Automatic Pairing of Chromosomes

Alisha DSouza

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. 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.

Objective

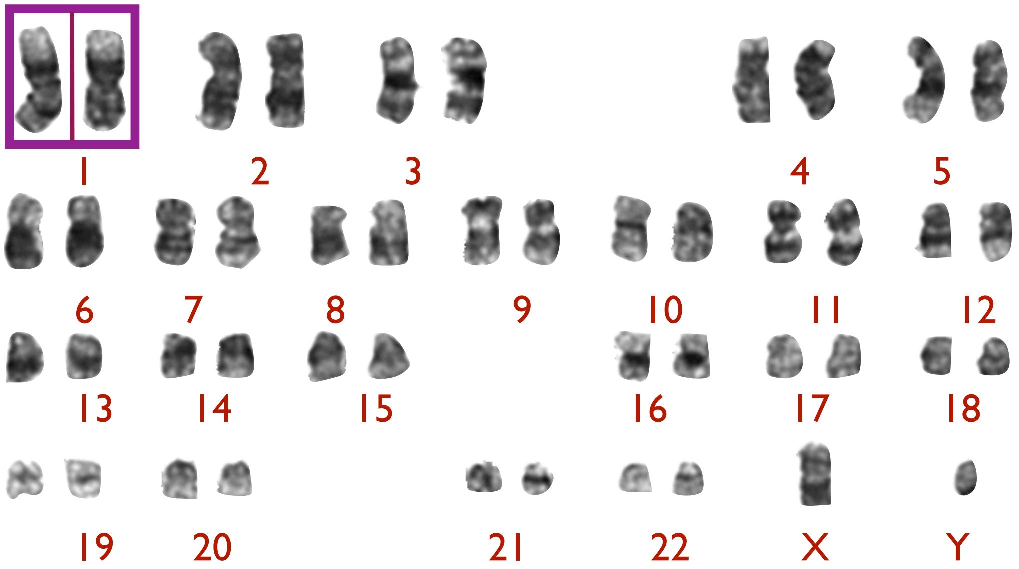
The goal of this project is to automatically pair chromosomes from a karyogram.

Dataset

The Lisbon-K1 dataset [3, 6], of chromosomes from bone marrow cells of leukemia patients, developed by the technicians of Institute of Molecular Medicine of Lisbon, will be 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 will be used.

Method

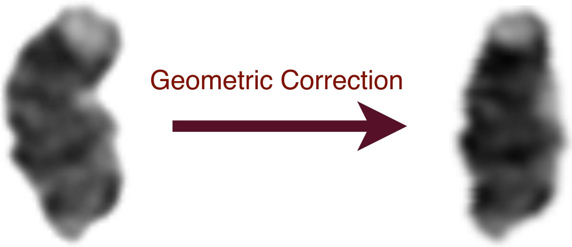
The chromosomes available in each karyogram are ordered and arranged according to the class to which they belong. The following karyogram image shows a highlighted pair and class numbers.



The adopted method 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 the method adopted for pairing and classification.

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].



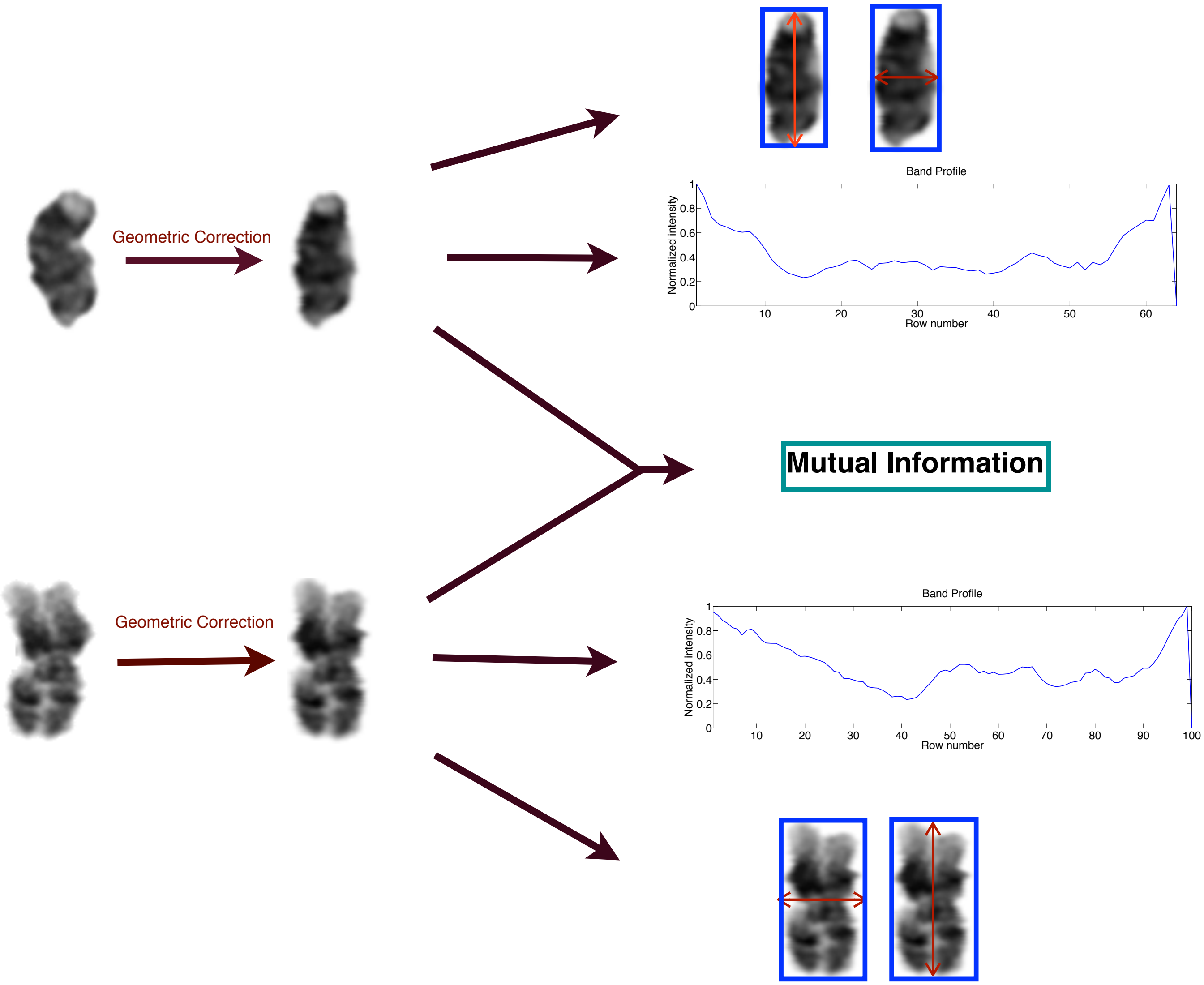
The features considered can be grouped into size-based – length, width and area of bounding box – and patter-based features – band profile and mutual information.

- *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.



2. Calculation of Distance between Chromosomes

As proposed by [5], 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.

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

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

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

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 w_i is computed by minimizing the sum of intraclass distances and maximizing the sum of interclass distances. This constrained optimization technique 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

ϕ_r = sum along columns **($\boldsymbol{\Theta}_r$) –** sum along columns**($\boldsymbol{\theta}_r$)**

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

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. $\mathbf{\theta_r}$ has a similar structure but involves all pairs from training set containing not more than one chromosome of class r. $\mathbf{w}_r$ is now the unit vector along the direction of ϕ_r [5],

$$\mathbf{w}_r \quad = \quad \Phi_r^T / \sqrt{\Phi_r \Phi_r^T}$$

Once all $\mathbf{w}_r$ $\forall$ r $\in [1, 22]$ are obtained, distance between chromosomes i and j is,

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

3. **Nearest Neighbor Classification and Pairing**

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 classified into the class of the chromosome from the test set to which it is nearest. Chromosomes of the test karyogram are paired in this way.

Milestone Achievements

The feature vectors associated with each chromosome have been obtained and preliminary pairing results are available, by the method described above. The accuracy of pairing by selecting karyogram 1 as the training set and testing on karyograms 2 and 3 averages to 30%. This low pairing accuracy may be due to various reasons – low quality of the Lisbon-K1 dataset, insufficient features or poor quality of feature extraction, naive classification.

Future Work

The following directions will be explored before the final presentation.

- Test the above method over all 33 karyograms.
- Increase the size of the training set.

- Improve quality of the feature set to improve pairing results
- Classification by a more sophisticated algorithm, possibly multiclass SVM.

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" (under review).

[5] 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

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