# CHROMOSOME PAIRING FOR KARYOTYPING PURPOSES USING MUTUAL INFORMATION

*Artem Khmelinskii, Rodrigo Ventura and João Sanches*

Institute for Systems and Robotics
Instituto Superior Técnico
Av. Rovisco Pais, 1; 1049-001 Lisboa, Portugal

**ABSTRACT**

Cytogenetics is the preferred tool in the diagnosis of genetic diseases such as leukemia and detection of aquired chromosomal abnormalities, such as translocations, deletions, monosomies or trisomies, etc. The *karyotyping* is a set of procedures, in the scope of the *cytogenetics*, that produces a visual representation of the 46 chromosomes, paired and arranged in decreasing order of size, observed during the metaphase step of the cellular division (*meiosis*).

The pairing is the procedure in the *karyotyping* process where the homologous chromosomes are paired according to dimensional, morphological and textural similarity criteria. This process is time consuming and is usually performed manually by experts. An automatic pairing algorithm is still an open problem.

In this paper we present new contributions to solve the automatic pairing problem in the scope of the *karyotyping* process for *leukemia* diagnostic purposes. Besides the traditional features used to compute the similarity between chromosomes, such as, normalized area, ellipsis axis length and banding profiles, we introduce the *Mutual Information* (MI) measure to assess the textural similarity between two chromosomes.

A supervised linear classifier is trained to combine the different features computed from each pair, aiming at the correct pairing (as given by experts). The resulting classifier is then employed, together with a combinatorial optimization algorithm based on A*, to compute the pairing for any given image.

Simulations using real images, obtained with a Leica$^{TM}$Optical Microscope DM 2500, were performed. These images were manually paired by experts and used as a ground truth for the pairing process to assess the performance of the proposed classifier. Furthermore, qualitative comparisons with the results obtained with a Leica$^{TM}$CW 4000 Karyo software were also performed.

***Index Terms***— Chromosome Pairing, Leukemia, Image processing, Classification, Mutual Information

## 1. INTRODUCTION

The karyotyping procedure is one of the most important steps in conventional *cytogenetic* analysis. The *karyogram* is an image representation of stained human chromosomes with the widely used Giemsa Stain metaphase spread (G-banding) in which homologous chromosomes are paired in 23 classes, arranged in order of decreasing size. A *karyotype* is the set of characteristics extracted from the *karyogram* that may be used to detect chromosomal abnormalities, such as, translocations, duplications, inversions, deletions, monosomies or trisomies (some of these abnormalities occur in leukemia cancerous cells). Fig.1 shows a typical image of a normal male *karyotype*. The *karyotyping* procedure is time consuming and technically demanding, when done manually. Automatic algorithms are needed but the difficulty of the problem makes it hard to design accurate and reliable automatic processing algorithm for *karyotyping* purposes. Namely, the essential process of chromosome pairing is still an open problem. For instance, the most widely used commercial packages available for *cytogenetic* analysis like Leica$^{TM}$, Metasystems$^{TM}$and Cytovision$^{TM}$are still very ineffective when it comes to chromosome classification and/or pairing.

The problem of the automated chromosome classification has been an important pattern recognition problem for more than 20 years and remains an active field of research [1–6].The more specific problem of chromosome pairing has been also investigated, namely by Wu et al [7,8].

The main issue in this paper is the chromosome pairing and not the classification itself. The ultimate goal is to design a pairing algorithm for *karyotyping* purposes in order to help the technical staff in this important step of the *cytogenetic* analysis.

The proposed algorithms use a linear classifier based on the traditional dimensional and morphological features extracted from the *karyogram* and a new one based on the *mutual information* (MI). The goal is to better characterize the textural information associated with each pair by adding discriminative power to the G-banding profiles information. The proposed algorithms characterize pairs of chromosomes rather than isolated ones.Therefore measures are extracted from each chromosome and combined in a pairwise basis to obtained features associated to candidate pairs.

The images were acquired with a Leica$^{TM}$Optical Microscope DM 2500 and the image pre-processing and chromosome segmentation were performed with Leica$^{TM}$CW 4000 Karyo software used in the Institute of Molecular Medicine of Lisbon (IMM). The pairing process has been performed in this institute mostly in a semi-manual fashion.

The images used in the *karyotyping* process for leukemia diagnostic purposes, in which we are interested in, present less quality than the ones used in the traditional genetic analysis that use the sets like Edinburgh, Copenhagen[1] and Philadelphia[9], namely with respect to the centromere, band profile description, and level of chromosome condensation.

Tests using real data have shown promising results when compared with the pairing results provided by the Leica$^{TM}$CW 4000 Karyo software.

This paper is organized as follows: section 2 formulates the problem, describing the features used in a pairwise basis, and sec-

Correspondent author: Artem Khmelinskii (aihmel@gmail.com). This work was supported by Fundação para a Ciência e a Tecnologia (ISR/IST plurianual funding) through the POS Conhecimento Program which includes FEDER funds.

tion 3 describes the classifier and the training procedure. Section 4 shows illustrative examples using several real data sets, and section 5 concludes the paper.

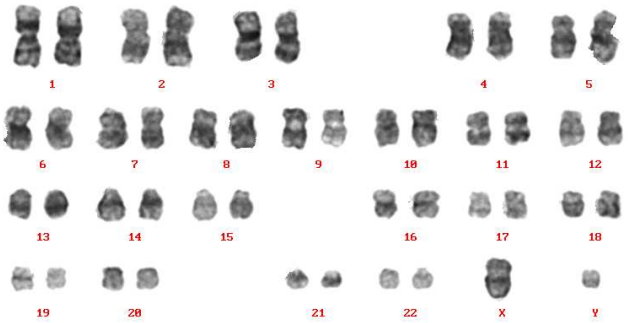

**Fig. 1**. Normal male karyotype.

## 2. PROBLEM FORMULATION

The data is composed by $2N$ chromosomes images obtained with a Leica$^{TM}$Optical Microscope DM 2500 and pre-processed by the Leica CW 4000 Karyo software where $N$ is the number of homologous pairs. In the pre-processing stage the images are denoised and the chromosomes are manually segmented (isolated) in a computer assisted basis. Therefore, in this paper, the data is composed by a set of isolated and unpaired chromosomes as shown in Fig.1.

The first step in the proposed methodology is a pre-processing algorithm where the following operations are performed: i) bounding box computation isolating each chromosome, ii) independent histogram equalization of the images contained in each bounding box (textural correction), and iii) chromosome axis determination and distortion correction (geometric correction).

This pre-processing step aims at reducing the effects of the textural and geometric distortions in the processing phases that follow. The brightness and contrast characteristics depend on the specific tuning of the microscope, and therefore the final results should be insensitive to it. On the other hand, the particular geometric distortion observed in a given chromosome is related with the particular methaphasic plaque under processing, and not with the intrinsic geometric characteristics of the chromosome, therefore it should also be compensated.

The proposed methodology consists of two phases. In the first phase, a classifier is trained based on a set of paired images provided by experts. This classifier is then used in the second phase to process new images. Then, the resulting pairing is obtained using a combinatorial optimization technique based on the A* algorithm [10].

The features used by the classifier to pair the chromosomes are distance metrics between each pair based on measures computed from the chromosomes such as: i) Size — axis dimensions of the ellipsis containing the chromosome, length proportion, perimeter, length, and area [2], ii) Shape — bounding box normalized area and iii) Texture — band profile [1] and mutual information.

The features related with texture, such as band profile and mutual information, are computed after resizing all bounding boxes to the same dimension by interpolation.

The distance between two chromosomes is computed as follows

$$d(i,j;\mathbf{w}) = \sum_{k=1}^{L} w(k)\, \|f_i(k), f_j(k)\|_k \qquad (1)$$

where $\|x,y\|_k$ denotes the metric function used to compare the $k$th features of the $x$th and $y$th chromosomes. All features except one use the Euclidean metric, i.e., $\|x,y\|_k = \sqrt{\sum_{r=1}^{n}[x(r)-y(r)]^2}$ where $x, y \in R^n$. The only exception is the *mutual information* (MI) feature, $\mathrm{MI}(i,j)$, whose values already are referred to each pair.

During the training step, described next, a set of vector weights, $\mathbf{w}_r$ with $1 \le r \le N = 22$ (in this paper the sexual chromosomes were not considered) are obtained from all pairs of chromosomes for each class $r$ in the training set.

The distance between two chromosomes is assumed to be the smallest one among all weight vectors $\mathbf{w}_r$,

$$D(i,j) = \min_{\mathbf{w}\in\{\mathbf{w}_1,\ldots\mathbf{w}_{22}\}} d(i,j;\mathbf{w}) \qquad (2)$$

The vectors $\mathbf{w}_r$ are obtained by minimizing an energy function under the constraint $\|w\| = 1$, $\mathbf{w}_r = \arg\min_{\mathbf{w}:\|\mathbf{w}\|=1} E(\mathbf{w})$. In this paper two energy functions are tested,

$$E_1(\mathbf{w}) = \left[\sum_{(i\wedge j)\in V(r)} d(i,j;\mathbf{w}) - \sum_{(i\vee j)\notin V(r)} d(i,j;\mathbf{w})\right]$$

$$E_2(\mathbf{w}) = \left[\sum_{(i\wedge j)\in V(r)} d^2(i,j;\mathbf{w}) - \sum_{(i\vee j)\notin V(r)} d^2(i,j;\mathbf{w})\right]$$

where $V(r)$ is the set of chromosomes of the $r$-th class. The algorithm that minimizes $E_1$ is called *method A* and the algorithm that minimizes $E_2$ is called *method B*.

In the *method A* each weight vector $\mathbf{w}_r$ is computed by minimizing the sum of intraclass distances and maximizing the sum of interclass distances. The *method B* is *mutatis mutandis* of *A*, except that squared distances are used instead.

Let us consider

$$\mathbf{F}_r = \begin{pmatrix} f_1(1) & f_1(2) & f_1(3) & \ldots & f_1(L) \\ f_2(1) & f_2(2) & f_2(3) & \ldots & f_2(L) \\ f_3(1) & f_3(2) & f_3(3) & \ldots & f_3(L) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ f_R(1) & f_R(2) & f_R(3) & \ldots & f_R(L) \end{pmatrix}. \qquad (3)$$

a $R \times L$ matrix where $L$ is the number of features and $R$ the number of different pairs of chromosomes in the training set from class $r$. Let us also consider the matrix $\tilde{\mathbf{F}}_r$ with the same structure of $\mathbf{F}_r$ but now involving all pairs of the training set where at least one of the indices does not belong to class $r$.

By using the Lagrange method both energies may be written as follows

$$\begin{aligned} E_1(\mathbf{w}_r) &= (\mathbf{1}^T\mathbf{F}_r - \tilde{\mathbf{1}}^T\tilde{\mathbf{F}}_r)\mathbf{w}_r + \gamma\mathbf{w}_r^T\mathbf{w}_r \\ &= \Phi\mathbf{w}_r + \gamma\mathbf{w}_r^T\mathbf{w}_r \qquad (4) \\ E_2(\mathbf{w}_r) &= (\mathbf{F}_r\mathbf{w}_r)^T(\mathbf{F}_r\mathbf{w}_r) - (\tilde{\mathbf{F}}_r\mathbf{w}_r)^T(\tilde{\mathbf{F}}_r\mathbf{w}_r) + \gamma\mathbf{w}_r^T\mathbf{w}_r \\ &= \mathbf{w}^T\Theta\mathbf{w} + \gamma\mathbf{w}_r^T\mathbf{w}_r \qquad (5) \end{aligned}$$

where $\mathbf{1}$ is a column vector of ones, $\Theta = \mathbf{F}_r^T\mathbf{F}_r - \tilde{\mathbf{F}}_r^T\tilde{\mathbf{F}}_r$, $\Phi = \mathbf{1}^T\mathbf{F}_r - \tilde{\mathbf{1}}^T\tilde{\mathbf{F}}_r$, and $\gamma$ is the Lagrange multiplier. The minimizers of $E_1(\mathbf{w}_r)$ and $E_2(\mathbf{w}_r)$ are respectively

$$\begin{aligned} (\mathbf{w}_r)_1 &= \Phi^T/\sqrt{\Phi\Phi^T} = \mathrm{vers}(\Phi) \qquad (6) \\ (\mathbf{w}_r)_2 &= u_\Theta \qquad (7) \end{aligned}$$

where $\mathrm{vers}(\Phi)$ is the unit length vector aligned with $\Phi$ and $u_\Theta$ is the unit norm eigenvector of $\Theta$ that minimizes $\mathbf{w}_r^T\Theta\mathbf{w}_r$.

The equations (6) and (7) are used to compute the set of vectors $\mathbf{w}_r$, with $1 \le r \le 22$, which are then used in turn to compute the distance between two chromosomes using the expression (2).

## 3. CLASSIFIER

Given a test set of $n$ chromosomes a $n \times n$ matrix of distances is computed by using the expression (2), $\mathbf{D} = \{D(i,j)\}$ where each element is the distance between the $i$th and $j$th chromosomes. This matrix is symmetric, and in the ideal case, the minimum entry of each row correspond to the right pairing (no column contains two entries that are minimum values in the respective rows).

The pairing process is a computationally hard problem because the final pairing matrix must minimize the overall distance which means that no local decisions could be taken. Next, a description of the algorithm used to solve this combinatorial optimization task is performed.

Considering $n$ chromosomes (for $n$ even), a pairing assignment $P$ is defined as a set of ordered pairs $(i,j)$, such that $i \neq j$ holds for any pair, and any given index $i$ appears in no more than one pair of the set. A pairing assignment is said to be *total* if, for any $i = 1, \ldots, n$, there is exactly one pair $(r,s)$ in the set such that either $i = r$ or $i = s$. The sum of distances implied by a pairing $P$ can be written as

$$C(P) = \sum_{(i,j) \in P} D(i,j) \tag{8}$$

The goal of the pairing process is then to find a total pairing $P$ that minimizes $C(P)$. To accomplish this goal, the well-known A* search algorithm is used [10]. The state space considered by the search algorithm consists of tuples in the form $\langle F, P \rangle$, where $F$ is a set of chromosomes (called the *free set*), and $P$ is a pairing assignment. Only valid solutions are considered by the search algorithm, thus, the set $F$ and the set of all indices in $P$ form disjoint sets, which union results in the set of all chromosomes $\{1, \ldots, n\}$. The successor function consist in adding pairs to $P$: after selecting a pivot $i \in F$, create a new state $\langle F', P' \rangle$ for each $j \in F \backslash \{i\}$, adding the pair $(i,j)$ to it ($P' = P \cup \{(i,j)\}$), and removing $i$ and $j$ from the corresponding free set ($F' = F \backslash \{i,j\}$). The choice of this successor function guarantees that no repeated states are generated, thus dispensing the need to deal with repeated states. Provided that the heuristic function is admissible, *i.e.,* the true cost from a given state and any solution state is never over-estimated, the A* is guaranteed to be complete and optimal. It is complete in the sense that, unless no solution exist, at least one solution is found, and it is optimal in the sense that, if more than one solution exists, the optimal one is returned. It is also optimally efficient, meaning that no other optimal search algorithm is guaranteed to expand fewer nodes [10]. The cost function $g$ is the sum of distances defined above, $g(\langle F, P \rangle) = C(P)$, with the heuristic function. The heuristic function $h$ is defined in the following way: given a state $\langle F, P \rangle$, build a matrix $D'$ containing only the rows and columns from the distances matrix $D$ which indices are in the set $F$; then, sum the minimum distance in each row (excluding the main diagonal) divided by 2 (otherwise, each distance would be accounted twice). Formally one can write

$$h(\langle F, P \rangle) = \sum_{i \in F} \min_{j \in F \backslash \{i\}} D(i,j)/2 \tag{9}$$

This heuristic function is admissible because the contribution of the pairs still to be made from $F$, to the sum of distances of any successor solution node, is greater or equal than $h(\langle F, P \rangle)$, for any valid state. The admissibility of this heuristic implies that the proposed algorithm finds the optimal solution for the pairing assignment problem [10].

## 4. EXPERIMENTAL RESULTS

In this section the results obtained with 19 real *karyograms* are presented. Two test sets with growing pairing difficulty are considered. The first set contains chromosomes only from classes 1, 10, 16, and 21, while the second, with higher pairing difficulty, contains chromosomes from classes 1, 3, 10, 12, 15, 16, 21 and 22. This means that the first testing set is composed by $19 \times 4 \times 2 = 152$ chromosomes and the second by 304 chromosomes. The chromosomes in each data set are classified and paired manually by experts, thus providing ground truth to assess the performance of the automatic pairing algorithms proposed in this paper.

These two test sets were used with both classifiers proposed in this paper and the results are listed in tables 1 and 2. These tables display the number of pairing errors in each experiment where 18 *karyograms*/pairs are used for training and the remaining one for testing. In these tables 19 lines are listed corresponding to the 19 possible tests using this strategy, that is, using all but one pair for training and using the remaining one for testing (leave-one-out cross-validation). The features used in these tests to compute the distance between pairs of chromosomes are the following: area, perimeter, length proportion, dimensions of the main axis of the ellipsis containing the chromosomes, normalized area, band profile, and mutual information (MI).

Besides the overall characterization of the classifier we are also interested in evaluating the improvement due to the introduction of the MI as a discriminative factor for the pairing process. Therefore, tables 1 and 2 also display the pairing results with and without MI for comparison purposes. In these tables the symbol $\surd$ is used to indicate that a correct pairing was obtained.

From tables 1 and 2 it is concluded that both methods provide almost the same results. The only exception in the set of all tests is the case of the test set 13 with MI where the *method A* performs better than *method B*. However, more tests are needed to clearly conclude whether *method A* is better than *method B*.

It is also concluded that the introduction of the MI in the set of features leads to an improvement in the pairing results in the case of experiments 6 and 12 with the data set with 8 classes. Again, additional tests are needed to confirm this result.

The classification time is dependent on the distance matrix $\mathbf{D}$ but in all tests performed here it is of the order of few milliseconds. Although combinatorial optimization problems are in general very hard to solve, these short solving times can be explained by the fact that the distances matrices are easy to handle, in the sense that the A* algorithm is able to proceed directly to the goal very quickly.

In the experiments presented here, listed in tables 1 and 2, it is observed a correct pairing with the *method A* and *mutual information* in 63% of the tests.

## 5. CONCLUSION

In this paper two pairing algorithms are proposed for pairing purposes in the scope of *karyotyping* process used in *cytogentic* analysis. The proposed algorithms are based on the traditional features extracted from the *karyogram*, such as, dimensions and banding profiles.

Here a new feature, the *mutual information* (MI), was introduced, to improve the discriminative power of the automatic pairing algorithm.

The ultimate goal of this work is to produce a reliable and accurate pairing method to be used in the scope of the *cytogenetics*, rather than a chromosome classifier.

Tests using 19 *karyograms* and a *leave-one-out cross-validation* (LOOCV) strategy allow to conclude that the proposed pairing al-

| | *Data set 1* | | *Data set 2* | |
|---|---|---|---|---|
| | without MI | with MI | without MI | with MI |
| 1 | √ | √ | √ | √ |
| 2 | √ | √ | √ | √ |
| 3 | 4 | 4 | 2 | 2 |
| 4 | √ | √ | √ | √ |
| 5 | √ | √ | √ | √ |
| 6 | 2 | 2 | 4 | 2 |
| 7 | √ | √ | 2 | 2 |
| 8 | √ | √ | √ | √ |
| 9 | √ | √ | 2 | 2 |
| 10 | √ | √ | 5 | 5 |
| 11 | 2 | 2 | √ | √ |
| 12 | √ | √ | 2 | √ |
| 13 | √ | √ | 4 | √ |
| 14 | √ | √ | 2 | 2 |
| 15 | √ | √ | √ | √ |
| 16 | √ | √ | √ | √ |
| 17 | 2 | 2 | 2 | 2 |
| 18 | √ | √ | √ | √ |
| 19 | √ | √ | √ | √ |

**Table 1**. Simulation results (number of mispairing pairs in a total of 19 pairs) using *method A* and two training/test sets: *Data set 1* with 4 types of chromosomes and *Data set 2* with 8 types of chromosomes. In both cases results with and without MI are presented.

| | *Data set 1* | | *Data set 2* | |
|---|---|---|---|---|
| | without MI | with MI | without MI | with MI |
| 1 | √ | √ | √ | √ |
| 2 | √ | √ | √ | √ |
| 3 | 4 | 4 | 2 | 2 |
| 4 | √ | √ | √ | √ |
| 5 | √ | √ | √ | √ |
| 6 | 2 | 2 | 4 | 2 |
| 7 | √ | √ | 2 | 2 |
| 8 | √ | √ | √ | √ |
| 9 | √ | √ | 2 | 2 |
| 10 | √ | √ | 5 | 5 |
| 11 | 2 | 2 | √ | √ |
| 12 | √ | √ | 2 | √ |
| 13 | √ | √ | 4 | 4 |
| 14 | √ | √ | 2 | 2 |
| 15 | √ | √ | √ | √ |
| 16 | √ | √ | √ | √ |
| 17 | 2 | 2 | 2 | 2 |
| 18 | √ | √ | √ | √ |
| 19 | √ | √ | √ | √ |

**Table 2**. Simulation results (number of mispairing pairs in a total of 19 pairs) using *method B* and two training/test sets: *Data set 1* with 4 types of chromosomes and *Data set 2* with 8 types of chromosomes. In both cases results with and without MI are presented.

gorithms, working with a limited number of classes ($\leq 8$), achieve 100% pairing accuracy in 63% of the tests (best case scenario).

Preliminary comparison with the results obtained with the Leica CW 4000 Karyo software, using the same data, have shown relevant and promising improvement. In the near future, detailed comparison with this software and other methods will be performed, in order to validate our algorithm.

This is an early stage toward a practical pairing software. Additional, more discriminative features, as well as more complex classifiers must be used. However, this work have shown that besides the huge difficulty of the problem, it is possible to design automatic classifiers to help the *cytogenetic* technicians during the pairing process of the *karyotyping* procedure.

## 6. ACKNOWLEDGMENTS

We would like to thank the clinical staff of the Cytogenetics/Virology Laboratory of the Institute of Molecular Medicine (IMM) - GenoMed of Lisbon (namely Sónia Santos) for providing us with the karyograms much needed for the testing of the implemented algorithms.

## 7. REFERENCES

[1] J. Piper and E. Granum, "On fully automatic feature measurement for banded chromosome classification," *Cytometry*, , no. 10, pp. 242–255, 1989.

[2] 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.

[3] 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.

[4] 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.

[5] P. Mousavi, R.K. Ward, P.M. Lansdorp, and S.S. Fels, "Multi-feature analysis and classification of human chromosome images using centromere segmentation algorithms," in *Proceedings of the International Conference on Image Processing*, Sep. 2000, vol. 1, pp. 152–155.

[6] 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.

[7] X. Wu, P. Biyani, S. Dumitrescu, and Q. Wu, "Globally optimal classification and pairing of human chromosomes," in *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, Sep. 2004, pp. 2789–2792.

[8] P. Biyani, X. Wu, and A. Sinha, "Joint classification and pairing of human chromosomes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 102–109, April-June 2005.

[9] 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.

[10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, second edition, 2003.