A STUDY FOR THE FEATURE SELECTION TO IDENTIFY GIEMSA-STAINED HUMAN CHROMOSOMES BASED ON ARTIFICIAL NEURAL NETWORK

Seung Yun Ryu, Jong Man Cho, Seung Hyo Woo
Department of Biomedical Engineering, Inje University, Kimhae, Korea

Abstract – **Many studies for computer-based chromosome analysis have shown that it is possible to classify chromosomes into 24 subgroups. In addition, artificial neural network (ANN) has been adopted for the human chromosome classification. It is important to select optimum features for training neural network classifier. We selected some features - relative length, normalized density profile (d.p) and centromeric index - used to identify chromosomes and trained neural network classifier changing the number of samples which used to get d.p. We found the fact that the classification error showed to be minimum when this number was equal to or greater than the length of No.1 human chromosome.**

Keywords **- Artificial neural network, Chromosomes, Feature selection**

I. INTRODUCTION

Human chromosome analysis is an essential task in cytogenetics, especially in prenatal screening and genetic syndrome diagnosis, cancer pathology research and environmentally induced mutagen dosimetry [3]. Cells used for chromosome analysis are taken mostly from amniotic fluid or blood samples. The stage at which the chromosomes are most suitable for analysis is the metaphase. One of the aims of chromosome analysis is the creation of a karyotype, which is a layout of chromosome images organized by decreasing size in pairs. The karyotype is obtained by all of procedures for cell culture, preparing slides, selection of best-observable chromosome image, analysis and classification. However, even today, chromosome analysis and karyotyping are manually performed in most cytogenetics laboratories in a repetitive, time consuming and therefore expensive procedure [2].

Therefore, automatic chromosome analysis has attracted much attention. So efforts to develop automatic chromosome classification techniques has been made during the last 20 years. ANN is suitable for automatic chromosome classification because the human chromosome images (see Fig. 1) have nonlinear properties. The purpose of this study is to find the best features for neural network classifiers.

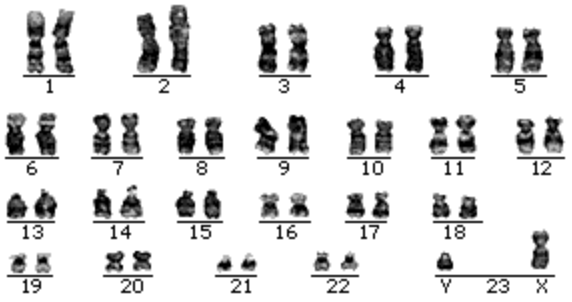

Fig. 1 The human chromosome images

This research focused on the feature selection to minimize the classification error.

II. METHODOLOGY

A. Chromosome Data

The suggested methodology is applied to Edinburgh database. The chromosome database was provided by Dr. Piper and exported from the Medical Research Council Human Genetic Unit, Edinburgh, UK.

B. Chromosome Feature Extraction

We got the skeleton of a chromosome by applying the medial axis transformation (MAT). The MAT is widely used as a convenient transformation for the representation of elongated objects, for example in character recognition or chromosome analysis, where the width of the objects contains little useful information [5]. In addition, we used the thinning algorithm that iteratively deletes edge points of a region subject to the constraints that deletion of these points does not remove end points, does not break connectedness, and does not cause excessive erosion of the region [4].

1) *Relative length*: One of the significant morphological features used to identify a chromosome is a length characteristic [1]. After applying MAT, we extended the skeletonized line to the boundary and got the length of the chromosome (the number of pixel in the skeletonized line). Because length varies according to the phase of the cell division, the length must be normalized. The relative length of the i-th chromosome (l_{ri}) can be obtained by normalizing the medial axis length using the following equation:

$$l_{ri} = \frac{l_i}{l_t} \tag{1}$$

where l_i (i = 1,2,…,24) is the length of i-th chromosome and l_t is the total length of all 46 chromosomes of one cell.

2) *Centromeric index*: The centromeric index (C.I.) is the ratio of the length of the short arm to the whole length of a chromosome. It is another significant morphological feature used to identify the chromosome.

$$C_i = \frac{\text{short arm length}}{\text{whole length of medial axis}} \tag{2}$$

where C_i is the C.I. This index, which indicates the location of the centromere of the chromosome, is obtained from the shape profile. The shape profile for a chromosome is obtained by measuring the width along a transverse line, perpendicular to the tangent of the medial axis and centered at unit distance along the medial axis [1].

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle A Study for the Feature Selection to Identify GIEMSA-Stained Human Chromosomes Based on Artificial Neural Network	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Department of Biomedical Engineering Inje University Kimhae, Korea	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
2

3) *Normalized density profile*: The Giemsa-stained human chromosome has a sequence of banding pattern that is perpendicular to the medial axis of the chromosome. Density profile is a one-dimensional graph of the banding pattern property of the chromosome computed at a sequence of points along the possibly curved chromosome medial axis. The density profile for a chromosome is obtained from measurements made along a transverse line, perpendicular to the tangent of the medial axis. Each profile value (I_i) results from the summing properties of points spaced at unit distance apart along each transverse line. To reduce the variation of the density values due to the different cell culturing conditions, this density profile is normalized three times by three different methods. First, this profile ($d_w(i)$) is normalized in the direction of the perpendicular to the medial axis of the chromosome to reduce the variation of width of a chromosome.

$$I_i = \sum_{j=0}^{m-1} d(i, j) \qquad (i=0,1,\ldots,n-1) \quad (3)$$

where m is the number of pixel on the line of perpendicular to the tangent of the medial axis and $d(i,j)$ is the pixel value on the line of perpendicular to the tangent of the medial axis.

$$d_w(i) = \frac{I_i}{w(i)} \qquad (i=0,1,\ldots,n-1) \quad (4)$$

where $w(i)$ is the width of i-th point in a chromosome.

Next, histogram equalization is applied to the profile to reduce the effect of nonhomogeneous illumination conditions of the microscope. Finally, the profile is normalized along the medial axis to obtain the same number of normalized density values ($d_N(i)$) regardless of chromosome length [1].

$$d_N(i) = \frac{d_w(i) - d_{wMIN}(i)}{d_{wMAX}(i)} \qquad (i=0,1,\ldots,n-1) \quad (5)$$

where $d_{wMIN}(i)$ is the minimum density value and $d_{wMAX}(i)$ is maximum.

In this study, five normalized density profiles were obtained and were compared the classification error to each case.

C. Chromosome Data Sets

Five data sets were prepared to select the best number of density value for one input pattern. These sets were the same (relative length + C.I + d.p) except for the number of density value. Each training set consisted of 460 input patterns, extracted from 460 chromosomes.

In addition to training sets, test sets were prepared to test the trained neural network classifier. Each test set had the same number of input data features as training set, but the features were extracted from chromosomes that were not used to prepare the training sets.

D. Neural Network Classifier

In this study, ANN as a classifier was examined. A two-layer neural network with an error backpropagation training algorithm was adopted in this study. The neural network classifier was implemented by software, in which the number of output nodes was fixed at 24, but the number of input output nodes was fixed at 24, but the number of input nodes, the number of processing elements in the hidden layer, the number of learning times, leaming constant, momentum term, and upper bound of error value were programmable.

III. RESULTS

We trained each neural network classifier with each feature set (25, 50, 70, 80, and 100 density values including C.I and relative length). We applied the test sets to the trained neural network and obtained each classification error shown in Table I.

The error showed to be minimum when the number of input node was 102 including relative length, C.I and 100 density values.

IV. DISCUSSION

We selected the features - relative length, C.I and d.p - to identify a chromosome and examined trained sets only changing the number of the density value to obtain the best number of density value of a chromosome. However, other features were not used in this study. We need to study to get other chromosome features that reduce the classification error.

V. CONCLUSION

We found the fact that the classification error showed to be minimum when the number of density values was equal to or greater than the length of No.1 human chromosome. The two-layer neural network trained with the error backpropagation training algorithm showed good potential in classification of Giemsa-stained.

TABLE I
CLASSIFICATION ERROR

	The number of density value				
	25	50	70	80	100
Error (%)	10.2	7.2	5.8	5.1	3.6

REFERENCES

[1] J. M. Cho, "Chromosome classification using backpropagation neural network," *IEEE Trans. Med. and Bio. Magn.*, vol. 19, pp. 28-33, January 2000.

[2] B. Lerner, "Toward a completely automatic neural-network-based human chromosome analysis," *IEEE Tran. Systems Man. and Cybernetics*, vol. 28, pp. 544-552, August 1998.

[3] A. Carothers and J. Piper, "Computer-aided classification of human chromosomes: A review," *Stat. Comput*., vol. 4, pp. 161-171, 1994.

[4] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Reading, MA: Addision-Wesley, ch. 8, 1992.

[5] B. Lemer, H. Guterman, I. Dinstein and Y. Romen, "Medial axis transform-based features and a neural network for human chromosome classification," *Pattern Recog*., vol. 28, No. 11, pp. 1673-1683, 1995.