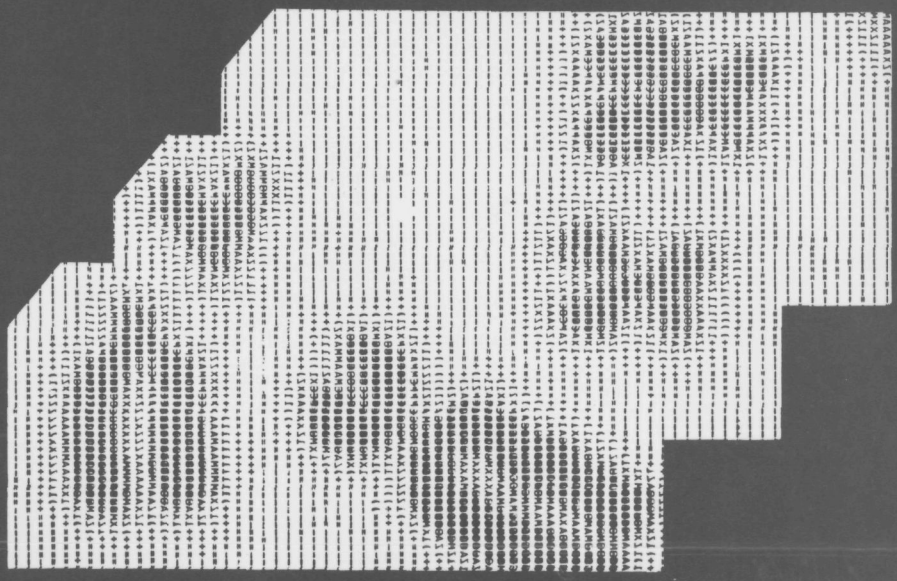
ANALYSIS OF DNA BASED MEASUREMENT METHODS APPLIED TO HUMAN CHROMOSOME CLASSIFICATION



F. C. A. GROEN

ANALYSIS OF DNA BASED MEASUREMENT METHODS APPLIED TO HUMAN CHROMOSOME CLASSIFICATION

ISBN 90 6231 023 0 soft-bound edition
ISBN 90 6231 024 9 hard-bound edition

ANALYSIS OF DNA BASED MEASUREMENT METHODS APPLIED TO HUMAN CHROMOSOME CLASSIFICATION

PROEFSCHRIFT

ter verkrijging van de graad van doctor in de technische wetenschappen aan de Technische Hogeschool Delft, op gezag van de rector magnificus Prof. Ir. L. Huisman, voor een commissie aangewezen door het college van dekanen te verdedigen op woensdag 23 februari 1977 te 16.00 uur

door

FRANCISCUS CORSTIAAN ARNOLD GROEN

natuurkundig ingenieur,
geboren te Wateringen

1977

Dutch Efficiency Bureau – Pijnacker

Dit proefschrift is goedgekeurd
door de promotor
PROF.DR.IR. C.J.D.M. VERHAGEN

Aan mijn ouders

CONTENTS

Chapter 1

GENERAL INTRODUCTION

1.1 SURVEY

The investigation of automated chromosome analysis is important. Not only because computer assisted karyotyping may be faster and more accurate than manual karyotyping but also because quantitative measurements give the possibility to detect statistically significant aberrations from the normal chromosome. A short introduction to chromosome analysis is given in Appendix A.

The choice which chromosome features have to be measured depends strongly on the staining method used. The first chromosome measurements were length measurements. These length measurements were based on the contour of a chromosome (Gallus et al. (1970), Ledley et al. (1964, 1965, 1966a, 1966b, 1968, 1969, 1972), Neurath et al. (1966, 1969)) or on the integrated density profile of a chromosome (Rutovitz (1967)). Which method is preferred depends partly on the shape of the chromosomes, which is influenced by the preparation technique used. For instance when both chromatids are close together, the profile method is most suitable.

The chromosomes contract during the metaphase. The contraction differs not only from cell to cell, but also within one cell from chromosome to chromosome. Even for one chromosome the contraction between the long arm and the short arm may differ (Gaillard (1970) and Fitzgerald (1965)). When the contour method is applied, the position of the arm ends depends on the definition of the contour of a chromosome. So length measurements are of limited use.

DNA based features are independent of contraction. DNA contents of chromosomes have been measured by Mendelsohn et al. (1966, 1969, 1973) Mayall (1974), Van der Ploeg et al. (1974), Bosman (1976). To compute DNA based features, the chromosomes are stained with a DNA specific dye. The integrated optical density is a measure for the mass of the chromophore present. The accuracy of these measurements is affected by several sources of error, due to the specimen and the measuring system.

The main emphasis in this thesis has been laid on the algorithms for the computation of DNA based features. In most of the final experiments, the cytologist used banding patterns for identification of the chromosomes (Caspersson et al. (1968)). To that end the chromosomes were stained with atebrine, before the DNA specific Feulgen staining. Banding patterns will be a subject of further quantitative research in the Pattern Recognition Group at the Applied Physics Department of the Delft University of Technology.

In this thesis algorithms are given to compute DNA based features. The accuracy attained in these DNA based features is compared to the accuracy in length measurements, which is influenced by the contraction. The length measurements were based on the integrated profile. In addition to the accuracy study also classification results with these features are given. Primarily we have tried to give a critical evaluation of some measuring, computing and pattern recognition techniques to problems associated with the field of chromosome analysis. Some of the results apply to a larger field of research.

Although banding patterns are a powerful tool, they are still dependent on the contraction of the chromosomes. DNA specific staining procedures give the possibility to compute DNA based features, which are independent of the contraction. Combining a banding technique with a DNA specific staining procedure on the same metaphase enables investigation of DNA based features of chromosomes already accurately classified according to the banding pattern.

Our research started with some preliminary investigations on the measurement of chromosome lengths based on the position of the arm ends and the centromere (Groen (1971)). These arm ends were computed from the curvature of the contour. The investigations led to a critical evaluation of the problems in the measurement of contours and their curvature, which is reported in chapter 2. The methods of Gallus/Aalderink and Ledley are compared and the errors in the curvature and the position of the arm ends of an artificial chromosome are given. To this end probability density functions of Freeman codes have been derived.

In chapter 3 a description is given of a program (CHRDNA) and corresponding subroutines to compute DNA based features. This program locates the chromosomes in complete scanned metaphases and computes the DNA content and the integrated density profile of the individual chromosomes. From this profile the DNA ratio length and centromeric index are computed. The algorithms used, are described and evaluated.

In chapter 4 several sources of error in the DNA measurement are investigated. The distributional error and the quantization error of a linear and of a logarithmic scale are examined in further detail. The experimentally determined errors in the features due to the scanning, photography and the homologue variations are given.

In chapter 5 the classification results with these features are given. The influence of the number of homologue pairs is investigated.

1.2 MATERIALS AND METHODS

The preparation, staining and scanning of metaphases were performed at the Department of Histochemistry and Cytochemistry of the University of Leyden. Details of the procedure are given by Van der Ploeg et al. (1974). A diagram of the procedure is given in figure 1.1.

CHEMICAL PREPARATION

Whole venous blood was obtained from healthy volunteers and cultured for 68-70 hours, after which colcemid was added to block the dividing cells in the metaphase (Bosman et al. (1975)). After centrifugation and osmotic expansion the cells were placed on object glasses and dried.

For the final measurements, the chromosomes were first stained with atebrine, according to Caspersson's technique and microphotographed. The atebrine was washed away by fixation in methanol: formaldehyde 35%: glacial acetic acid (85:10:5 volume parts). After hydrolysis, Feulgen staining was carried out with Schiff reagent prepared according to Duijndam et al. (1973) and the preparations were microphotographed again.

PHOTOGRAPHY

The metaphases were photographed on 35 mm Copex Ortho Rapid (Agfa-Gevaert) or Kodak High Contrast film, using a Dialux (Leitz) microscope with a Leica MDx camera body. In order to excite atebrine fluorescence, light of about 440nm was used. For Feulgen photography the light was filtered with an AL 559 filter. A grey-wedge was photographed together with the chromosomes for calibration purposes and to check whether the densities of the photographic negative were in the linear part of the Hurter-Driffield curve or not. (Den Tonkelaar et al. (1964)). The negatives were embedded in immersion oil between glass slides and scanned.

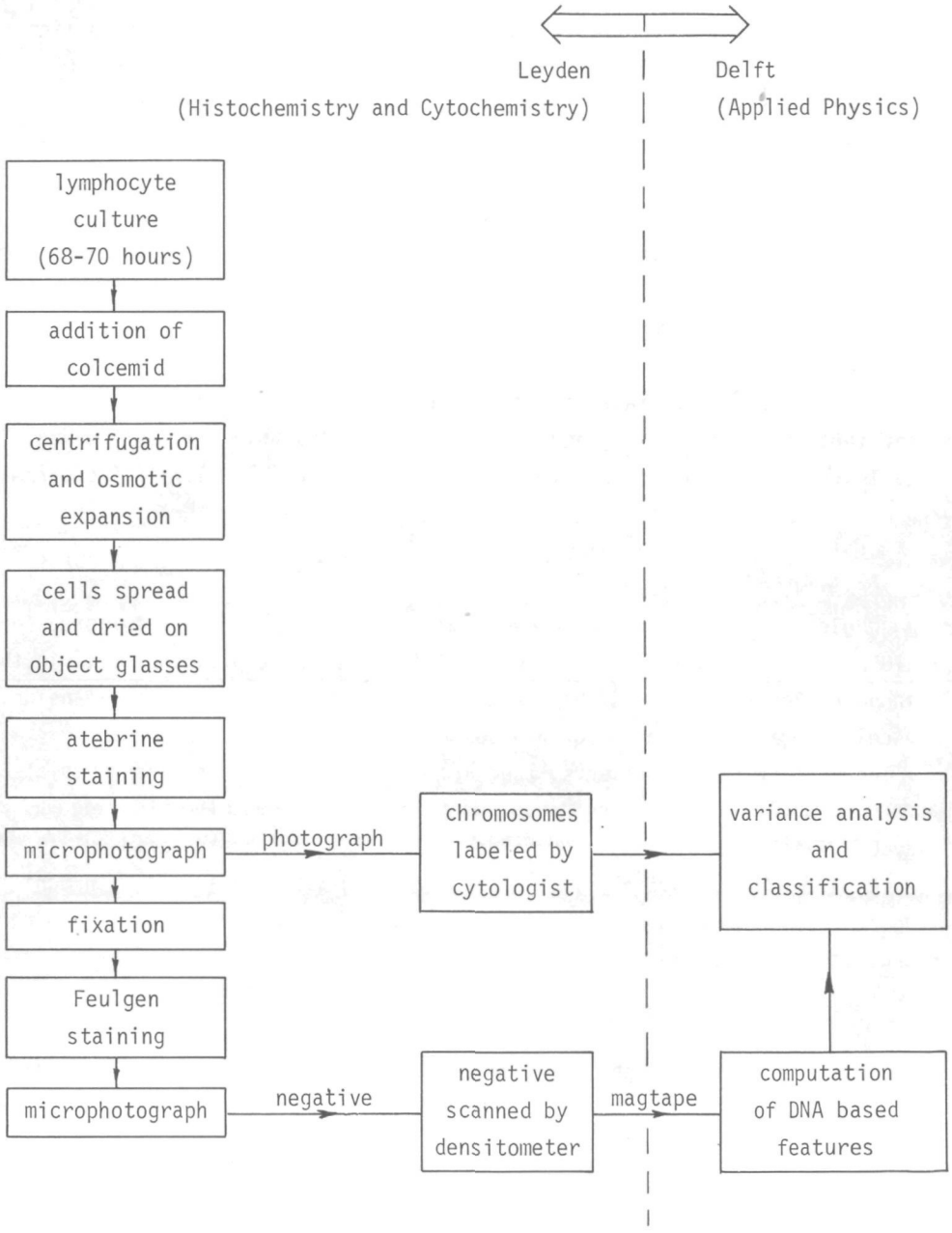


Figure 1.1 Diagram of the measurement procedure

SCANNING PROCEDURE

The microphotographs contain the image of a complete metaphase. The microphotographs of Feulgen stained metaphases are scanned with a SMP (Zeiss) cytophotometer interfaced to a PDP-12 computer. The scanning stage of the SMP cytophotometer has a stepsize of 10 μm with 200 steps per second. The intensities are measured at intervals, which are multiples of the stepsize. The diameter of the measuring diaphragm equals the measuring interval chosen. The illuminated field has a diameter of about 2 times the measuring diaphragm to reduce stray light errors. The intensities are quantized in 512 linear grey levels (9 bits).

The scanning is executed under control of the HISPAT program (Van der Ploeg et al. (1974)). The area scanned is a rectangle. Only one side of the rectangle is limited (320 points). When the metaphase image is too large, it is scanned in a number of overlapping rectangles.

The measured intensities are converted to densities and stored on 9 track magtape.

COMPUTATIONAL PROCEDURE

The computation is performed at the IBM 370/158 computer of the Delft University of Technology. The magtapes with scanned metaphases are analyzed with the CHRDNA program, which is discussed in detail in chapter 3. This program delivers DNA based features of the chromosomes.

For classification purposes and for the computation of the homologue variations, the chromosomes are labeled by their chromosome number (Paris conference, Hamerton (1973)). This labeling for the final experiments was done by cytologists according to the microphotographs of the atebrine stained metaphases. This label is added to the measured features.

The file system consists of the original scans of the metaphases, the computed profiles (projections of the densities on a principal axis or on a best fit polynomial) and the measured features.

Chapter 2

SOME QUANTITATIVE ASPECTS OF THE ANALYSIS OF CURVATURE MEASUREMENT

2.1 INTRODUCTION

The methods of Gallus (1970), Aalderink (1970) and Ledley (1964, 1965, 1966a, 1966b, 1968, 1969, 1972) to measure the curvature of quantized curves are investigated in this chapter. The influence of the contour-tracing algorithm on the measurement of the curvature has been discussed by Bennett et al. (1975), based on the noise characteristics of the frequency domain. The contour-tracing algorithm which we used (8 neighbour connectivity) appeared to have a good signal to noise ratio in Bennett's experiments.

In order to compare the measured curvature and the real curvature, we need a curve from which the real curvature can be computed. A function which more or less resembles the contour of a small chromosome is used, because one of our important applications of curvature measurement is chromosome analysis. The function ('analytical chromosome') is

$$r(\theta) = 1 + 0.25 \cos \omega\theta \qquad (2.1)$$

in which (r,θ) are polar coordinates and ω is a parameter. The function is plotted in figure 2.1. It resembles a small acrocentric chromosome for $\omega = 3$, and it shows some resemblance to a small median chromosome for $\omega = 4$. The curvature of this function has a symmetric character. When we introduce higher harmonics with a certain phase in $r(\theta)$, asymmetric shapes can be obtained as well.

A brief summary of the concepts of curvature is given here.
An extensive discussion has been given by Kreyszig (1959). Let $\underline{c}$ be a curve in $\mathcal{R}_3$ given in the allowable parametric representation:

$$\underline{c} = (x, y, z) = \underline{c}(t) \qquad (2.2)$$

with allowable parameter t.

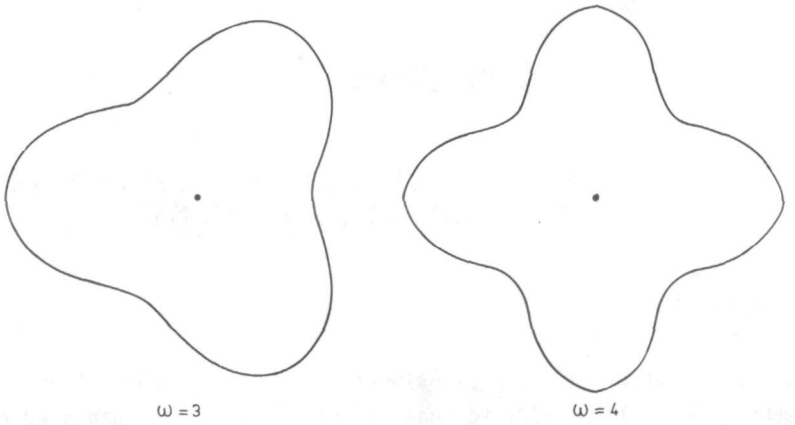


Figure 2.1 Analytical chromosome ($r(\theta) = 1 + 0.25 \cos \omega\theta$)

The length s(t) of an arc of $\underline{c}$ is

$$s(t) = \int_{t_o}^{t} \sqrt{(\underline{c}' \cdot \underline{c}')}\, dt \tag{2.3}$$

with $\underline{c}' = \dfrac{d\underline{c}}{dt}$, and t_o an arbitrary starting point.

The tangent $\underline{t}(s)$ to the curve $\underline{c}$ is defined as

$$\underline{t}(s) = \frac{d\underline{c}}{ds} \equiv \dot{\underline{c}}(s) \tag{2.4}$$

and the curvature $\kappa(s)$ of a curve is given as

$$\kappa(s) = |\dot{\underline{t}}(s)| = |\sqrt{\ddot{\underline{c}}(s) \cdot \ddot{\underline{c}}(s)}|. \tag{2.5}$$

We will restrict ourselves to a plane curve in a two-dimensional Euclidian space $\mathcal{R}_2$, and derive an expression for κ and the angle ϕ between the X-axis and the tangent vector $\underline{t}$ to the curve; now

$$\begin{aligned} \dot{\underline{c}}(s) &= (\dot{x}, \dot{y}) = (\cos \phi, \sin \phi) \\ \ddot{\underline{c}}(s) &= (\ddot{x}, \ddot{y}) = (-\dot{\phi} \sin \phi, \dot{\phi} \cos \phi). \end{aligned} \tag{2.6}$$

Using relation (2.5) we obtain

$$\kappa(s) = |\sqrt{\dot{\phi}(s)^2}| = |\dot{\phi}(s)| \tag{2.7}$$

So $\kappa(s)$ is the absolute value of $\dot{\phi}(s)$. The sign of $\dot{\phi}(s)$ determines whether the curve at position s is convex or concave.

2.2 QUANTIZED CURVES AND THEIR FREEMAN CODE

When a curve has to be processed by a digital computer it must be quantized. A square grid is superimposed on the curve. The intersections of curve and grid divide the curve into a large number of curve segments. For each intersection there are two grid nodes, one on either side of the curve. It depends on the quantization method used, which node is marked to be a point of the quantized curve.

Freeman (1961a, 1962, 1969) suggests the Grid Intersect Quantization (GIQ). In this method the node closest to the intersection is marked as a point of the quantized curve (see figure 2.2a). When the curves are the boundaries of objects an Object Boundary Quantization (OBQ) is used. See e.g. Gallus et al. (1970), Ledley et al. (1964, 1965, 1966a, 1966b, 1969), Aalderink (1970) and Freeman (1970). In this method the node which belongs to the object is marked as a point of the quantized curve (see figure 2.2b). Instead of marking the object nodes, the background nodes could be marked as well. This Background Boundary Quantization (BBQ) is illustrated in figure 2.2c.

In these quantization methods we must make the restriction, that the curve is quantized fine enough. It is not allowed, that a curve passes more than once between two neighbouring nodes of the grid. In the case illustrated in figure 2.3, such information will be lost in the quantization process. This matter is further discussed in Appendix B.

Young (1974) derived a quantization theorem for curves. This theorem requires for a curve with maximal curvature κ_{max} a grid constant

$$h \leqslant \frac{\pi}{\kappa_{max}} . \tag{2.8}$$

The intersections between the grid and the curve divide the curve into curve segments. The quantized curve consists of line elements from node to node. These are called Freeman vectors. A curve segment may be associated with a

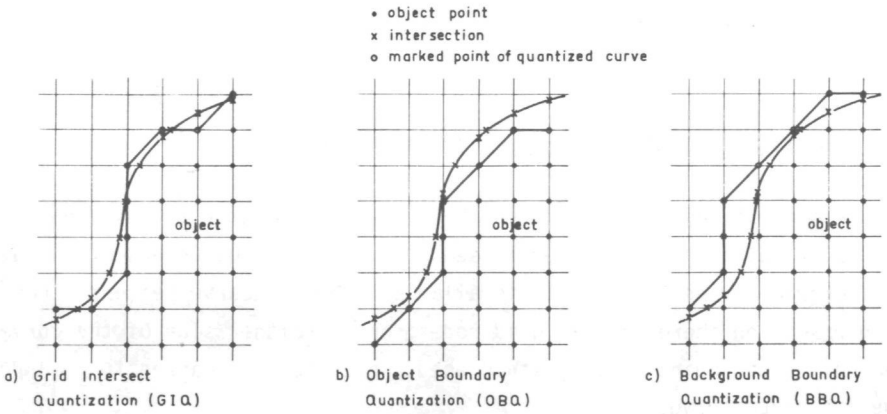


Figure 2.2 Curve quantization

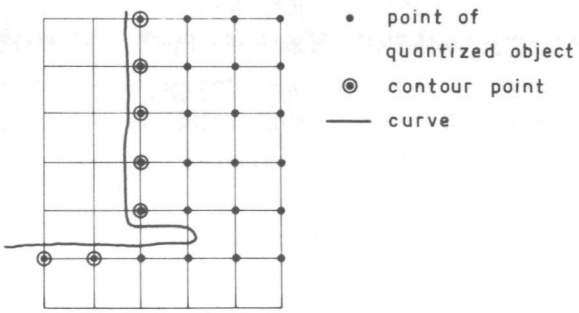


Figure 2.3 Not admitted case. The curve passes more than once between two nodes

Freeman vector or may reduce to a single node. This can be seen in figure 2.2 and will be further discussed in section 2.5.

A Freeman vector i can be presented by an integer f(i) from 0 to 7, where $\Phi = \frac{1}{4}\pi f(i)$ is the angle between the X-axis of the coordinate system of the grid and the Freeman vector. This code f(i) is given in table 2.1 and in figure 2.4. Properties of the Freeman code are given by Freeman (1961a, 1961b, 1962, 1969, 1970). Using the Freeman code one can represent a quantized curve by a string of numbers between 0 and 7. Freeman coding of straight lines has been discussed by Brons (1974).

Table 2.1 Freeman code

Freeman vector (x,y)	Freeman code
(1, 0)	0
(1, 1)	1
(0, 1)	2
(-1, 1)	3
(-1, 0)	4
(-1,-1)	5
(0,-1)	6
(1,-1)	7

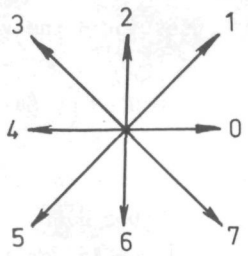


Figure 2.4 Freeman code

2.3 METHODS TO DETERMINE THE CURVATURE OF A QUANTIZED CURVE

Gallus et al. (1970) and Aalderink (1970) describe a method to determine concave and convex parts of a quantized contour by determining the difference f'(i) between two successive Freeman vectors f(i) and f(i+1) obtained by the method of OBQ. When f'(i) is out of the range [-3,4], due to the discontinuity between the Freeman code values 0 and 7, we have to add or subtract 8 so that f'(i) is made to lie within this range.

$$f'(i) = f(i+1) - f(i), \qquad -3 \leq f'(i) \leq 4. \tag{2.9}$$

The value f'(i) = -3 will never occur because of the Object Boundary Quantization process. This is explained in Appendix B.

The difference f'(i) is smoothed to suppress the noise caused by the quantization process. (This filter process may be described by z transforms. This gives no link, however, with the method of Ledley, investigated later on).

The smoothed signal v(i) is

$$v(i) = \sum_{j=-\infty}^{\infty} w(j)\, f'(i-j), \text{ in which} \tag{2.10}$$

$$\begin{aligned} w(j) &= 0 && \text{if} \quad |j| \geqslant B+n \\ w(j) &= (B+n-|j|)/A && \text{if} \quad B < |j| \leqslant B+n-1 \\ w(j) &= n/A && \text{if} \quad 0 \leqslant |j| \leqslant B \end{aligned} \tag{2.11}$$

and

$$A = n(n+2B).$$

Gallus uses for B the value 0 (triangular filter).
Aalderink uses for B the value 1 (trapezium filter).

The value of n used by Gallus and Aalderink depends on the number of points of the curve and is determined by the required frequencies in the signal.

A graphical representation of w(j) is given in figure 2.5a. The DFT of w(j), $W(\rho)$, is given in figure 2.5b.

A filter w(j) working on the difference of a variable may be described as a filter w'(j) working on the variable itself because, according to (2.9) and (2.10)

$$v(i) = \sum_{j=-\infty}^{\infty} w(j)\, f'(i-j) = \sum_{j=-\infty}^{\infty} w(j)\, f(i-j+1) - \sum_{j=-\infty}^{\infty} w(j)\, f(i-j) =$$

$$= \sum_{j=-\infty}^{\infty} w'(j)\, f(i-j) \tag{2.12}$$

$$\text{with } w'(j) = w(j+1) - w(j). \tag{2.13}$$

Using formulas (2.11) and (2.13) we obtain for the method of Gallus/Aalderink (G/A) for w'(j)

$$\begin{aligned} w'(j) &= 0 \quad \text{if} \quad j < -B-n \\ & \qquad\qquad -B \leqslant j \leqslant B-1 \\ & \qquad\qquad j > B+n-1 \end{aligned} \tag{2.14}$$

$$\begin{aligned} w'(j) &= 1/A \quad \text{if} \quad -B-n \leqslant j \leqslant -B-1 \\ w'(j) &= -1/A \quad \text{if} \quad B \leqslant j \leqslant B+n-1 \end{aligned}$$

$$A = n(n+2B).$$

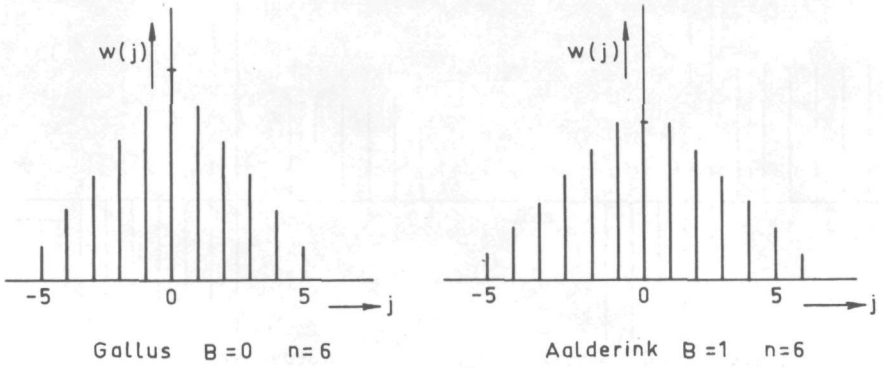


Figure 2.5a $w(j)$ *for the method of Gallus and Aalderink*

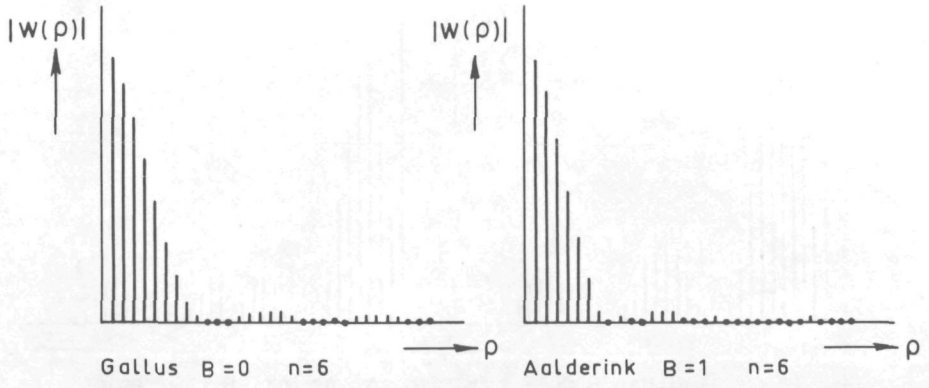


Figure 2.5b $|W(\rho)|$ *for the method of Gallus and Aalderink*

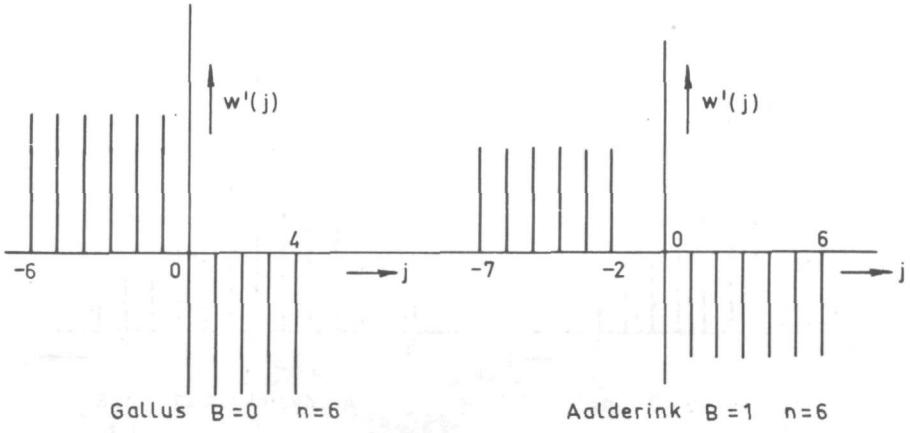


Figure 2.6a w'(j) *for the method of Gallus and Aalderink*

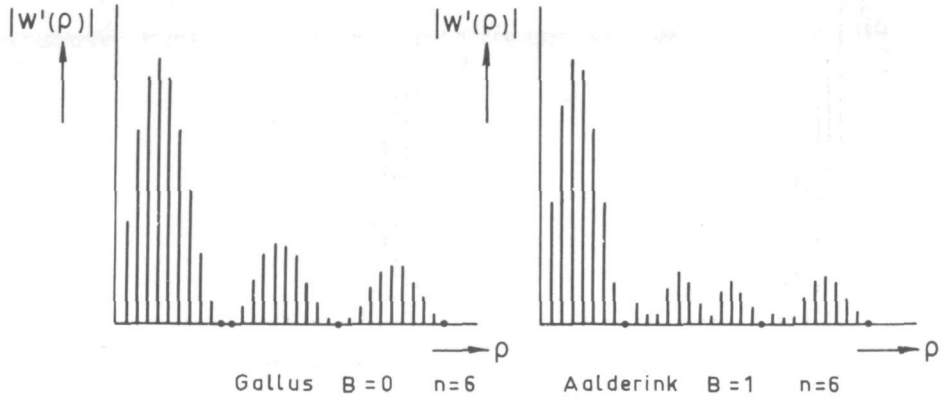


Figure 2.6b |W'(ρ)| *for the method of Gallus and Aalderink*

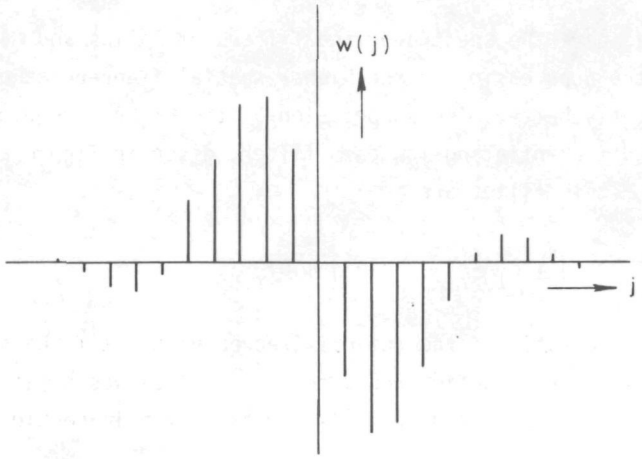


Figure 2.7a Coefficients of a differentiating low-pass filter ($B_1 = \frac{1}{4}$)

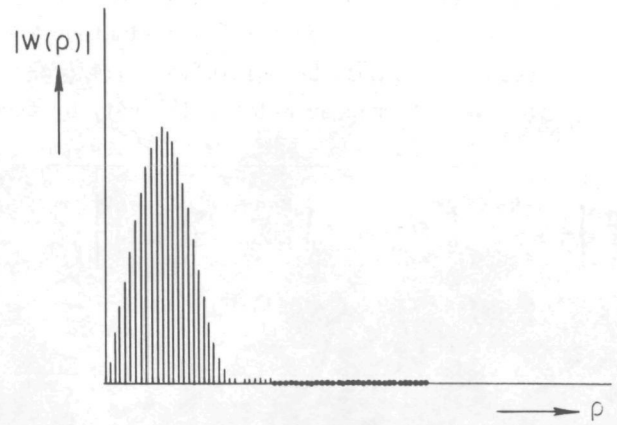


Figure 2.7b $|W(\rho)|$ *for a differentiating low-pass filter* ($B_1 = \frac{1}{4}$)

A graphical representation of w'(j) is given in figure 2.6a. The DFT of w'(j), W'(ρ) is given in figure 2.6b.

When we compare the spectra of the filters of Gallus and Aalderink (given in figure 2.6b) the suppression of the higher spatial frequencies of Aalderink's filter is slightly better. The suppression of the higher frequencies is poor compared to a differentiating low-pass filter, given in figure 2.7. The coefficients of this filter are

$$w(j) = \frac{B_1}{j} \cos j\pi B_1 - \frac{1}{\pi j^2} \sin j\pi B_1 \qquad (2.15)$$

in which B_1 is the ratio of the cut-off frequency and half the sampling frequency. A triangular windowing function was used. This filter was constructed according to Oppenheim (1975). Results of this filter are given in section 2.7.

The smoothed difference v(i) of the successive Freeman codes filtered according to Gallus/Aalderink can be described as

$$v(i) = \frac{1}{A}\left[\sum_{j=-B-n}^{-B-1} f(i-j) - \sum_{j=B}^{B+n-1} f(i-j)\right] . \qquad (2.16)$$

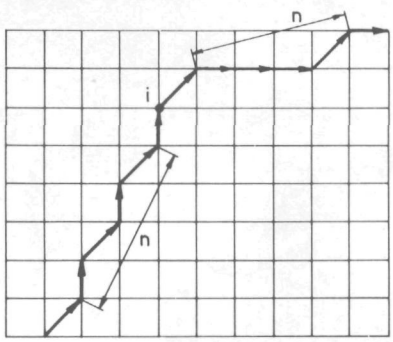
To obtain the angular direction $\Phi(i)$ of the Freeman vector, the Freeman code f(i) is multiplied by $\pi/4$. The smoothed difference v(i) is derived as a function of the number of Freeman code values. In Appendix C the real curve length associated with this number of codes is approximated. When this real curve length is taken into account v(i) must be multiplied by $4/(\pi\sqrt{2}h)$. An approximation for the curvature $\dot{\phi}_{G/A}(i)$ is obtained by multiplying v(i) by these two factors ($\pi/4$ and $4/\pi\sqrt{2}h$).

$$\dot{\phi}_{G/A}(i) = \frac{1}{A'}\left[\frac{1}{n}\sum_{j=-B-n}^{-B-1} \frac{\pi}{4} f(i-j) - \frac{1}{n}\sum_{j=B}^{B+n-1} \frac{\pi}{4} f(i-j)\right] \qquad (2.17a)$$

and

$$A' = \frac{(n+2B)\pi\sqrt{2}\,h}{4} . \qquad (2.17b)$$

A' can be interpreted as the approximated curve length between the centres of the leading and trailing curve segment. Equation (2.17a) denotes that the method of Gallus and Aalderink computes the average angular directions of the n Freeman vectors of the leading curve segment [i+B, i+B+n] and of the trailing curve segment [i-B-n, i-B] . The difference is taken as approximation for the curvature. In figure 2.8 an example of Aalderink's method is given for n=4, B=1, and h=1.



$$\dot{\phi}_A(i) = \frac{1}{\sqrt{2}.4.6}\,(0+0+0+1-(2+1+2+1)) = -0.15\ ^{rad}/_{m}$$

Figure 2.8 Example of Aalderink's method

The method of Ledley (1964, 1965, 1966a, 1966b) is based on curve segments with a centre, a trailing and a leading vector, illustrated in figure 2.9. Ledley approximates the curvature $\kappa_L(i)$ of the segment by Θ/L. The angle between the leading and the trailing vector is Θ and the length of the curve segment is L. This curvature is

$$\kappa_L(i) = \frac{\Theta}{L} = \frac{1}{L}\,(\phi_H - \phi_T) \qquad (2.18)$$

in which ϕ_H is the angle between the leading vector and the X-axis and ϕ_T is the angle between the trailing vector and the X-axis.

We use quantized curves, so length is expressed in the number of Freeman vectors. When the approximated real curve length associated with this number of

vectors is taken into account, we have to multiply the number of vectors by $\frac{1}{4}\pi\sqrt{2}\,h$ (Appendix C). The number of Freeman vectors of the leading and of the trailing vector is n; B is the number of Freeman vectors from the centre to the head of the trailing vector and also from the centre to the tail of the leading vector.

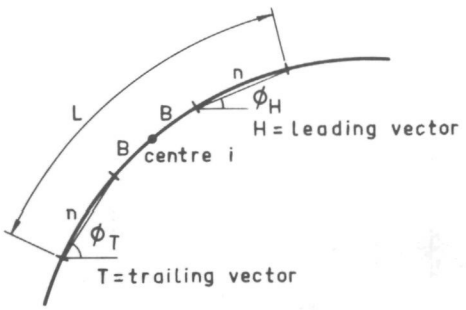


Figure 2.9 Contour segment according to Ledley

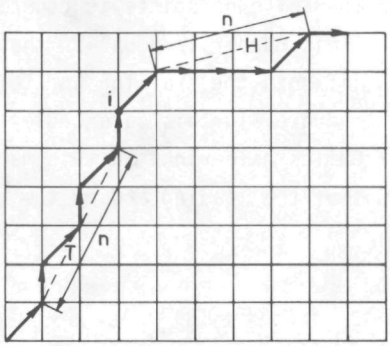
We assume that the angular direction of the leading or trailing vector is an approximation of the angular direction of the tangent to the curve in the middle of the curvesegment defined by the leading or trailing vector. The approximated curvelength A' between these two middles is given in equation (2.17b). Amending Ledley's method, we divide the difference between the angular directions of the leading and trailing vector by this distance to obtain the first order difference. For the curvature approximation $\dot{\phi}_L$ of Ledley we obtain

$$\dot{\phi}_L(i) = \frac{1}{A'} (\phi_H - \phi_T). \qquad (2.19)$$

In figure 2.10 an example of this method is given for h=1, B=1 and n=4.

When we compare the method of Gallus/Aalderink (2.17) to the method of Ledley (2.19) we see that the essential difference is the measurement of the angular direction of the leading and of the trailing vector. In the method of Gallus/Aalderink the angle is determined by the average angle between the n Freeman vectors and the X-axis. In Ledley's method, the angle between the vector itself (the chord) and the X-axis is computed.

In more recent literature Ledley (1968, 1969, 1972) approximates the angle



$$\dot{\phi}_L(i) = \frac{4}{\pi\sqrt{2.6}}\left(\arctan\frac{1}{4} - \arctan\frac{4}{2}\right) = -0.13\ ^{rad}/_{m}$$

Figure 2.10 Example of the method of Ledley

between the leading and the trailing vector by laying these vectors on a grid with their tails at the origin. The number of coordinate points which must be traversed along the outside of the rectangular grid in going from the head of the trailing vector to the head of the leading vector is counted. When the outside is traversed counter-clockwise the count is positive, when the outside is traversed clockwise the count is negative. In figure 2.11 this method is applied to the example of figure 2.10.

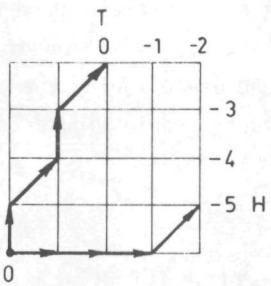


Figure 2.11 Approximation of Ledley

We assume that the curve segment defined by the leading or trailing vector may be approximated by a straight line. In this case the head of the leading and of the trailing vector will always be on a $(2n+1) \times (2n+1)$ square.

This square is given in figure 2.12a for n=4. An edge point of the square is uniquely coded by the sum of the Freeman code values from the origin to the edge point. In the approximation of Ledley the number of edge points is counted between the trailing and the leading vector. This number is equal to the difference in code value of the edge points, if both the trailing and leading vector reach the edge of the (2n+1) × (2n+1) square. In this case Ledley's approximation is identical to the method of Gallus/Aalderink. In this method the difference in the average Freeman code values of the leading and of the trailing vector is computed.

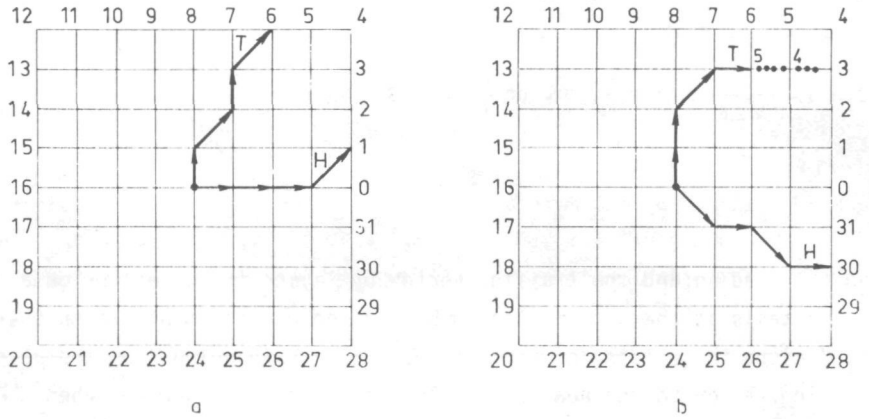


Figure 2.12 Square of edge points for n=4

When the contour is much curved, the leading and the trailing vector do not necessarily end up at the edge of the square. In the approximation of Ledley we partly count inside the square in this case instead of following the outside. This introduces an additional error, which is illustrated in figure 2.12b for the trailing vector T.

2.4 ERRORS PRESENT IN CURVATURE APPROXIMATION

We will investigate the errors in the method of Gallus/Aalderink and Ledley for the approximation of the curvature. The second method of Ledley will be left out of consideration because this method is identical to the method of Gallus/Aalderink for slightly curved contours. Part of the problem is closely related to numerical differentiation (Hildebrand (1956)).

We can distinguish two types of errors:

a) The error caused by the quantization of the grid, discussed in section 2.5.

b) The error caused by the approximation, which is related to the error in numerical differentiation. Finite differences are used instead of infinite small ones. This error is discussed in section 2.6.

In Ledley's method we compute the angular directions of the leading and trailing vector. The quantization error is due to the fact that the end points of these vectors are grid points, which generally are not located on the curve. For the second error we consider the original curve discarding the quantization error. As the curve may be locally described by a continuous and differentiable function $y = y(x)$, we can apply the mean-value theorem. So the angular direction of the chord equals the angular direction of the tangent to the curve in a point on the curve segment bounded by the leading (or trailing) vector. Discarding the quantization error we may write instead of equation (2.19)

$$\tilde{\dot{\phi}}_L(i) = \frac{1}{A'} \left[\phi(\xi_2) - \phi(\xi_1)\right] \tag{2.20}$$

and

$$\begin{aligned} s(i - B - n) &\leq \xi_1 \leq s(i - B) \\ s(i + B) &\leq \xi_2 \leq s(i + B + n). \end{aligned}$$

$s(i)$ is the position at the original curve, where the curve enters the grid element associated with contour point (i).

In the method of Gallus/Aalderink we sum the directions of n Freeman vectors of the leading and trailing vector. This may be described as summing n angular directions of the curve in the appropriate intervals, only roughly quantized in 8 values by the Freeman code, introducing the quantization error (a). For the second error (b) we again look at the original curve, discarding the quantization error. So for the leading and the trailing vector we average n angular directions of the original curve for those curve segments, which lead to a Freeman vector (section 2.2). Discarding the quantization error we may write instead of equation (2.17a)

$$\tilde{\dot{\phi}}_{G/A}(i) = \frac{1}{A'} \left[\frac{1}{n} \sum_{j=-B-n}^{-B-1} \phi(\zeta_{i-j}) - \frac{1}{n} \sum_{j=B}^{B+n-1} \phi(\zeta_{i-j}) \right]. \tag{2.21}$$

in which ξ_{i-j} is a point on the curve segment defined by the (i-j)th Freeman vector.

As the angular direction ϕ of the original curve is a continuous function of the arc length, we can again apply the mean-value theorem. So the average value of ϕ equals the value of ϕ in a point on the curve segment bounded by the leading (or trailing) vector. Equation (2.21) can be written as

$$\tilde{\phi}_{G/A}(i) = \frac{1}{A'}[\phi(\xi_4) - \phi(\xi_3)]$$

and (2.22)

$$s(i-B-n) \leq \xi_3 \leq s(i-B)$$
$$s(i+B) \leq \xi_4 \leq s(i+B+n).$$

Thus formula (2.17a) and formula (2.19) have been reduced to the same form. When we discard the quantization error, in both methods the curvature is approximated by the difference in angular direction of the tangent to the curve at two points ξ_1 and ξ_2 or ξ_3 and ξ_4. The points ξ_1 and ξ_2 or ξ_3 and ξ_4 are located at the curve segments defined by the trailing and leading vector. Generally $\xi_1 \neq \xi_3$ and $\xi_2 \neq \xi_4$.

2.5 QUANTIZATION ERRORS (OBQ) IN CURVATURE APPROXIMATION

A contour is divided by the intersections with the grid into a large number of curve segments. This can be seen in figure 2.2b for the Object Boundary Quantization process. A curve segment is not always represented by a Freeman vector, as it may reduce to a single node as well.

In figure 2.13 the possible cases are given for the Object Boundary Quantization process and a clockwise contour-tracing algorithm. We assumed that the grid is fine enough so that in one grid element the contour may be approximated by a straight line with angular direction φ (with the X-axis of the grid). The grid elements marked 1a, 2a, 3a and 4a do not lead to a Freeman code, the grid elements marked 1b, 2b, 3b and 4b have an even Freeman code and the grid elements marked 1c, 2c, 3c and 4c have an odd Freeman code. Rotation by a multiple of 90^0 transforms all cases a into each other. The same is valid for all cases b and c. The cases occurring for a certain value of φ are given in table 2.2.

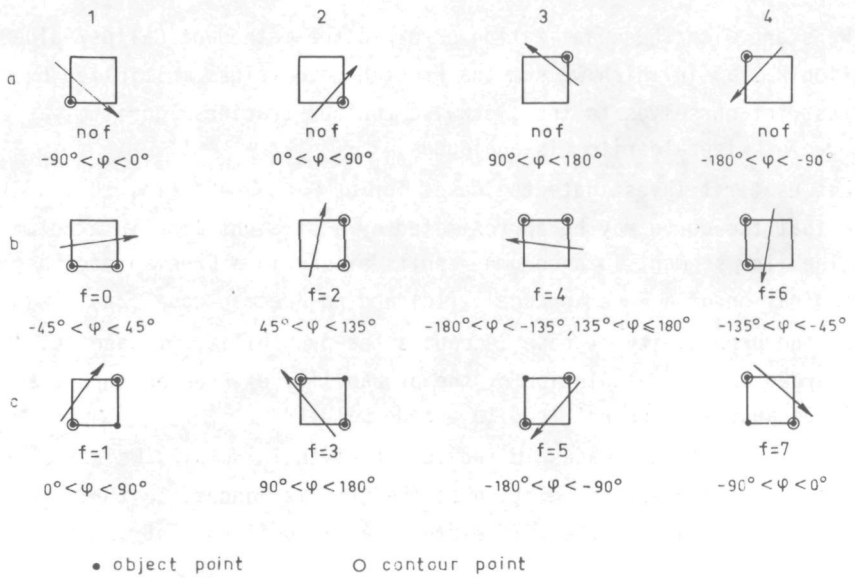


Figure 2.13 Clockwise OBQ contour-tracing algorithm

Table 2.2 Situations occurring for a given φ

φ	case
$\varphi = 0^{\circ}$	1b
$0^{\circ} < \varphi < 45^{\circ}$	1b, 1c, 2a
$\varphi = 45^{\circ}$	1c, 2a
$45^{\circ} < \varphi < 90^{\circ}$	1c, 2a, 2b
$\varphi = 90^{\circ}$	2b
$90^{\circ} < \varphi < 135^{\circ}$	2b, 2c, 3a
$\varphi = 135^{\circ}$	2c, 3a
$135^{\circ} < \varphi < 180^{\circ}$	2c, 3a, 3b
$\varphi = 180^{\circ}$	3b
$-180^{\circ} < \varphi < -135^{\circ}$	3b, 3c, 4a
$\varphi = 135^{\circ}$	3c, 4a
$-135^{\circ} < \varphi < -90^{\circ}$	3c, 4a, 4b
$\varphi = -90^{\circ}$	4b
$-90^{\circ} < \varphi < -45^{\circ}$	4b, 4c, 1a
$\varphi = -45^{\circ}$	4c, 1a
$-45^{\circ} < \varphi < 0^{\circ}$	4c, 1a, 1b

GALLUS/AALDERINK

We start with the quantization error in the method of Gallus/Aalderink (equation 2.17a) in which we sum the Freeman code values multiplied by $\pi/4$. We will restrict ourselves to the clockwise contour-tracing algorithm, as the counter-clockwise algorithm is analogous.

Let us first investigate the cases for $0^0 < \varphi < 45^0$ (1b, 1c and 2a). We assume that the curve may be approximated by a straight line in a column of the grid. The line segment in a column results either in a Freeman code 0 (1b) or in the combination of a Freeman code 1 (1c) and no Freeman code (2a).

So the probability of case 2a equals the probability of case 1c. The case 2a is discarded in the calculation of the probability of Freeman code 0 and Freeman code 1, because it does not lead to a code value.

In figure 2.14 the cases 1b and 1c are given in detail. We assume that the position of the contour in relation to the grid is random, so that the position y, where the contour enters the grid element has an uniform distribution p(y).

$$p(y) = \frac{1}{h} \quad \text{if} \quad 0 \leqslant y \leqslant h. \tag{2.23}$$

The probability that for a given φ ($0^0 < \varphi < 45^0$) a Freeman code value 1 will occur is:

$$p(f = 1|\varphi) = p(0 \leqslant x \leqslant h) = \int_0^{h \tan \varphi} p(y)dy, \tag{2.24}$$

$0^0 < \varphi < 45^0$.

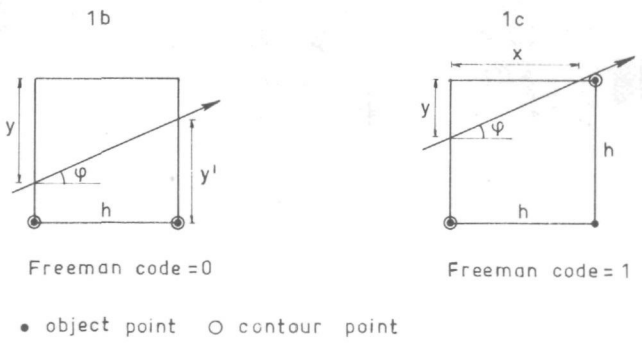


Figure 2.14 Two cases for $0^0 < \varphi < 45^0$

This gives

$$p(f = 1|\varphi) = \tan\varphi \quad \text{and} \quad p(f = 0|\varphi) = 1 - \tan\varphi, \quad 0^o < \varphi < 45^o. \tag{2.25}$$

Only Freeman code value 1 is present for $\varphi = 45^o$. Only the value 0 is present for $\varphi = 0^o$.

When $45^o < \varphi < 90^o$ a line segment contained in a row results in a Freeman code 2 (2b) or in the combination of no Freeman code (2a) and Freeman code 1 (1c). We assume that the position x, at which point the contour leaves the grid element has an uniform distribution. In the same way as above we obtain

$$p(f = 1|\varphi) = \text{cotan}\,\varphi \quad \text{and} \quad p(f = 2|\varphi) = 1 - \text{cotan}\,\varphi,$$

$$45^o < \varphi < 90^o \tag{2.26}$$

and so on. The distributions of $p(f = 0|\varphi)$ and $p(f = 1|\varphi)$ are given in figure 2.15. The distributions of all even codes and of all odd codes are identical, but they are shifted a multiple of 90^o.

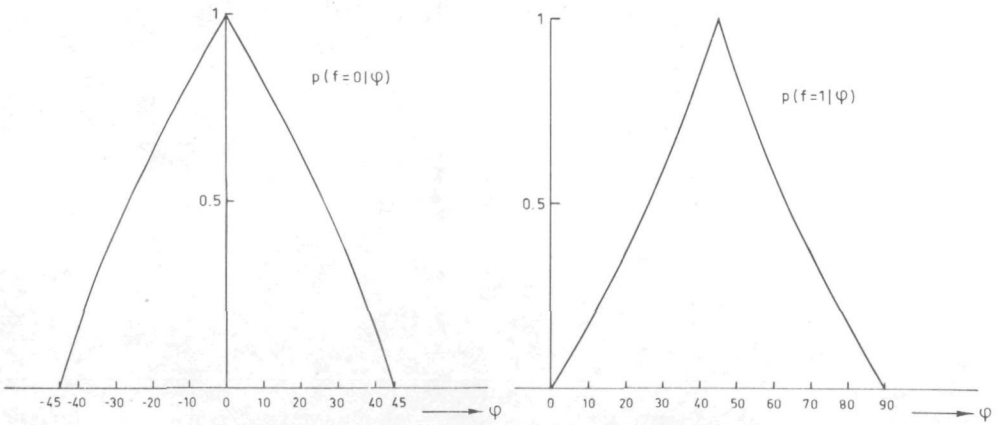


Figure 2.15 Distributions of $p(f = 0|\varphi)$ *and* $p(f = 1|\varphi)$

When the position of the contour in relation to the grid is random, the a priori probability $p(\varphi)$ that segmentation will create a curve segment with angular direction φ in a grid column (for $-45^o \leq \varphi \leq 45^o$, $-180^o < \varphi \leq -135^o$, $135^o \leq \varphi \leq 180^o$) or row ($45^o < \varphi < 135^o$, $-135^o < \varphi < 45^o$), is given by (Appendix C)

$$p(\varphi) = \frac{1}{8}\sqrt{2}\cos\varphi, \qquad -45^{o} \leq \varphi \leq 45^{o}. \tag{2.27}$$

So the a priori probabilities for an even and an odd Freeman code value can be computed as

$$p(f = \text{even}) = 4p(f=0) = \sqrt{2}\int_{0}^{\pi/4}(1-\tan\varphi)\cos\varphi\,d\varphi = 2 - \sqrt{2} = 0.5858$$

and (2.28)

$$p(f = \text{odd}) = 4p(f=1) = \sqrt{2}\int_{0}^{\pi/4}\tan\varphi\cos\varphi\,d\varphi = \sqrt{2} - 1 = 0.4142.$$

As the distributions for all even and odd codes are the same (only shifted), we will restrict ourselves to values of φ between 0^{o} and 45^{o}, because other functions can be obtained by shifting and reflection.

The bias in the expected value $\bar{\Phi}$ of the angular direction Φ of the Freeman vector for a given φ may be computed from equation (2.25) as

$$\text{bias}(\bar{\Phi}|\varphi) \equiv \mathcal{E}(\Phi - \varphi|\varphi) = \frac{\pi}{4}\tan\varphi - \varphi \qquad 0 \leq \varphi \leq 45^{o}. \tag{2.29a}$$

With the distribution $p(\varphi)$ given by equation (2.27) this bias is

$$\text{bias}(\bar{\Phi}) = \sqrt{2}\int_{0}^{\pi/4}\left(\frac{\pi}{4}\tan\varphi - \varphi\right)\cos\varphi\,d\varphi = \frac{\pi}{4}(\sqrt{2}-2) + (\sqrt{2}-1) =$$

$$= -4.59 \cdot 10^{-2}\ \text{rad} = -2.63^{o}. \tag{2.29b}$$

The variance for a given φ is

$$\text{var}(\Phi|\varphi) \equiv \mathcal{E}\{(\Phi-\bar{\Phi})^{2}|\varphi\} = \left(\frac{\pi}{4}\right)^{2}(1 - \tan\varphi)\tan\varphi, \quad 0 \leq \varphi \leq 45^{o}. \tag{2.30}$$

With the distribution $p(\varphi)$ given in equation (2.27) this variance can be computed as

$$\text{var}(\Phi) \equiv \mathcal{E}\{(\Phi-\bar{\Phi})^{2}\} = \frac{\pi^{2}\sqrt{2}}{16}[1-\ln(1+\sqrt{2})] = 0.1034\ \text{rad}^{2}. \tag{2.31}$$

The average angular direction of n Freeman vectors is used in the method of

Gallus/Aalderink as approximation of the angular direction of the tangent to the curve. Assuming, that the errors in the angular directions of the Freeman vectors are uncorrelated the variance in the curvature estimated by Gallus/Aalderink (equation 2.17a) is

$$\sigma^2_{G/A} = \mathrm{var}(\dot{\phi}_{G/A}) = \frac{\pi^2\sqrt{2}}{8nA'^2}\,[1-\ln(1-\sqrt{2})] \qquad (2.32a)$$

In the method of Gallus/Aalderink we take the difference between the leading and trailing vector. This difference introduces a factor 2 in equation (2.32a). In figure 2.16 $A'\sigma_{G/A}$ is plotted as function of n. In the case n=1, B=0, it is expected that the most correlated situation appears. Hence a new calculation has been performed for this case in Appendix D, where the correlation is taken into account.

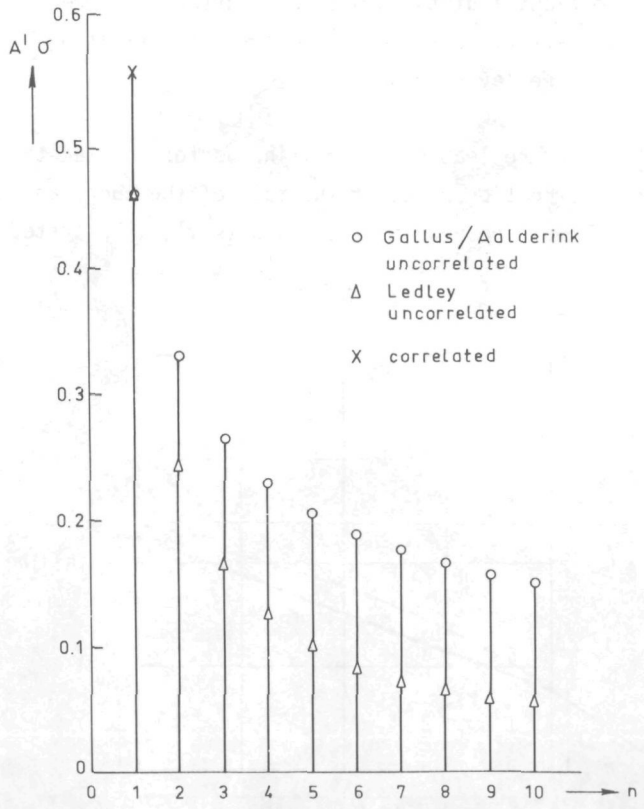


Figure 2.16 $A'\sigma$ *for the method of Gallus/Aalderink and Ledley*

In this correlated case the variance is

$$\sigma'^{2}_{G/A} = 2\left(\frac{\pi}{4A'}\right)^{2} [\sqrt{2} + 2 - \sqrt{10}] . \qquad (2.32b)$$

This results in an increase of about 20% in the standard deviation. This value of $A'\sigma'_{G/A}$ is also plotted in figure 2.16.

LEDLEY

In Ledley's method the angular directions Φ of the leading and the trailing vector are computed. We will again restrict ourselves to values of φ between 0^{o} and 45^{o}. Other cases can be obtained by shifting and reflection. When the curve segment length is n contour points, for slightly bent curves the horizontal distance between the head and the tail of the vector will be n times the grid constant h. This horizontal distance will be shorter for more sharply bent curves, introducing a larger error. So the error we are now estimating is the minimum error attainable with Ledley's method.

The angle between the leading or trailing vector (called the chord) and the X-axis is φ. The intersection point of the tail of the chord and the grid is y, the intersection point of the head of the chord is y', illustrated in figure 2.17.

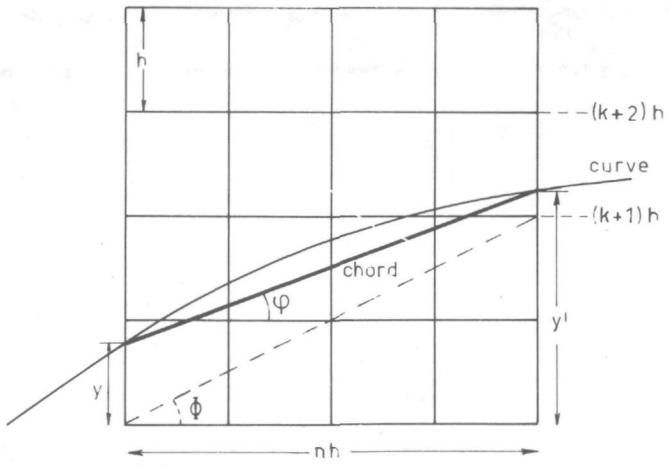


Figure 2.17 Quantization process in Ledley's method

The estimation of the angular direction Φ of the chord in Ledley's method is

$$\Phi = \arctan\left(\frac{\text{integer } (y')}{n}\right). \tag{2.33a}$$

For a given φ the intersection point y' is

$$y' = y + h\,n \tan\varphi. \tag{2.33b}$$

Assuming that y is uniformly distributed between 0 and h, y' will be uniformly distributed between $h\,n \tan\varphi$ and $h(n \tan\varphi + 1)$ for a given φ.

Let k be an integer, so that (k+1)h is in this interval

$$h\,n \tan\varphi \leqslant (k+1)h < h(n \tan\varphi + 1) \tag{2.33c}$$

Equation (2.33c) can be rewritten as

$$\arctan\frac{k}{n} < \varphi \leqslant \arctan\frac{k+1}{n}. \tag{2.33d}$$

We have two possible outcomes for Φ when φ is given

$$\begin{aligned} &\Phi = \arctan\frac{k}{n} \quad \text{if} \quad h\,n \tan\varphi \leqslant y' < (k+1)h \\ &\Phi = \arctan\frac{k+1}{n} \quad \text{if} \quad (k+1)h \leqslant y' < h(n \tan\varphi + 1). \end{aligned} \tag{2.34}$$

The probabilities for these two cases are given as

$$p\left(\Phi = \arctan\frac{k}{n}\right) = p[h\,n \tan\varphi \leqslant y' < (k+1)h] = \int_{h\,n \tan\varphi}^{(k+1)h} \frac{1}{h}\,dx =$$

$$= k + 1 - n \tan\varphi \tag{2.35a}$$

$$p\left(\Phi = \arctan\frac{k+1}{n}\right) = p[(k+1)h \leqslant y' < h(n \tan\varphi + 1)] =$$

$$= \int_{(k+1)h}^{h(n \tan\varphi + 1)} \frac{1}{h}\,dx = n \tan\varphi - k$$

in which

$$\arctan\frac{k}{n} < \varphi \leqslant \arctan\frac{k+1}{n},\ 0 \leqslant k < n-1. \tag{2.35b}$$

The bias in Ledley's method is calculated as

$$\text{bias}(\bar{\Phi}|\varphi) = \mathcal{E}(\Phi - \varphi|\varphi) = (n \tan\varphi - k)(\arctan\frac{k+1}{n} - \arctan\frac{k}{n}) + \arctan\frac{k}{n} - \varphi \tag{2.36}$$

with the restriction (2.35b).

As the situation here is analogous to that described in Appendix C (with column width nh instead of h) the distribution $p(\varphi)$ is given by equation (2.27). Hence the bias is

$$\text{bias}(\bar{\Phi}) = \sqrt{2}\sum_{k=0}^{n-1}(\arctan\frac{k+1}{n} - \arctan\frac{k}{n})(\sqrt{n^2+k^2} - \sqrt{n^2+(k+1)^2}) + (\sqrt{2}-1). \tag{2.37}$$

This bias is given in figure 2.18 as function of n. The variance for a given φ is

$$\mathcal{E}[(\Phi - \bar{\Phi})^2|\varphi] = (\arctan\frac{k+1}{n} - \arctan\frac{k}{n})^2(k + 1 - n\tan\varphi)(n\tan\varphi - k). \tag{2.38}$$

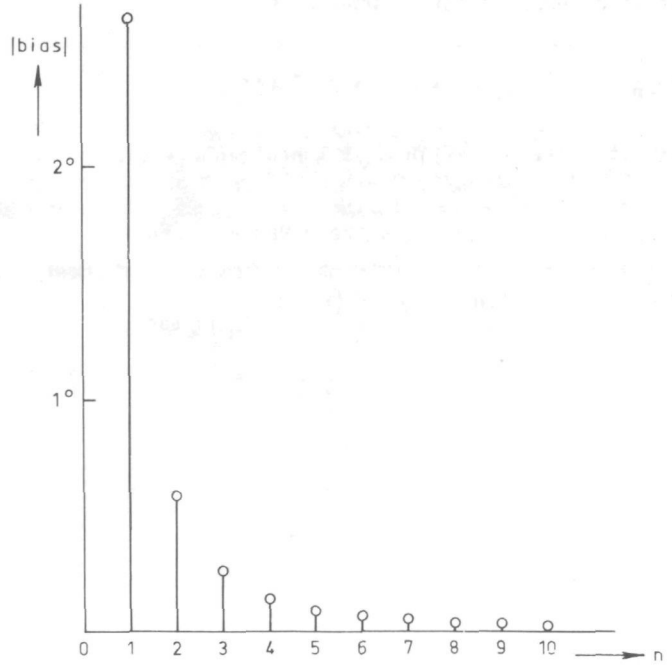


Figure 2.18 Bias in Ledley's method

With the distribution $p(\varphi)$ given in equation (2.27) this variance is

$$\mathcal{E}[(\Phi - \bar{\Phi})^2] = \sqrt{2} \sum_{k=0}^{n-1} (\arctan \frac{k+1}{n} - \arctan \frac{k}{n})^2 \left[(k+1)\sqrt{n^2+k^2} - k\sqrt{n^2+(k+1)^2} + \right.$$

$$\left. - n^2 \ln \frac{[k + 1 - n + \sqrt{n^2+ (k+1)^2}](k + n - \sqrt{n^2 + k^2})}{[k + 1 + n - \sqrt{n^2+ (k+1)^2}](k - n + \sqrt{n^2 + k^2})} \right]. \quad (2.39)$$

Now the quantization error in Ledley's method according to equation (2.19) follows from (2.39) through

$$\sigma_L^2 = \mathrm{var}(\dot{\phi}_L) = \frac{2}{A'^2} \mathcal{E}[(\Phi - \bar{\Phi})^2]. \quad (2.40)$$

In figure 2.16 $A'\sigma_L$ is plotted as a function of n. In the case n=1 Ledley's method is identical to the method of Gallus/Aalderink.

2.6 ERROR CAUSED BY NUMERICAL DIFFERENTIATION

This error is dependent on the shape of the curve and the place of ξ_1 and ξ_2 in the appropriate intervals (equation (2.20)). With methods of numerical analysis (Hildebrand (1956)) an upper bound can be given for this error, but for our investigation an upper bound is a much too rough approximation. The convergence of Taylor expansions (Hamming (1962)) in the region of interest is not sufficiently fast to allow an estimation of the error. So we have to restrict ourselves to an example as no general theory has been given. For the 'analytical chromosome' of section 2.1 this error is computed experimentally. The quadratic deviation E_b^2 between the curvature $\dot{\phi}$ of the analytical chromosome and the first order difference of ϕ is calculated as

$$E_b^2 = \frac{1}{n} \sum_{i=1}^{n} [\dot{\phi}(i) - \frac{1}{A'} (\phi(\xi_2) - \phi(\xi_1))]^2 \quad (2.41)$$

and

$$A' = \frac{(n + 2B)\sqrt{2}\,\pi h}{4}$$

in which n is the number of points involved.

The position of ξ_1 and ξ_2 in the intervals [s(i-B-n), s(i-B)] and

[s(i+B), s(i+B+n)] are not known (equation (2.20)).*) Two assumptions are made about the position of ξ_1 and ξ_2.

a) ξ_1 and ξ_2 are uniformly distributed in the intervals

b) ξ_1 and ξ_2 are in the middle of the interval.

In figure 2.19 the error E_b relative to the maximum real curvature $\dot{\phi}_{max}$ is given as function of the interval length nh (B=0), for the analytical chromosome with $\omega = 3$ and $\omega = 4$.

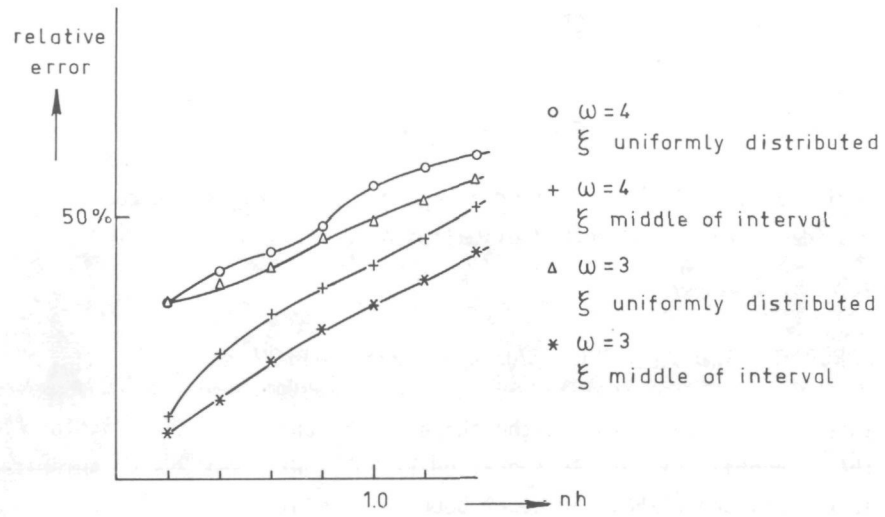


Figure 2.19 Relative error in the curvature due to the numerical differentiation
B = 0, h = 0.05

Figure 2.19 shows that the assumptions about the position of ξ_1 and ξ_2 have an important influence on the relative error. For an increasing interval nh the error due to a linear approximation of the curvature increases. This results in an increasing difference between the first order difference and the first derivative.

*) The non-equidistance of the arc lengths between the intersections of the contour and the grid also gives a contribution to the uncertainty of the position of ξ_1 and ξ_2. All statements in this chapter concerning ξ_1 and ξ_2 are assumed to hold for ξ_3 and ξ_4 as well.

2.7 EVALUATION OF THE ERRORS IN CURVATURE MEASUREMENT

The total error E is computed for the analytical chromosome ($\omega = 3$ and $\omega = 4$). Assuming that σ and E_b are uncorrelated, the total error E is calculated as

$$E = \sqrt{\sigma^2 + E_b^2} \tag{2.42}$$

in which σ and E_b are defined as in section 2.5 and 2.6 respectively. E_b is computed assuming that ξ_1 and ξ_2 are in the middle of the interval. The error E is theoretical as far as σ^2 is concerned. This error E relative to the maximum curvature $\dot{\phi}_{max}$ is given for the method of Gallus/Aalderink and the method of Ledley in figure 2.20. The experimentally found quadratic deviations between the real curvature and the estimated curvature in these two methods are also given in figure 2.20. The errors in the method of Gallus/Aalderink are given for different values of B in figure 2.21 and for different values of h in figure 2.22. These errors are computed for the analytical chromosome with $\omega = 4$. The errors show the same tendency in Ledley's method and for the analytical chromosome with $\omega = 3$.

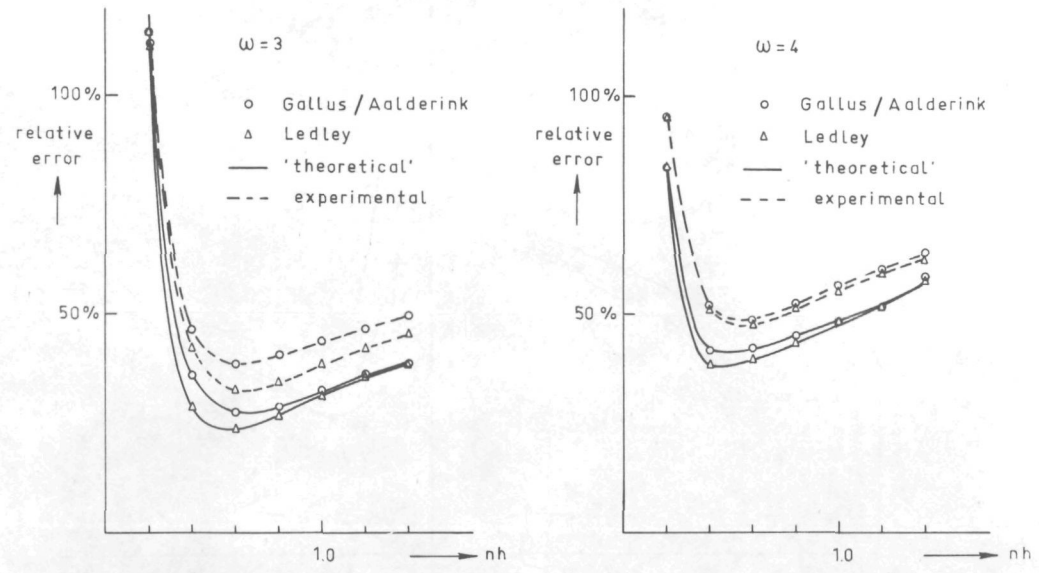


Figure 2.20 Relative error in the curvature for the method of Gallus/Aalderink and Ledley (B = 0, h = 0.2)

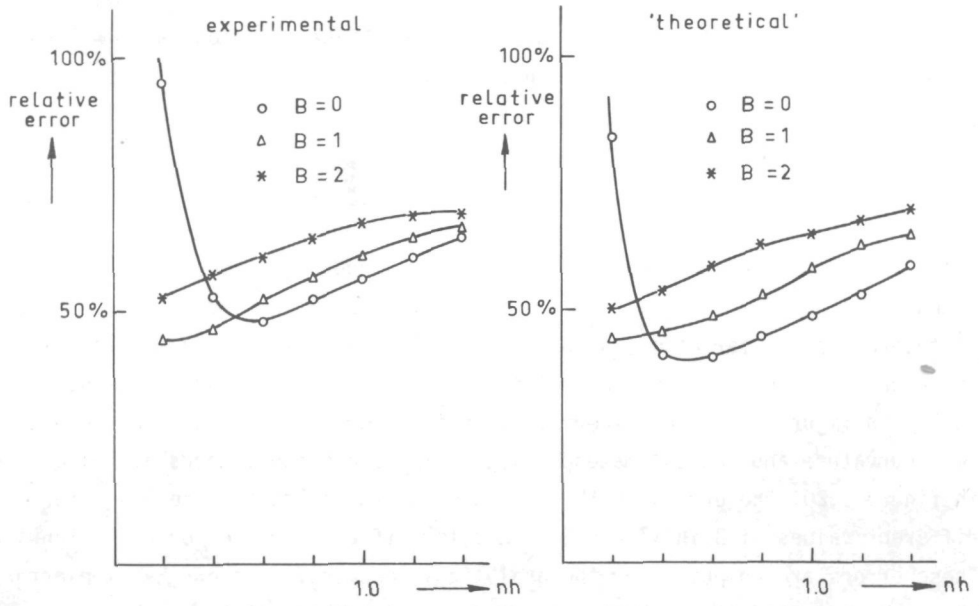


Figure 2.21 Relative error in the curvature for several values of B. *Method Gallus/Aalderink* $\omega = 4$, $h = 0.2$.

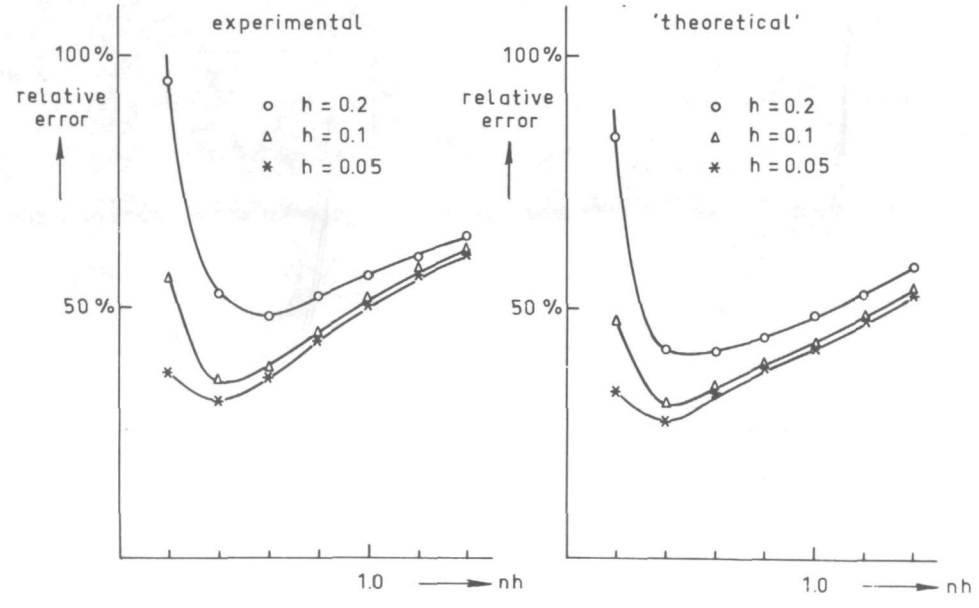


Figure 2.22 Relative error in the curvature for several values of h. *Method Gallus/Aalderink* $\omega = 4$, $B = 0$.

These figures show, that the shape of the 'theoretical' and experimental curves agree. This indicates that the theory about the quantization error provides good results. The difference between the 'theoretical' and experimental curve can be explained by the assumption we made in the computation of E_b. When we had assumed that ξ_1 and ξ_2 were uniformly distributed in the intervals, the 'theoretical' error would have been larger than the experimental error (see section 2.6).

The error consists of two parts: a quantization part (a) decreasing with increasing nh and a numerical differentiation part (b) increasing with increasing nh, giving an optimum, dependent on the shape of the curve. The minimum error is very high (between about 30% and 50%). A lower value of this minimal error can only be obtained when the minimum occurs at lower values of nh to reduce the numerical differentiation error (given in figure 2.19). This can be realized by decreasing grid constant h (and so increasing n) to reduce the quantization error. This agrees with figure 2.22. A minimum in the total error is not always present, because the quantization error may already be dominated by the numerical differentiation error at n=1. This is the case in figure 2.21 for B=1 and B=2.

Many cases with different values of B, n and h have been investigated. The minimal experimentally determined curvature error is in almost all cases at B=1. The difference in the minimal error between B=1 (Aalderink) and B=0 (Gallus) is very small.

In figure 2.23 the experimental error is given of the differentiating low pass filter described in section 2.3. When we compare this error with figure 2.20, it is clear that although the suppression of the higher frequencies is considerably better in this filter than in the method of Gallus/Aalderink, the resulting minimum error is almost identical.

The minimum error in Ledley's method is less than in the method of Gallus/ Aalderink, as can be expected from figure 2.16. This figure shows that the quantization error in Ledley's method is less. This is important for smaller values of h, as in this case the difference in the quantization error of both methods is obvious, and the error is not dominated by the numerical differentiation error.

In figure 2.24 the values of n are given for which the experimentally computed curvature error is minimum as function of the number of contour points for B = 0. This was done for the analytical chromosome with $\omega = 4$. The values for n experimentally found by Gallus for real chromosomes are also shown in this figure. They are below the optimal values of this investigation. As the analytical chromosome is only more or less representative for the smaller

chromosomes, the values given for B and nh are only valid for this type of chromosomes.

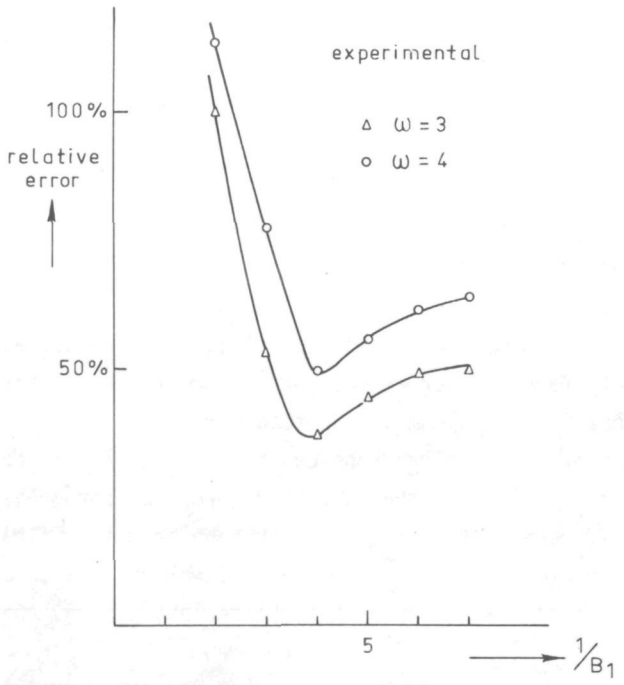


Figure 2.23 Relative error in the curvature. Differentiating low-pass filter
h = 0.2

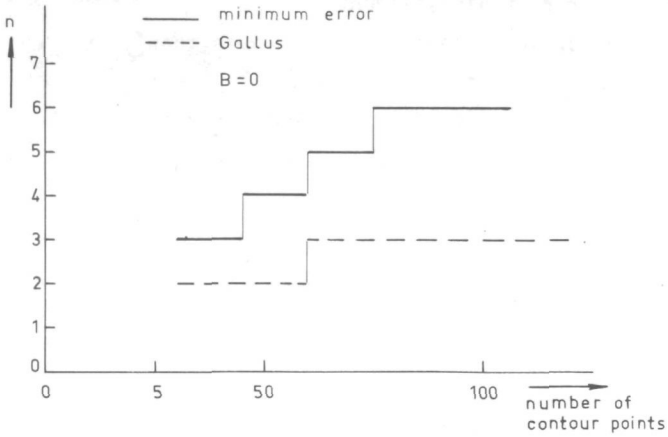


Figure 2.24 Recommended values for n

2.8 EVALUATION OF THE POSITION OF THE CURVATURE EXTREMA

Errors in the computed position of the curvature extrema are important in chromosome analysis, because these extrema are often used to locate arm ends. There again are two main error sources:

a) quantization errors,

b) errors caused by numerical differentiation.

The investigation of the numerical differentiation error presents the same problem as in section 2.6. As no general theory is available, we have to compute this error for the curvature experimentally. In section 2.6 we computed this error from the first order difference discarding the quantization effect. The error in the extrema is very much dependent on the proposition we make about the position of ξ_1 and ξ_2 in such an approach. So it is appropriate to investigate this error by experimental computation of the total quadratic deviation of the measured extrema in relation to the analytically computed extrema of the analytical chromosome . In this total quadratic deviation also the quantization error is present, which will be investigated first.

There are two quantization errors. The first quantization error is introduced by the grid. We have to take a grid point as extremum instead of the point on the contour, in which the curvature is extreme. This results in a quadratic deviation σ_1^2 which can be approximated as

$$\sigma_1^2 = \int_0^h \int_0^h \frac{1}{h^2} (x^2+y^2)dxdy = \frac{2}{3} h^2. \tag{2.43}$$

The second quantization error originates in the curvature measurement. The measured curvature in each curve point is contaminated by noise, as is discussed in section 2.5. So the point in which the measured curvature is extreme may differ from the point in which the real curvature is extreme. We assume that the measured curvature $\dot{\phi}_M(i)$ is equal to the real curvature $\dot{\phi}(i)$ to which the noise $\epsilon(i)$ is added, so

$$\dot{\phi}_M(i) = \dot{\phi}(i) + \epsilon(i). \tag{2.44}$$

We assume that the distribution of $\epsilon(i)$ is independent of i. So, no dependency is present in the noise of neighbouring points.

We will calculate the probability that curve point k is a curvature minimum.

This implies that

$$\dot{\phi}_M(k) < \dot{\phi}_M(l), \qquad \forall l \neq k. \tag{2.45}$$

Combination with equation (2.44) gives

$$\epsilon(l) > \dot{\phi}(k) - \dot{\phi}(l) + \epsilon(k), \qquad \forall l \neq k. \tag{2.46}$$

The probability that $\dot{\phi}_M(k)$ is minimum for a given value of $\epsilon(k)$ is the product of the probabilities that $\epsilon(l)$ fullfills equation (2.46) for all points $l \neq k$, so

$$p[\dot{\phi}_M(k) = \min|\epsilon(k)] = \prod_{\substack{l=-\infty \\ l\neq k}}^{+\infty} \int_{\dot{\phi}(k)-\dot{\phi}(l)+\epsilon(k)}^{\infty} p[\epsilon(l)]\, d\epsilon(l). \tag{2.47}$$

The unconditional probability that $\dot{\phi}_M(k)$ is minimum is

$$p[\dot{\phi}_M(k) = \min] = \int_{-\infty}^{+\infty} p[\epsilon(k)]\; p[\dot{\phi}_M(k) = \min|\epsilon(k)]\, d\epsilon(k) \tag{2.48}$$

and

$$\mathfrak{E}(k) = \sum_{k=-\infty}^{+\infty} k\; p[\dot{\phi}_M(k) = \min]\,.$$

$$\sigma_2^2(k) = \sum_{k=-\infty}^{+\infty} [k - \mathfrak{E}(k)]^2\; p[\dot{\phi}_M(k) = \min] \tag{2.49}$$

The total quantization error σ is given by

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}\,. \tag{2.50}$$

In figure 2.25 the quantization error σ is compared to the experimentally determined total quadratic error E in both methods for the analytical chromosome ($\omega = 3$ and $\omega = 4$). For the computation of σ_2 the distribution of $\epsilon(i)$ was approximated by a normal distribution with a standard deviation given in equation (2.32) (Gallus/Aalderink) and equation (2.40) (Ledley).

For small values of nh the curves do more or less agree. For larger values of nh the error of the numerical differentiation dominates in the experimentally determined error E and the curves can no longer be compared.

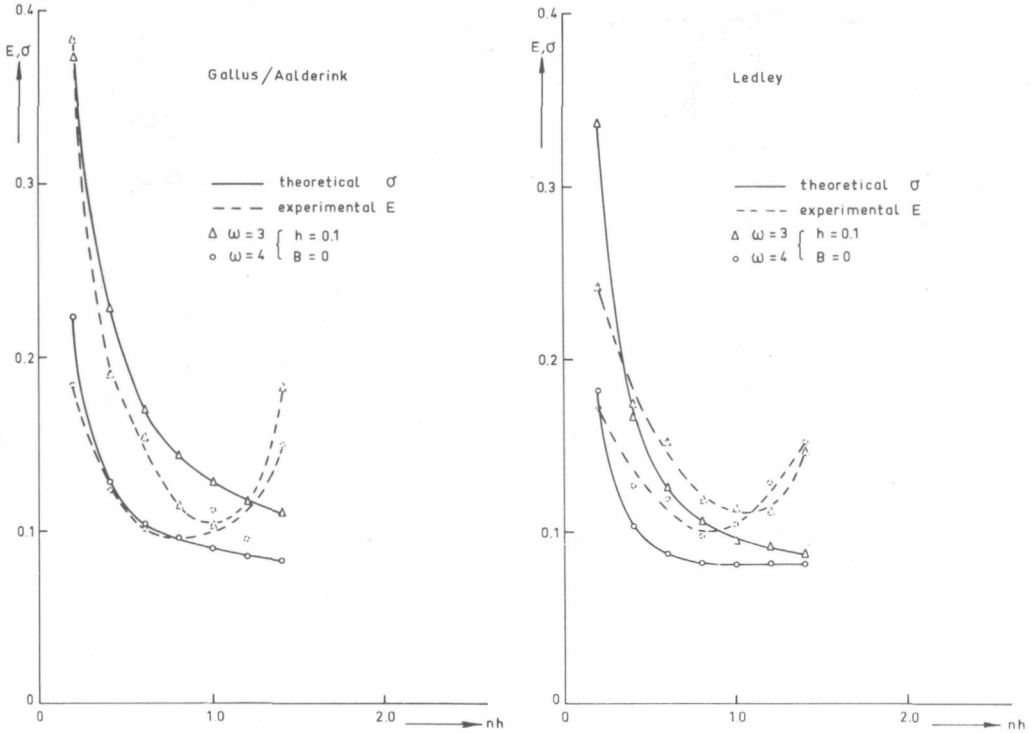


Figure 2.25 Quantization error σ and experimentally determined error E *in the position of the minima*

The quantization error σ in the position of the minima is greater for ω = 3 than for ω = 4. This may be expected as the curvature has a more flat minimum for ω = 3 than for ω = 4. The quantization error in the curvature is the same in both cases, resulting in a greater quantization error in the position for ω = 3.

In figure 2.26 the experimentally determined quadratic error in the position of the minima is given for both methods for two values of h. The difference between the value of the minima of these curves for the method of Gallus/Aalderink and Ledley is small.

When we compare figure 2.25 and figure 2.26 with the minima in the curvature error shown in figure 2.20, 2.21 and 2.22, we see that the minimum error in the extrema position lies at greater values of nh than the minimum in the curvature error. When we are only interested in the position of the curvature extrema, the optimal values of nh are greater than the optimal values, given in section 2.7

for the curvature error. The filtering still gives the low frequency information about the position of the extrema, introducing a large distortion in the curvature, as the higher frequencies are suppressed.

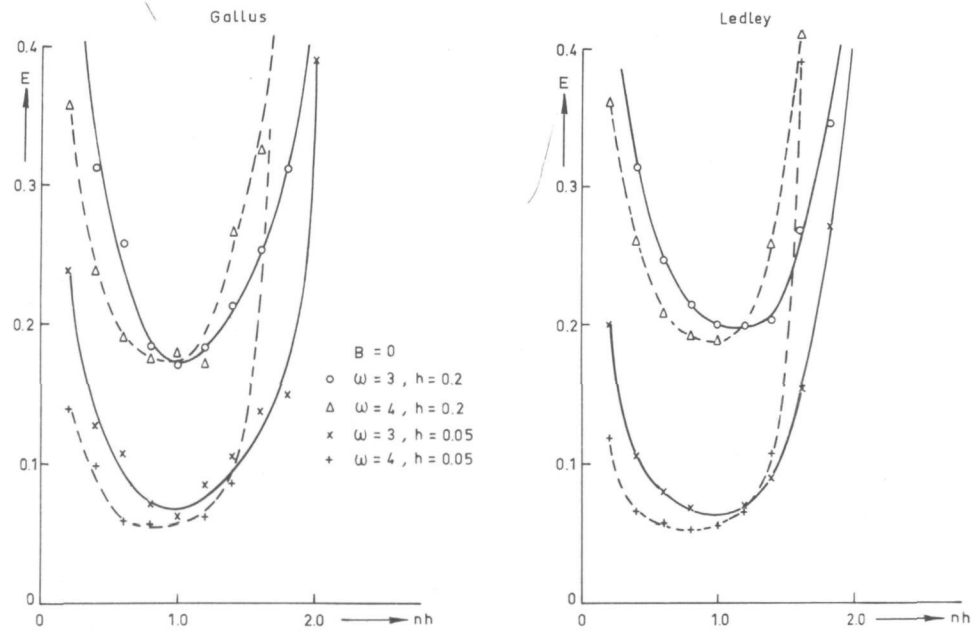


Figure 2.26 Quadratic error in the position of the minima for the method of Gallus/Aalderink and Ledley

So the filtering depends on the purpose for which we need the curvature. A filtering which gives a minimum error in the extrema position, introduces large (systematic) distortions in the curvature. A compromise might be to filter in such a way that the curvature error is about its minimum, as this only gives a moderate increase in the error of the extrema position, particularly for smaller values of h.

2.9 CONCLUSIONS

In this chapter we have seen that although the method of Gallus/Aalderink and the method of Ledley for the measurement of the curvature seem to be quite different, they only differ in the way in which the angular direction of the tangent to the curve is measured.

The errors in the measured curvature are theoretically and experimentically

examined for an 'analytical curve' resembling a small chromosome. The error as a function of the length of the leading or trailing vector n, consists of two parts, a decreasing quantization part and an increasing part caused by numerical differentiation. The minimum error for the investigated 'analytical chromosome' is high (between about 30% and 50%). The minimum error decreases with decreasing grid constant.

The difference in the curvature error between the method of Gallus (B=0) and Aalderink (B=1) is very small. The minimum error in the curvature is less in Ledley's method, than in the method of Gallus/Aalderink.

The minimum error in the position of the extrema lies at higher values of nh (about 1.0) than the minima in the curvature error. So the choice of the filter parameters depends on the purpose for which we need the curvature.

Chapter 3

THE COMPUTATION OF DNA BASED PARAMETERS OF FEULGEN STAINED HUMAN CHROMOSOMES

3.1 INTRODUCTION

This investigation has been carried out in collaboration with the Department of Histochemistry and Cytochemistry of the State University of Leyden It was here that the metaphase spreads were prepared from human lymphocyte cultures and here that staining, photography and scanning of the negatives took place.

A Zeiss cytoscan (SMP) controlled by a PDP 12 is used for the scanning. This mechanical moving-stage scanner has a smallest stepsize of 10 μm. In order to obtain sufficient spatial resolution (0.1 μm - 0.15 μm), photomicrographic negatives of the human metaphases are scanned. Details of this procedure are given by Van der Ploeg et al. (1974).

The scanned metaphases (on magtape) are analysed with the programs that will be described. A line-printer picture of such a scan is given in figure 3.1. The programs are modular, consisting of a main program, which calls subsequent subroutines for the necessary steps in the computation. In figure 3.2 a block diagram of the program for the computation of DNA based parameters is given. The program starts with the localization of the chromosomes by means of a histogram technique. The DNA content and the DNA profile are computed for each chromosome. The DNA profile represents the density integrated over narrow stripes across the chromosome. The centromere position is computed from this profile. The DNA ratio of a chromosome can be obtained by dividing the DNA content of the long arm by the total DNA content of the chromosome. The length of the chromosome and the centromeric index are also computed from the profile. The centromeric index is defined as the length of the long arm divided by the total length.

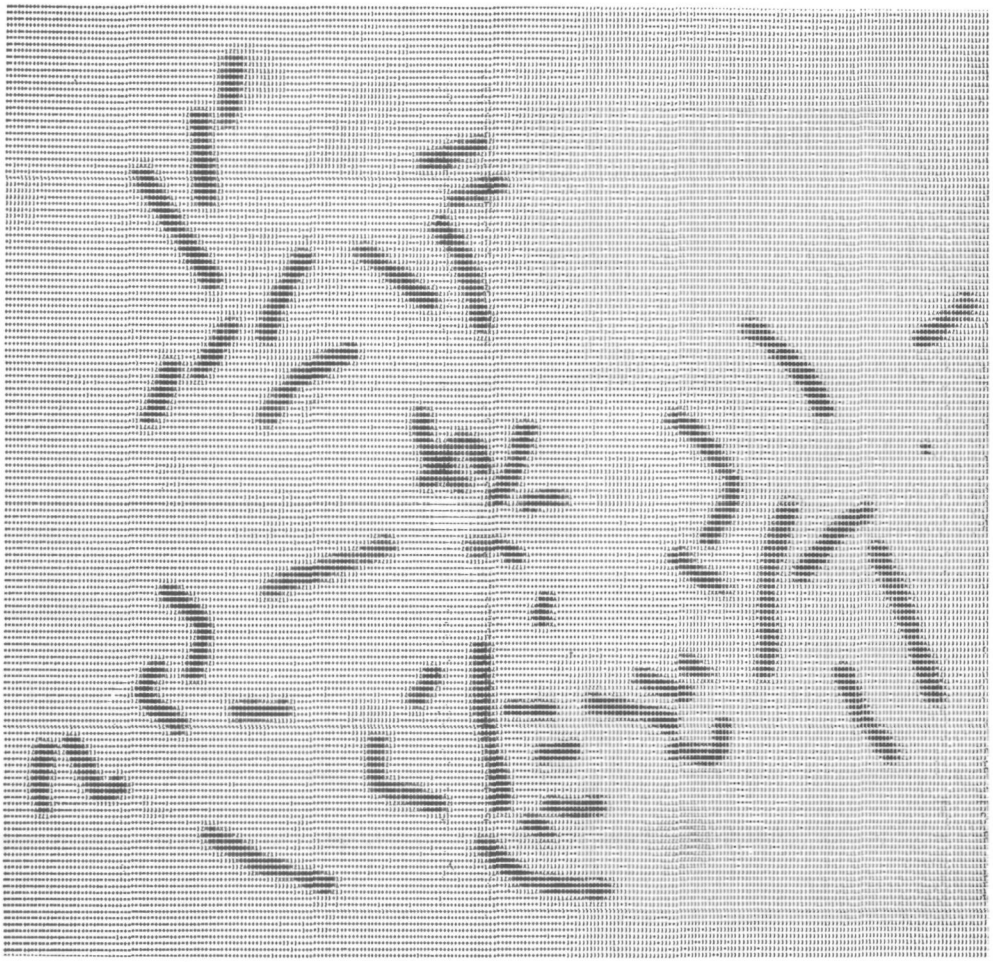


Figure 3.1 Line-printer picture of a scanned part of a metaphase

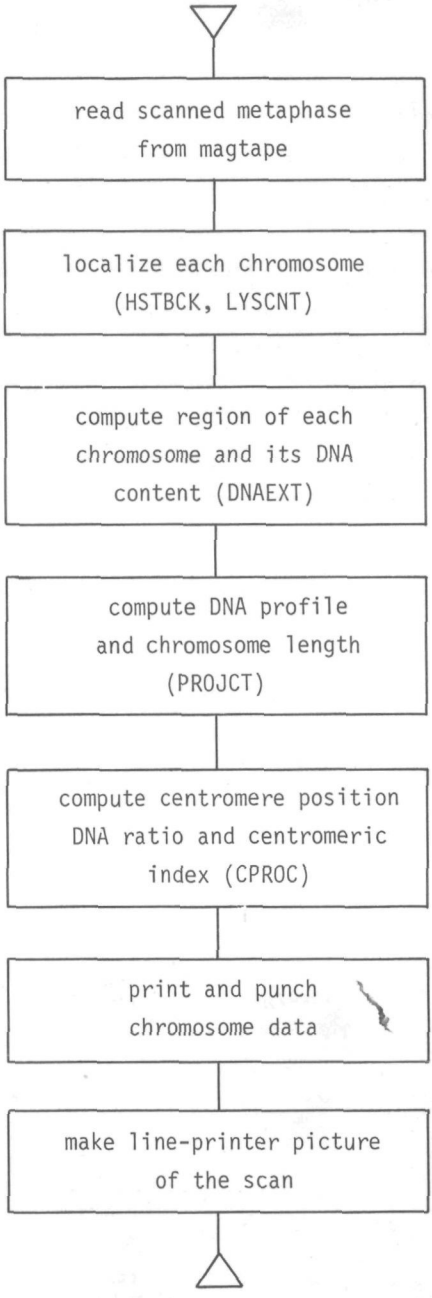


Figure 3.2 Block diagram of the program to compute DNA based features

3.2 LOCALIZATION OF THE CHROMOSOMES

First we have to determine the boundaries of the chromosomes, defined by a dissection level D_L. Points with a density higher than the dissection level are called background points. Points with a density less or equal to the dissection level are regarded as object points.

An image histogram is used to determine the dissection level. This image histogram represents the occurrence frequency of image samples as a function of the density. In figure 3.3 an image histogram of a metaphase is given. A pronounced background peak is always present in the histogram, because most points are background points. It is possible to base the value of the dissection level only on this background peak. It is better, however, to base the dissection level also on the chromosome part of the histogram, as the dissection level is also used to isolate chromosomes in difficult situations, where the chromosomes nearly touch.

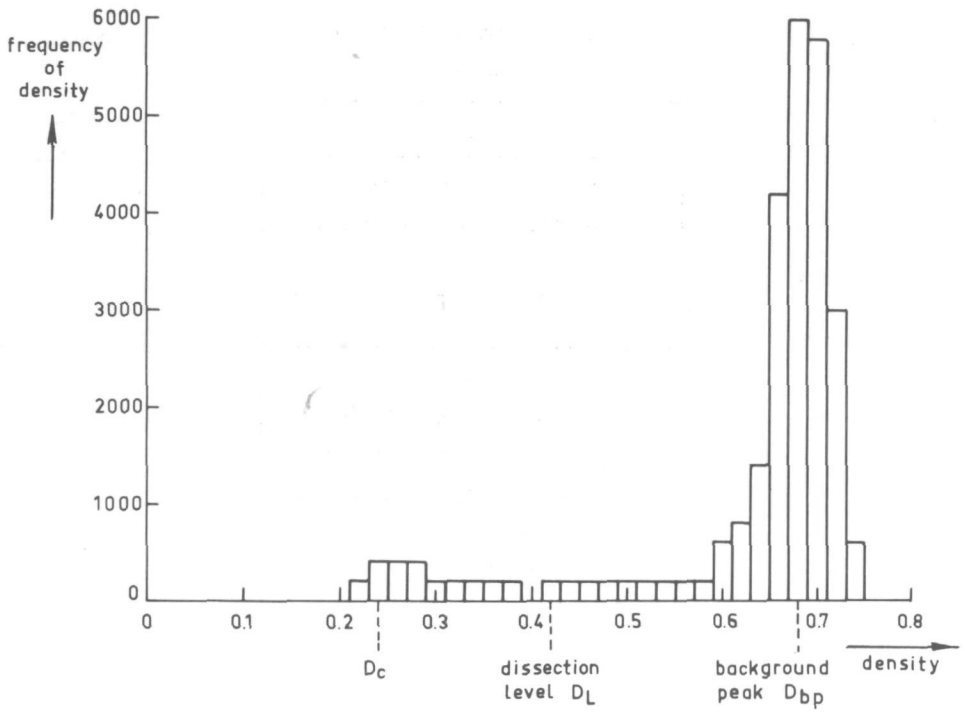


Figure 3.3 Histogram of the density values of a scanned negative of a metaphase (S01A)

When the image histogram has a bimodal structure, a common heuristic (Mendelsohn (1968, 1969)) defines the dissection level as the density corresponding to the internal minimum. Wall (1974) shows that when the noise is gaussian this procedure minimizes the variance of the estimated image area. In practice, however, the histogram does not always show a bimodal structure. Therefore a different method not based on bimodality will be used here. This method is described by Green (1970). The dissection level is obtained from the background peak value D_{bp} and a representative minimal chromosome density D_c. This minimal chromosome density D_c is the density for which the summed histogram values (cumulative histogram) equal a certain fraction f_1 of the total chromosome histogram area. The chromosome histogram area is the histogram area from which the background histogram area is subtracted (assuming that the background peak has a symmetric shape). The dissection level D_l is set at

$$D_L = D_{bp} - f_2(D_{bp} - D_c) \tag{3.1}$$

in which f_2 is a certain fraction. Both f_1 and f_2 are heuristically adapted.

In order to separate chromosomes, which are close together a low setting of the dissection level is used. The values of f_1 and f_2 used for Feulgen stained chromosome images are $f_1 = 0.1$ and $f_2 = 0.6$. These values were experimentally found by varying f_1 and f_2, until chromosomes which were close together could be separated without chromosomes being split up. This dissection level is computed with the subroutine HSTBCK.

When the dissection level is known, the contours of the objects are found by the Object Boundary Quantization method. The clockwise contour-tracing algorithm used tests the 8 neighbours of the last contour point found, until the first one of two successively scanned neighbours is a background point and the second one is an object point. This implies that the original contour must have intersected the grid between these two points. The object point is taken as contour point. This contour-tracing algorithm is programmed in the subroutine LYSCNT.

3.3 CALCULATION OF THE DNA CONTENT

The DNA content DNA of a chromosome C is computed as (chapter 4)

$$DNA = \frac{h^2}{k_a \gamma} \sum_{i,j \in C} \sum [D_b - D(i,j)] \qquad (3.2)$$

in which h is the grid constant, k_a is the specific absorptivity of the chromophore at the wavelength of the monochromatic light used for photography and γ is the gamma of the photographic material. D(i,j) is the measured density of the negative at spot position (i,j) and D_b is the average background density.

When we calculate the DNA content according to formula (3.2) for the chromosome points found with the subroutines of section 3.2, we introduce an error due to the relatively low setting of the dissection level, as is illustrated in figure 3.4. This low setting optimizes separation of chromosomes close together but cannot be used in the computation of DNA content.

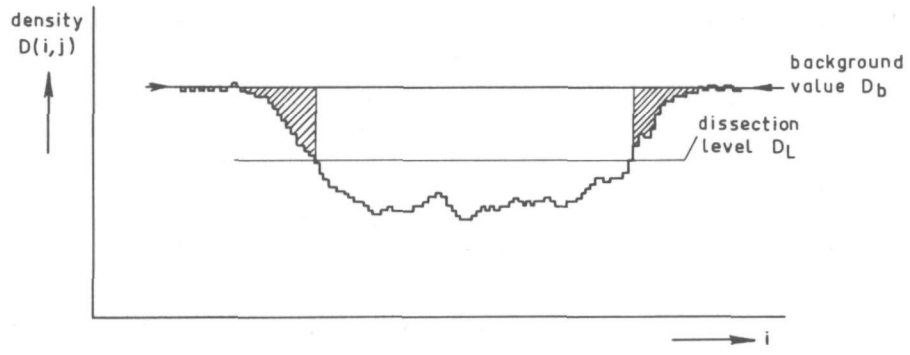


Figure 3.4 Section along a scanline of the density values

This problem is overcome by extending the regions of all chromosomes simultaneously, until the whole area of the metaphase is divided into chromosome regions. The first extension consists of the points which have four neighbour connectivity to an object point. The following extensions consist of those points which have four neighbour connectivity to the previous extension of an object. When the extensions of two different chromosomes come together, the extension stops locally at that point.

In figure 3.5 the average density values of the extensions for three chromosomes are given as a function of their distance from the original region

(as measured in the preparation). It can be seen, that after about 0.7 μm the background area is attained. To compute the DNA content the extensions up to 0.7 μm are regarded as part of the chromosome. The remainder of the extensions up to 1.4 μm is used to estimate the average background density D_b. In this way a local background value is determined for each chromosome.

The chromosome regions, the DNA content of the chromosomes, and the local background values are computed with the subroutine DNAEXT.

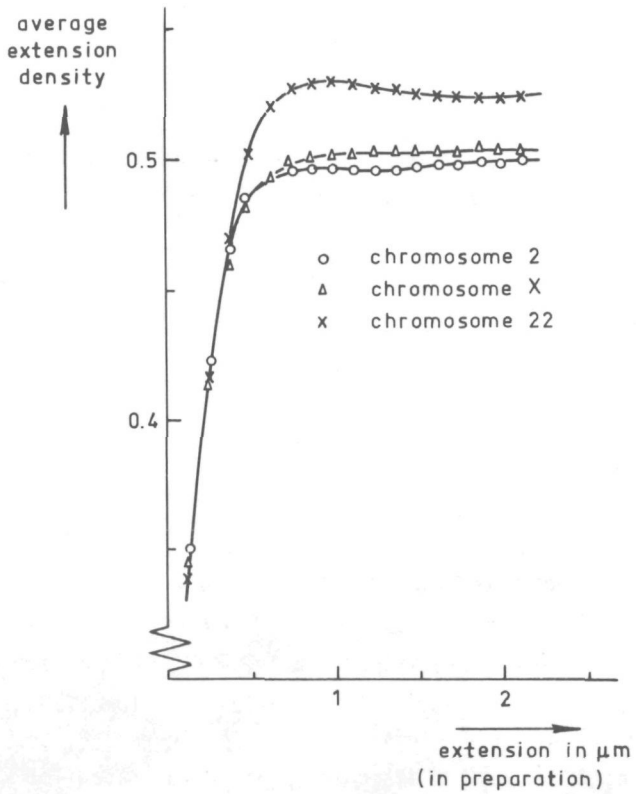


Figure 3.5 Average density values of the extensions (negative 744-10)

3.4 COMPUTATION OF THE DNA PROFILE

For chromosomes which are not curved the profiles are computed by a summation of the chromosome densities over stripes perpendicular to the principal axis. The principal axis passes along the centre of gravity. The angle θ between the principal axis and the X axis of the scan grid is

$$\theta = \tfrac{1}{2} \arctan \frac{2m_{11}}{m_{20} - m_{02}} \tag{3.3a}$$

with

$$m_{pq} = h^{p+q} \sum_{i,j \in C} \sum i^p j^q [D_b - D(i,j)] \; . \tag{3.3b}$$

See also Rutovitz (1967) and Ledley (1972). The summation stripes perpendicular to the principal axis constitute a new grid (requantization grid) rotated an angle θ from the scan grid.

PROFILES BY SUMMATION OF THE SCAN GRID POINTS

According to a not uncommon method,for each point (i,j) of the scan grid, the nearest point (k,l) in the requantization grid is computed. The profile P(k) is obtained by summation of all D(i,j) values assigned to rows of points along the y' direction of the requantization grid. When the grid constant of the requantization grid equals the grid constant of the scan grid, this profile is

$$P(k) = \frac{h^2}{k_a \gamma} \sum_{i,j \in C} \sum [D_b - D(i,j)] \tag{3.4a}$$

with the restriction that

$$k-\tfrac{1}{2} \leqslant i \cos \theta + j \sin \theta < k+\tfrac{1}{2}. \tag{3.4b}$$

This method introduces a significant error. The number of summed scan grid points (i,j) for a row (condition (3.4b)), varies as a function of θ and k as we have checked for a strip of constant width. In figure 3.6, the coefficient of variation in the number of summed elements is given (averaged for k) as a function of θ for such a strip of constant width. This coefficient of variation can be as large as 30%, clearly illustrating that this method can not be applied to the computation of profiles.

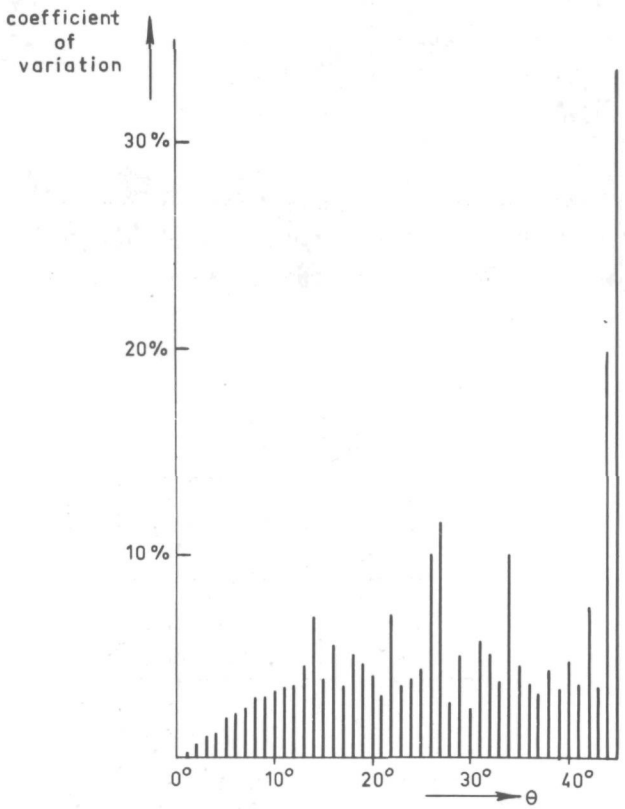


Figure 3.6 Coefficient of variation in the number of summed scan grid points for a strip of constant width (length = 50 h, width = 20 h)

PROFILES BY SUMMATION OF THE REQUANTIZATION GRID POINTS

It would have been best to compute the areas of the scan grid elements covered by each grid element of the requantization grid and use these areas as weight coefficients for the value of the scan grid points. From a computational point of view this is a very elaborate method.

A fairly good approximation can be obtained by using the following method illustrated in figure 3.7. For each point (k,l) of the requantization grid, the nearest point (i,j) in the scan grid is computed. The value of D(i,j) is given to D(k,l). These grid points are summed in the y' direction of the requantization grid to obtain the profile P(k)

$$P(k) = \frac{h'^2}{k_a \gamma} \sum_{l} [D_b - D(k,l)] \qquad (3.5a)$$

with the restriction that $i,j \in C$ and

$$\begin{aligned} i &= \text{integer}[(kh' \cos\theta - lh' \sin\theta)/h + \tfrac{1}{2}] \\ j &= \text{integer}[(lh' \cos\theta + kh' \sin\theta)/h + \tfrac{1}{2}], \end{aligned} \tag{3.5b}$$

with h' the grid constant in the requantization grid. Using this method for a chromosome of constant width the number of summed elements is constant, independent of k and θ. The number of times, however, that a scan grid element is sampled depends on k and θ.

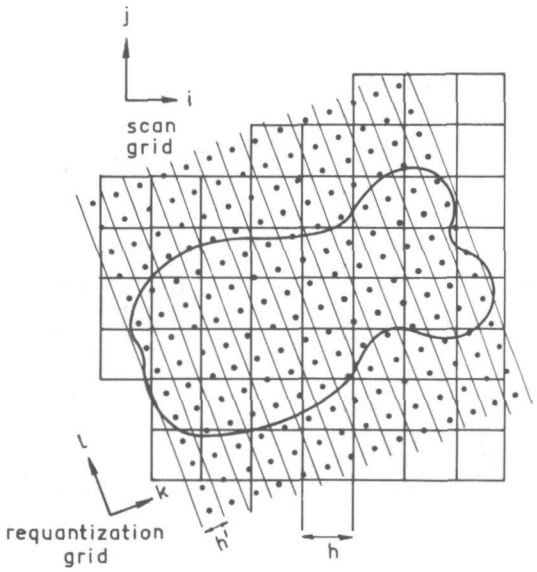


Figure 3.7 Summation of the requantization grid points

By taking a grid constant h' of the requantization grid, which is less than the grid constant of the scan grid, the grid elements of the scan grid are used a number of times. This number is a rough approximation of the covered area. Groen et al. (1976) have investigated the error in this method for a chromosome model, constructed with the aid of 9 two-dimensional Gaussian distributions on each chromatid. In figure 3.8 the maximum deviation in the profile of this chromosome model is given as function of h/h'. The error is less than about 2% if h/h' is larger than 4 ($\theta = 10^{o}$).

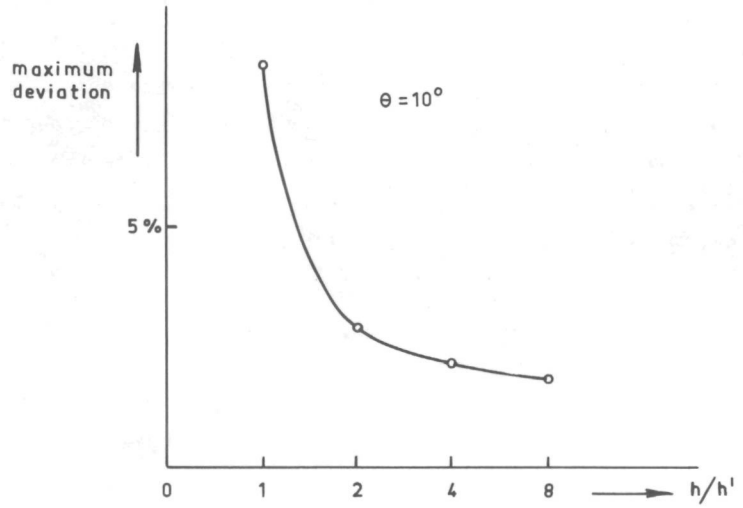


Figure 3.8 Maximum deviation as function of h/h' *for a chromosome model*

INTERPOLATION METHOD

The profile error caused by requantization can be reduced by interpolating between the corner points of the scan grid elements. Oosterlinck (1975) gives an interpolation in which the density in a grid element $D(x,y)$ is written as

$$D(x,y) = c_1 + c_2x + c_3y + c_4xy \tag{3.6}$$

in which c_1 is the density in the centre of the grid element and (x,y) is the position of the sample point measured from this centre (0,0). It is required that the function is exact at the four corner points $(-\frac{1}{2},\frac{1}{2})$, $(\frac{1}{2},-\frac{1}{2})$, $(\frac{1}{2},\frac{1}{2})$ and $(-\frac{1}{2},-\frac{1}{2})$. From these restrictions the coefficients of equation (3.6) can be solved and are given as

$$\begin{aligned}
c_1 &= \tfrac{1}{4}[\, D(-\tfrac{1}{2},-\tfrac{1}{2}) + D(\tfrac{1}{2},-\tfrac{1}{2}) + D(-\tfrac{1}{2},\tfrac{1}{2}) + D(\tfrac{1}{2},\tfrac{1}{2})] \\
c_2 &= \tfrac{1}{2}[-D(-\tfrac{1}{2},-\tfrac{1}{2}) + D(\tfrac{1}{2},-\tfrac{1}{2}) - D(-\tfrac{1}{2},\tfrac{1}{2}) + D(\tfrac{1}{2},\tfrac{1}{2})] \\
c_3 &= \tfrac{1}{2}[-D(-\tfrac{1}{2},-\tfrac{1}{2}) - D(\tfrac{1}{2},-\tfrac{1}{2}) + D(-\tfrac{1}{2},\tfrac{1}{2}) + D(\tfrac{1}{2},\tfrac{1}{2})] \\
c_4 &= [\, D(-\tfrac{1}{2},-\tfrac{1}{2}) - D(\tfrac{1}{2},-\tfrac{1}{2}) - D(-\tfrac{1}{2},\tfrac{1}{2}) + D(\tfrac{1}{2},\tfrac{1}{2})] \,.
\end{aligned} \tag{3.7}$$

The requantization error using this interpolation was experimentally investigated for the same chromosome model as was used in figure 3.8. The maximum deviation in the profile is 2% for h/h' equals 1 (independent of θ).

This interpolation method was used in the projection procedure.

CURVED CHROMOSOMES

When a chromosome is curved, summation of the densities perpendicular to the principal axis will introduce large errors. So a second order polynomial is fitted to the chromosome and the densities are summed perpendicular to this polynomial, as is described by Ledley (1972). This polynomial in the rotated grid (x', y') determined by the principal axis is

$$y'_p = g(x') = q_1 x'^2 + q_2 x' + q_3 . \tag{3.8}$$

When the distance between a point (x',y') and the polynomial is measured along the y' axis, this polynomial is the best fit in the RMS sense if it minimizes

$$E = \sum_{i,j \in C} \sum [y' - g(x')]^2 [D_b - D(i,j)] \tag{3.9}$$

with $x' = i\ h \cos\theta + j\ h \sin\theta$

$y' = j\ h \cos\theta - i\ h \sin\theta .$

The minimum of equation (3.9) is found by differentiation of E with respect to the coefficients of the polynomial q_1,q_2 and q_3 and setting these derivatives to zero.

The arc length s(x') of this polynomial measured from the top of the parabola ($x' = x'_o$) is calculated in Appendix E and found to be

$$s(x') = \frac{1}{4q_1}\left[2q_1(x'-x'_o)\sqrt{1+4q_1^2(x'-x'_o)^2} + \ln\left[2q_1(x'-x'_o) + \sqrt{1+4q_1^2(x'-x'_o)^2}\right]\right] \tag{3.10}$$

in which

$$x'_o = -\frac{q_2}{2q_1} .$$

Ledley uses as an approximation for this arc length

$$s_L(x') = (x'-x'_o)\left[1 + \frac{2}{3} q_1^2 (x'-x'_o)^2\right] . \tag{3.11}$$

In figure 3.9 the error in this approximation is given as a function of q_1 for a part of the parabola, which is symmetric around the top ($x' = x'_o$). The arc length of this part is 40, the grid constant equals 1 and $q_2 = 10\ q_1$. The distance $y'_e - y'_t$ from the top to the chord through the arc ends

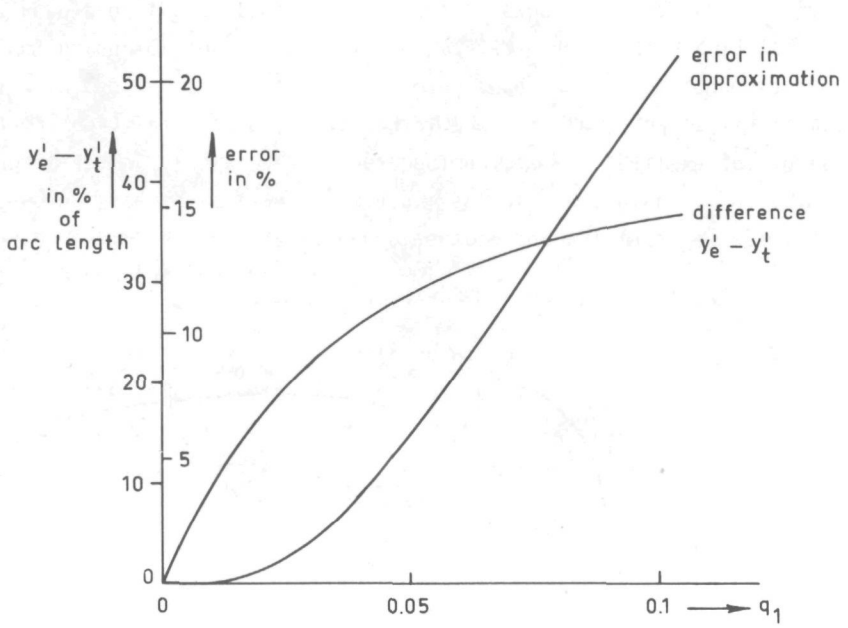


Figure 3.9 Relative error in Ledley's approximation of the arc length and the top distance $y'_e - y'_t$ *as a function of* q_1

(relative to the arc length) is also given in figure 3.9.
From this figure it is obvious that for more sharply curved chromosomes considerable deviations in the arc length may occur, therefore we have preferred to use equation (3.10). This implies, that the inverse calculation of x' from a given s has to be done numerically (Newton-Raphson method).

In order to sum, a new curvilinear grid (x",y") with y" perpendicular to the parabola is sampled. This summation is illustrated in figure 3.10, x" and y" correspond to the arc length and the distance from the parabola respectively. The grid constant in the y" direction equals the grid constant h" at the parabola. The points (x',y') in the coordinate system of the principal axis corresponding to the points (x",y") of the curvilinear grid are computed from the equations

$$\begin{aligned} x' &= s^{-1}(x") - y" \sin\varphi \\ y' &= g(s^{-1}(x")) + y" \cos\varphi \end{aligned} \qquad (3.12)$$

in which $\tan\varphi = 2q_1x' + q_2$. The angle between the tangent to the curve g(x') and the X' axis is φ. The sample points are weighed with the grid element area, because this area in a curvilinear coordinate system is not constant. This area a of a grid element in the curvilinear coordinate system is

$$a = h''^2 [1-2q_1 y''/(1 + \tan^2 \phi)^{3/2}] \qquad (3.13)$$

in which h" is the grid constant at the second order polynomial. A derivation of equation (3.13) is given in Appendix E.

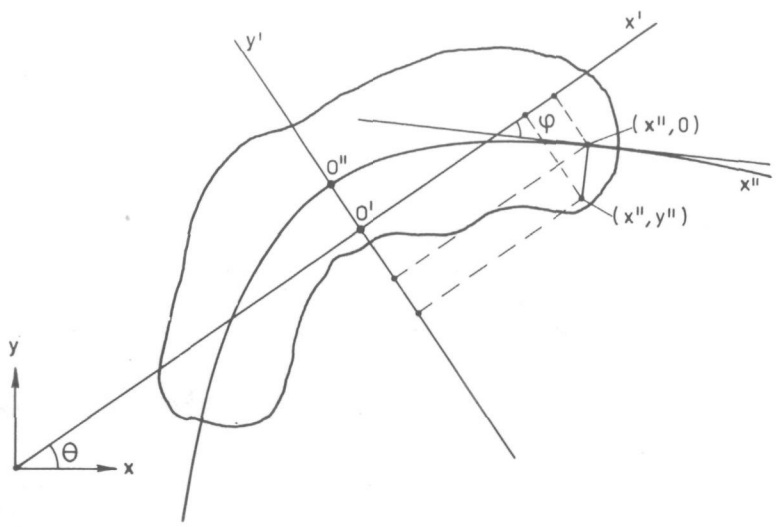


Figure 3.10 Summation perpendicular to a second order polynomial

The symmetry of the densities on both sides of the principal axis or the second order polynomial is obtained by summing the absolute differences of the projections for positive and negative y' or y". The decision whether the densities will be projected on a principal axis or on a second order polynomial is based on this symmetry measure. This measure is also used in chapter 4. The profiles are computed with the subroutine PROJCT, which normalizes the profile to a predetermine number of points along the axis. Therefore the projection is performed twice. The first time to compute the profile roughly in order to determine its length. The second time to obtain the profile with the required number of points by adjustment of the grid constant h' of the new grid. The end points of the profile are defined as the points, where the density drops below 0.1 of its maximum value.

In figure 3.11 the DNA profiles of two median, a submedian and an acrocentric chromosome are given.

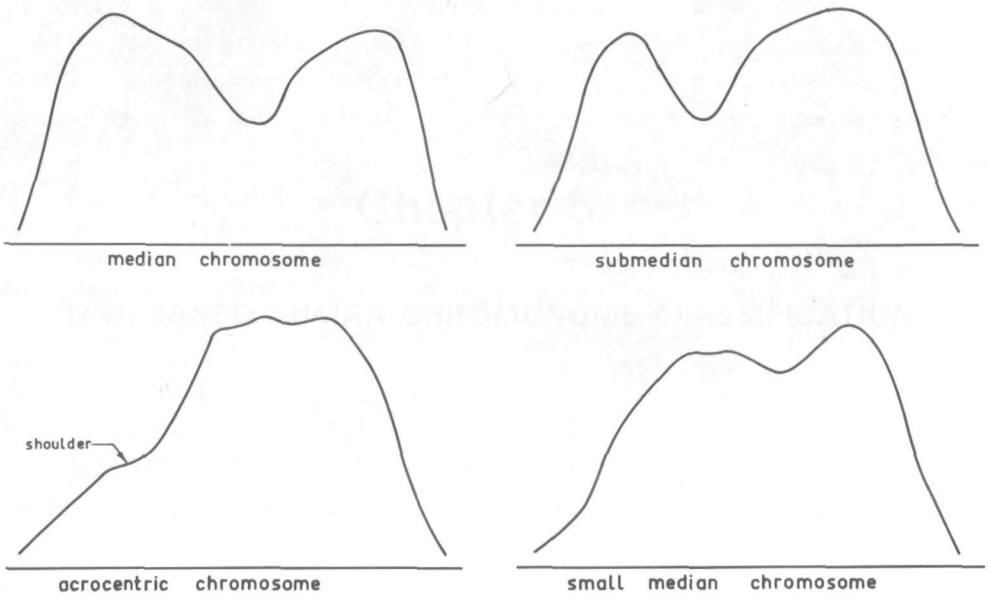


Figure 3.11 Profiles of different types of chromosomes with normalized length

3.5 THE COMPUTATION OF THE CENTROMERE POSITION FROM A DNA PROFILE

An important feature of a chromosome is the centromere position. For median and submedian chromosomes the centromere position is given by a minimum in the DNA profile. Generally no minimum can be observed in the profiles of acrocentric chromosomes, but the centromere position can be obtained from a shoulder (local change in slope) in either end of the profile. See also figure 3.11.

First of all the program has to distinguish between median or submedian chromosomes and acrocentric chromosomes. This is done by determination of the number of maxima in the profile. These profiles are contaminated with noise from e.g. the quantization process and the grid transformation. Hence precautions have to be taken to prevent introduction of spurious maxima by the noise. The suppression of spurious maxima is achieved by filtering the profile and by the restriction that maxima must be larger than a certain fraction of the average value of the profile.

The profiles are filtered digitally by fitting a local second or third order polynomial to the profiles. The width of the filters is chosen in such a way,

that the spurious maxima, which could occur, are suppressed. Van Zee (1974) has investigated these spurious maxima and the filter width experimentally. The spurious maxima were sufficiently suppressed and the filters did not influence the location of the maxima or the centromere when the filter width was between

Table 3.1 Filters for second an third order polynomials

point number \ number of points	11	9	7	3
-5	-36	-21		
-4	9	14	-2	
-3	44	39	3	- 3
-2	69	54	6	12
-1	84	59	7	17
0	89	54	6	12
1	84	39	3	3
2	69	14	-2	
3	44	-21		
4	9			
5	-36			
normalization factor	429	231	21	35

Table 3.2 Filters for the first derivative

second order polynomials

point number \ number of points	11	9	7	5
-5	-5			
-4	-4	-4		
-3	-3	-3	-3	
-2	-2	-2	-2	-2
-1	-1	-1	-1	-1
0	0	0	0	0
1	1	1	1	1
2	2	2	2	2
3	3	3	3	
4	4	4		
5	5			
normalization factor	110	60	28	10

third and fourth order polynomials

point number \ number of points	11	9	7	5
-5	300			
-4	-294	86		
-3	-532	-142	22	
-2	-503	-193	-67	1
-1	-296	-126	-58	-8
0	0	0	0	0
1	296	126	58	8
2	503	193	67	-1
3	532	142	-22	
4	294	-86		
5	-300			
normalization factor	5148	1188	252	12

7 and 9 points for profiles of 64 points. The weight coefficients of the filters used are given in table 3.1 and table 3.2. These filters are discussed by Savitzky et al. (1964).

The shoulder of the profile of acrocentric chromosomes is found by using a filter (table 3.2), which gives the first derivative of a profile. Then the same problem as with a median or submedian chromosome arises: to locate a minimum (due to the shoulder) between two maxima.

The number of maxima with a value larger than a certain fraction of the average value of the profile is computed for the smoothed profile. When two valid maxima exist, the chromosome is assumed to be median or submedian, and the internal minimum between these maxima is computed. This internal minimum is accepted when its value relative to the lowest of the two neighbouring maxima is less than a certain threshold. At this minimum a second order polynomial is fitted to the profile. The minimum of this polynomial is taken as the centromere position.

When no valid internal minimum or no two valid maxima can be found, we assume that the chromosome is acrocentric and in that case the first derivative of the profile is computed. The procedure is repeated, but now at the internal minimum of the derivative a third order polynomial is fitted to the profile. The minimum of the first derivative of this polynomial is taken as the centromere position. When still no valid internal minimum or no two valid maxima are found an error condition is raised.

The described computation of the centromere is realized with the subroutine CPROC. The value of the internal minimum divided by the lowest of the two neighbouring maxima is an output parameter, which gives an indication of the reliability of the centromere position determined.

Chapter 4

ERRORS IN THE MEASUREMENT OF DNA BASED FEATURES

4.1 INTRODUCTION

DNA based features of human chromosomes can be determined by cytophotometry, provided reliable preparation and staining methods are used. Extensive literature on cytophotometry and its possible errors is available. An excellent review is given by Mayall et al. (1970). Although cytophotometry still has some drawbacks, reliable results can be obtained under standardized conditions.

A survey of the possible errors is given in this chapter. The errors which are important for our investigation of DNA based features are discussed in further detail. The errors in these features are experimentally determined and evaluated for repeated scans and repeated photography and compared with the homologue variations.

To obtain sufficient spatial resolution (0.1 μm) with the SMP cytophotometer photomicrographic negatives of human metaphases are scanned. This requires two microscopic systems as illustrated in figure 4.1. In the first system a microscopic negative is obtained from the preparation. (In the description of this system light intensities*) will be denoted by an italic i.) In the second system (densitometer) the microscopic images are scanned. The diameter of the measuring spot equals the sample interval h.(Light intensities in this second system will be denoted by a capital I. For example i_0 and I_0 are the incident light intensities of the first resp. the second system). The errors in the second system are small compared with those in the first system, because only the scanned field of the negative is illuminated and the optical path is identical for all scanned fields.

---

*) We will use the word light intensity, because this is common practice in literature. According to the SI system it should be luminous flux.

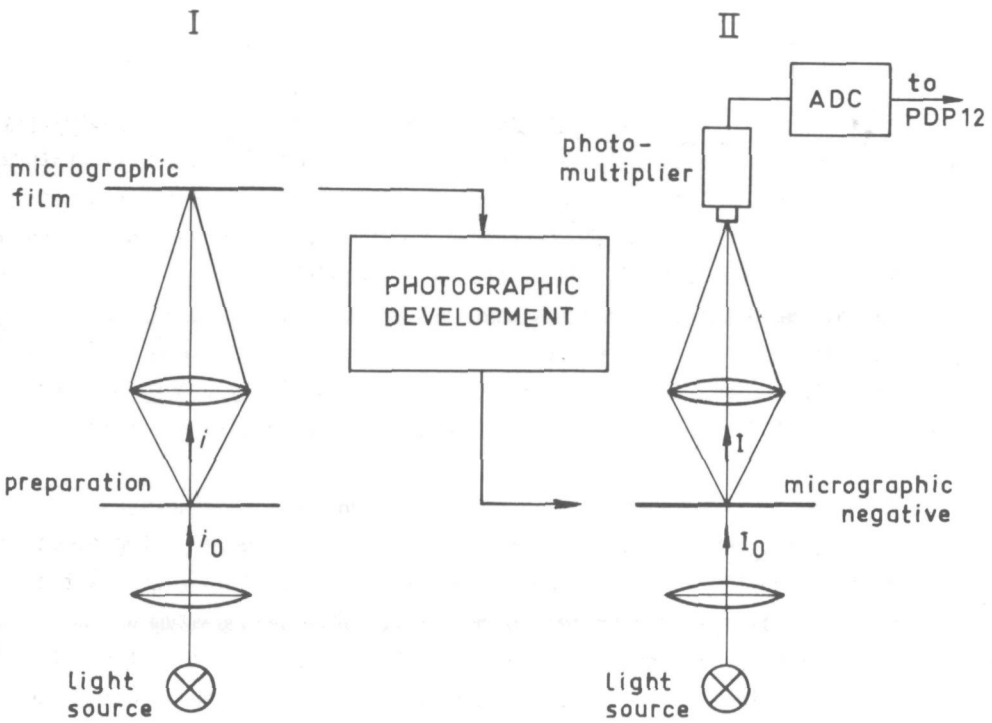


Figure 4.1 Microscopic imaging system

4.2 THE MEASUREMENT OF DNA BASED FEATURES

The law of Lambert and Beer relates the optical density d to the amount of chromophore present in the measuring field, when the chromophore is homogeneously distributed over the measuring field. This law was modified by Walker (1958) for cytophotometry. According to this modification the optical density d is

$$d = \frac{k_a M}{A} \tag{4.1a}$$

in which k_a is the specific absorptivity of the chromophore at the measuring wavelength, A is the area of the measuring field and M is the mass of the chromophore present in the measuring field.

The optical density in the first microscopic system is defined as

$$d = {}^{10}\log \frac{i_o}{i} \tag{4.1b}$$

in which i_o is the incident light intensity and i is the transmitted light intensity.

In our investigation photomicrographic negatives of chromosome preparations were scanned in order to obtain a measuring spot, which was small enough to assume a sufficient homogeneous chromophore distribution, as is discussed in section 4.5. Photography was performed at the wavelength of the absorbance maximum of the chromophore under standardized conditions. The properties of a photographic emulsion are commonly specified by the characteristic curve or Hurter-Driffield curve given in figure 4.2.

This curve relates the optical density D of the negative to the exposure H given as

$$H = E\, T_E = k_E\, i\, T_E \tag{4.2}$$

in which E is the intensity of illumination and T_E is the exposure time. k_E is a constant, which depends on the first microscopic system and i is the transmitted intensity. When T_E is extremely long or extremely short, the photographic densities produced are lower than the expected values (reciprocity failure). A degree of blackening always occurs at zero exposure. This is described as the fog density of the emulsion. In order to obtain an optimal response and a linear relation, we have to make sure that the optical densities produced by

chromosome and background lie within the linear part of the Hurter-Driffield curve (part B-C of figure 4.2). The relation between D and H of this linear part can be described as

$$D = \alpha + \gamma \, {}^{10}\log H = \alpha + \gamma \, {}^{10}\log k_E i T_E \tag{4.3a}$$

in which α is a constant and γ is the gamma of the photographic emulsion, defined as the slope of the linear portion of the Hurter-Driffield curve.

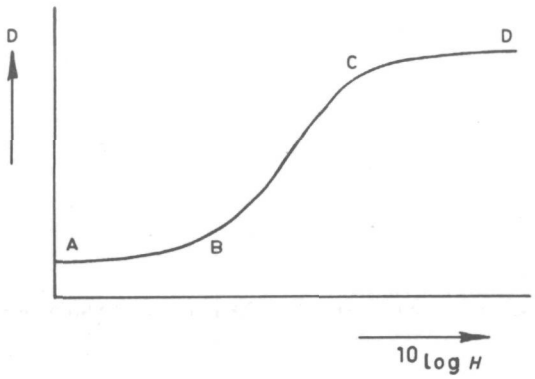


Figure 4.2 Hurter-Driffield curve of a photographic emulsion

The optical density of the negative measured in the second microscopic system is

$$D = {}^{10}\log \frac{I_o}{I} \tag{4.3b}$$

in which I_o is the incident intensity and I is the transmitted intensity.

The mass of DNA (DNA content) present in a chromosome C is obtained by summation of the densities measured over the measuring spot of the second microscopic system. Background correction is realized by subtraction of the background density obtained from the background extensions of the chromosome described in chapter 3. The DNA content is computed as

$$DNA = \frac{h^2}{k_a \gamma} \sum_{k,l \in C} \sum [D_b - D(k,l)] \tag{4.4}$$

in which D(k,l) is the density at grid position (k,l) of chromosome C and D_b is

the background density. The specific absorptivity k_a and γ are unknown constants. The grid constant is h. Both the grid positions and the grid constant are calculated as if the grid was present at the original preparation. As these densities are obtained from the photographic negative, by combination of equations (4.1), (4.3) and (4.4) DNA can be expressed in the intensities related to the preparation in the first microscopic system

$$DNA = \frac{h^2}{k_a} \sum_{k,l \in C} \sum {}^{10}\log \frac{i_b}{i(k,l)} = h^2 \sum_{k,l \in C} \sum \left(\frac{M(k,l)}{A} - \frac{M_b}{A} \right). \qquad (4.5)$$

The intensity at grid position (k,l) of the preparation is $i(k,l)$, when an incident intensity i_o is present. The mass of chromophore present at the measuring spot (grid position (k,l)) with area A is M(k,l). The background intensity of the chromosome in the preparation is i_b, when the same incident intensity i_o is present. The equivalent mass of chromophore present at the measuring spot in the background is M_b.

To compare the chromosome DNA content of different metaphases, the chromosomes were normalized as two unknown constants are present in equation (4.4). One method is to normalize the DNA content with respect to the DNA content of the chromosomes 2. In order to reduce noise in the normalized DNA values, it is better to take more than two chromosomes into account. In this investigation the chromosomes 2, 3 and 4 were used for normalization. The normalization factor η was obtained by linear regression of the measured DNA content against the expected normalized DNA contents of these chromosomes.

$$\eta = \frac{\sum_{i=2,3,4} DNA_i \, \Psi_i}{\sum_{i=2,3,4} \Psi_i^2} \qquad (4.6)$$

with measured DNA content DNA_i and an expected normalized DNA content Ψ_i of chromosome i *). The values of Ψ_i were first obtained from literature (Mendelsohn (1973)) and afterwards from our own experiments.

The variance σ^2_{DNA} in the DNA content, measured according to equation (4.4) is

*) One pair is present at most of each chromosome i.

$$\sigma^2_{DNA} = \frac{h^4}{k_a^2\gamma^2}\left(\left[\sum_{k,l\in C}\sum \sigma^2(D(k,l)\right] + \frac{n_c^2}{n_b}\sigma^2(D_b)\right) \tag{4.7a}$$

and

$$\sigma_{DNA} = \frac{h^2 n_c}{k_a\gamma}\sqrt{\frac{1}{n_c}\frac{\sum_{k,l\in C}\sum \sigma^2(D(k,l))}{n_c} + \frac{1}{n_b}\sigma^2(D_b)} \tag{4.7b}$$

in which n_c is the total number of points of chromosome C and D(k,l) the density at grid position (k,l). The background density is obtained from averaging n_b background points outside the chromosome. It is assumed that the errors in these background points are uncorrelated. So the variance in the average background density is $\sigma^2(D_b)/n_b$, in which $\sigma^2(D_b)$ is the variance in a single background point. When these errors are strongly correlated the value of n_b must be considered 1. The error in the average background density is the same for all points of the chromosome, introducing an error with a systematic character, contrary to the error in D(k,l) which has a stochastic character. For a number of errors like the quantization error (section 4.4) and the shot noise (section 4.3) $\sigma(D(k,l))$ is always less or equal to $\sigma(D_b)$. In that case the standard deviation in the DNA content σ_{DNA} will always be less or equal to ΔDNA given as

$$\sigma \leq \Delta DNA = \frac{h^2 n_c}{k_a\gamma}\sqrt{\frac{1}{n_c} + \frac{1}{n_b}}\,\sigma(D_b) \tag{4.8}$$

in which it is assumed that the errors in the density D(k,l) are uncorrelated. The relative error in the DNA content is

$$\frac{\Delta DNA}{DNA} = \frac{\sqrt{\frac{1}{n_c} + \frac{1}{n_b}}\,\sigma(D_b)}{D_a} \tag{4.9}$$

with $D_a = \frac{1}{n_c}\sum_{k,l\in C}\sum [D_b - D(k,l)]$, the average chromosome density.

4.3 ERRORS IN THE MEASUREMENT OF DNA BASED FEATURES

Generally there are two types of measuring errors: stochastic errors and systematic errors. Systematic errors are difficult to detect and introduce a bias in the experiment. When systematic errors occur as scaling factors, they are not

important in our investigation, because we are only interested in the relative magnitude of the measurements.

Errors in the specimen are due to:
a) *The staining.* (loss of substrate, non-specific staining, differences in the stoichiometry),
b) *The presence of other components* which absorb at the same wavelength.

Errors may occur in the microscope measuring system due to:
a) *The microscope.* The specimen is not an ideal amplitude object. This error is reduced by the embedding of the specimen in a mixture of Caedax: Cargille oil (15:1 mass parts) resulting in a refractive index of about 1.54. The focusing of the microscope is subjective and the question may be asked, whether the in-focus situation is the most pleasant image for the microscopist or not. In figure 4.3 the transfer function of the microscopic system according to Van den Berg (1974) is given (λ = 561 nm, numerical aperture (NA = 1.30). When a large numerical aperture is used, a high resolution is obtained but the depth of field is small. Figure 4.3 shows that an out of focus displacement of one wavelength gives drastic changes in the transfer function. Mendelsohn et al. (1972) showed, however, that the influence of focusing on the measured DNA content is small, because of the summation procedure over the whole chromosome.

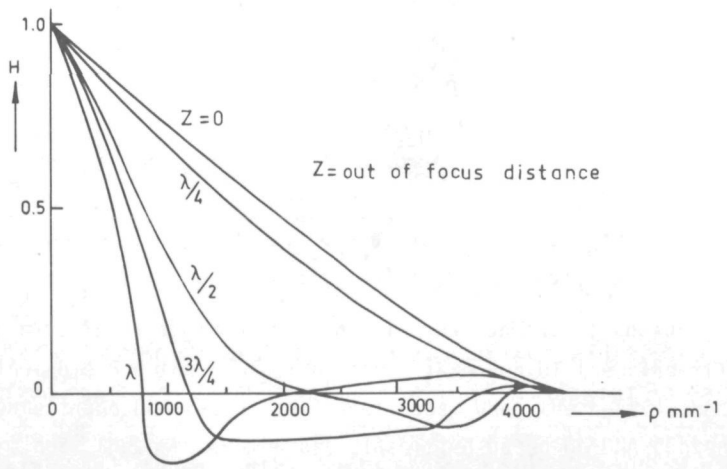


Figure 4.3 Optical Transfer Function (H) *of a microscopic system as function of the spatial frequency* ρ (λ = 561 nm, NA = 1.30)

b) *Glare*. Stray light is an important source of error in micro densitometry. It has extensively been discussed by Goldstein (1970). Light which passes through the preparation will be repeatedly reflected at glass-air surfaces in the microscope and by imperfections of other parts of the optical system. When the illuminated microscope field is larger than the specimen, some of the light passing through the specimen-free areas will end up at the image of the specimen by this error. So the apparent absorbance will be less than the real one. This error is slightly affected by the numerical aperture of the microscope condensor, but is closely related to the area of the illuminated specimen-free field. The true absorbance of the specimen d is according to Goldstein (1970)

$$d = {}^{10}\log [(1-J)/(i-J)] \qquad (4.10)$$

in which the intensity i_0 of the incident light is taken to be unity and i is the apparent transmittance of the specimen in the presence of glare J. Van der Ploeg et al. (1974) have investigated the glare of the system used, which appeared to be 0.9%. This low value for the glare is mainly due to the high quality of the optical system. The highest local absorbance in Feulgen stained chromosomes amounts to 0.25. The glare then results in a relative error in the measured absorbance of about 1%. In practice the error in the total integrated optical density value of a chromosome will be less since many of the local density values are lower.

c) *Distributional error*. This error is introduced by an inhomogeneous chromophore distribution over the measuring field. The distributional error in the total integrated optical density will be the sum of the distributional errors in the individual measurements, when the specimen is scanned. In section 4.5 we will investigate the influence of this error on the measured DNA content.

Mayall and Mendelsohn (1970) argue that there still is uncertainty about the unbiasedness of the law of Lambert and Beer for objects with dimensions approaching the optical resolution of the microscope. This should be a point of further research.

d) *Chromatic error*. This error can be expected when the product of specific absorptivity k_a of the chromophore and the sensitivity γ of the photographic emulsion changes in the spectral bandwidth of the measuring light. As in the photographic procedure an interference filter with a narrow bandwidth in the flat peak of the Feulgen DNA absorbance spectrum has been used, this error may be discarded.

e) *Condensor Aperture*. The conical illumination biases the average path length through the specimen and the effective area of the measuring field. As this error is a scaling factor, it is only important for absolute measurements.

Errors introduced by the photography are due to:

a) *Densities outside the linear part* of the Hurter-Driffield curve.

b) *Variations in the parameters* γ *and* α of the Hurter-Driffield curve.

c) *Graininess*. This limits the resolution in microphotographs, which have undergone excessive photographic magnification. Mees (1954) argues, that the origin of this effect is not only due to the individual grains of the emulsion but also to the random spatial distribution of quantum arrivals at the emulsion.

Van der Ploeg et al. (1974) have investigated the optimal combination of magnification, film type and stepsize of the scanning stage, with respect to this error.

Errors in the scanning densitometer are due to:

a) *The photosensitive device*. A photodiode (BPX42) is the photosensitive device in the densitometer. The shot noise of the photodiode is dominated by the thermal noise of the pre-amplifier. The measured standard deviation of the noise σ_{pd}, present in the amplified current is 0.14% of the maximum value of the scale and independent of the measured intensity.

Another source of error may be present in deviations from linearity in the characteristic of the photosensitive device. These deviations are small for the type of photodiode used.

b) *The quantization noise* and the linearity of the analog to digital convertor. The quantization noise of a linear and a logarithmic scale is investigated in more detail in section 4.4.

c) *The instability of the light intensity*. This error is negligible in the densitometer used.

d) *The measuring spot*. The size of the measuring spot does not only influence the distributional error, but also the optical transfer function of the densitometer. The optical transfer function for a circular spot is

$$H(\rho) = \frac{2J_1(\pi o\rho)}{\pi o\rho} \qquad (4.11)$$

in which o is the diameter of the spot, ρ is the spatial frequency and J_1 is the first order Bessel function (Van den Berg (1974)). In figure 4.4 the optical transfer function of the densitometer is given for several values of the spot-diameter. The spatial frequency is calculated as if the spot was present at the preparation in the first microscopic system. (The amplification of the first microscopic system is 322 times). In our experiments the spot diameter always equals the grid constant. At half the sampling frequency the transfer function decreases from 1 to 0.72 in this situation. This decrease due to the measuring spot is small compared with the influence of the focus in the first microscopic system (figure 4.3) on the transfer function of the complete system.

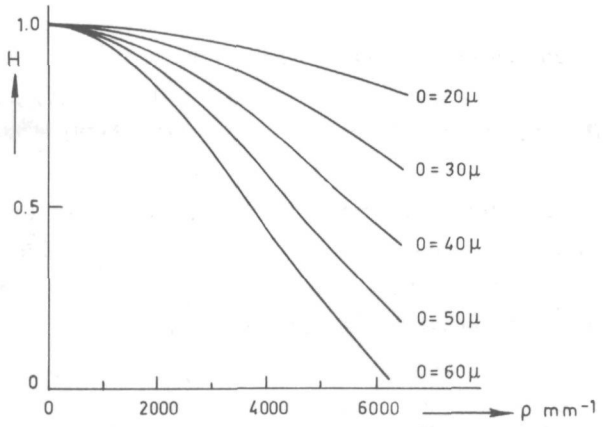


Figure 4.4 Optical transfer function for different spot diameters o. *Spatial frequency* ρ *of the first system (amplification 322 times)*

MEASUREMENTS IN RELATION TO DIFFERENT TYPES OF ERRORS

The errors present in repeated scans of the same negative are introduced by the stochastic errors in the scanning densitometer e.g. noise in the photo-sensitive device and quantization noise. The negative was remounted before each scan. Although the distributional error has a systematic character, it also introduces stochastic errors in repeated scans, because the position of the measuring spot will differ from scan to scan.

In order to investigate the photographic errors, repeated photographs were taken of one metaphase and these photographs were scanned. The errors now

include the scan duplication errors, the errors introduced by the photography and the errors of the first microscopic system as far as these errors have a stochastic character.

In our case a number of errors present in the first microscopic system are relatively small such as: glare, chromatic error, condensor aperture error and also the influence of non-ideal focusing. The distributional error is investigated in section 4.5. This error decreases with decreasing size of the measuring spot. The errors in the second microscopic system (densitometer) are even less, because the illuminated field is only two times the spot size and the optical path is identical for all scan points.

The scan duplication error, the photographic error and the homologue variations are experimentally investigated in section 4.6. The errors in the specimen, however, are indistinguishable from the homologue variations here.

4.4 QUANTIZATION ERRORS OF A LINEAR AND A LOGARITHMIC SCALE WITH RESPECT TO DNA MEASUREMENTS

In this section we shall exclusively discuss the quantization error in the measurement of the density in the second microscopic system (densitometer).

In a linear scale the intensity I is first quantized and then the logarithm is taken of the quantized intensity to obtain the density (equation (4.3b)). The maximum measured intensity I_{max} is set to the maximum N of the scale, so the quantization interval is I_{max}/N. The error in the density δD is in first order approximation given as

$$\delta D = \frac{dD}{dI}\,\delta I = \frac{^{10}\log e}{I}\,\delta I. \tag{4.12}$$

Assuming that the quantization error in the intensity I is uniformly distributed, the standard deviation $\sigma_{lin}(D)$ of the quantization error is in first order approximation

$$\sigma_{lin}(D) = \frac{^{10}\log e\; I_{max}}{\sqrt{12}\; I\; N}. \tag{4.13}$$

In a logarithmic scale first the logarithm is taken to obtain the density and then the density is quantized. Let the logarithmic scale consist of f_d density units, 10^{f_d} being the largest intensity rate that can be expressed within this scale. Hence we assume

$$f_d \geq {}^{10}\log \frac{I_o}{I_{max}} - {}^{10}\log \frac{I_o}{I_{min}} = {}^{10}\log \frac{I_{min}}{I_{max}} \tag{4.14}$$

in which I_{max} is the maximum intensity, and I_{min} is the lowest intensity to be measured. Assuming that the quantization error in the density D is in first order approximation uniformly distributed, we obtain for the standard deviation $\sigma_{log}(D)$ of the quantization error

$$\sigma_{log}(D) = \frac{f_d}{\sqrt{12}\ N} \tag{4.15}$$

in which N is the number of quantization levels of f_d. So when $I/I_{max} > ({}^{10}\log e)/f_d$, a linear scale gives a smaller quantization error.

The DNA content is computed according to equation (4.4). Since at a chromosome point (k,l) the density D(k,l) is less than the background density D_b, $\sigma_{lin}(D(k,l))$ will be less than $\sigma_{lin}(D_b)$ for a linear scale according to equation (4.3b). The standard deviation $\sigma_{log}(D(k,l))$ equals $\sigma_{log}(D_b)$ for a logarithmic scale. So the quantization error in the DNA content will be less or equal to ΔDNA given by equation (4.8). According to equations (4.9), (4.13) and (4.15) the upper bound for the relative error in the DNA content will be

$$\left(\frac{\Delta DNA}{DNA}\right)_{lin} = \frac{\sqrt{\frac{1}{n_b} + \frac{1}{n_c}}\ I_{max}\ {}^{10}\log e}{\sqrt{12}\ N\ I_b\ D_a} \tag{4.16}$$

for a linear scale and

$$\left(\frac{\Delta DNA}{DNA}\right)_{log} = \frac{f_d \sqrt{\frac{1}{n_b} + \frac{1}{n_c}}}{\sqrt{12}\ N\ D_a} \tag{4.17}$$

for a logarithmic scale, in which I_b is the background intensity and D_a is the average chromosome density.

According to equation (4.3a) and (4.3b) the intensities I measured by the densitometer in the negative, can be expressed in the intensities of the first microscopic system as

$$\frac{I_o}{I} = 10^{\alpha}(k_E \cdot T_E)^{\gamma}. \tag{4.18}$$

In the measuring process the illumination intensity I_o of the densitometer is adjusted in such a way, that the maximum occurring intensity I_{max} is the maximum N of the scale. The maximum intensity I_{max} and the background intensity I_b are given as

$$\frac{I_o}{I_{max}} = 10^{\alpha}(k_E \; i_{max} \; T_E)^{\gamma} \tag{4.19a}$$

and

$$\frac{I_o}{I_b} = 10^{\alpha}(k_E \; i_b \; T_E)^{\gamma} \; . \tag{4.19b}$$

From equation (4.19a) and (4.19b) γ can be expressed in the background intensity I_b as

$$\gamma = {}^{10}\log \frac{I_{max}}{I_b} \; / \; {}^{10}\log \frac{i_b}{i_{max}} \; . \tag{4.20}$$

The upper bound $\Delta DNA/DNA$ for the relative error in the linear case can be expressed as function of I_b by combination of equation (4.16), (4.3a) and (4.20) as

$$\left(\frac{\Delta DNA}{DNA}\right)_{lin} = \frac{n_c \left({}^{10}\log \frac{i_b}{i_{max}}\right) \sqrt{\frac{1}{n_b} + \frac{1}{n_c}} \; I_{max} \; {}^{10}\log e}{\sqrt{12} \; N \; I_b \left({}^{10}\log \frac{I_{max}}{I_b}\right) \sum\limits_{k,l \in C} \sum {}^{10}\log \frac{i_b}{i(k,l)}} \; . \tag{4.21a}$$

Combination with equation (4.1) gives:

$$\left(\frac{\Delta DNA}{DNA}\right)_{lin} = \frac{\sqrt{\frac{1}{n_b} + \frac{1}{n_c}} \; C_{max} \; I_{max} \; {}^{10}\log e}{\sqrt{12} \; N \; C_a \; I_b \; {}^{10}\log \frac{I_{max}}{I_b}} \tag{4.21b}$$

in which C_a is the average chromosome chromophore concentration given as

$$C_a = \frac{1}{n_c} \sum_{k,l \in C} \sum \left(\frac{M(k,l)}{A} - \frac{M_b}{A}\right) \tag{4.22a}$$

and C_{max} is the maximum chromosome chromophore concentration given as

$$C_{max} = \frac{M_{max}}{A} - \frac{M_b}{A} \; . \tag{4.22b}$$

In the same way ΔDNA/DNA is obtained by combination of equation (4.17), (4.3a), (4.20) and (4.1) for the logarithmic case

$$\left(\frac{\Delta DNA}{DNA}\right)_{log} = \frac{f_d \sqrt{\frac{1}{n_b} + \frac{1}{n_c}}\, C_{max}}{\sqrt{12}\, N\, C_a\, {}^{10}\log \frac{I_{max}}{I_b}} \,. \tag{4.23}$$

In figure 4.5 ΔDNA/DNA is given as function of I_b/I_{max}, which is a measure of the contrast present in the negative. The curves are given for both the linear and

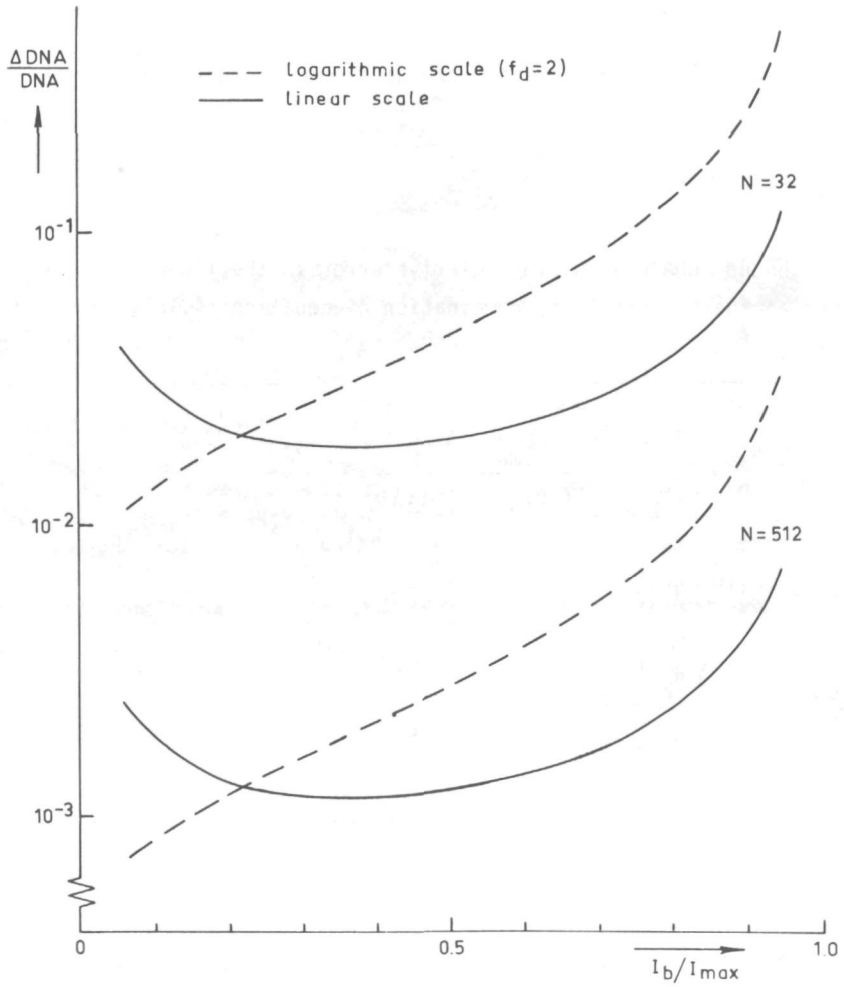


Figure 4.5 Relative error in the DNA content for a linear and a logarithmic scale (n_b = 100, C_a/C_{max} = 0.15)

the logarithmic case for 32 and 512 quantization levels. The curves are plotted for $f_d = 2$, $C_a/C_{max} = 0.15$, $n_b = 100$ and $n_c = 300$. These are values for an average chromosome of the So1A negative (see Section 4.6). $\Delta DNA/DNA$ is an upper bound for σ_{DNA}/DNA in the linear case. For $n_c \gg n_b$ this bound is close because in equation (4.7b) the average standard deviation of the chromosome densities is weighed with $1/n_c$. In the linear case the error has a minimum when I_b/I_{max} equals 1/e, but the minimum is flat. When the contrast present in the negative is of the magnitude of the density range f_d of the logarithmic scale, the logarithmic scale gives the smallest error. However, when the contrast is small compared to the density range, the error of a linear scale is less. When the contrast present in the different negatives to be scanned, varies to a large extent, a linear scale is preferable.

In figure 4.6 $\Delta DNA/DNA$ is given as a function of the number of quantization levels N computed from equations (4.21b) and (4.23) for I_b/I_{max} equals 0.05.

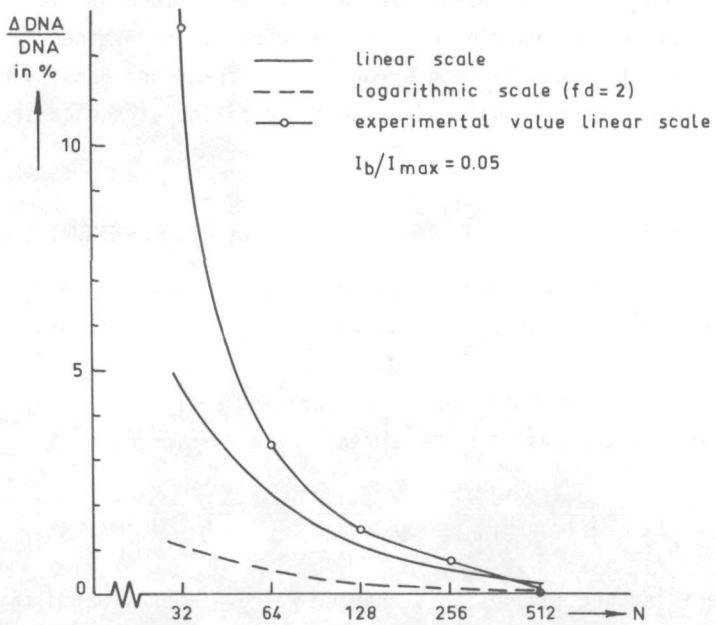


Figure 4.6 Relative error in the DNA content as function of $N(n_b = 100, n_c = 300, C_a/C_{max} = 0.15)$

The same parameter values are used as in figure 4.5. For negatives of Feulgen stained chromosomes I_b/I_{max} is between 0.05 and 0.2. When I_b/I_{max} equals 0.05, it is shown in figure 4.6, that the error in the logarithmic case ($f_d = 2$) at

32 quantization levels is about the same as with 128 quantization levels in the linear case.

In figure 4.6 also the experimentally determined relative root mean square error in the DNA content is given for a linear scale with N quantization levels. For one scan of the SolA negative the intensities were requantized and the relative root mean square error in the DNA content of the metaphase chromosomes was computed with respect to the situation with 512 quantization levels. For $N \geq 64$ quantization levels the experimental values are close to the upper bound, but not below it. This may be caused by the fact that all parameter values are approximations for an 'average chromosome' and that when $n_c \gg n_b$ the upper bound is very close. When $N < 64$ the quantization error in the background is so large, that the quantization errors in the background points may no longer be assumed to be uncorrelated.

Similar formulas are valid for the measurement of the profile. Only there the summation takes place over one stripe instead of over the whole chromosome. Since the position of the centromere is determined by the minimum of the profile, the quantization error of the background density having a systematic character does not influence the position of the centromere.

4.5 THE DISTRIBUTIONAL ERROR IN THE DNA MEASUREMENT

From section 4.2 on we have assumed that the chromophore is homogeneously distributed over the measuring spot. We shall now evaluate deviations from this assumption.

When we assume that the law of Lambert and Beer still holds for subresolvable dimensions, the DNA content of a chromosome is

$$\text{DNA} = \frac{1}{k_a \gamma} \iint_{x,y \in C} {}^{10}\log \frac{I(x,y)}{I_b} \, dxdy. \tag{4.24}$$

In the preceding sections of this chapter we have approximated this integral by (equation (4.4), (4.3b))

$$\text{DNA} = \sum_{k,l \in C} \sum \text{DNA}(k,l) = \frac{h^2}{k_a \gamma} \sum_{k,l \in C} \sum [D_b - D(k,l)] = \frac{h^2}{k_a \gamma} \sum_{k,l \in C} \sum {}^{10}\log \frac{I(k,l)}{I_b} \tag{4.25}$$

The intensity $I(k,l)$ is the average of the intensities $I(x,y)$ across the circular measuring spot. The distributional error is introduced by the inhomogeneous chromophore distribution over the measuring spot, because the

intensity I(x,y) is integrated across the spot instead of the density $^{10}\log I(x,y)$.

For a grid element (k,l) this distributional error is investigated. We also take into account that the intensity is not integrated across the grid element, but across the circular measuring spot contained in the grid element. In the grid element the intensity is locally expanded in a Taylor expansion. When we use only the first two terms, the intensity at a point (x,y) in the grid element (k,l) is

$$I(x,y) = (1+bx+cy)I_{kl} \tag{4.26}$$

with

$x = x-kh$, $y = y-lh$, $I_{kl} = I(kh,lh)$,

$$b = \frac{1}{I_{kl}} \left.\frac{\partial I}{\partial x}\right|_{\substack{x=0\\y=0}} \quad \text{and} \quad c = \frac{1}{I_{kl}} \left.\frac{\partial I}{\partial y}\right|_{\substack{x=0\\y=0}}.$$

I_{kl} is the intensity at the centre of the grid element (k,l).

The contribution DNA(k,l) of grid element (k,l) to the measured DNA content is

$$DNA(k,l) = \frac{h^2}{k_a\gamma}\, {}^{10}\log \frac{I(k,l)}{I_b} = \frac{h^2}{k_a\gamma}\, {}^{10}\log \frac{\iint_{x,y \leq r} I(x,y)\,dx\,dy}{I_b \pi r^2}. \tag{4.27}$$

With the approximation of equation (4.26) DNA(k,l) is

$$DNA(k,l) = \frac{h^2}{k_a\gamma}\, {}^{10}\log \frac{I_{kl}}{I_b}. \tag{4.28}$$

We should have computed for grid element (k,l) $\widetilde{DNA}(k,l)$, given as

$$\widetilde{DNA}(k,l) = \frac{1}{k_a\gamma} \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} {}^{10}\log \frac{I(x,y)}{I_b}\,dx\,dy \tag{4.29}$$

With the approximation of equation (4.26) $\widetilde{DNA}(k,l)$ is computed in Appendix F. For $-1 < bx + cy < 1$, in which x and y are contained in the grid element, $\widetilde{DNA}(k,l)$ is expanded as

$$\widetilde{DNA}(k,l) = \frac{h^2}{k_a\gamma}\, {}^{10}\log \frac{I_{kl}}{I_b} - \frac{^{10}\log e}{k_a\gamma} \sum_{i=2}^{\infty} \frac{h^{2i}}{2^{2i} i(i-1)(2i-1)} \cdot \frac{[(b+c)^{2i}-(b-c)^{2i}]}{bc}. \tag{4.30}$$

In first order approximation $\widetilde{DNA}(k,l)$ is

$$\widetilde{DNA}(k,l) \cong \frac{h^2}{k_a\gamma} \, {}^{10}\log \frac{I_{kl}}{I_b} - \frac{h^4(b^2 + c^2)}{24} \cdot \frac{{}^{10}\log e}{k_a\gamma} \, . \tag{4.31}$$

The distributional error $(\Delta DNA)_{dist}$ for grid element (k,l) is obtained from equation (4.31), (4.28) and (4.26) and is in first order approximation

$$(\Delta DNA)_{dist} = \widetilde{DNA}(k,l) - DNA(k,l) = - \frac{h^4(g_{kl}^2)}{24} \cdot \frac{{}^{10}\log e}{k_a\gamma} \tag{4.32a}$$

with $g_{kl}^2 = b^2 + c^2 = \left(\frac{\partial D}{\partial x}\right)_{kl}^2 + \left(\frac{\partial D}{\partial y}\right)_{kl}^2$. (4.32b)

The relative distributional error in the total DNA content is

$$\left(\frac{\Delta DNA}{DNA}\right)_{dist} = - \frac{h^2 \sum\limits_{k,l \in C} \sum g_{kl}^2}{24 \sum\limits_{k,l \in C} \sum ({}^{10}\log I_{kl} - {}^{10}\log I_b)} \, . \tag{4.33}$$

Equation (4.33) shows that in the first approximation the distributional error is proportional to the square of the grid constant h and to the sum of the squares of the gradients of the density.

An estimation of the distributional error in the negatives can be obtained from equation (4.33). The Sol A negatives (average density 0.2) have a maximum value of ΔD of about 0.2 for an extension of 0.18 μm in the specimen (70 μm in the negative). The 744 negatives (average density 0.1) have a maximum value of ΔD of about 0.06 for an extension of 0.13 μm in the specimen (40 μm in the negative). The gradient of the points located at the slope of the chromosome (about 1/3 of the chromosome points) is approximated by the maximum gradient. For the rest of the points it is assumed that the gradient is zero. This results in a distributional error for a stepsize of 0.2 μm in the preparation given as

Sol A negatives 0.8%
744 negatives 0.3%.

At the stepsizes which were used the distributional error is comparatively small. See also section 4.6.

4.6 FURTHER EXPERIMENTS AND RESULTS

To investigate the error present in repeated scans in relation to the grid constant, two metaphases (codenumber Sol A and 744) were photographed and repeatedly scanned with different grid constants. The diameter of the measuring spot was equal to the grid constant in all cases. One metaphase (744) was photographed nine times and scanned to investigate the photographic error. Ten different metaphases of one subject were photographed and scanned to obtain the coefficient of variation in homologue chromosomes. The measured features were DNA content, DNA ratio, length and centromeric index (C.I.). These features were measured with the methods described in chapter 3.

The standard deviation in a feature of an experiment is computed as follows: Let u_i^j be the value of feature u of chromosome i in metaphase scan j. The mean value of u_i^j for chromosome i is

$$\bar{u}_i = \frac{1}{n_m} \sum_{j=1}^{n_m} u_i^j \tag{4.34}$$

in which n_m is the total number of metaphase scans.
The variance in u_i for a chromosome i over the experiments involved is

$$s_i^2 = \frac{\sum_{j=1}^{n_m} (u_i^j - \bar{u}_i)^2}{n_m - 1} . \tag{4.35}$$

To determine the variance in a feature in a particular metaphase scan j, we have to group a number of different chromosomes in order to obtain some statistical significance. When we divide the 46 chromosomes in groups G_k, containing n_{G_k} chromosomes, we can compute the variance in group G_k in a certain metaphase scan j as

$$s_{G_k,j}^2 = \frac{\sum_{i \in G_k} (u_i^j - \bar{u}_i)^2}{n_{G_k}} . \tag{4.36}$$

To compute $s_{G_k,j}^2$ we assume that the variance in feature u is the same for the chromosomes of group G_k. In our experiment the 46 human chromosomes were divided into three groups: G_1 consisting of the chromosomes of the A and B group according to the Denver system, G_2 consisting of the C group and G_3 consisting of the D, E, F and G group.

Centromere positions which have not been placed correctly, result in outliers in the DNA ratio and the centromeric index. In experiments with samples from which the class is known, outliers can be detected by their deviation from the class mean. However, as methods were developed for the situation in which the class is not known,criteria based on individual measurements had to be found, to detect possible outliers. These criteria were evaluated with samples from which the class was known. Measurements were regarded to be outliers when their deviation from the class mean was over 20%. Four criteria were investigated:

a) The symmetry of the profile,
b) The quadratic deviation between the centromere position of the total profile and those of the profiles of the partial chromosome on each side of the principal axis (or the best fit polynomial),
c) The depth of the minimum in the profile or in the first derivative of the profile for acrocentric chromosomes,
d) A threshold for the DNA ratio of acrocentric chromosomes.

A combination of methods c and d detected the highest number of outliers. To accept a centromere position, the final conditions to be fulfilled are

1) the relative depth of the minimum in the profile or in the first derivative of the profile for acrocentric chromosomes must be less than 0.9
2) the DNA ratio must be larger than 0.7, when the centromere is located with the routine for acrocentric chromosomes

These two final conditions were used as outlier criterium for all experiments described in this thesis.

In table 4.1a the coefficient of variation is given in the feature values of the negative SolA for repeated scans with different grid constants. The coefficient of variation is calculated according to equation (4.36). Calculated as if the grid was present at the original specimen, the grid constants are: 0.13, 0.15, 0.18, 0.21 (twice), 0.26 and 0.31 μm. The number of chromosomes used is placed between brackets. In table 4.1b the coefficient of variation is given in the feature values of the negative 744-10 for repeated scans. The grid constants involved are 0.06, 0.09, 0.12, 0.16 and 0.22 μm. In table 4.1c the coefficient of variation in the feature values of the total metaphases is given for the grid constants, for which this coefficient of variation is minimum.

In table 4.2 the coefficient of variation is given in the features of nine different negatives of the metaphase 744 (grid constant 0.12 μm). In table 4.3a the homologue coefficient of variation is given for ten different metaphases of one subject labeled by the cytologist. In table 4.3b the homologue coefficient of

Table 4.1a Coefficient of variation for different grid constants, negative So1A

group	DNA content	DNA ratio	length	C.I.
A,B	0.8% (70)	0.4% (70)	0.6% (70)	0.6% (70)
C	0.9% (105)	0.5% (105)	0.5% (105)	0.6% (105)
D, E, F, G	1.7% (122)	1.4% (76)	2.5% (122)	2.4% (76)
total	1.2% (297)	0.9% (251)	1.7% (297)	1.4% (251)

Table 4.1b Coefficient of variation for different grid constants, negative 744-10

group	DNA content	DNA ratio	length	C.I.
A,B	0.8% (46)	2.1% (40)	1.2% (46)	2.5% (40)
C	1.1% (64)	0.9% (49)	1.6% (64)	1.0% (49)
D, E, F, G	1.4% (75)	1.9% (46)	2.0% (75)	2.6% (46)
total	1.2% (185)	1.7% (135)	1.7% (185)	2.1% (135)

Table 4.1c Grid constant for which the coefficient of variation of the total metaphase is minimum

		coefficient of variation			
negative	grid constant	DNA content	DNA ratio	length	C.I.
So1A	0.21 μm	0.6% (43)	0.7% (39)	1.2% (43)	1.1% (39)
744-10	0.16 μm	0.9% (39)	1.0% (25)	1.3% (39)	1.0% (25)

Table 4.2 Coefficient of variation for different negatives, metaphase 744 (stepsize 0.12 μ)

group	DNA content	DNA ratio	length	C.I.
A,B	1.5% (88)	2.7% (86)	1.5% (88)	2.1% (86)
C	2.0% (115)	2.9% (92)	1.5% (115)	2.2% (92)
D, E, F, G	2.8% (143)	3.4% (82)	2.1% (143)	4.4% (82)
total	2.2% (346)	3.0% (260)	1.8% (346)	3.0% (260)

Table 4.3a Coefficient of variation for the homologue chromosomes of 10 metaphases of one subject, classified by the cytologist

group	DNA content	DNA ratio	length	C.I.
A,B	6.0% (89)	6.7% (79)	5.8% (89)	5.7% (79)
C	6.9% (123)	6.5% (101)	7.3% (123)	5.9% (101)
D, E, F, G	7.3% (166)	12.9% (92)	10.4% (166)	12.5% (92)
total	6.9% (378)	9.2% (272)	8.6% (378)	8.7% (272)

Table 4.3b Coefficient of variation for the homologue chromosomes of 10 metaphases of one subject, classified automatically (Chapter 5)

group	DNA content	DNA ratio	length	C.I.
A,B	5.2% (87)	6.0% (78)	6.1% (87)	5.2% (78)
C	4.0% (126)	3.0% (102)	6.7% (126)	3.4% (102)
D, E, F, G	4.8% (165)	8.4% (92)	9.9% (165)	9.5% (92)
total	4.7% (378)	6.1% (272)	8.2% (378)	6.5% (272)

variation is given for the same metaphases, classified automatically based on DNA content and DNA ratio, with the methods described in chapter 5. So table 4.3a is affected by the possible errors of the cytologist and table 4.3b by the errors of the program.

First of all we consider the coefficient of variation in the DNA content. The errors present in repeated scans of the same negative (table 4.1a, 4.1b and 4.1c) are due to photodiode and pre-amplifier noise, quantization noise, the distributional error and the photographic error, as far as these last two errors have a stochastic character. Although the distributional error and the photographic error are systematic, these errors will have a stochastic component in repeated scans. In the repeated scans the negative is remounted, and the spot position will differ from scan to scan.

The quantization error was given in figure 4.6 of section 4.4. For the present situation (I_b/I_{max} between 0.05 and 0.2, D_a between 0.1 and 0.2, $n_c \cong 300$, $n_b \cong 100$, 512 quantization levels), this error will result in a coefficient of variation of about 0.2%. The measured noise of the photodiode and the pre-amplifier will result in a coefficient of variation of about 0.7% for the present situation (section 4.3). In section 4.5 it was shown, that the systematic distributional error is between 0.3% and 0.8%. In the remainder of this section it is shown, that the photographic error is about 2%. The stochastic components of the last two errors, however, are less. The coefficient of variation in repeated scans will be due to the combination of these four errors, and is according to tables 4.1a, 4.1b and 4.1c of the right order of magnitude.

The method of Hartley (1950) was used to test whether an essential difference in the variance for the different grid constants was present or not. With a confidence level of 5% the differences were just significant, showing a tendency to increase at the higher and the lower values of the grid constant. In section 4.5 it was shown, that the distributional error increases with increasing grid constant. The error given in equation (4.9) will also increase with decreasing n_b and n_c, so for higher values of the grid constant. For small values of the grid constant oversampling occurs, and the grain noise of the negative becomes important (Van der Ploeg et al. (1974)). In table 4.1c the minimum coefficient of variation is given for the complete metaphase.

The rank correlation test of Spearman (1904) was used to test, if the mean chromosome DNA content increases with increasing grid constant. According to equation (4.33) this tendency due to the distributional error can be expected. For the negative Sol1A a positive correlation was present with a confidence level of 5%.

For the negative 744-10 the distributional error was too small to be detected.

The coefficient of variation, given in table 4.2 for different negatives of the same metaphase, consists both of the errors mentioned earlier (photodiode and pre-amplifier noise, quantization noise, and the distributional error) and of the errors present in the photographic process and the first microscopic system. In section 4.3 it was already shown, that the errors in the first microscopic system were comparatively small. When we compare table 4.1b and 4.1c with table 4.2, we see that the photographic process does certainly contribute to the coefficient of variation in the DNA content (from 1.2% to 2.2%).

The negative of a grey-wedge was scanned to investigate the error due to the photographic process. In figure 4.7 the coefficient of variation in the measured intensity of this negative is given. The negative was scanned with a grid constant equivalent to 0.12 μm in the specimen. I_b/I_{max} is between 0.05 and

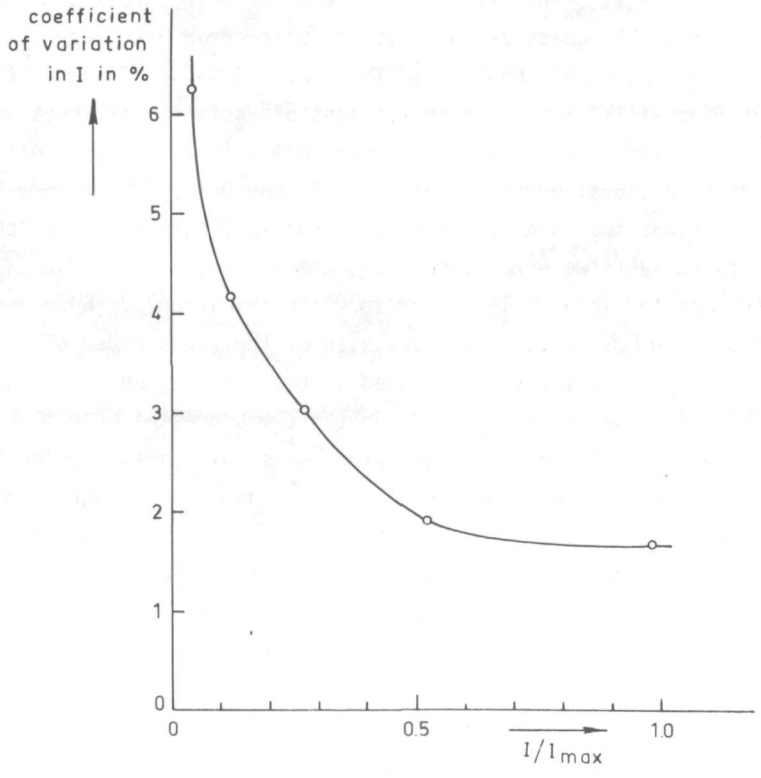


Figure 4.7 Coefficient of variation in the measured intensity for a scanned negative of a grey-wedge. (equivalent grid constant 0.12 μm)

0.2 for Feulgen stained metaphases. From equation (4.9) and figure (4.7) we may calculate the coefficient of variation in the DNA content due to the photographic process. For the present situation (D_a between 0.1 and 0.2, $n_c \cong 300$, $n_b \cong 100$) this photographic error results in a coefficient of variation in the DNA content of about 1.7%, which agrees with the values given in table 4.2.

Formula (4.9) shows that $\Delta DNA/DNA$ increases with decreasing average chromosome density D_a. So it is important to fit the regions for which the chromosome densities are summed as close as possible to the chromosome. For too large regions the summation over the added background densities will contribute to the amount of noise in the DNA content, but not to the DNA content itself. In figure 3.5 the average densities of the extensions (defined in chapter 3) of three chromosomes of negative 744-10 were given. This figure shows that the extensions reach the background at about 0.7 μm in the original preparation. To compute the DNA values of table 4.2 the extensions up to 0.7 μm were considered to belong to the chromosome, and the extensions over 0.7 μm were used to estimate the background density. An increase of the regions of the chromosomes to 1.2 μm gives an increase in the coefficient of variation in the DNA content to 3.4% (2.2% in table 4.2) illustrating the importance of a narrow region.

When we compare the homologue variation of the chromosomes classified by the cytologist (table 4.3a), or by the program (table 4.3b), with the variation present in the negatives, we see that the measuring errors are less than the homologue variation. Table 4.3a and table 4.3b show that automatic classification yields a smaller homologue variation in the DNA content and the DNA ratio than classification by the cytologist. The homologue variation, however, was calculated for the features that were also used for the automatic classification.

The DNA ratio R is computed as

$$R = \frac{DNA_L}{DNA} \tag{4.37}$$

with DNA content DNA_L of the longer chromosome-arm. The relative error in the DNA ratio is (when we assume independence between DNA_L and DNA)

$$\frac{\Delta R}{R} = \frac{\Delta DNA_L}{DNA_L} + \frac{\Delta DNA}{DNA} + \frac{\Delta x'_c \; P_c}{DNA_L} \tag{4.38}$$

with an error $\Delta x'_c$ in the centromere position and a value P_c of the DNA profile at the centromere position. As the systematic error in the background has the

same influence on DNA_L and DNA, this error will partly be compensated in the ratio.

When we compare table 4.1a, 4.1b and 4.1c, we see, that the coefficient of variation in repeated scans of negative So1A in the DNA ratio is less than the variation in the negative 744-10. As the variation in the DNA content is about the same, this indicates that the centromere position was more difficult to find in this metaphase 744 than in the metaphase So1A. The metaphase 744 had elongated bended chromosomes and had less contrast than the metaphase So1A.

The photographic process does also increase the specific variation in the DNA ratio (from 1.68% to 3.02%). The measuring error, however, is less than the homologue variation given in table 4.3a and 4.3b.

When the length of a chromosome is determined from the boundary, the length depends on the dissection level defining the boundary. In this investigation the length was obtained from the profile. In the profile method an analogous problem is present in the definition of the end points of the profile. The noise is reduced compared to that in the boundary defined length due to the summation over the points in a profile stripe. The end points of the profile were defined as the points where the profile density drops below 0.1 of its maximum value. The threshold of 0.1 was experimentally obtained by investigation of the influence of a variation in the end points. The lowest value of the threshold was taken at which the slope of the profile was steep enough to ensure a reliable detection of the end points. Interpolation between profile points has been used for more accurate determination of the end points.

Significant errors in the length are the errors in the position of the end points, caused by the noise present in the profile and the error in the fit of the principal axis (or best fit polynomial) to the chromosome. As within one metaphase the density distribution in the arm ends will hardly be dependent on chromosome length, the noise in the ends of the density profile will be the same too. So it was expected that the error in the length should be more or less constant, hence the coefficient of variation should increase with decreasing length. This tendency is present, but not distinct. This can be explained by the fact that the fit of the axis (or the polynomial) along which the length is measured will be better for the smaller than for the larger chromosomes and dominate the error in the end points for the latter. Table 4.2 shows that the coefficient of variation due to the photographic process in the length is of the same magnitude as in the DNA content.

As result of the contraction it might be expected, that the homologue variation in the length should be large. In table 4.3a, however, the homologue variation in the length is of the same magnitude as the homologue variation in the DNA content. This can be explained by the fact that length measurements are less sensitive to errors due to the specimen, as e.g. differences in stoichiometry and absorbing materials, than the measurement of DNA contents. The value of table 4.3b is limited in this respect, because the automatic classification was based on DNA content and DNA ratio.

The centromeric index C.I. is defined as the ratio between the length of the long arm and the total length of a chromosome, in the same way as the DNA-ratio is the ratio between the DNA contents. The relative error in the centromeric index is similar to equation (4.38) for the DNA ratio, with length instead of DNA content and P_c equal to 1. The error in the centromere position is again $\Delta x'_c$. So when P_c equals roughly the average profile density, the influence of the error in the position of the centromere $\Delta x'_c$ is the same in the DNA ratio and the centromeric index. In tables 4.1, 4.2 and 4.3 it is shown that the coefficient of variation in the DNA ratio and in the centromeric index is of the same magnitude.

4.7 CONCLUSIONS

In this chapter the different sources of error in the computation of DNA based features have been investigated.

The influence of the intensity quantization on the DNA content is compared for a linear and a logarithmic scale. The error in the DNA content for a linear scale has a minimum when the ratio between background and maximum intensity equals 1/e, but the minimum is flat. When the contrast present in the negative fits the density range of the logarithmic scale well, the logarithmic scale gives the smallest error. However, when the contrast is small compared to this range, the error of a linear scale is less. When the contrast present in the negatives varies to a large extent, a linear scale is preferable.

The magnitude of the distributional error for the grid constants (0.15 μm to 0.2 μm) proved to be relatively small compared to the other sources of error.

The photographic process does increase the coefficient of variation in the features with respect to the variation introduced by quantization noise and noise in the photosensitive device.

To keep the errors in the DNA content small, it is important to fit the region for which the densities are summed as close as possible to the chromosome.

The homologue variations in DNA content and DNA ratio are greater than the coefficient of variation introduced by the measuring process. The coefficients of variation introduced by the measuring process in the length features are of the same magnitude as those in the corresponding DNA features. The homologue variation in the length, however, is greater than that in the DNA content, while the homologue variation in the centromeric index is of the same magnitude as that in the DNA ratio.

Chapter 5

DNA BASED HUMAN CHROMOSOME CLASSIFICATION

5.1 INTRODUCTION

Many books and papers have been published on pattern recognition and classification theories. Well-known are the books of Duda et al. (1973), Meisel (1973), Fukunaga (1972), Patrick (1972) and Fu (1976). In statistical pattern recognition two approaches can be distinguished: parametric and non-parametric classification. In the parametric methods it is assumed, that the distribution of the features is known. From the learning samples the parameters of the distribution are estimated. In the non-parametric methods, the distribution is not known. The distribution may be estimated by e.g. density estimation or Parzen windows. An important non-parametric method in which no distribution is estimated, is the nearest neighbour method. This method assigns an unknown sample to the class of the nearest learning sample. Many variants of this nearest neighbour method exist.

The classification rule of some parametric and non-parametric methods can be expressed by a linear separation in the feature space.

In pattern recognition the classification system is based on the learning samples. When the number of learning samples is too small, the classification results with the learning samples are not representative for the performance of the classification system. Therefore a number of independent test samples must be used to evaluate the system. When the total number of samples is quite limited a so-called n-1 method might be used. In this method one sample of all the available samples is subsequently used as a test sample and all other samples are used as learning samples.

Classification of human chromosomes has some special characteristics. In a normal metaphase 46 chromosomes are present. Generally not all chromosomes can be analysed, because chromosomes may e.g. touch or overlap. So when normal

metaphases are scanned 46 chromosomes will be found at most. In a normal metaphase the 22 autosomes are present in homologue pairs. In addition to these autosomes a female has two X chromosomes and a male one X and one Y chromosome. The information that of each type of chromosome only one pair can be present in a scanned metaphase can be taken into account in the classification. This is only possible, however, when no aberrations are present in the number of chromosomes in the karyotype.

In this chapter classification results based on DNA content and DNA ratio are given. The classes we will consider consist of the 22 autosomes and the 2 sex chromosomes. In these experiments five subjects are involved. At least six metaphases of each subject are used. The influence of the constraint, that of each type of chromosome only one pair can be present is investigated. The classification results based on DNA content and DNA ratio are compared to the classification results based on length and centromeric index and combinations of these features.

The main errors in the DNA based features are treated in chapter 4. The DNA content and the DNA profile are obtained by summation of the densities of the grid elements for a chromosome or chromosome stripe. So the errors in the features consist of the sum of a large number comparable but not necessarily dependent errors. According to the central limit theorem, there is reason to expect, that the features are in approximation normally distributed. Therefore we used a parametric classification method, in which it is assumed that the features are normally distributed. The mean and the covariance matrix of the distributions were estimated from the learning samples.

5.2 BAYES DECISION THEORY

Suppose that one wants to classify individual objects i into classes Ω_j, based on the measured feature vectors $\underline{u}_i$. It is assumed that the conditional probability density functions $p(\underline{u}_i|\Omega_j)$ are known for the classes Ω_j. The Bayes rule relates the a posteriori probability $p(\Omega_j|\underline{u}_i)$ to the a priori probability $p(\Omega_j)$ as

$$p(\Omega_j|\underline{u}_i) = \frac{p(\underline{u}_i|\Omega_j)\ p(\Omega_j)}{p(\underline{u}_i)} \tag{5.1}$$

and

$$p(\underline{u}_i) = \sum_{j=1}^{Q} p(\underline{u}_i|\Omega_j)\ p(\Omega_j) \tag{5.2}$$

in which Q is the number of classes.

It can be shown, that the minimum error rate is obtained, when the object is classified in that class Ω_k for which the a posteriori probability is maximum (Duda et al. (1973)). So

$$\text{decide } \Omega_k \text{ if } p(\Omega_k | \underline{u}_i) > p(\Omega_j | \underline{u}_i), \text{ for all } k \neq j. \tag{5.3}$$

As the probability density function $p(\underline{u}_i)$ is independent of Ω_j, equation (5.3) can be rewritten as:

$$\text{decide } \Omega_k \text{ if } p(\underline{u}_i | \Omega_k)\, p(\Omega_k) > p(\underline{u}_i | \Omega_j)\, p(\Omega_j), \text{ for all } k \neq j. \tag{5.4}$$

A misallocation will produce a certain loss. Let $\Lambda(\Omega_k | \Omega_1)$ be the loss, when the true class is Ω_1 and the allocated class is Ω_k. The conditional risk Ξ of the allocation Ω_k is

$$\Xi(\Omega_k | \underline{u}_i) = \sum_{j=1}^{Q} \Lambda(\Omega_k | \Omega_j)\, p(\Omega_j | \underline{u}_i). \tag{5.5}$$

A minimum risk is obtained by the decision rule:

$$\text{decide } \Omega_k \text{ if } \Xi(\Omega_k | \underline{u}_i) < \Xi(\Omega_j | \underline{u}_i), \text{ for all } k \neq j. \tag{5.6}$$

5.3 CLASSIFICATION OF A SET OF OBJECTS

Instead of subsequently classifying the individual chromosomes, the chromosomes of one metaphase may be classified as a whole. The fact that of each type of chromosome only one pair can be present in a metaphase is then taken into account. This means that we have a set of 46 chromosomes of one metaphase, which must be assigned to the 22 pairs of autosomes and the sex chromosomes.

Generally we consider a set of N objects, originating from the classes Ω_i (i=1, ..., Q). The known number of objects in the set originating from class Ω_i is N_i. So $N = \sum_{i=1}^{Q} N_i$. The N-tuple $u = (\underline{u}_1, \ldots, \underline{u}_N)$ are the measured feature vectors of the N objects. It is assumed, that the N objects are independent. Let v_i indicate, that the object with measured feature vector $\underline{u}_i$ originates from class Ω_i, given by the N-tuple $\Upsilon = (v_1, \ldots, v_N)$. So $v_i \in (\Omega_1, \ldots, \Omega_Q)$ for i = 1, ..., N. The number of elements in Υ of class Ω_i is N_i(i=1, ..., Q). The N-tuple $\in = (e_1, \ldots, e_N)$ indicates the allocated class of the N objects.

BAYES METHOD

The number of misallocations χ is

$$\chi = \sum_{i=1}^{N} \delta(e_i, v_i) \tag{5.7a}$$

in which

$$\delta(e_i, v_i) = 0 \text{ if } e_i = v_i \qquad 1 \text{ if } e_i \neq v_i, \quad \text{for } i=1, \ldots, N. \tag{5.7b}$$

The mean number of misallocations for all possible Υ is

$$\overline{\chi} = \sum_{\forall \Upsilon} \sum_{i=1}^{N} \delta(e_i, v_i)\ p(\Upsilon|\underline{u}_i) \tag{5.8}$$

Slot (1976) has shown that the e_i's in equation (5.8) can be chosen independently and that the allocation criterion becomes:

decide e_i for $i=1, \ldots, N$ so that

$$\sum_{\forall \Upsilon} \delta(e_i, v_i)\ p(\underline{u}_i|\Upsilon) \text{ is minimum.} \tag{5.9}$$

There are $N!/(N_1!\, N_2!\, \ldots\, N_Q!)$ different N-tuples Υ. So for large values of N and Q, this classification rule is not useful in practice.

MAXIMUM LIKELIHOOD METHOD

Slot (1976) suggests a method in which the likelihood function L(u) is maximized. This likelihood rule is

decide e_i for $i=1, \ldots, N$ so that

$$L(u) = \prod_{i=1}^{N} p(e_i|\underline{u}_i) \text{ is maximum} \tag{5.10a}$$

with the restriction that

$$\sum_{i=1}^{N} \delta(e_i, \Omega_j) = N_j \text{ for all } j. \tag{5.10b}$$

An exhaustive search involves the same number of possibilities of Υ as the Bayes method. In this likelihood method an exhaustive search is not necessary.

The use of the logarithm of the likelihood function results in a sum to be maximized with restrictions. This can be achieved by linear programming, but it remains laborious.

For the classification experiments a very fast exchange algorithm was used instead of linear programming to reduce the computation time. This exchange algorithm does, however, not necessarily converge to the optimal solution. The exchange algorithm starts with an arbitrary classification of the chromosomes with the correct number of homologue pairs. The classification is improved by exchanging two chromosomes from different homologue pairs, if the likelihood function increases because of this action. This is repeated until no exchange which increases the likelihood function can be found.

5.4 CLASSIFICATION OF METAPHASES ON DNA BASED FEATURES

The chromosomes are classified with the Bayes method, described in section 5.2 and with the maximum likelihood method described in section 5.3. It is assumed that the chromosome features are normally distributed. The conditional probability density function of the features is

$$p(\underline{u}_i|\Omega_j) = \frac{1}{(2\pi)^{K/2}|\Sigma_j|^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2}(\underline{u}_i - \underline{\mu}_j)^t \ \Sigma_j^{-1} \ (\underline{u}_i - \underline{\mu}_j)\right] \qquad (5.11)$$

in which K is the number of features, $\underline{\mu}_j$ is the mean vector of class Ω_j and Σ_j is the covariance matrix of class Ω_j. For two features equation (5.11) is identical to

$$p(u_{1i}, u_{2i}|\Omega_j) = \frac{1}{2\pi\sigma_{1j}\sigma_{2j}\sqrt{1-\rho_j}} \exp\left[-\frac{1}{2(1-\rho_j)^2}\left\{\frac{(u_{1i} - \mu_{1j})^2}{\sigma_{1j}^2} + \right.\right.$$

$$\left.\left. - \frac{2\rho_j(u_{1i} - \mu_{1j})(u_{2i} - \mu_{2j})}{\sigma_{1j}\sigma_{2j}} + \frac{(u_{2i} - \mu_{2j})^2}{\sigma_{2j}^2}\right\}\right] \qquad (5.12)$$

in which ρ_j is the correlation coefficient between the two features of class Ω_j.

A n-1 method was used for the classification. In this method the metaphases of one subject are used as a test set and the metaphases of the other subjects as a learning set. This is repeated for all subjects, so the metaphases of all subjects are used as a test set once.

In the classification of chromosomes of a metaphase additional problems may arise

a) incomplete measured features

b) chromosomes of which no features could be measured, because the chromosomes did touch, overlap, or were not present in the scanned area. We will call these chromosomes: missing chromosomes.

INCOMPLETE MEASURED FEATURES

When the centromere position can not be determined, the DNA ratio and the centromeric index will be unknown. In this case the chromosomes may be classified, but only based on the features which could be measured. This implies that for all classes a covariance matrix must be inverted for all possible missing features (equation (5.11)). This is an elaborate method when a large number of features is involved. The method is used as at most four features were considered in the classification experiments.

A sub-optimal solution to the problem is the classification based on $p(\underline{u}_i|\Omega_j)$, in which the mean feature value in Ω_j is inserted for the missing feature. In this method the inversion of covariance matrices for each missing feature is avoided.

MISSING CHROMOSOMES

The missing chromosomes are only important for those classifications, in which the number of metaphase chromosomes is taken into account. The exchange algorithm used in this case starts from an arbitrary classification with the correct number of homologue pairs. In this classification dummy chromosomes are inserted to obtain a total number of 46 chromosomes. The dummy chromosomes are equally probable to all classes. These dummy chromosomes inserted for the missing chromosomes, reduce the influence of taking the number of homologue pairs into account.

5.5 CLASSIFICATION EXPERIMENTS AND RESULTS

The metaphases of five subjects were used in the experiments described. At least six metaphases of each subject were scanned. The mean value and the standard deviation in the measurement of the four features treated in the previous chapters were computed for each subject and for the total of all subjects. These values and the number of chromosomes involved are given for the

Table 5.1 Mean and standard deviation in the DNA content (times 1000)

chromosome	subject 1			subject 2			subject 3			subject 4			subject 5			total		
	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number
1	1036	57	9	1020	48	15	1095	77	17	1080	81	8	1013	39	9	1051	72	58
2	1005	26	7	991	35	12	998	74	19	1000	38	8	1012	26	12	1001	50	58
3	821	29	8	842	29	13	824	41	17	817	21	10	827	28	11	827	33	59
4	772	28	8	766	45	12	777	34	18	788	35	7	755	50	11	771	41	56
5	800	34	5	788	30	13	773	41	18	758	50	9	733	62	12	768	50	57
6	725	63	11	705	46	14	710	48	17	705	16	8	711	37	12	711	47	62
7	665	33	7	650	44	11	628	44	17	651	47	11	647	35	10	645	44	56
8	607	35	10	605	54	13	609	43	13	605	37	9	619	85	11	609	55	56
9	581	38	9	574	47	17	596	50	18	585	20	7	563	61	10	581	49	61
10	586	43	10	570	34	16	599	49	16	574	30	9	550	39	14	576	44	65
11	569	26	9	562	19	16	571	29	16	576	24	8	551	31	10	566	27	59
12	556	29	9	549	44	18	560	27	18	568	17	8	539	39	11	554	35	64
13	451	24	10	441	25	15	450	24	16	449	23	6	429	19	13	443	25	60
14	422	29	8	415	17	17	426	31	14	435	12	9	414	33	14	421	27	62
15	439	34	9	411	22	14	426	32	16	427	29	7	402	36	12	420	33	58
16	400	23	12	385	19	14	398	24	14	383	16	9	386	21	13	391	23	62
17	366	33	10	349	32	16	360	27	19	376	25	12	359	32	11	361	31	68
18	338	17	11	326	15	16	328	31	17	337	14	7	321	16	12	329	22	63
19	282	21	11	270	20	15	282	21	16	281	19	8	255	42	12	274	28	62
20	288	25	11	277	21	17	282	24	18	298	33	8	276	32	10	283	27	64
21	216	20	12	197	14	16	209	16	16	198	17	4	182	27	10	202	22	58
22	237	16	9	208	19	16	217	10	13	227	29	7	208	25	9	217	23	54
X	657	25	6	638	52	5	650	43	8	632	37	9	613	59	9	636	48	37
Y	188	24	5	246	20	9	223	15	7							224	30	21

Table 5.2 Mean and standard deviation in the DNA ratio (times 1000)

chromosome	subject 1			subject 2			subject 3			subject 4			subject 5			total		
	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number
1	522	17	8	526	17	15	533	23	14	540	19	5	533	21	9	530	21	51
2	620	13	5	602	32	12	615	48	16	601	51	4	616	43	12	612	42	49
3	548	30	8	546	29	13	554	55	16	536	23	7	536	11	10	546	37	54
4	718	28	7	718	31	12	727	35	17	686	67	4	727	8	10	720	35	50
5	727	39	5	717	48	11	714	35	16	722	43	8	722	52	12	719	43	52
6	644	39	9	643	41	13	639	48	15	663	19	5	625	22	8	642	40	50
7	608	52	7	621	43	11	638	35	13	632	42	7	622	31	9	626	42	47
8	659	36	10	672	41	10	635	46	12	666	43	4	691	32	11	664	45	47
9	665	32	8	644	39	15	645	42	15	674	14	6	680	94	8	652	38	52
10	709	28	8	690	13	13	677	44	14	632	44	5	696	36	13	686	40	53
11	598	13	9	625	48	14	623	46	14	613	14	3	598	31	10	613	40	50
12	733	36	9	737	57	15	703	36	12	697	40	7	733	17	8	721	45	51
13	836	65	4	887	56	7	901	34	9	890	84	4	907	82	7	889	67	31
14	883	28	4	840	100	7	924	20	6	820	140	6	911	17	2	869	96	25
15	871	8	4	887	35	7	837	64	11	874	39	3	884	22	3	863	51	28
16	605	38	9	615	33	12	600	81	13	594	41	6	606	24	11	604	50	51
17	676	64	9	670	79	8	700	120	10	609	40	5	677	34	6	673	85	38
18	761	17	3	810	130	4	749	53	5	700	180	2	774	56	4	764	99	18
19	620	110	11	590	38	8	620	120	9	534	22	4	556	21	10	592	88	42
20	680	160	6	610	120	13	610	140	8	558	18	3	571	41	4	620	130	34
21	825	75	4	850	68	11	867	67	11	772	22	3	865	79	5	848	73	34
22	830	100	5	855	59	6	841	82	9	821	80	3	857	47	8	845	75	31
X	628	26	6	628	26	4	637	30	6	681	49	8	639	18	9	645	43	33
Y	917	43	3	868	13	2	959		1							908	45	6

Table 5.3 Mean and standard deviation in the length (times 1000)

chromosome	subject 1			subject 2			subject 3			subject 4			subject 5			total		
	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number
1	1100	63	9	1076	57	15	1124	66	17	1092	64	8	1090	35	9	1098	62	58
2	985	47	7	1012	54	12	985	55	19	1009	54	8	1009	21	12	999	50	58
3	864	20	8	832	39	13	863	50	17	844	22	10	849	30	11	851	39	59
4	796	28	8	801	24	12	801	54	18	792	16	7	784	22	11	796	37	56
5	748	56	5	778	38	13	809	41	18	795	27	9	771	27	12	786	42	57
6	761	27	11	741	32	14	766	49	17	763	11	8	751	18	12	756	34	62
7	687	26	7	676	34	11	689	63	17	720	48	11	657	34	10	686	51	56
8	651	30	10	624	53	13	649	59	13	640	30	9	650	28	11	642	45	56
9	621	38	9	593	34	17	647	42	18	649	34	7	616	22	10	623	42	61
10	598	33	10	605	42	16	645	54	16	627	36	9	599	27	14	615	45	65
11	601	35	9	599	26	16	637	44	16	637	18	8	590	20	10	613	38	59
12	598	38	9	601	33	18	634	37	18	633	21	8	593	30	11	612	38	64
13	505	22	10	482	33	15	531	63	16	499	25	6	490	25	13	502	44	60
14	484	49	8	481	30	17	502	32	14	486	21	9	477	26	14	486	33	62
15	474	35	9	455	42	14	507	30	16	498	45	7	460	17	12	478	41	58
16	445	26	12	440	25	14	495	59	14	464	33	9	442	26	13	458	43	62
17	452	22	10	419	41	16	477	51	19	474	51	12	424	22	11	451	49	68
18	398	18	11	400	47	16	439	32	17	416	16	7	390	13	12	410	36	63
19	371	25	11	348	31	15	391	48	16	383	30	8	346	19	12	368	38	62
20	356	25	11	362	33	17	398	36	18	389	35	8	350	16	10	373	37	64
21	309	30	12	292	40	16	344	41	16	332	42	4	277	18	10	310	44	58
22	329	21	9	299	50	16	345	41	13	322	28	7	297	15	9	318	42	54
X	701	20	6	655	30	5	686	24	8	659	49	9	661	36	9	671	39	37
Y	292	18	5	334	40	9	345	51	7							328	46	21

Table 5.4 Mean and standard deviation in the centromeric index (times 1000)

chromosome	subject 1			subject 2			subject 3			subject 4			subject 5			total		
	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number	mean	st. deviation	number
1	508	9	8	514	18	15	522	23	14	515	6	5	516	22	9	515	20	51
2	611	15	5	601	20	12	598	22	16	578	34	4	598	50	12	598	32	49
3	525	21	8	530	32	13	540	48	16	522	24	7	526	20	10	530	35	54
4	700	12	7	686	31	12	687	30	17	660	55	4	705	9	10	690	31	50
5	694	26	5	681	27	11	671	36	16	686	40	8	683	44	12	680	37	52
6	614	31	9	610	35	13	604	47	15	630	27	5	600	13	8	610	37	50
7	578	37	7	603	38	11	606	26	13	608	28	7	599	25	9	600	33	47
8	627	40	10	629	30	10	611	39	12	633	34	4	654	26	11	630	38	47
9	630	19	8	613	32	15	612	30	15	647	12	6	651	84	8	621	30	52
10	669	22	8	652	15	13	638	47	14	601	38	5	657	23	13	647	36	53
11	580	6	9	603	39	14	596	31	14	579	11	3	576	33	10	590	33	50
12	685	26	9	689	52	15	653	29	12	658	28	7	685	15	8	674	39	51
13	745	71	4	792	55	7	821	42	9	820	100	4	835	94	7	807	78	31
14	794	42	4	766	88	7	843	19	6	740	100	6	824	17	2	789	81	25
15	788	17	4	792	36	7	737	61	11	785	53	3	798	23	3	770	54	28
16	577	32	9	584	30	12	576	60	13	573	39	6	577	13	11	577	40	51
17	631	67	9	624	54	8	660	110	10	580	22	5	634	23	6	632	74	38
18	669	28	3	760	13	4	676	38	5	660	140	2	699	42	4	697	89	18
19	586	66	11	571	35	8	588	91	9	524	14	4	539	20	10	566	62	42
20	640	130	6	592	89	13	590	130	8	550	19	3	556	30	4	600	120	34
21	733	78	4	754	74	11	777	99	11	678	27	3	770	100	5	755	89	34
22	750	120	5	756	79	6	755	95	9	734	84	3	764	45	8	755	86	31
X	606	22	6	598	21	4	610	25	6	657	55	8	612	17	9	620	39	33
Y	854	59	3	782	27	2	899		1							837	61	6

DNA content in table 5.1, for the DNA ratio in table 5.2, for the length in table 5.3 and for the centromeric index in table 5.4.

The automatic localization of the centromere position was impossible or did not meet the required criteria in about 29% of the cases. Some chromosomes gave difficulties in the localization of the centromere position, resulting in a higher standard deviation of the DNA ratio and centromeric index.

The DNA content of the Y chromosome of subject 1 is less than the DNA content of the Y chromosome of the other two male subjects. It was almost impossible to determine the centromere position of the Y chromosomes.

One of the two chromosomes 1 of the subject 3 had a significantly larger DNA content than the other. The same phenomenon is present in the metaphases of subject 4, but less obvious. This resulted in a larger DNA content of chromosome 1 of subject 3 and 4 in table 5.1.

The difference in length of the chromosomes 1 and 2 is larger than the difference in DNA content. This explains the difference in classification results of the A group (chromosomes 1-3) in the remainder of this section. Length and centromeric index gave better classification results than DNA content and DNA ratio for this A group.

The correlation between the features was calculated within each class of homologue pairs. The correlation between the DNA ratio and centromeric index is high (correlation coefficient larger than 0.8). This can be expected because these two features are based on the same centromere position. The correlation coefficient between DNA content and length is about 0.3. So some correlation is present between these two features.

Classification results of the test samples with the n-1 method are given in table 5.5. The number of homologue pairs is taken into account in this classification. The chromosomes are classified in the 24 individual classes and the error rate of the individual classification is listed, averaged for each of the seven groups of the Denver system and for the total metaphase. In table 5.5 the error rate is also given when the classes consist of the groups of the Denver system. In table 5.6 the confusion matrix is given of the classification based on DNA content and DNA ratio into 24 classes. When the centromere position is not determined, the chromosomes are classified based on DNA content or length only.

The error rate is defined as the percentage of differences between the classification by the program and by the cytologist. The error rate of classification into 24 classes is large. When the classification is based on DNA content and DNA ratio, this error rate is about 48%. The a priori probability of correct

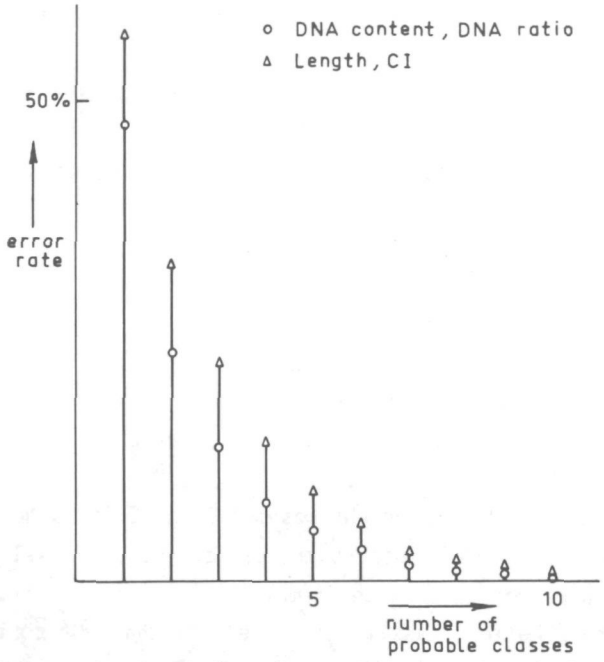


Figure 5.1 Error rate as a function of the number of probable classes. Classification with the n-1 method in 24 classes. The number of homologue pairs is not taken into account.

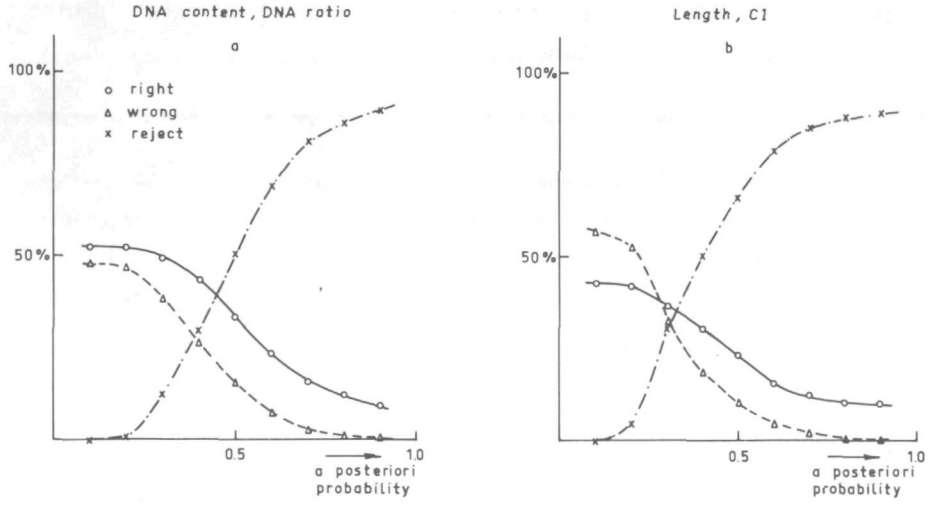


Figure 5.2 Reject rate and error rate of classification with the n-1 method in 24 classes. The number of homologue pairs is not taken into account.

Table 5.5 Error percentages in the classification (n-1 method, number of homologue pairs taken into account).

features	individual classification								classification in Denver system
	A	B	C	D	E	F	G	total	
DNA content, DNA ratio	8.0	54.0	60.0	57.8	34.2	54.8	56.4	48.2	7.7
length, CI	6.9	51.3	61.3	67.8	48.2	58.7	69.9	53.2	14.6
DNA content	33.7	67.3	72.4	60.0	50.8	55.6	56.4	59.3	13.3
length	22.9	67.3	74.3	72.8	71.0	68.3	69.9	65.6	21.2

Table 5.6 Confusion matrix of the classification based on DNA content and DNA ratio (n-1 method, number of homologue pairs taken into account).

real class	assigned class																							
	1	2	3	4	5	X	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Y
1	52	6																						
2	4	53					1																	
3			56	1	1		1																	
4			3	31	15		6				1													
5			1	30	21	1	1	1	1		1													
X					1	4	5	14	4	6	2	1												
6			2	1	10	5	34	9			1													
7						12	9	23	3	1	5	2	1											
8					1	6	2	10	8	9	11	6	3											
9					1	4	1	2	9	20	10	10	3		1									
10						2	1	6	8	10	17	3	17	1										
11						1		3	3	4	2	42	4											
12									2	4	14	6	36		2									
13													1	33	16	10								
14														16	22	16	6	1	1					
15													1	15	15	21	2	3	1					
16										1				1	2		50	8						
17														1	2	2	13	36	11	1	2			
18																1	1	11	41	1	7		1	
19																1			2	35	21		3	
20																			11	27	22		3	1
21																					1	31	17	9
22																				2	1	16	24	11
Y																				2	1	7	8	3

Table 5.7 Error percentages in the classification. Number of homologue pairs not taken into account. Learning set = test set.

features	individual classification								classification in Denver system
	A	B	C	D	E	F	G	total	
DNA content	28.0	70.8	72.4	67.8	45.6	58.7	48.9	58.8	14.5
DNA content DNA ratio	8.0	54.0	61.3	59.4	34.2	50.0	47.4	47.5	8.1
DNA content DNA ratio length	6.8	53.1	58.5	58.9	36.3	50.8	47.4	46.7	7.5
DNA content DNA ratio length, CI	6.3	54.0	60.0	55.0	42.0	49.2	60.9	48.6	7.7

Table 5.8 Error percentages in the classification based on DNA content and DNA ratio

method	individual classification								classification in Denver system
	A	B	C	D	E	F	G	total	
n-1 method pairs taken into account	8.0	54.0	60.0	57.8	34.2	54.8	56.4	48.2	7.7
n-1 method pairs not taken into account	8.0	54.0	61.1	60.0	35.8	51.6	51.9	48.3	8.1
learning set = testset pairs taken into account	8.0	48.7	58.3	54.4	31.6	51.6	48.9	45.4	7.4

classification into 24 classes is about 4%. The results of classification in the groups of the Denver system are much better although not satisfying. The classification based on DNA content and DNA ratio is better than the classification based on length and centromeric index. The error rate of classifying objects in the Denver system based on DNA content and DNA ratio is even a factor 1.9 lower. Only the individual classification in the A group (chromosomes 1-3) and the B group (chromosomes 4-5) of the Denver system is better for length and centromeric index than for DNA content and DNA ratio. Classification only based on the DNA content or length gives a considerable increase in the error rate.

Starting from the DNA content, we have added the features one by one. The error rates of the classification are given in table 5.7 (learning set = test set, homologue pairs not taken into account). Addition of the length to the DNA content and the DNA ratio results in a small decrease of the mean error rates. The mean error rate increases again with all four features.

In the experiments mentioned above, the chromosomes are assigned to the class with the highest a posteriori probability, when the number of homologue pairs is not taken into account. Instead of classifying a chromosome in one class a number of probable classes may be given. In this case the error rate is the percentage of chromosomes, for which the real class is not among the given probable classes. Figure 5.1 shows this error rate as a function of the number of probable classes. The chromosomes are classified with the n-1 method and the number of homologue pairs are not taken into account. A number of five classes is necessary to ensure an error rate of about 5%, when the chromosomes are classified based on DNA content and DNA ratio.

The error rate can be reduced by rejecting chromosomes with a low a posteriori probability. In figure 5.2 the classification results are given as a function of a threshold on the a posteriori probability. The chromosomes are classified with a n-1 method and the number of homologue pairs is not taken into account. The error rate of the classification based on DNA content and DNA ratio (figure 5.2a) can be reduced from about 48% to about 26% by a threshold of 0.4, introducing a reject rate of 30%. Reject rates up to 75% are necessary in order to obtain an error rate of less than 5%.

In table 5.8 the error rate is given when the test set is identical to the learning set and so no n-1 method is used. The improvement is limited, indicating that the number of learning samples is sufficient for the problem and the features used.

When the number of homologue pairs is not taken into account, the increase in the mean error rate is very small. This is also shown in table 5.8.

The improvement achieved by taking the number of homologue pairs into account, is investigated in a Monte Carlo experiment. In this experiment it is assumed, that the standard deviations are a factor f_3 times the measured standard deviations in the features. In figure 5.3 the mean error rates are given as a function of this factor f_3, when the chromosomes are classified based on DNA content and DNA ratio. The standard deviation in the metaphase error rate is between 3% and 7%. The absolute differences between the mean error rates are small. So for the classification based on DNA content and DNA ratio, there is little advantage in taking the number of homologue pairs into account. Figure 5.3 also shows that f_3 must be less than 1/8 to obtain an error rate less than 5%. Similar results may be obtained for length and centromeric index.

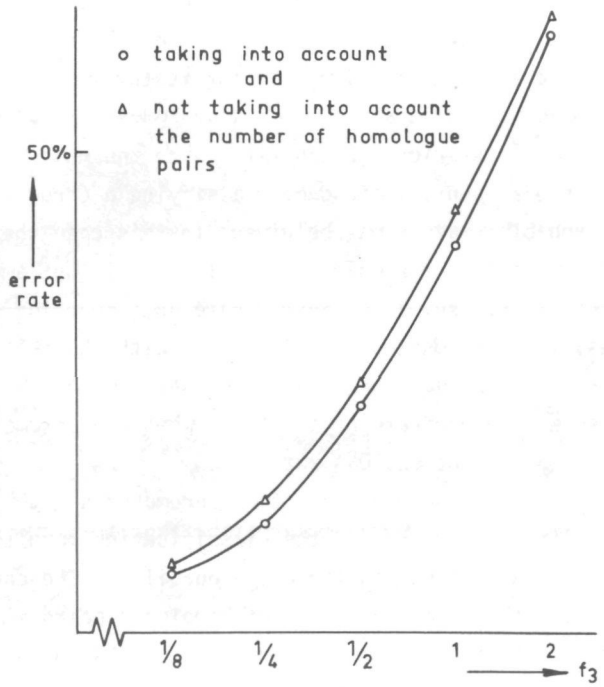


Figure 5.3 Error rate as a function of f_3. Monte Carlo experiment with DNA content and DNA ratio

The error rate at f_3=1 is slightly below the measured error rate in table 5.7. This can be explained by the fact, that in this Monte Carlo experiment no missing chromosomes or incomplete measured features were present.

5.6 CONCLUSIONS

This thesis deals with two topics concerning DNA based features: the mean and the standard deviation of the measured features (chapter 3 and 4) and the classification result with these features (this chapter). The minimum error rate depends on the distributions of the features, as these distributions and their parameters determine the overlap, which is present between the classes. The error rate of a particular classification method will be larger, because of the finite learning set and the assumptions made about the distributions. In our case the overlap between the classes - and thus the minimum error rate - is high. Therefore more emphasis has been given to the analysis of this minimum error rate than to the evaluation of classification methods.

In chapter 4 it is shown that the errors due to the measuring process are less than the variations present in homologue chromosomes. The latter variations are not only due to homologue differences, but also to errors in the classification of the cytologist, differences in stoichiometry and other absorbing materials in the specimen. These absorbing materials might be e.g. carbon particles present in the Schiff reagent used to stain the preparations or dirt adhering to the glassplates.

When only high quality metaphases are investigated, the standard deviations will be less than when also metaphases are taken into account in which e.g. other absorbing materials are present. We used metaphases that were not selected and that have the normal quality of the Histochemical Department in Leyden. We had, however, the drawback that the negatives had been stored too long before the metaphases were scanned. Due to this storage period small air bubbles had been formed in the embedding medium of the negatives, resulting in additional errors.

When we compare the homologue standard deviations of table 5.1 and table 5.2 to those given by Bosman (1976) (p. 64-67), we see that both measured standard deviations are of the same magnitude. This could be expected, because partly the same chromosomal material was used, be it with a different measuring method. When we compare these standard deviations with those given by Mendelsohn et al. (1973) and Mayall et al. (1975) we see that our standard deviations are slightly larger. Comparison between different authors is difficult, because the results depend on the material as well.

In our opinion automatic classification of chromosomes with DNA based features in the classes of autosomes and sex chromosomes may be an aid in an interactive karyotyping system. It gives an initial classification from which

the cytologist may start correcting the errors. From this point of view, classification results must be given for normal quality metaphases in which also outliers in the features (e.g. because of other absorbing particles) must be taken into account.

Classification results based on length and centromeric index are given by Neurath et al. (1975) and Ing et al. (1975). Error rates range from 10% to 35% for classification of the chromosomes in homologue pairs. The error rate we obtained is considerable higher (53%). In our case, however, no human intervention in the computation of the features occurred, only overlapping chromosomes were abandoned.

Classification results based on DNA content and DNA ratio are given by Mayall et al. (1975). Their error rate of classification in homologue pairs is 27%. The centromere position was, however, interactively adjusted by the operator in about 20% of the cases. The error rate in our experiment is about 48%. In our case no human interaction occurred and when the centromere position could not be located or did not meet the required criteria, the chromosomes were only classified on DNA content.

Compared to the classification with DNA based features, banding patterns seem to provide better classification parameters. Error rates of the classification with banding profiles reported by Ing (1975), Granlund (1973) and Møller (1973) range from 3% to 23%.

When DNA based features are used for classification in homologue pairs, it is clear that at this moment and given the culture and staining techniques used, human interaction in the computation of the features is necessary. Even the results obtained with human interaction as given in the literature are still insufficient. So classification with DNA based features at this moment is of limited use in clinical cytogenetics.

Classification is not the only or the main application of DNA based features. The determination of variant chromosomes as can be found in e.g. polymorphism, translocations or deletions, may be more important. Polymorphism is probably present in chromosome 1 of subjects 3 and 4. Other applications may be found in the field of prenatal diagnosis and chromosomal myelogeneous leukemia (see e.g. Mayall et al. (1975)).

SUMMARY

After a general introduction the measuring system is described in chapter 1. This system involves a photographic step to obtain sufficient spatial resolution .

In chapter 2 curvature measurement of quantized curves is evaluated. This chapter is the result of preliminary work on length measurements of chromosomes. Curvature measurement methods described by Gallus, Aalderink and Ledley are investigated. The methods seem to be quite different, but an analysis indicates that they only differ in the way in which the angular direction of the tangent to the curve is determined.

The errors in the measured curvature are theoretically and experimentally examined for an analytical curve resembling a small chromosome. The error consists of two parts, one quantization part and one part related to numerical differentiation. The error has an optimum as a function of the curve segment length. The probability density functions of Freeman codes have been calculated to evaluate the quantization error in the methods of Gallus and Aalderink.

The difference in the minimum curvature error between the methods of Gallus and of Aalderink is small. In Ledley's method this error is slightly less.

In chapter 3 a description is given of a program and subroutines to compute DNA based features. This program locates the chromosomes in scanned metaphases and computes the DNA content and the integrated density profile of the individual chromosomes. The chromosome regions for which the densities are summed are obtained by expansion of the original chromosome boundaries (up to 0.7 μm). The length, centromeric index and DNA ratio are obtained from the profile. The algorithms used are described and evaluated.

In chapter 4 the different sources of error in the computation of DNA based features are investigated. Two types, the quantization error and the distributional error, are examined in further detail.

The influence of the intensity quantization on the DNA content is compared for a linear and a logarithmic scale. The error in the DNA content for a linear scale has a minimum when the ratio between background and maximum intensity

equals 1/e, but the minimum is flat. When the contrast present in the photographic negative fits the density range of the logarithmic scale well, this scale gives the smallest error. When the contrast present in the negatives varies to a large extent, a linear scale is preferable.

The magnitude of the distributional error for the grid constants used, proved to be relatively small compared to the other sources of error. The photographic process does increase the coefficient of variation with respect to the variation introduced by the quantization noise and the noise in the photosensitive device. The homologue variations in the features are, however, larger than the photographic and other errors.

In chapter 5 the measured mean values and standard deviations of the features are determined for five subjects. At least six metaphases of each subject were scanned.

Classification results with these features are given both for a classification in homologue pairs, and for a classification in the Denver system. The error rate of classification based on DNA content and DNA ratio is 48% in the homologue pairs and 7.7% in the Denver system. The error rate of classification based on length and centromeric index is 53% in the homologue pairs and 14.6% in the Denver system. Compared to the literature these values are relatively high, but in our case there was no human interaction in the computation of the features.

Only a slight improvement is obtained when we take the number of homologue pairs into account for the classification.

We conclude that classification with these DNA based features may only be a first step in an interactive karyotyping system, given the culture and staining techniques used. A more important application of DNA based features may be the detection of variant chromosomes such as found in polymorphism, translocations or deletions.

SAMENVATTING

Na een algemene inleiding wordt in hoofdstuk 1 het meetsysteem beschreven. In dit systeem is een fotografische tussenstap aanwezig om voldoende spatiële resolutie te verkrijgen.

Hoofdstuk 2 behandelt een evaluatie van het meten van de kromming van gediscretiseerde contouren. Dit hoofdstuk vormt het resultaat van een inleidend onderzoek naar lengtemetingen aan chromosomen. In dit onderzoek zijn de methoden van Gallus, Aalderink en Ledley om de kromming te bepalen vergeleken. Ogenschijnlijk verschillen deze methoden sterk van elkaar, maar ze blijken na analyse slechts in de wijze waarop de hoekrichting van de raaklijn aan de contour wordt bepaald niet overeen te stemmen.

De fout in de gemeten kromming is theoretisch en experimenteel onderzocht voor een analytische kromme, die op een klein chromosoom lijkt. De fout kan gesplitst worden in een kwantiseringsfout en een fout die verband houdt met numerieke differentiatie. De fout heeft een minimum als functie van de lengte van het beschouwde contour segment. De kwantiseringsfout in de methode van Gallus/Aalderink is berekend met behulp van de verdelingsdichtheden van Freeman codes.

De minimale fout in de krommingsberekening volgens Ledley is iets kleiner dan die in de berekening volgens Gallus of Aalderink, waarvan de minimale fouten onderling zeer weinig verschillen.

In hoofdstuk 3 wordt een beschrijving gegeven van het programma met de bijbehorende subroutines om kenmerken te berekenen, die op DNA gebaseerd zijn. Dit programma localiseert de chromosomen in afgetaste metafasebeelden en berekent de DNA-inhoud en het geïntegreerde dichtheidsprofiel van de individuele chromosomen. De gebieden van de chromosomen, waarover de dichtheden worden gesommeerd, worden verkregen door uitbreiding van de oorspronkelijke begrenzingen van het chromosoom (tot 0.7 μm). De lengte, de centromeerindex en de DNA-ratio worden berekend uit het dichtheidsprofiel. De algorithmes, die zijn gebruikt, worden beschreven en geëvalueerd.

In hoofdstuk 4 worden de verschillende foutenbronnen in de berekening van

op DNA berustende kenmerken onderzocht. De kwantiseringsfout en de distributie fout worden nader bekeken.

De invloed van de intensiteitskwantisering op de DNA-inhoud is vergeleken voor een lineaire en logarithmische schaal. De fout in de DNA-inhoud voor een lineaire schaal heeft een minimum als de verhouding van de achtergrond en maximale intensiteit gelijk is aan 1/e. Dit minimum is vlak. Als het contrast, dat in de negatieven aanwezig is, goed overeenkomt met het gebied van de logarithmische schaal, geeft een logarithmische schaal de kleinste fout. Als het contrast tussen de negatieven aanzienlijk varieert, verdient een lineaire schaal de voorkeur.

De grootte van de distributiefout is betrekkelijk klein vergeleken met de andere foutenbronnen voor de gebruikte rastergroottes. Het fotografische proces vergroot de specifieke variatie, vergeleken met de specifieke variatie veroorzaakt door de kwantiseringsruis en de ruis in de fotodiode. De gevonden homologe variaties in de kenmerken zijn echter groter dan de variaties, die veroorzaakt worden door het fotografische proces en de andere optredende fouten.

In hoofdstuk 5 worden de gemeten gemiddelden en standaardafwijkingen in de kenmerken gegeven voor vijf proefpersonen. Van iedere proefpersoon zijn minimaal zes metafasen afgetast.

Klassificatieresultaten met deze kenmerken worden gegeven voor zowel klassificatie in de homologe paren als voor klassificatie in de groepen van het Denver systeem. Het foutenpercentage gebaseerd op DNA-inhoud en DNA-ratio bedraagt 48% voor klassificatie in homologe paren en 7,7% voor klassificatie in het Denver systeem. Het foutenpercentage gebaseerd op lengte en centromeerindex bedraagt 53% voor klassificatie in homologe paren en 14,6% voor klassificatie in het Denver systeem. Vergeleken met de literatuur zijn deze foutenpercentages betrekkelijk hoog, maar in ons geval is er geen menselijke tussenkomst in de berekening van de kenmerken toegepast.

Rekening houden met het aantal homologe paren dat in een metafase aanwezig is, geeft slechts een beperkte verbetering.

De konklusie is dat klassificatie gebaseerd op deze DNA-kenmerken alleen een eerste stap in een interaktief karyotyperend systeem kan vormen, bij de gebruikte kweek- en kleuringstechnieken. Het opsporen van variante chromosomen zoals die voorkomen bij polymorfisme, translocaties of deleties is wellicht een meer belangrijke toepassing van deze kenmerken.

APPENDICES

APPENDIX A

SOME ASPECTS OF HUMAN CYTOGENETICS

Chromosomes carry the genetic information in the DNA (deoxyribonucleic acid) molecule. DNA was isolated from salmon sperm cells for the first time in 1868 by Friedrich Miescher. In 1962 Crick, Watson and Wilkins (Wilkins (1964)) discovered the double helix structure of DNA. The bridges between the bases in the double helix consist of Cytosine (C) - Guanine (G) and Thymidine (T) - Adenine (A) combinations. The genetic information is present in the sequence of these bases in the DNA.

The meiosis preceeds the formation of the genetic cells. The 46 normal human chromosomes may be arranged into 22 homologue pairs of autosomes and two sex chromosomes. The chromosomes are distributed over the two resulting cells at the meiosis. If everything passes off correctly, each daughter cell receives a sex chromosome and one chromosome of each homologue pair, in total 23 chromosomes (haploid cell). During the impregnation again a cell of 46 chromosomes is formed. Of each homologue pair of chromosomes one chromosome is obtained from the father sperm cell and one chromosome was present in the egg cell of the mother.

Human chromosomes can be visualized during the metaphase of the mitotic division. The metaphase is that part of the normal cell cycle that immediately precedes the division of the cell into two daughter cells. In this phase the chromosomes condense into discrete objects with lengths from 2 to 20 μm.

Human cytogenetics became important between 1950 and 1960, when Tjio and Levan (1956) revealed the correct number of chromosomes in man and the first aberration: Down's syndrome with a trisomy of one chromosome was discovered by Lejeune (1959).

In the early sixties, human cytogenetics made rapid progress. Besides trisomy other aberrations were found like deletions ('cri du chat' syndrome) and translocations, where a part of one chromosome is transferred to another chromosome. The rapid progress seemed to come to an end, because no differentiating staining methods were available to identify and localize

individual regions in chromosomes. Nor were easily operational measuring devices available to study structural details rapidly and with high spatial resolution.

At that time it was only possible to determine numerical aberrations like chromosome - mosaicism, large structural aberrations and chromosome damages.

Although autoradiography gave some improvements in the identification of some special chromosome (e.g. X-chromosome), it did not yield essentially new developments in the chromosome analysis. The process of autoradiography, consisting of the labeling with radioactive bases, is difficult and laborious.

The identification of the chromosomes was based on the length measurements. Chromosomes were karyotyped according to the Denver conference (1960), where the human chromosomes were arranged into seven distinct groups. A further improvement of this classification was given at the London conference (1963) and the Chicago conference (1966).

Chromosome structures were also studied on a more fundamental basis. It was expected that cytochemical research would contribute much to the knowledge of cell differentiation and cell function, like the regulation of the gene function.

In 1968 a new impulse in human cytogenetics had come when Caspersson started a research with the idea that anti-leukemic chemotherapeutics could probably bind specifically to one of the bases in DNA. Applying quinacrine mustard to metaphase preparations, Caspersson et al. (1968) observed a banding pattern along chromosomes. It is assumed that the bands are AT-rich regions of the chromosome and that GC-rich regions are present between the bands.

Very soon after this discovery other reagents were found, which gave a banding pattern too, like the G, C, R and T banding techniques, described by Hsu (1973), Schnedl (1973) and Dutrillaux (1973). The importance of these banding patterns for cytodiagnosis was established at the Paris conference in 1971, where a new karyotyping of individual chromosomes was based on these patterns (Hamerton (1973)).

APPENDIX B

THE QUANTIZATION PROCESS (OBQ) AND THE VALUE OF THE FREEMAN CODE DIFFERENCE

In section 2.2 we made the restriction that a curve is never allowed to pass between two neighbouring grid nodes more than once (no narrow gap). The Object Boundary Quantization process with this restriction is identical to the contour-tracing algorithm used here. This contour-tracing algorithm scans the points of the quantized image. It tests the eight neighbours of the last found contour point in a clockwise direction, until the first one of two successive scanned neighbours is a background point and the next one is an object point. This implies that the original curve must have intersected the grid between these two nodes. The object point of the two is marked as contour point.

When we do not make the restriction in the OBQ, the original curve may pass twice between two nodes for a narrow gap. In this case a difference f'(i) in Freeman code values of -2 or -3 should be obtained, as is illustrated in figure B.1b and B.1c. When these situations occur the information about the gap is lost in the quantization process, because the quantized image of an object with a narrow gap is the same as the quantized image of an object without a gap (figure B.1a). The gap will not be detected by the contour-tracing algorithm, because in the algorithm it is assumed, that between two object points a curve will never pass.

In the Object Boundary Quantization process (with or without the restriction) it is impossible to obtain f'(i) = -4. For this narrow gap (illustrated in figure B.1d) both nodes (p and q) of the intersection with the grid will be selected as contour points because they are both object points. This results in different coming and going paths and for f'(i) = -4 these paths must be identical.

For f'(i) = -2, there is only one possible situation given in figure B.1e. As f'(i) = -4 and f'(i) = -3 will never occur, the possible values for f'(i) are

$$-2 \leq f'(i) \leq 4.$$

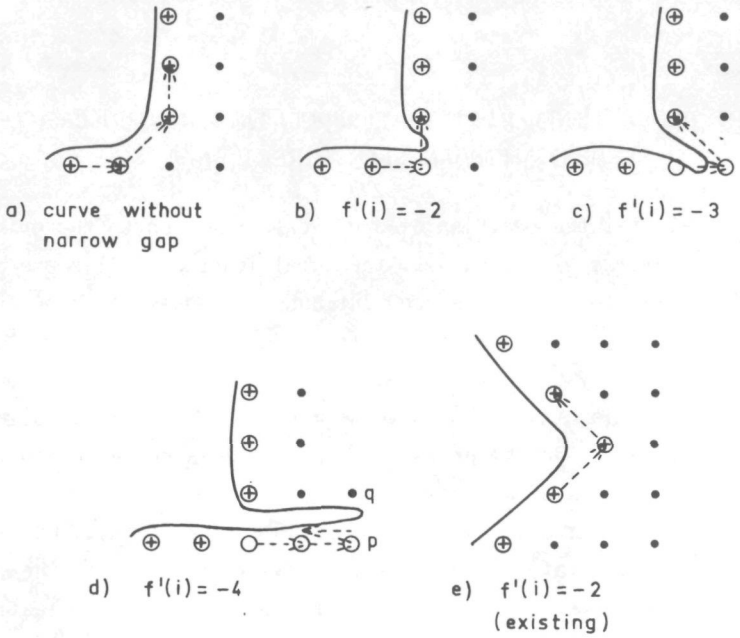


Figure B.1 (Not) occurring Freeman code differences for narrow gaps.

APPENDIX C

THE A PRIORI PROBABILITY $P(\varphi)$ THAT SEGMENTATION WILL CREATE A CURVE SEGMENT WITH ANGULAR DIRECTION φ

It is assumed that the position of a curve in relation to the grid is random. For the universe of all curves, the total length of all curve segments with an angular direction φ will be constant and independent of φ. A curve is divided by the grid intersections into a large number of curve segments. We assume that the curve may be approximated by a straight line in a column (or row) of the grid. The length of the curve segments contained in a column or row of the grid depends on φ. So the probability $p(\varphi)$ that a curve segment lies in a grid column or row depends on φ.

We will investigate the situation $0^{o} \leqslant \varphi \leqslant 45^{o}$. Other situations are similar and can be obtained by rotation and reflection. We will calculate the probability that the curve segment lies in a grid column for this situation illustrated in figure C.1.

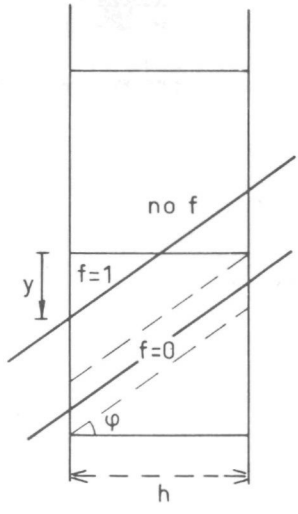


Figure C.1 Intersected line segment

The length of a line segment contained in a column is $h/\cos\varphi$. Such a line segment results in a Freeman code 0 or the combination of a Freeman code 1 and no Freeman code, until it intersects the next column. For the universe of all curves $p(\varphi)$ times this length must be constant. So $p(\varphi)$ is proportional to $\cos\varphi$. Integration of $p(\varphi)$ from 0^{o} to 45^{o} must give 1/8, so normalization of $p(\varphi)$ results in

$$p(\varphi) = \frac{\sqrt{2}}{8} \cos\varphi, \qquad (C.1)$$

$$0^{o} \leq \varphi \leq 45^{o}.$$

Each intersection between a column and a line segment results in a Freeman code. The length of the line segment contained in the column is $h/\cos\varphi$. So the expected length L of the real curve related to a Freeman code is

$$L = 8 \int_{0}^{\pi/4} p(\varphi) \frac{h}{\cos\varphi} \, d\varphi = \frac{\pi}{4} \sqrt{2}\, h. \qquad (C.2)$$

APPENDIX D

VARIANCE IN THE ESTIMATED CURVATURE WITH THE CORRELATION TAKEN INTO ACCOUNT (n=1, B=0)

The estimated curvatures by the method of Gallus/Aalderink and by Ledley's method are the same in the case n=1, B=0. The estimated curvature $\dot{\phi}(i)$ is given in this case as (equation (2.17b) and 2.19)

$$\dot{\phi}(i) = \frac{1}{A'} \left[\frac{\pi}{4} f(i) - \frac{\pi}{4} f(i-1)\right]. \qquad (D.1)$$

The variance in $\dot{\phi}(i)$ is given as

$$\mathrm{var}[\dot{\phi}(i)] = \left(\frac{\pi}{4A'}\right)^2 \left[2\ \mathrm{var}[f(i)] - 2\ \mathrm{cov}[f(i), f(i-1)\right]. \qquad (D.2)$$

In section 2.5 $\mathrm{var}(\Phi) = \left(\frac{\pi}{4}\right)^2 \mathrm{var}[f(i)]$ was given in equation (2.31) as

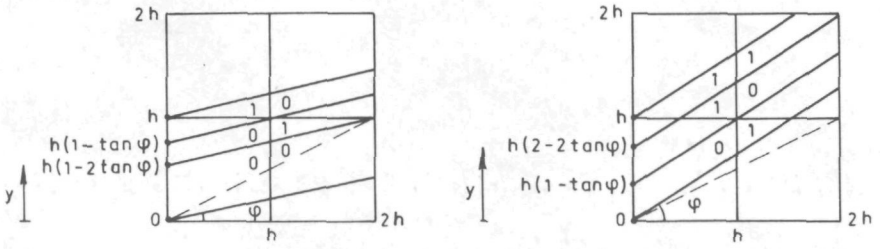
$$\mathrm{var}(\Phi) = \left(\frac{\pi}{4}\right)^2 \mathrm{var}[f(i)] = \left(\frac{\pi}{4}\right)^2 \sqrt{2}\,[1-\ln(1+\sqrt{2})]. \qquad (D.3)$$

Assuming that the curve segments may be approximated by a straight line, cov[f(i),f(i-1)] is calculated. We will restrict ourselves to $0^o \leq \varphi \leq 45^o$, as other cases can be obtained by rotation and reflection. When $0^o \leq \varphi \leq 45^o$ two essentially different situations exist, (a: $0 \leq \varphi \leq$ arctan 0.5, b: arctan $0.5 \leq \varphi \leq 45^o$), illustrated in figure D.1.

Assuming that y is uniformly distributed between 0 and h, the probability of occurrence of two Freeman code values is

a: $0 \leq \varphi \leq$ arctan 0.5

$$\begin{aligned} p[\ f(i-1)=0, f(i)=0 \mid \varphi\ ] &= 1-2\tan\varphi \\ p[\ f(i-1)=0, f(i)=1 \mid \varphi\ ] &= \tan\varphi \\ p[\ f(i-1)=1, f(i)=0 \mid \varphi\ ] &= \tan\varphi \\ p[\ f(i-1)=1, f(i)=1 \mid \varphi\ ] &= 0 \end{aligned} \qquad (D.4a)$$



a: $0 \leq \varphi \leq \arctan 0.5$ b: $\arctan 0.5 \leq \varphi \leq 45^{o}$

Figure D.1 Situations for two successive Freeman codes ($0 \leq \varphi \leq 45^{o}$)

b: $\arctan 0.5 \leq \varphi \leq 45^{o}$

$$\begin{aligned} &p[\ f(i-1)=0, f(i)=0 \mid \varphi\ ] = 0 \\ &p[\ f(i-1)=0, f(i)=1 \mid \varphi\ ] = 1-\tan\varphi \\ &p[\ f(i-1)=1, f(i)=0 \mid \varphi\ ] = 1-\tan\varphi \\ &p[\ f(i-1)=1, f(i)=1 \mid \varphi\ ] = 2\tan\varphi - 1 \end{aligned} \qquad \text{(D.4b)}$$

$$\mathrm{cov}[\ f(i), f(i-1) \mid \varphi\ ] = \&[\ f(i), f(i-1) \mid \varphi\ ] - \&[\ f(i) \mid \varphi\ ]^2. \qquad \text{(D.5)}$$

As only the case $f(i-1)=1, f(i)=1$ gives a contribution to $\&[\ f(i), f(i-1) \mid \varphi\ ]$, we obtain

$$\&[\ f(i), f(i-1) \mid \varphi\ ] = p[\ f(i-1)=1, f(i)=1 \mid \varphi\ ]. \qquad \text{(D.6)}$$

Combination of equations (D.6), (D.4a) and (D.4b) gives

$$\begin{aligned} &\&[\ f(i), f(i-1) \mid \varphi\ ] = 0 \qquad \text{if} \quad 0^{o} \leq \varphi \leq \arctan 0.5 \\ &\&[\ f(i), f(i-1) \mid \varphi\ ] = 2\tan\varphi - 1 \quad \text{if} \quad \arctan 0.5 \leq \varphi \leq 45^{o}. \end{aligned} \qquad \text{(D.7)}$$

$\&[\ f(i) \mid \varphi\ ]$ can be calculated from equation (2.25) of chapter 2. Combination with equation (D.5) and (D.7) gives

$$\begin{aligned} &\mathrm{cov}[\ f(i), f(i-1) \mid \varphi\ ] = -\tan^2\varphi \qquad \text{if} \quad 0^{o} \leq \varphi \leq \arctan 0.5 \\ &\mathrm{cov}[\ f(i), f(i-1) \mid \varphi\ ] = -(\tan\varphi - 1)^2 \quad \text{if} \quad \arctan 0.5 \leq \varphi \leq 45^{o}. \end{aligned} \qquad \text{(D.8)}$$

With $p(\varphi)$ given in equation (C.1) (Appendix C) $\mathrm{cov}[f(i), f(i-1)]$ is

$$\mathrm{cov}[\, f(i), f(i-1)\,] = \sqrt{2} \int_{\arctan 0.5}^{\pi/4} (2\tan\varphi - 1)\cos\varphi \, d\varphi \; - \sqrt{2} \int_{0}^{\pi/4} \tan^2\varphi \cos\varphi \, d\varphi \tag{D.9a}$$

or

$$\mathrm{cov}[\, f(i), f(i-1)\,] = -[2 + \sqrt{2}\,\ln(1+\sqrt{2}) - \sqrt{10}\,] = -8.42 \cdot 10^{-2}. \tag{D.9b}$$

Combination of equations (D.2), (D.9b) and (D.3) gives

$$\mathrm{var}[\dot{\phi}(i)] = \left(\frac{\pi}{4A'}\right)^2 2\,[\sqrt{2}+2-\sqrt{10}] = 0.3108/A'^2. \tag{D.10}$$

APPENDIX E

ARC LENGTH AND GRID ELEMENT AREA OF THE SECOND ORDER POLYNOMIAL

The arc length of a curve $\underline{c}$ is given by equation (2.3) of chapter 2

$$s(t) = \int_{t_o}^{t} \sqrt{\frac{d\underline{c}}{dt} \cdot \frac{d\underline{c}}{dt}}\, dt \tag{E.1}$$

with t_o an arbitrary starting point.

When the parameter t is substituted by x', we obtain for the second order polynomial given by equation (3.8)

$$\frac{d\underline{c}}{dx'} = (1,\ 2q_1x' + q_2) \tag{E.2}$$

and

$$s(x') = \int_{x'_o}^{x'} \sqrt{1 + (2q_1x' + q_2)^2}\, dx'. \tag{E.3}$$

This integral is

$$s(x') = \frac{1}{4q_1}\left[(2q_1x'+q_2)\sqrt{1+(2q_1x'+q_2)^2} + \ln\left[(2q_1x'+q_2)\ \sqrt{1+(2q_1x'+q_2)^2}\right]\right] \tag{E.4}$$

if $-q_2/2q_1$ is taken as arbitrary starting point x'_o.
When q_2 is substituted by $-2q_1x'_o$ in equation (E.4) s(x') is

$$s(x') = \frac{1}{4q_1}\left[2q_1(x'-x'_o)\sqrt{1+4q_1^2(x'-x'_o)^2} + \ln\left[\ 2q_1(x'-x'_o) + \sqrt{1+4q_1^2(x'-x'_o)^2}\right]\right]. \tag{E.5}$$

GRID ELEMENT AREA

The area a of a grid element at distance y" of the parabola is (Cf. figure E.1

$$a = \frac{\beta}{2}(r_1+r_2)(r_1-r_2) \tag{E.6}$$

in which $r_1 + r_2 = 2\,(R-y")$

$$r_1 - r_2 = h" \tag{E.7}$$

$$\beta = \frac{h"}{R}$$

so $a = h"^2\,(1 - \frac{y"}{R})$. (E.8)

The curvature κ of a parabola $y'_p = q_1x'^2 + q_2x' + q_3$ is

$$\kappa = \frac{1}{R} = \frac{2q_1}{[1+(2q_1x'+q_2)^2]^{3/2}}\,. \tag{E.9}$$

Combination of equation (E.8) and (E.9) gives

$$a = h"^2\left[1 - 2q_1y"/[1+(2q_1x'+q_2)^2]^{3/2}\right]. \tag{E.10}$$

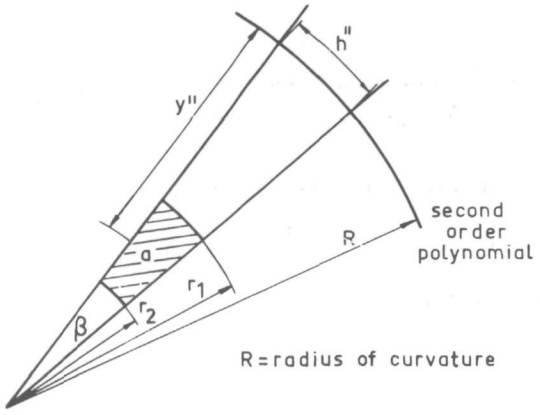


Figure E.1 Grid element of the second order polynomial

APPENDIX F

DISTRIBUTIONAL ERROR

The integral was given in section 4.5 by equation (4.29) and (4.26):

$$\widetilde{DNA}(k,l) = \frac{1}{k_a\gamma} \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} {}^{10}\log \frac{I_{kl}}{I_b} (1+bx+cy)\,dx\,dy \qquad \text{(F1)}$$

which may be split up into

$$\widetilde{DNA}(k,l) = D_1 + D_2 \frac{{}^{10}\log e}{k_a\gamma}$$

with

$$D_1 = \frac{1}{k_a\gamma} \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} {}^{10}\log \frac{I_{kl}}{I_b}\,dx\,dy = \frac{h^2}{k_a\gamma} {}^{10}\log \frac{I_{kl}}{I_b} \qquad \text{(F2)}$$

and

$$D_2 = \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} \ln(1+bx+cy)\,dx\,dy .$$

Integration of D_2 with respect to x gives:

$$D_2 = \frac{1}{b} \int_{-h/2}^{h/2} \Big[(1+bx+cy)\ln(1+bx+cy)-(1+bx+cy)\Big]_{-h/2}^{h/2} dy \qquad \text{(F3)}$$

which may be split up into

$$D_2 = D_3 + D_4 + D_5$$

with

$$D_3 = \frac{1}{b} \int_{-h/2}^{h/2} -bh \, dy = -h^2$$

$$D_4 = \frac{1}{b} \int_{-h/2}^{h/2} (1 + \frac{bh}{2} + cy) \ln(1 + \frac{bh}{2} + cy) dy$$

$$D_5 = -\frac{1}{b} \int_{-h/2}^{h/2} (1 - \frac{bh}{2} + cy) \ln(1 - \frac{bh}{2} + cy) dy .$$

Integration of D_4 with respect to y gives

$$D_4 = \frac{1}{2bc} \left[(1 + \frac{bh}{2} + \frac{ch}{2})^2 \ln (1 + \frac{bh}{2} + \frac{ch}{2}) + \right.$$

$$\left. - (1 + \frac{bh}{2} - \frac{ch}{2})^2 \ln (1 + \frac{bh}{2} - \frac{ch}{2}) \right] - \frac{1}{2bh} - \frac{h^2}{4} . \qquad (F.4)$$

A similar result may be obtained for D_5. Suppose $b' = \frac{bh}{2}$ and $c' = \frac{ch}{2}$, then D_2 may be rewritten as

$$D_2 = -\frac{3}{2} h^2 + \frac{1}{2bc} \left[(1+b'+c')^2 \ln (1+b'+c') - (1+b'-c')^2 \ln (1+b'-c') + \right.$$

$$\left. - (1-b'+c')^2 \ln (1-b'+c') + (1-b'-c')^2 \ln (1-b'-c') \right] . \qquad (F.5)$$

For $-1 < x < 1$ $\ln(1+x)$ may be expanded as

$$\ln(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} \frac{x}{i} . \qquad (F.6)$$

So assuming that $-1 < \pm b' \pm c' < +1$, D_2 may be expanded as

$$\begin{aligned} D_2 = -\frac{3}{2} h^2 + \frac{1}{2bc} \Bigg[& (1+2(b'+c')+(b'+c')^2)\left(\sum_{i=1}^{\infty} (-1)^{i+1} \frac{(b'+c')^i}{i} \right) + \\ & -(1+2(b'-c')+(b'-c')^2)\left(\sum_{i=1}^{\infty} (-1)^{i+1} \frac{(b'-c')^i}{i} \right) + \\ & -(1+2(-b'+c')+(-b'+c')^2)\left(\sum_{i=1}^{\infty} (-1)^{i+1} \frac{(-b'+c')^i}{i} \right) + \\ & +(1+2(-b'-c')+(-b'-c')^2)\left(\sum_{i=1}^{\infty} (-1)^{i+1} \frac{(-b'-c')^i}{i} \right) \Bigg] . \end{aligned} \qquad (F.7)$$

This expansion may be rewritten as:

$$D_2 \approx -\frac{3}{2}h^2 + \frac{1}{2bc}\sum_{i=1}^{\infty}\left[-\frac{1}{i}(b'+c')^{2i} - \frac{1}{i}(b'+c')^{2i+2} + \frac{4}{2i-1}(b'+c')^{2i} + \right.$$

$$\left. + \frac{1}{i}(b'-c')^{2i} + \frac{1}{i}(b'-c')^{2i+2} - \frac{4}{2i-1}(b'-c')^{2i}\right]. \qquad (F.8)$$

Equation (F.8) may be rewritten as

$$D_2 \approx -\frac{3}{2}h^2 + \frac{6b'c'}{bc} - \frac{1}{2bc}\sum_{i=2}^{\infty}\frac{1}{i(2i-1)(i-1)}\left[(b'+c')^{2i} - (b'-c')^{2i}\right]. \qquad (F.9)$$

Or combined with equation (F.2) to obtain $\widetilde{DNA}(k,l)$

$$\widetilde{DNA}(k,l) = \frac{h^2}{k_a\gamma}\,{}^{10}\log\frac{I_{kl}}{I_b} - \frac{{}^{10}\log e}{k_a\gamma}\sum_{i=2}^{\infty}\frac{h^{2i}}{2^{2i}i(i-1)(2i-1)}\cdot\frac{\left[(b+c)^{2i}-(b-c)^{2i}\right]}{bc}. \qquad (F.10)$$

LIST OF SYMBOLS

a	area of a grid element	m^2
A	constant	
A'	approximated length between the centres of the leading and trailing curve segment	
A	area of the measuring field	m^2
b	density gradient	m^{-1}
b'	constant	
B	half the number of Freeman vectors between the leading and trailing vector	
B_1	cutt-off parameter of the differentiating low-pass filter	
c	density gradient	m^{-1}
$\underline{c}$	curve	
c'	constant	
c_1/c_4	coefficients interpolation function	m/m^{-2}
C_a	average chromosome chromophore concentration	kg m^{-2}
C_{max}	maximum chromosome chromophore concentration	kg m^{-2}
d	optical density at preparation	
D	optical density at negative	
D_a	average chromosome density at negative	
D_b	background density at negative	
D_{bp}	background peak density at negative	
D_c	minimal chromosome density at negative	
D_L	dissection level density at negative	
D_1	constant	kg
D_2/D_5	constants	m^2

DNA	DNA content of a chromosome	kg
DNA_L	DNA content of the chromosome longer arm	kg
DNA(k,l)	DNA content contained in the measuring spot at position (k,l)	kg
$\widetilde{DNA}(k,l)$	DNA content contained in grid element (k,l) without distributional error	kg
ΔDNA	error in the DNA content of a chromosome	kg
ΔDNA_L	error in the DNA content of the chromosome longer arm	kg
$\in = (e_1, \ldots, e_N)$	N-tuple of the allocated classes	
E	total error	
E_b^2	quadratic error due to numerical differentiation	
E	intensity of illumination	lx
&	expectation	
f	Freeman code	
f'	Freeman code difference	
f_d	density units of the logarithmic scale	
f_1, f_2, f_3	fractions	
g	second order polynomial function	m
g_{kl}	density gradient at grid element (k,l)	m^{-1}
G_k	k^{th} group of chromosomes	
h	grid constant scan grid	m
h'	grid constant requantization grid	m
h"	grid constant at the second order polynomial	m
H	optical transfer function	
H	exposure	lx s
i	integer	
i	transmitted intensity (luminous flux) at preparation	lm
i_b	background intensity (luminous flux) at preparation	lm
i_{kl}	intensity (luminous flux) at position (k,l) of preparation	lm

i_{max}	intensity (luminous flux) at preparation corresponding to I_{max}	lm
i_o	incident intensity (luminous flux) at preparation	lm
I	transmitted intensity (luminous flux) at negative	lm
I_{kl}	intensity (luminous flux) at the centre of the grid element (k,l)	lm
I_{max}	maximum measured intensity (luminous flux) at negative	lm
I_o	incident intensity (luminous flux) at negative	lm
j	integer	
J	glare	lm
J_1	*first order Bessel function*	
k_a	specific absorptivity	$kg^{-1}\ m^2$
k_E	constant first microscopic system	m^{-2}
K	number of features	
l	integer	
L	likelihood function	
m_{pq}	p,q moment of the chromosome density distribution	m^{p+q}
M	mass of chromophore present in measuring field	kg
M_b	mass of chromophore present in measuring field in the background	kg
M(k,l)	mass of chromophore present in measuring field at position (k,l)	kg
n	number of Freeman code values of the leading (or trailing) vector	
n_b	number of background points	
n_c	total number of chromosome points	
n_m	number of metaphase scans	
n_{G_k}	number of chromosomes in group G_k	
N	number of quantization levels	
NA	numerical aperture	
N	number of objects in a set	

N_i	number of objects in class i	
o	diameter measuring spot	m
p	probability density function	
P	DNA based profile	
P_c	profile value at the centromere position	
q_1/q_3	coefficients of the second order polynomial	m^{-1}/m
Q	number of classes	
r, r_1, r_2	radii	m
R	radius of curvature	
R	DNA ratio	
ΔR	error in the DNA ratio	
s	arclength	m
s_L	Ledley's approximation of the arclength of a parabola	m
s^{-1}	inverse arclength	m
s^2	estimated variance	
$s^2_{G_k,j}$	estimated variance in group G_k of metaphase scan j	
t	allowable curve parameter	
$\underline{t}$	tangent to a curve	
t_o	starting value of t	
T_E	exposure time	s
$u = (\underline{u}_1, \ldots, \underline{u}_N)$	N-tuple of feature vectors	
$\underline{u}_i = (u_{1i}, \ldots, u_{Ki})$	feature vector of chromosome i	
u_i^j	feature value of chromosome i in metaphase scan j	
v	smoothed Freeman code difference	
w	weight coefficients	
w'	weight coefficient differences	
W	DFT of w	
W'	DFT of w'	

x, x, x', x"	coordinates	m
x'_o	x' coordinate of parabola top position	m
$\Delta x'_c$	error in the centromere position	m
y, y, y', y"	coordinates	m
y'_p	y' coordinates of the second order polynomial	m
z	coordinate	m
α	constant of photographic emulsion	
β	angle	
γ	gamma of photographic emulsion	
δ	Kronecker symbol	
$\epsilon(i)$	noise present in curvature at position (i)	rad m^{-1}
ζ_i	point on the curve segment bounded by the i^{th} Freeman vector	
η	normalization factor DNA content	kg
θ	rotation angle	rad
Θ	angle between leading and trailing vector	rad
κ	curvature	rad m^{-1}
κ_L	approximation of the curvature by Ledley's method	rad m^{-1}
κ_{max}	maximum curvature	rad m^{-1}
λ	wavelength	m
Λ	loss function	
$\underline{\mu}_j = (\mu_{1j}, \ldots, \mu_{kj})$	mean feature vector of class j	
ξ_1, ξ_3	points on the trailing curve segment	
ξ_2, ξ_4	points on the leading curve segment	
Ξ	conditional risk	
ρ	spatial frequency	
ρ_j	correlation coefficient of class j	
$\sigma(D)$	standard deviation of D	
$\sigma_{G/A}$	standard deviation of $\dot{\phi}_{G/A}$ without correlation	rad m^{-1}

$\sigma'_{G/A}$	standard deviation of $\dot{\phi}_{G/A}$ with correlation	rad m^{-1}
σ_L	standard deviation of $\dot{\phi}_L$ without correlation	rad m^{-1}
σ_1, σ_2	standard deviations of the curvature extrema positions	m
σ_{DNA}	standard deviation of the DNA content	kg
Σ_j	covariance matrix of the features of class j	
$\Upsilon = (\upsilon_1, \ldots, \upsilon_N)$	N-tuple of class indications	
φ	angular direction of a segmented line element	rad
ϕ	angular direction of the tangent to the curve	rad
ϕ_H	angular direction of the leading vector	rad
ϕ_T	angular direction of the trailing vector	rad
$\dot{\phi}_{G/A}$	approximation of the curvature by the method of Gallus/Aalderink	rad m^{-1}
$\dot{\phi}_L$	approximation of the curvature by Ledley's method	rad m^{-1}
$\dot{\phi}_M$	measured curvature	rad m^{-1}
$\dot{\phi}_{max}$	maximum curvature	rad m^{-1}
$\tilde{\dot{\phi}}_{G/A}$	$\dot{\phi}_{G/A}$ without quantization errors	rad m^{-1}
$\tilde{\dot{\phi}}_L$	$\dot{\phi}_L$ without quantization errors	rad m^{-1}
Φ	angular direction of a Freeman vector or a grid vector	rad
Ψ_i	normalized DNA content of chromosome i	
χ	number of misallocations	
ω	parameter analytical chromosome	
Ω_j	class j	

REFERENCES

AALDERINK, B.J. (1970). Automatische chromosoomanalyse. Inleidend onderzoek. (Automatic chromosome analysis, Introductory investigation). M.Sc. thesis, Delft University of Technology, 1970 (in Dutch).

BENNET, J.R., J.S. MAC DONALD (1975). On the measurement of curvature in a quantized environment. *IEEE Trans. on Computers* 24 (1975) 8, pp. 803-820.

BERG, H.J.N. van den (1974). Op het scherp van het beeld. (On the edge of the image). M.Sc. thesis, Delft University of Technology, 1974 (in Dutch).

BOSMAN, F.T., M. VAN DER PLOEG, A. SCHABERG, P. VAN DUYN (1975). Chromosome preparations of human blood lymphocytes. Evaluation of techniques. *Genetica* 45 (1975), pp. 425-433.

BOSMAN, F.T. (1976). DNA cyto-photometry of human metaphase chromosomes. Thesis Leiden, 1976.

BRONS, R. (1974). Linguistic methods for the description of a straight line on a grid. *Computer Graphics and Image Processing* 3 (1974), pp. 48-62.

CASPERSSON,T., S. FARBER, G.E. FOLEY, J. KUDYNOWSKI, E.J. MODEST, E. SIMONSSON, U. WAGH, L. ZECH (1968). Chemical differentiation along metaphase chromosomes. *Exp. Cell Res.* 49 (1968), pp. 219-222.

CHICAGO CONFERENCE (1966). Standardization in human cytogenetics. Birth defects: Original Article Series II 2 (1966).

DENVER CONFERENCE (1960). A proposed standard system of nomenclature of human mitotic chromosomes. *Am. J. Hum. Gen.* 12 (1960), p. 384.

DUDA, R.O., P.E. HART (1973). Pattern classification and scene analysis. New York, Wiley-Interscience, 1973.

DUTRILLAUX, B. (1973). Application to the normal karyotype of R-band and G-band techniques involving proteolytic digestion. In: *Chromosome identification*. (T. Caspersson and L. Zech, ed.). New York, Academic Press (1973), pp. 38-42.

DUYNDAM, W.A.L., P. VAN DUYN (1973). The dependence of the absorbance of the final chromophore formed in the Feulgen-Schiff reaction on the pH of the medium. *Histochemie* 35 (1973), pp. 373-375.

FITZGERALD, P.H. (1965). Differential contraction of large and small chromosomes in cultured leucocytes of man. *Cytogenetics* 4 (1965), p. 65.

FREEMAN, H. (1961[a]). On the encoding of arbitrary geometric configurations. *IRE Trans. on Electronic Computers* 10 (1961) 2, pp. 260-268.

FREEMAN, H. (1961[b]). Techniques for the digital computer analysis of chain encoded arbitrary plane curves. In: *Proceedings national electronic conference* 17, Chicago (1961).

FREEMAN, H. (1961[c]). A technique for the classification and recognition of geometric patterns. In: *Proceedings 3rd international congress on cybernetics, Namur* (1961), pp. 348-369.

FREEMAN. H. (1962). On the digital computer classification of geometric line patterns. In: *Proceedings national electronic conference* 18,Washington(1962).

FREEMAN, H. (1969). A review of relevant problems in the processing of line-drawing data. In: *Automatic interpretation and classification of images*. (A. Grasselli, ed.). London (1969), pp. 155-174.

FREEMAN, H. (1970). Boundary encoding and processing. In: *Picture processing and psychopictorics* (S. Lipkin and A. Rosenfeld, ed.) New York, Academic Press (1970), pp. 241-266.

FU, K.S. (1976). Digital pattern recognition. Berlin, Springer Verlag, 1976.

FUKUNAGA, K. (1972). Introduction to statistical pattern recognition. New York, Academic Press, 1972.

GAILLARD, J.L.J. (1970). Het cytofotometrisch bepaalde DNA-gehalte als identificatieparameter voor metaphase-chromosomen bij de mens. Thesis Leiden, 1970.

GALLUS, G., P.W. NEURATH (1970). Improved computer chromosome analysis incorporating preprocessing and boundary analysis. *Phys. Med. Biol.* 15 (1970) 3, pp. 435-445.

GOLDSTEIN, D.J. (1970). Aspects of scanning microdensitometry. *J. of Microscopy* 92 (1970), pp. 1-16.

GRANLUND, G.H. (1973). Use of distribution functions to describe chromosome profiles. In: *Chromosome identification*. (T. Caspersson and L. Zech, ed.). New York, Academic Press (1973), pp. 85-87.

GREEN, J.E. (1970). Computer methods for erytrocyte analysis. In: *IEEE Conference record of the symposium on feature extraction and selection in pattern recognition*. Argonne, Illinois, 1970.

GROEN, F.C.A. (1971). Onderzoek naar het herkennen van menselijke chromosomen. (Research into the recognition of human chromosomes In: *Proceedings IFAC symposium automatic control and computers in the medical field, Brussels, 1971* (in Dutch).

GROEN, F.C.A., P.W. VERBEEK, G.A. VAN ZEE, A. OOSTERLINCK (1976). Some aspects concerning the computation of chromosome banding profiles. *3rd international joint conference on pattern recognition, Coronado, 1976.*

HAMERTON, J.L. (1973). Chromosome band nomenclature. In: *Chromosome identification*. (T. Caspersson and L. Zech, ed.). New York, Academic Press (1973), pp. 90-96.

HAMMING, R.W. (1962). Numerical methods for scientists and engineers. New York, McGraw-Hill, 1962.

HARTLEY, H.O. (1950). The maximum F-ratio as a short-cut test for heterogenity of variance. *Biometrika* 37 (1950), pp. 308-312.

HILDEBRAND, F.B. (1956). Introduction to numerical analysis. New York, McGraw-Hill, 1956.

HSU, T.C. (1973). Constitutive heterochromatin (C-band) technique. In: *Chromosome identification*. (T. Caspersson and L. Zech, ed.). New York, Academic Press (1973), pp. 32-33.

ING, P.S., R.S. LEDLEY, H.A. LUBS (1975). Chromosome analysis at the national biomedical research foundation. In: *Proceedings of the Asilomar workshop on automation of cytogenetics*. Pacific Grove, California (1975), pp. 27-38.

KREYSIG, E. (1959). Differential Geometry. Toronto, University of Toronto Press, 1959.

LEDLEY, R.S. (1964). High-speed automatic analysis of biomedical pictures. *Science* 146 (1964), pp. 216-223.

LEDLEY, R.S., L.S. ROTOLO, T.J. GOLAB, J.D. JACOBSEN, M.D. GINSBERG, J.B. WILSON (1965). Fidac: film input to digital automatic computer and associated syntax-directed pattern-recognition programming system. In: *Optical and electro-optical information processing*, Cambridge, Massachusetts, MIT Press (1965), pp. 591-613.

LEDLEY, R.S., F.H. RUDDLE (1966[a]). Chromosome analysis by computer. *Scientific American* 214 (1966), pp. 40-46.

LEDLEY, R.S., L.S. ROTOLO, M. BELSON, J.D. JACOBSEN, J.B. WILSON, T. GOLAB (1966). Pattern recognition studies in the biomedical sciences. In: *Proceedings spring joint computer conference*, Boston Massachusetts, *(1966)*, pp. 411-430.

LEDLEY, R.S. (1969). Automatic pattern recognition for clinical medicine. *Proc. IEEE* 57 (1969) 11, pp. 2017-2035.

LEDLEY, R.S., M. LEGATOR, J.B. WILSON (1968). Automatic determination of mitotic index. In: *Pictorial Pattern Recognition*, Washington D.C., Thompson (1968), pp. 99-103.

LEDLEY, R.S. (1972). Analysis of cells. *IEEE Trans. on Computers* 21 (1972) 7, pp. 740-753.

LEJEUNE, J. (1959). Le mongolisme. Premier example d'aberration autosomique humaine. *Ann. Génét.* 1 (1959), p. 41.

LONDON CONFERENCE (1963). The London conference on the normal human karyotype. *Cytogenetics* 2 (1963), p. 264.

MAYALL, B.H., M.L. Mendelsohn (1970), Errors in absorption cytophotometry: Some theoretical and practical considerations. In: *Introduction to quantitative cytochemistry*, vol. 2, New York, Academic Press, 1970.

MAYALL, B.H. (1971). Digital image processing at Lawrence Livermore laboratory. Part II: Biomedical applications. *Computer* 7 (1974) 5, pp. 81-87.

MAYALL, B.H., A.V. CARRANO, D.H. MOORE II, L.K. ASHWORTH, D.E. BENNETT, E. BOGANT, J.L. LITTLEPAGE, J.L. MINKLER, D.L. PILUSO, M.L. MENDELSOHN (1975). Cytophotometric analysis of human chromosomes. In: *Proceedings of the Asilomar workshop on automation of cytogenetics*. Pacific Grove, California (1975), pp. 135-144.

MEES, C.E. (1954). The theory of the photographic process. New York, MacMillan Company (1954), pp. 460-479.

MEISEL, W.S. (1972). Computer-oriented approaches to pattern recognition. New York, Academic Press, 1972.

MENDELSOHN, M.L., T.J. CONWAY, D.A. HUNGERFORD, W.A. KOLMAN, B.H. PERRY, J.N.S. PREWITT. (1966). Computer-oriented analysis of human chromosomes. Part I:Photometric estimation of DNA content. *Cytogenetics* 5 (1966), pp. 223-242.

MENDELSOHN, M.L., B.H. MAYALL, J.N.S. PREWITT (1968). Digital transformation and computer analysis of microscope images. In: *Advances in optical and electron-microscopy* New York, Academic Press, 1968.

MENDELSOHN, M.L.,D.A. HUNGERFORD, B.H. MAYALL, B.H. PERRY, T. CONWAY, J.N.S. PREWITT (1969). Computer-oriented analysis of human chromosomes. Part II: Integrated optical density as a single parameter for karyotype analysis. *Ann. N.Y. Acad. Sci.* 157 (1969),pp. 376-392.

MENDELSOHN, M.L., B.H. MAYALL (1972). Computer-oriented analysis of human chromosomes. Part III: Focus *Computers in Biology and Medicine* 2 (1972) 2, pp. 137-150.

MENDELSOHN, M.L., B.H. MAYALL, E. BOGART, D.H. MOORE II, B.H. PERRY (1973). DNA content and DNA-based centromeric index of the 24 human chromosomes. *Science* 179 (1973), pp. 1126-1129.

MØLLER, A.R., H. NILSSON (1973). Computerized statistical analysis of banding patterns. In: *Chromosome identification*. (T. Caspersson and L. Zech, ed.). New York, Academic Press (1973), pp. 56-60.

NEURATH, P.W., B.L. BABLOUZIAN, T.H. WARMS, R.S. SERBAGI (1966). Human chromosome analysis by computer - an optical pattern recognition problem. *Ann. N.Y. Acad. Sci.* 128 (1966), pp. 1013-1028.

NEURATH, P.W., K. ENSLEIN (1969). Human chromosome analysis as computed from arm length measurements. *Cytogenetics* 8 (1969), pp. 337-354.

NEURATH, P.W., G. GALLUS, J.B. HORTON, W.D. SELLES (1975). Automatic karyotyping: Progress, perspective and economics. In: *Proceedings of the Asilomar workshop on automation of cytogenetics*. Pacific Grove, California (1975), pp. 17-26.

OOSTERLINCK, A. (1975). Contribution at the meeting on computer assisted chromosome analysis. Delft, April 23-24, 1975.

OPPENHEIM, A.V., R.W. SCHAFER (1975). Digital signal processing. Englewood Cliffs, New Jersey, Prentice Hall, 1975.

PATRICK, E.A. (1972), Fundamentals of pattern recognition, Englewood Cliffs, New Jersey, Prentice Hall, 1972 .

PLOEG, M. VAN DER, P. VAN DUYN, J.S. PLOEM (1974). High-resolution scanning densitometry of photographic negatives of human metaphase chromosomes. I: Instrumentation, II: Feulgen DNA measurement. *Histochemistry* 42 (1974),pp. 9-29, pp. 31-46.

RUTOVITZ, D. (1967). Machines to classify chromosomes? In: *Human radiation cytogenetics* (H.J. Evans, ed.). Amsterdam, North Holland Publ. Company (1967), pp. 58-93.

SAVITZKY, A., M.J.E. GOLAY (1964) Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36 (1964), pp. 1627-1639.

SCHNEDL, W. (1973). Giemsa banding techniques. In: *Chromosome identification* (T. Caspersson and L. Zech, ed.). New York, Academic Press (1973), pp. 34-37.

SLOT, R.S., (1976). On the profit of taking into account the known number of objects per class in classification methods. Delft, Report Pattern Recognition Group, 1976.

SPAERMAN, Ch. (1904). The proof and measurement of association between two things. *Am. J. of Psychology* 15 (1904), pp. 72-101.

TJIO, J.H., A. LEVAN (1956). The chromosome number of man, *Hereditas* 42 (1956) 1.

TONKELAAR, E.M. DEN, P. VAN DUYN (1964). Photographic colorimetry as a quantitative cytochemical method, I: Principles and practice of the method. *Histochemie* 4 (1964), pp. 1-9.

WALKER, P.M.B. (1958). Ultraviolet microspectrophotometry. In: *General cytochemical methods*, vol. 1, (J.F. Danielle, ed.). New York, Academic Press (1958), p. 163.

WALL, R. (1974). The gray level histogram for threshold boundary determination in image processing with applications to the scene segmentation problem in human chromosome analysis. Ph.D. Dissertation UCLA,1974.

WILKINS, M.H.F. (1964). The molecular configuration of nucleic acids. *Nobel lectures. Psychology or medicine 1942-1966*. New York Elseviers Publishing Company (1964), pp. 754-782.

YOUNG, I.T., J.E. WALKER, J.E. BOWIE (1974). An analysis technique for biological shape I. *Information and Control* 25 (1974), pp. 357-370.

ZEE, G.A. van (1974). Automatische chromosoomanalyse aan Feulgen-gekleurde preparaten. (Automatic chromosome analysis of Feulgen-stained preparations). B.Sc. Delft University of Technology, 1974 (in Dutch).

ACKNOWLEDGEMENT

It is a pleasure for me to thank dr. M. van der Ploeg of the Department of Histochemistry and Cytochemistry of the State University of Leyden for the work he did on the preparation and scanning of the metaphases, for the many stimulating discussions and for his critisism of the manuscript. I thank also the co-workers of the Department of Histochemistry and Cytochemistry, who contributed to the experiments.

I am indebted to the co-workers of the Pattern Recognition Group at the Applied Physics Department of the Delft University of Technology, who helped me in the course of time. In particular dr. P.W. Verbeek, dr.ir. J.C. Joosten and the students R.E. Slot and G.A. van Zee.

I am obliged to mrs. B.J.M. Scholten - van der Burg, who typed the manuscript and mr. M.G. Langen, who drew the diagrams.

STELLINGEN

behorende bij het proefschrift van

F.C.A. Groen

Delft, 23 februari 1977

1 Het valt te betwijfelen of klassificatie resultaten voor elk van de chromosomen, gebaseerd op DNA-inhoud en DNA-ratio, steeds beter zijn dan die gebaseerd op lengte en centromeer-index. (Dit proefschrift)

2 Bij het aftasten van negatieven met een sterk variërend kontrast is een lineaire kwantiseringsschaal te prefereren boven een logarithmische kwantiseringsschaal. (Dit proefschrift)

3 Even en oneven freemancodes zijn niet even waarschijnlijk. (Dit proefschrift)

4 Foutenpercentages van klassificatiemethoden zeggen soms meer over de gebruikte verzameling leer- en testobjecten dan over de klassificatiemethode.

5 Het gebruik van één verzameling leer- en testobjecten om klassificatiemethoden te vergelijken, is af te raden.

6 Een hechter samengaan van optische en digitale beeldbewerking is wenselijk, hetgeen als een nieuwe vorm van hybried rekenen kan worden beschouwd.

7 Het verdient aanbeveling om in het belichtingsregelsysteem van automatische fotocamera's met lange belichtingstijden en z.g. computerflitsers met zeer korte belichtingstijden een gemiddelde correctie voor de reciprociteitsafwijking op te nemen.

8 De afbeelding van kinderen op Romeinse munten heeft vaak een symbolische betekenis.

9 De feitelijke organisatiestruktuur van veel woningbouwverenigingen is een basis voor onverantwoorde beslissingen.

10 Zolang de reformatorische kerkgenootschappen in Nederland het muzikale gedeelte van de eredienst niet erkennen als een wezenlijke bijdrage in de uiting en de vorming van het geestelijke leven van de plaatselijke gemeenten, zullen zij niet komen tot het stellen van eisen, waaraan de kerkmusicus moet voldoen, hetgeen een noodzakelijke voorwaarde is voor de regeling van zijn positie met betrekking tot en binnen die kerkgenootschappen.

11 Uit oogpunt van geestelijke volksgezondheid is het niet wenselijk dat telefoonaansluitingen in nieuwbouwwijken sterk vertraagd worden aangebracht ten opzichte van de oplevering van de woningen.

12 Lange vrachtwagencombinaties dienen op grond van de verkeersveiligheid binnen de bebouwde kom van dorpen te worden geweerd.

13 Het in beschouwing nemen van het al dan niet gepoetst zijn van de schoenen van een sollicitant is een blijk van gebrek aan andere criteria.

ISBN 90 6231 023 0