© 2001 Wiley-Liss, Inc.
DOI 10.1002/cyto.10027

Cytometry 47:17–23 (2002)

Karyotyping of Comparative Genomic Hybridization Human Metaphases by Using Support Vector Machines

Zhenzhen Kou, Liang Ji,* and Xuegong Zhang
Department of Automation, Tsinghua University, Beijing, Peoples Republic of China

Received 20 March 2001; Revision Received 6 September 2001; Accepted 22 September 2001

Background: Comparative genomic hybridization (CGH) is a relatively new molecular cytogenetic method for detecting chromosomal imbalance. Karyotyping of human metaphases is an important step to assign each chromosome to one of 23 or 24 classes (22 autosomes and two sex chromosomes). Automatic karyotyping in CGH analysis is needed. However, conventional karyotyping approaches based on DAPI images require complex image enhancement procedures.

Methods: This paper proposes a simple feature extraction method, one that generates density profiles from original true color CGH images and uses normalized profiles as feature vectors without quantization. A classifier is developed by using support vector machine (SVM). It has good generalization ability and needs only limited training samples.

Results: Experiment results show that the feature extraction method of using color information in CGH images can improve greatly the classification success rate. The SVM classifier is able to acquire knowledge about human chromosomes from relatively few samples and has good generalization ability. A success rate of moe than 90% has been achieved and the time for training and testing is very short.

Conclusions: The feature extraction method proposed here and the SVM-based classifier offer a promising computerized intelligent system for automatic karyotyping of CGH human chromosomes. Cytometry 47:17–23, 2002. © 2001 Wiley-Liss, Inc.

Key terms: CGH; chromosome classification; karyotyping; SVM

Chromosome analysis is an essential procedure for detecting genetic abnormalities or damages due to environmental factors and for diagnosing cancer. Karyotyping, as a part of the task, is an important step to assign each chromosome to one of 23 or 24 classes (22 autosomes and a pair of sex chromosomes). Because manual karyotyping is tedious and time-consuming, great efforts have been made to develop computer-aided classifiers during the past years (1). So far as classification is concerned, commercial computerized chromosome analysis systems are far inferior to analysis done by human experts.

Most routine karyotyping is carried out on Giemsa-stained metaphase images, which appear as dark images on a light background and have a characteristic pattern of light and dark bands unique to each type of chromosome. Usually, a human chromosome is characterized by its size, banding pattern, and centromere position (1,2). Density profiles and width profiles are generated to represent these characteristics (1). Various approaches, such as Fourier decomposition, mixture of several Gaussian distributions, band transition (BT), 2D Laplace filtering, and Markov networks have been proposed for extracting features from profiles. Carothers and Piper (1) provided a thorough review of the approaches presented before the early 1990s. The "knock-out" algorithm (3), principal component analysis (3,4), the Kohonen network (5), and wavelet packets (6) were also used for chromosome feature extraction. These procedures of feature extraction are relatively complex and sometimes time-consuming.

Comparative genomic hybridization (CGH; 7) is a molecular cytogenetic method developed in the 1990s to detect chromosomal imbalances. It has great potential for a broad range of applications to basic research and clinical practice, such as detecting chromosome aberrations in cancer and mapping their locations on normal chromosomes. Unlike G-banded gray-scale chromosome images, the CGH images of human metaphases are darkfield images, with true colors (red, green, and blue images with an intensity of 0–255 or higher).

Conventionally, the banding patterns obtained from the DAPI image are used for chromosome identification, which can be facilitated by using the DAPI-inverse image resembling a Giemsa banding pattern (8–11). However, this approach requires a complex image enhancement

Grant sponsor: National Natural Science Foundation of China; Grant number: 39770227; Grant number: 69885004.

*Correspondence to: Dr. Liang Ji, Department of Automation, Tsinghua University, Beijing 100084, Peoples Republic of China.
E-mail: zc-sa@mail.tsinghua.edu.cn

algorithm before satisfactory results can be achieved because the resolution and contrast of DAPI images are much lower than those of G-banded metaphase images. To date, few studies about color karyotyping have been published (12,13). Although some studies on karyotyping have investigated spectral karyotyping (SKY; 12) and multicolor fluorescence in situ hybridization (M-FISH; 13), the approaches proposed are not suitable for CGH analysis.

In this paper, we propose a feature extraction approach that uses red and blue signals in original true color CGH images to generate integrated density profiles without image enhancement. After normalization, profiles are used directly as feature vectors without further treatment such as quantization and dimension reduction. Errington and Graham (14) used profiles as feature vectors. Unfortunately, due to the limitation of neural networks on the size of input vectors, the profiles had to be coarsely quantized to get small feature vectors. This would surely lose some useful information when applied to DAPI images. In contrast, the approach investigated here uses normalized profiles directly as features without quantization and the resolution of the R, B scale of the profiles is maintained. In this way, the complex procedure of further feature extraction is avoided. This method has been proven to be more straightforward and effective.

Various classification methods, such as the linear discriminant function, fuzzy subset theory, Bayesian theorem, and nearest-neighbor rule were applied for chromosome classification in the early years (1). Since the early 1990s, neural networks have been applied to chromosome classification (3–5,14–17). However, the input vectors to neural networks should not be too large and this requires a complex feature extraction or dimension reduction procedure. Rutovitz et al. (18) used a kernel-density method to estimate the likelihood function for each class, but this is computationally feasible only for small feature vectors.

A new learning algorithm has been proposed by Vapnik (19,20), which is based on statistical learning theory. This algorithm, called support vector machine (SVM), provides the largest margin between the optimal separation hyperplane (OHP) and the closest training vectors, which results in good generalization ability with limited training samples. SVM has shown excellent performance in a number of difficult learning tasks, such as handwritten digit recognition (21) and face detection (22). Because SVM has effectively solved the curse of dimensionality (19,20), there is little limitation of the dimensionality imposed on input vectors. This may circumvent the complex feature extraction and dimension reduction procedure.

Because SVM has little limitation on the size of input vectors and can achieve good generalization ability with limited training samples, effective feature vectors can be obtained easily from the original red and blue images in CGH human metaphases without complex further treatment such as quantization or dimension reduction. We present our approach for karyotyping CGH human metaphases, using R, B color signals for feature extraction and SVM for the design of the classifier.

MATERIALS AND METHODS

CGH and Image Acquisition

All the CGH images were kindly provided by Dr. Tommy Gerdes (University of Copenhagen, Denmark). The images are of the size 748 × 573 and are in true colors (red, green, and blue images with intensity of 0–256^2). Details about the CGH slide preparation and image acquisition were described in a study by Kirchhoff et al. (9). There are 71 metaphase images with 23 classes (22 autosomes and a pair of X chromosomes) because the reference DNA was obtained from a karyotypically normal female (9). The original classification of the chromosomes was determined by Dr. Mingrong Wang and Ms. Xin Xu, biologists from the Chinese Academy of Medical Science.

Chromosome Feature Selection and Extraction

Feature selection. Conventional chromosome classification is based on complex feature extraction procedures (1,2). However, the features of CGH images are quite different from those of G-banded images. For example, the banding patterns of DAPI images are not as clear as those of normal G-banded images and the banding patterns are not exactly the same in the three channels of R, G, B (Fig. 1). There is notable difference among profile R, profile G, and profile B. DAPI-inverse images are used for chromosome classification in CGH analysis (8–11). Unfortunately, a complex image enhancement algorithm is required to get clear banding patterns before achieving satisfactory results and color information from R, G signals is neglected. In CGH images, the reference hybridization (usually red) shows “hybridization banding,” which is not fully correlated with DAPI banding and is useful for chromosome classification. For the test DNA image (usually green), the pattern of test hybridization will vary from one CGH analysis to another because of the variety of chromosome imbalance. The information contained in green images might be unrelated to the chromosome class. In this study, the integrated density profiles generated from channel R, B are selected as features. The main reason is that the integrated density profiles can provide information on banding patterns, chromosome length, and cen-

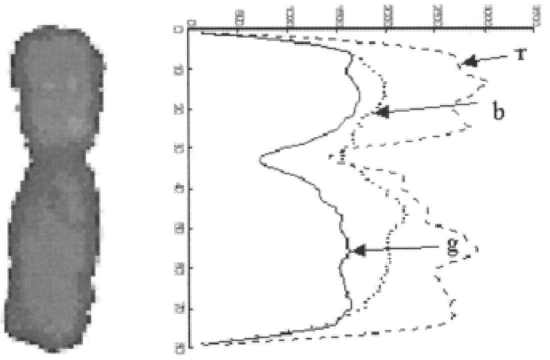

FIG. 1. A chromosome and its integrated density profiles of channels R, G, B.

tromere position. Another reason is that this feature extraction approach is relatively simple. To present the profile information consistently, all profiles are calculated from the end of the short arm to the end of the long arm. Figure 1 shows a chromosome in CGH images at the left, with its corresponding integrated density profiles at the right.

Digital Image Preprocessing

The preprocessing of CGH images aims at getting the boundary and the medial axis of each chromosome and calculating the profiles as depicted in Figure 2. CGH images are analyzed using the software developed by Liang Ji and Jiang Ni, namely, a CGH analyzer implemented on an IBM-compatible PC, manufactured by Legend Grove, Beijing, P.R. China, with the AMD Athlon 500-Mhz processor and 128 MB RAM. Some critical steps concerning karyotyping are listed below.

The RGB image is transformed into a gray image by calculating the average of the three components. The chromosomes are segmented according to the method described in Ji (23). The segmented chromosomes are then used as masks to obtain chromosomes from the original color image. The medial axis of each segmented chromosome is then determined by using the Hilditch skeleton (24). Along each normal of the medial axis, the red and blue signals are measured to get integrated density profiles.

The short arm of a chromosome is defined as the beginning part of its profiles. The following method is used to locate the centromere of a chromosome. In CGH images, the centromeric region appears blue because it is blocked by Cot-1 DNA during CGH experiments. This feature of CGH images is used to determine the position of the centromere. A gray image is obtained by the following formula

$$Gray = \begin{cases} B - \dfrac{R+G}{2} & if\ B - \dfrac{R+G}{2} \geq 0, \\ 0 & otherwise \end{cases}$$

where *Gray* is the gray value for the new image and *R, G, B* are the three color components of the corresponding CGH image. The average gray value is then measured along each normal of the medial axis. The centromere of a chromosome is determined as the position on the medial axis with the largest average gray value. If the largest value is zero, the centromere is found using the method described by Piper and Granum (2). Then, the end of the short arm is determined and the profiles are arranged in the same order.

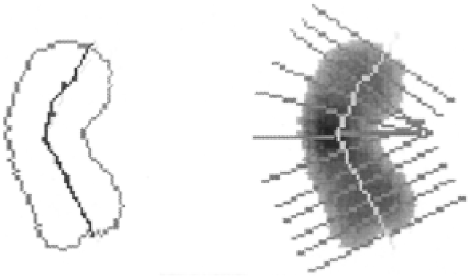

FIG. 2. The boundary and the medial axis of a chromosome and profile calculation.

The density profiles generated from channel R, B are labeled with d_r, d_b, respectively, and presented in the following vectors:

$$d_r = (d_1^r, d_2^r, \ldots, d_N^r) \qquad d_b = (d_1^b, d_2^b, \ldots, d_N^b),$$

where N is the length of the chromosome medial axis.

Normalization and feature vector. There are large variations in the appearance of metaphases due to various contractions and intensities of staining. In addition, some chromosomes may be missing from, or added to, a normal metaphase. Because such differences are not directly relevant to chromosome classification, their effects must be removed from feature measurements by appropriate normalization (1).

For length measurements, a simple approach of normalization is adopted by setting the length of the medial chromosome in each metaphase to a fixed length. The lengths of other chromosomes in the same metaphase can then be standardized by applying a linear transform:

$$\hat{N} = \frac{N * \hat{N}_{med}}{N_{med}},$$

where $\hat{N}$ is the length of a chromosome after normalization, N is the original length of the chromosome, N_{med} is the length of the medial chromosome in the metaphase, and $\hat{N}_{med}$ is a constant.

This method is chosen because the length of the medial chromosome in a metaphase is relatively steady and reliable in spite of missing chromosomes in some metaphases. The normalized length is chosen to be longer than any of the original ones of the same class. Interpolation is needed to obtain the normalized profile. Linear interpolation, Lagrange interpolation, and spline interpolation were tested and the best result was obtained by using the spline interpolation method. Therefore, the spline interpolation was chosen. After the spline interpolation, two normalized feature vectors $\hat{d}_r$, $\hat{d}_b$ are obtained.

Because input vectors of SVM should have the same dimension while the lengths of chromosomes are different, a standard length of profiles is needed. All lengths of profiles are extended to a maximum length, $N_{\max}$, and finally standard profiles are obtained as the following:

$$\bar{d}_r = (\bar{d}_1^r, \bar{d}_2^r, \ldots, \bar{d}_k^r, \ldots, \bar{d}_{N\max}^r)$$

$$\bar{d}_k^r = \begin{cases} \hat{d}_k^r & 1 \leq k \leq \hat{N} \\ 0 & \hat{N} < k \leq N_{\max} \end{cases}$$

$$\bar{d}_b = (\bar{d}^b_1, \bar{d}^b_2, \ldots, \bar{d}^b_k, \ldots, \bar{d}^b_{N_{\max}})$$

$$\bar{d}^b_k = \begin{cases} \hat{d}^b_k & 1 \le k \le \hat{N} \\ 0 & \hat{N} < k \le N_{\max} \end{cases},$$

where $\hat{N}$ is the normalized length of a chromosome and $N_{\max}$ is a constant.

This approach of normalization, expansion, and padding zeros also provides the information on the chromosome length, so that other features about chromosome length are unnecessary. The variations between metaphases are reduced after normalization.

Density profile. The normalized integrated density profiles d_r, d_b of each chromosome are calculated using the algorithm presented above.

Feature vector. The density profiles of a chromosome can be joined to get a feature vector of 2* $N_{\max}$ dimension. Because the use of kernels in SVM overcomes the curse of dimensionality and the fast algorithm of SVM (25) has been well developed, the training time of the SVM classifier is relatively short. The selection of such 2* $N_{\max}$ dimension vectors as features without any dimension reduction or quantization does not decrease much the efficiency of the classifier. This approach avoids loss of useful information during feature extraction and is simple and straightforward. Consequently, a chromosome can be characterized by the following feature vector:

$$X = (\bar{d}_r, \bar{d}_b).$$

SVM for Classification

The role of a classifier is to learn the classification rule from training samples and then to apply the rule to new samples to make decisions or predictions. Thus, for a classifier, one of the most important properties is its generalization ability or its ability to make correct predictions not only on the training data, but also on test data previously unseen in the training phase. The SVM algorithm contains a term to control the generalization ability so that optimal generalization can be obtained based on only limited training samples.

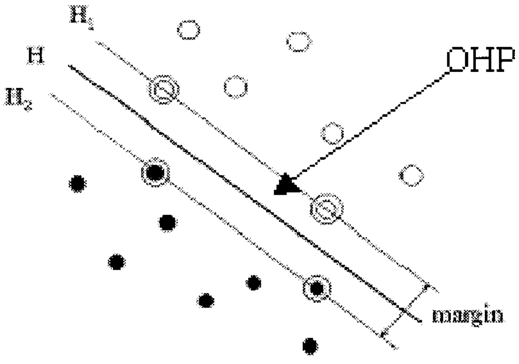

FIG. 3. The OHP and the margin. The dots and circles represent samples of class 1 and class 2, respectively.

For linearly separable two-class cases, it could be understood intuitively that the separation hyperplane that provides the largest margin from the plane to the closest training vectors would be optimal (Fig. 3). The basic idea of SVM is to seek the goal that not only all training samples are classified correctly, but that the separation margin is maximized. It has been proven that this will result in the best generalization ability (19,20,26).

For linearly nonseparable cases or cases where not all the training samples can be correctly classified simultaneously, SVM seeks a balance between minimal training errors and maximal separation margin. This goal can be realized by the following mathematical problem.

Solving for a linear separation function $y = \mathrm{sgn}[\mathbf{w}\cdot\mathbf{x} + b]$ that minimizes

$$(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right) \qquad (1)$$

subject to

$$y_i[(\mathbf{w}\cdot\mathbf{x}_i) + b] - 1 + \xi_i \ge 0, \quad i = 1, \ldots, n \qquad (2)$$

where $(\mathbf{w}\xi)$ is the optimization goal, $\mathbf{x} \in R^m$ is the feature vector, $y = \{-1, +1\}$ is the label of the two classes, $\frac{2}{\|\mathbf{w}\|^2}$ is equal to the margin (minimizing $\|\mathbf{w}\|$ corresponds to maximizing the separation margin), $\xi_i \ge 0$, $i = 1, \ldots, n$ are slack variables that control the training errors, and parameter C quantifies the trade-off between training error and system generalization ability (19,20,26).

The solution of this problem can be obtained by using the Lagrangian multipliers. This leads to the following quadratic programming (QP) problem. The detailed demonstration can be found in studies by Vapnik (19,20) and Zhnag (26). Maximize

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j (\mathbf{x_i}\cdot\mathbf{x_j}) \qquad (3)$$

subject to

$$0 \le \alpha_i \le C, \qquad i = 1, \ldots, n \qquad (4)$$

where α_i are the Lagrangian multipliers. The solution of original problem is then

$$f(\mathbf{x}) = y = \mathrm{sgn}\{(\mathbf{w}\cdot\mathbf{x}) + b\} = \mathrm{sgn}\left\{\sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i\cdot\mathbf{x}) + b\right\} \qquad (5)$$

where $f(\mathbf{x})$ represents the decision rule. There are several approaches to solve this QP problem. In Collobert and Bengio (25), a fast algorithm was proposed to solve large-scale problems.

Linear SVM can be generalized to nonlinear SVM. The basic idea is to transform the original input space $\{\mathbf{x}\}$onto a high-dimensional space by some nonlinear mapping $\Phi(\mathbf{x})$, and then find the optimal hyperplane in this feature space. Usually this kind of mapping would cause the problem known as the curse of the dimensionality. However, it has been discovered that in SVM, this kind of mapping need not be computed explicitly (19,20,26). Because only inner products are calculated in the high-dimensional feature space, provided certain conditions hold, these inner products can be calculated equivalently in the input space by some kernel $K(\mathbf{x}_1,\mathbf{x}_2) = \Phi(\mathbf{x}_1) \cdot \Phi(\mathbf{x}_2)$. The mathematical problem of linear SVM (described by formulae 3,4,5) can be modified to that of the nonlinear SVM:

$$\text{maximizing} \quad Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j}) \quad (6)$$

$$\text{subject to} \quad 0 \le \alpha_i \le C \quad (7)$$

and the obtained separation function is

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum y_i \alpha_i K(\mathbf{x}_\mathrm{i}, \mathbf{x}) - b \right) \quad (8)$$

Different nonlinear classifiers can be realized by different kernels. In our experiments, the Gaussian kernel $K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left\{ -\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2} \right\}$ is chosen. Here, σ is the parameter of the Gaussian kernel.

Because of the limitations of this study, the theoretical research results on SVM are not included. Further details of the SVM algorithm and its implementations are available in other studies (19,20,25,26).

Design of the Classifier

Multiclass classifier. As described in the previous section, the standard SVM was originally proposed for binary classification. Typically, a multiclass classifier is constructed by combining several binary classifiers, for example, the strategy of "one versus the other" and "pairwise" (27). How to effectively extend SVM to multiclass classification in one step is still an ongoing research issue. For the methods that consider all classes of data at once, a much larger optimization problem is required. Until now, experiments have been limited to small data sets (19). Experimental comparison between the one versus the other and pairwise strategies shows that the two approaches achieved approximate classification accuracy. However, the one versus the other strategy is much simpler and faster. Finally, we adopt the one versus the other scheme to build the n-class classifier based on two-class SVM (19). This is done by (1) constructing n two-class SVM classifiers $f_k(\mathbf{x})$, $k = 1,2,\ldots,n$, where rule $f_k(\mathbf{x})$ separates the data of class k from those of all other classes, i.e., $\mathrm{sgn}[f_{k(\mathbf{x}_\mathrm{i})}] = 1$ if data $\mathbf{x}_i$ belongs to class k, and $\mathrm{sgn}[f_k(\mathbf{x}_\mathrm{i})] = -1$ otherwise and (2) constructing the n-class classifier by choosing the class m corresponding to the maximal value of functions $f_k(\mathbf{x})$, $k = 1,2,\ldots,n$:

$$m = \arg\max\{f_1(\mathbf{x}_\mathrm{i}), \ldots, f_n(\mathbf{x}_\mathrm{i})\}$$

Context-dependent classifier. 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 (1). Therefore, a simple context-dependent classifier is developed based on the number constraint of the homologue chromosomes in a metaphase. In the karyotyping results of a metaphase, each class contains at most two chromosomes.

The constraint can be enforced as follows. First, chromosomes in a metaphase are classified using the multiclass classifier based on SVM. The karyotyping results are checked to find whether there exists any class containing more than two chromosomes. If such a class is found, it should be labeled. All the chromosomes having been classified into the labeled class are rearranged according to their decision function values on the classifier for the labeled class. Then, the two chromosomes with the first and second largest decision function values are finally classified to the labeled class. The rest of the chromosomes in the class are relabeled as rejected samples. If there exist some classes that contain more than two chromosomes, there must exist some classes that contain zero or one chromosome, and these are labeled as nonfull classes. All rejected samples are redistributed into the nonfull classes using the same scheme as used in the multiclass classifier described above. The check and redistribution procedures are carried out iteratively until there is no rejected sample.

RESULTS

In this study, the total experiment data set consists of 71 CGH images of human metaphases. It is a small set. All the metaphases are from females. Therefore, samples of only 23 classes (22 autosomes and the X sex chromosome) can be offered. Among all metaphases, 51 were selected randomly as the training set and the remaining ones were used as the test set.

Experiment 1: Using Integrated Density Profiles of B Signals as Features

Conventionally, DAPI images are used for chromosome identification in CGH analysis (8–11). Table 1 lists the experiment results by selecting the integrated density profiles of original DAPI signals in CGH images as feature

Table 1
Classification Results by Using Density Profiles of DAPI Signal as Features

Train data	Test data	Total errors	Success rate
2,273	893	217	75.70%

Table 2
Classification Results By Using Density Profiles of R, B Signals as Features

Train data	Test data	Total errors	Success rate
2,273	893	92	89.70%

Table 4
Classification Results By Using Density Profiles of R, G, B Profiles as a Feature and a Context-Free Classifier

Train data	Test data	Total errors	Success rate
2,273	893	83	90.71%

vectors. Choosing the optimal parameter (the one used when achieving the highest success rate) for the Gaussian kernel and using a context-free multiclass SVM classifier, a success rate above 75% was obtained.

Experiment 2: Using Integrated Density Profiles of R, B Signals as Features

To make use of the color information of CGH images, we chose normalized integrated density profiles generated from original R, B signals as feature vectors and used a context-free multiclass SVM classifier. The experiment results are listed in Table 2. A success rate above 90% was achieved. These results show that the color information from the R signal improves greatly the success rate.

Experiment 3: Using the Context-Dependent Classifier

A higher success rate was achieved by considering the number constraint of the homologue chromosomes in a metaphase. A higher success rate can also be obtained with the simple context-dependent classifier developed on the scheme described above. The experiment results of using density profiles from the two channels of R, B as feature vectors and the context-dependent multiclass SVM classifier are shown in Table 3. The whole algorithm is a time saver: at most, it takes 5 min to train 51 metaphases and 3 min to test 20 metaphases on an IBM PC as described previously.

CONCLUSIONS AND DISCUSSION

This study proposes a simple, fast, straightforward, and effective feature extraction method for CGH karyotyping. Different from the chromosome classification approaches that use enhanced DAPI-inverse images for classification, the method presented here uses the R, G, B signals in the original CGH images to determine the centromere position and to generate integrated density profiles. The use of kernels in SVM makes it possible to select whole normalized integrated density profiles directly as features without any further treatment such as quantization or dimension reduction. This method avoids the complex image enhancement, feature extraction and selection, or feature dimension reduction and the possible loss of useful information. The feature extraction approach is simple, straightforward, timesaving, and has been proved to be effective.

The SVM classifier offers a promising computerized intelligent system for automatic karyotyping of CGH human chromosomes. It is important for the SVM to easily use the information provided by the integrated density profiles, which makes the simple feature extraction method feasible. Good generalization ability and a success rate of above 90% can be achieved easily. Although there are only 71 metaphases available, the SVM classifier shows its good generalization ability with limited training samples. This is quite important in practice.

The Vysis CGH System, which uses enhanced DAPI-inverse images for karyotyping, achieved a 90% success rate (Dr. Jim Piper, personal communication). However, we have not seen any published results of the detailed DAPI-inverse image enhancement algorithm and its success rate. The feature extraction method and the SVM-based classifier offer a promising system for automatic karyotyping of CGH human chromosomes.

There has been discussion about whether the test DNA image (usually the green signal) can be used for chromosome classification in CGH analysis. It is true that the information carried by the test DNA image is completely unrelated to the chromosome class and that the pattern of test hybridization will have no consistency from one CGH analysis to another. Therefore, the green signals are completely unreliable when applied to chromosome classification. Because of the variety of chromosome imbalances in CGH analysis, the decision rule obtained by some data set may not have good generalization ability when applied to the unknown test data. However, experiments showed that better results could be obtained if all three profiles of R, G, B signals were used. Experiment results of selecting integrated density profiles of R, G, B signals as features are shown in Tables 4 (with a context-free classifier) and 5 (with a context-dependent classifier). In practice, some biologists use green images during manual karyotyping. For example, when the ends or centromeres are not clear in DAPI images, the red and green images are used for reference. The CGH images available refer to only two samples of test DNA (9). Further experiments are needed

Table 3
Classification Results By Taking Into Account the Number Constraint

Train data	Test data	Total errors	Success rate
2,273	893	82	90.82%

Table 5
Classification Results By Using Density Profiles of R, G, B Profiles as a Feature and a Context-Dependent Classifier

Train data	Test data	Total errors	Success rate
2,273	893	67	92.50%

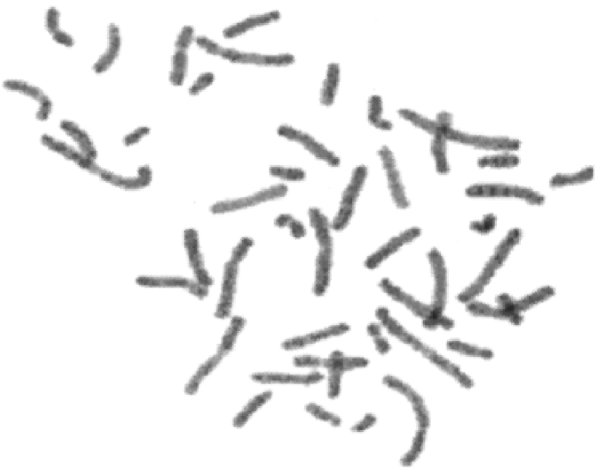

FIG. 4. A metaphase with overlapping chromosomes.

to determine whether the G signals can be generalized to wider applications. We are now collecting data from different laboratories for further research.

There is room for improvement. The selection of the kernel, its parameters such as σ of the Gaussian kernel, and the tradeoff parameter C lack sufficient theoretical guidance and finding the optimal parameter of the Gaussian kernel requires many experiments. More training data can surely increase the success rate. For example, a higher success rate has been achieved by using the Jackknife procedure. The context-dependent algorithm used here is simple and could be improved by modifying it. The images of CGH metaphases are not suitable for classification training. A lot of overlapping chromosomes can be found in the metaphases (Fig. 4). Part of the banding patterns in the profiles corresponding to the overlapping region cannot be restored completely. There are less than 42 chromosomes in some metaphases. These factors worsen the experiment results. Because all the profiles are generated from original CGH images, image enhancement before profile generation may improve the success rate.

ACKNOWLEDGMENTS

We thank Dr. Tommy Gerdes (University of Copenhagen) for providing the CGH images; Dr. Mingrong Wang and Ms. Xin Xu (Chinese Academy of Medical Science, Institute of Cancer Research, National Key Laboratory of Cancer Research, Beijing) for providing the original chromosome classification; Dr. Jim Piper, Dr. Ling Li, Kai Yu, and Haixin Ke for giving us valuable advice; and Jiang Ni for providing the original profiles of CGH images.

LITERATURE CITED

1. Carothers A, Piper J. Computer-aided classification of human chromosomes: a review. Stat Comput 1994;4:161-171.
2. Piper J, Granum E. On fully automatic feature measurement for banded chromosome classification. Cytometry 1989;10:242-255.
3. Lerner B, et al. Feature selection and chromosome classification using a multilayer perceptron network. Proc IEEE Int Conf Neural Networks 1994;6:3540-3545.
4. Ruan X. A classifier with the fuzzy hopfield network for human chromosomes. In Proceedings of the 3rd World Congress on Intelligent Control and Automation 2000;2:1159-1164.
5. Turner M, et al. Chromosome location and feature extraction using neural networks. Image Vision Comput 1993;11:235-239.
6. Wu Q, Castleman KR. Automated chromosome classification using wavelet-based band pattern descriptors. In Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems (CBMS 2000), 2000:189-194.
7. Kallioniemi A, Kallioniemi OP, Sudar D, et al. Comparative genomic hybridization for molecular analysis of solid tumors. Science 1992; 258:818-821.
8. du Manoir S, et al. Quantitative analysis of comparative genomic hybridization. Cytometry 1995;19:27-41.
9. Kirchhoff M, Gerdes T, Maahr J, et al. Automatic correction of the interfering effect of unsuppressed interspersed repetitive sequences in comparative genomic hybridization analysis. Cytometry 1997;28: 130-134.
10. Lundsteen C, Maahr J, et al. Image analysis of comparative genomic hybridization. Cytometry 1995;19:42-50.
11. Piper J, Rutovitz D, Sudar D, et al. Computer image analysis of comparative genomic hybridization. Cytometry 1995;19:10-26.
12. Schrock E, du Manoir S, et al. Multicolor spectral karyotyping of human chromosomes. Science 1996;273:494-497.
13. Speicher MR, et al. Karyotyping human chromosomes by combinatorial multifluor FISH. Nature Genet 1996;12:368-375.
14. Errington PA, Graham J. Application of artificial neural networks to chromosome classification. Cytometry 1993;14:627-639.
15. Cho JM. Chromosome classification using backpropagation neural networks: a process that overcomes nonlinearity problems to correctly classify Giemsa-stained human chromosomes. IEEE Eng Med Biol 2000;19:28-33.
16. Lerner B. Toward a completely automatic neural network based human chromosome analysis. IEEE Trans Systems Man Cybernet 1998;28:544-552.
17. Musavi MT, et al. Mouse chromosome classification by radial basis function network with fast orthogonal search. Neural Networks 1998;11:769-777.
18. Rutovitz D, Green DK, et al. Computer-assisted measurement in the cytogenetic laboratory. Pattern Recogn 1978;10:303-329.
19. Vapnik VN. Statistical learning theory. New York: John Wiley & Sons; 1998. 736 p.
20. Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag; 1995. 188 p.
21. LeCun Y, Jachel L, et al. Comparison of learning algorithms for handwritten digit recognition. In Proceedings of the International Conference on Artificial Neural Networks, 1995. p 53-60.
22. Edgar EO, et al. Support vector machines: training and applications. MIT AI Memo No.1602, 1997. p 28-30.
23. Ji L. Fully automatic chromosome segmentation. Cytometry 1994;17: 196-208.
24. Hilditch CJ. Linear skeletons from square cupboards. In: Melzer B, Michie D, editors. Machine Intelligence 4. Edinburgh: Edinburgh University Press; 1969. p 403-420.
25. Collobert R, Bengio S. Support vector machines for large-scale regression problems. 2000;IDIAP-RR-17.
26. Zhang X. Introduction to statistical learning theory and support vector machines. Acta Automatica Sinica (in Chinese) 2000;26:32-42.
27. Krebel U. Pairwise classification and support vector machines. In: Schlkopf B, Burges CJC, Smola AJ, editors. Advances in kernel methods: support vector learning. Cambridge, MA: MIT Press; 1999. p 255-268.