

# CLASSIFICATION OF HOMOLOGOUS HUMAN CHROMOSOMES USING MUTUAL INFORMATION MAXIMIZATION

*P. Mousavi, S.S. Fels, R.K. Ward, and P.M. Lansdorff*

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

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

## ABSTRACT

Multi-feature analysis of human chromosome images is a major step towards classification of homologous chromosomes. In this paper, for the first time, an automatic quantitative classification method is proposed for homolog differentiation using multiple features. This method is based on mutual information maximization applied to an unsupervised neural network architecture. The neural network consists of separate modules which are trained to classify homologs using independent features. Mutual information is then maximized between the outputs of the modules forcing them to produce the same classification results, for a given chromosome. The proposed method was successfully applied to classify homologs of chromosome 16 with 100% accuracy.

## 1. INTRODUCTION

The nucleus of cells in the human body contain 23 pairs of homologous chromosomes. In each pair, one chromosome is inherited from the mother known as the "maternal homolog" and the other chromosome is inherited from the father known as the "paternal homolog". Cancer is believed to be related to specific chromosome abnormalities. In order to study the characteristics of cancerous cells, it is essential to classify chromosomes into paternal and maternal homologous classes so as to have the ability to analyze separate homologs [1].

Figure 1 shows the structure of a metaphase chromosome consisting of two duplicate sister chromatids (with four arms known as P and Q arms). The central part of the chromosome is known as the "centromere". "Telomeres" form the ends of the P and Q arms. Telomere lengths are believed to have an important role in cell life-span and the development of cancerous cells.

The introduction of Fluorescence In Situ Hybridization (FISH) technology, high quality microscopes and novel Peptide Nucleic Acid (PNA) probes have started a new era in quantitative measurements of chromosome images for homolog classification purposes [2]. Multi-feature analysis of chromosome images seems to be the reasonable approach to homolog classification. This is due to the fact that there exists no robust mean by which the results of homolog classification can be verified. Thus, high correlation in the homolog classification results using several different features can be used as a verification tool.

To date, homolog classification algorithms have only used one feature at a time to differentiate homologs. In addition, attempts to classify homologs using multiple features employ ad hoc manual strategies in contrast to automatic algorithms for this purpose. In this paper, we propose a novel quantitative and automatic classification method for homologous chromosomes using multiple fea-

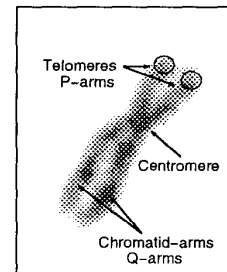


Fig. 1. Structure of a Human Chromosome

tures. This work was inspired by two previous independent classifiers developed by the authors in [3] and [4]. The classification method is based on mutual information maximization criterion applied to an unsupervised feed forward neural network architecture. We designed and tested our algorithm on chromosome 16 since for this chromosome, a cyto-technician can visually classify homologs providing us with ground truth. The neural network architecture is composed of two separate modules. Each module is designed to classify homologs of chromosome 16 based on independent features. The learning algorithm maximizes the mutual information between the outputs of the two modules training them to produce the same classification for different features. This neural network architecture is able to extract higher level features from the input data which cannot be detected otherwise. In addition, the architecture can be generalized to classify homologous chromosomes using as many features as available. This research opens a new venue for the application of unsupervised neural networks and mutual information maximization theory to homolog classification of human chromosomes.

The layout of the paper is as follows. In Section 2, we give an overview of the database preparation and basic feature extraction procedures. Section 3 discusses the proposed neural network architecture and the learning algorithm. In Section 4 we present the designed training experiments and the results obtained from these experiments for homolog classification of chromosome 16. Finally, the conclusions are drawn in Section 5.

## 2. DATABASE PREPARATION AND FEATURE EXTRACTION

FISH is based on utilizing different fluorescent probes that bind to specific substructures of chromosomes and examining these chromosomes under a fluorescence microscope. In this study, slides of metaphase chromosomes are treated with two different probes,

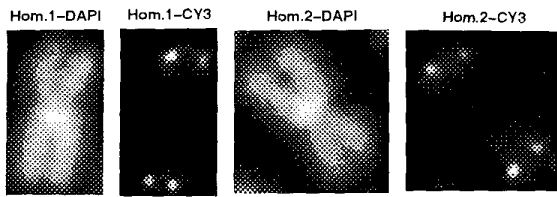


Fig. 2. A pair of homologous chromosome 16 in the database

DAPI and CY3. DAPI is a DNA probe that highlights the entire chromosome while CY3 is a PNA probe that only highlights the telomeres. Twelve images of the prepared slide are acquired using a digital camera attached to a fluorescent microscope. The database is created by isolating pairs of homologous chromosomes 16 from DAPI images as well as their corresponding CY3 image. An example of the prepared database for one pair of homologous chromosomes 16 is shown in Figure 2.

Once the database is ready, two intensity features are extracted from the DAPI and CY3 images for each chromosome 16. These features are the centromere intensities and the telomere lengths of chromosomes. As seen from the DAPI images in Figure 2, centromeres are brighter in intensity than the rest of the chromosome. Furthermore, chromosome 16 is a highly polymorphic chromosome, i.e. centromere intensity in one homolog is higher than that in the other homolog. To segment the centromeres from the DAPI images of chromosome 16, we developed a gradient method [3]. This method uses gradient histograms of DAPI images as well as chromosome intensity distributions to segment centromere areas. Centromere intensities are then measured over the segmented area for each homologous pair and saved as the first feature.

From the CY3 images of chromosomes, telomere intensities (lengths) are measured using a rank order filtering method [5]. A software developed in the Terry Fox Laboratory of BC Cancer Research Centre based on this method was used to perform telomere length measurements in the four chromosome arms namely P1, P2, Q1 and Q2 [5]. Since sister chromatids are duplicates, P1 and P2 as well as Q1 and Q2 are averaged to a single P value and a single Q value for each chromosome and normalized to zero mean with unity standard deviation over the whole database. Both the P and the Q arm telomere lengths form the second feature saved for each chromosome. In the next section, the centromere intensities as well as the P and the Q telomere lengths are used as inputs to the neural network for homolog classification.

### 3. NEURAL NETWORK ARCHITECTURE

In a previous study, two independent classifiers were designed to separate homologs of chromosomes 16 using the features extracted in Section 2 (centromere intensities and telomere lengths) [4]. The classification results from the two independent classifiers were perfectly correlated, i.e. both classifiers produced the same maternal and paternal homolog classes. These results were promising as each of the two classification methods verified the other one. In this study, we propose an unsupervised method to quantify the relationship between these two classifiers. A neural network is employed to capture the coherence in the feature set for classification purposes. A multi-layer feed forward neural network is able to represent a nonlinear function mapping between a set of input features and a set of output classes [6]. Therefore, these networks seem suitable for homolog classification problem.

The proposed network architecture consists of two separate

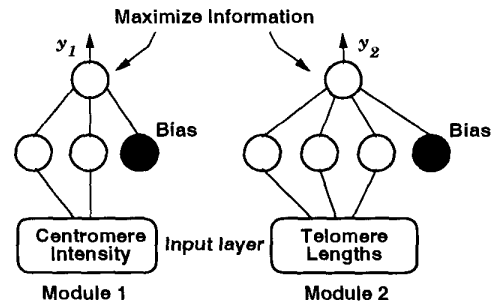


Fig. 3. Structure of the proposed neural network architecture

modules as shown in Figure 3. Each module is a three layer feed forward neural network that classifies homologs of chromosome 16 using one of the extracted features in Section 2. Our goal is to design an unsupervised learning algorithm which trains the two networks to present the same classification for a data point, based on different features.

A possible solution to this problem is to maximize the correlation between the outputs of the networks,  $y_1$  and  $y_2$ . Maximum correlation, however, is not a sufficient condition to solve our problem. Two variables can be perfectly correlated if, for example, they are nearly always zero however in this case, the outputs of the network will convey almost no information about the inputs.

A better method to train the networks is to maximize the mutual information between the outputs of the modules, instead. We apply this technique to design our learning algorithm. A brief explanation of the algorithm is as follows; In Figure 3, the first module receives centromere intensities as an input. It then learns to classify homologous chromosomes 16 based on this feature. The second module, on the other hand, receives normalized P and Q arm telomere lengths as inputs and learns to classify the homologs based on these lengths. The restriction of mutual information maximization is applied as an objective function between the outputs of the two modules. An error function is measured and back-propagated to the hidden layers of the networks through the gradients. This method trains the neural network architecture to produce the same classification results for the two modules.

As mentioned before, this architecture can be generalized to perform classification using more than two features by adding a new module to the network for each independent input feature. As a result, several well-known chromosome features (extracted from images prepared using different probes) which are speculated to be of little or no importance in homolog differentiation can be examined using our proposed architecture. Each of these features can be used as inputs to the network in combination with known "homolog differentiating" features, for specific chromosomes. The changes in classification results relative to the original classification performed with known-features only, can determine the value of the new feature in homolog differentiation.

Homolog classification using the extracted features in section 2, is not a straightforward task since these features are not linearly separable in most cases. On the other hand, the perception of the relationship between features in a pair of homologs is essential to the classification process. Therefore, the representation of the feature space to the neural network (the input structure), in addition to the training method, plays a major role in homolog classification.

### 3.1. Mutual Information Maximization

The mutual information between the outputs of the two modules of Figure 3 in continuous case is defined as:

$$I(y_1, y_2) = H(y_1) + H(y_2) - H(y_1, y_2) \quad (1)$$

$H(y_1) := E\{-\ln P(y_1)\}$  and  $H(y_1, y_2) := E\{-\ln P(y_1, y_2)\}$  are the entropies of  $y_1$  and the joint distributions of  $y_1$  and  $y_2$ , respectively [8]. From Equation (1), the mutual information can be rewritten as:

$$I(y_1, y_2) = E \left\{ \ln \frac{P(y_1, y_2)}{P(y_1)P(y_2)} \right\} \quad (2)$$

The concept of mutual information maximization has been extensively applied to signal and noise separation applications as well as to capture coherence in the input feature space [7]. In the same line, we assume the output of one module, e.g.  $y_2$ , to be a noisy version of the output of the second module,  $y_1$ , that is  $y_2 = y_1 + n$ . Therefore, in order to successfully predict  $y_1$ , the information that  $y_2$  conveys about  $y_1$  ( $I_{y_2, y_1}$ ) should be maximized. By making an assumption that both  $y_1$  and the additive noise have Gaussian distributions with zero mean,  $I_{y_2, y_1}$  can be written as [8]:

$$I_{y_2, y_1} = 0.5 \ln \frac{E\{y_2^2\}}{E\{n^2\}} = 0.5 \ln \frac{V(y_2)}{V(y_2 - y_1)} \quad (3)$$

where  $V(y_2)$  is the variance of  $y_2$  and  $V(y_2 - y_1)$  is the noise variance. In order to maximize  $I_{y_2, y_1}$ ,  $V(y_2)$  should be maximized, whereas  $V(y_2 - y_1)$  should be minimized. For symmetry and simplification, Equation (3) is generalized as:

$$I_{y_2, y_1} = \ln \frac{V(y_2)}{V(y_2 - y_1) + k} + \ln \frac{V(y_1)}{V(y_2 - y_1) + k} \quad (4)$$

According to Equation (4), each module maximizes the information its output conveys about the other module. Therefore, each module in the network will minimize  $V(y_2 - y_1)$ , maximizing both  $V(y_1)$  and  $V(y_2)$ . In order to prevent  $I$  from becoming infinite, a small constant  $k > 0$  is added to the denominators [7]. Error is then defined as  $\mathcal{E} = -I$  and is used as our objective function.

### 3.2. Error back-propagation and update rule

We use error back-propagation and batch training as our unsupervised learning algorithm. Starting with small random weights, all training cases are propagated forward through the networks to calculate outputs, activations of the hidden units and the error. For the error function defined in Section 3.1, the gradient of the error with respect to  $y_i$ , for the  $\alpha$  training case, is derived as:

$$\frac{\partial \mathcal{E}}{\partial y_i^\alpha} = -\frac{2}{N} \left[ \frac{y_i^\alpha - E\{y_i\}}{V(y_i)} - \frac{2(y_i^\alpha - y_j^\alpha) - E\{y_i - y_j\}}{V(y_i - y_j) + Nk} \right], \quad (5)$$

where  $y_i$  and  $y_j$  are the outputs of the  $i$ th and  $j$ th units in the top layer and  $N$  is the number of training cases. In order to calculate this gradient, two sweeps are made through the data. In the first sweep, the mean and variances are calculated and during the second sweep the gradient is computed. Error is then back-propagated to the hidden layers through the gradients and weight updates are measured using the update rule:

$$W_{j,i}^{new} = W_{j,i}^{old} - \eta \sum_N \delta_j N x_i^N - \beta W_{j,i}^{old} + \alpha \Delta W_{j,i}^{old} \quad (6)$$

Here  $W_{j,i}$  denotes the weight connecting the  $j$ th unit of a higher layer to the  $i$ th unit of a lower layer,  $\delta_j$  is the back-propagated

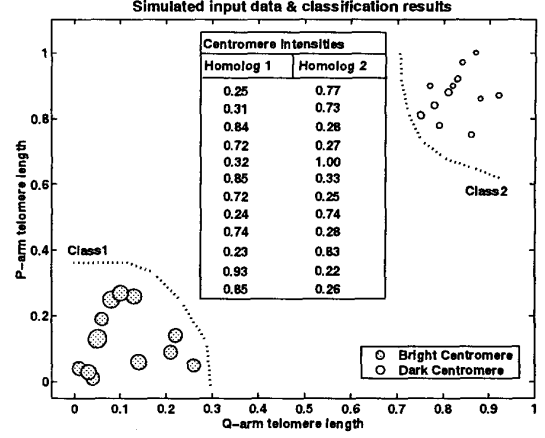


Fig. 4. Centromere intensities & telomere lengths; simulated data

gradient from the  $j$ th unit,  $x_i$  shows the output of the  $i$ th unit and  $\eta$ ,  $\beta$  and  $\alpha$  represent the learning rate, momentum coefficient and weight decay factor, respectively.

The gradient of the outputs are large at the beginning of the training and tend to get smaller as training proceeds. Therefore, we start training with very small learning rate and slowly increase it. Ideally, an adaptive algorithm should take care of the changes in the learning rate, however it is difficult to develop such an algorithm. We manually control the step size at different stages of the learning procedure. Momentum is introduced later in the training. A weight decay factor is also implemented in the update rule to prevent the network weights from blowing up.

## 4. TRAINING EXPERIMENTS DESIGN AND RESULTS

Two sets of experiments were designed to train and test the neural network. Simulated data was used in the first set of experiments where a rather simple classification problem was introduced to the network. In the second set of experiments, the real data with a more difficult classification problem was utilized.

### 4.1. Training using simulated data

In this experiment, centromere intensities and telomere lengths are simulated for twelve pairs of chromosomes 16 as illustrated in Figure 4. In this figure, the P and Q arm telomere lengths are shown as one point (circle) for each homolog. Moreover, the size of each circle is proportional to the centromere intensity of that particular homolog. Two distinct clusters of homologs are easily noticed in the figure based on the P versus Q arm telomere length distribution. On the other hand, the centromere data show that the difference in centromere intensities in one pair of homologous chromosomes is quite high. Furthermore, there is an evident distinction between homologs with bright centromeres and those with dark centromeres, i.e. the lowest intensity of the bright centered homolog in the entire database (the smallest grey circle) is well above the highest intensity of the dark centered homolog (the largest white circle). This data presents a relatively easy classification task to the neural network modules.

Training the neural network using this database, error converges to its desired minimum in a relatively low number of iterations. Studying the classification results also shown in Figure 4, the two modules classify chromosomes 16 into the same two

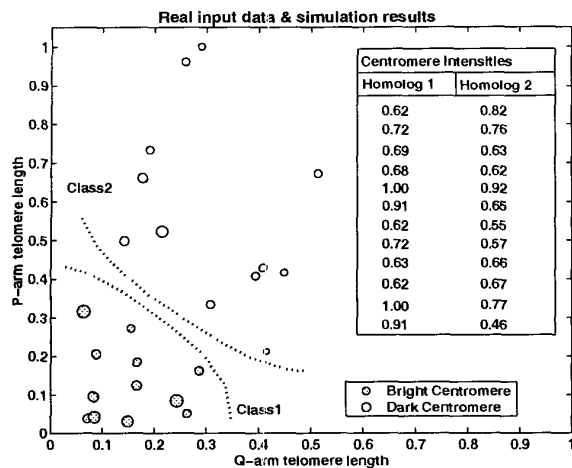


Fig. 5. Centromere intensities & telomere lengths; real data

homologous groups. Module 1 classifies chromosomes 16 into two groups of bright and dark centered homologs. On the other hand, module 2 classifies homologous chromosomes so that class 1 and 2 consist of homologs with P and Q telomere lengths distributed in the lower left corner and upper right corner of Figure 4, respectively. Mutual information between the outputs of the two networks is maximized at the end of the training period.

#### 4.2. Training using the real data

Centromere intensities and telomere lengths calculated for twelve pairs of homologous chromosomes 16 are used as training data for the neural network in this experiment. The distribution of the P and Q arm telomere lengths as well as the centromere intensities are shown in Figure 5. Here again, the size of the circles are proportional to the intensities of the centromeres. As seen, unlike the previous experiment, the distribution of the P and Q arm telomere lengths is not easily clusterable. In addition, the difference in centromere intensities within a homologous pair is rather low. Furthermore, the centromere data are not linearly classifiable as bright and dark centromeres. Module 1, in the neural network architecture proposed in Figure 3, will not be able to classify homologous chromosomes based on the current information. According to the current architecture, for training purposes, module 1 only finds a threshold and classifies the centromere values based on this threshold. Such a threshold does not exist for the real data, therefore training the network using this data will not yield satisfactory results. Module 1 in the network architecture has no knowledge of two chromosomes being homologs, i.e. if homolog 1 is classified in class 1, homolog 2 cannot be in the same class. In addition, the input structure does not provide an opportunity for this module to compare centromere intensities within a pair of homologs.

To overcome this problem, another unit is added to the input layer of module 1 in Figure 3. In the training phase, for each chromosome, the chromosome centromere intensity as well as its homologous pair centromere intensity are presented as inputs to module 1. Inputs of module 2 are still the P and Q arm telomere lengths of that chromosome. Training the network with this information, we expect that if pair  $(c_1, c_2, p_1, q_1)$  ( $c_1$ : centromere intensity of homolog 1 &  $c_2$ : centromere intensity of homolog 2) is classified in class 1 then  $(c_2, c_1, p_2, q_2)$  be classified in class 2. The clas-

sification results after training this network are shown in Figure 5 as "Class1" and "Class2". Naturally, training is slower than it was for the simulated data, however, the two modules successfully classify homologs into the same two classes. Furthermore, this architecture has the promise of being capable of classifying more complex databases.

#### 5. CONCLUSION

In this paper a new application of mutual information maximization and neural networks was introduced to classify homologs of human chromosomes. Using an unsupervised neural network architecture with two separate modules and a rather complex input feature set, homologs of chromosome 16 were successfully classified into maternal and paternal classes. The first module in the network classified the chromosomes based on the differences in their centromere intensities. The second module utilized the P arm and the Q arm telomere lengths distribution to classify homologous chromosomes 16. Mutual information between the outputs of the two modules was maximized resulting in the same classification for a particular chromosome. A unique input structure was suggested to enable the neural network to compare features within a pair of homologous chromosomes.

Our proposed classification method, can be generalized to solve complex homolog classification problems where more than two features are necessary as well as for classification of other chromosomes. More research is under way in this direction by testing the performance of the network in presence of random features in contrast to homolog-differentiating features. This network architecture can also be used to test the effectiveness of existing and new chromosome features in homolog classification. The results of this work constitute a major step towards multi-feature analysis of chromosome images. Moreover, such analysis will allow a comprehensive evaluation of centromere and telomere repeat content in various chromosome instability syndromes such as cancer.

#### 6. REFERENCES

- [1] U.M. Martens, et. al., "Short telomeres on human chromosome 17p," *Nature Genetics*, Vol.18, No.1, pp. 76-80, 1998.
- [2] P.M. Lansdorp, et. al., "Heterogeneity in telomere length of human chromosomes," *Human Mol. Gen.*, Vol.5, No.5, 1996.
- [3] P. Mousavi, R.K. Ward and P.M. Lansdorp, "Feature analysis and classification of chromosome 16 homologs using fluorescence microscopy image," *IEEE Can. J. Elec. & Comp. Eng.*, Vol.23, No.4, pp. 95-98, 1999.
- [4] P. Mousavi and R.K. Ward, "Analysis of telomere intensities in human chromosomes with applications to classification of chromosome 16 homologs," *Proc. IEEE PACRIM Conf. Comm., Comp. & Sig. Proc.*, pp. 205-208, August 1999.
- [5] S.S.S. Poon, U.M. Martens, R.K. Ward and P.M. Lansdorp, "Telomere length measurements using digital fluorescence microscopy," *Cytometry*, Vol.36, No.4, pp. 267-278, 1999.
- [6] C.M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 1995.
- [7] S. Becker and G.E. Hinton, "Spatial coherence as an internal teacher for a neural network," in *Backpropagation: Theory, Architectures and Applications*, Y. Chauvin and D. Rumelhart (eds), part of the series *Developments in Connectionist Theory*, Hillsdale, pp. 313-349, 1995.
- [8] A. Papoulis, *Probability, random variables, and stochastic processes*, McGraw-Hill, Third Edition, 1991.