

# IMAGE ANALYSIS FOR HIGH THROUGHPUT GENOMICS

*Suchendra M. Bhandarkar, Tongzhang Jiang, Kunal Verma and Nan Li*

Department of Computer Science, The University of Georgia  
Athens, Georgia 30602-7404, U.S.A.

## ABSTRACT

The design and implementation of a computer vision system called DNAScan for the automated analysis of DNA hybridization images is presented. The hybridization of a DNA clone with a radioactively tagged probe manifests itself as a spot on the hybridization membrane. A recursive segmentation procedure is designed and implemented to extract spot-like features in the hybridization images in the presence of a highly inhomogeneous background. Positive hybridization signals (hits) are extracted from the spot-like features using grouping and decomposition algorithms based on computational geometry. A mathematical model for the positive hybridization patterns and a pattern classifier based on shape-based moments are proposed and implemented to distinguish between the clone-probe hybridization signals.

## 1. INTRODUCTION

Generation and subsequent analysis of clone-probe hybridization data is the first step in high-throughput genomics experiments. With the increasing use of robotics for rapid generation of hybridization data, it has become imperative to be able to analyze the resulting data in an expeditious and reliable manner. The paucity of automated techniques for rapid and reliable analysis of hybridization data has proved to be a bottleneck in most high-throughput genomics experiments [2]. To this end, we present the design and implementation of a computer vision system called DNAScan for the automated analysis of DNA hybridization images.

In typical hybridization experiments, pieces of a single-stranded chromosome called *clones* are fixed at predetermined positions on a nylon membrane. The membrane is then exposed to radioactively tagged *probes*. A probe is a short distinguishable DNA fragment whose sequence is known. A probe attaches or *hybridizes* to a clone if there exists a site on the clone with a DNA sequence that is complementary to the DNA sequence of the probe. The clone-probe hybridization can be determined by the presence of radioactivity at the clone site on the membrane. After the membrane is washed to remove the excess probe residue, it is exposed to a film. The radioactively tagged probe leaves a signal in the form of a bright spot at the same location on the film as the clone site on the membrane. The film is scanned and converted to a digital format for further analysis.

Automated analysis of the hybridization signals on the film requires knowledge of the positions of the clones on

the membrane. The term *hybridization protocol* refers to the spatial arrangement of clones on the nylon membrane. In our case, a single film contains 2304 squares arranged in the form of a  $48 \times 48$  two-dimensional array. Each square contains  $4 \times 4 = 16$  cells where each cell contains a single clone. Each clone is spotted in duplicate within a  $4 \times 4$  square. A clone-probe hybridization thus results in two spots on the film corresponding to the positions of the two cells containing that clone. These two spots are referred to as a *positive hit*. A positive hit is characterized by the distance between the two spots and the orientation of the straight line joining the two spots. A positive hit is deemed to belong to a *pattern class* that is characterized by the distance and orientation measurements mentioned above. Determining the spatial locations of the clones in the  $4 \times 4$  array such that each pattern class can be distinguished from other pattern classes is the key issue in hybridization protocol design.

Depending on the number of probes that a given membrane is exposed to, the number of hybridizations in a  $4 \times 4$  square varies from 0 to  $n$ . The case where  $n = 1$  is called a single positive hit and the case where  $n \geq 2$  is called a multiple hit. We have developed algorithms to classify a single positive hit to a pattern class given a hybridization protocol. For the case where  $n \geq 2$ , we have developed algorithms to decompose the resulting signal into several single positive hits for further classification.

There have been several attempts to automate the process of analysis and interpretation of DNA hybridization images [1, 5, 6]. However, most of the techniques therein have employed ad-hoc heuristics that work well for high-resolution, noise-free images but not for images that have limited resolution and are corrupted by several hybridization artifacts. The segmentation and pattern classification algorithms developed and presented in this paper are robust to the presence of noisy artifacts and the accidental merging of spot-like features. The pattern classification is performed using shape-based moments [4] but our technique differs from traditional moment-based classifiers in that rotation- and scale-invariance is not a desirable characteristic in our case.

## 2. FEATURE EXTRACTION

A typical hybridization image (Figure 1) exhibits a highly inhomogeneous background and considerable variation in spot size. The inhomogeneous background is due to radioactive residue and the variation in spot size is due to variations in exposure times, spotting concentrations, tem-

perature and hybridization reaction strengths. The images are captured on film and converted into a digital format using a scanner. The typical size of these images is  $950 \times 950$  pixels where each pixel encodes an 8-bit gray scale value. Image preprocessing includes histogram equalization and rotation of the image to align it with the coordinate axes.

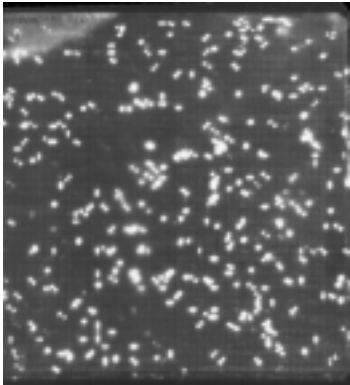


Figure 1: Typical hybridization image

The extraction of spot-like features involves three major steps: (1) segmentation of the hybridization signals from a highly inhomogeneous background, (2) grouping of spot-like features that constitute a single positive hit and (3) decomposition of spot-like features that are a result of multiple probe hits in a single cell.

### 2.1. Recursive Segmentation

Since the extraction of spot-like features from a highly inhomogeneous background cannot be achieved using a global thresholding technique, a recursive segmentation algorithm [1] is used to find a local threshold value. Initially, a very small threshold value  $i_t$ , typically 0, is applied to the entire image resulting in a set of regions of arbitrary shape. A bounding box is associated with each of these regions and a corresponding subimage extracted. A new threshold value  $i'_t$  is computed as  $i'_t = \max\{i_t + 1, i_t + \alpha(i_{max} - i_t)\}$  and used to further split the subimage into several new regions. Here  $i_{max}$  is the maximum pixel intensity in the subimage and  $\alpha = 0.1$  is a constant coefficient. The termination criterion for the recursion is that the region under consideration is identified as a *spot-like feature* or  $i'_t \geq i_{max}$ . The condition  $i'_t \geq i_{max}$  is self-explanatory. The key issue, therefore, is the proper definition of a *spot-like feature*.

In our case, a spot-like feature is deemed to fall in one of the following categories:

*Category 1:* A small spot which is one of the two spots comprising a positive hit. Typically, these small spots are weak in terms of intensity value.

*Category 2:* A large high-intensity region resulting from the merging of two or more spots.

Given the above two categories and the assumption that a spot is homogeneous in terms of gray level, the following criteria are used classify a region as a spot-like feature:

(a) A region is homogeneous with regard to intensity and one of the following conditions hold:

(b.1) The region area is less than the maximum spot size which is determined by the size of the cell (in pixels) on the nylon membrane. This criterion is designed for Category 1 spots.

(b.2) For at least two successive iterations of the procedure, a parent region always creates a single child region. This criterion is designed to retain the shape of Category 2 spots for further decomposition.

The spots in the two categories are differentiated using an area threshold which is determined by the membrane cell size.

### 2.2. Grouping

A grouping algorithm was designed to pair the Category 1 spots. A spot and its partner are assumed to be similar in terms of key properties such as area, intensity, perimeter and elongation. The distance (or difference) between the two spots in image space and in terms of the above key properties is used to determine a dissimilarity measure. The goal of the grouping algorithm is to find for every spot  $S_1$  in Category 1, a qualified Category 1 spot  $S_2$ , such that the dissimilarity measure  $DS(S_1, S_2)$  is minimized over all qualified pairs of Category 1 spots containing  $S_1$ . If no such Category 1 spot  $S_2$  exists, then  $S_1$  is considered a noisy artifact and removed from further consideration. A pair of Category 1 spots ( $S_1, S_2$ ) is deemed a qualified pair if  $DS(S_1, S_2)$  is below a threshold value.

The grouping algorithm for pairing similar spots is a greedy algorithm. The basic idea is to choose a pair of Category 1 spots with the smallest dissimilarity measure from all the qualified pairs at each step and repeat this process until no more qualified pairs remain. The grouping algorithm gives correct results in about 95% of the cases.

### 2.3. Decomposition

A decomposition procedure was developed to decompose large spots resulting from the merging of more than two spots. Typically, the large spots are the result of multiple probe hits within a  $4 \times 4$  square on the membrane. Two approaches to decompose a multiple-hit spot into several single-hit spots were developed, one based on non-convex polygon decomposition and the other based on the Hough transform.

The basic idea behind the first approach is to decompose a non-convex polygon into the smallest number of convex polygons; a classical problem in computational geometry. A single-hit spot is typically convex whereas a spot arising from multiple hits is typically non-convex. In our case, the spot is approximated by a polygon. Thus the two steps in non-convex polygon decomposition are first, determining the break points or non-convex corners on the contour of the polygon, and second, determining the proper sequence of break points for polygon decomposition. The sequence of breakpoints is determined using a greedy algorithm. At each stage we select a pair of breakpoints that result in (sub)spots with the maximum total compactness. The algorithm is halted when all the resulting subspots are convex.

In the Hough transform-based approach, the multiple-hit spots are modeled by the merging of several circles and the Hough transform for circle detection is used to extract

the circles underlying the single-hit spots. A circle of the form  $(x - a)^2 + (y - b)^2 = r^2$  is represented by the 3-D Hough accumulator denoted by  $(a, b, r)$ . At the end of the voting procedure, the cells in the  $(a, b, r)$  accumulator with a value greater than a certain threshold represent circles in image space. The threshold value used is  $r$ -dependent and in our case is  $2\pi r \times c$  where  $c$  is a constant such that  $0 < c < 1$ . The value of  $c$  used by us was in the range  $[0.4, 0.5]$ . Aggregation in the  $(a, b, r)$  accumulator is needed to account for noisy edge pixels. Cells  $\{P_1, P_2, \dots, P_n\}$  in the Hough accumulator are considered to constitute a group if the distance in  $(a, b, r)$  space between any two cells within the group is below a certain threshold and distance between a cell in one group and any other cell in another group is over the threshold. A vote-weighted average of the  $(a, b, r)$  parameters of the member cells is used to represent the group in parameter space.

### 3. PATTERN CLASSIFICATION

The pattern classification procedure consists of two steps: (a) choosing features that would enable class discrimination, and (b) designing a classifier to classify the positive hits to the clone class specified in the hybridization protocol.

#### 3.1. Feature Selection

Some of the desirable feature properties in our case include: (a) high discriminatory information in the chosen features, (b) low dimensionality of the feature vector, (c) scale invariance, (d) sensitivity to object shape and orientation, and (e) ability to compute the feature vector given an experimental protocol, instead of having to learn it from training examples.

For a 2-D shape  $S$  in a binary image, the absolute central moments  $u_{pq}$  are defined as:  $u_{pq} = \sum_S (|u - \bar{u}|)^p (|v - \bar{v}|)^q I(u, v)$  where  $\bar{u}$  and  $\bar{v}$  are the coordinates of the centroid of the 2-D shape and  $I(u, v)$  is the intensity at pixel  $(u, v)$ . For background pixels  $I(u, v) = 0$ , whereas for object pixels  $I(u, v) = 1$ . Using the hybridization protocol, each positive hit is approximated by two circles of similar size. Figure 2 shows two circles  $S_1$  and  $S_2$  with equal radii representing the two spots involved in a positive hit, where  $r$  is the circle radius and point  $(\alpha, \beta)$  is the midpoint of the line segment joining the two circle centers.

The absolute central moments for the continuous case can be computed as:  $u_{pq}(S_1, S_2) = \int_{-r}^r \int_{-r}^r (|u - \alpha|)^p (|v - \beta|)^q dudv$  where  $(u, v)$  is a pixel in either circle  $S_1$  or  $S_2$ . Since  $S_1$  and  $S_2$  are symmetric about  $(\alpha, \beta)$ , we have  $u_{pq}(S_1, S_2) = 2 \int_{-r}^r \int_{-r}^r (|u - \alpha|)^p (|v - \beta|)^q dudv$  where  $(u, v)$  lies in circle  $S_1$ . Based on the above equation, we can formulate the moments  $u_{21}, u_{20}, u_{12}$  and  $u_{02}$  which can be potentially used in the feature vector as:  $u_{20}(S_1, S_2) = \frac{8r^2}{3}(r^2 + 2\alpha^2)$ ,  $u_{02}(S_1, S_2) = \frac{8r^2}{3}(r^2 + 2\beta^2)$ ,  $u_{21}(S_1, S_2) = \frac{8}{3}\beta r^2(r^2 + 2\alpha^2)$ , and  $u_{12}(S_1, S_2) = \frac{8}{3}\alpha r^2(r^2 + 2\beta^2)$ .

From above equations, we can easily identify two candidate shape features  $\frac{u_{12}}{u_{02}} = \alpha$  and  $\frac{u_{21}}{u_{20}} = \beta$ . These features are independent of spot size  $r$ , sensitive to the spot shape but are not orientation sensitive. Thus, an orientation-

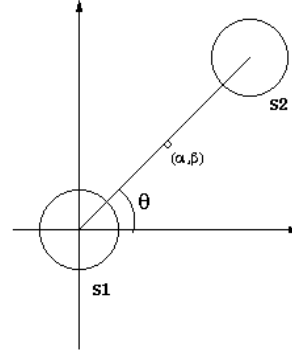


Figure 2: Geometric model for positive hits

dependent feature  $\tan^{-1} \left( \frac{u_{11}}{u_{20} - u_{02}} \right)$  needs to be included in the feature vector. We set the final feature vector to be  $\left[ \frac{u_{12}}{u_{02}}, \frac{u_{21}}{u_{20}}, \tan^{-1} \left( \frac{u_{11}}{u_{20} - u_{02}} \right) \right]$ .

Table 1 shows an experimental protocol ( $4 \times 4$  square on the hybridization membrane) designed at the University of Georgia where cells with the same label in the  $4 \times 4$  square contain the same clone. The clone placement pattern reflects the basic rules of protocol design: (a) if two patterns share the same orientation, they have different inter-spot distances (eg. patterns 1 and 5), and (b) if two patterns share same inter-spot distance, their orientations are different (eg. patterns 1 and 2).

Table 1: The hybridization protocol

1	4	8	2
3	5	8	3
7	7	5	6
2	4	6	1

#### 3.2. Classifier Design

The classifier was designed based on Bayes' decision theory. The feature vector used in the classifier is given by  $\left[ \frac{u_{12}}{u_{02}}, \frac{u_{21}}{u_{20}}, \frac{1}{2} \tan^{-1} \left( \frac{u_{11}}{u_{20} - u_{02}} \right) \right]$ . Instead of using the 3-D feature vector for classification, we used the orientation as a 1-D feature to first classify patterns into groups, where a group may contain more than one pattern class. Subvector  $\left[ \frac{u_{12}}{u_{02}}, \frac{u_{21}}{u_{20}} \right]$  is then used as a 2-D shape feature to further classify the patterns in each group. We prefer the second approach for the following reasons: (a) in both, protocol design and pattern classification by humans, the pattern orientation and pattern shape are examined separately; (b) feature space decomposition reduces the computational complexity significantly, and (c) orientation is computed using second-order moments which makes it more robust to noise than the shape subvector which entails the computation of third-order moments. In statistical terms, we assume that the orientation subvector and the shape subvector are statistically independent.

A 1-D orientation-based Bayesian classifier [7] was designed. The eight patterns described in our protocol (Table

1) can be divided into 4 groups, each containing 2 pattern classes, based on the pattern orientation  $\theta$  as shown in Table 2.

Table 2: Groups of pattern classes based on  $\theta$

Group	Pattern Class	$\theta$
1	3,7	0
2	2,6	$\frac{\pi}{4}$
3	4,8	$\frac{\pi}{2}$
4	1,5	$-\frac{\pi}{4}$

The group-conditional probability distribution for the measured orientation  $\theta$  was assumed to be Gaussian:  $p(\theta|g_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(\theta-\mu_i)^2}{2\sigma_i^2}}$  where  $\mu_i$  is the mean orientation of each group (Table 2) and  $\sigma_i^2$  is the variance associated with the orientation. The orientation variance for each group was estimated using positive hit samples from each pattern class. Using Bayesian formalism the a-posteriori probability distribution  $p(g_i|\theta)$  can be computed as  $p(g_i|\theta) = \frac{p(\theta|g_i)}{p(\theta)} p(g_i)$ . The Bayesian classifier assigns the pattern to the group with the highest a-posteriori probability  $p(g_i|\theta)$ . Under the assumption that all the groups  $g_i$  occur with equal probability and since  $p(\theta)$  does not depend on any particular group, maximizing  $p(g_i|\theta)$  is tantamount to maximizing the group-conditional probability  $p(\theta|g_i)$ . Since  $p(\theta|g_i)$  is Gaussian, maximizing  $p(\theta|g_i)$  is equivalent to minimizing the Mahalanobis distance  $D_M = \frac{(\theta-\mu_i)^2}{2\sigma_i^2}$ . Thus, the input pattern is assigned to the group  $g_i$  with the shortest Mahalanobis distance  $D_M$  [7].

For the shape-based classifier, we assume the class conditional probability density function to be a bivariate Gaussian distribution in  $x = \frac{u_{12}}{u_{02}}$  and  $y = \frac{u_{21}}{u_{20}}$ . The classification approach is one based on minimization of Bayesian risk [7]. For each group, a loss matrix is defined which quantifies the risk of misclassification. For example, Table 3 shows the loss matrix for Group 1 consisting of pattern classes 3 and 7. Class  $C_0$  refers to the *unknown* class.

Bayesian classification minimizes the overall risk of misclassification for every input pattern  $[x, y]$  within its corresponding group  $G$  determined by the orientation classifier. The overall risk of classifying an input pattern  $[x, y]$  to a certain class  $i$  in  $G$  is  $r_i(xy) = \sum_{j \neq i} P_{ij} \times p(C_j|xy)$  where  $P_{ij}$  is the penalty of misclassifying a pattern from class  $i$  to class  $j$  and  $p(C_j|xy) = \frac{1}{|G|} \times \frac{p(xy|C_j)}{\sum_{i \in G} p(xy|C_i)}$ . The pattern is assigned to the class which minimizes the overall risk. Here  $p(xy|C_i)$  is the class conditional probability density function which is assumed to be a bivariate Gaussian distribution. The mean vector and covariance matrix of the bivariate Gaussian distribution are estimated from the training samples.

#### 4. EXPERIMENT RESULTS

The DNAScan program was tested on approximately 100 real hybridization images. Manual classification of these

Table 3: Loss matrix for Bayesian classifier

	$C_3$	$C_7$	$C_0$
$C_3$	0	1	0.25
$C_7$	1	0	0.25

images was used as the ground truth. The deviation of the output of the program from the result of manual classification included (a) missing positive hits which could not be detected due to low background contrast, (b) wrong grouping of spot-like objects, and (c) misclassification of patterns. Experimental results showed the correct detection rate to be between 65% and 95% depending on the pattern class. It was noted that most of the misses (between 70% and 80%) were assigned to the unknown class which is a better situation than the one in which most of the misses are assigned to another pattern class. The patterns assigned to the unknown class could be referred to a human for further consideration.

#### 5. CONCLUSIONS

The paper presented a computer vision-based system called DNAScan for the analysis of DNA hybridization images. The image analysis was shown to consist of two stages, extraction of patterns denoting positive hits followed by pattern classification. Extraction of positive hits was shown to involve recursive segmentation, grouping and pattern decomposition for which algorithms were designed, implemented and tested. In the second stage which involved classifier design, a mathematical model for the hybridization protocol was proposed. The model was general and flexible enough to encompass a variety of experimental protocols. A two-stage Bayesian classifier based on the mathematical model was implemented and tested on several real hybridization images with satisfactory results.

#### 6. REFERENCES

- [1] S. Audic and G. Zanetti, Automatic reading of hybridization filter images, *CABIOS*, 11(5):489-495, 1995.
- [2] T. Brown, *Genomes*, John Wiley, New York, NY, 1999.
- [3] M. Friedman and A. Kandel, *Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy Logic Approaches*, World Sci. Pub. Co., New York, NY, 1999.
- [4] M.K. Hu, Visual pattern recognition by moment invariants, *IRE Trans. Info. Theory*, IT-8:179-187, 1962.
- [5] J. Piper, D. Rutovitz, D. Sudar, A. Kallioniemi, O. Kallioniemi, F. Waldman, J. Gray and D. Pinkel, Computer image analysis of comparative genomic hybridization, *Cytometry*, 19:10-26, 1995.
- [6] K. Roth, G. Wolf, M. Dietel and I. Peterson, Image analysis for comparative genomic hybridization based on a karyotyping program for Windows, *Anal. and Quant. Cytology and Histology*, 19(6):461-473, 1997.
- [7] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, San Diego, California, 1999.