Songklanakarin J. Sci. Technol.
32 (3), 255-262, May - Jun. 2010

Songklanakarin Journal of Science and Technology
http://www.sjst.psu.ac.th

Original Article

Chromosome image classification using a two-step probabilistic neural network

Sunthorn Rungruangbaiyok* and Pornchai Phukpattaranont

Department of Electrical Engineering, Faculty of Engineering
Prince of Songkla University, Hat Yai, Songkhla, 90112 Thailand.

Received 18 November 2009; Accepted 30 March 2010

Abstract

Chromosome image analysis is composed of image preparation, image analysis, and image diagnosis. General procedure of chromosome image analysis includes of image preprocessing in the first step, image segmentation, feature extraction, and image classification in the last step. This paper presents the preliminary results that use probabilistic neural network to classify chromosome image into 24 classes. Features of chromosome which were used in this paper are area, perimeter, band's area, singular value decomposition, and band profile. Chromosome images were grouped in two steps by probabilistic neural network. Six groups and twenty four groups are in the first and the second step, respectively. The result from the second step is twenty four chromosome classes. Density profile sampled at 10, 30, 50 and 80 were tested. The best classification result of female is 68.19% when density profile at 30 samples was used, and that of male is 61.30% when density profile at 50 samples was used.

Keywords: image analysis, chromosome image, image segmentation, feature extraction, neural network

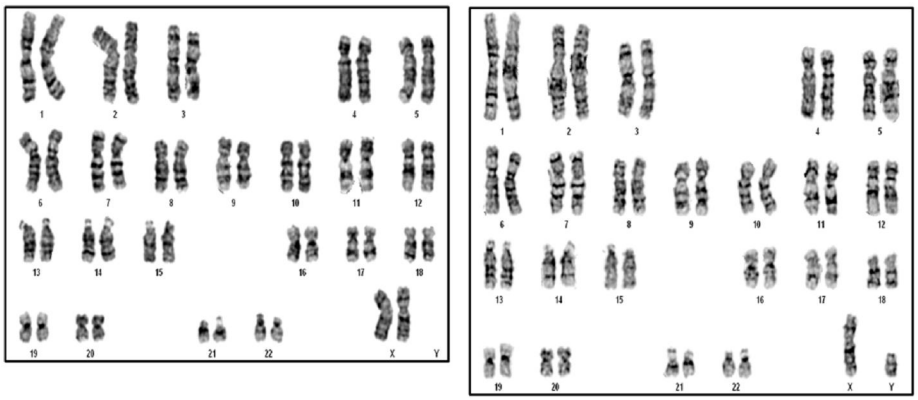
1. Introduction

Chromosome image classification is an important procedure for clinician, doctor and researcher to research and diagnose genetic disorder (Piper *et al.*, 1980), cancer (Hampton *et al.*, 1996, Truong *et al.*, 2004), and a variety of other diseases (Boehm *et al.*, 2004). In chromosome image diagnosis, at least fives images are usually used but at least twenty images are used if it is from bone marrow. Chromosomes are in every cell except red blood cell. Cells that used for chromosome image and analysis are usually taken from amniotic fluid, blood sample, and bone marrow. Normal human chromosome has 23 pairs: 22 pairs of body chromosome (called autosome chromosome) and one pair of sex chromosome, XX from female and XY from male.

* Corresponding author.
Email address: r_sunthorn@yahoo.com

Karyotyping is the most common procedure for analyzing and classifying banded chromosomes from images of a metaphase cell (Kyan *et al.*, 1999). This procedure defines the number and arrangement, size and structure of the chromosomes and assigns each chromosome to one of the 24 human chromosome classes as shown in Figure 1. In non-banding technique, construction of the karyogram the autosome chromosomes are numbered from 1 to 22 in decreasing order of length and sex chromosomes are referred to as X and Y. When the chromosomes are stained by methods that do not produce bands, they can be arranged into seven readily distinguishable groups(A-G) based on descending order of size and the position of the centromere (Shaffer *et al.*, 2005). Banding technique is added to increase in chromosome image classification because those former characteristic shape forms are variable.

Currently, supervised and unsupervised neural network have been used to classify chromosome classes for increasing efficiency and reducing processing time. A study indicated that processing time can be decreased by reducing



(a) (b)

Figure 1. Karyotype image, (a) female, (b) male.

size of output using binary code transform technique (Delshadpour *et al.*, 2003). On the other hand, increasing efficiency can be achieved by separating chromosome classification in two stages. In the first stage, chromosome images were classified into one of major group and each group was classified with its classifier in the second stage (Cho *et al.*, 2004). A variety of neural networks was used in chromosome classification including multi-layer perceptron neural network (Delshadpour, 2003), neural network, fuzzy logic rule based, and template matching (Badawi *et al.*, 2004), backpropagation neural network (Cho *et al.*, 2004), and probabilistic neural network (Sweeney *et al.*, 1994). Back-propagation neural network is generally used in classification but it takes time in training to learn relationship between input and output variables. In addition, sufficient datasets must be available to divide the data in a training set, a test set and a validation set to avoid over fitting (Yichu *et al.*, 2002).

Probabilistic neural network offers several advantages over back-propagation neural network. Its training time is much faster and usually a single pass (Wasserman, 1993). It was shown that its recognition rate is better than that from maximum likelihood and back-propagation neural network (Sweeney *et al.*, 1994). It allows true incremental learning where new training data can be added at any time without requiring retraining of the entire network (Masters, 1993, 1995; Musavi, 1992; Specht, 1990). However, karyotype is still a challenging problem due to variation of cell cultering conditions, chromosome staining, and microscopic illumination. Manually karyotyping is a labor-intensive and time consuming process (Xingwei *et al.*, 2005). Therefore, the computer-aid system has been developed to analyze chromosome image for karyotyping and improving these problems. Although users must approve and manually correct the final results, computer-aided system can help an expert to reduce labor and timing problem.

This paper was improved and changed some parts from our previous paper (Rungruangbaiyok *et al.*, 2009). Features which were added are band area, singular value decomposition and band profile. But the band ratio was removed. Back propagation neural network was changed to probabilistic neural network in image classification. Finally, chromosome images were grouped in two steps. There are six groups in the first step and twenty four classes in the second step.

2. Image analysis

Figure 2 shows a block diagram of chromosome image analysis. It consists of four processing steps, i.e., image preprocessing, image segmentation, feature extraction, and image classification. Details are given as follows.

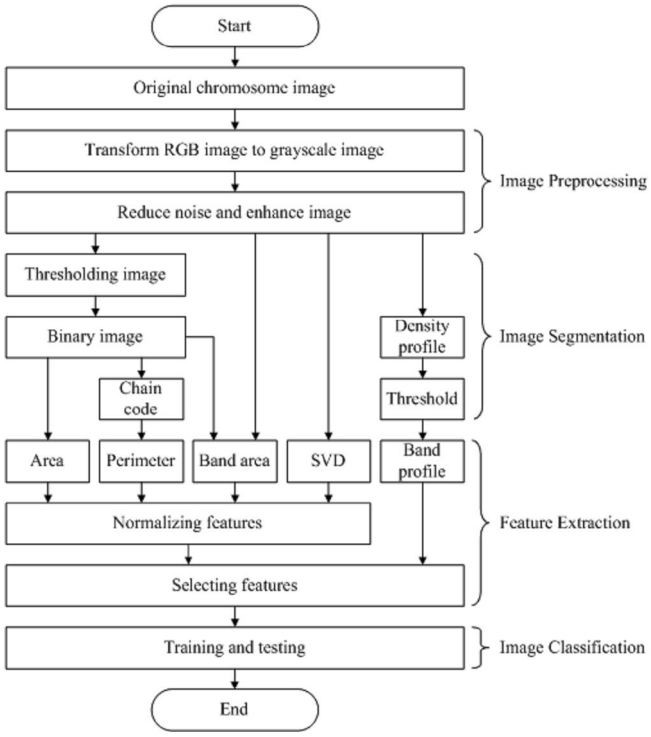


Figure 2. Image analysis block diagram.

2.1 Image preprocessing

Chromosome images were captured from microscope and digital camera. Some pictures are not sufficient clearly and some chromosomes in the pictures are overlapping because of cell culturing, chromosome shape, chromosome staining and illumination of microscope. Image preprocessing is a process that modifies and prepares the pixel values in the image for using in algorithm. The original chromosome metaphase cell pictures are RGB image. In this paper, RGB image is transformed to gray scale image by (McAndrew, 2004)

$$Ig = (0.2989 * R) + (0.5870 * G) + (0.1140 * R), \quad (1)$$

where *Ig* is gray scale image and *R, G, B* are red, green, blue components, respectively.

Noises in the image were suppressed by using the average filter. Histogram equalization was applied to improve contrast and image quality. Finally, each chromosome was arranged and saved for the next processing step.

2.2 Image segmentation

Image segmentation is a process that segments the interesting objects from a background. Thresholding technique is one of the widely used image segmentation technique. This technique segments object by seeking the grayscale value (0-255) in each element of matrix at threshold value and change them to a new one using Otsu's algorithm (Otsu, 1979).

After the binary image from thresholding is obtained, additional processing algorithms are applied in order to enhance the image. Dilation and erosion were performed using the structure element size 3x3. Every element in it is one. Average filter size 3x3 was used to remove noise. So, this process does not cause excessive dilation or erosion of the region.

2.3 Feature extraction

Each chromosome has characteristic identification. Their characteristic can be used as features. Effective features are quite difficult to find and select. They should be small and sufficiently optimized to classify chromosome in each class because they will affect efficiency and processing time.

2.3.1 Chromosome area

This algorithm brought each segmented chromosome image from the thresholding image in image segmentation step and calculated chromosome area by estimating summation of the area of each pixel in the image. The area of an individual pixel was determined by looking at its 2-by-2 neighborhood. It has six different patterns and details of area calculation from each pattern can be seen in Pratt *et al.* (1991).

2.3.2 Chromosome perimeter

A boundary may be represented by a chain of connected steps of known direction and length. The chain code is a concise way of recording a shape contour. In a two-dimensional image, array movement from one pixel to an adjoining pixel can only be undertaken in one of eight directions, so the eight compass points can be used as direction vectors. By giving a number to each direction, the outline of an object can be traced and coded as a sequence of numbers. The characteristic vector of odd number is like 45 degree from the X or Y axis, and even number is like the X or Y axis. As a result, chromosome perimeter is summation of even counting number and 1.414 times of odd counting number (Awcock *et al.*, 1979)

2.3.3 Singular value decomposition (SVD)

Singular value decomposition is a technique to normalize and reduce matrix size. This algorithm used SVD from grayscale chromosome image and used the maximum value in singular value matrix after normalization as features (Conroy *et al.*, 2000). These values and chromosome size are inversely proportional. However, this relationship can not be seen from singular values in other orders.

2.3.4 Band chromosome area

A band is a part of a chromosome that is clearly distinguishable from adjacent parts by appearing darker and lighter staining intensity with one or more banding techniques. Bands that stain darkly with one method may stain lightly with other methods. The chromosomes are visualized as a continuous series of light and dark bands. The bands are allocated to various regions along the chromosome arms, and the regions are delimited. Bands in chromosome have many resolutions in different stage of the cell cycle. Furthermore, the number of discernible bands depends not only on the stage of condensation but also on the banding technique used. The level of resolution is determined by the number of bands seen in a haploid set (22 autosomes + X and Y). The standard ideograms provide schematic representations of chromosomes corresponding to approximately 300, 400, 550, 700 and 850 bands (Shaffer *et al.*, 2005). This algorithm calculates band chromosome area by segmenting band from the chromosome image using thresholding technique. Threshold value is calculated by gray scale values from chromosome images where the same pixels position of binary image is one.

2.3.5 Band profile

Density profile presents band chromosome sequence. It was calculated from grayscale image using projection at

chromosome image. Mean of summation of grayscale value at projection axis is determined. They are sequence and their values are in the range of 0 and 255. After these values were obtained, they were thresholded to zero and one.

2.3.6 Normalization and feature selection

As a preprocessing step, chromosome images are normalized to be in the interval of zero to one before the features are extracted. In order to divide all chromosomes into 24 categories, seven classifiers are used as shown in Table 1. All features used in this paper include chromosome area (x_1), chromosome perimeter length (x_2), maximum value of singular value matrix from SVD (x_3), band chromosome area (x_4), and band profile ($\boldsymbol{x}_5$). However, we can see that features for each classifier are different. This is due to the fact that the selection of features for each classifier depends on the information from the statistical results of each feature in corresponding chromosome subgroup. For example, classifier G1 has x_2 as a feature because x_2 can separate chromosome 1 to chromosome 2 very well as shown in Figure 5(b).

2.4 Image classification

Artificial neural network is one of the widely used classification technique. It is a mathematic model. It calculates the predicted result from databases that were used in training. Probabilistic neural network was used in the classification. It consists of two layers network. The first layer is radial basis neural network. Weighted inputs are calculated with Euclidean distances function between the selected feature input vectors and training input vectors. Biases are set to 0.8326/spread in resulting in radial basis function that cross 0.5 at weighted input of spread. Spread is the deviation of radial basis function. The second layer is competitive neural network. Weighted input is set to target vector. And transfer function is compete function. Compete transfer function accepts the input vector that associated with that maximum probability of particular class input and this layer has no bias (Wasserman, 1993).

3. Materials and Methods

Chromosome pictures captured from microscope and digital camera have been used in this paper (60 metaphase cell pictures in JPEG form, 30 pictures of female and 30 pictures of male). All metaphase chromosome cells were stained with Giemsa- stained technique in several different stages of cell cycle.

This experiment starts from image preprocessing process. Original chromosome pictures were transformed from RGB image to gray scale image, reduced noises and improved images. The second process is image segmentation. This process segments chromosome images, band chromosome images and chromosome image perimeter using thresholding one and two levels and chain code technique, respectively. Singular value decomposition was implemented to calculate the diagonal matrix for seeking the maximum value in this matrix from gray scale images. Band profile was obtained by thresholding from density profile. Density profile was calculated from mean of summation of gray scale value in chromosome image using projection technique. The result from this process can be transformed as features. The features must be normalized and were selected optimized features when processing in feature extraction process. The final process is image classification. The probabilistic neural network was employed as a classifier. This process classifies the classification method in two steps. The first step classifies chromosome images to one of six groups. The second step classifier chromosome images in each group again with their classifier as shown in Figure 3. Finally, 2-fold cross validation was used for training and testing to classification method. This experiment compares the classification results when sampling the density profiles, 10, 30, 50 and 80 values.

Table 1. Features used in each classifier.

Classifier	Features
M	x_1, x_2, x_3
G1	$x_2, \boldsymbol{x}_5$
G2	$x_1, x_2, x_3, x_4, \boldsymbol{x}_5$
G3	$x_1, x_2, x_3, x_4, \boldsymbol{x}_5$
G4	$x_4, \boldsymbol{x}_5$
G5	$x_1, x_2, x_4, \boldsymbol{x}_5$
G6	$x_1, x_2, x_3, x_4, \boldsymbol{x}_5$

4. Results and Discussion

Figure 5(a) and 5(b) shows statistical results of chromosome area and chromosome perimeter. Their results are

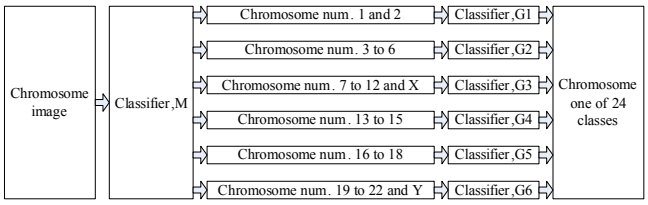


Figure 3. Block diagram of classifiers.

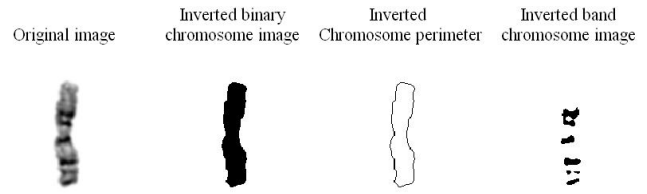


Figure 4. Example of segmented chromosome image and its features.

quite similar. Chromosome class number 1 to 22 are arranged from the large value to the small one. Chromosome class number 23 (chromosome X) is in about middle of range. Chromosome class number 24 (chromosome Y) is the smallest one.

Figure 5(c) shows statistical results of maximum values in diagonal matrix from singular value decomposition of grey scale image. They are reversed version of chromosome area and chromosome perimeter length features. In other words, they are arranged from the small value to the large one. Chromosomes class number 23 (chromosome X) is in about middle of range. Chromosomes class number 24 (chromosome Y) is the largest one.

Figure 5(d) shows statistical results of band chromosome area feature. Part of them is not in agreement with results from area and perimeter. For example, chromosome class number 19 and 20 can be separated by the band chromosome area feature but not by area and perimeter.

Figure 6 shows the density profiles. The density profiles can describe band chromosome. Chromosome class number 1 is shown in Figure 6 (a) and its density profile is shown in Figure 6(b). Figure 6 (c) shows band profile that is obtained from thresholding of the density profile values.

Table 2 shows a confusion matrix of sample classification. It is from 30 male chromosome images when the density profiles are sampled at 50. The best results are chromosome class number 1 but the worst ones are chromosome class number 22 and chromosome Y.

Table 3 shows preliminary efficiency result of probabilistic neural network from female and male pictures when the density profiles are sampled at 10, 30, 50 and 80. The best final result of female is 68.19% when 30 samples were used and that of male is 61.30% when 50 samples were used. Processing time of female and male is approximate 3.54 and 3.58 seconds, respectively.

Owing to the chromosomes that were classified mistaken from the first step can not be determined the efficiency of classifier in the second step. It means that the classifier in the second step was determined classification efficiency without chromosomes that were classified incorrectly. Efficiency of the first classifier of female and male is 92.39% and 89.49%, respectively, as shown in Table 3. The best efficiency in the second step is from classifier G1 in both female and male. But the worst efficiency is from classifier G6 as shown in Table 4.

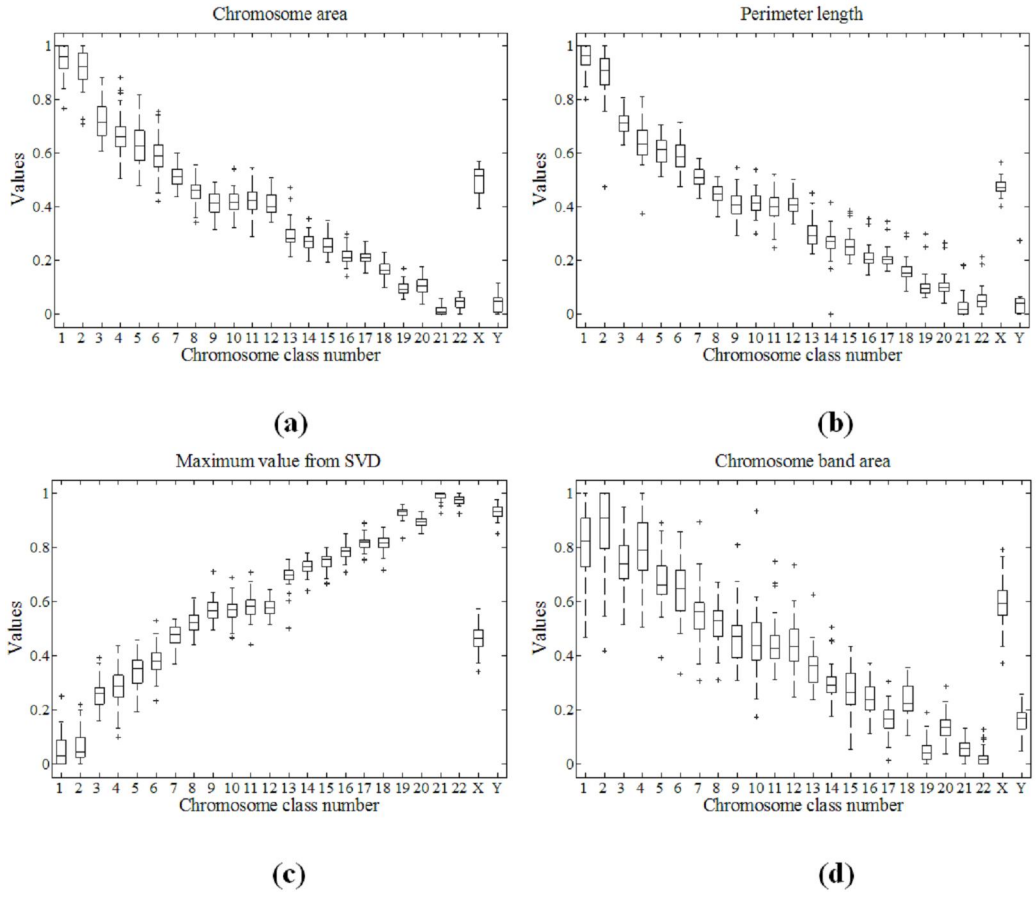
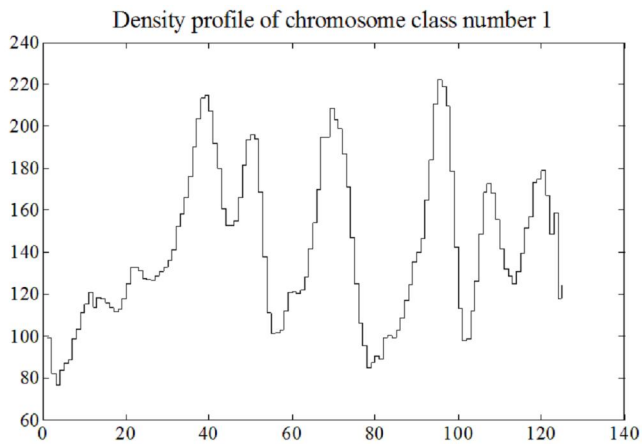
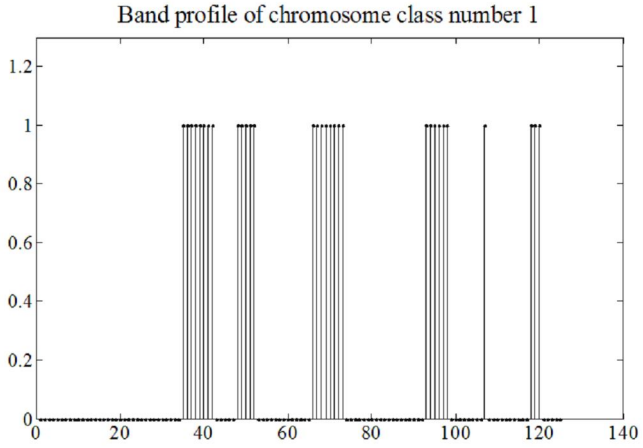


Figure 5. Plot of features statistics resulting from 40 metaphase chromosome pictures. X-axis is chromosome number and Y-axis is feature values after normalization: (a) chromosome area, (b) chromosome perimeter length, (c) maximum value in diagonal matrix from SVD, (d) band chromosome area.

(a)



(b)



(c)

Figure 6. (a) Chromosome number 1. (b) Density profile of chromosome number 1. (c) Band profile.

Table 2. Confusion matrix when band profile was sampled at 50 from 30 male chromosome images.

		Actual																						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
Predicted	1	58	18	4																					
	2	2	38																						
	3		4	48	10	4	7																	1	
	4			1	46	4	2					1													
	5			2	1	47	3	2																1	
	6			5	1	2	30	2																	
	7					2	7	44	5	11	5	3		1		2									
	8				1		3	3	42	13	1	1	1	4										1	
	9						5	5	9	28	1	2	1	2	1	1								3	
	10						2			1	48		1		2		1							1	
	11							1	2	2	2	47	4	1											
	12				1	1	1	1		2	1	3	52			2								1	
	13									1	1	1		49	2	4	9	2	1						
	14														44	2	11	3							
	15													2	1	40	8								
	16													1		4	23	3	1	1	4				
	17														6	1	2	50	1	1	2				1
	18														4	4	6	1	41	1	3				
	19																			24	38	40	14		
	20																	1	1	7	9	2	46		4
	21																		15	26	4	18			20
	22																						0		5
	X							2	2	2	1	2	1											22	
	Y																								0

Table 3. Performance of chromosome image classifier.

Density profile (sample number)	Female			Male		
	Efficiency (%)		Time (second)	Efficiency (%)		Time (second)
	1st classifier	2nd classifier		1st classifier	2nd classifier	
10	92.39	57.68	3.53	89.49	52.75	3.58
30	92.39	68.19	3.54	89.49	60.29	3.58
50	92.39	67.03	3.55	89.49	61.30	3.59
80	92.39	53.48	3.57	89.49	48.04	3.60

Table 4. Performance of each chromosome image classifier in the second step.

Density profile (sample number)	Group	Efficiency,Female (%)		Efficiency,Male (%)	
		(Real)	(Ignore)	(Real)	(Ignore)
10	1	83.33	84.03	79.17	82.61
	2	57.08	61.43	48.33	54.46
	3	62.14	64.29	54.10	55.82
	4	75.56	82.93	68.89	86.11
	5	71.67	92.14	65.56	92.19
	6	20.83	22.22	30.00	31.27
30	1	90.00	90.76	86.67	90.43
	2	77.50	83.41	69.17	77.93
	3	74.05	76.60	68.72	70.90
	4	84.44	93.83	72.78	90.97
	5	70.00	90.00	63.33	89.06
	6	24.17	25.78	18.15	18.92
50	1	88.33	89.08	80.00	83.48
	2	79.17	85.20	71.25	80.28
	3	75.95	78.38	72.05	74.34
	4	81.67	90.74	74.44	93.06
	5	71.11	91.43	63.33	89.06
	6	22.08	23.56	19.63	20.62
80	1	55.00	55.00	58.33	60.87
	2	50.42	54.26	42.08	47.42
	3	60.95	63.05	59.23	61.11
	4	67.78	75.31	57.22	71.53
	5	64.44	82.86	57.22	80.47
	6	24.17	25.78	19.63	20.54

5. Conclusions and Discussion

This paper proposed an algorithm for chromosome image classification. The technique consists of image pre-processing, image segmentation, feature selection, and image classification. Sixty chromosome pictures including 2760 chromosomes (60x46) were used in validation. Average filter was used in image preprocessing. Thresholding technique was implemented in image segmentation. The extracted features of chromosome are area, perimeter length, band's area, singular value decomposition, and density profile. Finally, probabilistic neural network was used in image classification. The classification method of chromosome images can be divided into two steps: Six groups in the first step and twenty four classes in the second steps.

Probabilistic neural network used in the algorithm consists of two steps. Efficiency of the first step for female is 92.39% and that for male is 89.49%. The best classification result of female is 68.19% when density profile at 30 samples was used and that of male is 61.30% when density profile at 50 samples was used. Processing time of female and male is 3.54 and 3.59 seconds, respectively. The classification results

of female are better than those of male because number of female's chromosome class is 23 and male is 24. In other words, female does not have chromosome Y.

The best classification rate of female chromosome from our previous work based on back-propagation neural network (Rungruangbaiyok *et al.*, 2009) is 66.41% and that of male is 64.78%. On the other hand, the best classification rate of female chromosome from the new proposed method based on probabilistic neural network is 68.19% and that of male is 61.30%. We can see that the classification rates of both methods are comparable. However, probabilistic neural network offers several advantages over back-propagation neural network. That is, its training time is much faster and usually a single pass. In addition, it allows true incremental learning where new training data can be added at any time without requiring retraining of the entire network. We added this comparison in the conclusion of the new manuscript.

Acknowledgement

Thank Ms. Jariya Khaiman from Department of Pathology, Faculty of Medicine, Prince of Songkla University for giving knowledge and chromosome karyotype pictures.

References

Awcock, G.J. and Thomas, R. 1996. Applied Image Processing, McGraw-Hill, Singapore, pp. 152-160.

Boehm, D., Herold, S., Kuechler, A., Liehr, T. and Laccone, F. 2004. Rapid detection of subtelomeric deletion/duplication by novel real-time quantitative PCR using SYBR-green dye. Human Mutation. 23(4), 368-378.

Cho, J., Ryu, S.Y. and Woo, S.H. 2004. A study for the hierarchical artificial neural network model for giemsa stained human chromosome classification. Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, CA, U.S.A., September 1-5, 2004, 4588-4591.

Conroy, J.M., Kolda, T.G., O'Leary, D.P. and O'Leary, T.J. 2000. Chromosome identification using hidden Markov models: comparison with neural networks, singular value decomposition, principal components analysis, and Fisher discriminant analysis. Laboratory investigation; A Journal of Technical Methods and Pathology. 80(11), 1629-1641"

Delshadpour, S. 2003. Reduced size multilayer perceptron neural network for human chromosome classification. Proceedings of the 25th Annual International Conference of the IEEE EMBS, Cancun, Mexico, September 17-21, 2249-2252.

Hampton, G.M., Larson, A.A., Baergen, R.N., Sommers, R.L., Kern, S. and Cavenee, W.K. 1996. Simultaneous assessment of loss of heterozygosity at multiple microsatellite loci using semi-automated fluorescence-based detection: subregional mapping of chromosome 4 in cervical carcinoma. Proceedings of the National Academy of Sciences of the United States of America. 93(13), 6704-6709.

Kyan, M.J., Guan, L., Amison, M.R. and Cogswell, C.J. 1999. Feature extraction of chromosomes from 3D confocal microscope images 1999 International conference on image processing Vol. 2, 24-28.

Masters, T. 1993. Practical Neural Network Recipes, Wiley, New York, Chapter 12.

Masters, T. 1995. Advanced Algorithms for Neural Networks: A C++ Sourcebook, Wiley, New York, Chapter 4.

McAndrew, A. 2004. Introduction to Digital Image Processing with MATLAB. Thomson Leaning, Inc., Massachusetts, USA, pp. 378-379."

Musavi, M.T., Chana, K.H., Hummelsa, D.M., Kalantria, K., and Ahmeda W. 1992. A probabilistic model for evaluation of neural network classifiers. Pattern Recognition. 25(10), 1241-1251"

Otsu, N. 1979. A Threshold Selection Method from Gray-Level Histograms, The Institute of Electrical and Electronics Engineers Transactions on Systems, Man, and Cybernetics, 9(1), 62-66.

Piper, J., Granum, E., Rutovitz, D. and Ruttledge, H. 1980. Automation of chromosome analysis. Signal Processing. 2(3), 203-221.

Pratt, William, K. 1991. Digital Image Processing, New York, John Wiley & Sons, Inc., pp. 634.

Rungruangbaiyok, S. and Phukpattaranont, P. 2009. Improvement in Chromosome Image Classification Based on Neural Network. The Seventh PSU Engineering Conference, Songkhla, Thailand, May 21-22, 2009, 235-240.

Shaff, L.G. and Tommerup, N. 2005. An International system for human cytogenetic nomenclature (2005), Karger Publishers. Basel, Switzerland. pp. 6-31.

Specht, D.F. 1990. Probabilistic neural networks and the polynomial adaline as complementary techniques for classification, The Institute of Electrical and Electronics Engineers Transactions on Neural Networks. 1(1), 111-121.

Specht, D.F. 1990. Probabilistic neural networks. Neural networks, 3(1), 109-118.

Sweeney, Jr., W.P., Musavi, M.T. and Guidi, J.N. 1994. Classification of chromosomes using a probabilistic neural network. Cytometry. 16(1), 17-24.

Truong, K., Gibaud, A., Zalcman, G., Guilly, M.N., Antoine, M., Commo F., Fouquet C., Cadranel J., Soussi T., Dutrillaux B., and Malfoy B. 2004. Quantitative fish determination of chromosome 3 arm imbalances in lung tumors by automated image cytometry. Medical science monitor: international medical journal of experimental and clinical research. 10(11), 426-432.

Wasserman, P.D. 1993. Advanced Methods in Neural Computing, Van, Nostrand Reinhold, New York, pp. 35-55.