Trainable grey-level models for disentangling overlapping chromosomes

Graham C. Charters, Jim Graham*

University of Manchester, Wolfson Image Analysis Unit, Department of Medical Biophysics, Stopford Building, Oxford Road, Manchester M13 9PT, UK

Received 11 September 1997; received in revised form 29 September 1998

Abstract

We propose and evaluate a mechanism for resolving the segmentation of overlapping chromosomes using trainable models of the expected banding appearance. The models consist of templates of sub-chromosome length band profiles. Candidate chromosome segments are classified according to their responses to the entire set of templates, and matched on the basis of the classifications. Evaluation of the models using a set of annotated banding profiles yields correct classification rates of 90.8% for isolated chromosomes, and 55.4% for chromosome fragments; 70.6% of overlapping chromosome pairs, simulated using the profile data set, are correctly resolved. © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Chromosome analysis; Trainable models; Template matching; Overlapping chromosomes; Chromosome banding patterns; Classification; Segmentation

1. Introduction

Cytogenetics is the study of the genetic constitution of individuals at a scale which is revealed by light microscopy. At this level, the genetic material of a cell can be seen as a number of distinct bodies – the chromosomes. Analysis of the appearance of chromosomes can provide information on inherited or acquired syndromes, exposure to genotoxic agents or the presence of cancers. There are 46 chromosomes in normal human cells. Appropriately stained, they can be made to exhibit a sequence of bands which, together with the chromosome size and the position of a characteristic constriction called the centromere (see Fig. 2), can be used to assign chromosomes visually into one of 24 classes (labelled 1-22, X and Y). The last two classes are the sex chromosomes, of which females have two in class X and males one X and one Y. All other classes contain two identical (homo-logous) chromosomes in normal individuals. It is often helpful to display chromosomes arranged in a karyotype – a tabular array in which the chromosomes are aligned in pairs (see Fig. 1).

The appearance of chromosomes depends on the stage of the cell division cycle at which they are viewed. For much of the cell cycle (interphase), individual chromosomes cannot be distinguished. They only appear as distinct bodies towards the end of the cycle, at prophase, when they are long string-like objects, contracting and separating at metaphase, just before cell division takes place. Although not biologically significant, it is common to refer to an intermediate stage of contraction between...
prophase and metaphase as prometaphase. Fig. 1 shows examples of prophase, prometaphase and metaphase chromosomes stained by a commonly used method (G-banding) together with a metaphase karyotype. Operationally, these stages in the division cycle are defined by the number of bands visible in the cell. The more elongated chromosomes exhibit more bands than shorter ones. Metaphase cells have around 450 bands and prometaphases are defined to have 550 bands and above. A cell is not considered to be in prophase unless at least 850 bands are visible. The greater number of bands at the more elongated stages provides a higher resolution description of the chromosome structure, which is advantageous for analysis. The analysis is, however, much more difficult due to the greater complexity of the banding patterns and the fact that longer chromosomes touch and overlap each other much more frequently than shorter ones (see Fig. 1).

The automation of chromosome analysis was first proposed in the 1960s [1]. Many years of effort have resulted in the development of commercial cytogenetics systems for analysis of banded chromosome preparations [2]. A great deal of attention in cytogenetics has recently been focused on molecular techniques and the application of image analysis to interpretation of fluorescence microscope images, including the location of specific hybridisation sites at interphase [3,4]. However, analysis of banded chromosomes remains of great importance particularly at the prometaphase or prophase stages of contraction. Most studies in automation, on the other hand, have concentrated on metaphase chromosomes, avoiding the segmentation difficulties arising from touches and
overlaps in the prophase and prometaphase cells. Graham and Piper [5] provide a review of methods used in automated chromosome analysis. Segmentation and classification of chromosomes into groups are important stages in the analysis and are generally taken to be separate tasks. Segmentation is usually performed by one or other of a number of thresholding methods [6–9]. Classification schemes use a representation of the banding pattern, generally derived from the integrated density profiles of the segmented object, together with size and centromere position [10–13] (see Fig. 2).

The straightforward segment-measure-classify strategy is inadequate for the analysis of images of chromosomes in their less contracted phases as it provides no mechanism for resolution of overlaps. Some attempts have been made to deal with clusters of touching (but not overlapping) chromosomes [6–9] where combinations of geometric and densitometric evidence have been used to resolve segmentation ambiguities.

Clearly, automatic separation of overlapping chromosomes is important for the analysis of prophase and prometaphase images, but has received relatively little attention compared to other aspects of the chromosome analysis problem, such as classification. Ji [7,14,15] has proposed methods for automatically segmenting both touching and overlapping clusters. His approach to segmentation of overlaps was to decompose a thresholded overlapping clusters. His approach to segmentation of overlaps was to decompose a thresholded overlapping clusters into individual components using geometric evidence, i.e. by reasoning about shapes. Agam and Dinstein [16] have similarly applied reasoning about boundary curvature to separating touching or slightly overlapping chromosomes. In this paper we propose an alternative, additional source of information for disentangling overlaps, namely the banding pattern. We suggest a strategy for using the banding pattern, which uses the same chromosome evidence for reasoning about segmentation solutions and for classifying segmented chromosomes. The method consists of identifying consistent pairings of short sections of banding pattern to make believable chromosomes. This approach conforms to the approach of human experts, who often piece together chromosomes in overlaps by matching short banding sequences to a mental model of the chromosome classes.

2. Outline of the method

We summarise our method here using, as an illustration, the schematic pair of overlapping chromosomes in Fig. 3a. The banding pattern in the overlapping region is obscured, but four short sections of banding pattern are visible. The uncertainty in segmentation arises because each of these segments could be matched to any of the other three to generate complete chromosomes (Fig. 3c, d and e). If the classes of the partial chromosome segments can be identified by their local banding pattern, then the segments belonging to the same class can be matched, and the segmentation uncertainty resolved.

The classification of the segments is performed by matching the unobscured sections of banding pattern to a set of templates previously selected to match preferentially to particular sub-chromosome sequences of the banding pattern. These are called Partial Chromosome Models (PCMs). Fig. 3b shows PCMs which match preferentially to the end sections of the schematic chromosomes. The match is determined using a measure of fit described in Section 4.2, and used to classify the visible chromosome segments.

This approach takes inspiration from the work of Lockwood et al. [17–19], who observed that certain sections of the banding pattern are highly characteristic and used in visual classification of prophase chromosomes. They described a method of classifying prophase chromosomes on the basis of a set of short sub-chromosome-length banding segments. By conducting an exhaustive band-by-band search of digitised prophase ideograms (stereotyped banding patterns) they identified a set of 94 “unique band sequences” [19].

The PCMs we describe here are derived by training. Sub-chromosome-length segments of banding profiles are extracted from a training set of profiles in such a way as to give a large number of templates with a range of lengths (from half to three times the length of the shortest chromosomes), located at positions covering the banding profiles of all classes. For each of the templates, corresponding segments of banding profile are taken from the homologous chromosomes in all cells in the training set.

![Fig. 2. Schematic representation of features used in chromosome classification. The bands are arranged linearly along the chromosome. Integrating density normal to the medial axis gives a density profile which is characteristic of the chromosome. The overall length of the chromosome and the position of the centromere are also important features. The centromere divides the chromosome into two “arms”, conventionally labelled p and q.](image-url)
to provide a statistical description of the segment, which forms the PCM. Homologous chromosomes from different cells may have significantly different lengths, introducing a plastic deformation into the profile. This is accounted for by resampling each training profile to have the same number of samples, effectively length-normalising the segments (see Section 4.1 and Fig. 4). A collection of PCMs is created, derived from different segments of different classes, and each is matched at every position on the unobscured sections of banding pattern in an overlap to be resolved (the “target” segments). A measure of fit is calculated between each PCM and each target segment (the response of the target to the PCM).

We need a set of templates which is highly discriminating. That is, the collection of responses from the set of PCMs should provide accurate assignment of the target segment to the correct chromosome class. We assume that this requires the PCMs to be specific for their “own” chromosomes. That is to say the response of a PCM should be higher when matched to a target segment corresponding to the one from which it was trained than when matched elsewhere. Depending on how characteristic and consistent the banding pattern is at different locations, the specificity of the PCMs will vary. We select from the very large set of possible PCMs, a smaller “specific” set. If the specificity of each of the selected PCMs were very high, it would be sufficient to classify each target segment on the basis of the best matching PCM. It turns out that even the selected PCMs are not that specific (Fig. 6), and we classify the target segments on the basis of the response vector – the vector of maximum responses from each PCM.

The reason for classifying segments is to provide a mechanism for deciding which segments should be matched in resolving the overlap. The response vector is used as a feature vector in calculating probabilities of the segment being part of a chromosome from each of the classes. By treating the class assignments of each of the segments as independent events, we calculate the probability of assigning each pair of segments to the chromosome classes. The pairings that give the maximum combined probability for classification of both resulting chromosomes provide the resolution of the segmentation uncertainty and the final classification.

Assuming that the class probabilities assigned to each segment are independent, we apply Bayes’ formula. The use of Bayes formula in contextual assignments is well covered in standard sources (e.g. [20]). Here we summarise the steps in arriving at a resolution of the overlap.

The probability, $P(u_k)$, of each segment, $x_i$, being a member of class $k$ from the 24 classes is given by

$$P(u_k) = P(g_k|x_i) = \frac{P(g_k) P(x_i|g_k)}{\sum_{j=1}^{24} P(g_j) P(x_i|g_j)} \quad k = 1, 2, \ldots, 24.$$  \hspace{1cm} (1)
\(P(g_k)\) is the prior probability of obtaining class \(g_k\) (equal for all classes except the sex chromosomes), \(P(x_i \mid g_k)\) is the probability of obtaining the features \(x_i\), given the class \(g_k\). This probability is given by application of a suitable classification function to the feature vector. We use a linear classifier (Section 5.4).

The probabilities that candidates should be matched to form a single (identified) chromosome are calculated according to

\[
P(c_{ij}) = \max_k (P(u_{ik}, u_{jk})) = \max_k (P(u_{ik})P(u_{jk})),
\]

\[
k = 1, 2, \ldots, 24.
\]

\(P(c_{ij})\) is the maximum probability that segments \(i\) and \(j\) belong to a single chromosome from any class.

The resolution of the overlap is the configuration which gives the maximum combined probability for both chromosomes

\[
P(s_{ijlm}) = P(c_{ij}, c_{lm}) = P(c_{ij})P(c_{lm}),
\]

where \(P(s_{ijlm})\) is the probability that segments \(i\) and \(j\) form a single chromosome and segments \(l\) and \(m\) form another.

3. Chromosome data and experimental approach

For our study we have used a digitised set of prometaphase chromosomes which have been used in a number of previous classification studies [21–24]. This data set (known as the 600-band data set) consists of 6177 chromosome banding profiles from 136 G-banded blood cells, fixed and stained at the 600-band stage. Each chromosome is represented by its density profile (Fig. 2) together with its centromere position and class specified by a trained cytogeneticist. Isolated images of each chromosome are also available, although not used in this study. The profiles in the data set carry with them an identifier for their cell of origin. In some circumstances (such as the simulation experiments of Section 6) it is useful to consider the chromosomes in the context of complete cells, as would occur in “live” analysis.

Approximately 34\% of the chromosomes in the data set were touching neighbouring chromosomes in the original images. For the purposes of the earlier studies (for which the data were collected), cells were selected which contained very few overlaps; no information is retained concerning which of the chromosomes were involved in these.

In the following section, we describe some details of PCM representation, how the fit value is calculated and the method of assigning segments to classes. The method of selecting a “specific” set of PCMs from an initially large set of candidates is described in Section 5. Cross-validation experiments are used to evaluate the specificity of individual PCMs. In Section 6 we evaluate the selected PCMs for their ability to classify whole chromosomes and chromosome segments, and to resolve overlaps simulated from the 600-band data. Paradoxically, the “clean” nature of the data is an advantage for this study, as we can use simulation to generate a large set of overlaps with known correct resolution.

4. Trainable template matching

As we use the templates to generate probabilities of banding segments belonging to each chromosome class, the classes need to be defined in terms of observations made on the image features. While the banding patterns of chromosomes are characteristic of individual classes, there is considerable within-class variability. Furthermore, variations in the protocols for chromosome sample preparation lead to differences in the appearance of chromosomes imaged in different cytogenetic laboratories. It is important that templates represent the range of banding patterns that are likely to be observed in chromosome images, both in terms of an average banding pattern and the allowed variability. We capture the appearance and variability of the chromosomes by defining the PCMs with respect to a training set of banding profiles. PCMs are trainable templates. We require a representation of the banding sequence which can be determined by training and which includes a description of variability. We also need a method of assessing how closely the template fits to the density profile of a target chromosome segment. Having established a set of templates we require a method for using these templates for assigning the segments to chromosome classes.

4.1. Template representation

There are many possible ways of representing a banding profile, and several of these have been used in constructing chromosome classifiers [11,12]. Errington and Graham [13] have noted that the sequence of profile samples itself is as effective a representation as any. The banding profile illustrated in Fig. 2 is sampled at single pixel intervals along the chromosome axis. For classification purposes, the sampling interval can be much larger than a single pixel, and Errington and Graham used coarsely sampled profiles in their neural network classifier. This representation has also been used by Nivall [24] in classifying chromosome profiles, including the 600-band data set, and we adopt it for this study (see Fig. 4). The most appropriate density of profile sampling for template matching is a parameter to be determined empirically (Section 5.2).

The PCM template is a mean profile constructed from the corresponding profile segments in each cell in the training set, together with the covariance matrix. The different “raw” profiles in the training set corresponding
Fig. 4. The stages of training a PCM. (a) A training set of \( n \) homologous chromosome profiles is selected. The “raw” profiles are sampled at intervals of a single pixel along the chromosome axis (Fig. 2). Homologous chromosomes from different cells vary in the detailed form of the profile as well as in length, and hence in their number of samples. A particular PCM corresponds to a section of the profile, indicated by the heavier line. The start and end positions are specified in fractions of the chromosome length (in this case 0.15–0.6). (b) The appropriate sections are extracted from each training example and resampled more coarsely to give a fixed number of samples in each. (c) The variation in corresponding sample values is used to calculate a mean profile segment and covariance matrix. Templates of this kind are generated for a number of partial profile segments in each class, of different lengths and positions.

4.2. Matching method

As chromosome profiles are highly variable, there are several possible matching methods which might be

to the same chromosome class are represented by a different number of samples, due to differing degrees of contraction of different cells and small variations in segmentation parameters. These differences can be quite marked – up to 50% in profile length. If the samples are to be used as classification features, a constant length is required. In generating the coarsely sampled partial profiles, a fixed number of samples is used for each template, the start and end positions of the segments being expressed as a fraction of the chromosome length. This is illustrated in Fig. 4. A large set of PCMs is created in this way, each specified by differing start and end positions on chromosomes of all classes (see Section 5.1).
4.3. Chromosome discrimination

As we shall see in Section 5.4, 182 “specific” PCMs are defined with a range of lengths and starting points distributed along the profiles corresponding to all chromosome classes. For chromosome segmentation and identification, we seek a response from each PCM to each segment of chromosome as illustrated in Fig. 3. The unobscured segments of chromosome are easily identified, and each PCM is matched to all available locations on each segment. The response for each PCM is the maximum value of the measure of fit generated over all locations on a segment. The responses for all PCMs form the response vector for the segment. Any appropriate classifier may be used for assigning the segments to chromosome classes using the response vector. Because of constraints on the size of our training set we use a linear classifier.

5. Identifying the set of partial chromosome models

There is a very large number of candidate sub-chromosome profile segments of different lengths at different locations, each of which could form a PCM. Not all of these will provide specific matches against target profiles. We wish to identify a set of templates that collectively gives best discrimination between chromosome fragments. We assume that this set will be contained within the set of templates that match with high specificity to their “own” chromosome segments. We determine this specific set empirically using cross-validation experiments. The data are split into two sets, A and B. PCMs trained using set A are used to identify chromosomes in set B and vice versa. In splitting the data set (here and in the evaluation experiments of Section 6), assignment to subsets is made on the basis of cells, i.e. chromosomes from the same cell are always placed in the same data set.

For each trained PCM, a successful identification in the evaluation set is counted if the correct chromosome appears in one of its top two scores after matching to all possible locations. (There are two potential correct fits to each template. In the case of chromosomes with only one example in a cell, such as the sex chromosomes, a success is scored only when the top fit is correct.) Each candidate PCM is evaluated according to its number of successes (or recognition rate).

5.1. Generation of candidate sequences

In principle, we need to generate all possible candidate templates for evaluation (all lengths of template with starting points all along the chromosome density profiles derived from each class). This is a dauntingly large task. The size of the task was reduced by limiting the range of template lengths tested and evaluating the matches at points separated by more than a single sample.

The experiments of Lockwood et al. [18] showed that, for prophase chromosomes, their unique band sequences ranged from slightly less than the length of the shortest chromosome class to approximately twice the length of the shortest class. Taking this result into account, we chose to test six sizes of sequences ranging from one half to three times the length of the shortest class.

Working along the profiles, we used a separation of three samples between candidate template positions, reducing the number of tests required by approximately one-third. It will become clear later that this was a reasonable separation due to the fact that adjacent sequences produced similar results (Fig. 6). This resulted in 1308 candidate sequences for evaluation.

5.2. Optimisation of sample density

Errington and Graham [13] have noted that, for classification, the optimum number of samples used to describe the profiles is rather less than the number in the “raw” profiles obtained by single pixel sampling along the axis. The use of more coarsely sampled profiles not only reduces the computation required, but also increases classification rates. We can therefore maximise the specificity of the PCMs by selecting the correct sampling density for the templates. We select the sampling density empirically by measuring the recognition rate (as defined above) at different sampling densities. Profiles are resampled from the “raw” profiles by local averaging.

The “optimum sampling density” is the one that gives the highest recognition rate. Fig. 5 shows the results for classes 1 and 2. The graphs show the optimum number of
Fig. 5. Selecting the number of samples used to represent the template. Sample densities giving optimal performance for class 1 chromosomes (a) and class 2 chromosomes (b). Consistent results are obtained for training on the two halves of the data split, and the relationship between the optimum number of samples and the measured length of the profile is approximately linear.

5.3. Candidate template evaluation

The 1308 candidate templates generated as described in Section 4.1 recognise their “own” banding sequences with different specificities. In this section we describe the process of selecting the set of the most specific PCMs. Fig. 6 shows the variation in specificity for PCMs of different lengths derived from different locations on chromosome 1. Specificity is assessed by the recognition rate. Results from the separate evaluation experiments are shown, indicating that the independent training sets are broadly comparable. The average density profile for class 1 chromosomes is shown in Fig. 6a. The rest of the figure shows the results of matching the six different lengths of PCM (b-g). In each of the graphs, the horizontal axis represents the location of the PCM along the profile. (The structure of the profile at relevant points can be determined from the density curve at (a).) The vertical axis of each curve is the recognition rate achieved for a PCM trained at that location. The short numbered bars indicate the positions and lengths of the PCMs finally selected, and give an indication of the length of PCMs evaluated at each level. The differences in the recognition rates of templates of different lengths generated from different locations are clear, and correspond to intuition. For example, the long lightly stained region at the right-hand end of Fig. 6a is a readily recognisable feature of chromosome 1, and the templates derived from this region achieve high recognition rates at all template lengths. Conversely, the region in the middle of chromosome 1 is not very characteristic and gives poor recognition rates at all lengths. Generally, PCMs of different lengths at the same location give similar responses, although sometimes a high level of specificity is obtained at one particular length which captures a locally characteristic banding appearance. An example of this is the section labelled 5 in Fig. 6c, where a locally high response occurs, in contrast to the responses of longer and shorter PCMs at that location. Notice that the recognition rate for any individual template is never very high (about 75% at best); it is their use in combination which gives specificity. Notice also that specificity varies slowly along the chromosome, justifying the strategy of generating templates centred on every third sample.

Similar specificity diagrams were generated for all classes, and templates were selected from the complete set of 1308 according to the following criteria.

1. The most specific templates of each class were chosen (as in Fig. 6).
2. Where several templates of the same length shared substantial sections of profile, only the most specific of them was used. Short templates overlapping with longer ones were retained, even though they may be less specific, on the grounds that they may be useful in resolving overlaps where sections of longer templates might be obscured.

Application of these selection criteria reduced the set to 182 templates, of which the fourteen selected for chromosome 1 are shown in Fig. 6.

5.4. Using PCMs for segment classification

This set of PCMs, selected for their individual specificity in matching to their "own" chromosomes, is used to generate a feature vector for classifying target segments. In this way, the response of each PCM to a target contributes to the segment's classification. We attempted to reduce the dimension of this feature vector using Forward and Backward stepwise selection [26,27]. However, removal of features consistently resulted in reduced classification performance, so the full set of 182 PCMs was retained.

Any suitable classifier might be used for classifying segments. The dimension of the feature vector is quite large (182). Although the data set is substantial, there is also a large number of classes, so that the quantity of training data for each class is limited. As a consequence we use a linear classifier. We refer to this classifier as the linear PCM classifier.

6. Evaluation of PCMs for resolving overlaps

In this section we evaluate the performance of the linear PCM classifier in classifying whole chromosomes,
classifying chromosome segments and resolving overlaps.

6.1. Cross-validation strategy

We wish to reduce bias in the estimation of our classifier performance by cross-validation. This is often achieved by splitting an annotated data set into two, using each half in turn to act as a training set for classifying the other, as in Section 5. In our evaluation experiments we require to train two sets of models: the PCM templates themselves, and the linear PCM classifier. The PCM templates used are those selected as described in Section 5.3 They are trained by gathering statistics from the chromosome fragments as described in Section 4.1. The linear PCM classifier based on these templates is itself trained using the response vector of identified profile fragments. To reduce bias, these different models should be trained on separate data, with yet further data being used for evaluation. We therefore split the data into thirds, using each subset of the data in turn for (i) training templates (the PCM training set), (ii) training the linear thirds, using each subset of the data in turn for (i) training the other, as in Section 5. In our evaluation experiments we require to train two sets of models: the PCM templates themselves, and the linear PCM classifier. The PCM templates used are those selected as described in Section 5.3 They are trained by gathering statistics from the chromosome fragments as described in Section 4.1. The linear PCM classifier based on these templates is itself trained using the response vector of identified profile fragments. To reduce bias, these different models should be trained on separate data, with yet further data being used for evaluation. We therefore split the data into thirds, using each subset of the data in turn for (i) training templates (the PCM training set), (ii) training the linear thirds, using each subset of the data in turn for (i) training

6.2. Simulation of overlaps

The chromosomes in the 600-band data set have been selected so that occlusion by overlapping was kept to a minimum. We use this set of clean profiles to conduct experiments on resolving overlaps by simulating the density profiles from overlapping chromosomes. The advantage of this approach over the use of genuine overlaps is that any number of overlapping configurations may be created, each with a known true resolution, for both training and evaluation. The disadvantage is that the appearance of the density profile at the overlap may not be totally realistic. This is not particularly problematic as overlapping regions in real chromosome images can be identified by a number of straightforward criteria [14].

We simulate overlaps by obscuring short sections of profile at randomly selected positions on pairs of chromosomes. The number of overlaps in any cell was selected at random from a range determined from the observed numbers in prophase images (about 26% of chromosomes contain at least one overlap). The positions of obscured sections of profile were selected randomly along the lengths of the profiles; the widths of the obscured sections were generated from the observed distributions of chromosome widths. Profile densities in the obscured region were set to a value darker than the normal maximum density. We use the information on the cell of origin of the chromosomes to simulate the analysis of a complete cell at a time.

6.3. Experiments

We performed the three following experiments using simulated overlaps. In each case the PCM templates were trained as described in Section 4.1 using the PCM training sets. The experiments differ in the evaluation sets used and the corresponding classifier training sets.

Among the conditions to be varied in the experiments, the training data and the evaluation data may consist of “clean” profiles (containing no overlaps) or “representatively overlapped” profiles (simulated overlaps occurring in about 26% of chromosomes). In the latter case, to obtain sufficient numbers of overlaps for training and evaluation, the overlap simulation procedure was applied to the data in three passes, generating around 11 000 segments. To evaluate the effect of training set size, a set of about twice that number was generated from six passes. The larger numbers of overlaps were generated by multiple passes, rather than a single pass with a higher overlap rate, so that the distribution of sizes of unoverlapped segments would remain representative. We will refer to the three-pass or six-pass training or evaluation sets.

Experiment 1 (Classification of Whole Chromosomes). In this experiment we sought to obtain a measure of classification performance assuming all overlaps have been correctly resolved. Three different evaluations were carried out using different regimes of profile simulation.

(i) Evaluation of “clean” profiles: no simulated overlaps introduced.
(ii) Evaluation of “representative overlaps”: each cell contained a number of overlaps as described in Section 6.2.
(iii) Evaluation on “wholly overlapped” profiles: isolated chromosomes which were not involved in simulated overlaps were excluded from the evaluation.

To obtain sufficient evaluation examples, experiments (ii) and (iii) were conducted using a three-pass evaluation set.

Linear PCM classifiers were trained on each of the three types of simulated data, and each classifier used in turn to classify evaluation data of each type: nine classification experiments in all.

Experiment 2 (Classification of Chromosome Segments). In this experiment we tested the PCM classification performance when applied to the classification of chromosome segments (uncorrupted sub-chromosome length sections of profile extracted from overlapping chromosomes in the three-pass evaluation set). From this we obtained a measure of the ability to classify chromosome fragments, distinct from the results on overlap resolution (below).
Lengths of segments varied from under 10% of chromosome length to complete chromosomes. The shortest segments were 15 profile samples long, corresponding to the shortest templates generated. Classification was tested using linear PCM classifiers trained in three different ways to compare different training regimes.

(i) Trained on whole chromosomes.
(ii) Trained on segments from the three-pass training set.
(iii) Trained on segments from the six-pass training set.

**Experiment 3 (Resolution of Overlaps).** One hundred and thirty-six-overlapping pairs of chromosomes (one pair from each cell in the data set) were simulated. Each overlapping pair consisted of four segments. We performed overlap resolution experiments using linear PCM classifiers, trained using three-pass and six-pass training sets respectively. Overlap resolution was conducted as described in Section 2 (Eqs. (1)–(3)).

### 6.4. Results

**Experiment 1.** Table 1 shows the results for classification of whole chromosomes using PCMs. Each column corresponds to one of the three forms of data used to train the classifier, and each row corresponds to the data classified.

**Experiment 2.** Table 2 shows the results for the classification of chromosome segments according to the training data used to generate the classifier. Training on chromosome segments is clearly superior to training on whole chromosomes, and some advantage is gained from the larger training set.

**Experiment 3.** Table 3 shows the results of using the PCM templates to resolve overlapping pairs of chromosomes. The rows correspond respectively to the smaller and larger training set for the linear PCM classifier. Each row shows the percentage of overlaps correctly resolved and the percentage of the correctly resolved overlaps which were correctly classified.

### 7. Discussion and conclusions

Experiment 1 confirms the expected result that classifying "clean" chromosomes gives better results than classifying chromosomes with overlaps. The absolute classification rate for complete clean chromosomes is encouraging, and compares well with previously published classification methods. Table 4 shows a comparison with the results of previous classification studies using the 600-band data. The first column shows the

#### Table 1
Whole chromosome classification using linear PCM classifiers. Correct classification rates are shown for training and evaluation on clean, representative and wholly overlapped data (see text). Best classification is obtained when the appropriate data are used for classifier training.

<table>
<thead>
<tr>
<th>Training data</th>
<th>Clean</th>
<th>Representative</th>
<th>Overlapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification data</td>
<td>Clean</td>
<td>90.8%</td>
<td>88.8%</td>
</tr>
<tr>
<td></td>
<td>Representative</td>
<td>77.5%</td>
<td>83.7%</td>
</tr>
<tr>
<td></td>
<td>Overlapped</td>
<td>52.6%</td>
<td>64.4%</td>
</tr>
</tbody>
</table>

#### Table 2
Classification of chromosome segments. Correct classification rates of isolated chromosome segments using the linear PCM classifier trained on whole chromosomes (without simulated overlaps) and on fragments extracted from the training set. Training on fragments is clearly superior, and the larger number of fragments in the six-pass training set gives advantage (see text).

<table>
<thead>
<tr>
<th>Classifier training</th>
<th>Percentage correct classifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole (clean) chromosomes</td>
<td>32.8</td>
</tr>
<tr>
<td>Chromosome segments (3 passes)</td>
<td>51.7</td>
</tr>
<tr>
<td>Chromosome segments (6 passes)</td>
<td>55.4</td>
</tr>
</tbody>
</table>

#### Table 3
Results for resolving overlapping chromosome pairs. Correct overlap resolution rates for linear PCM classifiers trained on segments derived from three passes and six passes of segment generation from the training set. The rightmost column shows the number of correctly identified chromosomes which were also correctly classified.

<table>
<thead>
<tr>
<th>Segment training set</th>
<th>Correctly resolved overlaps</th>
<th>Correctly classified chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three-pass training set</td>
<td>66.5%</td>
<td>79.6%</td>
</tr>
<tr>
<td>Six-pass training set</td>
<td>70.6%</td>
<td>82.6%</td>
</tr>
</tbody>
</table>
result of using trainable templates to classify clean chromosomes with whole (rather than partial) chromosome templates and a quadratic (i.e. multivariate gaussian), instead of a linear, classifier. The classification performance is almost identical with that shown in Table 1, indicating that the PCMs provide as complete a description of the banding pattern as the full profiles. If anything, they do slightly better. The remainder of Table 4 sets this performance in context by comparing the template matching classifier of column 1 with the results of Kleinschmidt et al. [23] and Nivall [24], who have previously achieved the best classification performance on the 600-band data. Both of these studies used, in addition to the banding pattern, the powerful features of chromosome length and centromere position (Fig. 2), which are not available to the PCM classifier. The second column of Table 4 shows the result of the template matching classifier of column 1 when these additional features are used. Template matching achieves an improvement in classification over both Kleinschmidt and Nivall. Although these improvements are significant (at the 1.4% and 5% levels respectively), they are small. The object of the comparison is not to achieve a better classifier, but to demonstrate that the form of the template and the matching method are capable of creditable results in recognising banding patterns on prometaphase chromosomes, and that PCMs adequately represent the banding information.

Tables 1 and 2 show that the performance of the linear PCM classifier is best when trained with data of the same type as is being classified. It is not surprising that the classification of “clean” chromosomes is best done using “clean” templates, but it is more difficult to see why the converse should be true. At the moment, we have no explanation for this observation. However, using the figures of Table 1, we propose that if we have no idea whether a chromosome to be classified is isolated or overlapped then the best results for classifying representative overlapped data is 83.7%. In analysing a chromosome image, it is usually possible to know which chromosomes are isolated and which are involved in overlaps. In which case an appropriately trained classifier could be used for each chromosome. Given our observation that 26% of chromosomes are typically involved in an overlap, then the best result we can obtain is approximately 84.8% (0.74 × 90.8 + 0.26 × 67.7). We could improve this rate to 86.0% if we were to use the quadratic discriminant template classifier to classify the clean chromosomes (Table 4).

The classification rates presented here will be to some extent underestimated. The assumption that the chromosomes in the 600-band data set contain no overlaps is not entirely true. There is a small (unknown) level of residual overlap in the data, providing an element of noise in the measurements, which will result in a slight depression of the classification performance. This does not affect our conclusions, as the same assumption has been made in all studies making use of this data set.

Experiment 2 indicates that chromosome fragments can also be classified fairly well. The fragments range in length from about 10% to about 90% of the chromosome length. It is unsurprising that the correct classification rate is much lower than for isolated chromosomes. Experiment 3 shows that about 70% of simulated overlaps can be correctly resolved using banding information alone. Of those correctly resolved 82.6% are also correctly classified. For overlap resolution we can tolerate the modest classification performance for segments observed in Experiment 2, and it is possible to resolve the overlap without correctly identifying each of the chromosomes. If two chromosomes overlap, it is sufficient to have good evidence for identifying one of them, provided there is no strong evidence for an alternative erroneous interpretation.

The approach adopted in this study expands on the “unique band sequences” of Lockwood et al. [17–19]. They identified a set of 94 such sequences [18], which were used as templates to be matched to candidate chromosomes. We have extended this idea in two important ways to provide the banding evidence for chromosome segmentation and classification. Firstly, we incorporate knowledge of profile variability into the choice and use of

<table>
<thead>
<tr>
<th>Template matching by quadratic classifier (banding data only)</th>
<th>Template matching by quadratic classifier (including length and centromere position)</th>
<th>Kleinschmidt et al.</th>
<th>Nivall</th>
</tr>
</thead>
<tbody>
<tr>
<td>90.2%</td>
<td>92.4%</td>
<td>91.3%</td>
<td>91.6%</td>
</tr>
</tbody>
</table>
sequence selection is determined by measured specificity of the templates, rather than by an intuitive assessment of “uniqueness”. Lockwood et al. did not evaluate their approach fully for classification, concentrating on evaluating matching specificity. Using cross-correlation to match templates to about 20% of the possible profile positions on 850-band chromosomes they correctly identified 88% of template sequences [18]. Our results demonstrate that this approach, applied to somewhat more condensed chromosome material, can achieve results comparable to classification on whole chromosome profiles.

The principal motivation for this study was the resolution of overlapping chromosomes, which we have considered as quite separate from the case of clusters of touching chromosomes. Previous work [14,16] has demonstrated that geometric evidence concerning local boundary shape can lead to fairly accurate extraction of individual chromosomes from touching or slightly overlapping configurations. In principle, banding information could also be brought to bear to resolve touching clusters, but in that case, the classification of complete hypothesised chromosomes could be used. We have sought here to concentrate on the case of total overlaps, where complete banding information is not available. The most successful previous study with this aim is that of Ji [14], who used purely geometric reasoning. Geometric cues are often powerful for resolving overlaps (see Fig. 1), and Ji achieves a correct resolution rate of 94.6% in resolving 46 overlaps. He subsequently showed how his overlap resolution method can be combined with splitting touching chromosomes for successful automatic segmentation of unband ed chromosomes [15] (stained to be uniformly dark – used in counting chromosome aberrations for environmental monitoring).

On the face of it, Ji’s approach gives significantly better overlap resolution. However two points can be made. Firstly, Ji’s study [14] illustrates a difficulty in carrying out this type of investigation: his methods were tested on relatively small numbers of overlaps. This arises from the difficulty of identifying a sufficient number of configurations of chromosomes which are suitable for the analysis, and for which “correct” solutions are known for both training and evaluation. The approach taken here overcomes this difficulty by simulating the appearance of overlaps from “clean” data. We can therefore generate as many partial chromosomes, with known classifications, as we wish. There is a potential criticism of such an approach, in that it requires the appearance of the simulated overlaps to be realistic. Since we use only the banding profile information away from the obscured regions, we feel safe that nothing of significance is lost in pretending that these sections came from genuinely overlapping chromosomes.

Secondly, the method we describe here uses the banding pattern as the only source of evidence for segmentation. Rather than being seen as an alternative, the banding pattern provides additional evidence which can be used in conjunction with geometric information to provide a more informed basis for the assessment of segmentation hypotheses. We have demonstrated that this source of evidence alone can provide a useful contribution to resolving the segmentation uncertainty. The issue of trainability is important here: using trainable models means that the methods are not tied specifically to the properties of a given type of material, nor are they critically dependent on the setting of arbitrary heuristic parameters. Furthermore, basing features on a training set results in measures of compatibility between segments which approximate to true probabilities and which could, in principle, be used in combination with geometric cues to improve the performance of both approaches. Problems of relative scaling between disparate sources of evidence can be overcome using Bayesian methods if all evidence is expressed as probabilities. We have investigated [25] how PCMs may be combined with a trainable geometric method, using the images of isolated chromosomes available with the 600-band data, and will describe this in a later publication.

The disadvantage of using trainable models is in the necessity for a large training set. The 600-band data set used in this study was adequate in terms of numbers of samples and length of chromosomes to demonstrate feasibility. The methods would be used to best effect in the segmentation and classification of chromosomes from the prophase stage of contraction. Previous investigations into segmenting overlapping chromosomes have relied on reasoning about chromosome shapes to resolve the ambiguities in interpretation. Information contained in the chromosome banding pattern can also be used for this purpose. We propose and evaluate a mechanism of using the banding information based on trainable grey level models. The models, referred to as Partial Chromosome Models, consist of a set of templates corresponding to banding sequences of sub-chromosome length, selected so that they provide good discrimination between chromosome classes. Candidate profiles are matched to templates using a quadratic classification function. Chromosome segments are assigned to chromosome classes on the basis of their responses to the entire set of templates, by using the matching scores.
as features in a linear classifier. The classifications of the segments are then used to propose matches to identify complete chromosomes within a composite object. We evaluate the method using a set of chromosome banding profiles derived from prometaphase chromosomes, whose classes have been expertly identified. The form of the model and the matching method are shown to be capable of high specificity, achieving correct classification results on whole chromosomes, using whole chromosome models, of 92.4% which improves on previously published classification results on this set of data. Using Partial Chromosome Models, a correct classification rate of 90.8% is obtained for isolated whole chromosomes and 55.4% for chromosome fragments, some of which represent less than 10% of the chromosome length. We test the ability of the models to resolve overlaps by simulating overlapping pairs of chromosomes using the profile data set. Despite the rather low rate of correct classification for chromosome fragments, 70.6% of simulated overlaps are correctly resolved. We discuss the possibility of combining the use of grey-level cues with geometric cues for untangling overlapping chromosomes.

Acknowledgements

The data used in this study were made available as part of the Concerted Action of Automated Cytogenetics Groups, Project No. II.1.1/13 and the Concerted Action on Automated Molecular Cytogenetic Analysis, Project No. BMH1-CT92-1307, supported by the European Community. We are grateful for the co-operation of colleagues within these projects.

References

About the Author—GRAHAM C. CHARTERS received his B.Sc. in Computer Science in 1990, his M.Sc. in Numerical Analysis and Computing in 1991 and his Ph.D. in Machine Vision in 1995, all from the University of Manchester, England. He is currently with the IBM U.K. Scientific Centre in Winchester, England, where his research interests are Medical Imaging and Pattern Recognition.

About the Author—JIM GRAHAM received the B.Sc. degree in Physics from the University of Edinburgh, Scotland in 1974 and his Ph.D. in Structural Biology from the University of Cambridge, England in 1978. He joined the Wolfson Image Analysis Unit at the University of Manchester in 1978, where he is currently Senior Lecturer in Medical Biophysics and Honorary Lecturer in Computer Science. His research interests are biomedical and industrial applications of Machine Vision and Pattern Recognition.