Medial Axis Transform Based Features and a Neural Network for Human Chromosome Classification

Lerner, B., Rosenberg, B., Levinstein, M., Guterman, H., Dinstein, I. and Romem, Y.*
Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel 84105

* The Institute of Medical Genetics, Soroka Medical Center
Beer-Sheva, Israel 84105

Abstract

Medial axis transform (MAT) based features and a two-layer feedforward neural network were used in this study for human chromosome classification. Two approaches to the MAT, the "skeleton" and the piecewise linear (PWL), were examined. The medial axis based on the "skeleton" approach, as well as, the chromosome classification results based on this approach were slightly better than these of the PWL approach. Several chromosome features, like the density profile, the centrometric index and the length of the chromosome, as well as, combinations of them, were tested. The probability of correct training set classification using all the available features and the neural network classifier was almost perfect (99.3-99.6%). The probability of correct test set classification was greater than 97% using features based on the "PWL" approach and over 98% using features based on the "skeleton" approach.

1. Introduction

Human chromosome inspection is a vital task in cytogenetics, especially in clinical prenatal analysis, genetical syndrome diagnosis (e.g., Down's syndrome), cancer pathology research and environmentally induced mutagen dosimetry [7], [10]. Cells used for chromosome inspection are taken mostly from amniotic fluid or blood samples. One of the inspection aims is to detect 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 is performed manually in most of the cytogenetic laboratories in a time consuming, repetitive and expensive procedure [9], [10].

Efforts to develop automatic chromosome classification techniques have been made through the last 40 years. However, all the efforts to make the chromosome analysis automatic had limited success and poor classification results compare to those of a trained cytotechnician [2], [7], [9], [10]. Some of the reasons for 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 [7], [10].

Neural networks make it possible 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 some 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.

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Beer-Sheva, Israel.

2. Feature description

Appropriate feature description is considered to be one of the most important part of classification procedures, and in human chromosome classification it is probably the most important one. In some studies, global features, like the histogram of gray levels [3] or the 2D Fourier transform components [4], have been used. In this study, we have employed 3 types of features: the density profile (d.p) along the medial axis [1], [5], [7], the centrometric index (c.i) (the ratio of the short arm length to the whole chromosome length) [2], [5], [7] and the length (lng) of the chromosome [5], [7]. The Medial Axis Transform (MAT) is almost always required for the extraction of these features.

2.1 The MAT

The MAT is widely used as a convenient transformation for elongated objects, e.g., in character recognition or chromosome analysis where the width of the objects contains little (if any at all) useful information. The MAT of an object cannot only reduce storage and time requirements, but also to preserve the topological properties of the object.

Two different approaches to MAT were used in this work, namely, the "skeleton" and the PWL approaches [5]. The "skeleton" approach is based on finding a preliminary medial axis of chromosome via the realization of the fire front's propagation and extinction [11]. This preliminary medial axis is further processed to get one extended continuous medial axis. Removing irrelevant points of the preliminary medial axis on one hand and completion of necessary points on the other hand complete the postprocessing of the medial axis in this approach. The second approach employs a piecewise linear (PWL) approximation [2], [5] to the medial axis. The PWL is preferred over the use of existing polynomial approximation techniques whenever a chromosome is not straight [2].

2.2 Feature extraction

The MAT in both approaches enables us to transform the 2D image of the chromosome to 1D representation. By calculating lines perpendicular to the medial axis points we can integrate (or average) the intensities (gray levels) of all the image pixels along these lines and to obtain a density profile (d.p).

The method we have used in this study to calculate the centrometric index (c.i) is based on searching for the closest pair of opposite contour points on the clipped contours of a chromosome [5], similarly to the method described in [2]. However, instead of using an exhaustive search for the closest pair we searched for the closest pair along the lines perpendicular to the medial axis. No fundamental difference in results of the two methods is expected. However, our method is faster than the method in [2] (there is no exhaustive search of all the pairs of opposite points). The length of the chromosome was calculated along the medial axis.

All the features were further normalized. The d.p feature vector was normalized both in length and in value. Normalizing in length yields suitable feature representation (all classified vectors are in the same dimension) and invariance to scale change. The length of the normalized d.p vector was set to be 64 both from chromosome length and from practical considerations. The 64 values of the d.p vector, the centrometric index and the chromosome length were normalized into the [-0.5,0.5] range, in agreement with the MLP requirements.

3. The neural network classifier

In this research, a two-layer feedforward neural network trained by the backpropagation (bp) learning algorithm [8] was chosen for the chromosome classification. The bp algorithm is an error driven parameter estimation algorithm where the objective is to minimize the output squared error function by

adjusting interconnection weights and node thresholds. The network was initialize using random weights in the [-1,1] range. The number of hidden units of the network was set according to the Principal Components Analysis (PCA), applied to the feature vectors. The number was set to be the number of the largest eigenvalues, the sum of which accounts for more than a pre-specified percentage of the sum of all the eigenvalues [6]. This pre-specified percentage has been called by us "var". In the implementations, the "var" parameter was set to values of 70-90%.

4. Data set

Images of amniotic fluid cells were acquired from the Institute of Medical Genetics of Soroka Medical Center, Beer-Sheva. The pictures were obtained with the aid of a light microscope and captured by a CCD camera (Cohu). The pictures were digitized with a frame grabber (VISIONplus-AT). 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 done manually using a graphical software package on a 486 PC computer. Chromosomes of 5 different types, namely types "2", "4", "13", "19" and "x" were extracted [3], [4] from more than 150 different cells.

For each chromosome the MAT was extracted and the 66 features (64 d.p + c.i + lng) were computed using the procedure described in [5]. Several variations of features were tested to evaluate their importance to the classification procedure, e.g., d.p alone, d.p + c.i, d.p + c.i + lng and c.i + lng. The d.p features were extracted both using the integral representation and the average representation and in both approaches: "skeleton" and PWL.

5. Results

The input vector to the neural network was either 2 or 64-66 dimensional (depend on the type of the features). 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.

Optimization of the neural network parameters regarding the chromosome data is described elsewhere [6]. The learning rate (μ) was set to be 0.026, the momentum constant (α) to be 0.97 and the training cycle was set to be 4000 epochs, although only 500-1000 epochs were required to get almost the best results. Training and test vectors were chosen randomly from the same data set where the number of training vectors was 70-90% of all the vectors (depending on the experiment) and the remaining vectors were reserved for testing. All the simulations were repeated (at least) 3 times, with the same network parameters but with different sets of randomly chosen training vectors, and the results were averaged.

5.1 The PWL vs. the "skeleton" approach to the MAT

Two major conclusions can be made [5] while comparing the PWL and the "skeleton" approaches.

(a). (b).

Figure 1. A comparison of the (a). "skeleton" and (b). PWL approaches to the MAT.

The first is that the medial axis of the "skeleton" approach is finer than the axis of the PWL approach and follows very accurately the chromosome band pattern (Figure 1). The second conclusion, which can be concluded from Figure 2, is that while the probability of correct training set classification is similar in both approaches, the probability of correct test set classification is larger using the "skeleton" features (in about 3-5%). Both conclusions seem to be very close related. Figure 2 depicts the classification results of an experiment in which the percentage of training vectors ("per") is 70-90% of all the vectors and the "var" parameter is set to be 70-90%.

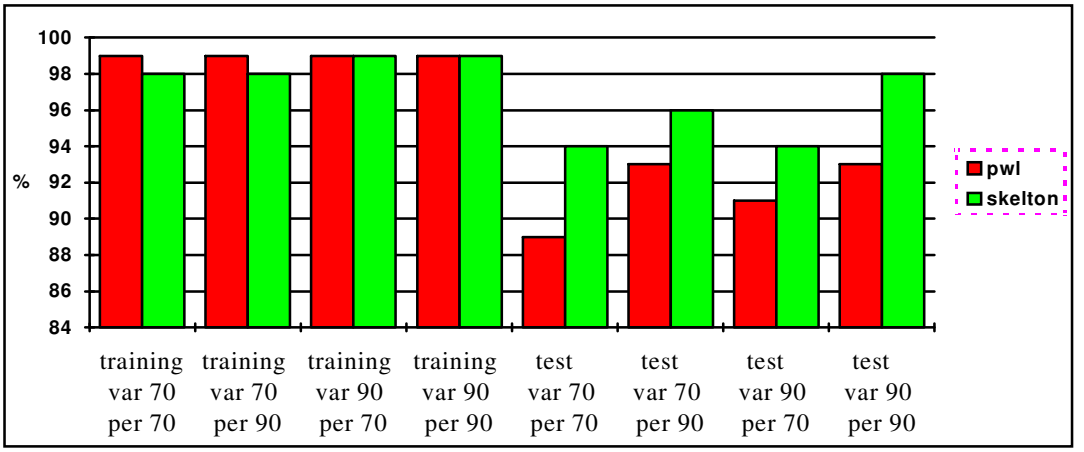

Figure 2. Classification based on the density profile features.

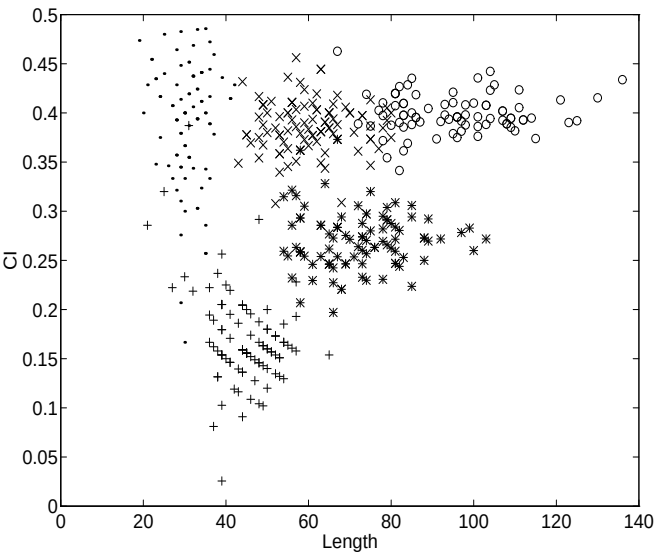

Figure 3. Chromosome clustering into a 2-dimensional feature space spanned by the centrometric index (c.i) and the chromosome length (lng). ("o"- chromosome type "2", "∗"- chromosome type "4", "+"- chromosome type "13", "•"- chromosome type "19" and "x"- chromosome type "x").

5.2 Feature evaluation

The relative importance for the classification procedure of four sets of features was examined. The first set includes only the density profile (d.p) features, while the second set includes, in addition, the centrometric index (c.i). The length of the chromosomes (lng) is the additional feature in the third set. The forth set includes only the (c.i) and the (lng) features. To learn about the significance of these two last features, we have plotted in Figure 3 the two of them one against the other for the entire data set. We can see that these two important features are almost sufficient for the classification of the chromosome data into it 5 types. However, these two features would not be enough when we will try to classify the chromosome data to all its 24 types.

The probability of correct classification of the neural network, using the 4 sets of features, for the PWL approach, is given in Figure 4. The probability of correct classification in the training and in the test sets using the first set of features (d.p) was 99.15-99.5% and 89.3-92.9%, respectively, for various combinations of the two parameters- "per" and "var". The probability of classification of the second set of features (d.p + c.i) was 99.3-99.5% and 92.1-96.45% for training and test, respectively, and this of the third set (d.p + c.i + lng) was 99.3-99.6% and 94.2-97.2% for training and test, respectively. The probability, using only the (c.i) and the (lng) features, was 93.05-94.4% for training and 86.9-92.9% for the test. These results indicate that the 2 features are almost equally important as the 64 d.p features for the classification of the 5 particular classes (types of chromosomes). This conclusion will be definitely changed whenever all the 24 chromosome types will be used. The probabilities achieved using the "skeleton" approach were equal or little higher compare to these of the PWL. It can be clearly seen from the figure the importance of combining different features, especially whenever the "var" is relatively low (small amount of information is retained by the PCA).

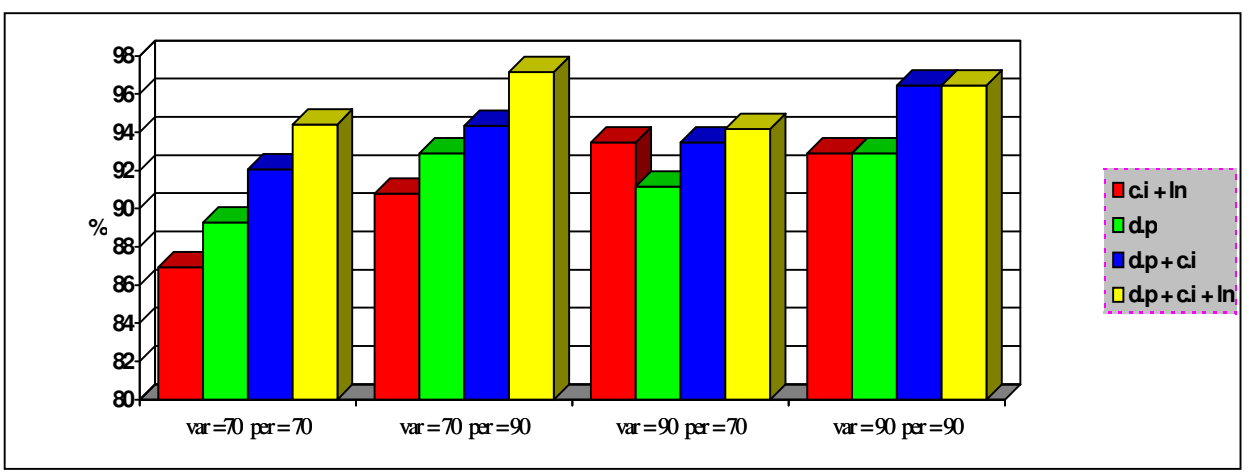

Figure 4. The probability of correct test classification using the 4 sets of features and the PWL approach to the MAT.

6. Discussion and Conclusions

Medial axis transform based features and a two-layer feedforward neural network were used in this study for human chromosome classification. Two approaches to the MAT, the "skeleton" and the PWL, were examined. The medial axis based on the "skeleton" approach, as well as, the chromosome classification results based on this approach were slightly better than these of the PWL approach. Several typical chromosome features, like the density profile, the centrometric index and the length of the

chromosome, as well as, combinations of them, were tested. When classifying only 5 types of chromosomes, as was done in this study, the relative importance of the centrometric index and of the length of the chromosome is very high. The probability of correct training set classification using all the available features and the neural network classifier was almost perfect (99.3-99.6%). The probability of correct test set classification was greater than 97% using features based on the PWL approach and over 98% using features based on the "skeleton" approach.

7. References

1. Granlund, G.H. (1976). Identification of human chromosome by using integrated density profile. *IEEE Transactions on Biomedical Engineering*, **BME-23**, 182-192.
2. 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.
3. 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.
4. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Classification of human chromosomes by two-dimensional Fourier transform components. *WCNN'93*, Portland, July 11-15, 793-796.
5. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Medial axis transform based features and a neural network for human chromosome classification. (Submitted for publication).
6. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Human chromosome classification using multilayer perceptron neural network. (Submitted for publication).
7. Piper, J., Granum, E., Rutovitz, D. & Ruttledge, H. (1980). Automation of chromosome analysis. *Signal Processing*, **2**, 203-221.
8. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning internal representations by error propagation. In Rumelhart, D.E., McClelland, J.L. and the PDP research group, *Parallel Distributed Processing*, vol. 1, chap. 8, Cambridge: MIT Press.
9. 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.
10. Wu, Q., Suetens, P. & Oosterlinck, A. (1987). Toward an expert system for chromosome analysis. *Knowledge-Based Systems*, **1**, 43-52.
11. Xia, Y. (1989). Skeletonization via the realization of the fire front's propagation and extinction in digital binary shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 1076-1086.