

# Karyotyping of Comparative Genomic Hybridization Human Metaphases Using Kernel Nearest-Neighbor Algorithm

Kai Yu and Liang Ji\*

State Key Laboratory of Intelligent Technology and Systems, Department of Automation, Tsinghua University, Beijing, People's Republic of China

Received 28 November 2001; Revision Received 4 May 2002; Accepted 3 June 2002.

**Background:** Comparative genomic hybridization (CGH) is a relatively new molecular cytogenetic method that detects chromosomal imbalances. Automatic karyotyping is an important step in CGH analysis because the precise position of the chromosome abnormality must be located and manual karyotyping is tedious and time-consuming. In the past, computer-aided karyotyping was done by using the 4',6-diamidino-2-phenylindole, dihydrochloride (DAPI)-inverse images, which required complex image enhancement procedures.

**Methods:** An innovative method, kernel nearest-neighbor (K-NN) algorithm, is proposed to accomplish automatic karyotyping. The algorithm is an application of the "kernel approach," which offers an alternative solution to linear learning machines by mapping data into a high dimensional feature space. By implicitly calculating Euclidean or Mahalanobis distance in a high dimensional image feature space, two kinds of K-NN algorithms are obtained. New

feature extraction methods concerning multicolor information in CGH images are used for the first time.

**Results:** Experiment results show that the feature extraction method of using multicolor information in CGH images improves greatly the classification success rate. A high success rate of about 91.5% has been achieved, which shows that the K-NN classifier efficiently accomplishes automatic chromosome classification from relatively few samples.

**Conclusions:** The feature extraction method proposed here and K-NN classifiers offer a promising computerized intelligent system for automatic karyotyping of CGH human chromosomes. *Cytometry* 48:202–208, 2002.

© 2002 Wiley-Liss, Inc.

**Key terms:** comparative genomic hybridization; chromosome classification; karyotyping; kernel; nearest-neighbor

Comparative genomic hybridization (CGH) is a relatively new molecular cytogenetic method that detects chromosomal imbalances (1). In CGH, two genomic DNA samples are hybridized simultaneously in situ to normal human metaphase spreads and detected with different fluorochromes. The intensity ratio of the two fluorescence signals is a measure of the copy number ratio between the two genomic DNA samples. CGH has great potential for a broad range of applications and has made a great impact on the analysis of tumor chromosomes.

Karyotyping assigns each chromosome to one of 23 or 24 classes (22 autosomes and a pair of sex chromosomes). This is an important step in CGH analysis because the precise position of the chromosome abnormality must be located before the quantity of the intensity ratio imbalance can be calculated. However, manual karyotyping is tedious and time-consuming. Great efforts have been made to develop computer-aided chromosome classifiers. Two aspects of karyotyping are crucial. One is the design of the classifier and the other is the set of features used in the classification.

Conventionally, biologists do manual CGH chromosome classification by checking the color metaphase images. However, computer-aided karyotyping of CGH metaphases uses only the 4',6-diamidino-2-phenylindole, dihydrochloride (DAPI)-inverse images, which are similar to G-banded images. In this way, all sophisticated methods for G-banded chromosome classification can be directly used for DAPI images. Computer-aided karyotyping treats DAPI-inverse images as G-banded images (2–5). Therefore, many of the classification methods used in Giemsa-stained chromosome images are applicable. A thorough review of chromosome classification is provided by Carothers and

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

\*Correspondence to: Liang Ji, Institute of Information Processing, Department of Automation, Tsinghua University, Beijing (100084), Peoples Republic of China.

E-mail: zc-sa@mail.tsinghua.edu.cn or yukai99@mails.tsinghua.edu.cn

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/cyto.10130

Piper (6). In general, a human chromosome is characterized by its size, centromere position, and banding pattern (7). Other methods for chromosome feature extraction include the “knock-out” algorithm (8), principal components analysis (PCA; 8,9), Kohonen network (10), and wavelet packets (11).

G-banded images have clear banding patterns that appear as bright and dark bands on the chromosomes. The banding patterns of DAPI-inverse images are similar to G-banded images, but they are far less clear than the G-banded patterns. As a result, automatic chromosome identification is difficult in CGH analysis. This may also explain why the success rate of CGH karyotyping has rarely been reported. Image enhancement has to be used to make the banding patterns clearer. The highest success rate achieved by such a method so far is about 90% by Vysys (unpublished data). When only DAPI-inverse images are used for karyotyping, useful information for identifying the chromosomes is lost. For example, the multicolor information provided by CGH images is used by cytogeneticists to locate the centromeres of the chromosomes, but this information cannot be revealed completely by DAPI image alone. Only a few studies have been published on multicolor karyotyping (12,13). For example, spectral karyotyping (SKY; 12) and multicolor fluorescence in situ hybridization (M-FISH; 13) are not suitable for CGH analysis. Four-color CGH (14) is concerned with quality control, not karyotyping. To our knowledge, the only attempt to utilize color information for karyotyping CGH images was made by Kou et al. (15). However, we use a different approach in this study.

In this study, we use multicolor information of CGH chromosomes for chromosome feature extraction and multistep classification. A novel intuitive learning method, kernel nearest-neighbor (K-NN) algorithm, is applied to automatic chromosome classification. The “kernel approach” has received much attention from researchers in the machine learning field because of the rapid development of the support vector machine (SVM). The kernel approach can be studied in a general way and extended to different learning systems. In this study, the conventional nearest-neighbor algorithm is extended using the kernel approach. Simple substitution of kernel distance for norm distance leads to the common K-NN algorithm. Considering sample distribution and using the kernel PCA (KPCA) technique, the K-NN algorithm with kernel Mahalanobis distance (KMD) is achieved ultimately. The two K-NN algorithms are new applications of the kernel approach and have not been used for solving the pattern analysis problems of the real world. A high success rate (91.5%) was obtained by applying a combination of the two K-NN algorithms. The classification result suggests that the K-NN algorithms can be used efficiently for real-world pattern analysis problems, especially nonlinear small-size problems.

## MATERIALS AND METHODS

### CGH and Image Acquisition

All CGH images were provided by Dr. Tommy Gerdes (University of Copenhagen, Copenhagen, Denmark). The

images are of the size  $748 \times 573$  and are in TIFF format, containing three-color channels, each with  $256^2$  resolution. The red, green, and blue channels represent Texas Red, fluorescein isothiocyanate, and DAPI, respectively. Details about the CGH slide preparation and image acquisition were described by Kirchhoff et al. (3). There are 71 metaphase images with 23 classes (22 autosomes and a pair of X chromosomes) because the target metaphase was obtained from a karyotypically normal female (3). Some metaphase images are incomplete, containing less than 46 chromosomes.

The original classification of the chromosomes was determined by Dr. Mingrong Wang and Ms. Xin Xu, both of whom are biologists at the Institute of Cancer Research, the Chinese Academy of Medical Science (Beijing, China).

### Digital Image Preprocessing

Preprocessing of the CGH images was accomplished using the software developed by Dr. Liang Ji (the second author) and Ms Jiang Ni (Dept. of Automation, Tsinghua University, Beijing, China), named “CGH analyzer” implemented on an IBM-compatible PC, manufactured by Legend Grove, Beijing, China, with the AMD Athlon 500-MHz processor and 128 MB RAM. Critical steps involved in karyotyping include the following: (1) the RGB image is transformed into a gray image by calculating the average of the three components; (2) the chromosomes are segmented according to the method described by Ji (16); (3) the segmented chromosomes are used as masks to select chromosomes from the original color images; (4) the medial axis of each segmented chromosome is determined by using the Hilditch skeleton (17); and finally, the red and blue signals are measured along each normal of the medial axis to obtain integrated density profiles.

### Chromosome Feature Extraction and Normalization

Generally, the chromosomes in a metaphase are divided into seven “Denver” groups, labeled A to G, and a pair of sex chromosomes. Ten features extracted from the shape, banding pattern, intensity, and color information are used for chromosome classification. The extracted features were selected based on information provided by human experts and by the manual classification process.

Two shape-related features are length and centromere index. Length is obtained easily by measuring the length of the medial axis of each chromosome. Because different metaphases may have different scales, they should be normalized to make the lengths in different metaphases comparable. The standardized length  $Len$  is defined by the following formula,

$$Len = L/\bar{L}$$

where  $L$  is the length of the chromosome and  $\bar{L}$  is the average length of the metaphase.

Conventionally, the second-order moment along the medial axis of the chromosome in the DAPI-inverse image

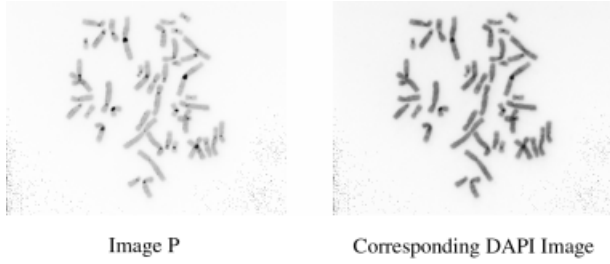


FIG. 1. Comparison of image P and DAPI image.

is measured and the position of the smallest extreme is obtained as the centromere position (7). However, sometimes this does not work well. In contrast, the centromeres in a CGH metaphase are recognized easily due to the block of Cot-1 DNA used in the hybridization. Very low red and green signals appear at the centromere positions and the pericentromeric areas appear blue. That is, a linear transformation takes place among the red, green, and blue images:

$$P = B - (R + G)/2$$

where R, G, and B represent the red, green, and blue signals, respectively, and P is the result of the transformation. The image P is important for extraction of other features. Figure 1 shows an example of image P and the corresponding DAPI image for comparison.

In the image P, along the medial axis of each chromosome, the mean density profile is calculated and the position of the maximum is recognized as the centromere position. Occasionally, the P values in the whole chromosome are zero or negative. In this case, the centromere position will be found by using conventional method. The centromere separates a chromosome into two arms. The long and short arms are called q and p, respectively. Let  $L_p$  and  $L_q$  be the lengths of the p and q arms, respectively. The centromere index *Centro* is defined as:

$$Centro = \frac{L_p}{L}$$

There are seven banding pattern-related features extracted as below. One is the mass of the pericentromeric area, *Peri*. Mass is the sum of total intensity values in a specific region of an image. The pericentromeric region is defined as the region around the centromere in the P image, where the mean density profile decreases from the maximum to 80% of the value. The standardized result is the mass value divided by the median mass value of the metaphase, denoted by *Peri*.

Four features represent the rough information of the banding pattern. Let  $\hat{d}_r$  and  $\hat{d}_b$  be the integrated density profiles of red and blue signals, respectively. The ratio  $\hat{d} = \hat{d}_r/\hat{d}_b$  is then determined, and  $\hat{d}_{med}$  represents the median value in the  $\hat{d}$  profiles. Let  $n_p$ ,  $n_q$  represent the number of

points whose values are greater than  $\hat{d}_{med}$  in the p and q arms of the  $\hat{d}$  profile, respectively, and let  $pos_p$ ,  $pos_q$  represent the center of gravity of the  $n_p$  and  $n_q$  points, respectively. Then, four features are defined as following:

$$fqr = \frac{n_q}{L_q} \quad fpr = \frac{n_p}{L_p} \quad fpc = \frac{pos_p}{L_p} \quad fqc = \frac{pos_q}{L_q}$$

Another band-related feature is *fD*, which described the intensity balance of the q arm in the DAPI image. The q arm of a chromosome in the DAPI image is divided into two halves lengthwise. The median intensity values of each half is calculated as  $Top_{med}$  and  $Bottom_{med}$ . The average gray value  $gray_b$  of the chromosome in the DAPI image is used for standardization. *fD* is defined as:

$$fD = \frac{Top_{med} - Bottom_{med}}{gray_b}$$

Similar to *fD*, *mpq* represents the intensity difference between the p and q arms in the DAPI image. It is calculated as:

$$mpq = \frac{|T_p - T_q|}{gray_b}$$

where  $T_p$  and  $T_q$  are the average intensity values of the p and q arms in  $\hat{d}_b$ , respectively.

A color-related feature, *r2b*, the ratio of the red and the blue signal, is also used. Let  $mass_r$  and  $mass_g$  be the masses of the chromosome in the red and blue images and let  $gray_r$ ,  $gray_b$  be the average gray values of the chromosome in the red and blue images respectively. The standardized feature is denoted as:

$$r2b = \frac{mass_r \cdot gray_b}{mass_b \cdot gray_c}$$

### Multistep Classification Process

Two-step classification is used in automatic karyotyping to imitate cytogeneticists. The chromosomes are divided into 10 groups, which are A-1 (1#), A-2 (2#), A-3 (3#), B

Table 1  
Feature Combinations

Group	Feature	Sequence no. of classified chromosomes
First step	Len, Centro, Peri	A-1(1#), A-2(2#), A-3(3#), E-16(16#), B, C+X, D, E (17,18), F, G
B	fqr	(4#, 5#)
C+X	Len, Centro, Peri, fpr, fqc, fpc	(6#~12#, 23#)
D	Peri, fD, r2b	(13#~15#)
E	Len, Centro, Peri, r2b	(17#, 18#)
F	Peri, mpq, r2b	(19#, 20#)
G	Len, Centro, Peri, r2b	(21#, 22#)

(4#, 5#), C+X (6#-12#, 23#), D (13#-15#), E-16 (16#), E (17,18)(17#, 18#), F (19#, 20#), G (21#, 22#), and then they are assigned to a specific class. The number (or numbers) in parenthesis denotes the class to which a chromosome belongs. Because A-1, A-2, A-3, and E-16 are already specific classes, only six groups remain to be classified in the second step. The feature combinations are shown in Table 1.

**K-NN Algorithm for Classification**

**Kernel approach.** People often use a “generalized linear classifier” to enhance the common linear classifier. As Figure 2 shows, it is impossible to distinguish the two sets using a linear classifier in the original feature space. By using certain nonlinear feature mapping, the two sets are linearly classifiable in a high dimensional image space. However, in many cases, the dimension of the image feature space is much larger than that of the original space and is too large to perform operations, such as the inner product.

The kernel approach offers a solution to calculate the inner product in a high dimensional image feature space without increasing the number of parameters. It provides one of the main blocks of SVM and has peaked the interest of researchers in the machine learning field (18). Besides linear learning machines, the approach has been extended to other learning systems, such as KPCA (19,20), and has shown much promise. Consider a case of mapping an  $n$ -dimension feature space to an  $m$ -dimension image feature space:

$$\begin{aligned} \vec{x} &= (x_1, \dots, x_n) \xrightarrow{\text{feature mapping}} \psi(\vec{x}) \\ &= (\phi_1(\vec{x}), \dots, \phi_m(\vec{x})), \vec{x} \in S_1, \psi(\vec{x}) \in S_2 \end{aligned}$$

where  $S_1$  is the original  $n$ -dimension feature space and  $S_2$  is the new  $m$ -dimension image feature space.  $\vec{x}$  is an arbitrary vector in  $S_1$ ,  $\psi(\vec{x})$  is the corresponding image vector in  $S_2$ .  $\psi$  is an arbitrary nonlinear mapping function from the original space to a possibly high dimensional space  $S_2$  and  $\phi_i, i = 1. . m$  are feature mapping functions.

A kernel denotes a function  $K$ , such that for all  $\vec{x}, \vec{y} \in S_1$

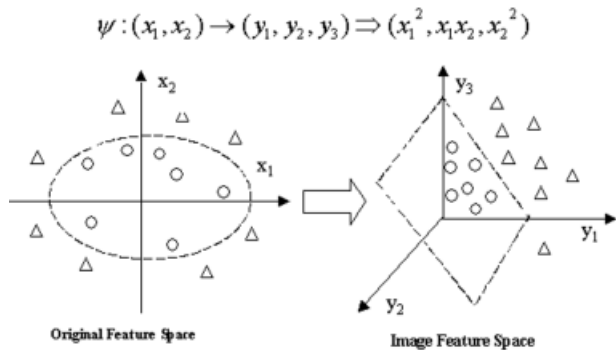


FIG. 2. From common linear classifier to generalized linear classifier.

$$K(\vec{x}, \vec{y}) = (\psi(\vec{x}), \psi(\vec{y}))$$

where  $\langle \vec{x}, \vec{y} \rangle$  denotes the inner product of  $\vec{x}$  and  $\vec{y}$ .

The definition of the kernel implies that the inner product in the new image feature space can be computed without actually carrying out the mapping  $\psi$ . A specific choice of kernel function might then correspond to an inner product of samples mapped by a suitable nonlinear function  $\psi$ . According to the Hilbert-Schmidt theory,  $K(\vec{x}, \vec{y})$  can be an arbitrary symmetric function that satisfies the Mercer condition (18). Three kernel functions are commonly used (18): polynomial kernel:  $K(\vec{x}, \vec{y}) = (1 + \langle \vec{x}, \vec{y} \rangle)^p$ ; radial basis function (RBF) kernel:  $K(\vec{x}, \vec{y}) = \exp\{-\|\vec{x} - \vec{y}\|^2/\sigma^2\}$ ; and sigmoid kernel:  $K(\vec{x}, \vec{y}) = \tanh(\alpha \langle \vec{x}, \vec{y} \rangle + \beta)$ .  $p, \sigma, \alpha$ , and  $\beta$  are adjustable parameters of the above kernel functions. For a sigmoid kernel, only partial parameters are available (18).

**K-NN algorithm.** The NN algorithm is an intuitive classification algorithm (21). It assigns a test sample to the class to which the nearest training sample to it belongs. A norm distance metric, such as Euclidean distance (ED), is used to determine which training sample is the nearest to the test sample. Suppose that the NN algorithm is implemented in a high dimensional image feature space, the norm distance in such a space should be computed, i.e., a feature mapping is applied as above. The square of the norm distance in the image feature space can be decomposed into the sum of several inner products and can be obtained by applying the kernel approach (22).

When the kernel norm distance is computed as above and the NN algorithm is applied in the image feature space, a K-NN classifier is obtained. Extensions of the conventional NN algorithm, such as a multineighbor NN, edited NN, can find corresponding versions of the K-NN algorithm (22).

**KMD and the K-NN algorithm.** In distance-based learning systems, besides the ED, the Mahalanobis distance (MD) is commonly used (23). The MD takes into account the distribution of samples. Suppose there are some samples  $y_i, i = 1. . M$  (column vectors) in an  $n$ -dimensional feature space,  $\bar{\mu}$  is the center vector of samples (mean value), the centered samples  $\tilde{x}_i, i = 1. . M$  is denoted as:

$$\bar{\mu} = \frac{1}{M} \sum_{i=1}^M \tilde{y}_i \quad \tilde{x}_i = \tilde{y}_i - \bar{\mu}$$

The ED and MD between the two samples can be calculated as follows:

$$ED^2(\tilde{y}_i, \tilde{y}_j) = (\tilde{y}_i - \tilde{y}_j)^T (\tilde{y}_i - \tilde{y}_j) = (\tilde{x}_i - \tilde{x}_j)^T (\tilde{x}_i - \tilde{x}_j)$$

$$\begin{aligned} MD^2(\tilde{y}_i, \tilde{y}_j) &= (\tilde{y}_i - \tilde{y}_j)^T C_x^{-1} (\tilde{y}_i - \tilde{y}_j) \\ &= (\tilde{x}_i - \tilde{x}_j)^T C_x^{-1} (\tilde{x}_i - \tilde{x}_j) \end{aligned}$$

where

$$C_x = \frac{1}{M} \sum_{i=1}^M \tilde{x}_i \tilde{x}_i^T$$

is the covariance matrix of centered vectors. It describes the location and dispersion of the population.

By definition,  $C_x$  is a real symmetric and positive semidefinite matrix. Considering that  $C_x$  is a positive definite, a new form of the MD distance can be obtained based on specific data set. Equation 1 is an orthogonal decomposition of  $C_x$  and Eq. 2 is the new form of MD.

$$C_x = V\Lambda V^T \quad (1)$$

$$\begin{aligned} MD^2(\tilde{x}, \tilde{y}) &= (\tilde{x} - \tilde{y})^T V\Lambda^{-1}V^T(\tilde{x} - \tilde{y}) \\ &= (\tilde{x}^V - \tilde{y}^V)^T \Lambda^{-1}(\tilde{x}^V - \tilde{y}^V) \\ &= (\sqrt{\Lambda^{-1}}(\tilde{x}^V - \tilde{y}^V))^T (\sqrt{\Lambda^{-1}}(\tilde{x}^V - \tilde{y}^V)) \end{aligned} \quad (2)$$

where  $\tilde{x}, \tilde{y}$  are arbitrary  $n$ -dimensional vectors in feature space and  $\tilde{x}^V, \tilde{y}^V$  are the corresponding projections onto the eigenvector matrix  $V = [\tilde{v}_1, \dots, \tilde{v}_n]$ . The  $i$ th element of  $\tilde{x}^V$  is the corresponding projection component onto  $\tilde{v}_i$ ,  $\sqrt{\Lambda^{-1}}$  is the diagonal weighting matrix with rank  $n$ . The  $i$ th diagonal element is  $1/\sqrt{\lambda_i}$  and  $\lambda_i, i = 1 \dots n$  are eigenvalues of  $C_x$ .

Equation 2 associates the MD with PCA. PCA is a powerful technique for extracting structure. It is an orthogonal transformation of the coordinate system in which the data set is described. The projection values in the new coordinate system are called principal components (PC). The orthogonal eigenvectors that span the PC space are called PC axes. PCA describes some intrinsic geometric properties of the specific data set. If each coordinate axis of the PC space is weighted by the reciprocal of the square root of the variance, the PC space is normalized. In the normalized PC space, PCs are orthogonal and all have unit variance. Equation 2 shows that the MD in the original feature space is equivalent to the ED in the normalized PC space. We can select some of those "useful" PC axes by eliminating PC axes that have zero or nearly zero eigenvalues and we can calculate the ED in the selected normalized PC space. This means that matrix  $V$  (Eq. 2) does not need to be square and the rank of  $\sqrt{\Lambda^{-1}}$  does not need to be  $n$ . The above calculation procedures can be formulated in a way that inner products are exclusively used (24), which implies that the kernel approach can be used here.

KPCA is a kernel version of linear PCA (20). It is an orthogonal transformation of the coordinate system in a potentially high dimensional image feature space using the kernel approach. KPCA does not concern PCs in the original feature space, but PCs in the image feature space that are nonlinearly related to the input samples. If the MD is calculated in an appropriate nonlinear image space, the high-order correlations of a specific data set may be eliminated. The K-NN algorithm with such a distance metric

might obtain better results. As indicated above, the MD is equivalent to the ED in a selected normalized linear PC space. By applying KPCA, the KMD can be introduced (19). In summary, the KMD can be calculated by selecting the kernel function and calculating the centered kernel matrix  $K$  (20); diagonalizing  $K$  and obtaining all the eigenvalues and eigenvectors; selecting some eigenvectors as orthogonal basis vectors of a new KPC space according to a certain criterion; computing all KPC of each sample, i.e., projections onto the selected eigenvectors, and normalizing all KPC, i.e., divide KPC by the square root of corresponding eigenvalues; and calculating the ED between projections of different samples in the selected normalized KPC space.

The K-NN with KMD provides a novel kernel-based learning system. By performing KPCA and normalizing the selected KPC space, the distribution of samples of each class might be reshaped to an approximate unit hypersphere, which makes K-NN more applicable and powerful especially in classification problems with samples highly overlapping in the feature space.

In Step 3, the criterion for selection of eigenvalues after performing KPCA should be set to get satisfactory results. In this study, the min-max-ratio (MMR) is used as an eigenvalue selection criterion. A predetermined ratio is set, which is the smallest eigenvalue in the selected set divided by the largest eigenvalue in the whole set. For example,

$$MMR = \frac{\min_{j \in SelectedSet} (\lambda_j)}{\max_{i \in WholeSet} (\lambda_i)}$$

Any eigenvectors will be eliminated if the ratio of the corresponding eigenvalue to the largest one is smaller than the MMR.

## RESULTS

We used 71 metaphases. The training set comprised 51 metaphases selected randomly and the test set comprised the remaining metaphases. Due to the absence of chromosomes, some metaphases comprised 45 chromosomes or less. Therefore, there were 2,273 samples in the training set and 893 samples in the test set. In addition, there were 23 classes.

We used a combination of common K-NN and K-NN with KMD to accomplish classification. Parameter selection is an important step in the classification procedure of K-NN algorithms. The following K-NN algorithms and parameters were used for experimenting.

First step

Common K-NN, polynomial kernel,  $p = 10$ , three neighbors

Second step

B: K-NN with KMD, RBF kernel,  $\sigma = 1$ ,  $MMR = 10^{-6}$ , nine neighbors

C+X: common K-NN, polynomial kernel,  $p = 2$ , seven neighbors

Table 2  
Success Rate of the First Step

Training data	Test data	Error data	Success rate (%)
2,273	893	24	97.31

D: K-NN with KMD, RBF kernel,  $\sigma = 0.5$ , MMR =  $10^{-7}$ , five neighbors

E (17,18): common K-NN, polynomial kernel,  $p = 1$ , one neighbor

F: K-NN with KMD, polynomial kernel,  $p = 1$ , MMR =  $10^{-3}$ , 15 neighbors

G: K-NN with KMD, polynomial kernel,  $p = 1$ , MMR =  $10^{-3}$ , 25 neighbors

The above algorithms and parameters were experimentally determined. Because the K-NN algorithms will reshape the distribution of the samples, they are sensitive to the training samples. The parameters, including the number of neighbors, are very dependent on the training set. Considering different features have different data range, all data were normalized to [0,1] before classification.

The success rate of the first step is listed in Table 2. Table 3 shows the original success rate of the second step. The experiment in Table 3 was done separately to evaluate algorithms used for different groups in the second step. In this experiment, some types of data (B, C+X, D, E [17,18], F, G) were selected from the whole data set to form six new pairs of data sets. For example, to test the performance of the algorithm and the parameters used for group B, the chromosomes of 4# and 5# were selected from the original training and test sets to form new training and test sets. The new sets contained only 4# and 5# and the features were selected as shown in Table 1. The success rates of the six groups after the multistep classification are shown in Table 4, i.e., the algorithms described above were applied to the resulting groups of the first step. The success rate of the first step is shown in Table 2. Table 5 includes the final overall success rate of all experiments. The overall success rate using the conventional NN algorithm is also listed in Table 5 for comparison.

Figure 3 shows a representative example of chromosomes that are assigned incorrectly by the algorithm. The example shows that the pattern of overlapping chromosomes was not restored perfectly, which may lead to

Table 3  
Original Success Rates of the Second Step\*

Group	Training data	Test data	Error data	Success rate (%)
B	198	79	5	93.67
C+X	789	310	38	87.74
D	299	117	6	94.87
E (17,18)	199	78	0	100
F	194	77	6	92.21
G	194	76	1	98.68

\*Use the original data.

Table 4  
Actual Success Rates of the Second Group\*

Group	Total no.	Error data	Success rate (%)
B	80	7	91.25
C+X	310	40	87.10
D	114	8	92.98
E (17,18)	80	2	97.50
F	82	14	82.93
G	71	2	97.18

\*Use the results of the first step.

unsuccessful classification. Most incorrect samples are chromosomes overlapping in the original color images.

## DISCUSSION AND CONCLUSIONS

This study proposes a straightforward and effective feature extraction method for CGH karyotyping. Different from the chromosome classification approaches that use enhanced DAPI-inverse images for classification (2,5), the method presented here makes full use of the R, G, and B signals in the original CGH images to determine the centromere position and to generate integrated density profiles. Simple and robust features are then extracted.

A novel kernel-based algorithm, the K-NN algorithm, was also employed to accomplish the classification. The K-NN algorithm is a new application of the kernel approach and has two forms. By substituting the kernel distance for the norm distance and introducing KPCA, common K-NN and K-NN with KMD can be achieved, respectively. The K-NN algorithm is an extension of the conventional NN algorithms. It inherits the good ability of generalization and is especially suitable for small-size classification. Because the K-NN algorithm has many variations, such as edited and multineighbor K-NN, and the kernel approach depends on the selection of kernel function as well as the parameters, the K-NN algorithms should be examined thoroughly for a specific problem. The solution still depends on experimentation. There has been no ideal theoretical guidance yet. Once a suitable combination of K-NN algorithms and kernel parameters is set according to a specific training data set, this classifier will achieve satisfactory success rates and good ability of generalization.

In this CGH karyotyping problem, we used a multistep classification strategy. A combination of the common K-NN algorithm and the K-NN with KMD algorithm was used for classification. We selected suitable variations of the K-NN algorithms and parameters of kernel function according to distribution of the samples in the training set

Table 5  
Overall Success Rate After Two Steps

Algorithm	Training data	Test data	Error data	Success rate (%)
K-NN	2,273	893	76	91.49
Conventional NN	2,273	893	138	84.55

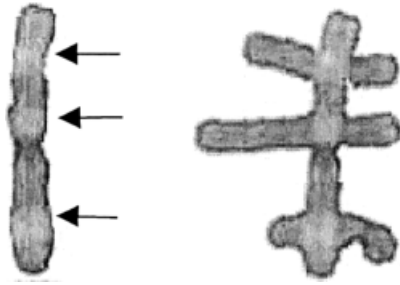


FIG. 3. Representative example of incorrectly assigned chromosomes. Arrows indicate the incompletely restored pattern, which leads to incorrect classification (left). The chromosome in the original CGH image, which overlaps with other chromosomes (right).

by experimenting. Finally, a high success rate of 91.5% was obtained.

However, there is a room for improvement in chromosome classification. Because the selection of the kernel parameter and the K-NN variations lack sufficient theoretical guidance, better results may be found after more experiments. The normal human karyotype consists of 22 pairs of autosomes and a pair of sex chromosomes (6). A simple context-sensitive classifier based on the number constraint of the homolog chromosomes in a metaphase may improve the success rate (25). A lot of overlapping chromosomes are found in the metaphases. The patterns in the overlapping regions cannot be restored completely by the segmentation method (16). We should improve the segmentation method rather than the classifier alone. In addition, further image enhancement of CGH images before feature extraction may improve the success rate.

#### ACKNOWLEDGMENTS

We thank Dr. Tommy Gerdes (University of Copenhagen) for supplying the CGH images, Dr. Mingrong Wang and Ms. Xin Xu (the Institute of Cancer Research, National Key Laboratory of Cancer Research, Chinese Academy of Medical Science, Beijing) for providing the original chromosome classification, Jiang Ni for performing the image preprocessing and feature extraction, and Zhenzhen Kou for giving valuable advice.

#### LITERATURE CITED

1. Kallioniemi A, Kallioniemi OP, Sudar D, et al. Comparative genomic hybridization for molecular analysis of solid tumors. *Science* 1992; 258:818-821.
2. Du Manoir S, et al. Quantitative analysis of comparative genomic hybridization. *Cytometry* 1995;19:27-41.
3. Kirchoff 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.
4. Lundsteen C, Maahr J, et al. Image analysis of comparative genomic hybridization. *Cytometry* 1995;19: 42-50.
5. Piper J, Rutovitz D, Sudar D, et al. Computer image analysis of comparative genomic hybridization. *Cytometry* 1995;19:10-26.
6. Carothers A, Piper J. Computer-aided classification of human chromosomes: a review. *Stat Comput* 1994;4:161-171.
7. Piper J, Granum E. On fully automatic feature measurement for banded chromosome classification. *Cytometry* 1989;10:242-255.
8. Lerner B, et al. Feature selection and chromosome classification using a multilayer perception network. *IEEE international conference on neural networks, 1994. IEEE World Congress on Computational Intelligence*; 6:3540-3545.
9. Ruan X. A classifier with the fuzzy hopfield network for human chromosomes. *The 3rd world congress on intelligent control and automation. 2000. p 1159-1164.*
10. Turner M, et al. Chromosome location and feature extraction using neural networks. *Image Vision Comput* 1993;11:235-239.
11. Wu Q, Castleman KR. Automated chromosome classification using wavelet-based band pattern descriptors. *13th IEEE symposium on computer-based medical systems (CBMS 2000). 2000. p 189-194.*
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 multi-fluor FISH. *Nature Genet* 1996;12:368-375.
14. Karhu R, Rummukainen J, Lörch T, Isola J. Four-color CGH: a new method for quality control of comparative genomic hybridization. *Genes Chromosomes Cancer* 1999;24:112-118.
15. Kou ZZ, Ji L, Zhang XG. Karyotyping of CGH human metaphases by using support vector machines. *Cytometry* 2002;47:17-23.
16. Ji L. Fully automatic chromosome segmentation. *Cytometry* 1994;17: 196-208.
17. Hilditch CJ. Linear skeletons from square cupboards. In: Melzer B, Michie D, editors. *Machine intelligence 4*. Edinburgh: Edinburgh University Press; 1969. p 403-420.
18. Vapnik VN. *The nature of statistical learning theory*. New York: Springer-Verlag; 1995. 188 p.
19. Ruiz A, López-de-Teruel PE. Nonlinear kernel-based statistical pattern analysis. *IEEE Trans Neural Network* 2001;12:16-32.
20. Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 1998;10:1299-1319.
21. Duda RO, Hart PE. *Pattern classification and scene analysis*. New York: Wiley; 1973. 482 p.
22. Yu K, Ji L, Zhang XG. Kernel nearest-neighbor algorithm. *Neural Processing Lett* 2002;15:147-156.
23. Maesschalck RD, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. *Chemometrics Intel Lab Syst* 2000;50:1-18.
24. Kirby M, Sirovich L. Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE Trans Pattern Anal Machine Intel* 1990;12:103-108.
25. Piper J. Classification of chromosomes constrained by expected class size. *Pattern Recog Lett* 1986;4:391-395.