Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 19–28
Pattern Recognition Letters
www.elsevier.com/locate/patrec

New features for automatic classification of human chromosomes: A feasibility study

Mehdi Moradi [a,*], S. Kamaledin Setarehdan [b]

[a] School of Computing, Queen's University, Kingston, Ontario, Canada K7L 3N6
[b] Control and Intelligent Processing Center of Excellence, Electrical and Computer Engineering Department, Faculty of Engineering, University of Tehran, Tehran, Iran

Received 27 August 2004; received in revised form 18 March 2005
Available online 29 August 2005

Communicated by L. Goldfarb

Abstract

Karyotyping, a standard method for presenting pictures of the human chromosomes for diagnostic purposes, is a long standing, yet common technique in cytogenetics. Automating the chromosome classification process is the first step in designing an automatic karyotyping system. The main aim in this study was to define a new group of features for better representation and classification of chromosomes. Width, position and the average intensity of the two most eye-catching regions of each chromosome (that we call characteristic bands) are the new proposed features. The concept of a characteristic band is based on the expert cytogeneticists' method in classification of the chromosomes. The length, centromeric index (CI) and an index of overall darkness or brightness of the image (NAGD) were also included in the final nine-dimensional feature vectors describing each chromosome. To automatically find the characteristic bands and calculate the new features, different windows in chromosome's density profile were scored based on their intensity and width. As a feasibility study, our work was focused on classification of chromosomes in group E. Three layer artificial neural networks were employed to classify each chromosome in one of the three possible classes (chromosomes 16, 17 and 18). The best results obtained were accurate classification of up to 98.6% of chromosomes. Particularly a six-dimensional subset of the features showed reproducibly high performances in classification experiments. The results of this feasibility study show that new features inspired from human expert's classification method are potentially capable of improving the accuracy of the karyotyping systems.

Keywords: Chromosomes; Classification; Feature extraction; Characteristic regions; Artificial neural networks

1. Introduction

Many genetic disorders or possible abnormalities that may occur in the future generations can be predicted through analyzing the shape and morphological characteristics of the chromosomes. In addition to some well-known genetic abnormalities like aneuploidy (improper number of chromosomes), translocation, and deletion, some of the fatal pathological conditions like lukemia are also correlated with chromosome defects (Hong, 2000). Karyotype, a standard table presenting pictures of the 46 human chromosomes obtained from a single cell either by drawing or by photography using a light microscope (Hong, 2000), is often used to analyze the shape and morphological characteristics of the chromosomes by a specialist for diagnostic purposes.

To develop a karyotype, a cell is photographed under a light microscope during the metaphase stage (one of the four stages of the cell division). Laboratory staining techniques applied to the samples create a unique band pattern for each chromosome. A band is a region along the chromosome axis with a distinct intensity from its adjacent.

* Corresponding author.
E-mail addresses: moradi@cs.queensu.ca (M. Moradi), ksetareh@ut.ac.ir (S.K. Setarehdan).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.06.011

In the next step, each of the chromosomes (22 autosomal pairs and a pair of sex chromosomes) should be identified. This process is usually carried out manually by expert clinicians who view the pictures, identify the chromosomes, and cut and place them in their specified locations in the karyotype.

Despite the development of the banding techniques Karyotyping is still a difficult and time consuming task which must be done by an experienced operator or a cytogenetic expert. The tedious nature of manual karyotyping has encouraged many computer vision and medical image processing researchers to investigate automatic or semi-automatic techniques for Karyotyping in the last three decades (Carothers and Piper, 1994). However, automatic karyotyping is still considered as a difficult task mainly due to the shape variability caused by the non-rigid nature of the chromosomes that gives them unpredictable appearances within the pictures.

Chromosome classification can be viewed as a pattern recognition problem, where the aim is to assign each chromosome to one of the 24 possible classes. The feature vector commonly used to describe a chromosome includes the *length*, the *centromeric index* (the ratio of the short arm of the chromosome to its long arm, which are separated by the narrowest part of the chromosome known as the centromere), and a one-dimensional vector obtained by intensity sampling of the chromosome along its longitudinal axis, which is known as the *density profile* (Carothers and Piper, 1994; Lerner et al., 1995; Sweeney and Becker, 1997; Shin and Pu, 1990). In some studies, a reduced version of the density profile (Lerner et al., 1995) or features extracted from its Fourier or wavelet transformation have been used (Sweeney and Becker, 1997). Using wavelet packet transformation for extraction of features that represent the chromosome shape has recently been reported (Guimaraes et al., 2003). The resulting feature vector is then used with a classification method like the Bayesian classifier (Carothers and Piper, 1994; Qiang and Castleman, 2000), neural network classifier (Cho, 2000; Lerner et al., 1995; Sweeney and Becker, 1997; Lerner, 1998; Graham et al., 1992), or fuzzy classifier (Vanderheydt et al., 1980), nearest neighbor (Groen et al., 1989).

Although the results reported in these studies are encouraging, the karyotyping process in daily laboratory routine still needs the human interaction. A human expert can identify each chromosome in the picture using a hierarchical chromosome identification and classification method. He/she uses some geometric and morphologic features such as the length of the chromosomes for initially classifying them into a small number of groups. Then, applying some simple rules such as the location of the centromere, the location and width of the characteristic bands and their position relative to the centromere and/or relative to each other, the human expert can effectively recognize and identify each chromosome. The concept of characteristic band is very important in this process. Based on the survey conducted in this research, the level of importance of a band is mainly based on the following three factors:

(1) Width of the band.
(2) Intensity of the band.
(3) Relative position of the band.

If a wide and dark band is repeated in the same position for the same chromosome in different images, it is considered as a characteristic band. In this study we have defined a set of features that describe the important characteristic bands for each chromosome. These features include the width, position and the average intensity of the most noticeable characteristic bands of the chromosomes. For a quick reference these features are named (db1W) and (db2W) for the Width of the first and the second dark bands, respectively, (db1P) and (db2P) for the Position of the first and the second dark bands, and (db1I) and

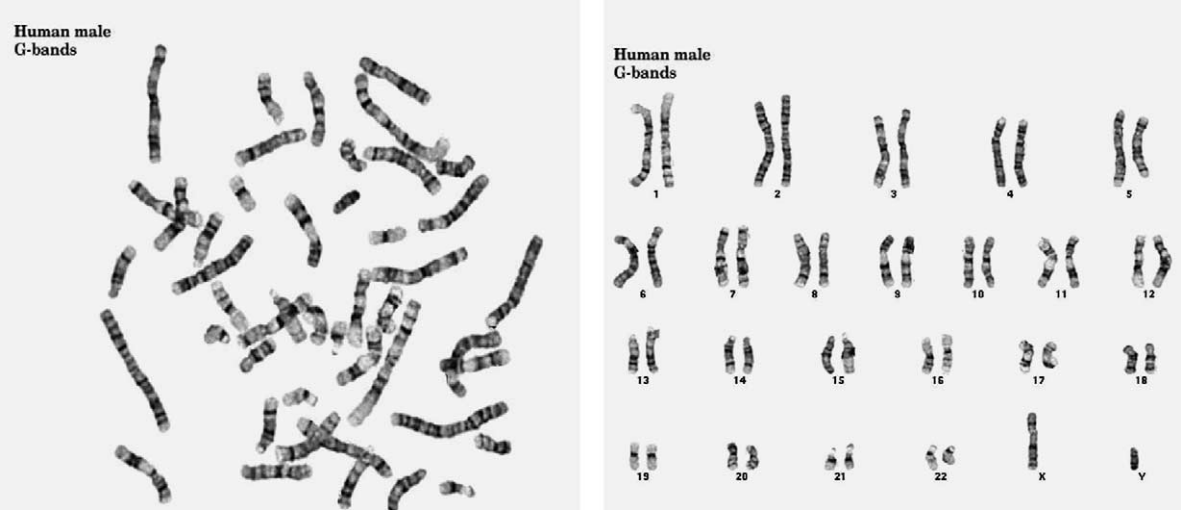

Fig. 1. G-banded chromosomes as seen under a microscope (left) and after karyotyping (right).

(db2I) for the gray level Intensity of the first and the second dark bands (Fig. 1).

As a feasibility study, our work has been limited to the chromosomes in group E. Since the chromosomes in this group have very close lengths, the intensity-based features are more important in their classification process. The choice of number of the characteristic bands (two in our case) was based on the consultation with the experts and analyzing the ideogram of the chromosomes in group E (the standard ideogram of chromosome 16 is shown in Fig. 2). Table 1 summarizes the typical values of the proposed features for chromosomes 16, 17 and 18 calculated by an expert using the standard ideograms (Fig. 2). The intensity-based features (db1I, db2I) are not included in this table, because ideograms do not suggest typical values for them.

This study is aimed to simulate the human expert's knowledge and design a robust chromosome identification and classification algorithm. We have used a medial axis transformation (MAT) based technique to extract the density profile of the chromosomes and used a wavelet based denoising method for identifying the characteristic bands. Multi Layer Perceptron networks are used for classification. The results confirm the efficiency of the new set of features. The rest of the paper is organized as follows. Section 2 describes the dataset used in this study for testing the proposed algorithm. Section 3 illustrates the feature extraction process including automatic extraction of the density profile of the chromosomes and tuning process used to automatically extract the features mimicking human expert knowledge. Section 4 discusses the classification method and presents results of the application of the proposed algorithm to the dataset. Finally Section 5 concludes the paper.

2. The dataset

The images used in this study were produced in the Cytogenetic Laboratory of Cancer Institute, Imam Hospital, Tehran, Iran. The images were acquired by a conventional photography system using a light microscope (Leitz, ortholux) with a magnification factor of 100×. The chromosomes were segmented from the pictures by an expert in the Cytogenetic Laboratory and then scanned by a scanner (Microtek, ScanPlus 6) with a resolution of 300 dpi. The gray scale resolution of the resulting digitized pictures was set to 256 levels. The dataset includes 303 chromosomes from 76 patients (three pairs of chromosomes from 25 patients and three single chromosomes from 51 patients).

3. Chromosomes in the feature domain

Conventionally, a chromosome is described by its *length*, its *centromeric index* (CI) and its *density profile*. In this work, length and CI are used together with the new features developed based on the human expert classification method. The process of feature extraction will be

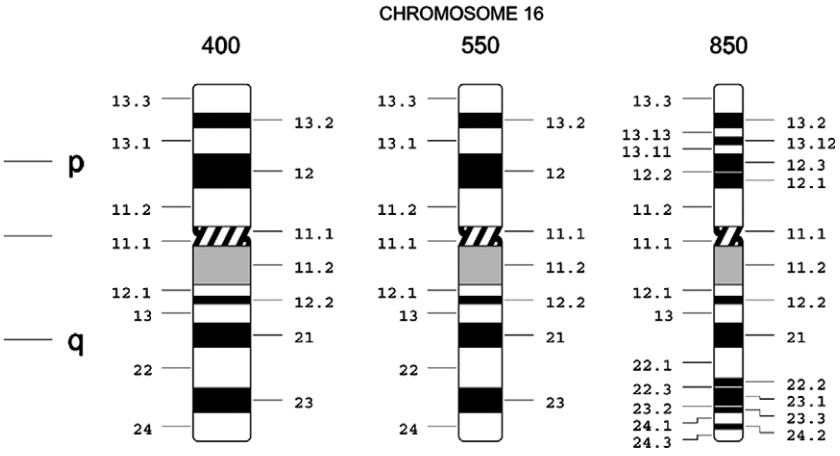

Fig. 2. Ideograms of chromosome 16 in different resolutions.

Table 1
Typical values of features used in this study, these values are extracted from the ideograms (Fig. 2) by an expert

Chr. no.	Relative length (L_R)	CI	First characteristic band	db1W	db1P	Second characteristic band	db2W	db2P
16	0.345	0.75	q21[a]	0.068	0.68	q23	0.82	0.84
17	0.331	0.5	q22–24	0.26	0.75	p12	0.08	0.16
18	0.322	0.312	q12	0.146	0.38	q22	0.137	0.69

[a] In practice, the region beginning from q21 and ending in centromere was considered as the first characteristic band of chromosome 16 (db1W = 0.23, db1P = 0.6).

discussed in this section. Due to the non-homogeneous illuminating conditions in the microscopic images, an intensity normalizing procedure is necessary before the calculation of any feature depending on the intensity of the images. For this purpose, the histogram of the image was modified using histogram stretching technique (Pratt, 1999).

3.1. Medial axis transformation (MAT)

First introduced in 1967 (Blum, 1967), medial axis transformation (MAT) has been frequently used for calculating geometric features and density profile of the chromosomes (Piper and Granum, 1989; Lerner et al., 1995). MAT gives the skeleton of the object which can be used as a one dimensional presentation of a two dimensional shape in pattern classification applications. From the geometrical point of view, the medial axis of a solid object can be formulated as the locus of the center of a maximal disk as it rolls around the interior of the object. The coordinates of the center points along with the radii of the circle in each position give the MAT and locations of the center points mark the medial axis.

In the current application, the medial axis gives a curve passing through the middle of the chromosome along its longitudinal direction (Vernon, 1991) and branches to two parts at the two ends. To obtain the medial axis, a binary version of the image is needed in which the object and the background are separated. The binary version can be obtained by appropriately thresholding the intensity of the image.

In this study, the medial axes of the chromosomes were computed based on the Euclidian distance transform of the binary image. A primary version of the medial axis was extracted from the distance transform based on the algorithm described in (Shin and Pu, 1990) and then a thinning process yielded the medial axis. Fig. 3 illustrates these steps.

The length of the chromosome was defined as the length of the central curve which is coincident to the medial axis in most parts, except for two ends of the chromosome. At these ending parts the medial axis is branched into two parts and the central curve was approximated by the median of the triangle formed by these two branches and the chromosome border. For more details on thresholding process, extraction of the central curve from the medial axis and localizing the centromere see our previous publications which are devoted to the image processing and computational geometry algorithms that we used for feature extraction from the chromosome images (Moradi et al., 2003a,b).

3.2. Density profile (DP)

The density profile (DP) of a chromosome was defined in Section 1. In this work, each sample of the DP signal is defined as the average intensity of the pixels lying on a line perpendicular to the approximation of the central curve of the chromosome. For this purpose, at each pixel belonging to the approximation of the central curve, a perpendicular line was considered and the intensities of the object pixels belonging to this line were averaged to produce one sample of the DP signal (a number in 0–255 range). As it is usually desired to attribute the peaks of the DP signal to darkest regions of the chromosome, each sample of the DP signal (P_i) was replaced by its mirror (255-P_i). The DP signals are illustrated in Fig. 4 for a set of chromosomes in group E.

3.3. Extraction of features describing the characteristic bands

As discussed in Section 1, a new set of features are defined and used for classification of the chromosomes of group E in this work. These features are the position, the width and the intensity of the two most obvious characteristic regions of the chromosomes. The DP signal is used to automatically identify these bands. This process will be discussed in more details in this section.

3.3.1. Preprocessing of the density profile signal

The DP signal is a one dimensional signal representing the intensity variations along the main axis of the chromo-

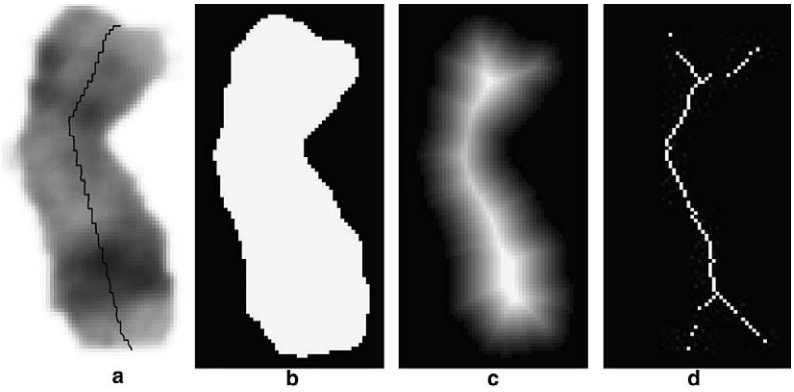

Fig. 3. (a) A typical chromosome 17 on which the central curve is identified, (b) the binary image of the same chromosome, (c) the distance transform of the binary image, and (d) the medial axis of the chromosome.

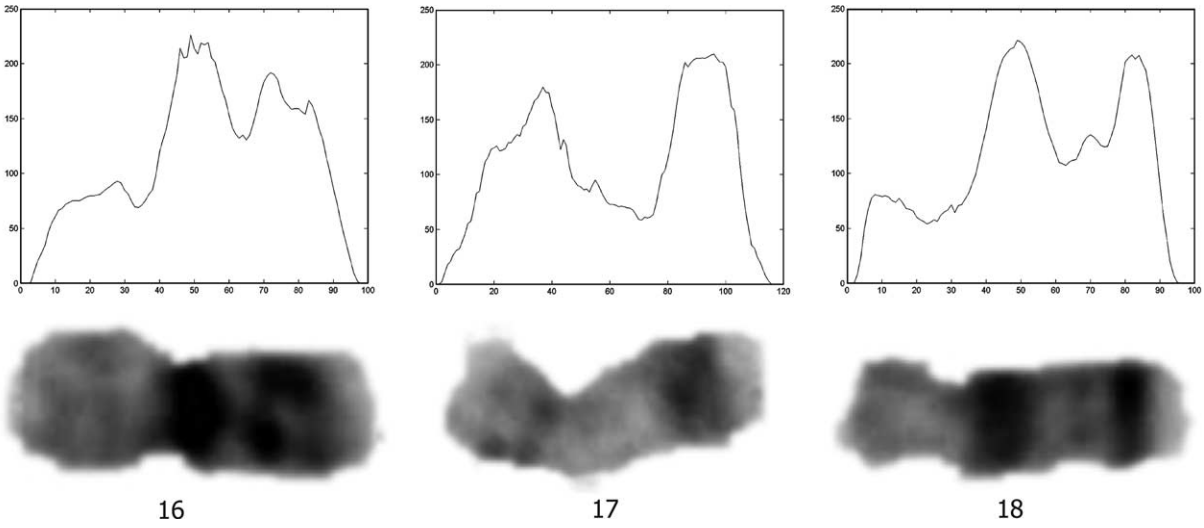

Fig. 4. Typical chromosomes of group E and the density profile signals extracted for them.

some. The characteristic bands are represented by relatively long standing flat peaks in this signal. To determine the beginning and ending points of these bands, an algorithm was developed to score all the peaks with reasonable lengths in the DP signal. Sharp and transient peaks can have a misleading effect on the process, therefore, it is necessary to eliminate the unnecessary details of the DP signal. Two well-studied tools for this purpose are the digital low-pass filters (Oppenheim and Schafer, 1989) and the wavelet denoising method (Aldroubi and Unser, 1996), both of which were applied in this work, but the wavelet denoising method was found to be more effective.

Wavelet denoising was applied to the DP signal based on Donoho's soft thresholding algorithm described in (Aldroubi and Unser, 1996). First a wavelet transformation with the Haar wavelet function was applied to the DP signal and a set of approximation coefficients (a_k) and a set of detail coefficients $\{d_{j,k}, k = 1, \ldots, 2^j, j = L, \ldots, J\}$ were generated (L is the number of decomposition levels). Then a soft thresholding rule was applied to the detail coefficients:

$$d_{ij} = \begin{cases} d_{ij} - T & d_{ij} \geqslant T \\ 0 & |d_{ij}| < T \\ d_{ij} + T & d_{ij} < -T \end{cases} \quad (1)$$

where T is the threshold value. The approximation coefficients were left unchanged to avoid the loss of the main structure of the signal. The DP signal was then reconstructed using this modified set of coefficients producing the denoised signal. Donoho's suggested threshold value, which was also used in this work, is defined as follows:

$$T = \sigma\sqrt{2\log n} \quad (2)$$

where n is the number of samples in the signal and σ is the standard deviation of the noise. Since σ is not known, it is substituted by its approximation:

$$\hat{\sigma} = \mathrm{MAD}/0.6745 \quad (3)$$

where MAD is the absolute median value of the detail coefficients of the first level of the wavelet decomposition computed as (Aldroubi and Unser, 1996):

$$\mathrm{MAD} = \left(\sum_{i=1}^{2^j} |d_{1i} - \mathrm{med}|\right) \Big/ 2^L \quad (4)$$

where med is the median of the detail coefficients of the first level and 2^L is the total number of these coefficients. Fig. 5 shows the results of the application of the wavelet preprocessing method on the DP signal of a typical chromosome 16.

3.3.2. Proposed scoring algorithm for the extraction of the characteristic regions

As discussed in Section 1, from the human expert point of view the width and the intensity of a characteristic region (band) are important parameters for identification of the chromosomes. Thus these are the parameters that should be taken into consideration in order to identify the real characteristic bands among the possible candidate regions. In our algorithm, a set of windows (with various reasonable lengths) were slide over the processed $S = w_1 \times W_R + w_2 \times I_R$ DP signal, and for each one a composite score was calculated as follows:

$$S = w_1 \times W_R + w_2 \times I_R \quad (5)$$

where W_R is the width of the window, I_R is the average intensity of the DP signal within the window and w_1 and w_2 are the tuning parameters ($w_1,\ w_2 < 1,\ w_1 + w_2 = 1$). The window with the highest score was considered as the first characteristic band. A few points need to be clarified in this process. The sizes of the scored windows, the tuning process and the method for determination of the second band will be discussed in the following paragraphs.

3.3.2.1. The window sizes. Considering the standard ideograms of G-banded chromosomes (Fig. 2), one can expect the lengths of the characteristic bands to be close to a

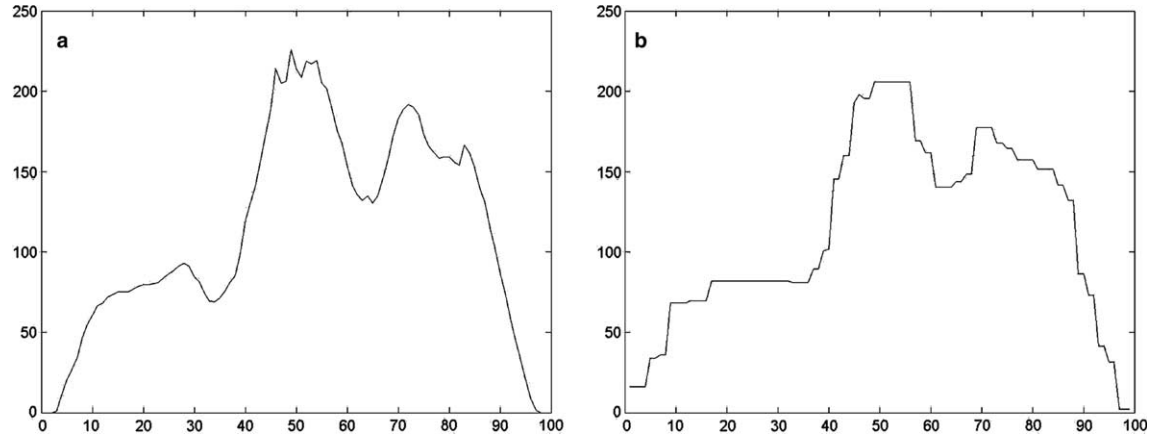

Fig. 5. (a) DP signal of a chromosome 16 and (b) result of the application of the wavelet denoising method.

known portion of the whole length of the chromosome. Therefore, the ratio of the number of the samples in the window of the DP signal that represents a specific band to the number of all samples in the signal is also expected to be in a known range. Based on this fact, the window size was determined using the typical values of the characteristic regions (Table 1). As the shortest ideal characteristic dark band in group E covers about 8% of the whole chromosome (db2 of chromosome 17) and the widest one covers about 26% of the chromosome (db1 of chromosome 17), the widths of the candidate scored windows were limited between 5% and 35% of the length of the DP signal. All the definable windows with the number of samples in the range of 5% to 35% of the whole signal were scored based on Eq. (5).

3.3.2.2. Determination of the tuning parameters w_1, w_2. The tuning parameters were selected in an iterative process based on closeness of the mean of the band feature values extracted using a specific set of $\{w_1, w_2\}$ to the values of reference features presented in Table 1. The weigh parameters where changed with the steps of length 0.1. The scoring algorithm was repeated for sets of $\{w_1, w_2\}$ and the results were sorted based on the mean square error of the average of computed feature values over the dataset (the error was defined as the difference of the computed values and the reference values in Table 1). Because there is no reference value available for intensity related feature of the characteristic bands—dbI—it was not considered in the process. The set of weighs $\{w_1 = 0.3,\ w_2 = 0.7\}$ resulted in feature values with closest average value over the dataset to the reference values. The process of tuning the parameters was performed using only the features describing the first characteristic band.

3.3.2.3. Determination of the window in the DP signal representing the second characteristic band. As explained, the scoring was first applied to the whole signal and the first region was selected as the region with the highest score. As the first and the second regions have to be non-overlapping, the second region was found by applying the algorithm separately on the two remaining parts of the signal at the left and the right side of the previously defined first band. The second characteristic band was determined as one of the two resulting windows with higher score. The same process for determining the tuning parameters was followed only this time using the features describing the second characteristic band.

In summary, in order to determine the beginning and ending points of the windows in the DP signal that represent the characteristic bands of the chromosomes, a set of windows with reasonable number of samples were scored based on their average intensity and number of samples. The windows having high scores correspond to flat peaks in the signal that most likely represent the characteristic bands of the chromosome. Fig. 6 shows the windows corresponding to the first and the second characteristic bands on DP signals of a set of group E chromosomes.

3.3.3. Computing the new features

After extracting the positions of the characteristic bands, enough data for computation of the band describing features is available. Assume that a characteristic dark band is extended from the ith to the jth sample of the DP signal. The relative position of this region is given by

$$\mathrm{dbP} = \frac{(i+j)/2}{N_{\mathrm{P}}} \tag{6}$$

where N_{P} is the total number of the samples in the DP signal. According to this equation, regions closer to the end (bottom) of the chromosome will have a greater dbP. The relative width of the characteristic dark region is given by

$$\mathrm{dbW} = \frac{(j-i)}{N_{\mathrm{P}}} \tag{7}$$

Finally, the intensity of the characteristic region is defined as the mean magnitude of the samples of the DP signal in that region:

$$\mathrm{dbI} = \frac{\sum_{n=i}^{j} P_n}{j-i} \tag{8}$$

where P_n is the nth sample of the DP signal.

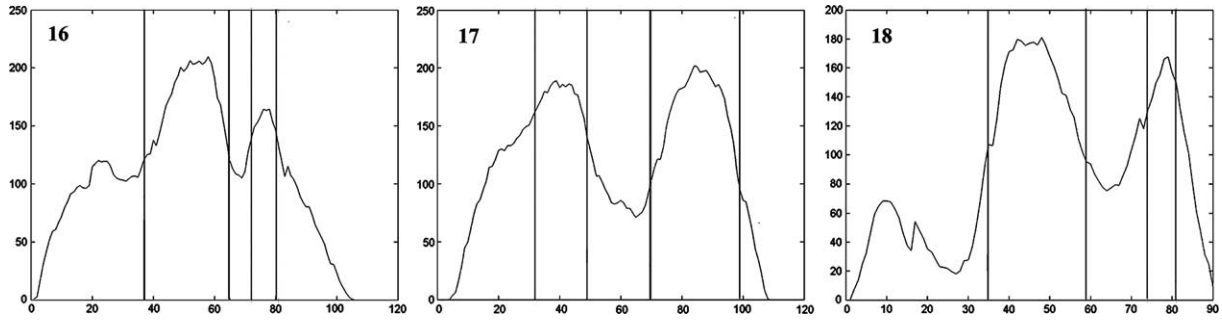

Fig. 6. Characteristic regions in the DP signal determined by grading algorithm for a set of group E chromosomes.

3.4. Normalized average gray density (NAGD)

The last feature in our nine-dimensional feature vector was the relative NAGD which is computed as follows:

$$\mathrm{NAGD} = \sum_{i=1}^{n_j} I_i / n_j \tag{9}$$

$$\mathrm{NAGD}_{\mathrm{R}_j} = \mathrm{NAGD}_j \Big/ \sum_{i=16}^{i=18} \mathrm{NAGD}_i \tag{10}$$

where n_j is the number of the image pixels belonging to the chromosome object and I_i is the intensity level of ith pixel of the image. The relative NAGD (Eq. (10)) is an index of overall darkness of the chromosome. Among chromosomes in group E, chromosome 18 is usually darker than the other two.

4. Classification method and results

Utilization of artificial neural networks (ANN's) for classification of chromosomes has been intensively studied in the past. Lerner (1998) has suggested that ANN's are the best chromosome classifiers, especially when the number of classes is small. When the number of classes increases, the efficiency of Bayes piecewise linear classifier approaches to the ANN based classifier. In the present study, the number of classes was limited to three. Therefore, ANN was employed for classification. Three layer feed-forward perceptron neural networks with different number of neurons in the hidden layer were trained by the backpropagation learning rule and used for the classification of the chromosomes.

One problem that can occur during neural network training is over-fitting which reduces the generalization capability of the network. We used an early stopping strategy to validate the learning process. In this technique, the available data is divided into three subsets: training, validation and testing sets. The training subset is used for updating the ANN parameters; the testing subset is used for final assessment, and the classification error on the validation set is monitored during the training process to avoid over-fitting. The validation error will normally decrease during the initial phase of training similar to the training set error. However, when the ANN begins to over-fit the training data, the error on the validation set typically begins to rise. When the validation error increases for a specified number of iterations, the training is stopped, and the weights that produced the minimum error on the validation set are retrieved. The results of our classification experiments are presented in this section.

4.1. Classification using the nine-dimensional feature vectors

In the first round of the experiments, the nine-dimensional feature vector (Length, CI, NAGD, db1P, db1W, db1I, db2P, db2W, db2I) was used. Three layer perceptrons with 1, 2, 3, 5, 7, 10 and 13 neurons in the hidden layer were used as classifiers. It was found that the networks with 13 and 10 neurons in the hidden layer had no clear advantage over the networks with five and seven neurons. Therefore, architectures with more than 13 neurons in the hidden layer were not tested. On the other hand, it was noticed that the network with one neuron in the hidden layer did not have the capability of proper classification of the patterns.

The process of training and testing with early stopping strategy (with maximum possible epochs of 150) was conducted 20 times for each network. For each round of experiments, 102 chromosomes were randomly selected to be included in the training set and the remaining 201 chromosomes constituted the training set. To avoid over-fitting, a subset of the test set, including 30 feature vectors was used as the validation set. The results are presented in Table 2, including the average success of the network on the training and test sets. The standard deviation, the best and the worst performance of the architectures are also reported in the table. It can be inferred that the results are acceptable in terms of average and best performance. However, all architectures have shown some poor performances as well. These results show that the best networks acquired can obviously classify the test dataset with high accuracy. However, they might not perform well enough as a general classifier. Feature selection might improve the results; this theory was tested and confirmed in the second round of experiments that will be described later in this section.

Fig. 7 demonstrates the progress diagram of the networks with five and seven neurons in the hidden layer.

Table 2
The results of classification of chromosomes with nine-dimensional feature vectors

Number of neurons in the hidden layer	The average of correct classification results on training set (%)	The average of correct classification results on test set (%)	The standard deviation of the classification results over 20 experiments (%)	Worst classification result on test set (%)	The best classification results on test set (%)
3	97.6	90.4	**2.6**	80.1	**97.1**
5	99.3	88.3	3.19	75.1	92.9
7	98	88	3.05	80.1	94.2
10	99	90.4	3.29	77.4	95.6

Each network was trained and tested 20 times; the average, the best, the worst and the standard deviation of the results are presented.

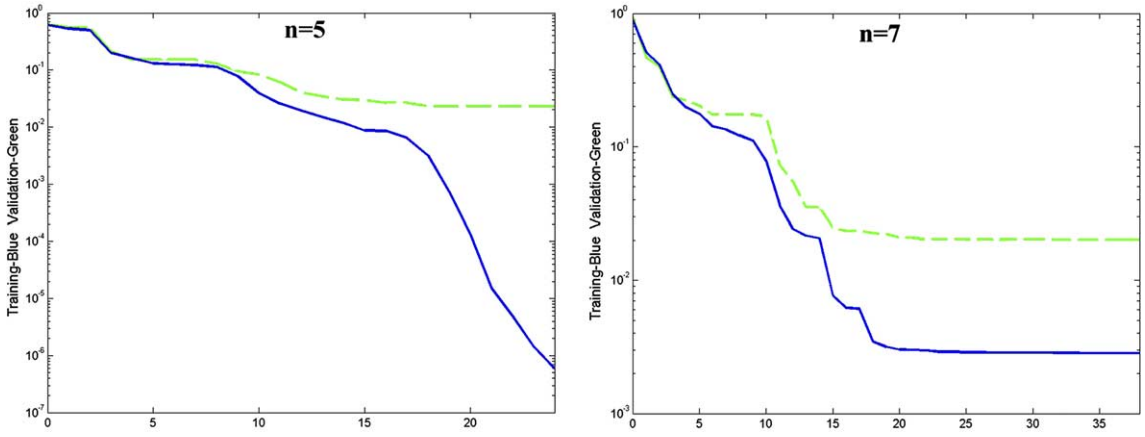

Fig. 7. Progress diagram of the training and validation process of the neural networks with five and seven hidden layer neurons in the first round of the experiments: the dashed diagram represents the error on the validation data and the continuous line represents the error on the training set. The validation criterion has caused early stopping of the training process because the error on the validation set has become constant.

The dashed line represents the error on the validation set while the continuous line shows the error on the training set. In both cases, the early stopping criterion has been satisfied before the maximum number of the epochs (150) is reached.

4.2. Exclusion of NAGD, db1I and db2I from the feature vectors

It is always desirable to use smaller feature vectors in order to decrease the complexity of the classifiers. Different feature selection procedures have been applied to the traditional chromosome feature vectors in the past (Cho, 2000; Lerner et al., 1995). In this study, we followed a knowledge based approach: intensity-based features (i.e., NAGD, db1I and db2I), are subject to the imaging and illumination conditions and also the standard ideograms do not suggest typical values for them. Therefore, given our feature extraction procedure, these features are prone to the maximum error. The effect of exclusion of these features on the accuracy of the classifications was examined in the second round of experiments.

The same settings in terms of random splitting of the dataset before experiments, and maximum possible number of epochs (150) were used. The reported results are acquired using 20 rounds of training–testing experiments on each one of the architectures. The results indicate that the reduced feature vector leads to a more robust classification performance (Table 3). Although the maximum performances are in the same range (98.6 compared to 97.1 in the first round), the standard deviations have decreased and the difference between the best and the worst performances of the networks are meaningfully less than the first round. A combination of network structure and feature vector that has constantly shown high performances in 20 rounds of experiments (with different combinations of the vectors in training–testing subsets) can be confidently considered as an efficient classifier method. As the table indicates, architectures with five and seven neurons in the hidden layer have shown the highest average and highest best performances, respectively.

The analysis of the misclassified cases (confusion matrix analysis) showed some interesting results. In the network architecture with five neurons in the hidden layer (which showed the highest average performance), the average inaccuracy is about 5%. From this amount, about 1.35% was caused by misclassification of chromosome 16 as 18. The rest of meaningful misclassifications were rare and constitute less than 1% of the average error. Most of the error cases were caused by the undefined outputs (many of these

Table 3
The results of classification of chromosomes with six-dimensional feature vectors (db1W, db2W, db1P, db2P, CI, L)

Number of neurons in the hidden layer	The average of correct classification results on training set (%)	The average of correct classification results on test set (%)	The standard deviation of the classification results over 20 experiments (%)	Worst classification result on test set (%)	The best classification results on test set (%)
3	100	94.1	1.41	91.3	97.1
5	100	94.9	1.28	91.3	97.1
7	100	94.1	1.73	88	**98.55**
10	100	94.88	1.23	91.3	97.1

Each network was trained and tested 20 times; the average, the best, the worst and the standard deviation of the results are presented.

cases can be avoided: if two or three neurons in the output layer are activated instead of one, the winner is the neuron with the highest output value). A similar pattern of confusion matrix elements was also observed in the other examined network structures with a meaningfully higher value of error caused by mistaking chromosome 16 as 18. The similarity of the positions of the characteristic bands of chromosomes 16 and 18 might partly justify this above average error rate.

5. Discussions and conclusions

Automatic human chromosome classification is one of the most widely investigated stages of the karyotyping process (Lerner, 1998). Over the past few years, several classification methods have been developed and tested for this purpose. Most of these classifiers have two main flaws (Groen et al., 1989; Piper and Granum, 1989): poor performance compared to the human expert (70–80% compared to 99.7%) and the requirement for an operator interaction to correct the misclassifications. The main source of these shortcomings might lie in using low level or inappropriate features compared to the powerful feature synthesis mechanism of the human expert's brain (Lerner, 1998). The goal of this study was to emulate this process. A new representation of human expert's knowledge about the appearance of the chromosomes in class E was developed for this purpose. This representation is in form of a new set of features that include position, width and intensity of the two most important characteristic regions (bands) in the chromosomes.

The results obtained show that the new set of features used with an MLP including only 3–10 neurons in the hidden layer leads to a success rate of about 97.1%. A simple feature selection process increased the maximum classification rate to 98.6 and resulted in much more reproducible results. The neural networks trained with our six-dimensional feature vectors constantly resulted in very high classification rates (each ANN structure was trained and tested 20 times each time with a different combination of the training and testing data). Compared to Lerner et al. (1995), who have applied the same classification approach as in the present study to solve a five class chromosome classification problem, our results are slightly more accurate and reproducible (Lerner et al. have used a feature vector consisting of CI, length and samples of the density profile signal and achieved the maximum classification rate of 98% on an MLP). These results show the potential capability of the new features for classification of different chromosomes. More attention should be paid to the feature extraction process to ensure that the maximum separability of the classes by the new set of features is obtained. Particularly, we found out that the accuracy of the values extracted for intensity related features is limited both by the illumination conditions and our feature extraction process. This is probably the reason for the increase in the accuracy of the classifications after exclusion of these features.

Similar to many other investigators, we have considered a problem with three classes for the feasibility study phase. A prerequisite for the generalization of this method to the 24-class problem is defining characteristic regions for the other groups of chromosomes. Number of these regions may not necessarily be "two" in all cases. Also, for a 24-class problem, feature vectors of higher dimensions might be needed. In that situation, the intensity-based features would have an important role and the feature extraction process should be modified to ensure the validity of their values. We suggest a hierarchal classification process for the 24-class problem: in the first step chromosomes can be divided into some major subsets. Geometric and morphologic features can be used for this step. Then a different set of characteristic regions can be defined for chromosomes in each subset and the features describing these regions along with length and CI can be used for final classification of the chromosomes.

Acknowledgements

The authors would like to thank Dr. S.R. Ghaffari and Ms. F. Farzanfar from the Cytogenetic Laboratory of the Cancer Institute of Imam Hospital in Tehran, Iran for their help in providing images and useful comments.

References

Aldroubi, A., Unser, M., 1996. Wavelets in Medicine and Biology. CRC Press, FL, USA.

Blum, H., 1967. A transformation for extracting new descriptors of the shape. In: Proceedings of the Symposium on Models for Perception of Speech and Visual Form, pp. 362–380.

Carothers, A., Piper, J., 1994. Computer-aided classification of human chromosomes: A review. Statistics and Computing 4, 161–171.

Cho, J.M., 2000. Chromosome classification using backpropagation neural networks. IEEE Engineering in Medicine and Biology 19, 28–33.

Graham, J., Errington, P., Jennings, A., 1992. A neural network chromosome classifier. Journal of Radiation Research 33, 250–257.

Groen, F.C.A., 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.

Guimaraes, L.V., Schuck, A., Elbern, A., 2003. Chromosome classification for karyotype composing applying shape representation on wavelet packet transform. In: Proceedings of the 25th Annual International Conference of the IEEE EMBS, 1, pp. 941–943.

Hong, L.M., 2000. Medical Cytogenetics, first ed. Marcel Dekker, NY, USA.

Lerner, B., 1998. Towards a completely automatic neural-network-based human chromosome analysis. IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics 28, 544–552.

Lerner, B., Guterman, H., Dinstein, I., Romem, Y., 1995. Medial axis transform-based features and a neural network for human chromosome classification. Pattern Recognition 28, 1673–1683.

Moradi, M., Setarehdan, S.K., Ghaffari, S.R., 2003a. Automatic locating the centromere on human chromosome pictures. In: Proceedings of the 16th IEEE Symposium on Computer-based Medical Systems, pp. 56–61.

Moradi, M., Setarehdan, S.K., Ghaffari, S.R., 2003b. Automatic landmark detection on chromosomes' images for feature extraction purposes. In: Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis (ISPA 2003), pp. 567–570.

Oppenheim, A.V., Schafer, R.W., 1989. Discrete-Time Signal Processing. Prentice Hall, NJ, USA.

Piper, J., Granum, E., 1989. On fully automatic feature measurement for banded chromosome classification. Cytometry 10, 242–255.

Pratt, W.K., 1999. Digital Image Processing, second ed. John Wiley and Sons, NY, USA.

Qiang, W., Castleman, K.R., 2000. Automated chromosome classifier using wavelet-based descriptors. In: Proceedings of the IEEE Symposium on Computer Based Medical Systems (CBMS2000), pp. 189–194.

Shin, F.Y., Pu, C.C., 1990. Medial axis transformation with single-pixel and connectivity preservation using euclidean distance computation. In: Proceedings of the 10th International Conference on Pattern Recognition, pp. 723–725.

Sweeney, N., Becker, R.L., 1997. A comparison of wavelet and Fourier description for a neural network chromosome classifier. In: Proceedings of IEEE International Conference on Engineering in Medicine and Biology, pp. 1359–1362.

Vanderheydt, L., Oosterlinck, A., Van Daele, J., Van Den Berghe, H., 1980. Design of graph-representation and a fuzzy-classifier for human chromosomes. Pattern Recognition 12, 201–210.

Vernon, D., 1991. Machine Vision, first ed. Prentice-Hall, UK.