Published in *Pattern Recognition* 34 (2001) 923–938

Using dominant points and variants for profile extraction from chromosomes[1]

Gunter Ritter[2] and Gernot Schreib

Fakultät für Informatik und Mathematik
Universität Passau, Germany

Abstract: The most accurate methods for automatic analysis 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. Ambiguities are dissolved by the recently developed 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 0.61%.

Key words: Analysis of chromosomes, image processing, profile extraction, shape recognition, dominant points, longitudinal axes, method of variants

Contents

[1]Research supported by Deutsche Forschungsgemeinschaft, Ri477/4
[2]Corresponding author, ritter@fim.uni-passau.de

1 Introduction

Classification of the chromosomes in a cell of a eukaryotic organism under a light microscope is a clear-cut task that lends itself to automation on computers. Automation has indeed begun in the early 1960's. It combines methods from engineering (image processing, pattern recognition), statistics (data and discriminant analysis), mathematics (discrete and continuous optimization, functional analysis, geometry, numerics, approximation theory), and computer science (algorithms, computational geometry, data bases, visualization tools); moreover, all this is in front of a cytogenetical background.

The most accurate automatic methods available today on the basis of *correctly segmented cells* follow a number of consecutive steps. These are

(i) extraction of *primitive numeric features* from the chromosome such as area and mean grey value;

(ii) identification of the *oblong shape* of the chromosome. This means finding a *longitudinal axis* by methods of 2D image processing. This axis is a continuous curve connecting the two tips of the chromosome thereby dividing it into its two chromatids, the essentially congruent, longitudinal halves;

(iii) representation of *band pattern* and *shape* by so-called *profiles*, i.e., univariate functions along the axis;

(iv) extraction of *profile-dependent numeric features* by applying methods from signal processing (Fourier or other coefficients) to the univariate profiles;

(v) *normalization* of features;

(vi) design and identification of a suitable *statistical model* of the feature sets by methods of statistical data analysis;

(vii) constrained *classification* of the feature sets into the biological classes.

Approaching the error rate of automatic chromosome classification to that of the human expert is a challenge for pattern recognition and image processing. The error rate had been continually lowered over the past two decades and the goal is coming into reach now. This progress is mainly due to application of three principles:

(α) exploitation of prior knowledge,

(β) precise data modeling including proper outlier handling and robust parameter estimation,

(γ) The Principle of Least Commitment.

Let us briefly discuss (α) – (γ). The method of classification most often used and yielding the most accurate results today is the Bayesian. Habbema [1, 2] and Slot [3] proposed taking into account the correct number of chromosomes in each class by constraining the optimization problem contained in Bayes' method in the way prescribed by Nature. This idea reduces the

error rate by a factor of almost 2. Tso and Graham [4] discovered an intimate connection between constrained classification of statistically independent objects and the linear assignment problem known from operations research; this led to efficient algorithms for solving the problem proposed by Habbema and Slot. Tso et al. [5] subsequently refined this method to an efficient way of context-dependent karyotyping.

Up to a few years ago the classical normal statistical model was used in order to describe feature sets. Today, more refined data models [6, 7], outlier handling [8], and robust methods of parameter estimation [9, 8] are being applied, resulting in another reduction of the error rate by a factor of 2.5. This latter improvement is due to accommodating outliers contained in the feature sets.

Marr's [10] "Principle of Least Commitment" postulates that no early decision be taken during a process of identification or recognition unless this decision is safe. We realize it here by the recently developed method of *variants* [11]. In (statistical) object identification one considers a number of observations and some statistical model, e.g. in the form of a likelihood function. The observations (*variants*) may either come from different physical objects or they may be different measurements of the same. In the latter case the observations foreign to the model may be – deterministic or random – perturbations of the correct observation of the object, sometimes just spurious ones. Assume that *exactly one* variant belongs to the model. Two problems arise: *selection* of this *regular* variant and *classification* of an object into one of several classes in the presence of variants of the object.

It is intuitively appealing to use the maximum of the products "likelihood times prior probability" of all variants as selection rule.[3] Note that this "Simple Selector" depends essentially on the statistical model of the *regular* variant alone and may not be optimal. However, a recent analysis [11] of the Bayesian estimator given the joint distribution of *all* variants reveals various sufficient conditions so that it equals the Simple Selector. One of them is involutive relationship between two variants and another one independence of all variants and identical distribution of the irregular ones.

The method of variants can be combined with classification and has many applications. Ritter and Pesch [12] used it for handling polarities in chromosome classification. The correct polarity of a chromosome is defined by the position of its centromere. This is, however, not always easily detected under a light microscope. In order not to rely exclusively on the centromere for estimating polarity one may consider two variants, one feature set for each polarity. Information at a higher level is then collected for both of them. The variants thus generated satisfy the first optimality condition stated above. Various *classifiers*, called "Simple Classifiers" and "Simple Classifier–Selectors", which efficiently and often optimally exploit all this information are due to Gallegos and Ritter, cf. Sect. 6.1; they defer the decision about the correct variant until the classification process, automatically choosing the most prospective variant as the basis for classification. The "polarity-free" chromosome classifier [12] is a special case; it reduces the error rate by another 25%.

Outliers in the feature data result mainly from errors committed during image processing, steps (ii) and (iii). Whereas their *accommodation*, cf. (β), is certainly helpful, their *avoidance* would be far better. We, therefore, take another look at image processing here. In particular, we employ the method of variants again in order to find the tips of the chromosome and we propose a new method for constructing a longitudinal axis. In our present application, the variants come from ambiguities in recognizing the shape of a chromosome. Note that there is only one correct shape interpretation whereas there may be many wrong ones!

A common approach in the literature [13] for constructing a longitudinal axis uses the so-called *medial axis* [14], at least in the case of bent chromosomes. This approach causes two problems: First, it yields a *set* of pixels rather than a parametrized *curve* and, second, this set is usually

[3] Despite the superficial similarity, variant selection must not be mistaken for hypothesis testing or classification. In some sense, variant selection is even converse. In the former case, several *observations* compete for one statistical model and in the latter several *statistical models* compete for one observation.

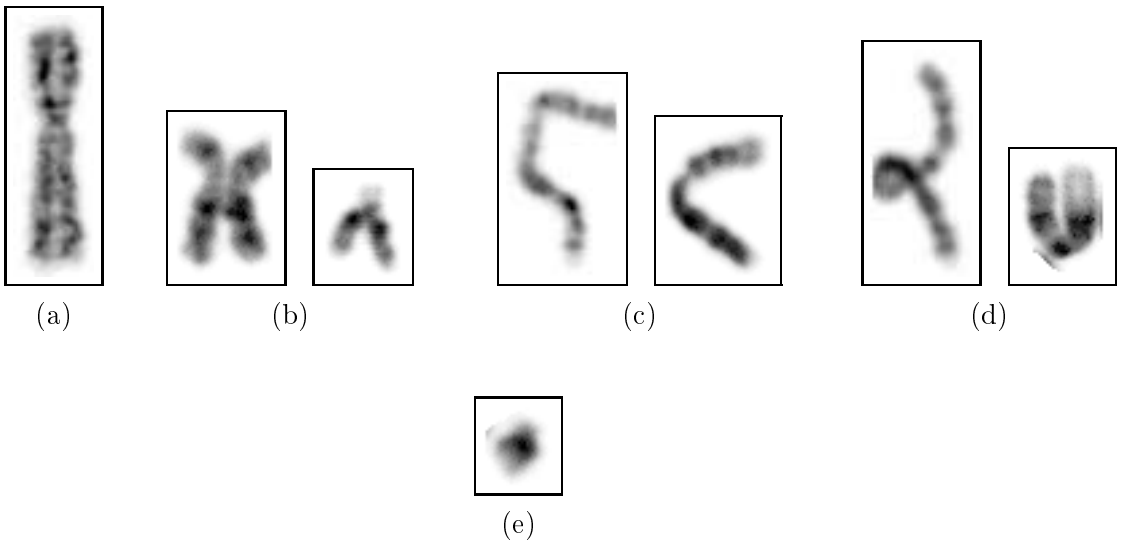


Figure 1: Various chromosome shapes. (a) Straight chromosome; (b) meta- and acrocentric chromosomes in late metaphase; (c) mildly bent chromosomes; (d) severely bent (twisted and U-shaped) chromosomes; (e) small, almost circular chromosome.

branched close to the two ends of the chromosome making it look like two Y's connected at their feet. More diffuse point sets may result from randomness of the shapes of chromosomes. The branching has to be removed and the medial axis has to be parametrized subsequently. Both these operations carry uncertainty into the method; they are necessary even with the best-shaped chromosomes. We feel that this is a disadvantage of the method of medial axes although the method proposed in [13] is often successful. A wrong longitudinal axis, however, leads to severely erroneous features and often to a misclassification.

We avoid this disadvantage by proposing an approach based on *dominant points*, Sect. 3. A dominant point in image processing commonly means a singularity or a point of conspicuous curvature along a curve or the contour of an object. The dominant points of a chromosome, i.e., the points of highest contour curvature, indicate the positions of its two tips. To our knowledge, the first authors to use the curvature for finding the tips, albeit with a different aim, were Ledley et al. [15]. Our method, too, tries to find these first. Of course, due to the random character of the chromosome images, these points are sometimes equivocal and we may get more than two of them, in particular with small, bent, X- or Y-shaped, or badly shaped chromosomes. For these cases, where the method finds more than two dominant points, we design some rules first trying to reduce their number to two. If there is doubt about the correct pair we resort to the method of variants described above, processing all possible pairs of dominant points. The classifier decides which variant is the correct one corresponding to the tips of the chromosome. These variants approximately satisfy the second optimality condition of the Simple Selector stated above.

We next connect each pair of dominant points to be processed by a cubic spline as described in Section 4.2. This is the longitudinal axis required. The parametrization of the axis flows from the natural parametrization of the chromosome's contour and, as a rule, there are no problems at the ends of the spline. Let us call a chromosome *severely bent* if two remote parts touch or overlap. Otherwise bent chromosomes will be called *mildly bent*. The method of dominant points combined with variants yields good results except in cases of severely bent or short, almost circular chromosomes where our method, too, may fail to produce correct axes.

After having constructed the longitudinal axis we extract band pattern profiles along the axis subdividing the chromosome in perpendicular slices. This is performed by means of a Voronoi diagram with respect to equidistant subdivision points on the axis.

The experienced cytogeneticist classifies chromosomes at an error rate between 0.1% (images of good quality) and 0.3% (clinical applications) with respect to chromosomes. To the best of our knowledge, the most accurate automatic classifier reported to date is a context-dependent, polarity-free Bayesian classifier [12] based on an accurate distributional model for the class-conditional distributions, cf. [6, 8, 7]. This classifier was applied to 29 features extracted from the profiles of the large Copenhagen image data set Cpr computed on the Edinburgh MRC chromosome analysis system [13]; it attained a cross-validation error rate of 0.92%. This is the first automatic classification of human chromosomes into their 24 classes achieving an error rate below 1%. The present profiles based on the method of dominant points and variants improve the error rate mentioned above by 15%, cf. Table 2 in Section 6. This progress must be attributed to superior axes, in particular in bent chromosomes. These cross-validation error rates refer to automatic classification of correctly segmented cell images without manual orientation of chromosomes. Section 6 contains also a study of parameter sensitivities. In the final Section 7 we discuss remaining problems and some future directions to be followed in order to further reduce the error rate by means of methods of image processing and pattern recognition.

Notation. The symbols $\mathbb{Z}$ and $\mathbb{Z}^2$ denote the set of integers and the discrete plane, respectively, and $\|\cdot\|$ stands for the Euclidean norm on the Euclidean plane $\mathbb{R}^2$. The notation $x \cdot y$ means the scalar product of $x, y \in \mathbb{R}^2$. A star $*$ denotes convolution of two functions on $\mathbb{Z}$ or on a cyclic group $\mathbb{Z}/N\mathbb{Z}$. Given a real valued function f, f^+ denotes its positive part. The neighborhood $\mathcal{N}_5$ of the origin in $\mathbb{Z}^2$ is the set $\{(0,0),(1,0),(-1,0),(0,1),(0,-1)\}$; other neighborhoods, such as $\mathcal{N}_9$, are defined in a similar way.

2 State of the art

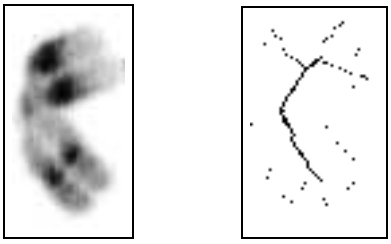
2.1 Profiles

By the early seventies biologists had developed a procedure to display AT- and CG-rich domains in chromosomes by using an enzymatic reaction and subsequent staining. AT-rich domains appear as dark bands perpendicular to the longitudinal axis; they are characteristic of the biological class allowing its recognition. A representation of the band pattern or other information (e. g. width) along the chromosome is called a *profile*; it is a univariate function whose construction needs a longitudinal axis. Three profiles are customary: the *density* profile, the *gradient* profile, and the *shape* profile.

In practice, the longitudinal axis is discretized for profile extraction. In [13], a line perpendicular to the longitudinal axis is drawn at each discretization point. These transverse lines are again discretized by equidistant points which are, in general, located between image pixels. These points are assigned gray values by means of two-dimensional linear interpolation between the four neighboring pixels. The value of the density profile corresponding to a perpendicular line is the sum of the gray values along the line. The shape profile represents the moment of inertia with respect to the normalized mass of the gray values along the transverse line. We will develop a simpler method based on Voronoi diagrams, cf. Sect. 5.

2.2 Longitudinal axis

Extraction of a *band-pattern profile* from a chromosome is based on a longitudinal axis as which its morphological *medial axis* [14, 16] can serve [13]. Stated in terms of mathematical morphology, the medial axis of an object in the Euclidean plane is the set of local maxima of the function which assigns to each point of the object its distance to the complement. The medial axis is a popular means for shape recognition, representation, and description and for image compression. There exist several thinning algorithms for extracting medial axes by means of morphological operations, both in the continuous and discrete cases, cf. [14, 17, 18] and the literature cited there.



(a) (b)

Figure 2: Medial axis. (a) original image; (b) medial axis

The method of medial axes was applied to chromosome analysis first by Hilditch [19], later by Piper and Granum [13]. A disadvantage of the medial axis in chromosome analysis is the fact that it is a point set rather than a parametrized curve which is eventually needed for profile extraction. Moreover, it tends to ramify near the tips of the object; ramifications may also be caused by noise in the object boundary. Another disadvantage was missing connection of the extracted axis; this was later amended [20, 21]. The tendency to produce ramified structures which have to be suitably parametrized, however, remains a problem, in particular with bent and badly-shaped chromosomes, cf. Fig. 2. This is the reason why we resort to a method that transforms the contour of the chromosome into a parametrized longitudinal axis right away. As a first step we need to estimate the tips of the chromosome; we will use a method based on dominant points.

2.3 Dominant points

Conspicuous points of a plane curve $B : \mathbb{R} \to \mathbb{R}^2$, e.g., points of high or low curvature or vertices, are usually called *dominant points* (or *landmarks*). Dominant points, although being still an active field of research, are by now a well–established method for representing, recognizing, analyzing, and describing the shape of an object or of a signal or for detecting an object in an image. In chromosome analysis, the dominant points of the contour curvature turn out to indicate possible candidates of the two tips.

In the case of a *discrete* curve $b : \mathbb{Z} \to \mathbb{R}^2$, a common algorithm for detecting dominant points uses *deflection angles* between adjacent points b_{i-1}, b_i, b_{i+1} or distant points $b_{i-\kappa}$, b_i, $b_{i+\kappa}$ [22] and [23]. Using distant points with subsequent smoothing by means of a filter is indispensable in the presence of boundary noise. Other algorithms are based on a numerical discretization of the curvature $\left(B'_x B''_y - B''_x B'_y\right) / \left({B'_x}^2 + {B'_y}^2\right)^{3/2}$ of a *plane* curve B, cf., e.g., Ansari and Delp [24]. Anderson and Bezdek [25] apply the method of minimum-mean-square approximation for computing deflection angles and curvatures of noisy curves; overviews with various suggestions for improvement, such as automatic adaptation of smoothing widths, are found in [26] and [27].

There exists a wealth of algorithms for extracting dominant points. An iterative method based on minimal polygonal approximation is proposed and applied to noiseless boundaries in Arcelli [28]. A method suitable in the presence of noise is the scale-space approach [29, 30, 31, 32, 33]; it attempts to eliminate dominant points generated by boundary noise and works on a scale of different smoothings of the curvature. Wavelets are used in [34] and morphological operations in [35]. A comparison of methods is presented in [36]. For our application we use the method proposed in [23] with a scale parameter adapting to size which turns out to be sufficiently accurate, robust, and efficient. It efficiently exploits the prior information that a chromosome has two tips.

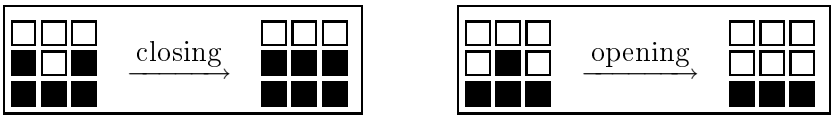


Figure 3: Smoothing by mathematical morphology. Black pixels represent the chromosome, white ones the background.

3 Extraction of dominant points

The basis for this and the following section is a gray-value image represented by a function $g : \mathbb{Z}^2 \to 0..255$ and containing a single chromosome in the form of the subset $I_{\text{raw}} = \{x \in \mathbb{Z}^2 / g(x) > 0\}$. The set $\{x \in \mathbb{Z}^2 / g(x) = 0\}$ represents the background of the image.

At first we have to recognize the oblong shape of the chromosome. To this end we use its tips which we estimate in four steps; these are

- extraction of the object boundary and of the associated contour,
- estimation of the contour curvature,
- determination of its "essential" maxima, and
- extraction of dominant points related to the essential maxima.

These procedures are applied to *all* chromosomes without distinction. The dominant points indicate the positions of the tips.

3.1 Boundary

As a first step towards constructing dominant points and detecting tips the $\mathcal{N}_5$-boundary of the chromosome is needed. A metaphase chromosome under a light microscope is a, usually oblong, 2D object with a noisy boundary. In order to remove some of the boundary noise in the form of indentations and protrusions it is suitable to first apply methods from mathematical morphology [14] to the set I_{raw}. In the language of mathematical morphology, the subsequent application of a dilation and an erosion is called *closing* whereas the application of the two operations in reverse order is called *opening* (Fig. 3). Both, openings and closings, are idempotent operations. In our case we apply a closing followed by an opening with respect to the neighborhood $\mathcal{N}_9$ to I_{raw}. Let I_{co} denote the resulting subset. We use a standard contour-following algorithm [16, Ch. 11] for transforming its boundary into a closed, discrete curve $(b_0, \ldots, b_{N-1})$ which we extend for convenience to an N-periodic sequence $(b_n)_{n\in\mathbb{Z}}$ in $\mathbb{Z}^2$ and which we call its *contour*. The chromosome used for further processing is the set $I = I_{co} \setminus \{b_0, \ldots, b_{N-1}\}$. All following operations are relative to I, cf. Fig. 4.

3.2 Contour curvature

The local maxima of the contour curvature indicate possible positions of the two tips of the chromosome. To each index $i \in \mathbb{Z}$ we associate the (possibly degenerate) triangle defined by the three points $b_{i-\kappa}$, b_i, $b_{i+\kappa}$ on the contour for some natural number $\kappa \geq 1$ to be specified and the rotation angle

$$(1) \quad c(i) = \arccos\left(\frac{(b_i - b_{i-\kappa}) \cdot (b_{i+\kappa} - b_i)}{\|b_i - b_{i-\kappa}\| \, \|b_{i+\kappa} - b_i\|}\right) \operatorname{sgn} \det(b_i - b_{i-\kappa} \quad b_{i+\kappa} - b_i).$$

The parameter κ should be adapted to scale and carefully chosen such that c yields high values for all indices i near the tips and low values otherwise. This is best achieved if 2κ lies between

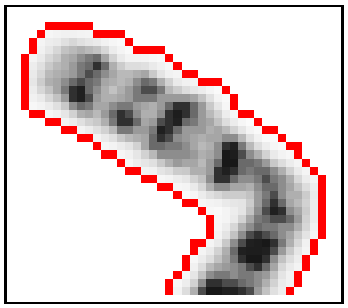


Figure 4: Chromosome boundary after closing and opening

the width and the length of the chromosome. If the mean of these two quantities is taken then κ is just $N/8$. The parameter κ also serves to reduce the influence of boundary noise. It turns out that a linear combination of the circumference N and the mean $\bar{N}$ of the circumferences of all chromosomes in the cell is more stable than a real multiple of N alone. We, thus, define

$$\kappa = \max\left\{\left\lfloor q_1\, N + q_2\, \bar{N}\right\rfloor, 1\right\}, \tag{2}$$

where q_1 and q_2 are real parameters. These, and all other parameters q_i appearing in the sequel, have to be determined by calibration. As a second means to reducing boundary noise the function c is smoothed by convolution with the triangular kernel

$$\begin{aligned} \Delta_\kappa : \mathbb{Z} &\to \mathbb{R} \\ z &\mapsto \begin{cases} \dfrac{1}{\kappa}\left(1 - \dfrac{|z|}{\kappa}\right), & \text{if } |z| < \kappa, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

We found in a series of assays that the optimal smoothing width κ is again as in Eqn. (2). The convolution $K := \Delta_\kappa * c$ serves as the "contour curvature" for further processing. Examples of curvature functions are shown in Figs. 5 and 6.

3.3 Essential maxima and dominant points of the contour

A *vertex* of a plane curve is a local minimum or maximum of its curvature. The following algorithm is inspired by Mukhopadhaya's (1909) four-vertex theorem of differential geometry in the large, cf., Osserman [37]: A simple, closed, regular $\mathcal{C}^2$-curve in the plane has at least four vertices, two maxima and two minima. Applied to the contour of a chromosome, the local maxima are, of course, candidates for its tips.

Because of remaining noise in the contour curvature K we cannot associate dominant points with all its local maxima. This is the reason why we define and look for *essential* maxima (and minima). Excluding the case of a constant curvature (this occurs, e.g., if the shape of the chromosome is $\mathcal{N}_4$ or $\mathcal{N}_5$ for $\kappa = 1$ and $\mathcal{N}_{12}$ for $\kappa \in \{1, 2\}$ — situations not occurring in practice) we proceed as follows.

Definition: Given a threshold $T \geq 0$, we say that the (N-periodic) function K has a *T-essential maximum* at $M \in \mathbb{Z}$, if there exist two indices $j_1 < M < j_2$ such that

(i) $K(M) = \max\limits_{j_1 < l < j_2} K(l)$

(ii) $\max\{K(j_1), K(j_2)\} \leq K(M) - T$.

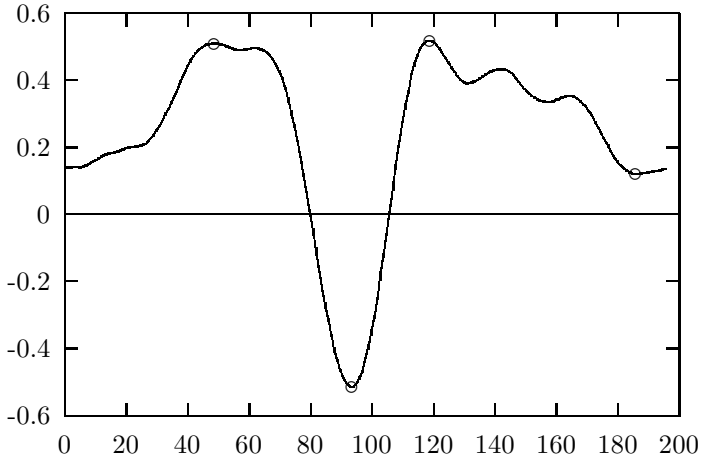


Figure 5: Curvature function of the chromosome shown in Fig. 2 with its T-essential extrema

By duality, we call a T-essential maximum of $-K$ a *T-essential minimum* of K. T-essential maxima and minima are illustrated in Fig. 5.

Because of (i), a T-essential maximum (minimum) is a local maximum (minimum). It is easy to see that 0-essential maxima (minima) are exactly the local ones and the global maxima (minima) are $\left(\max_{0\le l<N} K(l) - \min_{0\le l<N} K(l)\right)$-essential. Thus, a value T such that

$$(3) \quad 0 \le T \le \max_{0\le l<N} K(l) - \min_{0\le l<N} K(l)$$

should be chosen.

The properties of T-essential extrema are not obvious and we, therefore, include a few formal statements and proofs in the appendix. The proposition there shows that the Algorithm `essential_extrema` in Table 1 correctly computes all T-essential maxima (and minima which will be used later, too) of a periodic function K and that it terminates. Its inputs are K and the threshold

$$T = q_3 \left(\max_{0\le l<N} K(l) - \min_{0\le l<N} K^+(l)\right),$$

$0 \le q_3 \le 1$. Note that T is in harmony with (3). In general, a T-essential maximum is not yet a good estimate of a tip of the chromosome; if, e.g., the tip is flat then the maximum may not be centered, cf. Fig. 6. It may be located at an arbitrary point of its associated mountain and we rather propose to use the center of this mountain. More precisely, let m_i and m_{i+1} be two consecutive T-essential minima and let M_i, $m_i < M_i < m_{i+1}$, be the T-essential maximum located in between. If $K(M_i) > 0$ then define the height

$$h := K(M_i) - \max\{K^+(m_i), K^+(m_{i+1})\},$$

the cut level

$$h' := q_4\, h + \max\{K^+(m_i), K^+(m_{i+1})\}$$

and the indices j_1 and j_2

$$j_1 := \max\{j < M_i / K(j) \le h'\} \quad \text{and} \quad j_2 := \min\{j > M_i / K(j) \le h'\}.$$

We finally use $D_i := (j_1 + j_2)/2$ as the dominant point corresponding to M_i.

Table 1: Algorithm `essential_extrema`

Specification: $(n, (m_i)_{0\le i\le n}, (M_i)_{0\le i<n}) \leftarrow$ `essential_extrema`(K, T)
Given: a nonconstant, N-periodic function $K : \mathbb{Z} \to \mathbb{R}$,
a real number T satisfying (3).
Find: a number $n \ge 1$ and two finite sequences $(m_i)_{0\le i\le n}$, $(M_i)_{0\le i<n}$ of integers such that m_i is a T-essential minimum, M_i is a T-essential maximum and $m_i < M_i < m_{i+1}$ for all $0 \le i < n$. Within the period $m_0..m_n$, both sequences have maximal lengths.

```
begin
    n ← 0
    m_0 ∈ argmin_{0≤j<N} K(j)
    s ← max{i < m_0 / K(i) − K(m_0) ≥ T}
    m_0 ← min{i ≥ s / K(i) = K(m_0)}
    (m_0, l_0) ← search_next(m_0, K, T)
    do
       (M_n, L_n) ← search_next(l_n, −K, T)
       n ← n + 1
       (m_n, l_n) ← search_next(L_{n−1}, K, T)
    while m_n < m_0 + N od
    return(n, (m_i)_{0≤i≤n}, (M_i)_{0≤i<n})
end
```

Subroutine `search_next`

Specification: $(m, l) \leftarrow$ `search_next`(i, K, T)
Given: an index $i \in \mathbb{Z}$, an N-periodic function $K : \mathbb{Z} \to \mathbb{R}$,
a real number T satisfying (3).
Find: the pair of indices (m, l) such that $l > i$ is minimal with the property $K(l) - \min_{i\le j<l} K(j) \ge T$ and m is minimal in the set $\operatorname{argmin}_{i\le j<l} K(j)$.

```
begin
    l ← i + 1
    m ← i
    while K(l) − K(m) < T do
          l ← l + 1
          if K(l − 1) < K(m) then
             m ← l − 1
          fi
    od
    return(m, l)
end
```

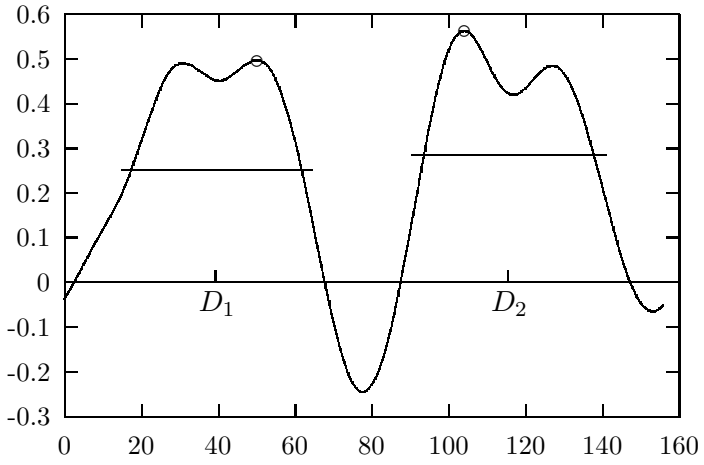


Figure 6: T-essential maxima and their dominant points D_1 and D_2 determined as the midpoints of the two associated mountains.

3.4 Rules for potential tips

If T meets the requirements of Algorithm `essential_extrema` and if the curvature function is not constant then the proposition in the appendix shows that we obtain at least one dominant point. We expect two dominant points in the case of an oblong, I-shaped chromosome. These indicate its tips. In order to justify the following rules it is useful to first discuss the various reasons why a chromosome may exhibit a number deviating from two. The first reason is biological. An important task of a chromosome is to divide into its two chromatids. It is, therefore, quite natural that chromosomes in late metaphase show three or four dominant points in the acrocentric and metacentric cases, respectively, cf. Fig. 7(d) and Fig. 8(c). Other sources of a deviating number of dominant points are artefacts such as bent and circular chromosomes and overlappings. Therefore, we developed the following set of rules in order to manage the various cases of one, two, three,... dominant points. Let $(D_1, \ldots, D_k)$ be the increasingly ordered sequence of dominant points constructed above.

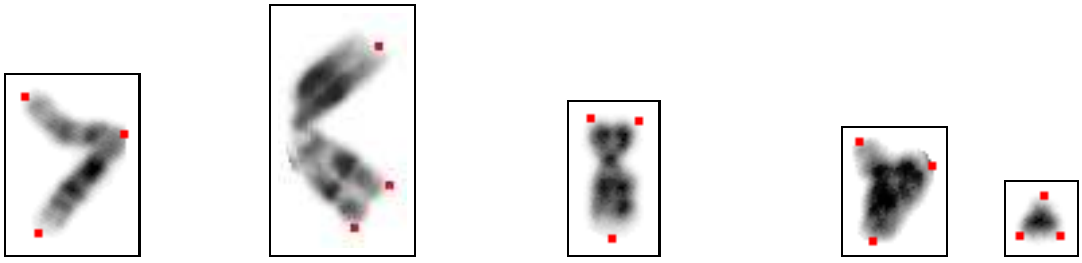
– $k = 1$: If there is only one dominant point D_1 then the most prospective complement is the opposite point on the boundary; these two points are the estimated tips.

– $k = 2$: We use D_1 and D_2 as tips.

– $k = 3$: In this case, we first try to reduce the number of dominant points to two by using the first of the the following two rules that applies.

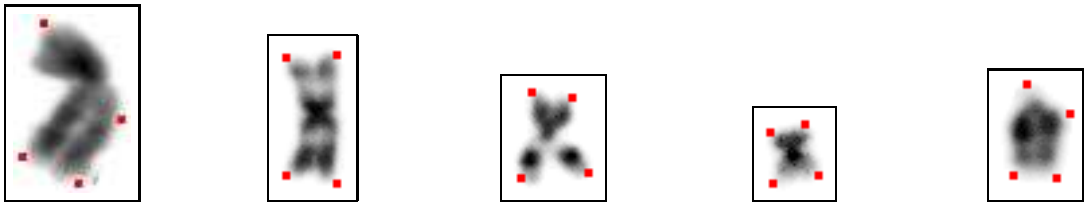
(i) *Deletion of a dominant point at the knee of a bent chromosome.* We first check whether the chromosome is bent. Let m_1, m_2, m_3 be the T-essential minima such that $D_1 < m_3 < D_2 < m_1 < D_3 < m_2$. Observe that m_i is the local minimum located opposite the dominant point D_i along the contour. Suppose that there is exactly one $l \in \{1, 2, 3\}$ such that $K(m_l) \leq -q_5\, K(D_l)$. Let D_a and D_b, $a, b \in 1..3$, be the two dominant points different from D_l. The chromosome is likely to be bent if the "double ratio" $R(D_a, D_b, D_l, m_l)$ is close enough to 1. In this case, we delete D_l from the list of dominant points, cf. Fig. 7(a). Otherwise, this chromosome may be bent and in the process of division, cf. Fig. 7(b).

(ii) *Replacing two neighboring dominant points with one.* Let $r_1 \leq r_2 \leq r_3$ denote the increasing rearrangement of the three distances $D_2 - D_1$, $D_3 - D_2$, $N + D_1 - D_3$ between the dominant points along the contour. If $r_1/r_2 \leq q_6$ then we have found a very short edge of the triangle,



(a) (b) (c) (d)

Figure 7: Three dominant points. (a) bent chromosome; (b) bent chromosome in late metaphase; (c) triangle with a short edge; (d) method of variants needed



(a) (b) (c) (d) (e)

Figure 8: Four dominant points. (a) bent chromosome; (b) rectangular chromosome; (c) X-shaped chromosome; (d) almost quadratic chromosome; (e) other artefact

cf. Fig. 7(c). Since the vertices adjacent to the edge of length r_1 are close together we replace them with their midpoint[4].

If none of the previous rules applies (cf. Fig. 7(d)) then no early decision about the pair of tips is taken. We rather resort to the method of *variants* described in the introduction and in Section 6.1. This method processes all six potential pairs of tips as variants and leaves it to the classifier to deal with all variants at the same time in order to estimate the class of the chromosome. These pairs are (D_1, D_2), (D_1, D_3), (D_2, D_3), $(D_1, \lfloor (D_2 + D_3)/2 \rfloor)$, $(D_2, \lfloor (N + D_1 + D_3)/2 \rfloor)$, $(D_3, \lfloor (D_1 + D_2)/2 \rfloor)$.

– $k = 4$: In this case, we determine local deflection angles at each of the four dominant points. If none is much smaller than 90° we have an indication that the chromosome is in late metaphase. In this case, we take the two pairs of potential tips formed by the midpoints[4] of opposite edges as variants, cf. Fig. 8(b). In the opposite case we delete the dominant point belonging to the smallest angle sending this chromosome to the case $k = 3$, cf. Fig. 8(a),(e).

We tested rules for replacing two neighboring dominant points and for detecting bent chromosomes also in this case. However, contrary to the case $k = 3$, these rules did not improve results.

– $k > 4$: We keep the dominant points corresponding to the four largest T-essential maxima and proceed with the case $k = 4$.

[4]Our implementation uses midpoints along the *contour*. If a data set contains many Y- and X-shaped chromosomes in late metaphase then *Euclidean* midpoints may be preferable.

4 Longitudinal axis

4.1 Supporting points

We next compute a longitudinal axis for each pair of dominant points (b_m, b_n) found in the preceding section. Without loss of generality we may assume $m = 0$ and $n \geq N/2$. Running at unit speed through the path $b_0, \ldots, b_n$ and, simultaneously, through the opposite path $b_0, b_{N-1}, \ldots, b_n$ at speed $t = (N-n)/n$ one arrives at the same time at the destination b_n. Let

$$p_k = \frac{b_k + b_{\lfloor N-kt \rfloor}}{2}, \quad 0 \leq k \leq n,$$

be the Euclidean midpoints of the contour points simultaneously reached. Obviously, $p_0 = b_0$ and $p_n = b_n$. Next define a natural number l and a subsequence $(p_{k_0}, \ldots, p_{k_l})$ of $(p_0, \ldots, p_n)$ of *supporting points* by putting recursively $k_0 = 0$ and

$$k_{j+1} = \min\{i > k_j / p_i \in I \text{ and } \|p_i - p_{k_j}\| \geq q_7\}$$

until no further points p_i are left. Let l be the largest index j. We finally replace p_{k_l} with b_n. This includes the point b_n as a supporting point and avoids that two adjacent supporting points lie too close together. Put $s_j = p_{k_j}$, $0 \leq j \leq l$. The sequence $P = (s_0, \ldots, s_l)$ serves as the basis for the transition to a parametrized curve in the plane $\mathbb{R}^2$, cf. Fig. 9(a).

4.2 Spline interpolation and extrapolation

In view of the following profile extraction it is desirable to construct a longitudinal axis as smooth as possible approximating the supporting points. Several classical two-dimensional spline algorithms, like B-spline curves [38] and Bernstein polynomials, are available. However, it turns out that the one-dimensional spline approximation designed by Reinsch [39] yields better results. Given a finite sequence $t_0, \ldots, t_l \in \mathbb{R}$ and an error bound S, Reinsch constructs a cubic spline $\bar{T} : [0, l] \to \mathbb{R}$ with minimal $\mathbb{L}^2$-norm of its second derivative such that

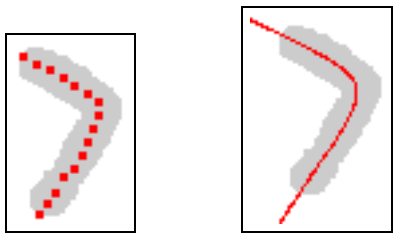
$$\sum_{i=0}^{l} \left(\frac{\bar{T}(i) - t_i}{\sigma_i} \right)^2 \leq S,$$

σ_i being the standard deviation of the random variable associated with t_i. The method itself takes care of distributing the approximation errors over the set of supporting points. We apply it to the two projections P_1 and P_2 of the sequence P of supporting points onto the coordinate axes obtaining a two-dimensional spline $\bar{S} = (\bar{T}_1, \bar{T}_2)$. We assume here that $\sigma_1 = \cdots = \sigma_{l-1} = q_8$, $\sigma_0 = \sigma_l = q_9$, and $S = l + 1$. If the only source of noise were the discretization of the image we would expect the parameters $q_8 = q_9 \approx 0.2$ to be optimal. Since there is an additional boundary noise and uncertainty about the position of the tips we rather expect optimal parameters $q_9 \gg q_8 > 0.2$. We finally extend this spline tangentially and linearly beyond both ends obtaining a curve $S : \mathbb{R} \to \mathbb{R}^2$ which serves as an extended longitudinal axis of the chromosome, cf. Fig. 9(b).

5 Profile extraction and features

5.1 Slices

Profiles are univariate functions along the longitudinal axis obtained from measurements in perpendicular direction. They are constructed by dividing the chromosome in slices of equal



(a) (b)

Figure 9: (a) Supporting points; (b) longitudinal axis

width perpendicular to the longitudinal axis. As a first step, equidistant subdivision points on the longitudinal axis are chosen by putting $t_0 = 0$ and

$$\int_{t_h}^{t_{h+1}} \|S'(\tau)\|\, d\tau = q_{10}, \quad h \in \mathbb{Z}.$$

Where $\|S'\|$ is large enough, the approximation

$$t_h = \begin{cases} t_{h-1} + \dfrac{q_{10}}{\|S'(t_{h-1})\|}, & \text{if } h > 0, \\ t_{h+1} - \dfrac{q_{10}}{\|S'(t_{h+1})\|}, & \text{if } h < 0, \end{cases}$$

can be used. The sequence of *subdivision points* is $M = (S(t_h))_{h\in\mathbb{Z}} \subseteq \mathbb{R}^2$. If the subdivision is fine enough, i.e., if q_{10} is small enough then the number $(t_{h+1} - t_h)\|S'(t_h)\|$ is sufficiently close to the distance $\|S(t_{h+1}) - S(t_h)\|$ along the longitudinal axis.

Next, from the set M of subdivision points we construct the Voronoi diagram $\bar{\Phi} : M \to 2^I$ defined by

$$\bar{\Phi}(m) = \left\{ x \in I \;/\; \min_{y\in M} \|y - x\| = \|m - x\| \right\};$$

ties are broken on a first-come-first-served basis. The Voronoi sets $\bar{\Phi}(m)$ are the slices required; their union is I, cf. Fig. 10. Moreover, by finiteness of I, there exist two numbers $\bar{h}_0, \bar{h}_1 \in \mathbb{Z}$, $\bar{h}_0 < \bar{h}_1$, such that $\bar{\Phi}(S(t_{\bar{h}_0})) \neq \emptyset$, $\bar{\Phi}(S(t_{\bar{h}_1})) \neq \emptyset$, and $\bar{\Phi}(S(t_h)) = \emptyset$ for all indices h outside the integral interval $\bar{h}_0..\bar{h}_1$. The number $\bar{h}_1 - \bar{h}_0$ is a first approximation of the length of the chromosome.

Since boundaries of chromosomes are noisy it turns out that it is favorable to restrict the chromosome to a narrower domain around the longitudinal axis, cf. [13]. We, therefore, use the narrowed slices

$$\Phi(m) = \left\{ x \in \bar{\Phi}(m) \;/\; \|m - x\| \le q_{11}\, \frac{\#I}{\bar{h}_1 - \bar{h}_0} \right\}$$

instead of $\bar{\Phi}(m)$. The quotient $\#I/(\bar{h}_1 - \bar{h}_0)$ is an approximation of the mean width.

5.2 Profiles

Following Piper and Granum [13] we extract three profiles: A profile describing the local mass close to each point on the axis (called *density profile*), the modulus of the derivative of the

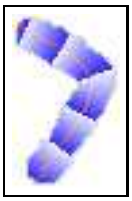


Figure 10: Slices represented by different gray values.

density profile (*gradient profile*), and a certain *shape profile* containing information on width and centromere. The *raw density profile* [40] is defined as

$$D_{\mathrm{raw}}(h) = \sum_{x \in \Phi(m)} \mathrm{g}(x), \qquad h \in \mathbb{Z},\ m = S(t_h),$$

where, as in Sect. 3, g is the function assigning gray values to pixels; by convention, $\sum_{x\in\emptyset} \mathrm{g}(x) = 0$. It is very noisy since our parameter q_{10}, cf. Sect. 5.1, is of subpixel size causing the masses contained in the slices to fluctuate heavily. This sum may be empty, even if $\bar{h}_0 < h < \bar{h}_1$. Just as the curvature function c, the function D_{raw}, too, is smoothed with a triangular kernel $\Delta_{q_{12}}$ in order to eliminate noise; this yields the smoothed density profile $D_{\mathrm{smooth}} = \Delta_{q_{12}} * D_{\mathrm{raw}}$.

In the same way as we narrowed the chromosome it is also favorable to shorten it at both ends in order to achieve better centering and to eliminate artifacts near the tips. Since the support of the function D_{smooth} is bounded and nonempty there exist the numbers $h_0 = \min\{h \in \mathbb{Z}/D_{\mathrm{smooth}}(h) \geq q_{13}\ \gamma\}$ and $h_1 = \max\{h \in \mathbb{Z}/D_{\mathrm{smooth}}(h) \geq q_{13}\ \gamma\}$, where γ is the mean mass of a chromosome in the cell and q_{13} is another strictly positive parameter. Let d denote the smaller of the two numbers $D_{\mathrm{smooth}}(h_0)$ and $D_{\mathrm{smooth}}(h_1)$. The density profile is finally the function D defined by

$$D(h) := D_{\mathrm{smooth}}(h) - d, \quad h_0 \leq h \leq h_1;$$

moreover, the number $h_1 - h_0$ is an estimate of the length of the chromosome.

The *gradient profile* is the modulus of differences of the density profile. Since this operation is analogous to forming a derivative the density profile has to be further smoothed with a triangular kernel $\Delta_{q_{14}}$. The gradient profile is then

$$G(h) = |\Delta_{q_{14}} * D(h+1) - \Delta_{q_{14}} * D(h)|, \quad h_0 \leq h < h_1.$$

Important information is also contained in the so-called *shape profile*, in particular information on the position of the centromere. We adopt the method introduced in [13] combining the width with the gray-value information contained in the pixels. Generally speaking, the shape profile is the moment of inertia of the *normalized* gray–value distribution in each slice relative to the tangent to the longitudinal axis at the subdivision point of the slice.

Robust estimation of the moment of inertia needs some care. We use slices $\Psi(m)$, $m = S(t_h)$, constructed in the same way as $\Phi(m)$ but with a different narrowing parameter q_{15} instead of q_{11}. If the tangential vector $v = S'(t_h)$ does not vanish then the normal vector $n_m = (-v_2, v_1)/\|v\|$ at $m = S(t_h)$ points in the main direction of the slice $\Psi(m)$. In the, unlikely, opposite case we replace this normal vector with the angular bisectrix of the triangle $S(t_{h-1}), S(t_h), S(t_{h+1})$. Note that $S(t_h)$ is different from $S(t_{h\pm1})$. Let $\pi_m(x) = (x - m) \cdot n_m$ be the projection of $x \in \Psi(m)$

onto the normal to the spline at m. Because of discretization, the quantities

$$M_0(h) := \sum_{x \in \Psi(m)} \mathrm{g}(x),$$
$$E(h) := \sum_{x \in \Psi(m)} \pi_m(x)\, \mathrm{g}(x), \quad \text{and}$$
$$M_2(h) := \sum_{x \in \Psi(m)} \pi_m^2(x)\, \mathrm{g}(x)$$

are only crude estimates of the total mass (in the *original, continuous,* image) along the normal at m, of the barycenter of the mass distribution along this normal, and of its moment of inertia with respect to the tangent, respectively. Due to the fluctuations explained above, some smoothing must first be performed and more robust estimates of these quantities are the convolutions $(\Delta_{q16} * M_0)$, $(\Delta_{q16} * E)$, and $(\Delta_{q16} * M_2)$. After normalizing, we finally find the robust estimate

$$h \mapsto \frac{(\Delta_{q16} * M_2)(h)}{(\Delta_{q16} * M_0)(h)} - \left(\frac{(\Delta_{q16} * E)(h)}{(\Delta_{q16} * M_0)(h)} \right)^2, \quad h_0 \le h \le h_1, \tag{4}$$

of the (length-corrected) shape profile. Reversing the order of smoothing and the normalization contained in Eqn. (4) would be rather disastrous.

5.3 Features

Each pair of dominant points found in Section 3.4 defines a spline as constructed in Section 4.2. Each spline in turn yields two feature sets, one for each polarity. This results in between two and twelve variants for each chromosome.

Since we are dealing with variants coming from several "axes" of the same chromosome some caution is necessary concerning the normalization of some of the features. Normalization of a feature needs the sum of the corresponding feature values across all chromosomes of the cell. If there are variant axes then the correct features are unknown at this stage. This makes it necessary to take a preliminary choice of a suitable variant for each chromosome in order to approximate this sum well. In the case of three dominant points the longest of the point-to-edge axes is the most prospective one and in the case of four dominant points it is the longer of the two.

6 Classification and parameter sensitivity

6.1 Classification in the presence of variants

We have explained the idea of variants already in the introduction. In the present paper, besides the two polarities, the axes arising from uncertainties during the image processing are treated as variants. In the former case, the wrong polarity is a perturbation of the correct one, in the latter case wrong axes give rise to spurious observations, cf. Sect. 3.4. The task is not only to find the regular variant, i.e., the correct axis in its correct polarity, but also to recognize the biological class despite the presence of variants. Moreover, the method of recognition should be context sensitive. Let us describe an algorithm for *constrained* classification in the presence of variants due to M. T. Gallegos and the first-named author, the "Simple Constrained Classifier". A theoretical Bayesian analysis of this estimator will appear elsewhere; the special case of polarities was treated by Ritter and Pesch [12].

Let f_j be the class-conditional density function (of the regular variant) corresponding to the biological class $j \in 1..n$ and let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,b_i}) \in \mathbb{R}^{b_i d}$ be the array of variants of chromosome

$i \in 1..m$ in the cell to be classified. Let p_j be the prior probability of class j, let $q_{i,h}$ be the prior probability of the observation $h \in 1..b_i$ to be the regular variant of chromosome i, and let α_j be the prior probability for class j to be missing, cf. [8, 7].

Simple Constrained Classifier Let $\hat{\sigma}$ be the solution to the linear assignment problem

$$\min_{\sigma} \sum_i d_{i,\sigma(i)},$$

where the minimum ranges over all permutations σ of n elements and where

$$d_{i,j} = \begin{cases} -\ln\left(\sum_{h=1}^{b_i} f_j(x_{i,h}) q_{i,h}\right) - \ln p_j, & i \le m, \\ -\ln \alpha_j, & j > m, \end{cases}$$

$1 \le j \le n$. The MAP-classifier given the observation $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is the restriction of $\hat{\sigma}$ to $1..m$. □

The point is that the Simple Constrained Classifier estimates the class of $\mathbf{x}_i$ on the basis of the f_j's alone, i.e., without knowledge of the densities of the irregular variants. These are not at hand if the irregular variants arise from wrong axes. However, it can be shown that, under the conditions on the joint distribution of all variants of each object mentioned in the introduction, it equals the Bayesian estimator given the whole statistical information. The estimator implicitly uses the regular variant for classification. If, besides the class, the regular variant together with its correct polarity is to be estimated then summation over $1 \le h \le b_i$ has to be replaced with $\operatorname{argmax}_{1 \le h \le b_i}$; this is required for producing a karyotype.

Since the Simple Constrained Classifier amounts to solving a linear assignment problem it has an efficient solution. On the other hand, various statistical models f_j such as elliptical symmetry [6], quadratic asymmetry [7], and mixture models [8] have been shown to be useful; these gave rise to the definition of a series of classifiers constrained on the correct number of chromosomes in each class. In the cases without variants, they were given the acronyms IEC, IQA, IECO, and IQAO, the letter "O" standing for a mixture model with **o**utliers. The relative Simple Constrained Classifiers for the different statistical models are dubbed VIEC, VIQA, VIECO, and VIQAO. These algorithms are also applicable to the cases of missing chromosomes and trisomies, cf. [8, 7].

6.2 Classification results

We used the large Copenhagen image data set Cpr as a benchmark. It consists of 2 804 karyotyped human cells. This size is necessary to allow the estimation of a large number of parameters for identifying accurate statistical data models. (Data sets formerly – and still – used as benchmarks, such as the Copenhagen, the Edinburgh, and the Philadelphia data sets, consist of between a hundred and two hundred cells, only.) Most of the images contain between 150 and 250 bands. As mentioned in the introduction, Ritter and Pesch [12] recently tested four polarity-free classifiers with 29 features extracted from Cpr by means of the axes described in [13]. In the present paper, we apply the method of variants also to variant axes. In Table 2, we compare cross-validation (test-set) results for almost the same 29 features but now extracted by means of the axes described in Section 5.3 with the former results. Prior probabilities p_j and α_j are given in [8] and the $q_{i,h}$'s are uniform.

The error rates of our best classifiers are approaching that of the expert cytogeneticist which, in everyday clinical data sets, was estimated as 0.3%. The amount of manual misclassifications contained in Cpr is of this size. It can be expected that most of these errors appear as "training-set errors" no matter which classifier is used. All cells containing such deviating classifications which could not be explained by obvious defects in the images were shown to an expert cytogeneticist (without telling him the differences). In one out of three cases he decided that the manual

	$\text{VIEC}_{\text{normal}}$	$\text{VIEC}_{\text{Pareto}}$	$\text{VIECO}_{\text{Pareto}}$	$\text{VIQAO}_{\text{Pareto}}$
	1.94/27.2	1.27/18.3	0.99/14.9	0.92/14.2
Cpr		λ=33.0	λ_{OUT}=33.0	λ_{OUT}=34.0
Edinburgh axes			cut_{BAS}=10.0	cut_{BAS}=6.5
			cut_{OUT}=7.3	cut_{OUT}=7.1
				Q-factor=0.4
	1.42/21.4	0.97/14.5	0.81/12.3	0.78/11.9
Cpr		λ=34.0	λ_{OUT}=34.0	λ_{OUT}=34.0
dominant points			cut_{BAS}=6.5	cut_{BAS}=6.4
and variant axes			cut_{OUT}=7.6	cut_{OUT}=6.9
				Q-factor=0.5
	1.24/18.9	0.80/12.0	0.64/9.6	0.61/9.2
Cpa		λ=34.0	λ_{OUT}=34.0	λ_{OUT}=34.0
dominant points			cut_{BAS}=7.5	cut_{BAS}=6.0
and variant axes			cut_{OUT}=7.3	cut_{OUT}=6.9
				Q-factor=0.4

Table 2: Cross-validation error rates with respect to the data sets Cpr and Cpa for MAP-classifiers based on various statistical models f_j, cf. [6, 8, 7, 12]. The specification p/q means that p percent of chromosomes were misclassified in q percent of all cells. The meanings of the classification parameters λ, λ_{OUT}, cut_{BAS}, cut_{OUT} and Q-factor are explained in [8] and [7]. Variances and quadratic asymmetries robustly estimated. First row taken from [12].

classification was wrong; the class assignments of these about 200 misclassified chromosomes were corrected. About 100 wrong polarities, mainly in classes 1, 3, and 19, appearing as outliers in the process of parameter estimation, were corrected, too. Since our error rates are always understood after correct segmentation of cell images four faulty segmentations were improved. The corrected image data set was named Cpa. The error rates obtained for Cpa are included in Table 2. The error rates of the two data sets differ by 0.17%. The best test set result for Cpa reported here is an error rate of 0.61% with respect to chromosomes. Ten out of eleven cells are completely correctly recognized.

6.3 Parameter sensitivity

We finally tested the sensitivities of classification results to variations of the parameters q_i. The results are shown in Table 3. The parameters q_1 and q_2 influence the curvature function, q_3 and q_4 are responsible of the positions of the T-essential maxima, q_5 and q_6 control the rules for constructing dominant points from essential maxima, and q_7–q_9 optimize approximation and smoothness of the spline. The remaining parameters are used for profile extraction.

The parameter q_4 determines the position of the dominant points. Because of local minima of the contour curvature this position may react in a discontinuous way to changes of q_4 which substantiates the relatively high sensitivity of the parameter q_4. Also the parameter q_{10} merits special attention; it is the granularity of discretization of the longitudinal axis. This discretization causes fluctuations of the number of pixels in the Voronoi sets which explain the high sensitivity of this parameter. The narrowing parameter q_{15} must not be much smaller than 0.5 since, otherwise, the shape cannot be recognized. All other parameters are fairly insensitive.

i	q_i	δ	Sensitivities				
			$\Delta_i(-2\delta)$	$\Delta_i(-\delta)$	$\Delta_i(0)$	$\Delta_i(\delta)$	$\Delta_i(2\delta)$
1	0.035	0.005	0.03	0.02	0.00	0.02	0.02
2	0.055	0.005	0.04	0.02	0.00	0.01	0.03
3	0.12	0.04	0.02	0.02	0.00	0.02	0.04
4	0.5	0.05	0.04	0.06	0.00	0.05	0.03
5	0.8	0.1	0.02	0.01	0.00	0.01	0.01
6	0.5	0.1	0.01	0.01	0.00	0.00	0.01
7	5	0.4	0.04	0.05	0.00	0.05	0.06
8	0.29	0.05	0.04	0.03	0.00	0.01	0.04
9	2.3	0.2	0.01	0.01	0.00	0.03	0.03
10	0.2	0.025	0.12	0.06	0.00	0.04	0.06
11	0.39	0.04	0.03	0.03	0.00	0.02	0.03
12	5	1	0.05	0.01	0.00	0.02	0.02
13	$4.0\ 10^{-4}$	$4.0\ 10^{-5}$	0.03	0.02	0.00	0.02	0.01
14	5	1	0.02	−0.01	0.00	0.00	0.02
15	0.48	0.05	0.07	0.02	0.00	0.04	0.06
16	8	1	0.01	0.00	0.00	0.00	0.01

Table 3: Parameter sensitivities of error rates with respect to the classifier $\text{VIECO}_{\text{Pareto}}$ and the data set Cpr. The second column contains the optimal parameter values determined by calibration. The error rate with respect to chromosomes for the optimal parameters referred to in the column labeled $\Delta_i(0)$ is 0.81%. Parameters are varied by the quantities shown in the third column and the resulting differences of the error rates in the remaining columns. They are given in % relative to the total number of chromosomes.

7 Discussion

Dominant points in combination with variants were shown to be an effective and natural tool for recognizing the oblong shape of a chromosome and for laying the ground for feature extraction. Their application reduces the number of outliers contained in the feature data substantially.

Visual inspection of chromosomes with high Mahalanobis distances from their class centers shows that remaining outliers in the feature set come from outlier images and have mainly four causes:

- chromosome severely bent, cf. Fig. 1(d);
- overlappings, cf. Fig. 11(a);
- Giemsa stains at or on the chromosome, cf. Fig. 11(b),(c);
- blurred or indistinct images, cf. Fig. 11(d).

We did not pay particular attention to avoid classification errors of severely bent chromosomes. This could be done by applying ad hoc methods but would not add much insight into pattern recognition. Overlappings give rise to perturbed profiles and, thus, to large Mahalanobis distances. Stains are most confusing since they damage both axes and profiles; they are difficult to remove. Finally, blurred and indistinct images puzzle even the expert.

The efficacy of the method of variants, theoretically established in [11], is confirmed in this paper by an application to variant axes. All variants are offered to the classifier which deals with the different variants in a smooth way. This encouraging observation suggests to apply variants in other contexts as well. It is conceivable to apply the method of variants again to overlappings; it may also be used for making effective use of the position of the centromere. In the present

(a) (b) (c) (d)

Figure 11: Outlier images. (a) Overlapping; (b) two homologous chromosomes, the right one exposing stains; (c) chromosome with a stain; (d) blurred and indistinct image.

paper, as in [12], the position of the centromere is not explicitly used, neither as a feature nor for the purpose of orientation. It is, however, implicitly contained in the shape profile.

Acknowledgments. We thank Dr. Jim Piper of Vysis Inc. for providing us the image data set Cpr and for a number of stimulating discussions. We also thank Herr Harth, Passau, for his assistance with biomedical questions.

Appendix: T-essential extrema

Let $K : \mathbb{Z} \to \mathbb{R}$ be a nonconstant, N-periodic function. We first show that T-essential extrema of K as defined in Section 3.3 preserve the alternating property that local extrema enjoy, viz., if $M_1 \neq M_2$ are T-essential maxima of K such that $K(M_1) \neq K(M_2)$ then there exists a T-essential minimum in between. If $K(M_1) = K(M_2)$ then there may be no T-essential minimum between M_1 and M_2. In fact, there may even be more than two consecutive maxima of equal height with no T-essential minima in between. We will need this property for proving correctness of Algorithm `essential_extrema` displayed in Table 1.

Lemma: If $M_1 < M_2$ are T-essential maxima of K such that $K(M_1) \neq K(M_2)$ then there exists a T-essential minimum m such that $M_1 < m < M_2$. An analogous statement holds for T-essential minima.

Proof. Without loss of generality, let $K(M_1) > K(M_2)$. Let j_1, j_2, $j_1 < M_2 < j_2$, be two points for M_2 as in the definition. By (i) and since $K(M_1) > K(M_2)$ we have $M_1 < j_1$. Let $m := \operatorname{argmin}_{M_1<l<M_2} K(l)$. We conclude $K(j_1) \geq K(m)$ which, together with (ii), yields

$$K(M_1) - K(m) \geq K(M_2) - K(m) \geq K(M_2) - K(j_1) \geq T.$$

This proves that $K(m)$ is a T-essential minimum. □

We are now in the position to show that Algorithm `essential_extrema` terminates and is correct.

Proposition: Let T be as in (3). Then the Algorithm `essential_extrema` of Table 1 computes an integer $n \geq 1$ and two sequences $(m_i)_{0\leq i\leq n}$ and $(M_i)_{0\leq i<n}$ of indices such that m_i is a T-essential minimum, M_i is a T-essential maximum, and $m_{i-1} < M_{i-1} < m_i$ for all $1 \leq i \leq n$. Moreover, we have $m_n = m_0 + N$ and the sequences are of maximal length within the interval $m_0..m_n$.

Proof. We first discuss termination and correctness of Subroutine `search_next`. Since $T \leq \max_{0\leq j<N} K(j) - \min_{0\leq j<N} K(j)$ and K is N-periodic the condition of the while-loop fails if m is a global minimum and l a global maximum. This state is reachable upon which `search_next` terminates. Turning to correctness of `search_next` note that, after each call of the while-loop,

the variable m contains the smallest index m such that $i \le m < l$ and $K(m) = \min_{i \le j < l} K(j)$; this is minimality of m. The loop terminates with the first l such that $K(l) - K(m) \ge T$, hence minimality of l.

We now deal with termination and correctness of Algorithm `essential_extrema`. The initial call of `search_next` yields a pair of indices (m, l) such that $m_0 \le m < l$ and $K(m) = \min_{m_0 \le j < l} K(j) = K(m_0)$, since m_0 is a global minimum. By minimality of m we can conclude $m = m_0$. Since the algorithm visits the while-loop at least once we have $n \ge 1$. The successive calls of `search_next` yield sequences (m_j), (l_j) and (M_j), (L_j) such that the relations

$$m_j < l_j \le M_j < L_j \le m_{j+1} < l_{j+1}$$

hold after the $(j+1)$th run through the while-loop. Therefore, (m_j) is a strictly increasing sequence and the while-loop terminates.

We next show that the M_j's are T-essential maxima. By definition of M_j we have $K(M_j) \ge K(i)$ for all i, $M_j \le i \le L_j$. By construction of M_j we have $K(M_j) \ge K(i)$ for all i, $l_j \le i \le M_j$, and from minimality of l_j we infer $K(l_j) \ge K(i)$ for all i, $m_j \le i \le l_j$. Hence, M_j maximizes K in the interval $m_j..L_j$, i.e., we have (i) for M_j with $j_1 = m_j$ and $j_2 = L_j$. Observing $K(l_j) - K(m_j) \ge T$ and $K(M_j) \ge K(l_j)$ we obtain

$$K(M_j) - K(m_j) \ge K(l_j) - K(m_j) \ge T.$$

Since $K(M_j) - K(L_j) \ge T$ is ensured by `search_next` we have also (ii) so that M_j is a T-essential maximum. A dual argument shows that m_j is a T-essential minimum.

The lemma above justifies the alternating search of T-essential minima and maxima, since subroutine `search_next` correctly handles consecutive minima and maxima, respectively, of equal height. After termination we have $m_n \ge m_0 + N$; in fact, equality holds. Indeed, note first $m_{n-1} < m_0 + N \le m_n$. If we had $m_{n-1} < m_0 + N < M_{n-1}$ then there were no T-essential maximum located between the T-essential minima m_{n-1} and $m_0 + N$ and the lemma would imply $K(m_{n-1}) = K(m_0)$. This would contradict the construction of m_0, cf. the initialization of the algorithm. Hence $M_{n-1} < m_0 + N \le m_n$ and by minimality of L_{n-1}, we have $M_{n-1} < L_{n-1} \le m_0 + N$. Since m_0 is a global minimum, it follows from the definition of m_n

$$K(m_0 + N) = K(m_0) \le K(m_n) = \min_{L_{n-1} \le j \le m_n} K(j) \le K(m_0 + N) = K(m_0),$$

in particular $K(m_0) = K(m_n)$ and $m_n = m_0 + N$ again by minimality of m_n.

Finally, maximality of the length of the two computed sequences follows from minimality of (M_n, L_n) and (m_n, l_n), respectively. □

References

1. J.D.F. Habbema, A discriminant analysis approach to the identification of human chromosomes, *Biometrics* **32**, 919–928 (1976).

2. J.D.F. Habbema, Statistical methods for classification of human chromosomes, *Biometrics* **35**, 103–118 (1979).

3. R. E. Slot, On the profit of taking into account the known number of objects per class in classification methods, *IEEE Trans. Inf. Theory* **25**, 484–488 (1979).

4. M.K.S. Tso and J. Graham, The transportation algorithm as an aid to chromosome classification, *Patt. Rec. Lett.* **1**, 489–496 (1983).

5. M.K.S. Tso, P. Kleinschmidt, I. Mitterreiter, and J. Graham, An efficient transportation algorithm for automatic chromosome karyotyping, *Patt. Rec. Lett.* **12**, 117–126 (1991).

6. Gunter Ritter, María Teresa Gallegos, and Karl Gaggermeier, Automatic context–sensitive karyotyping of human chromosomes based on elliptically symmetric statistical distributions, *Patt. Rec.* **28**, 823–831 (1995).

7. Gunter Ritter and Karl Gaggermeier, Automatic classification of chromosomes by means of quadratically asymmetric statistical distributions, *Patt. Rec.* **32**, 997–1008 (1999).

8. Gunter Ritter and María Teresa Gallegos, Outliers in statistical pattern recognition and an application to automatic chromosome classification, *Patt. Rec. Lett.* **18**, 525–539 (1997).

9. Jim Piper, I. Poole, and A. Carothers, Stein's paradox and improved quadratic discrimination of real and simulated data by covariance weighting, in *Proceedings of the 12 th IAPR International Conference on Pattern Recognition*, Los Alamitos, CA, IEEE Comput. Soc. Press, Jerusalem, Israel, 1994, 529–532.

10. D. Marr, *Vision*, Freeman, San Francisco, 1982.

11. Gunter Ritter and María Teresa Gallegos, Bayesian object identification: variants, *Journal of Multivariate Analysis* **81**, 301–334 (2002).

12. Gunter Ritter and Christoph Pesch, Polarity–free automatic classification of chromosomes, *Computational Statistics and Data Analysis* **35**, 351–372 (2001).

13. Jim Piper and Erik Granum, On fully automatic feature measurement for banded chromosome classification, *Cytometry* **10**, 242–255 (1989).

14. Jean Serra, *Image Analysis and Mathematical Morphology*, Academic Press, London, San Diego, New York, Berkeley, Boston, Sydney, Tokyo, Toronto, third edition, 1989.

15. Robert S. Ledley, Herbert A. Lubs, and Frank H. Ruddle, Introduction to chromosome analysis, *Comput. Biol. Med.* **2**, 107–128 (1972).

16. Azriel Rosenfeld and Avinash C. Kak, *Digital Picture Processing*, volume 2, Academic Press, Orlando, San Diego, New York, Austin, Boston, London, Sydney, Tokyo, Toronto, second edition, 1982.

17. Annick Montanvert, Medial line: Graph representation and shape description, in *Proc. Eighth Int. Conf. on Pattern Recognition*, Paris, 1986, 430–432.

18. Narendra Ahuja and Jen-Hui Chuang, Shape representation using a generalized potential field model, *IEEE Trans. Patt. Anal. Mach. Int.* **19**, 169–179 (1997).

19. C. Judith Hilditch, Linear skeletons from square cupboards, *Machine Intelligence* **4**, 403–420 (1969).

20. Liang Ji and Jim Piper, Fast homotopy-preserving skeletons using mathematical morphology, *IEEE Trans. Patt. Anal. Mach. Int.* **14**, 653–664 (1992).

21. S. Hazout and N. Q. Nguyen, Image analysis by morphological automata, *Pattern Recognition* **24**, 401–408 (1991).

22. Azriel Rosenfeld and Emily Johnston, Angle detection on digital curves, *IEEE Trans. Computers* **22**, 875–878 (1973).

23. Azriel Rosenfeld and Joan S. Weazka, An improved method of angle detection on digital curves, *IEEE Trans. Computers* **24**, 940–941 (1975).

24. Nirwan Ansari and Edward J. Delp, On detecting dominant points, *Pattern Recognition* **24**, 441–451 (1991).

25. Ian M. Anderson and James C. Bezdek, Curvature and tangential deflection of discrete arcs: A theory based on the commutator of scatter matrix pairs and its application to vertex detection in planar shape data, *IEEE Trans. Patt. Anal. Mach. Int.* **6**, 27–40 (1984).

26. Cho-Huak Teh and Roland T. Chin, On the detection of dominant points on digital curves, *IEEE Trans. Patt. Anal. Mach. Int.* **11**, 859–872 (1989).

27. José Manuel Iñesta, Mateo Buendía, and María Ángeles Sarti, Reliable polygonal approximations of imaged real objects through dominant point detection, *Pattern Recognition* **31**, 685–697 (1998).

28. Carlo Arcelli and Giuliana Ramella, Finding contour-based abstractions of planar patterns, *Pattern Recognition* **26**, 1563–1577 (1993).

29. Farzin Mokhtarian and Alan Mackworth, Scale-based description and recognition of planer curves and two-dimensional shapes, *IEEE Trans. Patt. Anal. Mach. Int.* **8**, 34–43 (1986).

30. Farzin Mokhtarian and Alan K. Mackworth, A theory of multiscale, curvature-based shape representation for planar curves, *IEEE Trans. Patt. Anal. Mach. Int.* **14**, 789–805 (1992).

31. Soo-Chang Pei and Chao-Nan Lin, The detection of dominant points on digital curves by scale-based filtering, *Pattern Recognition* **25**, 1307–1314 (1992).

32. Anothai Rattarangsi and Roland T. Chin, Scale-based detection of corners of planar curves, *IEEE Trans. Patt. Anal. Mach. Int.* **14**, 430–449 (1992).

33. Bingcheng Li, Repeatedly smoothing, discrete scale-space evolution and dominant point detection, *Pattern Recognition* **29**, 1049–1059 (1996).

34. Jiann-Shu Lee, Yung-Nien Sun, Chin-Hsing Chen, and Ching-Tsorng Tsai, Wavelet based corner detection, *Pattern Recognition* **26**, 853–865 (1993).

35. Xintong Zhang and Dongming Zhao, A parallel algorithm for detecting dominant points on multiple digital curves, *Pattern Recognition* **30**, 239–244 (1997).

36. Tokuhisa Kadonaga and Keiichi Abe, Comparison of methods for detecting corner points from digital curves, in *Graphics Recognition*, volume 1072 of *Lecture Notes in Computer Science*, eds. R. Kasturi and K. Tombre, Springer, 1996, 23–34.

37. R. Osserman, The four-or-more-vertex theorem, *AMM* **92**, 332–337 (1985).

38. William J. Gordon and Richard F. Riesenfeld, B-spline curves and surfaces, in *Computer aided geometric design*, eds. Robert E. Barnhill and Richard F. Riesenfeld, Academic Press, San Diego, 1974, 95–126.

39. Christian H. Reinsch, Smoothing by spline functions, *Numer. Math.* **10**, 177–183 (1967).

40. S. Hazout, J. Mignot, M. Guiguet, and A. J. Valleron, Rectification of distorted chromosome image: automatic determination of density profiles, *Comput.-Biol.-Med.* **14**, 63–76 (1984).