journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A modular framework for the automatic classification of chromosomes in Q-band images☆

*Enea Poletti, Enrico Grisan, Alfredo Ruggeri**

Department of Information Engineering, University of Padova, Via Gradenigo 6/a, 35131 Padova, Italy

ARTICLE INFO

Article history:
Received 28 January 2011
Received in revised form
11 May 2011
Accepted 24 July 2011

Keywords:
Alternating decision trees
Artificial neural networks
Automatic karyotyping
Chromosome classification
Chromosome polarization
Image classification

ABSTRACT

The manual analysis of the karyogram is a complex and time-consuming operation, as it requires meticulous attention to details and well-trained personnel. Routine Q-band laboratory images show chromosomes that are randomly rotated, blurred or corrupted by overlapping and dye stains. We address here the problem of robust automatic classification, which is still an open issue. The proposed method starts with an improved estimation of the chromosome medial axis, along which an established set of features is then extracted. The following novel polarization stage estimates the chromosome orientation and makes this feature set independent on the reading direction along the axis. Feature rescaling and normalizing techniques take full advantage of the results of the polarization step, reducing the intra-class and increasing the inter-class variances. After a standard neural network based classification, a novel class reassignment algorithm is employed to maximize the probability of correct classification, by exploiting the constrained composition of the human karyotype.

An average 94% of correct classification was achieved by the proposed method on 5474 chromosomes, whose images were acquired during laboratory routine and comprise karyotypes belonging to slightly different prometaphase stages. In order to provide the scientific community with a public dataset, all the data we used are publicly available for download.

1. Introduction

Chromosome karyotype analysis [1] is an important screening procedure routinely performed in clinical cytogenetic labs with the aim of early diagnosing cancers and genetic aberrations. Chromosomes are first stained with a fluorescent dye, and then imaged through a microscope for subsequent analysis and classification. Each chromosome in the image has to be identified and assigned to one of 24 classes. The result of this procedure is the so-called karyotype image, in which all chromosomes in a cell are graphically arranged according to an international system for cytogenetic nomenclature (ISCN) [1] classification. Fig. 1 shows a typical PAL resolution (768 × 576 pixels, 8 bits/pixel) Q-banding prometaphase image and its karyotype.

The appearance of chromosomes depends on the stage of the cell division cycle at which they are imaged. For much of the cell cycle (interphase), individual chromosomes cannot be distinguished. They appear as distinct bodies only towards the end of the cycle, at prophase, when they are long string-like objects, contracting and separating at metaphase, just before cell division. Prometaphase is the intermediate stage of contraction between prophase and metaphase. These different

☆ Preliminary results were presented at the 30th IEEE EMBS International Conference, Vancouver (British Columbia, Canada) August 2008. This work was partially supported by TesiImaging, Milan, Italy.

* *Corresponding author*. Tel.: +39 049 827 7624; fax: +39 049 827 7699.
E-mail addresses: enea.poletti@dei.unipd.it (E. Poletti), enrico.grisan@dei.unipd.it (E. Grisan), alfredo.ruggeri@unipd.it, alfredo.ruggeri@dei.unipd.it (A. Ruggeri).

0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.07.013

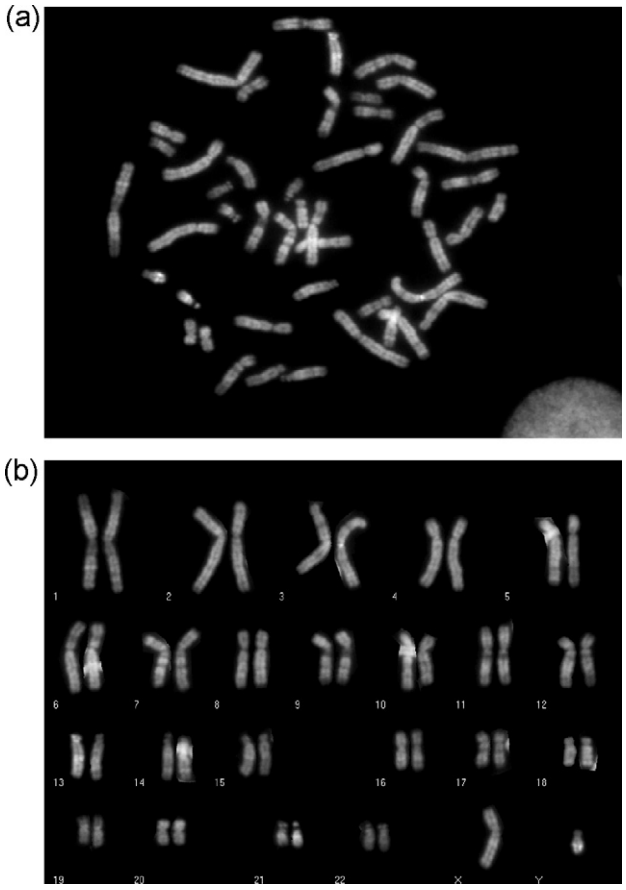

Fig. 1 – (a) Typical Q-band prometaphase image acquired with PAL resolution and (b) the corresponding manual karyotype.

stages in the division cycle are characterized by the number of bands visible in the chromosome body, and by the elongation of chromosomes: the most elongated ones, with the highest resolution in the structure, are also those that present the greatest difficulty in chromosome analysis, due to the fact that longer chromosomes touch and overlap one another much more frequently than shorter ones.

The conventional framework for an automatic karyotyping system consists of a number of different steps: image enhancement, segmentation, features extraction, and classification. Different solutions were also proposed, e.g., with a classification driven segmentation [2], in which both tasks are simultaneously carried out through an iterative scheme. Nevertheless, segmentation and classification of chromosomes are generally assumed to be separate tasks. Following this approach, this work focuses on the classification, assuming the segmentation was already carried out, e.g., by the algorithm proposed in [3]. The aim is to assign each chromosome to one of the 24 possible classes, by exploiting chromosome features extracted from the image.

Several chromosome features have been taken into consideration in the literature, e.g., the ones directly computed from the chromosome outline: area, relative density, perimeter of the convex hull [4]. Two more features are of particular importance: the density profile and the contour function. The former is a representation of the banding pattern of each chromosome class, whereas the latter measures the way the distance between the chromosome contour and its axis varies. These two features can be obtained only after the axis of the chromosome has been estimated [5]. Most of the axis extraction methods [2,6,7] are based on skeletonization, which has the drawback of producing spurious branches, which are difficult to correctly prune, particularly at the chromosome tips.

The dark and thin region that connects the two sister chromatids of a chromosome is called centromere and it is used by cytologists to align chromosomes belonging to the same class [1]. The two arms connected by the centromere are labeled p (the shorter of the two) and q (the longer): the p and q length ratio is called centromeric index (CI) and has been reported to be one of the most discriminant features for karyotyping [2,6,8,9]. Usually the CI is computed by searching the thinnest and/or darkest region of the chromosome. Both the analysis of the vertical and horizontal projection vectors of the chromosome binary image [9] and empirical rules based on morphological chromosome features and grayscale information have been used to estimate the position of the centromere [10]. Their main drawback is the impossibility to deal with bent chromosomes, which show a distorted centromere region, and with chromosomes that have an overlap or a touch which corrupt banding pattern and contour function close to the centromere.

The asymmetry of chromosome shape and banding with respect to the centromere makes the ordering of the features extracted from the axis depending on its reading direction (polarization), but only some methods deal with this problem [6,10]. They both propose empirical rules that aim to estimate the position of the centromere in order to correctly polarize the chromosomes. Other methods either assume that the polarization is given [11–13], or disregard the issue, accepting to have the directional features of the same chromosome randomly aligned, to the detriment of the classification accuracy [7,14].

Although additional features have been proposed, such as the variants-based profile extraction [15], the wavelet and Fourier descriptors derived from the density profile [16], or the wave packet transform [17], the most discriminating features for classification appear to be length, CI, density profile, and contour function [6–8].

As regards classification itself, the most common method used in the literature is the multi-layer perceptron trained with a back-propagation algorithm [2,7,11,14]. More recently, a two-stage classification platform was proposed in [18], where a genetic algorithm was implemented to optimize the topology of several ANNs. In the first layer of the scheme, a single ANN was employed to classify the 24 chromosome classes into seven groups with similar image characteristics, while in the second layer seven different ANNs were adaptively optimized for each of the seven groups to identify individual chromosomes. Other classifiers have been applied, like the Inferred Markov Network Models [12], maximum likelihood [6], a similarity profile classifier performing a subsequence matching [19], or dynamic time warping [13].

Even if the classification of chromosomes is a constrained problem, since a normal karyotype consist of 22 pairs of autosomal chromosomes and one pair of sex chromosomes, most