome. The length of the arm is calculated by dividing this on the basis of centromere and among them the lager one is defined to large arm and the smaller one to small arm and the ratio of large arm to small arm is defined to arm ratio. The total sum of pixels in chromosomal range is identified to area of chromosome.

## 4. Chromosome Pairing and Polyploidy Determination using a Neural Networks and Cluster Analysis

Chromosome pairing and polyploidy determination have done using a neural networks in which 5-dimentional characteristic vector of two chromosomes is input and cluster analysis which uses neural network output value.

Algorithm for chromosome pairing and determination of polyploidy using a neural networks and cluster analysis is as following.

**Stage 1** Standard several chromosome images are prepared.

In these images chromosome numbering and determination of chromosomal characteristics are performed. Each chromosomal character is saved as a vector being composed of five parameters.

In this vector length, arm ratio, area is a real numbers and the value of satellite and second constriction is 1 or 0 according to its presence.

**Stage 2** Next neural network is composed.

Neural network is composed of the Multi-Layer Neural Network with Back-Propagation algorithm. This Neural network number of input is equal to number of chromosome and input signal is set as a chromosomal characteristic vector. Hidden layer is composed of 2 layers and neural cells is about 20 and evaluation function is used of sigmoid function. The signal of input layer is gone to hidden layer and the signal of hidden layer is again gone to output with weighted links.

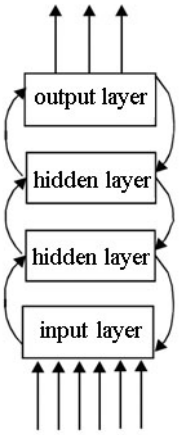

Fig. 6. Architecture of neural network

Input is two chromosomes, pairing is performed when output is 1, and not when output is 0.

**Stage 3** This training is performed on standard images.

If input is characteristic vector of homologous chromosomes then output value is set to 1, and if input is characteristic vector of nonhomologous chromosomes then output value is set to 0.

**Stage 4** chromosome paring using cluster analysis

Estimate paring possibility by output values of neural network where input is characteristic vectors of all possible chromosome pairs except B chromosome. The nearer output value is to 1, the higher is paring possibility, while 0, lower.

Based on these values chromosome paring is processed.

Consume ploidy as 2, 3, 4, and so on, and perform paring in order to obtain homologous chromosome cluster corresponding to ploidy number and estimate the overall paring mark by summing of all paring possibilities in each cluster. Then in order to determine diploid, tetraploid, hexaploid add compensation value to this mark.

Finally determine ploidy and chromosome paring from the result with the highest mark. In the case of B chromosomes, chromosome paring are not carried.

**Stage 5** from the final chromosome paring result determine karyotype formula and draw ideogram.

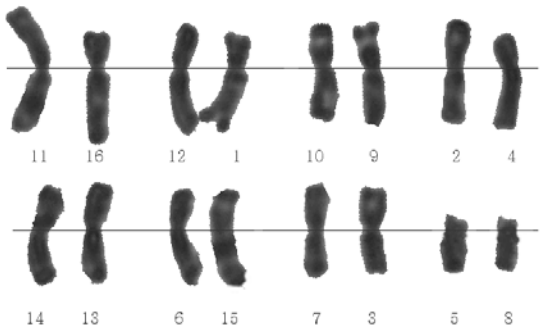

Fig. 7. Karyogram

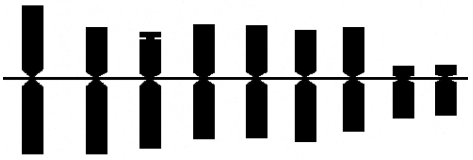
Fig. 8. Ideogram

Karyotype formula: $2n$=14+2B=14m(2sat)+2st

## 5. Discussion and Conclusion

To develop an automatic karyotyping analysis system is important in cytogenetic study. Recently the development of an automatic karyotyping analysis system based on image processing in chromosome study has made a great advance. Specially, in human chromosome study many algorithms for chromosome classification and homologous chromosome pairing have been developed using band pattern as well as length and arm ratio. However, in plant many chromosome classification works using band pattern have not been performed and it has not been reported in the case of the presence of B chromosomes presented in a few lines of some species.

In this paper in rye chromosome image including B chromosomes the algorithms for chromosome classification based on the extraction of skeletal lines and homologous chromosome pairing and ploidy determination unrelated to the number of B chromosome.

1) We present automatic counting algorithm by which a correct separation between touching and overlapping chromosomes is possible and considered even the presence of satellite and secondary constriction in the determination of chromosomal characteristic indices such as the length, the arm ratio, the area based on the methods of chromosome image analysis developed by this time.

2) And also, we have developed new algorithms for homologous pairing and/or determination of polyploidy by the chromosomal characteristic indices such as the length, the arm ratio, the total pixel number, the presence of satellite and secondary constriction.

## References

[1] Boaz Lerner et al.; IEEE Transactions on Signal Processing, 46, 10, 2841, 1998.

[2] C. U. Garcia et al.; Mach. Vision Appl., 14, 145, 2003.

[3] Charles Lee et al.; Am. J. Hum. Genet., 68, 1043, 2001.

[4] C. C. Graham et al.; IEEE Transactions on Signal Processing, 50, 8, 2080, 2002.

[5] M. M. Ibrahiem El Emary; Journal of Computer Science, 2, 1, 72, 2006.

[6] Jau-hong Kao et al.; Pattern Recognition, 41, 77, 2008.

[7] L. V. Guimaraes et al.; Proceedings of the 25th Annual International Conference on IEEE EMBS, 941～943, 2003.

[8] Mehdi Moradi et al.; Pattern Recognition Letters, 27, 19, 2006.

[9] M. Moradi et al.; Proceedings of the 16th IEEE Symposium on Computer-based Medical Systems, 56～61, 2003.

[10] M. Moradi et al.; Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, 567～570, 2003.

[11] N. Jones et al.; Trends in Plant Science, 8, 9, 417, 2003.

[12] V. Gajendran et al.; International Conference on Image Processing, 2929～2932, 2004.

[13] X. Wang et al.; J. Phys., D 38, 2536, 2005.