Entropy Estimation for Segmentation of Multi-Spectral Chromosome Images

Wade Schwartzkopf, Brian L. Evans, and Alan C. Bovik
Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin, TX 78712-1084
{wade,bevans,bovik}@ece.utexas.edu

Abstract

In the early 1990s, the state-of-the-art in commercial chromosome image acquisition was grayscale. Automated chromosome classification was based on the grayscale image and boundary information obtained during segmentation. Multi-spectral image acquisition was developed in 1990 and commercialized in the mid-1990s. One acquisition method, multiplex fluorescence in-situ hybridization (M-FISH), uses five color dyes. We previously introduced a segmentation algorithm for M-FISH images that minimizes the entropy of classified pixels within possible chromosomes. In this paper, we extend this entropy-minimization algorithm to work on raw image data, which removes the requirement for pixel classification. This method works by estimating entropy from raw image data rather than calculating entropy from classified pixels. A successful example image is given to illustrate the algorithm. Finally, it is determined that entropy estimation for minimum entropy segmentation adds a heavy computational burden without contributing any significant increase in classification performance, and thus not worth the effort.

1. Introduction

Chromosomes are the cell structures that contain genetic information. When chromosomes are photographed, the images contain much information about the health of an individual. The images are useful for diagnosing disorders and studying various diseases. In the past, it has been necessary for laboratory technicians to examine these images visually to collect the useful information contained in these images. However, since many images often have to be inspected and since visual inspection is time consuming and expensive, many attempts have been made to automate chromosome image analysis. For example, automated segmentation algorithms for grayscale chromosome images have been able to correctly decompose about 80-90% of touching and overlapping chromosomes [1, 2, 3]. These automated procedures rely on chromosome shape and texture.

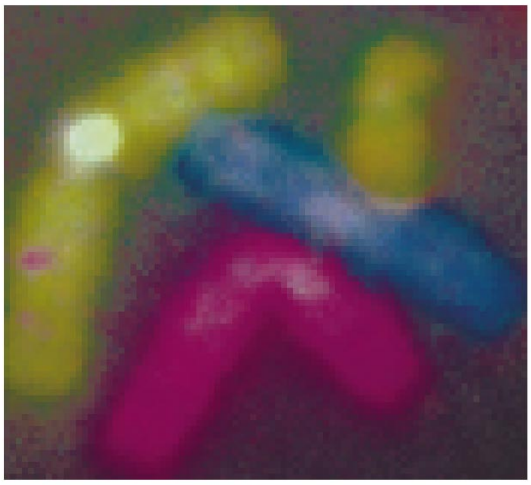
In the 1990's, new techniques were developed to dye chromosomes with multiple colors so that each chromosome class appears to be a distinct color. This makes analysis of chromosome images easier, not only for human inspection, but also for computer analysis. This research focuses on one such dying technique, known as M-FISH (multiplex fluorescence in-situ hybridization). This work takes advantage of the multi-spectral information in M-FISH images to improve past methods of computer segmentation and analysis of chromosome images.

2. Multi-spectral M-FISH images

A new way to image chromosomes came about with the invention of chromosome painting [4], combinatorial labeling [5] and ratio labeling [6]. These techniques made use of fluorophores (dyes) that attach to a single class of chromosomes, parts of chromosomes, or specific sequences of DNA. Using these techniques, one could create a combination of fluorophores such that each class of chromosomes absorbed a different combination of these fluorophores [7, 8, 9]. Therefore, each chromosome class would appear to be a different color and would be visually distinguishable from all of the other classes.

An image of each fluorophore can be obtained by employing appropriate optical filters. This way, each pixel could be represented as a vector, where each element represents the intensity of the response to one fluorophore. Instead of obtaining a grayscale image by traditional chromosome imaging techniques, such as Giemsa banding [10], a multi-spectral image could now obtained in which the spectral composition at each point

a) Boundary of cluster



b) Multi-spectral information in cluster

Figure 1: Comparison of two types of cluster information

reveals the combination of fluorophores and, thus, the chromosome class of the matter at that point. Using this combinatorial labeling, known as M-FISH, one can determine the chromosome class at every pixel.

Such an imaging technique has a couple of obvious advantages. First, the task of chromosome classification is greatly simplified. Instead of determining and then comparing the chromosome lengths, centromere positions, and banding patterns, one only has to look at the spectral information within that chromosome. The second advantage is that it is possible to detect smaller translocations and rearrangements than were discernible with grayscale chromosome banding patterns only [11].

3. Multi-spectral entropy estimation for segmentation

Traditional chromosome segmentation methods use shape information from the boundary of the chromosomes to detect and decompose clusters. Cut points are then found by examining the shape of boundary of the cluster [2, 3, 12]. Occasionally, grayscale information from inside the chromosome clusters is also used. One popular method is "valley searching" [13] where a minimum cost algorithm attempts to locate low gray-value valleys running through the cluster to locate separation between the chromosomes.

With M-FISH images, a new source of information is available for segmentation. There are examples in which even a trained observercannot determine the correct segmentation just by looking at boundary or greyscale information in a cluster of chromosomes, but seeing the multispectral information makes it obvious (See Fig. 1).

To use the multi-spectral information available in M-FISH, we previously introducedan objective function of minimum entropy that uses this multi-spectral information to evaluate possible cut lines [14]. For the objective function, we used a measure of entropy. In particular, we used Shannon's definition of entropy for a discrete random variable [15]

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

where n is the number of possible classes. Probability p_i was calculated as the percentage of class i pixels within the object. In a perfectly classified and segmented image, the entropy of each segment will be zero, since all the pixels in each segment will be classified into the same class. The larger the number of different classes that are found within a segment, the higher the entropy will be.

However, to use this entropy measure, it was necessary to employ a step of pixel classification before segmentation. We propose calculating entropy from raw image data, using a differential entropy estimation technique, and thus avoiding this classification step. In particular, we have used the nearest neighbor estimation technique [16].

Let $\rho_{n,i}$ be the nearest neighbor distance of X_i and its nearest neighbor X_j: $\rho_{n,i} = \min_{j \neq i, j \leq n} \left\| X_i - X_j \right\|$. Then the nearest neighbor estimate is given by

a) M-FISH image b) Connected components c) Entropy-segmented image

Figure 2: Example of an M-FISH image segmented with entropy segmentation

$$H_n = \frac{1}{n}\sum_{i=1}^{n} \ln\left(n\rho_{n,i}\right) + \ln 2 + C_E \qquad (2)$$

where C_E is the Euler constant: $C_E = -\int_0^\infty e^{-t} \ln t \, dt$. This estimator can be shown to have mean square consistency. It has been chosen because it is defined for multiple dimensions and because it is less computationally complex than other methods such as minimal spanning tree methods [17].

Aside from the benefit of being able to remove the step of pixel classification from the process of chromosome image segmentation, the hope with this algorithm was that performance might be gained over past minimum entropy methods [14] on chromosomes whose pixels lay near the decision boundary of two classes for some pixel classifier. In such a case, the classified pixels for that chromosome might come from some distribution of two different classes, while the vector values of the pixels within that chromosome might actually be very close to each other. So in this case, a minimum entropy algorithm based on classified pixels might incorrectly split the chromosome, whereas the entropy estimation technique would correctly recognize the object as a single chromosome.

4. Example

Fig. 2 segments an M-FISH image using the entropy method. The original chromosome image is in Fig. 2a. Fig. 2b shows the connected components of the image after thresholding. In this image, several groups of chromosomes (those labeled 16, 19, 23, 29, and 35) are labeled as a single object. Fig. 2c shows the entropy-segmented chromosomes. All touching chromosomes were correctly split. The two overlapped chromosomes (40 and 43) were correctly identified. In these chromosomes, both ends are labeled as one chromosome.

5. Conclusion

This paper extends the idea of entropy as a criterion for selecting cut lines to decompose groups of chromosomes that touch and overlap each other. This algorithm uses nearest neighbor distances to estimate entropy from raw image data to accomplish minimum entropy segmentation without requiring pixel classification.

We tested the entropy estimation technique on selected images from a public database of 200 hand-segmented M-FISH images. This database is available from Advanced Digital Imaging Research at

http://www.adires.com/projects/mfish_db.shtml

This database contains 200 hand-segmented M-FISH images, or approximately 9000 individual chromosomes.

For comparison, a simple pixel classification algorithm was run on the entire ADIR MFISH dataset. Performance of the pixel classifier was around 70% for the entire database, although it varied widely across the dataset from images with 30% classification accuracy to 95%. Images used for the test had pixel accuracy rates of 70-80%. We then performed minimum entropy segmentation using the algorithm found in [14].

The entropy estimation algorithm worked on many images, but its performance was very sensitive to its parameters, such as the entropy threshold that

distinguished chromosomes from chromosome clusters, and the entropy difference threshold, that is the drop in entropy necessary for a cut to be considered valid. Often a set a parameters could be found that would work for an image or two, but no general set of parameters worked for a large number of the database images. Some images had no set of parameters that segmented the image correctly. In general, performance did not exceed that of entropy segmentation via pixel classification [14].

Furthermore, the computation time of the entropy estimation algorithm was prohibitive, often taking hours to days to segment a single image. This drawback made it impossible to test the algorithm on all 200 images in the dataset, so only a few images were tried, but the performance of the algorithm on these images never exceeded that of the pixel-classification-based algorithm. Even though integrating entropy estimation into the algorithm avoids the step of pixel classification, the added computational load of entropy estimation for each possible cut point is much larger than the step of pixel classification, which it replaced.

Thus it seems that entropy estimation adds the unnecessarily burdensome computation while providing no increment in performance. This is likely because rarely are chromosome's pixel vector-values smooth and still found on a pixel classifier's decision boundary. Apparently, more often what happens is that high variance is found within chromosome pixel vector values, but this variance still falls within pixel classifier bounds, so its pixels are still classified well and thus segmented correctly.

Because of its computational complexity and its poor segmentation performance, we conclude that entropy estimation is a poor method for minimum entropy MFISH segmentation.

6. References

[1] C. Lundsteen and J. Piper, *Automation of Cytogenetics*, Berlin, Springer-Verlag, 1989.

[2] J. Liang, "Intelligent Splitting in the Chromosome Domain," *Pattern Recognition*, vol. 22, no. 5, pp. 519-532, 1989.

[3] G. Agam and I. Dinstein, "Geometric Separation of Partially Overlapping Nonrigid Objects Applied to Automatic Chromosome Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 11, pp. 1212-1222, 1997.

[4] D. Pinkel, T. Straume, and J. W. Gray, "Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization," *Proc. National Academy of Sciences of the United States of America*, vol. 83, pp. 2934-2938, 1986.

[5] P. M. Nederlof, S. van der Flier, J. Wiegant, A. K. Raap, H. J. Tanke, J. S. Ploem, and M. van der Ploeg, "Multiple fluorescence in situ hybridization," *Cytometry*, vol. 11, pp. 126-131, 1990.

[6] P. M. Nederlof, S. van der Flier, J. Vrolijk, H. J. Tanke, and A. K. Raap, "Fluorescence Ratio Measurements of Double-labeled Probes for Multiple in Situ Hybridization by Digital Imaging Microscopy," *Cytometry*, vol. 13, pp. 839-845, 1992.

[7] M. R. Speicher, S. G. Ballard, and D. C. Ward, "Karyotyping Human Chromosomes by Combinatorial Multi-fluor FISH," *Nature Genetics*, vol. 12, pp. 368-375, 1996.

[8] M. M. Le Beau, "One FISH, two FISH, red FISH, blue FISH," *Nature Genetics*, vol. 12, pp. 341-344, 1996.

[9] T. Ried, A. Baldini, T. C. Rand, and D. C. Ward, "Simultaneous visualization of seven different DNA probes by in situ hybridization using combinatorial fluorescence and digital imaging microscopy," *Proc. National Academy of Sciences of the United States of America*, vol. 89, pp. 1388-1392, 1992.

[10] M. Seabright, "A rapid banding technique for human chromosomes," *Lancet ii*, pp. 971-2, 1971.

[11] T. Veldman, C. Vignon, E. Schröck, J. D. Rowley, and T. Ried, "Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping," *Nature Genetics*, vol. 15, pp. 406-410, 1997.

[12] Q. Wu, *Automated Identification of Human Chromosomes as an Exercise in Building Intelligent Image Recognition Systems*, Catholic University of Leuven, Leuven, Belgium, Doctoral Dissertation, 1987.

[13] A. M. Vossepoel, *Analysis of Image Segmentation for Automated Chromosome Identification*, University of Leiden, Leiden, Netherlands, Doctoral Dissertation, 1987.

[14] W. Schwartzkopf, B. L. Evans, and A. C. Bovik, ``Minimum Entropy Segmentation Applied to Multi-Spectral Chromosome Images", *Proc. IEEE Int. Conf. on Image Processing*, Oct. 7-10, 2001, vol. II, pp. 865-868, Thessaloniki, Greece.

[15] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Sys. Tech. Journal*, vol. 27, pp. 379-423, 623-656, 1948.

[16] J. Beirlant, E. J. Dudewicz, L. Györfi, E. C. van der Meulen, "Nonparametric Entropy Estimation: An Overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17-39, June 1997.

[17] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 1921-1938, September 1999.