Classification of Human Chromosomes by Two-Dimensional Fourier Transform Components

Boaz Lerner, Hugo Guterman and Its'hak Dinstein

Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel 84105

Abstract

The two-dimensional Fourier transform (FT) components of a chromosome image were used as features of a neural network classifier. The amplitudes and phases of the FT components within windows of different sizes were used as input to a 2-layer feedforward neural network trained by a modified backpropagation algorithm. Several combinations of windows of size 2x2 to 6x6 for both the amplitude and the phase were tested. The data base consisted of 481 images of 5 different types of human chromosomes obtained from 150 cells. For all the combinations of windows and features, correct classification probabilities for training and for testing were 94-100% and 81.3-91.7%, respectively. The best results were accomplished using a 4x4 window for both amplitude and phase. With only 1770 epochs for training, correct classification of 99.3% for training and 91.7% for testing were achieved.

Introduction

Most people have 46 chromosomes in nearly every nucleated cell of their body. The chromosomes carry the genetic information used in the development of the individual. In most people the set of chromosomes is identical in almost all cells, and they appear as 22 pairs of autosomes and 2 sex chromosomes, one of each pair being inherited from the father and the other from the mother [1].

Human chromosome inspection is a vital task in cytogenetics especially in clinical prenatal analysis, genetically syndrome diagnosis (e.g, Down's syndrome), cancer pathology research and environmentally induced mutagen dosimetry [1], [2]. Cells used for chromosome inspection are taken mostly from amniotic fluid or blood samples. The stage at which the chromosomes are most suitable for analysis is the metaphase (Fig. 1). One of the aims of chromosome analysis is the creation of a karyotype, which is a layout of images of chromosomes taken from a cell, organized by decreasing size in pairs (Fig. 2). The karyotype is obtained by cutting a microscopic photograph of the cell and arranging the chromosomes into their appropriate places on the layout according to their visual classification by the cytotechnician. Karyotyping is a useful tool for detecting deviations from normal cell structure. Abnormal cells can have an excess or deficit of a chromosome and/or structural defects like breaks, fragments or translocations (exchange of genetic material between chromosomes). However, even today this inspection and karyotyping are performed manually in most of the cytogenetic laboratories in a time consuming, repetitive and expensive procedure [2], [3].

Efforts to automatically classify the chromosomes have been made through the last 40 years. Computerized chromosome analysis basically consists of several stages: pre-processing, segmentation, intermediate processing and feature extraction and finally, classification. The pre-processing stage aim is to improve the cell image by techniques of noise removal, edge enhancement and/or contrast improvement. The segmentation purpose is to isolate the metaphase chromosomes from the background and from undivided cell nuclei and irrelevant biological materials within the image. After segmentation, some intermediate processing is needed. This processing is usually limited to local algorithms like median axis transformation, centromere finding, etc. Feature extraction can be further performed on the basis of these algorithms. The result of the feature extraction is a more condensed representation that still retains most of the important chromosome information. It is easier and faster to classify

based on features than on the picture itself and usually it yields more information. Different features were used to describe chromosomes, e.g, the features of size (length, area, ...) [1], the centromeric index (the ratio of the short arm size to the whole chromosome size) [1], [4] and the band pattern descriptors (like density profile along the medial axis [1], [4], Fourier or Gaussian decomposition of the density profile [5], [6], histogram of gray levels [7], etc.) . In the final stage, the classification is generally made by statistical methods (e.g, linear or quadratic discriminator, distance functions, Bayes rule) [1], [2].

All the efforts to make the chromosome analysis automatic had limited success and poor classification results compare to those of a trained cytotechnician [1]-[4]. Some of the reasons to the poor performances are the inadequate use of the expert knowledge and experience and the insufficient ability to make comparisons and/or eliminations among chromosomes within the same metaphase. In addition, the systems always require the operator interaction to separate touching and/or overlapping chromosomes and to verify the classification results [1], [2].

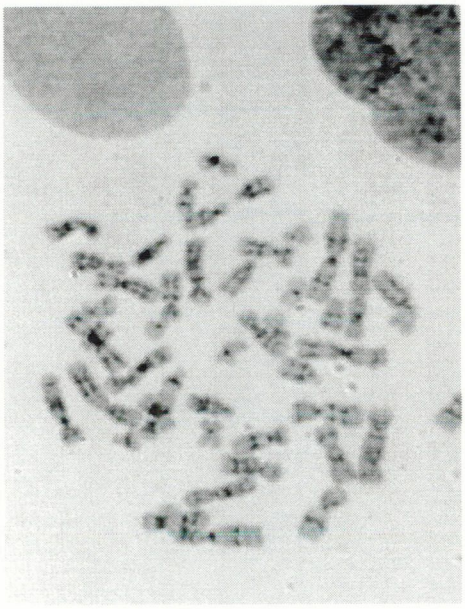

Figure 1. A human metaphase cell, x1000

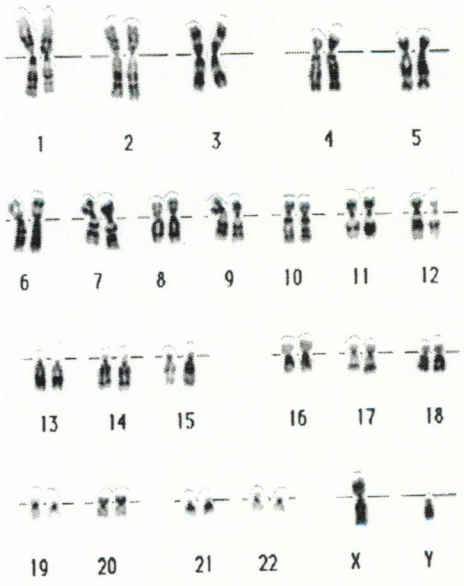

Figure 2. A karyotype of a metaphase cell

Neural networks have the potential to overcome most of these limitations. This is mainly because they permit application of expert knowledge and experience through network training. Furthermore, human chromosome classification based on neural networks requires no a-priori assumptions or knowledge of the data to be classified as conventional methods need. Finally, it is well known that the problems best solved by neural networks are those that humans do well, and classification of chromosomes is one of them.

Methodology

In this work, images of amniotic fluid cells were acquired from the Institute of Medical Genetics of Soroka hospital, Beer-Sheva. The pictures were obtained with the aid of a light microscope and captured by a CCD camera (Cohu). The pictures were further digitized with a frame graber (VISIONplus-AT) (Fig. 3). The size of the digitized picture was 512 X 768 pixels and each pixel was represented by 1 byte (256 gray levels). No pre-processing techniques were applied. The segmentation was made manually using a graphical software on a 486 PC computer. Each chromosome was stored on a file of size 150 X 64 pixels (this file size was set due to the largest chromosome size and to some system requirements). Chromosomes of 5 different types, namely types "2", "4", "13", "19" and "x" were extracted from more than 150 different cells. One set of chromosomes included 481 chromosomes that were arranged so their positions and orientations in the files were almost the same (e.g, at file

top-left corner with the "head" of the chromosomes up). The second set included 600 chromosomes with random positions and orientations.

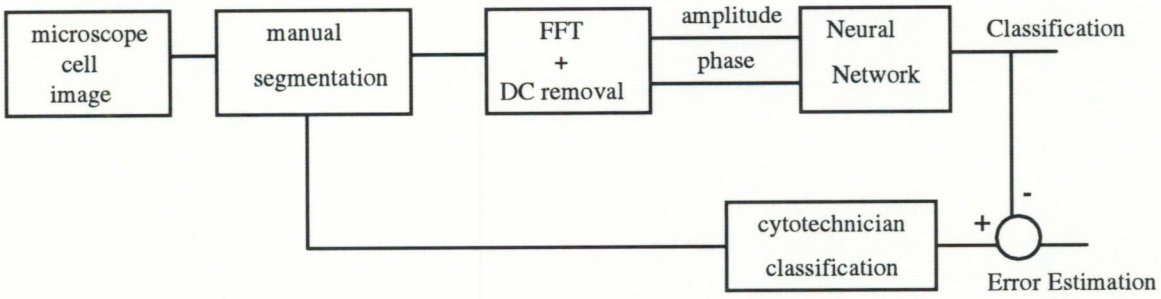

Figure 3. Layout of the methodology

For chromosome images, the primary features are sets of bands along the chromosome, perpendicular to the long axis [1], [5]. This band pattern of the chromosomes suggests that the Fourier transform components can represent some important information of the chromosomes. More than that, when analyzing pictures, the complexity and the time required to extract features are crucial factors. Therefore, two-dimensional Fourier transform components have some advantages in representing a picture. They are easily computed and do not require any previous calculations.

There is an optimum for the number of components (features) required for correct reconstruction of the pattern from its components. Caspersson et al. [6] found out that in the one-dimensional case up to 8 amplitude and 4 phase components of the FT were required to synthesize all chromosome patterns with only very slight deviations from the original pattern. In this study we examined the amplitude and phase, as well as their combinations, of the first 2 to 10 two-dimensional Fourier transform components. These components were arranged in a "window" around the dc component. The dc component did not carry any important information about the picture and it was removed. No further use of components of higher order is needed, as was empirically found in this study and previously confirmed [5], [6].

A two-layer feed forward neural network trained by the backpropagation learning rule was chosen for the classification (Fig. 3). The network that used in this research involved two modifications to the backpropagation algorithm [8]. The first is the application of a Principal Component Analysis (PCA) for finding the network structure (number of hidden units) and providing the initial weights. The second is the use of a conjugate gradient algorithm for searching the minima. In this way we can design the network structure by employing the a-priori information contained in the data covariance matrix. The input vector was based on the Fourier components linearly normalized in the [-0.5, 0.5] range. The output vector was 5-dimensional with one component set to "1" (actually 0.9) for the correct classification and "0" (actually 0.1) elsewhere. Training and test vectors were chosen randomly from the same data set. The number of training vectors was 70-90% of all the vectors and the remaining vectors were reserved for testing.

Results

Several simulations were made with different features, window sizes, data sets and network configurations. Classification probabilities were higher when features based on both amplitude and phase were employed. Combinations of 2x2 to 6x6 windows with phase information and 2x2 to 4x4 windows with amplitude information gave best results. These results correlate well with other studies on the one-dimensional case [5] ,[6]. Also, the classification was better when only chromosomes from the first set were used. The number of the hidden units was varied in the range 5-22 (depend on the PCA results).

The outcomes for all sizes of windows were chosen by a voting procedure. Using only 1000-2000 epochs for training, 94-100% and 81.3-91.7% correct classification probabilities were achieved for training and testing, respectively. Fig. 4 outlines the classification probabilities for a window of size 4x4 with features based on amplitude, phase and their combination. Best results of all the simulations were 99.3% for training and 91.7% for

testing. These results were achieved for both amplitude and phase within a window of size 4x4 and with only 6 hidden units and 1770 epochs for training.

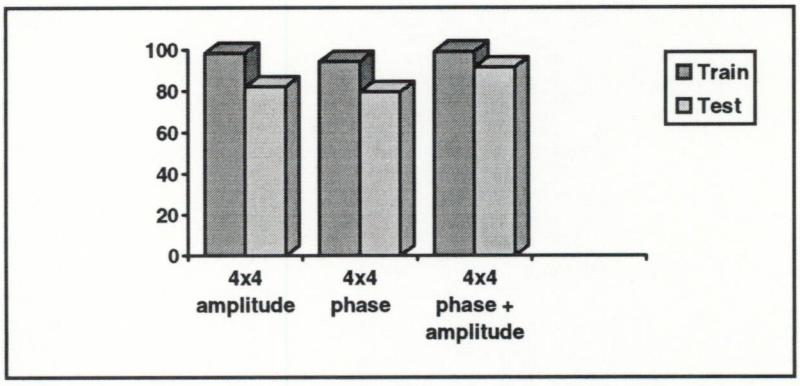

Figure 4. Correct classification probabilities for a 4x4 window

Conclusions

Human chromosome classification based on features of the 2D Fourier decomposition yield very promising results. The number of FT components essential to comprehensively represent images of chromosomes is relatively small. Hence, the feature extraction and both neural network training and testing are relatively simple and fast. The major disadvantage of the Fourier features is that they are global and not invariant to rotation and scale change. Therefore, some rotation of the data may be necessary. Larger data bases of chromosomes in a variety of orientations and scales may help to overcome the invariance problem.

Correct classification probability of 91.7% while testing, based on global features as the FT components and after only 1770 epochs for training, is very encouraging, especially when compared to the months required to train an expert cytotechnician.

Acknowledgment

The authors thank Mrs. Z. Reizina from the Institute of Medical Genetics, Soroka hospital, Beer-Sheva, for her cooperation in data preparation and for discussions concerning clinical aspects.

References

1. Piper, J., Granum, E., Rutovitz, D. & Ruttledge, H. (1980). Automation of chromosome analysis. *Signal Processing*, **2**, 203-221.
2. Wu, Q., Suetens, P. & Oosterlinck, A. (1987). Toward an expert system for chromosome analysis. *Knowledge-Based Systems*, **1**, 43-52.
3. Vanderheydt, L., Oosterlinck, A., Van Daele, J. & Van Den Berghe, H. (1980). Design of a graph-representation and a fuzzy-classifier for human chromosomes. *Pattern Recognition*, **12**, 201-210.
4. Groen, F.C.A., ten Kate, T. K., Smeulders, A. W. M. & Young, I. T. (1989). Human chromosome classification based on local band descriptors. *Pattern Recognition Letters*, **9**, 211-222.
5. Granlund, G. H. (1976). Identification of human chromosome by using integrated density profile. *IEEE Transactions on Biomedical Engineering*, **BME-23**, 182-192.
6. Caspersson, T., Lommaka, G. & Moller, A. (1971). Computerized chromosome identification by aid of the quinacrine mustard fluorescence technique. *Hereditas,* **67**, 103-110.
7. Lerner, B., Guterman, H. & Dinstein, I. (1992). On classification of human chromosomes. *Neural Networks for Learning, Recognition and Control*, a research conference at Boston University, May 14-16.
8. Guterman, H. Application of Principal Component Analysis to the Design of Neural Networks (submitted).