Segmentation of 2D Gel Electrophoresis Spots Using a Markov Random Field

Christopher S. Hoeflich and Jason J. Corso
csh7@cse.buffalo.edu, jcorso@cse.buffalo.edu

Computer Science and Engineering
University at Buffalo, Buffalo, NY, USA

ABSTRACT

We propose a statistical model-based approach for the segmentation of fragments of DNA as a first step in the automation of the primarily manual process of comparing two or more images resulting from the Restriction Landmark Genomic Scanning (RLGS) method. These 2D gel electrophoresis images are the product of the separation of DNA into fragments that appear as spots on X-ray films. The goal is to find instances where a spot appears in one image and not in another since a missing spot can be correlated with a region of DNA that has been affected by a disease such as cancer. The entire comparison process is typically done manually, which is tedious and very error prone. We pose the problem as the labeling of each image pixel as either a spot or non-spot and use a Markov Random Field (MRF) model and simulated annealing for inference. Neighboring spot labels are then connected to form spot regions. The MRF based model was tested on actual 2D gel electrophoresis images.

Keywords: Segmentation, Statistical Methods, Pattern Recognition

1. INTRODUCTION

Automatic matching of 2D gel electrophoresis images resulting from the Restriction Landmark Genomic Scanning (RLGS) method, first developed by Hatada, et. al.,[1] is a very challenging yet practical problem. A sample of DNA is separated in a gel-like material based on two different properties, such as molecular composition and molecular weight, and an X-ray film is placed over these gels to capture the resulting separated fragments of DNA. Two different images, one corresponding to a normal sample of DNA and another corresponding to an abnormal sample of DNA, are typically compared to one another. The task is to first detect and segment spots within each image that correspond to fragments of DNA and then perform matching, ultimately determining which spots appear in one image and not in the other. Missing spots correspond to specific fragments of DNA that have been methylated as the result of exposure to a disease such as cancer. In this paper, we propose a method to solve the problem of spot detection and segmentation using a Markov Random Field (MRF) model and simulated annealing.

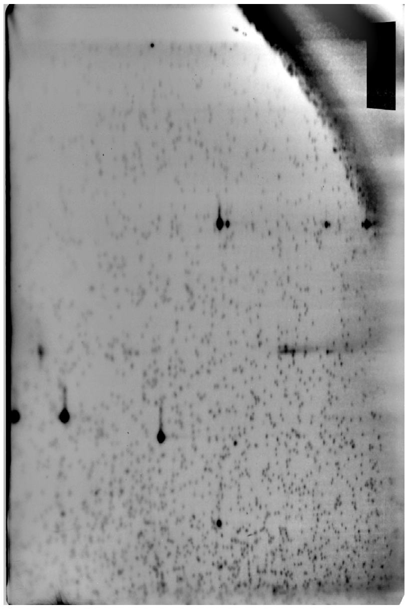


Figure 1: RLGS Image

The entire process of matching 2D gel electrophoresis images is widely done completely manually, in which a technician sits at a light table and visually detects spots that occur in one image and not in the other. As seen in Figure 1, there are typically hundreds of spots contained in a single image, and it is not possible to directly overlay images due to non-linear distortions and variations in corresponding spots. Automatic segmentation of these spots is problematic because the intensity of spots across a single image varies greatly, so basic thresholding techniques can not be employed. Furthermore, a single spot is typically composed of a wide range of pixel intensities because these images are created by placing an X-ray film over the gels, and not every spot fully saturates the film.

2. PREVIOUS WORK

Previous developments in this area have employed a wide range of techniques, and a majority of applications address the matching of 2D gel electrophoresis images for proteins as opposed to DNA. This is because 2D gel electrophoresis images for proteins have a relatively more uniform background and are somewhat easier to work with than the 2D gel electrophoresis images for DNA. Kim et. al.[2] proposed a hierarchical segmentation based on thresholding and the detection of watersheds. They first pre-process the images to remove noise and enhance contrast, then thresholding is applied which produces large regions. A watershed detection algorithm is then applied recursively on these regions until only a single blob is detected which is considered to be a spot. Their method relies on setting several parameters and is sensitive to noise, and 2D gel electrophoresis images typically contain noise.

Sugahara et. al.[4] smoothed image regions by averaging pixel intensities using an $m \times m$ window and performed a thresholding operation which ultimately subtracted the background, and then created a binary image for spot detection. This method relies heavily on the selection of a proper threshold value which can cause either an over-segmentation of spots in some regions as well as an under-segmentation of spots in other regions. Takahashi et. al.[5] performed image enhancement and smoothing before defining local maxima in order to label the spots. This method also relies on the definition of threshold values in order to function properly.

More recently, Morris et. al.[3] developed a very accurate and robust method of detecting spots in 2D gel electrophoresis images. Their process involves an "average gel" which is created by first using registration software to create an alignment of all gels being used. The pixel intensities are then averaged across the aligned gels. The gels are each de-noised using the average gel, and pinnacles (regions that are a local maximum in both the horizontal and vertical directions and above a certain threshold) are detected which denote the spot locations. A disadvantage of this method is the need to perform image registration as a pre-processing step and the need to define a threshold in order to determine which regions are pinnacles.

3. METHODS

In order to automatically segment the DNA fragments in 2D gel electrophoresis images, we propose a Markov Random Field (MRF) based segmentation that uses simulated annealing for inference in order to label each pixel in the image as either a spot or non-spot. The simulated annealing process introduces a temperature term to a Gibbs distribution:

$$p(y) = \frac{1}{Z} \exp\left(\frac{-H(y)}{T}\right), \tag{1}$$

where Z is a normalizing constant, y are the MRF variables, $H(y)$ is an energy function, and T is the temperature. T is typically referred to as the *inverse temperature* because the energy function is being multiplied by $\frac{1}{T}$. Note that as $T \to 0$, $\frac{1}{T} \to \infty$. Winkler[7] shows that given an energy function $H(y)$, as $T \to 0$ a Gibbs distribution will converge to its maximal modes. Thus, sampling from a Gibbs distribution that slowly decreases T will eventually yield maximal modes with high probability.

3.1 MRF Model

Our goal is to assign a single label of either spot of non-spot to each pixel location. An image I consists of N pixels such that $I = \{i | i = 1, \ldots, N\}$, and x_i denotes the intensity of pixel i such that $x_i = \{0, \ldots, 255\}$. We define the set of possible labels as $\gamma = \{-1, +1\}$ where -1 represents the location of a non-spot pixel and $+1$ represents the location of a spot pixel. The label assigned to each pixel location of the image is that label which maximizes the probability $p(\gamma|I)$. Brute force maximization of this probability is computationally intractable since for N pixels there are 2^N possible label assignments. As a result, we model this maximization problem using a Gibbs distribution:

$$p(\gamma|I) = \frac{1}{Z} \exp\left(-\beta_1 \sum_{i \in I} U_V(N(i), \gamma_i) - \beta_2 \sum_{i \in I} U_I(x_i, \gamma_i) - \beta_3 \sum_{i \sim t} U_S(\gamma_i, \gamma_t)\right) .$$

The variables β_1, β_2, and β_3 are tunable model parameters, $N(i)$ denotes the first-order and second-order neighbors of pixel i, and $i \sim t$ denotes the first-order neighbors of pixel i. Our model takes into consideration local

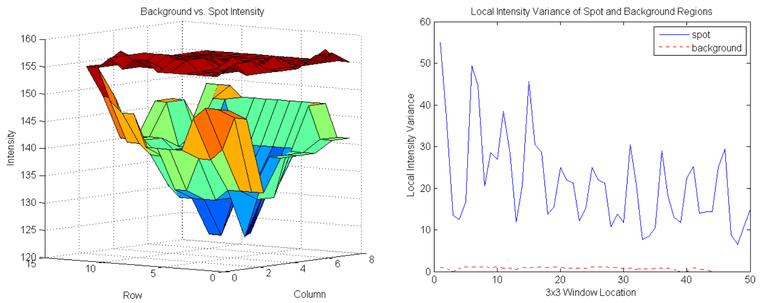


Figure 2: Spot Regions Have Higher Variance

variance of pixel intensities, the intensity value of each pixel, and the labels assigned to neighboring pixels through the use of the following terms.

Variance. The term $U_V(N(i), \gamma_i)$ models the variance of pixel i and its eight-connected neighbors $N(i)$ and is given by:

$$U_V(N(i), \gamma_i) = \gamma_i \frac{1}{|\{i, N(i)\}| - 1} \sum_{j=\{i,N(i)\}}^{N} (x_j - \mu_i)^2 \ . \tag{2}$$

In this equation, μ_i is the local mean of the intensity values of pixel i along with its eight neighbors $N(i)$. Our model encourages a lower energy when the variance is high and the label is $+1$ for a spot region and encourages a higher energy when the label is assigned to be -1 for a non-spot region. Spot regions have a much higher variance of pixel intensity than the non-spot regions surrounding them. Figure 2 shows a three-dimensional plot of a typical spot region and a typical non-spot region on the same set of axis. The (x, y, z) values correspond to the $(row, column, intensity)$ values of the spot and the non-spot regions, with the spot intensities appearing as an inverted peak and the background intensities appearing as a relatively smooth sheet above the inverted peak. The graph on the right shows a plot of the local variance as a $3x3$ window is moved over a spot region and then a non-spot region. The average local spot variance of the intensity was found to be 21.48 and the average local non-spot variance of the intensity was found to be 0.69. It is therefore evident that spot regions have a higher variance of intensity than the variance of intensity of non-spot regions.

Intensity. The term $U_I(x_i, \gamma_i)$ models the the intensity of a pixel i given its intensity x_i and its current label γ_i. This term is given by:

$$U_I(x_i, \gamma_i) = \begin{cases} \frac{(x_i - \mu)^2}{2\sigma^2} & \text{if } \gamma_i = +1 \\ -\log\left(1 - \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right) & \text{if } \gamma_i = -1 \end{cases} \ . \tag{3}$$

We model the intensity term as the negative log probability of a Gaussian (see Figure 3 for justification). The normalizing term $\frac{1}{\sqrt{2\pi\sigma^2}}$ is not used because it does not rely on the intensity of the pixel. The parameters of the Gaussian $\{\mu, \sigma^2\}$ were learned from actual data by manually segmenting spots and calculating the mean and variance of their intensities.

Smoothing. The term $U_S(\gamma_i, \gamma_t)$ promotes the labeling of pixels that are similar to the labels of their neighboring pixels. This term is given by a standard Ising model:

$$U_S(\gamma_i, \gamma_t) = \gamma_i \gamma_t \ . \tag{4}$$

This smoothing term is used to make the model more robust to noise that commonly occurs throughout 2D gel electrophoresis images.

Cooling Schedule. The simulated annealing process prefers a cooling schedule that starts with a high temperature T that slowly decreases to a temperature with a relatively small value. We model the cooling

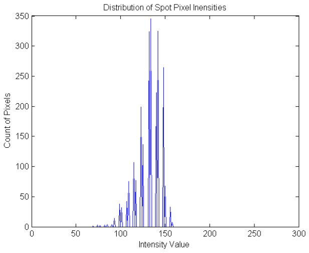


Figure 3: Spot Intensity Distribution

schedule as an inverse logarithm using the equation:

$$T^{(t)} = T^{(0)} \left(\frac{T^{(K)}}{T^{(0)}} \right)^{\frac{t}{K}}, \tag{5}$$

where $T^{(t)}$ is the temperature at time t, $T^{(0)}$ is the initial temperature, $T^{(K)}$ is the final temperature, and K is the desired number of temperature values that $T^{(t)}$ should have. K is the number of sweeps of the Gibbs sampler that are performed.

3.2 Simulated Annealing

Figure 4 is a representation of the simulated annealing process. The general procedure begins as labels are randomly assigned to each pixel location and a suitable cooling schedule is created which has a high initial starting temperature which lowers slowly to a relatively low, final temperature. One sweep of the Gibbs sampler is then performed for each pixel location with the current value of the temperature T, which in turn updates all pixel labels. If T has not yet reached its final value then T is slightly decreased and another sweep of the Gibbs sampler is performed for each pixel location. The entire process continues until T is at its final and lowest value, at which point each pixel has been assigned a label which is a good approximation of a globally optimum labeling. The claim of globally optimal labeling is based on proofs of the simulated annealing process provided in *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*.[7]

3.3 Connected Components

The result of the simulated annealing process in conjunction with the underlying MRF model is a set of labels for each pixel location in the image. The last step in the segmentation process is to use the two-pass connected components algorithm in order to connect all of the +1 labels which denote spot regions such that each individual pixel is associated with a connected spot region.

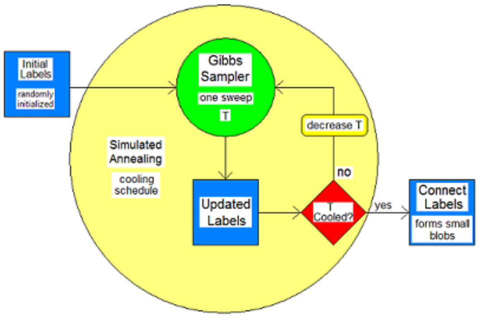
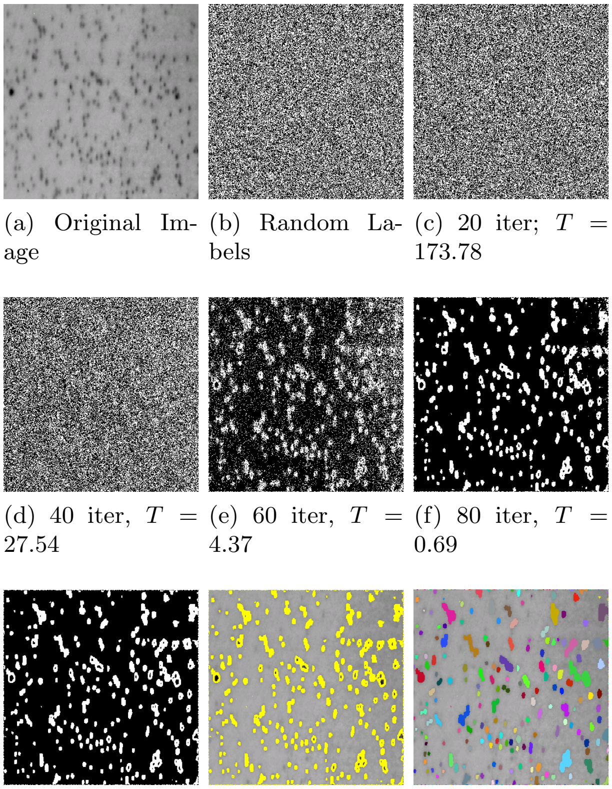


Figure 4: Simulated Annealing Process

4. RESULTS

We implemented the proposed model and tested it on actual 2D gel electrophoresis images. We learned the parameters of the Gaussian distribution used in the second term of the Gibbs distribution, $\{\mu, \sigma^2\}$, off-line from sample images, and we set the model parameters β_1, β_2, and β_3 by hand to $0.2, 0.8$, and 0.3, respectively. We performed a total of 100 iterations in the simulated annealing process, setting the parameters of the cooling schedule to $T^{(0)} = 1000$, $T^{(K)} = 0.1$, and $K = 100$. The sample image selected has dimensions 346 × 346.

Figure 5 is a graphical representation of our entire proposed process. 5a shows the original image, and 5b shows the initial randomly assigned labels, where white denotes a spot pixel and black denotes a non-spot pixel. 5c through 5g show the eventual convergence of the simulated annealing process. The current sweep number and temperature which resulted in the labeling shown are found below each respective image. 5h shows the final labeling of pixels as spots and non-spots superimposed on the original image, with the yellow color showing which pixels have been labeled as spots. The final image 5i shows the result of connecting neighboring spot pixels in order to form small blobs. A different color has been assigned to each spot in order to show a distinct labeling of spots. The labels undergo a "burning in" process up to approximately iteration 50 in which the assigned labels are more sporadic due to the higher values of T. The initially sporadic labeling helps to insure that the likelihood of the labels assigned avoids local minima. By approximately iteration 60 it becomes evident that the labeling is beginning to converge to the locations of the spot pixels. Then, through iterations 80 and 100, the noisy labeling along the right hand side of the image is removed as the temperature reaches its lowest, stable value.



(a) Original Image

(b) Random Labels

(c) 20 iter; $T = 173.78$

(d) 40 iter, $T = 27.54$

(e) 60 iter, $T = 4.37$

(f) 80 iter, $T = 0.69$

(g) 100 iter, $T = 0.01$

(h) Final Labels

(i) Connected Blobs

Figure 5: Convergence Process

5. CONCLUSION

We have developed a model that segments and labels regions of high variance of intensity as belonging to the same class of spot pixels. We have proposed a model for a Gibbs distribution that characterizes a spot of a 2D gel electrophoresis image by taking into account local variance of pixel intensities, the intensity of each individual pixel, and the labels of neighboring pixels. By incorporating this Gibbs distribution into the simulated annealing process along with an appropriately chosen cooling schedule, each pixel is assigned a label the contributes the least amount of energy to the image as a whole.

As seen in Figure 5, nearly all spots have been properly labeled and independently grouped together as small blobs. A problem with the independent blobs is that spot regions that touch one-another have been labeled as one single spot. A post processing step thus needs to be incorporated which separates touching spots into two or more distinct spots. We can accomplished this using a shape model since each spot can be represented by an ellipse.

In our future work, we are investigating the second step of the process: matching DNA spots across two images. We plan to use Grauman and Darrel's Pyramid Match Kernel.[6] The addition of the matching process would automate the matching of 2D gel electrophoresis images.

6. ACKNOWLEDGMENTS

The authors wish to thank Dr. Hiroki Nagase and his research laboratory of Roswell Park Cancer Institute for providing the RLGS images used in this research. Christopher Hoeflich was funded through the Integrative Graduate Education and Research Training (IGERT) Fellowship Program of the University at Buffalo.

REFERENCES

1. Hatada et. al. A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences of the USA*, 88(21):9523–9527, 1991.
2. Kim et. al. Segmentation of protein spots in 2d gel electrophoresis images with watersheds using hierarchical threshold. *LNCS - Computer and Information Sciences - ISCIS 2003*, 2869:389–396, 2003.
3. Morris et. al. Pinnacle: a fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data. *Bioinformatics*, 24(4):529–536, 2008.
4. Sugahara et. al. An automatic image analysis system for rlgs films. *Mammalian Genome*, 9:643–651, 1998.
5. Takahashi et. al. Dnainsight: An image processing system for 2-d gel electrophoresis of genomic dna. *Genome Informatics*, 8:135–146, 1997.
6. Kristen Grauman and Trevor Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, 2007.
7. Gerhard Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer, Berlin, 2nd edition, 2003.