Automatic Pairing of Chromosomes

Alisha DSouza

Abstract

Karyogram is a visual depiction of chromosomes as a pair-wise ordered arrangement. Chromosomes from 30 karyograms of the Lisbon-K1 dataset are automatically paired by a multiclass k-Nearest Neighbor classifier and the performance is discussed.

Final Project Report
CS 174

1. **INTRODUCTION**

Karyotype [1] is a set of characteristics that describe the chromosomes in a cell. An ordered depiction of the karyotype, as an image, in a standard format, is called a karyogram; chromosomes are arranged in pairs by size (decreasing order) and centromere position. Study of karyograms is at the heart of cytogenetics. These analyses contribute greatly to the study of chromosomal abnormalities and aberrations, genetic disorders, taxonomical relationships etcetera.

In humans, somatic cells have 23 classes of chromosomes (22 autosomes and 2 sex chromosomes), and a total of 46 chromosomes per cell; 22 pairs of chromosomes are present in each cell. In order to develop a karyogram, cells arrested at the metaphase stage of cell division are stained, by a dye, such as Giemsa [2] and imaged. The chromosomes then need to be arranged in pairs in order of decreasing size. As a result of staining, each chromosome has a unique banding pattern that aids classification. This process of pairing and karyotyping is usually done manually and requires considerable time of an expert. Automating these is an active field of research [3] and is highly desirable. Figure 1 shows a karyogram from the Lisbon-K1 dataset, where chromosomes are arranged in the order of their class. Figure 2 shows a stained chromosome with distinct banding pattern.

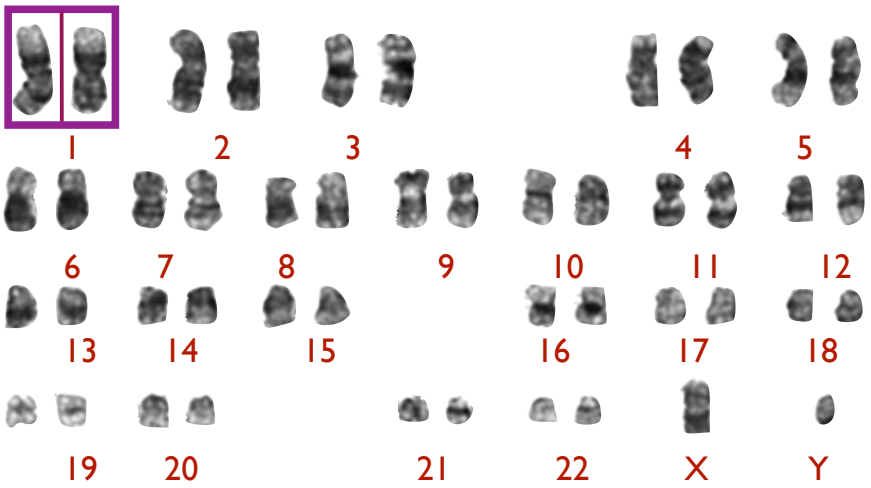

Fig.1 Karyogram 1 from Lisbon-K1 dataset. The chromosome class is indicated by the red numbering and the first pair is highlighted by a box.

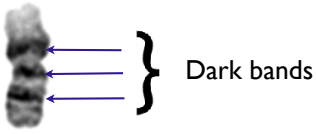

Fig.2 Visible banding pattern

1.1 DATASET

The Lisbon-K1 (LK_1) dataset [3, 15], of chromosomes from bone marrow cells of leukemia patients, developed by the technicians of Institute of Molecular Medicine of Lisbon, was used for this project. The dataset contains 200 karyograms (9200 chromosomes). For the purpose of this project a subset of 33 Karyograms from this dataset was used. This dataset

is of much lower quality than other more widely used datasets. Figure 3 shows a comparison between the LK_1 dataset and Copenhagen dataset.

Fig.3 Comparison of chromosomes of low-quality LK_1 dataset (left) from bone-marrow cells and chromosomes from high quality Copenhagen dataset (right) [16]

1.2 RELATED WORK

For pairing based on classification, numerous methods of classifier design have been proposed in literature. For example, hidden markov models [5], template matching [6], neural network and multilayer perceptron [7] – [12], wavelet [13], fuzzy [6] and Bayes [9] classifiers have been proposed. Classification success is usually in the range of 70% to 80% with these (on high quality datasets), which is much lower than the accuracy of 99.70% achieved by a human expert [3]. Khmelinskii et al propose an algorithm that pairs chromosomes directly without accurately classifying them and assistance from a rough classification, performed using Support Vector Machine (SVM) classifier is used [14].

2. **METHOD**

The chromosomes available in each karyogram are ordered and arranged according to the class to which they belong. Figure 1 shows a karyogram image. The adopted methods for pairing uses the distance between feature vectors associated with each chromosome. The distances of a given chromosome from each chromosome in the training set are calculated and the chromosome is classified to the class that is nearest to it. The following steps describe two methods adopted and tested for pairing and classification.

2.1 FEATURE EXTRACTION

In order to build a metric for calculating distance between two chromosomes, some features need to be extracted. Preceding this the chromosome images are pre-processed and geometrically corrected so that their boundaries are more-or-less parallel and an axis of symmetry if drawn would be parallel to the lateral boundaries[4].

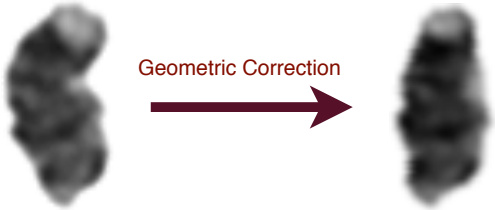

Fig. 4 Geometric Correction

The features considered can be grouped into size-based – length, width, ratio of length of width and area of bounding box – and patter-based features – band profile and mutual information.

❖ *Band profile :* Average intensity along each row of the corrected chromosome image.

❖ *Mutual Information :* This feature is always measured for pair of chromosomes and cannot be calculated for a single chromosome. The mutual information *MI* between a pair of chromosome images I_A and I_B is:

$$MI(I_A, I_B) = \sum_{a,b} p_{AB}(a,b) \log \left[\frac{p_{AB}(a,b)}{p_A(a) p_B(b)} \right]$$

where $p_{AB}(a,b)$ is the joint histogram of the images I_A and I_B and $p_A(a)$ and $p_B(b)$ are the histograms of each image respectively.

The following figure summarizes the above.

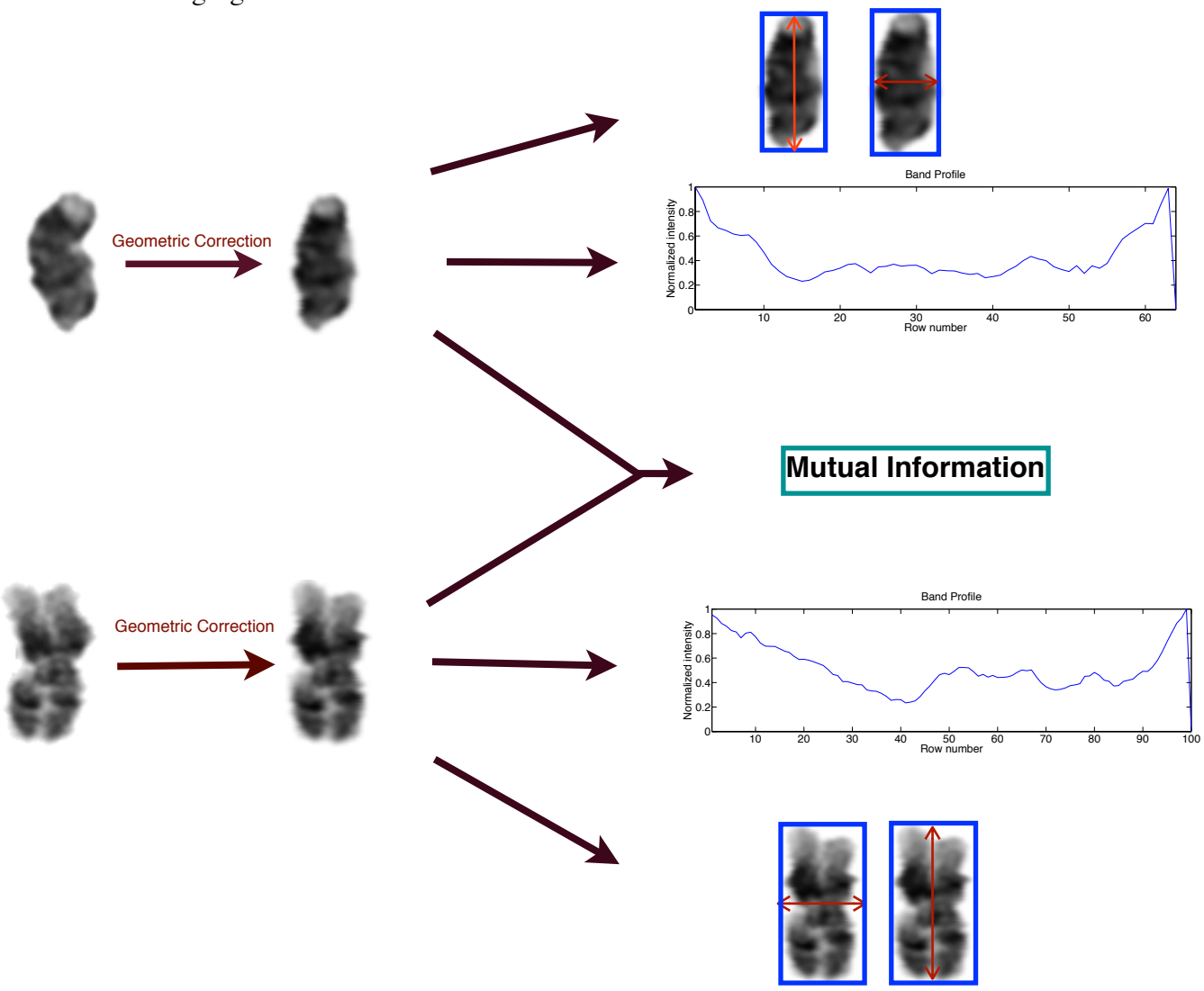

Fig. 5 Summary of feature extraction

2.2 CALCULATION OF DISTANCE BETWEEN CHROMOSOMES

Two approaches were adopted for the calculation of distance between pairs of chromosomes. The first was a weighted-distance approach and the second was a Euclidean-distance approach. Both of these are discussed.

✤ Weighted-Distance Approach

As proposed by [16], the distance between two chromosomes *i* and *j* with respect to the k^{th} feature is,

$$D(i,j;\mathbf{w}) = \sum_{k=1}^{L} w(k) d_k(i,j)$$

where *w(k)* is the weight associated with the k^{th} feature and **w** represents the weight vector.

$$\mathbf{w}_r = \underset{\mathbf{w}:\|\mathbf{w}\|=1}{\arg\min}\, E(\mathbf{w}).$$

The weights **w** are obtained during the training step by a constrained optimization of the following objective,

$$E(\mathbf{w}_i) = \underbrace{\sum_{(a,b)\in V(i)} D(a,b;\mathbf{w}_i)}_{\textbf{intra}\text{class distance}} - \underbrace{\sum_{(a,b)\in U(i)} D(a,b;\mathbf{w}_i)}_{\textbf{inter}\text{class distance}} \qquad s.t.\ \|\mathbf{w}_i\| = 1$$

Where *V(i)* is the set of chromosomes of the i^{th} class and *U(i)* is the set of chromosomes containing no more than one chromosome from the i^{th} class. So each $\mathbf{w}_i$ is computed by minimizing the sum of intraclass distances and maximizing the sum of interclass distances. This constrained optimization problem is approached using the method of Lagrange multipliers and the cost function ***E (w)*** is then,

$$E(\mathbf{w}_r) \quad = \quad \Phi_r \mathbf{w}_r + \gamma \mathbf{w}_r^T \mathbf{w}_r$$

where γ is the Lagrange multiplier and $\Phi_r = 1^T \Theta_r - 1^T \tilde{\Theta}_r$

$$\mathbf{\Theta}_r = \begin{pmatrix} d_1(1) & d_1(2) & d_1(3) & ... & d_1(L) \\ d_2(1) & d_2(2) & d_2(3) & ... & d_2(L) \\ d_3(1) & d_3(2) & d_3(3) & ... & d_3(L) \\ ... & ... & ... & ... & ... \\ d_R(1) & d_R(2) & d_R(3) & ... & d_R(L) \end{pmatrix}$$

Here each element $d_i(k)$ is the distance between the i^{th} pair of chromosomes from training set associated with the k^{th} feature such that all pairs belong to class *r*. $\mathbf{\Theta}_r$ thus represents

intraclass distances. $\tilde{\Theta}_r$ has a similar structure but involves all pairs from training set containing not more than one chromosome of class *r*. $\mathbf{w}_r$ has the closed form solution and is now the unit vector along the direction of Φ_r [16],

$$\mathbf{w}_r = \Phi_r^T / \sqrt{\Phi_r \Phi_r^T} \qquad \mathbf{w}_r \forall r \in [1, 22]$$

Once we have all $\mathbf{w}_r$'s distance between chromosomes *i* and *j* is,

$$\mathcal{D}(i,j) = \min_{r \in \{1,\ldots,22\}} D(i,j;\mathbf{w}_r)$$

✤ Euclidean-Distance Approach
Here the distance between chromosomes *i* and *j* is,

$$\mathcal{D}(i,j) = \sum_{k=1}^{L} \|d_k(i,j)\|^2$$

where $d_k(i,j)$ is the distance between the said pair with respect to the k^{th} feature.

2.3 k-NEAREST NEIGHBOR CLASSIFICATION

For a given chromosome *i* the distances $\mathcal{D}(i,j)$, where j represents all chromosomes from the training set, are calculated. The chromosome is grouped into the class which is the majority. Chromosomes of the test karyogram are paired in this way.

3. RESULTS

We see an average accuracy of classification of 38.41% with the weighted-distance approach and 52.65% with the Euclidean-distance approach for k=1. Figure 6 shows the error observed with change of k.

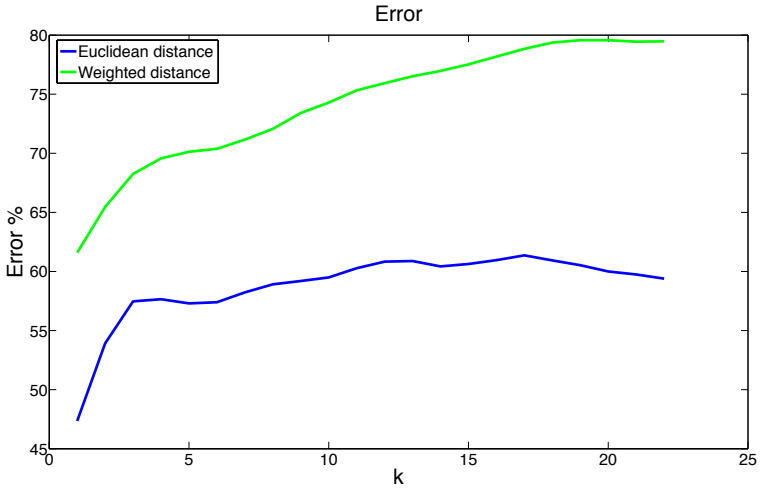

Fig.6 Error versus k

4. DISCUSSION

The pairing problem is approached as a multinomial classification problem with 22 classes. Greater accuracy is obtained by using the simpler Euclidean-Distance based k-NN classifier. This may be explained by the following:
✤ The optimization of weights in the weighted-distance based classifier requires that band profiles are of the same length for all pairs of chromosomes and cross-correlation between each pair is not necessarily maximized.

✤ The Euclidean-distance based classifier does not require such a constraint so the relative position of band profiles for a pair of chromosomes can be adjusted so that cross-correlation is maximized.

The accuracy of 52.65% is acceptable and is comparable to an accuracy of < 50.50% achieved with a Nearest Neighbor classifier on the LK_1 dataset reported by Khemlinskii et al [16].

5. CONCLUSIONS

Although the accuracy of classification is poor, it is promising. The observed trends are as expected. Use of a larger dataset, will enable a better understanding.

6. REFERENCES

[1] http://en.wikipedia.org/wiki/Karyotype#cite_note-3

[2] http://en.wikipedia.org/wiki/Giemsa_stain

[3] A. Khmelinskii, R. Ventura and J. Sanches, "Chromosome Pairing for Karyotyping Purposes using Mutual Information," *Proceedings of the 5th IEEE International Symposium on Biomedical Imaging,* 14-17 May 2008, pp 484-487.

[4] S. Khan, A. DSouza, J. Sanches and R. Ventura, "Geometric Correction of Deformed Chromosomes for Automatic Karyotyping" (accepted for publication).

[5] J. M. Conroy, R. L. Jr. Becker, W. Lefkowitz, K. L. Christopher, R. B. Surana, T. O'Leary, D. P. O'Leary, and T. G. Kolda, "Hidden markov models for chromosome identification," in *Proceedings of the 14th IEEE Symposium of Computer-Based Medical Systems,* July 2001, pp. 473– 477.

[6] A. M. Badawi, K. G. Hasan, E. A. Aly, and R. A. Messiha, "Chromosomes classification based on neural networks, fuzzy rule based, and template matching classifiers," in *Proceedings of the 46th IEEE International Midwest Symposium on Circuits and Systems,* Dec. 2003, vol. 1, pp. 383–387.

[7] J. R. Stanley, M. J. Keller, P. Gader, and W. C. Caldwell, "Data-driven homologue matching for chromosome identification," *IEEE Transactions on Medical Imaging,* vol. 17, no. 3, pp. 451–462, 1998.

[8] M. Zardoshti-Kermani and A. Afshordi, "Classification of chromosomes using higher-order neural networks," in *IEEE International Conference on Neural Networks,* Nov.-Dec. 1995, vol. 5, pp. 2587–2591.

[9] B. Lerner, H. Guterman, I. Dinstein, and Y. Romem, "A comparison of multilayer perceptron neural network and bayes piecewise classifier for chromosome classification,"

IEEE World Congress on Neural Networks, IEEE International Conference on Computational Intelligence, vol. 6, pp. 3472–3477, June-July 1994.

[10] B. Lerner, M. Levinstein, B. Rosenberg, H. Guterman, L. Dinstein, and Y. Romem, "Feature selection and chromosome classification using a multilayer perceptron neural network," *IEEE World Congress on Computational Intelligence., IEEE International Conference on Neural Networks,* vol. 6, pp. 3540–3545, Jun.-Jul. 1994.

[11] B. Lerner, "Toward a completely automatic neural-network-based human chromosome analysis," *IEEE Transactions on Systems, Man and Cybernetics, Part B,* vol. 28, no. 4, pp. 544–552, Aug. 1998.

[12] J. M. Cho, "Chromosome classification using backpropagation neural networks," *IEEE Engineering in Medicine and Biology Magazine,* vol. 19, no. 1, pp. 28–33, Jan.-Feb. 2000. [13] Q. Wu and K. R. Castleman, "Automated chromosome classification using wavelet-based band pattern descriptors," in *13th IEEE Symposium on Computer-Based Medical Systems,* June 2000, pp. 189–194.

[14] A. Khmelinskii, R. Ventura, J. Sanches, "Classifier-assisted metric for chromosome pairing," *IEEE International Conference of Engineering in Medicine and Biology Society,* 2010, 6729-6732.

[15] http://mediawiki.isr.ist.utl.pt/wiki/Lisbon-K_Chromosome_Dataset

[16] A. Khmelinskii, R. Ventura, J. Sanches, "A Novel Metric for Bone Marrow Cells Chromosome Pairing," *IEEE Transactions on Biomedical Engineering,* vol.57, no.6, pp. 1420-1429, June 2010