

# MULTI-FEATURE ANALYSIS AND CLASSIFICATION OF HUMAN CHROMOSOME IMAGES USING CENTROMERE SEGMENTATION ALGORITHMS

*P. Mousavi, R.K. Ward, P.M. Lansdorp<sup>†</sup> and S.S. Fels*

Department of Electrical and Computer Engineering,  
University of British Columbia, Vancouver, B.C.

<sup>†</sup>Terry Fox Laboratory, B.C. Cancer Research Center, Vancouver, B.C.

## ABSTRACT

Classification of homologous human chromosomes is essential to advanced studies of cancer genetics. This paper describes novel segmentation and classification algorithms to extract multiple features, from microscopy images of chromosomes, for classification purposes. Multicolour images of metaphase chromosomes prepared applying PNA probes are used for this purpose. Centromeres are segmented using an iterative fuzzy algorithm as well as a gradient method. Moreover, telomere length measurements are performed on chromosome images and normalized for the image database. Multiple intensity features are calculated as a result of the developed algorithms. Heteromorphic chromosomes (such as 16 and 22) are then successfully classified into their parental homologues, based on the calculated multiple features, and used to verify the developed methods.

## 1. INTRODUCTION

Image processing techniques have always played a major role in advancement of cancer research (using microscopy images of chromosomes). This role has been specifically emphasized after the introduction of novel Peptide Nucleic Acid (PNA) probes which provide chromosome images with high quantitative information [2]. Since then, several studies have focused on analyzing homologous chromosomes in microscopy images prepared using Fluorescence In Situ Hybridization (FISH) technology. [3, 4, 5].

The nucleus of cells in the human body are made up of 23 pairs of chromosomes [1]. Each chromosome in one pair is inherited from one parent. Scientists believe cancer is related to specific chromosome abnormalities. To study the characteristics of cancer, it is essential to have the ability to analyze separate homologues [3]. At present, there is no biological method to classify homologous chromosomes except in rare circumstances with interference of an expert technician. In this paper we give an overview of three methods

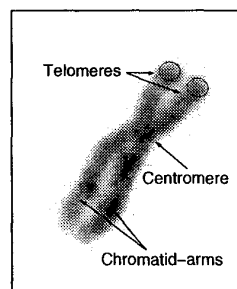


Fig. 1. Structure of a Human Chromosome

developed to extract multiple features from images of chromosomes. The first method is based on fuzzy sets theory and is inspired by gradient method and back propagation. It successfully segments centromeres from FITC images of chromosomes. The second proposed method segments centromeres from DAPI images and is based on thresholding the 2D gradient of the chromosome image. In the third method, telomere lengths are measured, normalized in the database and clustered into two groups using fuzzy c-means algorithms. Once centromere segmentation and telomere length measurements are performed, several intensity based features are calculated as a result. Homologue classification is then performed for polymorphic chromosomes 16 and 22 based on the extracted features. The layout of the paper is as follows. In section 2, we discuss the database preparation and image acquisition procedure. Section 3 describes three developed methods for multiple feature extraction from images of chromosomes. In section 4, the results of homologue classification for heteromorphic chromosomes are presented and discussed. The conclusion of the paper is presented in section 5.

## 2. DATABASE PREPARATION

Images of chromosomes studied in this research are generated using Fluorescence In Situ Hybridization (FISH) technology. FISH is based on fluorescence probes bind-



**Fig. 2.** A metaphase image using DAPI (left), FITC-centromere (middle) and CY3 (right) probe

ing to certain repeats on chromosomes and examining these chromosomes under a fluorescence microscope. A slide of metaphase chromosomes is prepared using multiple probes. Twelve images of the prepared slide are then acquired by a fluorescence microscope and preprocessed prior to being used. The probes used for preparing our database are DAPI, CY3 and FITC-centromere probes. DAPI highlights the whole chromosome, CY3 binds to telomeres on chromosomes and FITC makes only centromeres of the chromosomes visible (Figure 2). The database is created by cutting every homologous chromosome pair (from the DAPI image) as well as its corresponding centromeres and telomeres (from FITC and CY3 images) as separate images.

### 3. MULTIPLE FEATURE EXTRACTION FROM CHROMOSOME IMAGES

Since there is no biological method to verify homologue classification, multiple features of chromosome images are extracted and used for classification purposes. The correlation of classification results employing different features is used to validate our developed methods. The extracted features described in this paper are:

- Integrated Fluorescence Intensities (IFI) of centromeres segmented from FITC images,
- IFI of centromeres segmented from DAPI images,
- telomere data classification from CY3 images,
- morphological differences in heteromorphic chromosome 22.

In the following subsections we will give an overview of the algorithms developed for extracting the above features.

#### 3.1. Centromere Segmentation - FITC images

Although centromeres in FITC images have higher intensities than their background (Figure 2-middle), segmentation is not a straight forward task. This is due to

the fact that boundaries of centromeres, in FITC images, have extremely gradual transitions. We propose a segmentation method based on fuzzy sets approach [6]. As the centromeres have unclear boundaries, this approach makes the process of assigning pixels to centromere/background more accurate.

In this method, a fuzzy membership function is applied to the centromere image assigning a membership value,  $O(i, j)$ , to each pixel of the image,  $(i, j)$  [6]. Generally, there is a small spectral overlap of DAPI images in the FITC image. In order to classify a pixel as part of the centromere, we need to examine the intensity of the surrounding pixels as well as the gray level value of that particular pixel. Therefore a 3x3 neighborhood mask with different weight coefficients is defined for each pixel.

Inspired by Gradient method and Back-Propagation algorithm, an iterative process is developed which determines the membership of each pixel to a region. The error function and update rule are defined as:

$$E = \sum_{i,j} O(i, j) (1 - O(i, j)) \quad (1)$$

$$O^+(i, j) = O(i, j) \quad (2)$$

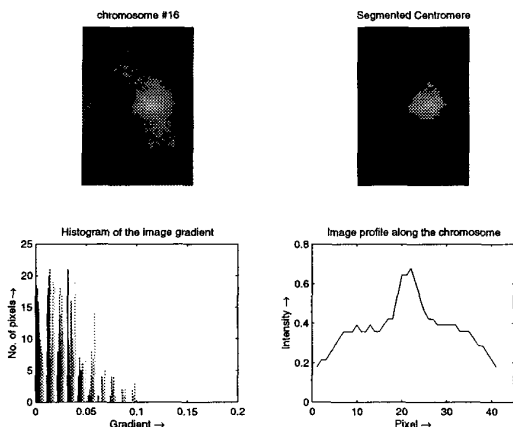
$$+ 2\eta \left[ \sum_{k,l \in N^0} W(k, l) (O(k, l) - 0.5) \right] \cdot O(i, j) (1 - O(i, j))$$

where  $\eta$  is the learning rate and  $W(i, j)$  are the weight coefficients of the neighborhood mask,  $N^0$ , around  $O(i, j)$ . Using this method centromeres of chromosomes are successfully segmented from the FITC images. Integrated Fluorescence Intensities (IFI) are computed over the centromere area for the database and will be later used for homologue classification. An example of centromere segmentation for chromosome 22 is shown in Figure 4.

#### 3.2. Centromere segmentation - DAPI images

Certain chromosomes display heteromorphic properties after staining with DAPI probe. In chromosome 16, for example, the centromere intensity in one homologue is brighter than that of the other homologue. This property can be used as a feature to classify homologues of chromosome 16. Therefore, we developed an algorithm to segment centromeres from DAPI images of chromosomes and measured IFI values over the segmented areas.

The algorithm is based on a gradient method where, the two dimensional gradient of the chromosome image is calculated and the gray level mode histogram of this gradient image is formed (Figure 3). The valleys towards the right end of the histogram are the maximum



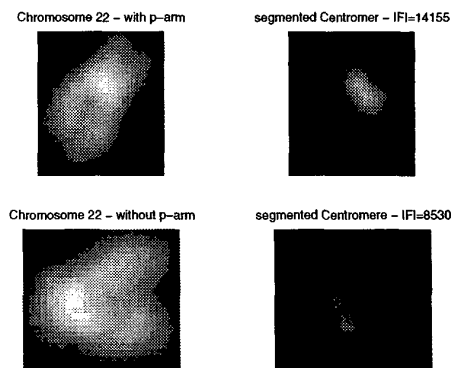
**Fig. 3.** Centromere Segmentation, from DAPI image, for chromosome 16

difference regions. Since the centromere pixels of chromosomes are brighter and have higher gradients than the rest of the chromosome image, the centromere is segmented by choosing 1) pixels that lie in the maximum difference region of the image gradient histogram and 2) pixels that have an intensity value more than a certain threshold. Applying this method, centromeres of chromosomes 16 are segmented from DAPI images and IFI values of the centromere areas are calculated for all these chromosomes. In Figure 3, we show the image profile, histogram of the 2D gradient image and the result of centromere segmentation for an example of chromosome 16.

### 3.3. Telomere data classification - CY3 images

Telomeres form the end parts of the chromosome and are believed to have a role in cancer. Telomere lengths are important features in homologue classification. A software has been developed at the Terry Fox laboratory to measure telomere lengths in each chromosome image. Applying this software to our database a table is generated with four telomere values for each chromosome image, namely P1, P2, Q1 and Q2 (a telomere length for each arm). The average values of P1 and P2 (P) and Q1 and Q2 (Q) are calculated and normalized to a zero mean and unity standard deviation. Normalized P and Q values are plotted against each other in order to study the clustering characteristics of their distribution.

Fuzzy c-means algorithm is used to classify the telomere data into two clusters. This algorithm starts with an initial guess for the cluster centers. By iteratively updating the cluster centers and the membership values for each data point, the algorithm moves the clus-



**Fig. 4.** Centromere Segmentation and Classification of Chromosome 22

ter centers to the correct location. This iteration is based on minimizing the distance function from any data point to a cluster center. Homologues of chromosome 16 are classified into two classes using this algorithm (Figure 5). In figure 5, the solid line shows the classification of the relative-telomere-length points into two separate homologue classes.

### 3.4. Morphological differences in heteromorphic chromosome 22

As mentioned before, in almost 50% of the population, heteromorphic chromosomes (such as 16 and 22) show obvious differences in their appearances once stained with DAPI probe. In chromosome 22, one homologue has short p-arms while the other homologue has no p-arms. We used this differentiating feature and classified homologues of chromosome 22 into two initial parental classes. In the next section, we will combine this morphological feature with IFI values computed from previous sections to re-classify homologues of chromosome 22 and verify our developed methods.

## 4. RESULTS AND DISCUSSION

In the previous sections, we discussed three algorithms to extract features from images of chromosomes. We now use chromosome 22 to verify the feature extraction results from sections 3.1 and 3.4, and chromosome 16 to verify results from sections 3.2 and 3.3.

Once the centromere regions in FITC images are segmented for all chromosomes 22, the total integrated fluorescence intensities are calculated over each region. Homologues of chromosome 22 are classified into two groups using the differences in their IFI values. On the other hand, the two homologues of chromosome 22

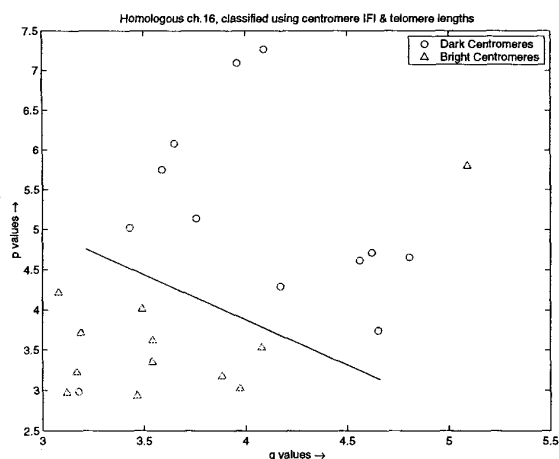


Fig. 5. Classification of Chromosome 16

show apparent differences in their p-arms on DAPI images. This feature is also used to classify chromosomes 22 into two homologous groups. Studying the classification results using these two features (Figure 4), it is concluded that chromosomes 22 with a p-arm have higher centromere intensities than chromosomes 22 without a p-arm. The classification results (using both features) are identical and chromosome 22 is successfully classified into its parental homologue classes.

Chromosome 16 is heteromorphic, i.e. its two homologues have different centromere intensities in the DAPI image. After all centromeres of chromosomes 16 are segmented from DAPI images, IFI values are calculated over the segmented areas and chromosomes 16 are classified into parental homologue classes accordingly. On the other hand, in section 3.3, telomere values of these chromosomes are calculated, normalized and plotted. Fuzzy c-means algorithm is used to classify chromosomes 16 in two classes of parental homologues, using the telomere data. Comparing the results of the two classification methods (Figure 5), it is concluded that chromosome 16 homologues with brighter centromeres tend to have smaller ratios of  $\frac{P}{(c-Q)}$  (where P and Q are normalized p-arm and q-arm telomere lengths and c is a constant). In addition, homologue classification using telomere and centromere information are almost identical and chromosomes 16 are correctly classified into two homologous groups.

## 5. CONCLUSION AND FUTURE WORK

Cancer is a somatic genetic disease and studies suggest predispositions to specific cancers are inherited as abnormal genes. Classification of homologous chro-

mosomes is essential to studies of cancer genetics. In this paper, we gave an overview of three developed algorithms which successfully extract intensity features from images of chromosomes. Combining these features with morphological and intensity features of heteromorphic chromosomes, we both classified chromosomes 16 and 22 into their parental homologues and verified our developed algorithms. The results of homologue classification in this paper will help track the responsible genes for cancer in generations.

Every time slide preparation procedures change, practical aspects of image acquisition introduce new challenges to our image processing algorithms. Future work should be directed towards improving the algorithms to make them robust and accommodating enough to be able to compensate for these variations.

## Acknowledgments

The authors would like to thank E. Chavez for her valuable help in acquiring and karyotyping the chromosome images. They would also wish to thank the National Institute of Health for providing funding for this research.

## 6. REFERENCES

- [1] A.J. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, W.M. Gelbart, *An introduction to genetic analysis*, Sixth edition, Freeman, 1996.
- [2] P.M. Lansdorp, V. Dragowska, N. Rufer, T. Brummendorf, S.S.S. Poon, P. Mousavi, T. Duncan, U. Martens, "Applications of Peptide Nucleic Acid Probes in Cytometry," *Proc. of the ISAC XIX International Congress*, 1998.
- [3] P. Mousavi, R. Ward, P.M. Lansdorp, "Feature analysis and classification of chromosome 16 homologs using fluorescence microscopy image," *IEEE Can. Journal of Elec. & Comp. Eng.*, Vol.23, No. 4, 1999.
- [4] S.S.S. Poon, *Telomere length measurements using fluorescence microscopy*, Ph.D. thesis, Dept. of Elec. & Comp. Eng., University of British Columbia, 1997.
- [5] P.M. Lansdorp, N. Werwoerd, F. Van de Rijke, et. al., "Heterogeneity in telomere length of human chromosomes," *Human Molecular Genetics*, Vol.5, No.5, 1996.
- [6] M. Sameti, R. Ward, "A fuzzy segmentation algorithm for mammogram partitioning," *Digital Mammography'96*, Elsevier Science B.V., Newyork, 1996.