Profile and feature extraction from chromosomes

Gunter Ritter and Gernot Schreib
Universität Passau
Fakultät für Mathematik und Informatik
94 030 Passau
Germany

Abstract

The most accurate methods for automatic classification of chromosomes under a light microscope today extract numerical features from band-pattern profiles along their longitudinal axes. The construction of a reliable axis is a crucial step in this process. We propose a new way based on the dominant points of the contour and cubic splines. The dominant points serve as candidates for the tips of the chromosome or its chromatids. Ambiguities are dissolved by the recently proposed method of variants for object identification. A Voronoi diagram decomposes the chromosome in slices for profile extraction. The method improves the currently best classification results significantly yielding a test-set error rate of 0.6% applied to a data set of the band level 200.

1. Introduction

Classification of the chromosomes of a eukaryotic cell under a light microscope in their biological classes is a clear–cut task that lends itself to automation on computers. In the last decade, feature–oriented classification methods in combination with the Bayesian paradigm have turned out to attain error rates approaching those of the human expert, cf. [2, 8, 7, 10]. This progress is mainly due to the application of three principles: (α) exploitation of prior knowledge, (β) precise data modeling including proper outlier handling and robust parameter estimation, (γ) proper application of Marr's [4] "Principle of Least Commitment."

Main steps to feature extraction are (i) identification of the, usually *oblong*, *shape* of the chromosome; (ii) representation of *band pattern* and *shape* by so-called *profiles*, i.e., univariate functions along the axis; (iii) extraction of *numeric features* by applying methods from signal processing (Fourier or other coefficients) to the univariate profiles.

Unlike methods of local band and shape description [1], we do not use *positions* of special objects on the chromosome, such as its darkest band or its centromere, as features. The detection of these objects is often unsafe; a serious outlier usually results if the position of a different object is erroneously used for a measurement. In particular, we do not try to explicitly detect the position of the centromere for determining polarity. It is, however, implicitly contained in the shape profile, cf. Sect. 3.2.

A prerequisite for feature extraction is recognition of the shape of the chromosome. It is now classical [3] to use dominant points and longitudinal axes. But, contrary to more traditional methods, we avoid taking an early decision on the shape unless this decision is safe. We rather resort to the recently proposed method of variants that lays the grounds for an efficient Bayesian classifier, the Simple Constrained Classifier–Selector, cf. Sect. 4 and [9].

2. Longitudinal axes

2.1. Dominant points

At first on our way to profiles and features we try to recognize the oblong shape of the chromosome. To do this, we use its tips which we estimate in four steps; these are

- extraction of the object boundary and the associated *contour* by a standard contour-following algorithm,
- estimation of the *contour curvature* by local deflection angles and triangular smoothing,
- computation of the *"essential" maxima* of the contour curvature, and
- determination of *dominant points* related to the essential maxima.

Intuitively, an *essential maximum* [10] of a function $f\colon \mathbb{Z} \to \mathbb{R}$ is the highest peak lying between two sufficiently deep valleys. By duality, an *essential minimum* is

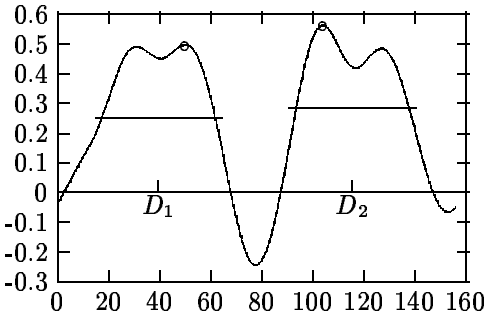

Figure 1. Essential maxima and their dominant points D_1 and D_2 determined as the centers of the two associated mountains.

an essential maximum of $-f$. In general, an essential maximum of the contour curvature is not yet a precise estimate of a tip of the chromosome. This may be located at an arbitrary point of its associated mountain and we rather propose the center of this mountain as a dominant point, cf. Fig. 1.

2.2. Axes

By the *longitudinal axis* of a chromosome we mean a continuous curve connecting the two tips of the chromosome thereby dividing it into its two chromatids, the essentially congruent longitudinal halves.

One might expect that the method of Sect. 2.1 yields mostly two dominant points which may serve as the tips. But there are various reasons for a number different from two. One reason is biological, other sources are artefacts such as bent or circular chromosomes and overlappings. For constructing a reliable longitudinal axis it is first necessary to understand why and when the various cases of one, two, three,... dominant points arise. Biologists divide mitosis (the process of nuclear division) into four stages called *prophase*, *metaphase*, *anaphase*, and *telophase*. At the beginning of the prophase the chromosomes become distinct for the first time. They get progressively shorter through a process of contraction or condensation. Next, in the metaphase, the pairs of sister chromatids begin to separate. The anaphase begins, when the chromatids are separated. The last stage, the telophase, finishes the process of nuclear division.

For analysis, prophases and metaphases, only, are used. The reason is that the band pattern blurs when chromosomes get shorter and shorter. It is not surprising that chromosomes of the same cell may be in different phases at the same time. In fact, modern preparation techniques try to synchronize the phases of the division process. But application of these methods is restricted and the classifier should be able to handle prophases and metaphases simultaneously.

Let $(D_1, \ldots, D_k)$ be the increasingly ordered sequence of dominant points constructed in Sect. 2.1. Our method normally results in two, three or four dominant points. If $k = 2$ then we construct a longitudinal axis in the following way. We run from D_1 to D_2 along both sides of the chromosome at constant speed so as to arrive at the destination D_2 at the same time. We mark the Euclidean midpoints of the contour points simultaneously reached. A subsequence of them serves as *supporting points* for transition to a plane cubic spline which we extend tangentially and linearly beyond both ends. This is the (extended) longitudinal axis required.

The cases $k = 3$ and $k = 4$ need a more detailed discussion. The two chromatids of the originally elongated chromosome begin to split in the late metaphase. Y-shaped acrocentric and X-shaped metacentric chromosomes with three and four essential maxima and dominant points, respectively, are characteristic at this stage, cf. Figs. 2(b),(e) and 3(b),(c),(d). In the case of *three* dominant points, D_1, D_2, D_3, we consider six longitudinal axes. The three "point-to-midpoint" axes $(D_1, (D_2, D_3))$, $(D_2, (D_1, D_3))$, and $(D_3, (D_1, D_2))$ connect a dominant point with the Euclidean midpoint of the opposite edge. One of these axes is correct if the chromosome is Y-shaped. Unfortunately, other sources of maxima of the contour curvature are artefacts such as bent chromosomes, cf. Fig. 2(a). Therefore, "point-to-point" axes corresponding to the pairs (D_1, D_2), (D_1, D_3), and (D_2, D_3) have to be considered as alternative axes.

Supporting points of a point-to-point axis are computed as in the case $k = 2$. A point-to-midpoint axis $(D_1, (D_2, D_3))$ is computed by simultaneously running through the paths $D_1, \ldots, D_2$ and $D_1, \ldots, D_3$ for constructing the supporting points.

The presence of *four* dominant points D_1, D_2, D_3, D_4 indicates an X-shaped chromosome; we suggest to use the "midpoint-to-midpoint" axes $((D_1, D_2), (D_3, D_4))$ and $((D_2, D_3), (D_4, D_1))$ which connect midpoints of opposite edges. As in the previous case, artefacts may create dominant points and we may again consider the point-to-point axes determined by the opposite points (D_1, D_3) and (D_2, D_4) and eight point-to-midpoint connections of a vertex to a non-incident edge $(D_1, (D_2, D_3))$, $(D_1, (D_3, D_4))$, ... We use the paths $D_1, \ldots, D_2$ and $D_4, \ldots, D_3$ for constructing the supporting points of a midpoint-to-midpoint axis $((D_1, D_2), (D_3, D_4))$.

In some cases it is sufficient to use a subset of the axes suggested above. In the case of four dominant points, e.g., the two midpoint-to-midpoint axes often suffice. Although visual inspection of chromosome images shows that the longest diagonal or a point-to-midpoint axis is sometimes correct, cf. Fig. 3(e),(f), admitting these axes does not significantly improve classification error rates. This means that these cases occur only rarely.

2.3. Rules

Sometimes, an early rule-based reduction of the number of dominant points or axes is possible. As a rule for deleting *dominant points*, one may determine local deflection angles at each of the dominant points. In the case $k = 4$ we have an indication that the chromosome is in the late metaphase if none is far away from $90°$, cf. Fig. 3(c),(d). In the opposite case there is a large and a small angle; we may then delete the dominant point belonging to the largest angle and send this chromosome to the case $k = 3$, cf. Fig. 3(a).

Concerning the choice of *axes*, the crux is to decide whether three or more dominant points are caused by a late metaphase or by artefacts. Assuming a rectangular shape, a crude estimate of the mean width of the chromosome, based on area A and boundary length N, is $\bar{w} = (N - \sqrt{N^2 - 16A})/4$. The mean width of a chromosome in a cell is almost independent of its class. Hence, the arithmetic mean of these widths across a cell is a stable estimate of $\bar{w}$. If an edge is short and its length is of the order of the mean width its midpoint is likely to lie on the correct axis, cf. Figs. 2(c),(d) and 3(c).

We finally try to detect point-to-point axes in the case $k = 3$. A *bent* chromosome, cf. Fig. 2(a), often meets the requirements of this case. It is indicated by the existence of a deep essential minimum which suggests to delete the opposite dominant point. But some caution is in order. If the remaining two dominant points define a short edge then the chromosome might be bent and in the process of division, cf. Fig. 2(c). Otherwise, it could also be an acrocentric chromosome in late metaphase, cf. Fig. 2(b). In both cases, deletion would be wrong. Let us point out that the choice of these rules depends on the specific preparations and cell types (amnion, chorion, blood, and bone marrow). Of course, the more wrong cases the rules can exclude the better the classification results will be, provided the correct axis is still there.

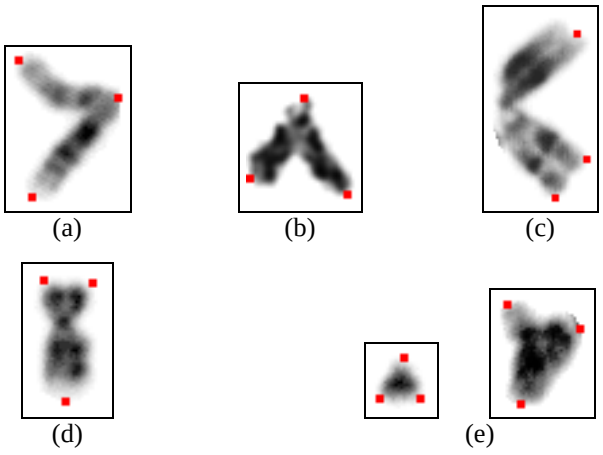

Figure 2. Three dominant points. (a) Bent chromosome; (b) acrocentric chromosome in late metaphase; (c) bent chromosome in late metaphase; (d) triangle with a short edge; (e) almost equilateral triangles

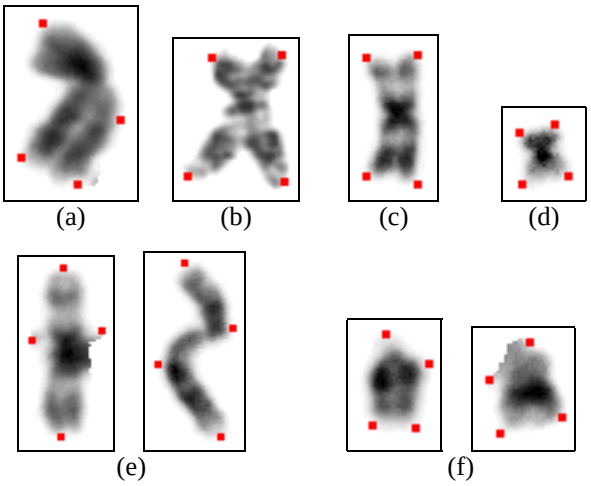

Figure 3. Four dominant points. (a) Bent chromosome; (b) X-shaped chromosome; (c) rectangular chromosome; (d) almost quadratic chromosome; (e) point-to-point axis correct; (f) acro- and metacentric chromosomes with similar shapes

As stated above, our method normally results in two, three or four dominant points. A different number may be created by strange artefacts such as overlappings and occurs only rarely. In the case of one dominant point, we use the opposite point as the most prospective complement. Finally, in the case of five or more, we remove all points but the ones corresponding to the four largest essential maxima and proceed with the case $k = 4$.

We have thus defined our axes for further processing.

3. Profile and feature extraction

3.1. Slices

Profiles are univariate functions along a longitudinal axis obtained from local measurements. They are constructed by dividing the chromosome in slices of equal width perpendicular to the longitudinal axis. As a first step, we subdivide this axis by a set M of equidistant subdivision points. For the next step, let us denote by I the set of pixels making up the chromosome image. We now use M

to construct the Voronoi diagram $\Phi : M \to 2^I$ defined by $\Phi(m) = \{x \in I \,/\, \min_{y \in M} \|y - x\| = \|m - x\|\}$, $m \in M$. The Voronoi sets $\Phi(m)$ are the slices required; their union is I.

3.2. Profiles and features

Following Piper and Granum [6], we extract up to three profiles: The *density profile* describes the *local mass* close to each point on the axis. It assigns the sum of the gray values in each slice to its supporting point. This raw version of the density profile is very noisy since the numbers of chromosome pixels assigned to the subdivision points fluctuate heavily, in particular if slices are thin. Just as the curvature function, the density profile, too, is smoothed with a triangular kernel in order to eliminate noise. The *gradient profile* is the modulus of differences of the density profile. Information on the position of the centromere is contained in the so–called *shape profile*. It is defined as the *moment of inertia* of each normalized slice relative to the tangential direction of the longitudinal axis at the relative subdivision point. Again, suitable smoothing is crucial.

Piper and Granum [6] propose a set of 30 features, mainly certain linear functionals of the profiles above. We adopt these features with minor changes. In particular, we use Fourier coefficients instead of "wdd"–coefficients. Their numbers were determined by calibration.

4. Classification

Each spline constructed in Sect. 2 yields two feature sets, one for each polarity. This results in between two and 12 feature sets per chromosome due to unknown polarity and shape information. The "Simple Constrained Classifier–Selector" [9] is a Bayesian estimator for the shape, the polarity, as well as the class assignment at the same time. It uses all feature sets as inputs and can be applied in combination with all statistical models that have proved to be useful such as elliptical symmetry, quadratic asymmetry, and mixture models with outliers, cf. [8, 7]. Moreover, it is efficient since it amounts to solving a *Hitchcock problem*.

Let us finally compare the error rates of some statistical classifiers applied to the two data sets Cpr and Pki of *correctly segmented* cells described in Table 1. Both data sets consist of everyday clinical human cells and all classifiers are *constrained* to the correct number of chromosomes in each of the 24 biological classes. Kleinschmidt et al. [2] fed an ML–estimator based on the normal density function with the features described in [6] extracted from Cpr and achieved an error rate of 3.1% relative to chromosomes. In the same paper, they also applied an ML–classifier with a modified likelihood function to the same feature sets improving the error rate to 2.0%. Subsequently, our work

Data set	year	tissue	# cells
Cpr [5]	1988–90	amnion	2804
Pki	1999	amnion, chorion, blood	971

Table 1. Data sets for comparative studies

group [8, 7] designed various MAP–estimators based on the statistical models mentioned above attaining error rates down to 1.2%. The present method of profile and feature extraction further reduces this rate to 0.8%. It turned out that a good number of "errors" of the automatic classifiers were due to erroneous manual classifications. Therefore, the manual classification of Cpr was corrected by an experienced cytogeneticist. The error rate of the best of our MAP–classifiers applied to this data set is the currently best error rate of 0.6%, cf. [10].

The band level of Cpr lies between 150 and 200 and does not meet modern medical requirements. We also began to collect a new data set of clinical cells which we name "Pki", cf. Table 1. Presently, it consists of 971 pro– or metaphases at the band level 350-450; it is likely to grow. Here, our methods attain an error rate of 1.8%.

References

[1] F. C. Groen, T. K. ten Kate, A. W. M. Smeulders, and I. T. Young. Human chromosome classification based on local band descriptors. *Patt. Rec. Lett.*, 9:211–222, 1989.

[2] P. Kleinschmidt, I. Mitterreiter, and J. Piper. Improved chromosome classification using monotonic functions of Mahalanobis distance and the transportation method. *ZOR–Math. Meth. Oper. Res.*, 40:305–323, 1994.

[3] R. S. Ledley, H. A. Lubs, and F. H. Ruddle. Introduction to chromosome analysis. *Comput. Biol. Med.*, 2:107–128, 1972.

[4] D. Marr. *Vision*. Freeman, San Francisco, 1982.

[5] J. Piper. Variability and bias in experimentally measured classifier error rates. *Patt. Rec. Lett.*, 13:685–692, 1992.

[6] J. Piper and E. Granum. On fully automatic feature measurement for banded chromosome classification. *Cytometry*, 10:242–255, 1989.

[7] G. Ritter and K. Gaggermeier. Automatic classification of chromosomes by means of quadratically asymmetric statistical distributions. *Pattern Recognition*, 32:997–1008, 1999.

[8] G. Ritter and M. T. Gallegos. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Patt. Rec. Lett.*, 18:525–539, 1997.

[9] G. Ritter and M. T. Gallegos. A Bayesian approach to object identification in pattern recognition. In A. S. et al., editor, *Proceedings of the 15th International Conference on Pattern Recognition*, volume 2, pages 418–421, Barcelona, 2000.

[10] G. Ritter and G. Schreib. Using dominant points and variants for profile extraction from chromosomes. *Pattern Recognition*, 34:923–938, 2001.