Statistics and Computing 4 (1994) 161–171

Computer-aided classification of human chromosomes: a review

ANDREW CAROTHERS and JIM PIPER

MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Received February 1994 and accepted February 1994

Computer-aided imaging systems are now widely used in cytogenetic laboratories to reduce the tedium and labour-intensiveness of traditional methods of chromosome analysis. Automatic chromosome classification is an essential component of such systems, and we review here the statistical techniques that have contributed towards it. Although completely error-free classification has not been, nor is ever likely to be, achieved, error rates have been reduced to levels that are acceptable for many routine purposes. Further reductions are likely to be achieved through advances in basic biology rather than in statistical methodology. Nevertheless, the subject remains of interest to those involved in statistical classification, because of its intrinsic challenges and because of the large body of existing results with which to compare new approaches. Also, the existence of very large databases of correctly-classified chromosomes provides a valuable resource for empirical investigations of the statistical properties of classifiers.

Keywords: Automation, chromosome, classification, karyotype, neural network, pattern recognition, review

1. Introduction

The importance of chromosome abnormalities in human disease has become widely recognized over the past 30 years. They are present in perhaps 20% of all conceptions, in 50% of early spontaneous abortions, in 10% of mentally retarded individuals and in many, perhaps most, cancers (Speed *et al.* 1976, Bond and Chandley 1983). They may occur either in all the cells of the body, that is as *constitutional* abnormalities, or only in certain specific cell lines. The former can either be inherited from a parent or arise as a new mutation during formation of the sperm or egg, whereas the latter result from a mutation occurring at any time from conception to old age, often as a consequence of exposure to environmental hazards. Some are abnormalities of *number*, with one or more entire chromosomes additional to or missing from the normal complement; others are of *structure*, with pieces of some chromosomes missing (deletions), turned back-to-front (inversions) or shifted on to another chromosome (translocations). Because of this complexity, chromosome analysis has evolved into a specialized discipline with widespread applications in both research and clinical practice, including prenatal screening, genetic counselling, oncology, radiation dosimetry and toxicology.

0960-3174/94 © 1994 Chapman & Hall

The need for automation arises from the fact that the 'traditional' (i.e. manual) methods of analysis are tedious and labour-intensive. Because chromosomes are frequently lost or obscured during preparation, several cells must usually be analysed until the observer is satisfied as to their chromosome constitution, or 'karyotype'. However, cells at the stage of division (metaphase) when the chromosomes are most easily analysed are relatively sparse, so that finding the required number may take time. If a hard copy is needed, the cell is first photographed, then the individual chromosomes are cut out and pasted together in pairs in a standard format known as a 'karyogram' (Fig. 1). An experienced operator takes about an hour to carry out a typical analysis requiring, say, 15 metaphase cells to be found and to have their chromosomes counted and 5 cells to be fully karyotyped, and a further 30 minutes for each karyogram produced. Within the last 10 years computers have become sufficiently cheap and fast to make full or partial automation of several stages in this process a practical prospect. Today, there are about a dozen companies marketing automated or semi-automated metaphase-finding and/or karyotyping systems, of which several hundred are in use throughout the world.

Our present purpose is to provide a comprehensive overview of the methods used by computer-aided systems to

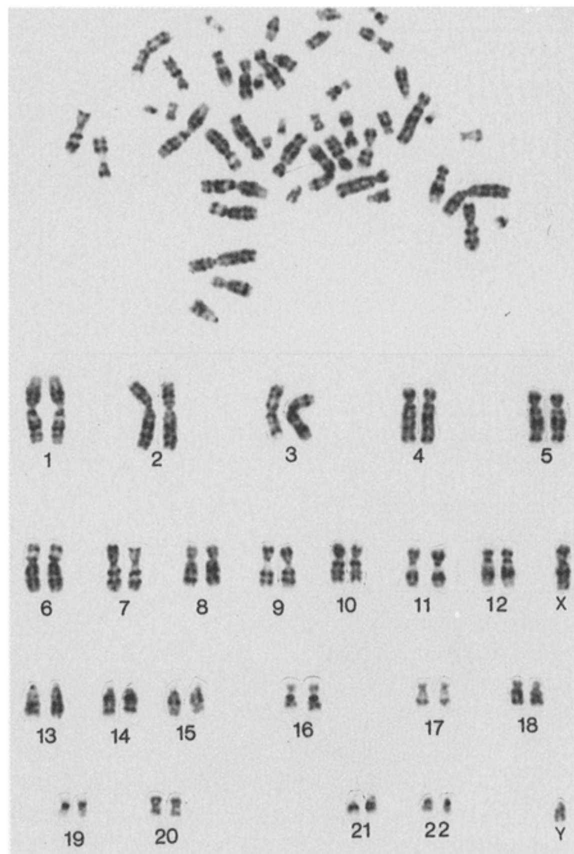

Fig. 1. *G-banded metaphase cell and karyogram*

classify human chromosomes fixed on slides. We shall therefore not be concerned with other aspects of automation, such as slide preparation and metaphase-finding, nor with the extensive literature on the techniques of flow cytometry (for further references on these topics, see for example Fantes and Green, 1989; Martin *et al.*, 1989; Vrolijk *et al.*, 1989; Korthof and Carothers, 1991). As will be seen, the complete armoury of statistical classification methods has been applied to this problem at various times, and it therefore provides an excellent exemplar for comparative studies of different approaches. Furthermore, the widespread routine use of automation in cytogenetic laboratories in recent years has led to the existence of exceptionally large databases of correctly classified chromosomes, making it possible, for example, to explore empirically the asymptotic behaviour of classifiers. This review is based largely on the bibliography provided by Lundsteen and Piper (1989), together with other and more recent references from our personal collections.

We start by reviewing methods of extracting features from processed images for input into statistical classifiers (Section 2). We then consider techniques for *normalizing* feature measurements to allow for the potentially large differences between cells and between specimens, for example in the state of contraction or density of staining (Section 3). In Section 4 we review what we refer to as 'context-free' methods of classification, in which individual chromosomes are classified independently of each other. We take into account the fact that the normal human karyotype consists of an ordered set of 22 pairs of autosomes and a pair of sex chromosomes leads to so-called 'context-dependent' classifiers (Section 5). Widening the context further, we then consider the problem of how best to infer the common underlying karyotype of a population of cells, given that the karyotype of any individual cell may be subject to error ('multiple-cell' karyotyping, Section 6). In Section 7 we review a number of more specialized applications of automated chromosome analysis, including aberration scoring for radiation dosimetry, and the particular problems and opportunities involved in the analysis of the very long overlapping chromosomes characteristic of prophase and prometaphase cells. Finally, in Section 8 we attempt to summarize the progress made to date, and the prospects for future progress in this field.

2. Image processing and feature extraction

In the early days, chromosomes were stained uniformly and as a result could be distinguished only on the basis of size and shape. Seven size–shape groups, denoted by the letters A–G inclusive, were defined at a conference in Denver, Colorado (Denver Conference, 1960). These so-called Denver groups are still useful for some purposes. However, nowadays most routine karyotyping is carried out on Giemsa-stained chromosomes. These appear as dark (absorption) images on a light background and have a characteristic pattern of light and dark bands unique to each type of chromosome, and are referred to as G-banding (Fig. 1). Automated image processing follows the usual steps of scanning, digitization, thresholding, segmentation and feature extraction. Since none of these, apart from the latter, is of present concern, the reader interested in technical details is referred to Rutovitz *et al.* (1978), Piper *et al.* (1980), Nickolls *et al.* (1981), van Vliet *et al.* (1990).

The features used for classification are related to size, shape and banding pattern. To represent the latter, the first step is to estimate the medial axis of the chromosome, then to generate a profile by taking the integrated density at right angles to the axis at each point (Fig. 2) (Hilditch, 1969; Groen *et al.*, 1976; Piper *et al.*, 1980; Piper and Granum, 1989). Another profile can be generated by taking the absolute differences of the density profile. A shape profile can be obtained in similar fashion by computing the *width*, or some transformation of it, at right angles to the medial axis (Piper and Granum, 1989). This is particularly useful for identifying the position of the centromere, which usually corresponds to a point of minimum width. Various methods have been proposed for extracting features from profiles.

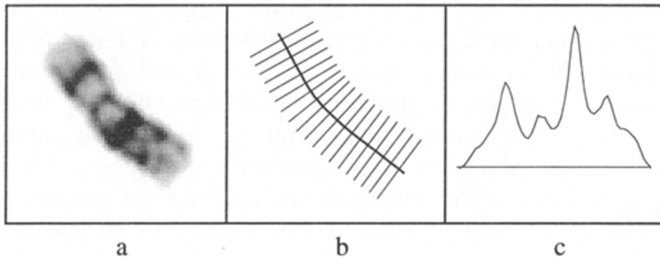

Fig. 2. *(a) Segmented chromosome; (b) medial axis and normal chords; (c) profile of density normal to the medial axis*

Caspersson *et al.* (1971) produced a *global* description of the density profile by means of a Fourier decomposition, of which the first eight or so harmonics were found to be useful for discrimination. However, this approach does not permit identification of *local* features such as missing, additional or displaced bands. It also encounters difficulties in handling differential contraction of chromosomes, and the resulting effects on the fundamental frequency and all higher harmonics. Granlund represented profiles as a mixture of several Gaussian distributions each characterized by height, width and position (Granlund, 1971; 1974; 1976). The procedure is limited by the fact that bands are not intrinsically Gaussian in nature, and may show extreme kurtosis or other distortions. Lundsteen and Granum (1975) reduced each profile to a sequence of numbers representing the position of each peak, its density and the density difference between it and the adjacent valley. By producing ideograms of 'artificial' chromosomes based only on these data, and presenting them for visual classification, they were able to show that these band transition (BT) sequences contained all the essential discriminatory information present in the original profile. Granum *et al.* (1981) derived a set of global features by taking a product between the density profile, $f(x)$, and each of a series of weighting functions, $w_i(x)$ $(i = 1, \ldots, K)$ (see Fig. 3), to give

$$\text{feature } i = \int_0^1 f(x)\, w_i(x)\, \mathrm{d}x \qquad (i = 1, \ldots, K)$$

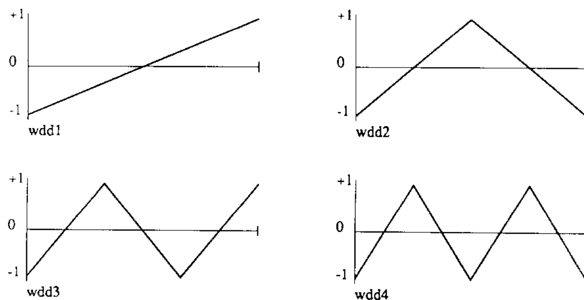

Fig. 3. *Weighting functions used to extract features of the density profile according to the method of Granum et al. (1981)*

where the $w_i(x)$ can be either symmetric (polarity-independent) or asymmetric (polarity-dependent). Groen *et al.* (1989) proposed a local band descriptor based on two-dimensional Laplace filtering of the image, followed by determination of the position of certain key bands (for instance the largest, the darkest, the first on the p-terminal, etc.). Granum and Thomason (1990) represented the profiles from chromosomes of the same class by means of a Markov network in which the probability of a particular transition from a band or feature to its neighbour was equated to the frequency with which it occurred in a learning set. An unknown profile could then be fitted to the Markov network of each chromosome class in turn by dynamic programming, and a measure of goodness-of-fit obtained by taking the appropriate product of transition probabilities. Errington and Graham (1993) input the banding profile directly to a multilayer perceptron neural network, and thereby avoided the need to extract intuitively-defined features from it.

Piper and co-workers proposed that features be grouped according to how much *a priori* information is needed to measure them (Piper *et al.*, 1980; Piper and Granum, 1989). Level 1 features can be measured directly from the chromosome image (e.g. area, density, convex hull perimeter); level 2 features require the medial axis (e.g. length, 'even-valued' functions of the density profile); level 3 features require both the medial axis and the correct polarity (e.g. 'odd-valued' functions of the density profile); finally, level 4 features require not only the medial axis and polarity, but also the position of the centromere (e.g. the ratio of short arm to long arm, the centromeric index, for length, area or density). Clearly, the lower the level of features on which a classifier is based the more sensitive it will be to errors at higher levels of processing.

3. Feature normalization

There are large variations in the appearance of cells, even from the same slide, because of differential contraction and intensity of staining and because some chromosomes may be unrecognizable, or missing from, or additional to, the normal complement. Typically, for example, chromosomes of the same class may vary in size by a factor of 2–3 between different metaphases (for comparison, chromosomes of different class *within* a metaphase differ in size by a ratio of 5 : 1). Since these differences are not directly relevant to chromosome classification, their effects must be removed from feature measurements by appropriate *normalization.* Given the correct identities of individual chromosomes, the problem is trivial since appropriate correction factors can be applied to produce the desired mean feature values. However, an automatic system faces a chicken-and-egg situation in which normalization cannot be correctly applied until chromosomes have been

correctly classified and vice versa. Hilditch and Rutovitz (1972) therefore proposed an iterative procedure in which a crude normalization is first applied, followed by a tentative classification, followed by a refined normalization and so forth until (normally) convergence. For size measurements, Piper and Granum (1989) adopted the simpler approach of normalizing by a multiplicative transformation, based on setting the median-sized chromosome in each cell to a value of 1000, claiming that the normalized measurements should be little affected by missing or additional chromosomes or by undetected composites of two or more chromosomes. They standardized other measurements by applying a linear transformation:

$$f'_{ij} = A_i + B_i f_{ij}$$

where f'_{ij}, f_{ij} denote respectively the transformed and untransformed values of feature i on the jth chromosome, and A_i and B_i denote cell-specific constants chosen to give a predetermined mean and standard deviation to the transformed features. The latter approach may be regarded as unprincipled, in contrast to size normalization which depends on prior knowledge that chromosomes within the same metaphase have the same degree of contraction. We suspect that a better understanding of the causes of variation in feature measurements would lead to more sensible and powerful normalization strategies. Granum (1982) showed that normalization generally improved classifier performance. For further discussion see Ledley *et al.* (1972) and Moore (1975).

4. Context-free classification

If individual chromosomes are considered as independent objects, without regard to their context as components of a karyotype, then the problem of assigning them to classes becomes a familiar one of statistical discrimination and is covered by many standard texts (see for example Choi 1986 for an excellent overview). Suppose we denote the feature vector of the ith chromosome from the jth cell by $\boldsymbol{x}_{ij}$, then the posterior probability that it belongs to class C_k is, by Bayes' theorem,

$$\Pr(C_k \mid \boldsymbol{x}_{ij}) \propto \Pr(\boldsymbol{x}_{ij} \mid C_k) \Pr(C_k)$$

Assigning each chromosome to the class for which this probability is maximized is an optimal decision rule, in the sense of minimizing the probability of error. However, this assumes that the two quantities on the right-hand side are known, or can be estimated without error. Since this is in fact not the case, the rule is not necessarily optimal but is assumed to be nearly so. The prior probability $\Pr(C_k)$ is usually known from the context. Thus, all autosomes occur with equal frequency in the normal karyotype, but the expected frequencies of the sex chromosomes depend on what is known of the sex of the subject. For example in prenatal screening, where the sex is not known *a priori*, one might assume prior probabilities of 2/46, 1.5/46 and 0.5/46 for each autosome, and the X and Y chromosomes respectively. Another complication concerns how to deal with structurally abnormal or grossly mismeasured chromosomes. The obvious solution is to have a 'reject' class with *a posteriori* probability fixed at a value such that the fraction of normal chromosomes assigned to it is less than some small predetermined amount (Paton, 1969). Essentially, then, the problem of assigning chromosomes to classes reduces to that of estimating the likelihood $\Pr(\boldsymbol{x}_{ij} \mid C_k)$. Most often, this has been done by assuming a multivariate Gaussian distribution for $\boldsymbol{x}_{ij}$, that is that

$$\Pr(\boldsymbol{x}_{ij} \mid C_k) = (2\pi)^{-n/2} \mid V_k \mid^{-1/2} \exp\left[-\tfrac{1}{2}(\boldsymbol{x}_{ij} - \mu_k)^{\mathrm{T}} V_k^{-1} (\boldsymbol{x}_{ij} - \mu_k)\right]$$

where n is the number of features and μ_k, V_k, representing respectively the mean feature vector and covariance matrix for chromosomes of class C_k, must be estimated from a learning set of correctly-classified chromosomes (Lundsteen *et al.*, 1986). A difficulty with this approach concerns the large number of parameters involved. With 24 classes and n features the full model has $12n(n+3)$ parameters, and requires an appropriately large training set. Several workers have therefore considered ways of reducing the complexity, either by discarding features that are so highly correlated with others that they contribute little to effective discrimination, or by simplifying the covariance matrices in various ways (Granum, 1982; Kirby *et al.*, 1991; Theobald and Kirby 1994). Piper (1987) showed that, with moderate-sized training sets, it was possible to achieve great simplification without loss of accuracy by ignoring all off-diagonal terms in the covariance matrices, i.e. by effectively treating the features as independent within each class. This was a somewhat surprising result, since the absolute magnitudes of the feature correlations were generally quite large. He also developed a method of selecting features based on a combination of the discriminating power of each feature taken in isolation, together with its lack of correlation with the set of features already selected. The method was simple to compute and appeared to work well particularly, as might be predicted, with the zero-covariance classifier model. Subsequently, he showed that considerable improvements in accuracy could be achieved using the full covariance matrices, but only if the size of the training set was about 10 times the total number of parameters in the model (Piper, 1992). This implies, for example, that with 24 features a training set of about 75 000 chromosomes would be needed to realize the potential of a fully-parametrized model. He also presented some results on the estimation of error rates which can be summarized as follows. Let $\boldsymbol{E}_{\mathrm{R}}(N)$ represent the estimated error rate obtained by resubstituting the training set, of size N, into

the classifier that was trained on it. Let $\boldsymbol{E}_{\mathrm{T}}(N)$ represent the estimated error rate obtained by submitting an independent test set, also of size N, to the same classifier. Then it can be inferred from Piper's results that the quantity

$$\tfrac{1}{2}\boldsymbol{E}_{\mathrm{R}}(N) + \tfrac{1}{2}\boldsymbol{E}_{\mathrm{T}}(N)$$

is an approximately unbiased estimate of the asymptotic error rates $\boldsymbol{E}_{\mathrm{R}}(\infty)$ and $\boldsymbol{E}_{\mathrm{T}}(\infty)$ (which are of course equal). Although this is no more than an empirical observation, it has previously been noted in an altogether different context by Toussaint and Sharpe (1975), and we therefore conjecture that it may be true under very general conditions. If so it would provide an obvious and straightforward method for unbiased estimation of the asymptotic error rates from a finite sample of any size. However, a formal proof is lacking. This is an example of the usefulness of the very large data sets available from automated chromosome analysis in providing insights into the properties of multi-class, multi-feature classifiers.

Other approaches have also been investigated. Rutovitz *et al.* (1978), working with a feature vector comprising size and centromeric index only, used a kernel-density method to estimate the likelihood function for each class. They replaced each point in the feature space of the learning set by a multivariate Gaussian spread function. However, this is computationally feasible only for small feature vectors. A group at Leuven used a classifier based on fuzzy subset theory, the outcome of which was a numerical measure of the similarity of an unknown chromosome to a known class (Vanderheydt *et al.*, 1979). The Athena system adopted a non-parametric approach in which features were assumed to be distributed independently of each other and the marginal feature distributions were estimated from histograms derived from the learning set (van Vliet *et al.*, 1989), so that

$$\Pr(\boldsymbol{x}_{ij} \mid C_k) = \prod_{m=1}^{M} h_{km}(x_{ijm})$$

where $h_{km}(x)$ denotes the empirically derived relative frequency of value x of the mth feature for class C_k, and x_{ijm} denotes the mth component of $\boldsymbol{x}_{ij}$. Shepherd *et al.* (1988) found that a decision-tree method gave consistently higher misclassification rates than a simple linear classifier, although it was much faster and required less memory. Kirby and Theobald (1993) investigated the performance of various 'two-stage' procedures, in which a single feature was first used to eliminate some candidate classes, before applying all features for a final classification. They reported greatly reduced overall allocation times, with a negligible penalty in terms of increased misclassification rates. Errington and Graham (1993) used a multilayer perceptron neural network, taking as inputs the chromosome size and centromeric index, and a coarsely quantized representation of the banding profile. They found that the performance compared favourably with that of a standard parametric classifier. The advantage of a neural network is that it reduces the need for intuitively defined features. However, networks perform best if their architecture and various controlling parameters are 'customized' for particular applications, so that in practice they may be no easier to implement than any other type of classifier.

5. Context-dependent classification

It seems intuitively obvious that misclassification error rates could be reduced by taking into account the fact that the normal human karyotype consists of 22 pairs of autosomes and a pair of sex chromosomes. In particular, human karyotypers rely strongly on between-chromosome comparison, and this has been shown to reduce error rates by at least an order of magnitude in manual karyotyping (Lundsteen *et al.*, 1976). Also, unpublished experiments at this laboratory have shown that when chromosomes are initially *forced* into homologous pairs using knowledge of their correct classes, and the pairs are then classified automatically, the error rates are typically approximately halved. This contextual knowledge constitutes a constraint that penalizes, say, the allocation of three chromosomes to one class and one to another. A method of allowing for such a constraint would be to incorporate it formally into a parametric model. For example, the likelihood for assigning to a particular class three chromosomes, each having two features, could be obtained from a multivariate Gaussian model with dispersion matrix of the form:

$$\begin{matrix} v_1 & c_{11} & c_{11} & b_{12} & c_{12} & c_{12} \\ c_{11} & v_1 & c_{11} & c_{12} & b_{12} & c_{12} \\ c_{11} & c_{11} & v_1 & c_{12} & c_{12} & b_{12} \\ b_{12} & c_{12} & c_{12} & v_2 & c_{22} & c_{22} \\ c_{12} & b_{12} & c_{12} & c_{22} & v_2 & c_{22} \\ c_{12} & c_{12} & b_{12} & c_{22} & c_{22} & v_2 \end{matrix}$$

where v_1 denotes the variance of feature 1 (assumed equal for all homologues of this class), v_2 the variance of feature 2 (likewise), b_{12} the covariance between features 1 and 2 on the same chromosome, c_{11} the covariance between homologues for feature 1, c_{22} the covariance between homologues for feature 2 and c_{12} the covariance between feature 1 and feature 2 on different (homologous) chromosomes. Parameters of type c_{xx} are therefore additional to those required for the independent assignment model, and presumably the size of the required training set would have to be increased accordingly.

In principle, for a complete karyotype of 46 chromosomes, an optimal (maximum likelihood) solution could be found by computing the likelihood for each of the

Table 1. *Percentage error rates for three classifiers on each of three data sets*

	Copenhagen	Edinburgh	Philadelphia
Context-independent ML classification	6.5	18.3	22.8
Rearrangement classifier RC3	5.7	16.4	20.6
Transportation procedure	4.4	15.5	19.9

$46!/2^{23}$ possible allocations to 23 pairs (Slot, 1979). Clearly this is far too many to be practical, though in reality the number of alternative allocations of a particular chromosome is much less than 23. Also there may be many alternative solutions with likelihoods close to the global maximum, any of which may be equally useful. Such considerations have led several researchers to seek practical though sub-optimal solutions. Habbema (1976) proposed an exhaustive-search approach for those classes that were particularly prone to confusion, for example chromosomes 4 and 5, or X and 7. Others have implemented what are essentially relaxation methods (Rosenfeld, 1978), attempting to permute the class assignments iteratively and in parallel to arrive at a 'best' solution. For example, Rutovitz (1977) described a method based on a Bayesian classifier for producing a shift from a class with too many chromosomes to one with too few in a Denver classification of homogeneously stained chromosomes. The cost of moving a chromosome from a class G to another H was defined as the negative of the maximum probability of any chromosome assigned to G being a member of H. A shift was usually implemented as a *cascade*, a chromosome being moved from class C_1 to C_2, another from C_2 to $C_{3,\dots}$, another from C_{n-1} to C_n. The cost of a cascade was the maximum of the costs of each individual move. Minimum-cost cascades, with costs below some threshold of plausibility, were chosen to correct the original karyotype in an iterative procedure which chose and implemented the lowest cost cascade first. Other 'rearrangement' classifiers were described and compared by Piper (1986) who concluded that the best overall performance was given by a version (designated RC3) of Rutovitz's method which included a penalty for 'implausible' assignments and in which the cost of allocating chromosome i, with feature vector x_i, to class C_k was defined as

$$c_{ik} = -\log L(x_i \mid C_k)$$

where L denotes the likelihood. However, a practical method of finding a globally optimal solution to the constrained allocation problem was proposed by Tso and Graham (1983), who noted that it could be formulated as a special case of the well-known 'transportation' problem in linear programming. In its classical form, the problem is that of minimizing the total cost of sending items from a set of suppliers, each of whom has a certain number of items available, to a set of customers, each of whom demands a certain number, given the costs of sending an item from any supplier to any customer. In the present context, 'suppliers' correspond to chromosome classes, each of which in a normal cell has two 'labels' corresponding to 'items', and 'customers' correspond to individual chromosomes, each of which demands a single label in order to be allocated. With costs c_{ik} (as for RC3) it can be shown that the solution to the transportation problem corresponds to the global maximum likelihood. Tso and Graham also showed how the method could be extended to cover situations in which some chromosomes were additional to, or missing from, the normal complement. Subsequently, they and their co-workers devised a rapid and efficient algorithm for the special case of unit demands and compared the results with those of the RC3 classifier (Tso *et al.* 1991). They found that error rates using the transportation algorithm were indeed reduced, but by rather a small amount, as shown in Table 1.

6. Multiple-cell karyotyping

When, as is often the case, all the cells in a sample can be assumed to have an identical karyotype, then that karyotype can be determined with great confidence by combining information from several cells. The basic principle is that, although karyotyping errors are made in individual cells as a result of distorted, touching, overlapping or missing chromosomes, such errors should be random and independent from one cell to another. Hence, a reliable composite can be built up from a sample of imperfect or incomplete cells, rather as one might reconstruct a scene from a number of random, independent and incomplete glimpses of it. An advantage of this approach is that it makes it possible to take from each cell only chromosomes that are well-segmented and undistorted. Granlund and co-workers used this principle to design a system to compare the characteristics of the set of chromosomes taken from many different cells and assigned to a particular class with those of a reference set of the same class, and hence to determine whether abnormalities of either number or structure were likely to be present (Granlund, *et al.* 1976; Granlund, 1978). The theoretical properties of a multiple-cell karyotyping system were worked out by Carothers *et al.* (1983) who derived a necessary and sufficient condition for consistent estimation of the correct number, t, of

chromosomes in a particular class, C, expressed as:

$$\sum_j p_{tj}[\ln(p_{tj}) - \ln(p_{ij})] > 0 \qquad (\text{all } i \neq t)$$

where

$p_{ij} = \Pr(j$ chromosomes are assigned to $C \mid i$ are actually present in C and all other classes have the normal complement)

and showed that the condition was easily met in practice. By calculating how many cells would have to be examined in order to achieve target false-negative and false-positive rates for given levels of misclassification error in individual cells, they concluded that a fully-automatic system for detecting abnormalities of chromosome number or gross structural rearrangements was entirely feasible, even with the relatively low levels of accuracy currently achievable in individual cells. A practical system for multiple-cell karyotyping was implemented by Lundsteen and co-workers at the Rigshospital in Copenhagen (Lundsteen *et al.*, 1989). This required no special hardware but simply displayed the results of fully-automatic karyotyping of individual cells in the form of a spreadsheet, in which each column represented the chromosomes from a different cell, and each row those assigned to a particular class. Chromosomes that could not be reliably assigned were placed in a 'reject' class at the bottom of the column. The cytogeneticist could then inspect the display and rapidly decide whether any departures from the expected pattern were real or a result of processing error. An example is shown in Figure 1 of Lundsteen *et al.* (1989) where, although there is apparently a No. 1 chromosome missing from the third cell, it is easily seen that the other three cells have a normal complement of No. 1s and also that the 'missing' chromosome is in the reject class at the bottom, almost certainly as a result of being severely bent. It can then be confidently concluded that this cell line has a normal complement of No. 1 chromosomes.

7. Special applications

We here review a number of applications of chromosome analysis other than routine karyotyping in which the role of automation has been investigated. However, the total effort expended in these areas has been small compared with that devoted to karyotyping.

7.1. *Aberration scoring for genetic toxicology*

By 'aberration scoring' we mean the estimation of the frequency of particular chromosome abnormalities that occur randomly as the result of certain environmental influences. Examples are: chromosome rearrangements resulting from exposure to ionizing radiation (Lörch *et al.*, 1989; Bayley *et al.* 1991); chromosome breaks and gaps induced by chemical mutagens (Turner *et al.*, 1993); and sister chromatid exchanges (SCEs) (Zack *et al.*, 1977; Shafer *et al.* 1986; Garcia-Sagredo *et al.* 1994). Such aberrations occur randomly and uniformly throughout the genome and, in the first two cases instanced, are relatively rare. For example, the incidence of dicentric chromosomes (Fig. 4), which have two constrictions instead of the usual one and are characteristic of exposure to ionizing radiation, is approximately one per 1000 metaphases (or 1 per 50 000 chromosomes) in an unexposed individual. The frequency increases with acute dosage according to a linear-quadratic relationship. However, at the low doses at which no other biological effects are visible (say around 0.2 Gy) the frequency may be increased by a factor of only 2–3 over background levels. Reliable estimation of the dose therefore requires the analysis of many thousands of metaphases. The final outcome of such analyses is a frequency estimate, which is subject to various sources of uncertainty. Thus, in contrast to karyotyping, there is no absolute requirement for the machine analysis to be as near error-free as possible. The machine contribution to the uncertainty is a parameter that can be included in a cost-benefit analysis of a system. Typically, these systems function as pre-screeners, presenting potential aberrations to an operator for visual review and rejection of false positives. The problem is to identify a particular structural feature of the chromosome (centromeres, in the case of dicentric example; or breaks in the case of chemical toxicology) with sufficient reliability that the rare cases of chromosomes with an incorrect number of the characteristic feature can be identified.

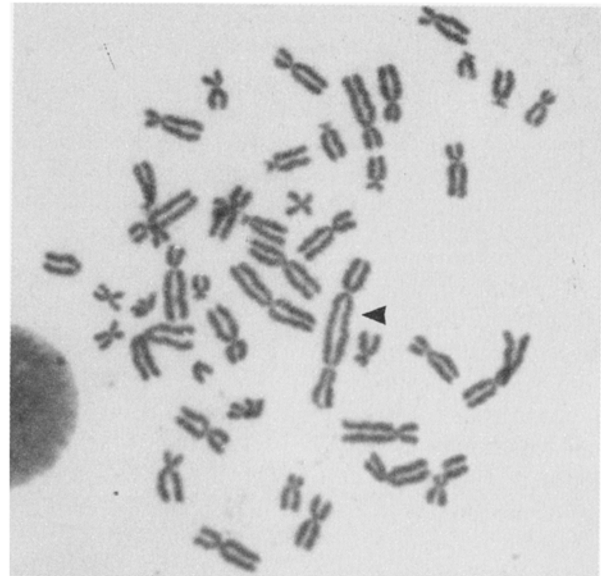

Fig. 4. *Metaphase cell with dicentric chromosome (arrowed)*

In the example of dicentric chromosomes, two approaches have been used. The first was similar to karyotyping, in that an attempt was made to extract numerical features that discriminated between dicentric and normal chromosomes and that were invariant to the biological variability between metaphases or between chromosomes (Lörch *et al.*, 1989). A classifier trained on a large set of metaphases was then used in the conventional fashion. The approach thus depended critically on an effective normalization procedure. The second method used a model of the expected number of centromeres per metaphase to train a centromere classifier *within each metaphase individually*, and that was specific to that cell, thus avoiding the problem of inter-cell normalization (Piper and Sprey 1992). Because there was no human supervision of the analysis, the training was based on data which was presumed, but not known for certain, to have a correct allocation. Approaches to classification in the presence of 'imperfect supervision' have been considered by Chhikara and McKeon (1984) and Krishnan (1988).

Generally similar problems arise in scoring SCEs (Fig. 5), except that the background frequency is many orders of magnitude higher (about 6 per metaphase cell) so that visual review of the machine analysis is not cost-effective. Zack *et al.* (1977) used a Parzen-estimated classifier based on two features to determine the probability P_{sce} that an observed apparent crossover event was a true SCE, and a similar classifier to estimate the probability P_c that the image region being analysed was a single chromosome (and not a chromosome cluster). Under the assumption that these decisions were independent, the joint probability of the event being a true SCE in a single chromosome was given by the product $P_{sce} P_c$. The SCE frequency for the cell was then estimated as the sum of the joint probabilities over all detected events, normalized by the estimated proportion of the metaphase that consisted of fully segmented chromosomes. More information on automation of aberration scoring can be found in a survey paper (Piper, 1991).

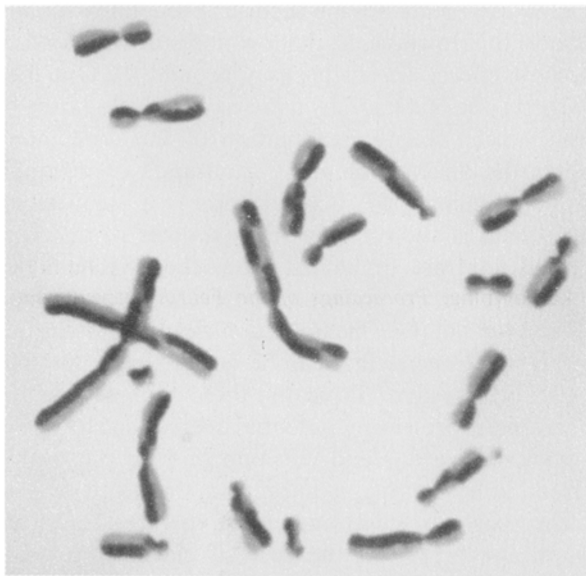

Fig. 5. *Metaphase chromosomes stained to show sister chromatid exchanges (SCEs)*

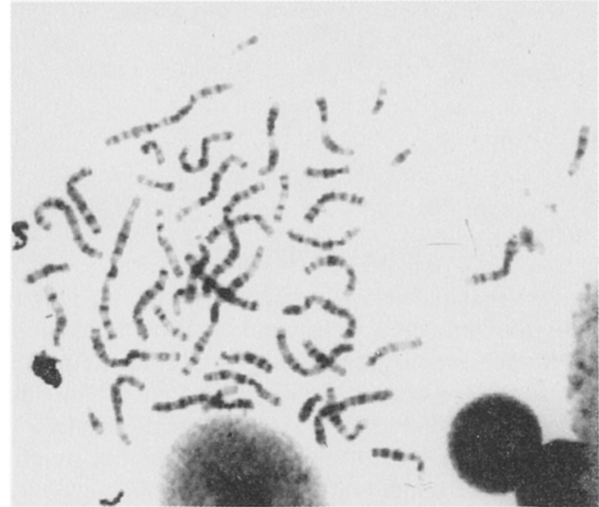

Fig. 6. *Chromosomes at the prometaphase stage of cell division*

7.2. *Prometaphase chromosomes*

When a small chromosomal lesion is suspected, cytogeneticists typically extend their analysis to cells at the prometaphase stage of cell division, where the chromosomes are much less contracted and many more bands are visible (Fig. 6). In fact, the trend of the last decade has been to use longer chromosomes, to the extent that the majority of the established databases are now atypical of best modern cytogenetic practice. High-accuracy classification of these materials has not been demonstrated. One problem is that prometaphase cells tend to contain large numbers of overlapping or severely bent chromosomes. Furthermore, the centromere is often not clearly visible as a distinct morphological feature.

7.3 *Other banding patterns and fluorescence*

Although G-banding is the most commonly-used pattern in the UK and US, other methods of differentially staining chromosomes are used for some purposes or in other countries. An example is R-banding which produces a pattern approximately complementary to G-banding. The different banding methods do not normally affect the overall morphology of the chromosomes, and similar methods of classification can be used for all of them. However, there is considerable variation between laboratories even using the same banding method, so that classifiers must be trained on material from the same laboratory (Piper and Granum, 1989). Increasingly important for a variety of

research and diagnostic purposes is the use of fluorescent staining techniques. These generally produce patterns resembling G- or R-banding, though often of lower contrast, and the chromosomes remain morphologically similar to those from conventionally stained preparations. Hence, the conventional classification techniques described above should work well, and what little data there is at present tends to confirm this. A genuine alternative is the work of Arndt-Jovin and Jovin (1990) who used two-channel total fluorescence to classify the chromosomes by their position in a two-dimensional chart, in a manner closely resembling the techniques used for flow cytogenetics (Fantes and Green, 1986).

8. Conclusions and future developments

The development of automation for chromosome analysis has been a success both scientifically and commercially. Furthermore, the pattern recognition and classification problems have presented unique and fascinating challenges for imaging scientists and statisticians. These have included the definition and extraction of discriminating features, the development of appropriate statistical models of inter- and intra-cell variation, the effective use of contextual information and constraints, the reduction of computational complexity (becoming less important with increases in computing power), the question of how best to present results for interactive interpretation, and the potential insights into the asymptotic behaviour of classifiers obtainable from the very large data sets available. However, some cautionary words are also appropriate. The first is that, in spite of the expenditure of much effort and ingenuity in developing sophisticated statistical models, relatively simple-minded approaches often seem to have been almost as effective. An example is the assumption of independence between features in a parametric model estimated from small training sets (Section 4). Also, the use of contextual information has to date produced only rather small reductions in error rates. We suspect that this is largely because most errors result from poorly segmented or severely distorted chromosomes, which may be hard to interpret even by the human eye. If this is so, then further progress is most likely to result from improvements at a more basic level of image processing (or of biology) than we have considered here. A second reservation concerns the rapid progress of genetical research which is constantly presenting new opportunities and, by implication, diminishing the significance of earlier developments. At present, for example, the use of fluorescent probes and 'painted' chromosomes is revolutionizing the ways in which chromosome abnormalities are detected and analysed. The challenges for automation are no less than before, but often very different. Fortunately, many lessons learnt from earlier experiences with automation can be adapted to the new technology.

References

Arndt-Jovin, D. J. and Jovin, T. M. (1990) Multivariate chromosome analysis and complete karyotyping using dual labeling and fluorescence digital imaging microscopy. *Cytometry*, **11**, 80–93.

Bayley, R., Carothers, A., Chen, X., Farrow, S., Gordon, J., Ji, L., Piper, J., Rutovitz, D., Stark, M. and Wald, N. (1991) Radiation dosimetry by automatic image analysis of dicentric chromosomes. *Mutation Research*, **253**, 223–235.

Bond, D. J. and Chandley, A. (1983) *Aneuploidy*, Oxford University Press.

Carothers, A. D., Rutovitz, D. and Granum, E. (1983) An efficient multiple-cell approach to automatic aneuploidy screening. *Analytical and Quantitative Cytology*, **5**, 194–200.

Caspersson, T., Lomakka, G. and Moller, A. (1971) Computerised chromosome identification by aid the quinacrine mustard fluorescence technique. *Heraditas*, **67**, 103–109.

Chhikara, R. S. and McKeon, J. (1984) Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association*, **79**, 899–906.

Choi, S. C. (ed.) (1986) *Statistical Methods of Discrimination and Classification: Advances in Theory and Applications*. Pergamon Press, New York.

Denver Conference (1960) A proposed standard system of nomenclature of human mitotic chromosomes. *Lancet*, **1**, 1063–1065.

Errington, P. A. and Graham, J. (1993) Application of artificial neural networks to chromosome classification. *Cytometry*, **14**, 627–639.

Fantes, J. A. and Green, D. K. (1989) The flow cytometry approach to automated chromosome analysis. In *Automation of Cytogenetics*, ed. C. Lundsteen and J. Piper, pp. 103–111, Springer-Verlag, Berlin.

Garcia-Sagredo, J. M., Piper, J., Rutovitz, D., Vaquero, J. J. and Vazquez, Y. (1994) Automatic scoring of sister chromatid exchanges by image analysis in a dose response experiment. *Environmental and Molecular Mutagenesis*, in press.

Granlund, G. H. (1971) The use of distribution functions to describe integrated density profiles of human chromosomes. *Journal of Theoretical Biology*, **40**, 573–589.

Granlund, G. H. (1974) Statistical analysis of chromosome characteristics. *Pattern Recognition*, **6**, 115–126.

Granlund, G. H. (1976) Identification of human chromosomes by using integrated density profiles. *IEEE Transactions on Biomedical Engineering*, **23**, 182–192.

Granlund, G. H. (1978) The structure of a system for multiple-cell karyotyping. *Proceedings of the Fourth International Joint Conference on Pattern Recognition, Kyoto, Japan*, 837–841.

Granlund, G. H., Zack, G. W., Young, I. T. and Eden, M. (1976) A technique for multiple-cell chromosome karyotyping. *Journal of Histochemistry and Cytochemistry*, **24**, 160–167.

Granum, E. (1982) Application of statistical and syntactical methods of analysis and classification to chromosome data. In *NATO ASI Series No. C.81: Pattern Recognition Theory and Applications*, ed. J. Kittler, pp. 373–398. Reidel, Dordrecht.

Granum, E. and Thomason, M. G. (1990) Automatically inferred Markov network models for classification of chromosomal band pattern structures. *Cytometry*, **11**, 26–39.

Granum, E., Gerdes, T. and Lundsteen, C. (1981) Simple weighted density distributions, WDDs, for discrimination between G-banded chromosomes. *Proceedings of the Fourth European Chromosome Analysis Workshop*, Edinburgh.

Groen, F. C. A., Verbeek, P. W., van Zee, G. A. and Oosterlinck, A. (1976) Some aspects concerning the computation of chromosome profiles. *Proceedings of the 3rd International Conference on Pattern Recognition, Coronado, California*, 547–550.

Groen, F. C. A., ten Kate, T. K., Smeulders, A. W. M. and Young, I. T. (1989) Human chromosome classification based on local band descriptors. *Pattern Recognition Letters*, **9**, 211–222.

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

Hilditch, C. J. (1969) Linear skeletons from square cupboards. In *Machine Intelligence* 4, ed. B. Meltzer and D. Michie, pp. 403–420. Edinburgh University Press.

Hilditch, C. J. and Rutovitz, D. (1972) Normalisation of chromosome measurements. *Computers in Biology and Medicine*, **2**, 167–179.

Kirby, S. P. J. and Theobald, C. M. (1993) Some two-stage procedures for the calculation of discriminant scores in the automated allocation of human chromosomes. *Pattern Recognition Letters*, **14**, 221–227.

Kirby, S. P. J., Theobald, C. M., Piper, J. and Carothers, A. D. (1991) Some methods of combining class information for the classification of human chromosomes. *Statistics in Medicine*, **10**, 141–149.

Korthof, G. and Carothers, A. D. (1991) Tests of performance of four semi-automatic metaphase-finding and karyotyping systems. *Clinical Genetics*, **40**, 441–451.

Krishnan, T. (1988) Efficiency of learning with imperfect supervision. *Pattern Recognition*, **21**, 183–188.

Ledley, R. S., Lubs, H. A. and Ruddle, F. H. (1972) Introduction to chromosome analysis. *Computers in Biology and Medicine*, **2**, 107–128.

Lörch, T., Wittler, C., Stephan, G. and Bille, J. (1989) An automated chromosome aberration scoring system. In *Automation of Cytogenetics*, eds, C. Lundsteen and J. Piper, pp. 19–30. Springer-Verlag, Berlin.

Lundsteen, C. and Granum, E. (1975) Description of chromosome banding patterns by band transition sequences. *Clinical Genetics*, **15**, 418–429.

Lundsteen, C. and Piper, J. (eds) (1989) *Automation of Cytogenetics*. Springer-Verlag, Berlin.

Lundsteen, C., Lind, A.-M. and Granum, E. (1976) Visual classification of banded human chromosomes I. Karyotyping compared with classification of isolated chromosomes. *Annals of Human Genetics*, **40**, 87–97.

Lundsteen, C., Gerdes, T. and Maahr, J. (1986) Automatic classification of chromosomes as part of a routine system for clinical analysis. *Cytometry*, **7**, 1–7.

Lundsteen, C., Gerdes, T. and Maahr, J. (1989) Cytogenetic analysis by automatic multiple cell karyotyping. In *Automation of Cytogenetics*, eds, C. Lundsteen and J. Piper, pp. 263–274. Springer-Verlag, Berlin.

Martin, A. O., Shaunnessy, M., Sabrin, H., Maremont, S., Dyer, A., Cimino, M. C., Rissman, A., McKinney, R. D., Cohen, M. M., Jenkins, E. C., Kowal, D. and Simpson, J. L. (1989) Evaluation and development of a system for automated preparation of blood specimens for cytogenetic analysis. In *Automation of Cytogenetics*, eds, C. Lundsteen and J. Piper, pp. 149–173. Springer-Verlag, Berlin.

Moore, D. H. (1975) Normalisation of chromosome measurements: a new method. *Computers in Biology and Medicine*, **5**, 21–28.

Nickolls, P., Piper, J., Rutovitz, D., Chisholm, A., Johnstone, I. and Robertson, M. (1981) Pre-processing of images in an automated chromosome analysis system. *Pattern Recognition*, **14**, 219–229.

Paton, K. (1969) Automatic chromosome identification by the maximum-likelihood method. *Annals of Human Genetics*, **33**, 177–184.

Piper, J. (1986) Classification of chromosomes constrained by expected class size. *Pattern Recognition Letters*, **4**, 391–395.

Piper, J. (1987) The effect of zero feature correlation assumption on maximum likelihood based classification of chromosomes. *Signal Processing*, **12**, 49–57.

Piper, J. (1991) Automated cytogenetics in the study of mutagenesis and cancer. In *Advances in Mutagenesis Research 2*, ed. G. Obe, pp. 127–153. Springer-Verlag, Berlin.

Piper, J. (1992) Variability and bias in experimentally measured classifier error rates. *Pattern Recognition Letters*, **13**, 685–692.

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

Piper, J. and Sprey, J. (1992) Adaptive classifiers for dicentric chromosomes. *Journal of Radiation Research*, **33**: Supplement, 159–170.

Piper, J., Granum, E., Rutovitz, D. and Ruttledge, H. (1980) Automation of chromosome analysis. *Signal Processing*, **2**, 203–221.

Rosenfeld, A. (1978) Relaxation methods in image processing and analysis. *Proceedings of the Fourth International Joint Conference on Pattern Recognition, Kyoto, Japan*, pp. 181–185.

Rutovitz, D. (1977) Chromosome classification and segmentation as exercises in knowing what to expect. In *Machine Intelligence 8*, eds, E. W. Elcock and D. Michie, pp. 455–472. Harwood, London.

Rutovitz, D., Green, D. K., Farrow, A. S. J. and Mason, D. C. (1978) Computer-assisted measurement in the cytogenetic laboratory. In *Pattern Recognition*, ed. B. G. Batchelor, pp. 303–329. Plenum, London.

Shafer, D. A., Mandelberg, K. I. and Falek, A. (1986) Computer automation of metaphase finding, sister chromatid exchange, and chromosome damage analysis. In *Chemical Mutagens: Principles and Methods for their Detection*, **10**, ed. F. J. de Serres, pp. 357–380. Plenum Press, New York.

Shepherd, B., Piper, J. and Rutovitz, R. (1988) Comparison of ACLS and classical linear methods in a biological application. In *Machine Intelligence 11*, ed. J. Richards, pp. 423–434. Oxford University Press.

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

Speed, R. M., Johnston, A. W. and Evans, H. J. (1976) Chromosome survey of total population of mentally subnormal in North-East of Scotland. *Journal of Medical Genetics*, **13**, 295–306.

Theobald, C. M. and Kirby, S. P. J. (1994) Discrimination using covariance selection models for the automated allocation of human chromosomes. Submitted.

Toussaint, G. T. and Sharpe, P. M. (1975) An efficient method for estimating the probability of misclassification applied to a problem in medical diagnosis. *Computers in Biology and Medicine*, **4**, 269–278.

Tso, M. and Graham, J. (1983) The transportation problem as an aid to chromosome classification. *Pattern Recognition Letters*, **1**, 489–496.

Tso, M., Kleinschmidt, P., Mitterreiter, I. and Graham, J. (1991) An efficient transportation algorithm for automatic chromosome karyotyping. *Pattern Recognition Letters*, **12**, 117–126.

Turner, M., Austin, J., Allinson, M. and Thompson, P. (1993) Matching an elastic model of chromosomal shape to features on a self-organising map. *Proceedings of the British Machine Vision Conference, Guildford, UK, 21–23 September 1993*, pp. 499–508.

Vanderheydt, L., Dom, F., Oosterlinck, A. and van den Berghe, H. (1979) An application of fuzzy subset theory to the classification of human chromosomes. In *Proceedings of the IEEE Conference on Pattern Recognition and Image Processing, Chicago*, pp. 466–472.

van Vliet, L. J., Young, I. T., ten Kate, T. K., Mayall, B., Groen, F. C. A. and Roos, R. (1989) Athena: a Macintosh-based interactive karyotyping system. In *Automation of Cytogenetics*, eds, C. Lundsteen and J. Piper, pp. 47–66. Springer-Verlag, Berlin.

van Vliet, L. J., Young, I. T. and Mayall, B. H. (1990) The Athena semi-automated karyotyping system. *Cytometry*, **11**, 51–58.

Vrolijk, J., Korthof, G., Vletter, G., van der Geest, C. R. G., Gerrese, G. W. and Pearson, P. L. (1989) An automated system for the culturing and harvesting of human chromosome specimens. In *Automation of Cytogenetics*, eds, C. Lundsteen and J. Piper, pp. 135–148. Springer-Verlag, Berlin.

Zack, G. W., Rogers, W. E. and Latt, S. A. (1977) Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry and Cytochemistry*, **25**, 741–753.