32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Classifier-assisted metric for chromosome pairing

Rodrigo Ventura*, Artem Khmelinskii and J. Miguel Sanches

Abstract— ***Cytogenetics* plays a central role in the detection of chromosomal abnormalities and in the diagnosis of genetic diseases. A *karyogram* is an image representation of human chromosomes arranged in order of decreasing size and paired in 23 classes. In this paper we propose an approach to automatically pair the chromosomes into a *karyogram*, using the information obtained in a rough SVM-based classification step, to help the pairing process mainly based on similarity metrics between the chromosomes. Using a set of geometric and band pattern features extracted from the chromosome images, the algorithm is formulated on a Bayesian framework, combining the similarity metric with the results from the classifier. The solution is obtained solving a mixed integer program. Two datasets with contrasting quality levels and 836 chromosomes each were used to test and validate the algorithm. Relevant improvements with respect to the algorithm described by the authors in [1] were obtained with average paring rates above 92%, close to the rates obtained by human operators.**

Index Terms— **Chromosome, Pairing, Classification, Mixed Integer Programming, Image processing, Optical Microscopy, Support Vector Machines**

I. Introduction

Karyotyping is a set of procedures, in the scope of the *cytogenetics*, that produces a visual representation of the 46 chromosomes, arranged in decreasing order of size and paired in 22 classes of homologous elements plus two sex-determinative chromosomes. This sorting and pairing process of chromosomes, extracted from the *metaphase plate*, is difficult, time consuming and most of the times performed manually. An automatic procedure is still needed.

A significant number of approaches have been proposed and used in the design of classifiers, *e.g.:* neural network and multilayer perceptron [2], [3], [4], [5], Bayes [6], hidden Markov models (HMM) [7], template matching [8], wavelet [9] and fuzzy [8].

However, when the quality of the images is very poor, which is the case of the chromosomes extracted from *bone marrow* cells used in the diagnosis of leukemia, or the geometrical distortions are too severe, the available classification strategies do not work properly and the classification rates obtained with automatic classifiers, typically in the range of 70%-80%, are still far from the performance reached by the human operator, typically with an approximate classification rate of 99.70% [4]. To overcome those issues, in [1] the authors have proposed a different method where the pairing is performed without classification and the pairing criteria are similarity measures computed over all possible pair-combinations of chromosomes. Since the application of the *G-banding* [10] procedure to the chromosomes generates a distinct transverse banding pattern characteristic of each class, that pattern is the utmost important feature for chromosome classification and pairing. To fully utilize the band profile of each chromosome, the mutual information feature was used together with other geometrical and band pattern features extracted from the chromosome images from a given *metaphase plate*.

In this paper, one step further is proposed to improve the accuracy of the algorithm described in [1]. A rough classification, performed with a *support vector machine* (SVM) classifier [11] is used to help the pairing procedure. The result of this classification is then combined with the similarity measures presented in [1], using a Bayesian framework. This results in a mixed integer program (MIP) that can easily be solved using a standard MIP solver.

The contributions of this paper are threefold:

- The chromosome pairing method presented in [1], based on an energy minimization principle is recast as a maximum likelihood problem, thus paving the way to the extension here presented
- A novel, two step classifier-assisted metric for automatic chromosome paring is presented, using a Bayesian framework
- The presented method further improves ($\approx$ 16 percentage points) the approach first introduced in [1] (where the best mean classification rate (MCR) obtained was 76.10%) achieving a MCR of 92.8%

II. Chromosome data

To test and validate the proposed algorithm two chromosome datasets with different quality levels were used: Grisan *et al.* [12] and Lisbon-K1 chromosome dataset (*LK1*) [1], [13]. The difference in quality is related to the centromere position, band profile description/discrimination and level of condensation of the chromosome. The first one is of the same nature and "high" quality as the classic Philadelphia, Edinburgh and Copenhagen datasets [14], [3], [15] because the images are based on cells extracted from the amniotic fluid and choroidal villi (pre-natal *cytogenetics*). *LK1* is a dataset of "low" quality since it is based on *bone marrow* cells used for *leukemia* diagnosis. It presents much less quality than the former ones, used in the traditional *cytogenetics*. From each dataset, 19 *karyograms*, each one composed by

*Corresponding author: rodrigo.ventura@isr.ist.utl.pt

This work was supported by the Portuguese government - FCT (ISR/IST plurianual funding)

Rodrigo Ventura, Artem Khmelinskii `aihmel@gmail.com` and J. Miguel Sanches `jmrs@ist.utl.pt` are with the Institute for Systems and Robotics at the Instituto Superior Técnico, Lisbon Portugal. Artem is also with the Division of Image Processing, Department of Radiology, LUMC, Leiden, the Netherlands.

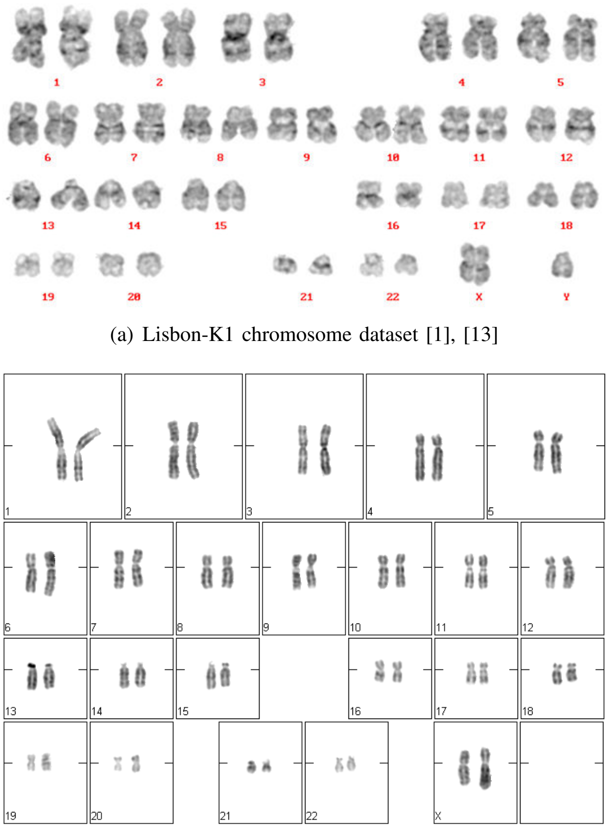
978-1-4244-4124-2/10/$25.00 ©2010 IEEE

$2N$ chromosomes images where $N = 22$ is the number of homologous pairs/classes, are used. In total, each test set consists of 836 chromosomes (Table I). All chromosomes were manually segmented, correctly oriented, ordered and annotated by the clinical staff to be used as ground truth data. The sex chromosomes were put aside and only *karyograms* that present no numerical or structural abnormalities were used at this stage of the work. Figure 1 presents a *karyogram* example for each dataset.

Dataset	Lisbon-K1 [1], [13]	Grisan *et al.* [12]
Tissue	bone marrow	amniotic fluid, choroidal villi
# Classes	22	22
# Chromosomes	836	836

TABLE I

CHROMOSOME DATASETS USED TO TEST AND VALIDATE THE PROPOSED ALGORITHM: LK_1 [1], [13] AND GRISAN *et al.* [12]



(a) Lisbon-K1 chromosome dataset [1], [13]

(b) Grisan *et al.* chromosome dataset [12]

Fig. 1. *Karyogram* examples for 2 chromosome datasets with different levels of quality: Lisbon-K1 [1], [13]-“low” & Grisan *et al.* [12]-“high”

III. Problem Formulation

This paper addresses the problem of automating the chromosome pairing procedure, by extracting a set of features from each chromosome found in the *karyogram* image. Given a feature vector extracted from each individual chromosome, the *pairing problem* consists in identifying the pairs of homologous chromosomes. We consider here 44 chromosomes, forming 22 pairs.

The *classification problem* identifies the correct type of each one of the 44 chromosome from a set of 22 classes. Each class contains homologous chromosomes that are supposed to be grouped. In a normal karyotype it is expected to exist exactly two chromosomes for each class.

At first sight, if all chromosomes were correctly classified into one of the 22 possible types, the pairing result would follow trivially, since there would be exactly two chromosomes classified in each type. However, automatic classification of chromosomes is still an open problem, and thus, errors in classification prevent a correct pairing. This is the motivation behind looking to the pairing of chromosomes as a separate problem, since their exact classification into classes is of no importance, as far as pairing is concerned.

IV. Classification and Pairing

In [1], a chromosome pairing method was proposed based on a matrix $\mathbf{D}$ of distances between chromosomes of a karyogram. The entries d_{ij} of this matrix are computed from a metric function that yields a distance measure between any given pair of chromosomes i and j (for $i \neq j$). Each distance is obtained by minimizing a weighted sum of features w.r.t. a set of weight vectors $\{\mathbf{w}_r, r = 1, ..., 22\}$

$$d_{ij} = \min_{r \in \{1,\ldots,22\}} f(i,j;\mathbf{w}_r), \tag{1}$$

$$\text{where} \quad f(i,j;\mathbf{w}) = \sum_{k=1}^{L} w(k) d_k(i,j) \tag{2}$$

The weight vectors are obtained by minimizing an energy function over a training set of manually paired chromosomes (ground truth). While most of these distances $d_k(i,j)$ are euclidean distances between features of individual chromosomes (area, perimeter, normalized area, bounding box dimensions, chromosome length proportion, band profile), one of them, the mutual information, is a pairwise feature [1].

Given the distance matrix $\mathbf{D}$, the resulting pairing is obtained by solving an integer programming problem [1], using a standard solver. The solution corresponds to the pairing that minimizes the sum of the distances of each one of the pairs found

$$C(\mathcal{P}) = \sum_{(i,j) \in \mathcal{P}} d_{ij}, \tag{3}$$

where $\mathcal{P}$ is the set of chromosome pairs.

This approach can be re-formulated using a probabilistic framework in the following way: any valid pairing solution can be represented by a binary symmetric matrix, denoted $\mathbf{X}$, with entries x_{ij} where $x_{ij} = x_{ji} = 1$ if chromosome i pairs with j, and $x_{ij} = x_{ji} = 0$ otherwise. Each binary matrix $\mathbf{X}$ represents a valid pairing if and only if the following (linear) constraints are satisfied:

$$\begin{array}{l} \mathbf{X} = \mathbf{X}^T \\ \forall_i \; x_{ii} = 0 \\ \forall_i \; \sum_j x_{ij} = 1 \end{array} \tag{4}$$

because a chromosome cannot be paired with itself, and each chromosome (line in matrix $\mathbf{D}$) can only be paired with other single chromosome. The optimal pairing configuration solution is the maximum likelihood estimation problem formulated as

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} P(\mathbf{D}|\mathbf{X}), \tag{5}$$

where the range of variation of $\mathbf{X}$ is the space of binary matrices satisfying (4). Assuming conditional independence of each distance d_{ij} given a pairing $\mathbf{X}$, the distribution $P(\mathbf{D}|\mathbf{X})$ can be expanded into the product $\prod_{(i,j)} P(d_{ij}|\mathbf{X})$. This product can be factored into two sets, depending on the value 0 or 1 of the corresponding x_{ij}. Here we disregard the terms for $x_{ij} = 0$, considering that the useful information for maximizing $P(\mathbf{D}|\mathbf{X})$ comes mostly from the distances for which $x_{ij} = 1$. Thus,

$$P(\mathbf{D}|\mathbf{X}) = \eta \prod_{(i,j)|x_{ij}=1} P(d_{ij}|x_{ij} = 1), \tag{6}$$

where η is a normalizing constant. Adopting an exponential distribution for the distance for each pair of chromosomes

$$P(d_{ij}|x_{ij} = 1) = \lambda e^{-\lambda d_{ij}}, \quad \text{for } d_{ij} \geq 0 \tag{7}$$

parametrized by λ and taking the negative logarithm of expression (6) the following maximum likelihood estimator is obtained

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \sum_{(i,j)|x_{ij}=1} d_{ij}. \tag{8}$$

This expression provides a formal justification to the pairing method based on the minimization of the pair distances published in [1].

Consider now that an individual classification algorithm obtained a tentative classification, expressed as a 44×22 matrix $\hat{\mathbf{C}}$, where each entry $\hat{c}_{ik}$ is one if chromosome i was classified in class $k \in \{1 \ldots, 22\}$ and zero otherwise. Thus, the following maximum a posteriori estimator can be written:

$$\begin{aligned} \hat{\mathbf{X}} &= \arg\max_{\mathbf{X}} P(\mathbf{X}|\mathbf{D}, \hat{\mathbf{C}}) \\ &= \arg\max_{\mathbf{X}} P(\mathbf{D}|\mathbf{X}, \hat{\mathbf{C}}) P(\mathbf{X}|\hat{\mathbf{C}}) / P(\mathbf{D}|\hat{\mathbf{C}}) \\ &= \arg\max_{\mathbf{X}} P(\mathbf{D}|\mathbf{X}) P(\mathbf{X}|\hat{\mathbf{C}}), \end{aligned} \tag{9}$$

where it was considered that the distances matrix distribution given the true pairing does not depend on the classification results, *i.e.,* $P(\mathbf{D}|\mathbf{X}, \hat{\mathbf{C}}) = P(\mathbf{D}|\mathbf{X})$. That is to say that, for each combination of two chromosomes, its distribution only depends on whether they form a pair or not. Now, note that the $P(\mathbf{D}|\mathbf{X})$ is the same as the one in (5), and therefore it can be computed using (6). The term $P(\mathbf{X}|\hat{\mathbf{C}})$, which can be seen as a prior on the distribution of $\mathbf{X}$ after classification, is here estimated in the following way: first, $P(\mathbf{X}|\hat{\mathbf{C}})$ is factorized, assuming conditional independence for each pair proposed in $\mathbf{X}$ (*i.e.,* the (i,j) pairs such that $x_{ij} = 1$)

$$P(\mathbf{X}|\hat{\mathbf{C}}) = \prod_{(i,j)\,|\,x_{ij}=1} P(x_{ij} = 1\,|\,\hat{\mathbf{C}}), \tag{10}$$

where $P(x_{ij} = 1\,|\,\hat{\mathbf{C}})$ is the probability of chromosomes i and j forming a pair, given a classification output $\hat{\mathbf{C}}$. Taking the negative log probability of the $\arg\max$ argument in (9) and assuming an exponential distribution (7), the final estimator is derived:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \sum_{(i,j)\,|\,x_{ij}=1} \left(d_{ij} - l_{ij|\hat{\mathbf{C}}} \right), \tag{11}$$

where $l_{ij|\hat{\mathbf{C}}} = \frac{1}{\lambda} \log P(x_{ij} = 1\,|\,\hat{\mathbf{C}})$. Note that the summations in (11) have a fixed amount of terms and the result is invariant to summing a constant to $l_{ij|\hat{\mathbf{C}}}$. Thus, for the $l_{ij|\hat{\mathbf{C}}}$ this simple approach is considered:

$$l_{ij|\hat{\mathbf{C}}} = \begin{cases} L & \text{if } \hat{c}_{ik} = \hat{c}_{jk} \neq \hat{c}_{mk} \text{ for } i \neq m \neq j, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

meaning that a constant non-zero value is used whenever two *and only two* chromosomes are classified in the same class.

Dataset	SVM	LCD	CaLCD	p-value
Lisbon-K1	58.6	41.4	**70.8**	< 0.008
Grisan *et al.*	67.9	76.1	**92.8**	$< 10^{-15}$

TABLE II

COMPARATIVE RESULTS OF THE PAIRING METHODS EVALUATED ($L = 0.17$), EXPRESSED IN TERMS OF THE MEAN PAIRING RATES (AVERAGE OF THE RATIO BETWEEN THE NUMBER OF CORRECTLY PAIRED CHROMOSOMES AND THE TOTAL NUMBER OF PAIRS OF CHROMOSOMES CONSIDERED (22), IN PERCENTAGE). THE LAST COLUMN CONTAINS THE ONE-SIDED P-VALUE OF STATISTICAL SIGNIFICANCE

V. Experimental Results

Experiments[1] comprised a comparative study of the following pairing methods:

- **Linear Combination of Distances (LCD)**, the pairing method based on distance matrices, previously described in [1], using estimator (8);
- **Support Vector Machines (SVM)**, the pairing method based on the individual classification of features using the SVM classifier alone;
- **Classification-assisted LCD (CaLCD)**, the pairing method proposed in this paper, that uses the estimator (11).

Table II summarises the results obtained for the two datasets described in section II. The results are expressed in terms of percentage of correctly identified pairs, using a leave-one-out cross validation (LOOCV) approach[2]. The results were obtained by setting the L parameter to 0.17 (manually selected by trial-and-error, at this stage of the work). The statistical significance concerns the rejection of the null hypothesis that the amount of errors of the proposed CaLCD method is greater or equal than the one of LCD, for the same

[1] The MIP solver used was the GNU Linear Programming Kit (GLPK), and the SVM classifier was the LIBSVM.

[2] For each karyogram in the dataset, the test set includes that karyogram, while the training set includes all of the others.

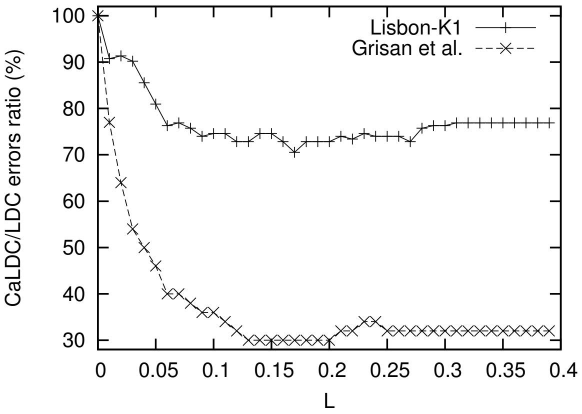


Fig. 2. Sensitivity of the error ratio with respect to the parameter L

test karyogram. From these results one can conclude that the new method CaLCD here proposed effectively improves the pairing performance, when compared with the previously presented LCD.

Note that when $L = 0$, the pairing results are insensitive to the prior classification, the estimators (8) and (11) become equivalent. Figure 2 shows the obtained results in terms of the error ratio for each one of the datasets considered, in function of L. This error ratio was computed by dividing the number of incorrect pairs with CaLCD by the one with LCD. It expresses the fraction of the error obtained by the proposed method, with respect to the LCD pairing method. From these results we observe that all plots show a roughly convex behaviour, with a minimum for L in the vicinity of 0.17. This suggests the existence of an optimal value for L, maximizing pairing performance. The difference in the classification rates between the *LK1* and Grisan *et al.* datasets is due to the large quality difference between both datasets described in section II.

VI. Conclusions and future work

This paper presents a new chromosome pairing method, consisting of two steps: a *classification step*, using a SVM classifier with the goal of identifying chromosome pairs, and a *pairing step,* based on distance measures between chromosomes, which combines information from the classifier output with the pairwise features. The classifier output is only partially used, *i.e.,* restricted to the cases where two and only two chromosomes are classified into the same class. Experimental results using two datasets from different origins were performed, and the results show a significative improvement of this method over the (previously published) pairing step alone, and over a pure classification-based approach with a maximum average paring rate of 92.8%. The influence of the L parameter, specifying the degree of influence of the classification step on the pairwise chromosome distances, was also evaluated. Empirical evidence points towards a convex behaviour of the pairing performance with respect to this parameter.

For future work, we intend to further explore this combined classification and pairing approach. Outstanding questions include: (1) what is the theoretical reason behind the observed improvement, since both classification and pairing make use of essentially the same features extracted from individual chromosomes; (2) if partial classification results improves pairing, can partial pairing results improve classification? Finally, the issue of selecting the L parameter has to be addressed. We believe that the answer to question (1) above will shed some light on this problem.

VII. Acknowledgments

The authors thank Prof. Maria do Carmo Fonseca - Institute of Molecular Medicine (IMM) in Lisbon and the clinical staff of the Cytogenetics/Virology Laboratory of the IMM - GenoMed of Lisbon (namely Sónia Santos) for providing the *karyograms* used to build the *LK1* chromosome dataset.

References

[1] A. Khmelinskii, R. Ventura, and J. Sanches, “A novel metric for bone marrow cells chromosome pairing,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 6, pp. 1420–1429, 2010.

[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] M. Z. 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Swansbury, *Cancer Cytogenetics: Methods and Protocols (Methods in Molecular Biology)*, Humana Press, 2003.

[11] Corinna Cortes and Vladimir Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] E. Grisan, E. Poletti, and A. Ruggeri, “Automatic segmentation and disentangling of chromosome in q-band prometaphase images,” *IEEE TITB*, vol. 13, no. 4, pp. 575–581, 2009.

[13] “Lisbon K1 - Chromosome Dataset,” http://mediawiki.isr.ist.utl.pt/wiki/Lisbon-K_Chromosome_Dataset, March 31, 2010.

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

[15] 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.