30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

Automatic Chromosome Pairing Using Mutual Information

Artem Khmelinskii, Rodrigo Ventura and João Sanches

***Abstract*— Cytogenetics is a key tool in the detection of acquired chromosomal abnormalities and in the diagnosis of genetic diseases such as leukemia. The *karyotyping* is a set of procedures, in the scope of the *cytogenetics*, that produces a visual representation of the** 46 **chromosomes (called *karyogram*), paired and arranged in decreasing order of size.**

The pairing procedure aims to identify all pairs of homologous chromosomes. The pairing criterion is based on dimensional, morphological, and textural features similarity. This process is time consuming when performed manually, and demanding from a technical point of view. An automatic pairing algorithm would thus bring benefits, but it remains an open problem to date.

In this paper a new strategy for automatic pairing of homologous chromosomes is proposed. Besides the traditional features described in the literature, the *Mutual Information* (MI) is used to discriminate chromosome textural differences. A supervised non-linear classifier is trained by using manual classifications provided by expert technicians, combining the different features computed from each pair.

Simulations using 836 **real chromosome images, obtained with a Leica^{TM} Optical Microscope DM 2500, in a leave-one-out cross-validation fashion, were performed for training and testing the algorithm. Promising and relevant results were found, despite the poor quality of the original chromosome images, contrasting with state-of-the-art algorithms and datasets found in the literature.**

***Index Terms*— Chromosome, Pairing, Leukemia, Image Processing, Optical Microscope, Mutual Information, Optimization, Classification, Integer Programming**

I. Introduction

The *cytogenetic* is used for detection of chromosomal abnormalities occurring in several genetic diseases such as *Down syndrome* or *leukemia*. The *karyogram* is an image representation of stained human chromosomes with the widely used Giemsa Stain metaphase spread (G-banding) in which chromosomes are paired in 22 classes of homologous chromosomes and two sex-determinative chromosomes (XX for the female or XY for the male), arranged in order of decreasing size. The *karyotype* of a patient is the set of characteristics extracted from the *karyogram* that may be used for diagnosis purposes. Fig. 1 shows a typical normal male *karyotype* where the chromosomes are observed during the *metaphase* stage of the cellular division called *mitosis*. In this stage the chromosomes are at their most condensed state appearing better defined than in all others stages of the cellular cycle.

To form the *karyogram*, the chromosome images, extracted from the *metaphase plate*, must be segmented and paired. Very often this pairing procedure is done by visual inspection which is a time consuming and technically demanding task. Automatic pairing is needed but it is a difficult 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 with respect to chromosome classification and/or pairing. In fact, the problem of the automated chromosome classification has been an active field of research in the last 20 years and it still is an open problem [2]–[7].

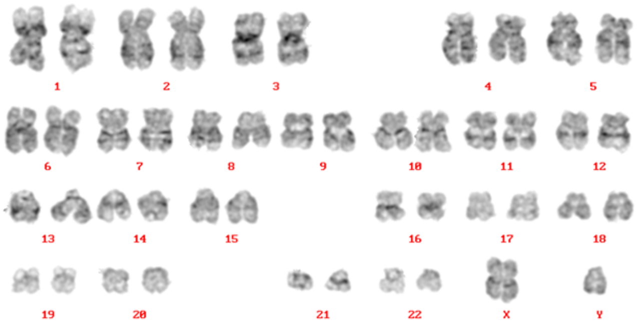

Fig. 1. Normal male karyotype.

This paper is focused in the specific task of pairing in the scope of *karyotyping* procedure. The motivating goal is to design an automatic pairing algorithm in order to improve the productivity of the technical staff of the Institute of Molecular Medicine of Lisbon (IMM) in this type of the *cytogenetic* exams.

This paper presents improvements with respect to the algorithm described in [1] where the *mutual information* (MI) is added to the traditional dimensional and morphological features to compare chromosomes. 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 algorithm computes similarity of chromosomes instead of classifying them individually.

In this paper, improvements with respect to [1] on the optimization algorithm used in the pairing procedure itself are described. Additionally, refinements and tuning on the feature extraction process makes it possible to use a complete set of 22 pairs of chromosomes with a classifier performance comparable with the state-of-the-art algorithms described in the literature. A new extended set of *karyograms* is used in this paper for training and testing the classifier allowing for a better characterization and validation of the pairing algorithm. Notice that the *karyograms* used in leukemia diagnosis, obtained from *bone marrow* cells in which we are interested in, present much less quality than the ones used in the traditional genetic analysis that uses the sets like Edinburgh [2,4], Copenhagen [2,3] or Philadelphia [2] where the images have better quality and less variability than the chromosomes images obtained from the *bone marrow* cells used in this work.

This paper is organized as follows: section II contains a problem formulation description and section II-B describes the classifier and the training procedure. Section III shows the experimental results and section IV concludes the paper.

II. Problem Formulation

The problem treated in this paper may be formulated as follows: considering n chromosomes (for n even, and $n \leq 44$)[1], a pairing

Corresponding author: Rodrigo Ventura (yoda@isr.ist.utl.pt)

Affiliation: Institute for Systems and Robotics / Instituto Superior Técnico, 1049-001 Lisbon, Portugal

Partially supported by FCT, under ISR/IST plurianual funding (POSC program, FEDER).

[1] The pairing process can also be applied to any subset of chromosome classes.

978-1-4244-1815-2/08/$25.00 ©2008 IEEE.

assignment $\mathcal{P}$ is defined as a set of ordered pairs (i, j), corresponding to the pairs of homologous chromosomes, satisfying two trivial constraints: (1) $i \neq j$ holds for any pair, and (2) any given index i appears in no more than one pair of the set. A pairing assignment is said to be *total* iff, 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 $\mathcal{P}$ can be written as

$$C(\mathcal{P}) = \sum_{(i,j)\in P} d(i,j) \tag{1}$$

where $d(i, j)$ is the distance function between chromosomes i and j. *The goal of the pairing process is then to find a total pairing $\mathcal{P}$ that minimizes $C(\mathcal{P})$.* To accomplish this task, the problem is divided in two sub-problems: (1) feature extraction, described in the following section, and (2) classification, explained in section II-B.

A. Feature extraction

An image pre-processing procedure is first performed before the feature extraction stage. All chromosome images are histogram equalized in order to minimize the effects of the contrast and bright differences in the classification process. Furthermore, geometric compensation is also performed in order to compensate for geometrical distortions occurring in the *metaphase*. This is done by computing the skeleton medial axis of each chromosome and distort it to make this medial axis straight. For the textural feature extraction a geometrical scaling is also performed by interpolation in order to obtain normalized images with the equal dimensions.

Three types of features are used to compare two chromosomes, i) dimensional, ii) shape and iii) textural. The dimensional features aim to discriminate the dimensions, and correspond to: the axis dimensions of the ellipsis containing the chromosome, the chromosome length proportion (which corresponds to dividing the length of all the chromosomes by the major lenght of all, which ideally would be the lenght of the chromosome of the first class), the border perimeter, the length, and the area. The shape is only discriminated by the normalized area of the chromosome image. The textural features are the banding profile and the *mutual information* (see details in [1]). The banding profile is obtained by integrating along the pixels of each chromosomal image line.

For the pairing process it does not matter the absolute values of the measures computed for each chromosome, but rather their relation within each one of the total of n^2 pairs. Thus, for the dimensional features, the associated pair distance feature is $d(m_a, m_b) = |m_a - m_b|$ where m_a and m_b are the measures extracted from the first and second chromosomes of the pair. For the banding profile the distance feature is $d_{bp}(p_a, p_b) = \min_d \|p_a - p_b(d)\|_2$ where $\|.\|_2$ denotes the Euclidean distance between the profile p_a of the first chromosome and the profile $p_b(d)$ of the second chromosome shifted by d samples. The goal of this procedure is to accurately compare both profiles even if they are misaligned. The *mutual information* is a single scalar that provides a (textural) similarity level measure between every two chromosomes.

The distance between two arbitrary chromosomes is defined by

$$d(a,b) = \min_i \sum_{k=1}^{L} w_i(k) d_k(a,b) \tag{2}$$

where $1 \leq i \leq 22$ is the index of the vector $\mathbf{w}_i = \{w_i(1), w_i(2), ..., w_i(L)\}^T$ and L is the number of measures.

The weight vectors $\mathbf{w}_i$ are obtained in the training step from all pairs of chromosomes for each class i in the training set by minimizing an energy function under the constraint $\|w\| = 1$

$$\mathbf{w}_r = \arg \min_{\mathbf{w}:\|\mathbf{w}\|=1} E(\mathbf{w}) \tag{3}$$

In this paper two energy functions are tested:

$$E_1(\mathbf{w}_i) = \sum_{(a,b)\in V(i)} d(a,b;i) - \sum_{(a,b)\in U(i)} d(a,b;i) \tag{4}$$

$$E_2(\mathbf{w}_i) = \sum_{(a,b)\in V(i)} d^2(a,b;i) - \sum_{(a,b)\in U(i)} d^2(a,b;i) \tag{5}$$

where $d(a,b;i) = \sum_{k=1}^{L} w_i(k) d_k(a,b)$, $V(i)$ is the set of all pairs of chromosomes of the i^{th} class and $U(i)$ is the set of all chromosomes where at most one chromosome in each pair belongs to the i^{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. For details on the minimization of (4) and (5) see [1].

Each *karyogram*, with n somatic chromosomes, gives rise to a $n \times n$ distance matrix, $\mathbf{D} = \{d(i,j)\}$ where $d(i, j)$ is the distance between the i^{th} and j^{th} chromosomes of the *karyogram*. The goal is then to estimate the total pairing $\mathcal{P}$, defined in section II, that minimizes (1). The next sub-section describes the procedure to obtain $\mathcal{P}$.

B. Classification

The pairing process is a computationally hard problem because the optimal pairing must minimize the overall distance, meaning that the solution must correspond to a global minimum of the cost function. This problem can be stated as a combinatorial optimization problem, which can be solved using standard integer programming techniques, since the cost function, as well as the constraints, are linear.

The distances computed using expression (2) form a symmetric matrix of distances $\mathbf{D}$, where each element is the distance between the i-th and the j-th chromosomes, $\mathbf{D}_{ij} = d(i,j)$.

Note that the cost function (1) can be reformulated as a matrix inner product between the distance matrix $\mathbf{D}$ and a pairing matrix $\mathbf{X}$, such that

$$\mathbf{X}_{ij} = \begin{cases} 1 & (i,j) \in \mathcal{P} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Thus, (1) can be re-written as $C(\mathcal{P}) = \mathbf{D} \cdot \mathbf{X}$ where '$\cdot$' denotes the usual matrix inner product:

$$\mathbf{D} \cdot \mathbf{X} = \sum_i \sum_j \mathbf{D}_{ij} \mathbf{X}_{ij} \tag{7}$$

The cost function becomes then linear with the pairing matrix $\mathbf{X}$, which becomes the parameters with respect to which (7) is to be minimized.

The constraints referred above can be easily re-written as linear constraints in the following way: constraint (1) is equivalent to state that the main diagonal of $\mathbf{D}$ is all zeroes, and constraint (2) corresponds to having one and only one entry equal to 1 in each row, as well as in each column. Constraining the domain of the matrix entries to be boolean (*i.e.,* $\mathbf{X}_{ij} \in \{0, 1\}$), the latter is the same to say that

$$\forall_i \sum_j \mathbf{X}_{ij} = 1 \qquad \text{and} \qquad \forall_j \sum_i \mathbf{X}_{ij} = 1 \tag{8}$$

The combinatorial optimization problem can then be restated as a integer programming problem, consisting in

$$
\begin{array}{rl}
\text{minimize} & \mathbf{D} \cdot \mathbf{X} \\
\text{where} & \forall_i \, \forall_j \; \mathbf{X}_{ij} \in \{0, 1\} \\
\text{subject to} & \forall_i \, \forall_j \; \mathbf{X}_{ij} = \mathbf{X}_{ji} \\
 & \forall_i \; \mathbf{X}_{ii} = 0 \\
 & \forall_i \; \sum_j \mathbf{X}_{ij} = 1 \\
 & \forall_j \; \sum_i \mathbf{X}_{ij} = 1
\end{array}
\tag{9}
$$

To solve this integer programming problem, the GNU Linear Programming Kit[2] (GLPK) was employed.

This method has proved significantly faster than other state-space search methods used previously by the authors [1], since those methods did not exploit the linear nature of the problem. All results presented in this paper were obtained after negligible execution times (average: 0.285s; compare with A* average: >257s and with branch-and-bound average: 11.2s).

III. Experimental Results

In this section the results obtained using 3 test sets with growing pairing difficulty are presented. 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 (Fig. 1). Both these test sets were built upon 27 *karyograms*. The third test set was obtained from 19 *karyograms* and contains chromosomes from all of the 22 classes (the sex chromosomes are not considered here). This means that the first testing set consists of $27 \times 4 \times 2 = 216$ chromosomes, the second of 432 chromosomes, and the third one of 836 chromosomes (Tab. I). 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 algorithm. All the chromosomes were correctly oriented in a computer assisted basis by the clinical staff. Only karyograms that present no numerical or structural abnormalities were used. Regarding the quality of metaphase plates, and subsequently of the *karyograms* used in this stage of the work, only the "best" *karyograms* were included in the data sets, i.e., *karyograms* where the chromosomes are not in the highest stage of condensation and bending, and where it is possible for a non-expert to discriminate the band profile, like the ones presented in the Fig. 1

	Data set 1	Data set 2	Data set 3
Tissue of origin	bone marrow cells		
Nr. of chromosome classes	4	8	22
Total nr. of chromosomes	216	432	836

TABLE I

Chromosome data sets used to evaluate the implemented algorithms.

These test sets were used with both classifiers proposed in this paper and the results are listed in tables II, III, and IV. These tables display the number of pairing errors in each experiment where each line corresponds to the possible test of the leave-one-out cross-validation strategy (LOOCV), i.e., where all but one *karyogram* are used for training and the remaining one is used for testing.

Besides the overall characterization of the classifier we are also interested in evaluating the improvement of the algorithm due to the introduction of the MI as a discriminative factor in the distance function. Therefore, tables II, III and IV also display the pairing results with and without MI for comparison purposes. In these tables the symbol $\surd$ is used to indicate that a completely correct pairing was obtained (0 errors).

	Data set 1		*Data set 2*	
	w/out MI	w/ MI	w/out MI	w/ MI
1	$\surd$	$\surd$	2	2
2	$\surd$	$\surd$	$\surd$	$\surd$
3	$\surd$	$\surd$	4	2
4–9	$\surd$	$\surd$	$\surd$	$\surd$
10	$\surd$	$\surd$	$\surd$	2
11	$\surd$	$\surd$	$\surd$	$\surd$
12	$\surd$	$\surd$	2	$\surd$
13–14	$\surd$	$\surd$	$\surd$	$\surd$
15	2	2	$\surd$	$\surd$
16	$\surd$	$\surd$	2	2
17	$\surd$	$\surd$	$\surd$	$\surd$
18	2	2	$\surd$	$\surd$
19	$\surd$	$\surd$	$\surd$	$\surd$
20	2	2	$\surd$	$\surd$
21–25	$\surd$	$\surd$	$\surd$	$\surd$
26	$\surd$	$\surd$	4	2
27	$\surd$	$\surd$	$\surd$	$\surd$
total	6	6	14	10

TABLE II

Simulation results (number of pairing errors in a total of 27 karyograms) using method A and two training/test sets: data set 1 with 4 classes of chromosomes and data set 2 with 8 classes of chromosomes. In both cases, results with and without MI are presented.

In the results presented here, the mean classification rate corresponds to the average of the ratio between the number of correctly paired chromosomes and $n/2$ (the total number of pairs of chromosomes considered).

From table V it is concluded that overall, *method A* performs better then *method B* (the only exception among all tests is the case of the test set 12 with MI where the *method B* performs better than *method A*, Tab. IV), and although not presented here, this behavior is consistent across various experiments performed throughout the conducted work.

It is also concluded that the introduction of the MI in the set of features often leads to an improvement in the pairing results. This can be observed in the case of experiments 3, 12 and 26 in the *Data set 2* with 8 classes (Tab. II) and 2, 13, 14 and 16 in the *Data set 3* with 22 classes for *method A* (Tab. IV). For *method B* this can be observed in the case of experiments 12 in the *Data set 2* (Tab. III) and 2, 18 and 19 in the *Data set 3* (Tab. IV).

The classification time is dependent on the distance matrix $\mathbf{D}$ but in all tests performed here it ranges from few milliseconds (4 classes test sets) up to few tenth of a second for the 22 classes test sets which is a major improvement against much higher execution times of the previously used classification method by the authors [1] that could reach up to minutes when performed on the 22 classes test sets (see II-B).

A 70.10% mean classification rate is observed with the *method A* and *mutual information* for the most realistic *Data Set 3* with the 22 classes of chromosomes (Tab. V). And although below the 85.8% and 98.10% classification rates presented by X. Wu et. al in [5,6] (which can be explained by the very low quality of the chromosome images used in our dataset, as described above) those are very optimistic and promising results.

[2] http://www.gnu.org/software/glpk

	Data set 1		*Data set 2*	
	w/out MI	w/ MI	w/out MI	w/ MI
1	√	√	2	2
2	√	√	√	√
3	√	√	4	4
4–9	√	√	√	√
10	√	√	2	2
11	√	√	√	√
12	√	√	2	√
13–14	√	√	√	√
15	2	2	√	√
16	√	√	2	2
17	√	√	√	√
18	2	2	√	√
19	√	√	√	√
20	2	2	√	√
21–25	√	√	√	√
26	√	√	2	4
27	√	√	√	√
total	6	6	14	14

TABLE III

SIMULATION RESULTS (NUMBER OF PAIRING ERRORS IN A TOTAL OF 27 *karyograms*) USING *method B* AND TWO TRAINING/TEST SETS: *Data set 1* WITH 4 CLASSES OF CHROMOSOMES AND *Data set 2* WITH 8 CLASSES OF CHROMOSOMES. IN BOTH CASES RESULTS WITH AND WITHOUT MI ARE PRESENTED.

IV. CONCLUSIONS

In this paper a pairing algorithm is proposed for pairing purposes in the scope of *karyotyping* process used in *cytogentic* analysis. The proposed algorithm is based on the traditional features extracted from the *karyogram*, such as, dimensions and banding profile and on a new feature, based on the *mutual information* (MI), 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.

New refinements in the algorithms already presented in [1] and a new, better and quicker optimization technique introduced here were crucial to proceed to the next experimental level up to 22 classes of chromosomes. Indeed, tests using 19 *karyograms* and a *leave-one-out cross-validation* strategy allow us to conclude that in the proposed pairing algorithms, working with 22 classes of chromosomes, *method A* has an overall better performance, achieving a 70.10% pairing accuracy when executed together with *mutual information*.

Preliminary qualitative comparison with the results obtained with the Leica$^{\text{TM}}$ CW 4000 Karyo software, using the same data, have shown a relevant and promising improvement. In the near future, detailed comparison with this software and other methods and datasets will be performed, in order to better validate our algorithm.

V. 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. We would also like to thank João Xavier for helping us formulating the classification as a integer programming problem.

	Data set 3			
	method A		*method B*	
	w/out MI	w/ MI	w/out MI	w/ MI
1	13	13	13	13
2	10	8	10	8
3	2	2	2	2
4	2	2	2	2
5	5	5	5	5
6	13	13	13	13
7	√	√	√	√
8	8	8	8	8
9	2	2	2	2
10	8	10	8	10
11	5	5	5	5
12	10	10	9	9
13	7	6	6	6
14	15	13	13	13
15	√	√	√	√
16	4	2	4	4
17	2	2	2	2
18	13	13	14	13
19	11	11	12	11
total	130	125	128	126

TABLE IV

SIMULATION RESULTS (NUMBER OF PAIRING ERRORS IN A TOTAL OF 19 *karyograms*) USING BOTH *method A* AND *method B* AND THE THIRD TRAINING/TEST SET: *Data set 3* WITH 22 CLASSES OF CHROMOSOMES. FOR BOTH METHODS RESULTS WITH AND WITHOUT MI ARE PRESENTED.

	Data Set 1		*Data Set 2*		*Data Set 3*	
method	(w/o)	(w)	(w/o)	(w)	(w/o)	(w)
A	94.4%	94.4%	93.5%	95.4%	68.9%	**70.1%**
B	94.4%	94.4%	93.5%	93.5%	69.4%	69.9%

TABLE V

COMPARISON OF THE MEAN CLASSIFICATION RATES BETWEEN *method A* AND *method B* FOR ALL THE TRAINING/TEST DATA SETS: WITHOUT (*w/o*) AND WITH (*w*) MI).

REFERENCES

[1] 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

[2] J. Piper and E. Granum On Fully Automatic Feature Measurement for Banded Chromosome Classification. *Cytometry J.*, 1989, nr. 10, pp 242-255

[3] J.R. Stanley, M.J. Keller, P. Gader and W.C. Caldwell Data-Driven Homologue Matching for Chromosome Identification *IEEE Transactions on Medical Imaging*, 1998, vol. 17, nr. 3, pp 451-462

[4] B. Lerner Toward a completely automatic neural-network-based human chromosome analysis *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 1998, vol. 28, nr. 4, pp 544-552

[5] X. Wu, P. Biyani, S. Dumitrescu and Q. Wu Globally Optimal Classification and Pairing of Human Chromosomes *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, 2004, pp 2789-2792

[6] P. Biyani, X. Wu and A. Sinha Joint Classification and Pairing of Human Chromosomes *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2005, vol. 2, nr. 2, pp 102-109

[7] P.P.S. Karvelis, A.A.T. Tzallas, D.D.I. Fotiadis and I.I. Georgiou A Multichannel Watershed-Based Segmentation Method for Multispectral Chromosome Classification *IEEE Transactions on Medical Imaging*, Accepted for publication in a future issue