Automated Spot Type Identification on NAPPA Arrays

Robert Rivera
Department of Biomedical Informatics
Arizona State University
13212 East Shea Boulevard
Scottsdale, AZ 85259
rdriver1@asu.edu

Jie Wang
Center for Personalized Diagnostics
The Biodesign Institute
Arizona State University
PO Box 875001
Tempe, AZ 85287-5001
Jie.Wang.1@asu.edu

Ji Qiu
Center for Personalized Diagnostics
The Biodesign Institute
Arizona State University
PO Box 875001
Tempe, AZ 85287-5001
Ji.Qiu@asu.edu

Joshua LaBaer
Center for Personalized Diagnostics
The Biodesign Institute
Arizona State University
PO Box 875001
Tempe, AZ 85287-5001
Joshua.LaBaer@asu.edu

Garrick Wallstrom
Department of Biomedical Informatics and CPD, The Biodesign Institute
Arizona State University
PO Box 875001
Tempe, AZ 85287-5001
Garrick.Wallstrom@asu.edu

ABSTRACT

Biomarker discovery studies that utilize NAPPA protein microarrays must identify spots on each array with significant protein binding. One approach is to identify spots for which the signal extends out beyond the spot, which we refer to as the *halo effect*. The identification of spots exhibiting the halo effect is a cumbersome process that requires humans to adjust the contrast and brightness of the image of the microarray and scan through 2352 spots per image. The goal of our research is to create a set of attributes using pixel intensity data within the image that will accurately classify halo spots on the microarray in order to automate identification of halo spots, reduce time spent on classifications, and prevent user-to-user variability between classifications. This paper describes how we generated relevant attributes and used data mining techniques to create a model that would classify each spot on a protein microarray. With our approach we were able to classify halos with a recall, precision, and f-measure of .818, .655, and .727 respectively.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Application – *data mining.*

I.4.0 [**Image Processing and Computer Vision**]: General – *image processing software.*

I.5.2 [**Pattern Recognition**]: Design Methodology – *classifier design and evaluation, feature evaluation and selection*

KDD-DMH'13, August 11, 2013, Chicago, Illinois, USA.

General Terms

Algorithms, Measurement, Performance, Design, Reliability, Human Factors, Standardization.

Keywords

Halo, Microarray, NAPPA, Biomedical Image, Data Mining

1. INTRODUCTION

NAPPA- Nucleic Acid Programmable Protein Array is a method of producing protein microarrays that was designed to avoid the problems associated with using purified proteins for microarrays by instead printing cDNA and producing fresh proteins *in situ* on the microarrays [5]. NAPPA technology has been used to discover biomarkers for the early detection of breast cancer [2] and is currently being used to search for biomarkers in other diseases including type 1 diabetes, colon cancer, ovarian cancer, and tuberculosis.

Traditional analysis of NAPPA data utilizes the sum of pixel intensity within each spot on the array as a measure of protein binding. A promising alternative approach is to identify halo spots in which the pixel intensity is elevated in a region even outside of the spot itself. Using ELISA (Enzyme-linked Immunosorbent Assay) and LIPS (Luciferase Immuno-Precipitation System) as two orthogonal methods to NAPPA, we have successfully confirmed that the halo effect of P53 spots indicate positive autoantibody responses in basal like breast cancer patients (data not shown). Currently, halo identification has been adopted as a complementary method for analyzing data generated from autoantibody biomarker screening as well as other studies utilizing NAPPA technology, such as protein post-translational modifications and protein-protein interaction.

However, the time it takes for a user to evaluate a microarray image for halos ranges from 10-20 minutes per image. In order to automate halo classification and eliminate user-to-user variability, we developed a Java program that can automate this process. The program incorporates the ImageJ API [1] to generate attributes and utilizes the Weka API [4] to learn a model based on these attributes that would be able to classify each of the spots on a microarray image. Spot types include halos, weak halos, comets, regular spots, and blank spots (see Figure 1 and 2). We show that through the use of our program we can significantly decrease the amount of time spent on manual classification of halos. It would take users 16 hours to make halo classifications on 100 images. Using our statistical attribute model, which is described in the methods section, 100 images can be classified in approximately 3 minutes.

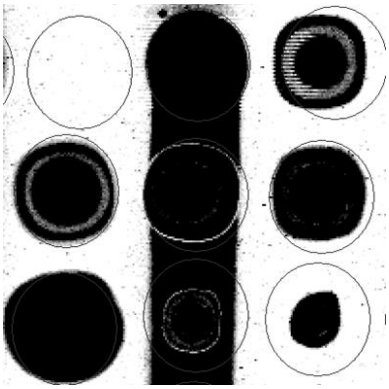

Figure 1. Image of comets, regular spots, and blank spots. The spot in the upper left corner is a blank. The spots in the middle column are all considered comets. All other spots are regular spots.

2. METHODS

The process by which our program classified spot types on microarrays can be broken down into three main parts: spot recognition and starter data generation using Array-Pro Analyzer (Media Cybernetics, Inc.), attribute generation using our program, and classification of spot types using data mining techniques.

2.1 Spot Recognition and Generation of Startup Data

In order to generate the attributes used to classify the spots, each image was manually loaded into Array-Pro. Using Array-Pro, the users generate a grid of 2352 circles, each circle encapsulating a spot in the microarray, except in the case where the circle was in a location where no binding occurred. After generating the grids, Array-Pro is capable of providing image data for each spot on the grid. For the purpose of this experiment, position (row:column), mean background intensity, and horizontal and vertical pixel location of the center of the grid circle were generated and saved to be used in the attribute generation portion.

2.2 Attribute Generation

The program that was written for this research is capable of generating two different sets of attributes: a statistical set and a pixel intensity set.

2.2.1 Statistical Attribute Set

This attribute set contains 33 relevant attributes corresponding to each spot on the microarray image. Each attribute (with the exception of the mean background intensity) is either a pixel's distance from the center of the spot (this attribute will be referred to as radius), the intensity of a specific pixel, or a statistical computation using these two variables. The radius is not the standard physical metric that is used in everyday language, but rather a measure of how many pixels the pixel of interest is shifted vertically or horizontally from the location of the center pixel. For example, if the pixel of interest is located 20 pixels up and 20 pixels right of the center pixel, the radius would be consider 20. Note that the pixels considered for attribute generation always had an equal shift vertically and horizontally, thus resulting in only four directions for analysis (see Figure 2). These directions were chosen in order to avoid quick intersection with neighboring spots.

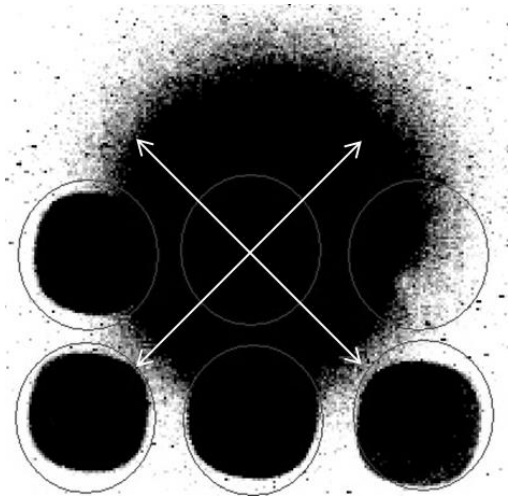

Figure 2. Example of halo and the radial directions

The attribute space was chosen based on the differences in pixel intensity transition in the regions just outside the grid circles of the spots for the different types of spots. Figures 3, 4, and 5 show changes in pixel intensities of the regions outside individual spots as the pixels move further away from the center. It was believed that creating an attribute set that would evaluate the strength and direction of the relationship between these two variables, along with some other values, would allow for the differentiation of the spot types.

The following is a description of each attribute:

Correlation- The value of the correlation attribute was obtained by computing Pearson's correlation of the radius of a pixel against its intensity. Beginning at a radius of 21, the pixel intensity was acquired for each radius up to 40 creating 20 ordered pairs: (radius r, pixel intensity of pixel at radius r). The correlation coefficient was computed using these ordered pairs. Since there are four different radial directions, four sets of 20 ordered pairs were generated, resulting in 4 correlation attributes: upper right, upper left, bottom right, and bottom left correlations. In addition, the mean of the upper correlations and the mean of all the correlations were computed and served as two additional attributes, resulting in a final total of six correlation

related attributes. These attributes were chosen because it was believed that halo spots would have linearly decreasing pixel intensities as the pixels moved further away from the center of the spot while for other spot types this relation would have little correlation.

Slope- Using the ordered pairs generated for the correlation attributes, the slopes of two lines of best-fit were acquired for each radial direction. The first slope was generated by computing the slope of the best fit line using ordered pairs: (radius r, pixel intensity at radius r) where r ranged between 21 and 30. The second slope was generated by computing the slope of the best fit line using ordered pairs: (radius r, pixel intensity at radius r) where r ranged between 31 and 40. For each pair of slopes, the ratio between the two was computed (note that a value of .0001 was added to the slopes to account for the case in which the second slope was equal to 0). These three measures were computed for each of the considered directions resulting in 12 slope related attributes. These attributes were chosen to distinguish comets from halos when correlation failed to do so. The difference between the two slopes of comets is predicted to be greater than those of other spots because the comets should show a drastic decrease in pixel intensity once the radius moves out of the comet strip.

Cutoff Radius- Beginning at a radius of 0 (the center pixel), the pixel intensity of pixels were acquired until either the pixel intensity dropped below 1000, or the radius reached 40. If the pixel intensity dropped below 1000, the radius at which this occurred was recorded. 1000 was chosen as the stopping intensity because it was an intensity that would stop approximately at the edge of a normal spot for most images. If the radius reached 40 before the intensity dropped below 1000, then 40 was recorded. 40 was chosen as the maximum radius in order to prevent the search from traveling into another spot. The cutoff radius was recorded for each of the four considered directions of pixel shifts resulting in four attributes. In addition, the means and medians of the four direction's cutoff radii were computed creating two additional attributes for a final total of six cutoff radius attributes. These attributes were chosen because it was believed that halos would have higher intensities outside the grid circle and thus would result in larger cutoff radii.

Pixel Intensity- For each of the four considered directions, the pixel intensity of the pixel at a radius of 40 from the center of each spot was obtained and served as an attribute. Each of these values was also multiplied by a scaling factor described below. The pixel intensities and the scaled pixel intensities account for 8 total attributes. These attributes were chosen for the same reason as the cutoff radius; halos should have higher pixel intensities outside the grid circle.

Background Intensity- This attribute is the only value generated in Array-Pro that is used for the purpose of classification. It is computed by taking the mean pixel intensity of a two pixel thick ring around the grid circle of each spot. It accounts for one attribute of the statistical attribute set. Again, this attribute was chosen because halos should have higher pixel intensities outside the grid circle.

Whenever pixel intensity was extracted from the image, its value was scaled by a factor of 247 divided by the median intensity of the image to adjust for overall intensity differences between the images. All slopes, correlations, and cutoff radii were determined using intensity values that were adjusted by the scaling factor.

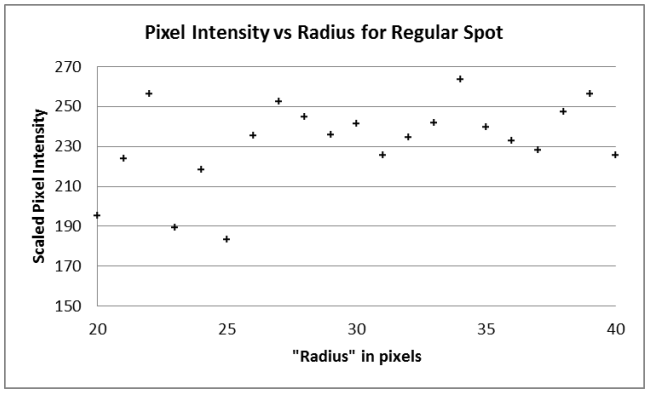

Figure 3. Plot of pixel intensities along radial direction for a regular spot. The plot shows a weak, positive relationship between the two variables.

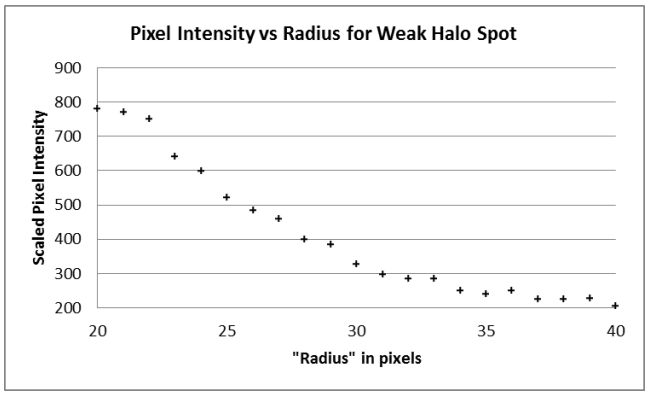

Figure 4. Plot of pixel intensities along radial direction for a spot with a weak halo. The plot shows a strong, negative relationship between the two variables.

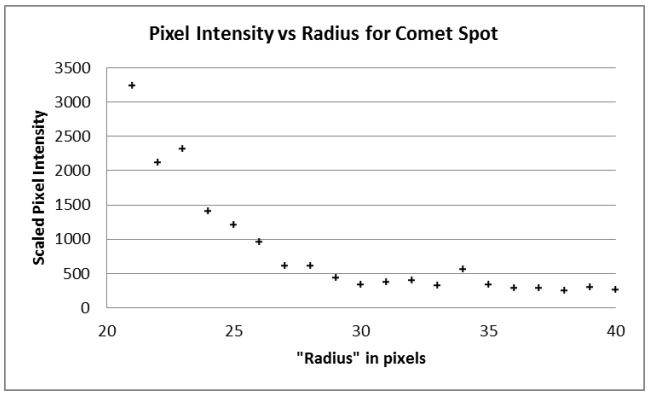

Figure 5. Plot of pixel intensities along radial direction for a spot demonstrating the comet effect. The plot shows a relatively strong, negative relationship between the two variables; however, the descent of the first ten pixels is much sharper than that of the weak halo. The pixel intensity value for the pixel located at radius of 20 is approximately 30800.

2.2.2 Pixel Intensity Attribute Set

This attribute set contains 4225 attributes, each representing a pixel intensity of a different location. The attributes were generated by taking a 65 pixel by 65 pixel box around the center of each grid circle and obtaining the pixel intensities for each pixel within the box. This was done for each spot on the microarray image. The dimensions of the box were chosen to contain regions both outside and inside each spot without cutting into other spots. The intensity values were scaled by the factor described in the previous section.

2.3 Model Building and Classification

In order to classify the spots on each microarray image, it was necessary to build a training data set and a test data set that would be used to generate a model and evaluate the performance of the model, respectively.

2.3.1 Training and Test Data Sets

A total of 88 images were available to train and test on. 23 of these images were randomly selected to be used for training data. Within each of these images, all (or almost all) of the halos and weak halos were classified by eye and added to the training data set. Due to a heavily skewed class distribution, comets, regular spots, and blanks spots were arbitrarily classified by eye in quantities that would keep the class distribution of the training data set relatively even. Only 20 blanks were used because we were confident that the classifier would be able to handle this class even with small distributions. As there were two attribute sets, there were also two training data sets, each one corresponding to one of the attribute sets. The statistical attribute training set contained 493 spots in total: 22 blank spots, 142 comets, 101 halos, 145 regular spots, and 83 weak halos. The pixel intensity attribute set contained 491 spots in total: 22 blanks, 141 comets, 101 halos, 137 regular spots, and 90 weak halos. A few values from each set were removed in order to improve model performance.

58 of the remaining images were used for the validation set. Just as in the training images, all (or almost all) of the halos and weak halos within the images were classified by eye and added to the test data set. For the purpose of this project, it was not necessary to classify any other type of spot specifically. All spots that were not classified as halos or weak halos were classified as non-halos. Again, a test data set was built for each of the attribute sets. Both sets contained 287 halos, 225 weak halos, and 140611 non-halos.

7 images were neither used in the training nor test set as they were of poor quality. Poor quality images include those with unclear spot visibility and scattering of high intensity regions in the background of the slide. These images would also likely be excluded from traditional analysis of NAPPA array data.

2.3.2 Model Building

In order to build the model used for classification, several iterations of various Weka classifiers using five-fold cross-validation were performed.

The final model for the statistical training set was built using a cost sensitive RandomForest classifier [3] that increased the cost of classifying weak halos as comets by a factor of five, the cost of classifying comets as weak halos by a factor of ten and regular spots as weak halos by a factor of 20. The number of trees within the ensemble was 200 and 6 random attributes were compared at each split of the tree building process.

The final model for the pixel intensity training set was built using a cost sensitive RandomForest classifier that increased the cost of classifying comets as weak halos and regular spots as weak halos by a factor of ten. The number of trees within the ensemble was 200 and 13 random attributes were compared at each split of the tree building process. Cost sensitive analysis was used in both cases in order to improve the precision when the models were applied to the validation set.

2.3.3 Classification and Correction

Using the Weka API that was incorporated into our program, we applied each of the models to their corresponding test data set. After the classification was completed, correctional post-processing was performed in order to improve the results.

For the statistical test set, three rules incorporated into our program were used to reclassify certain spots. The rules were performed in the following order:

1. If the mean upper correlation value was less than -.9 and the spot was not classified as a halo or a weak halo, it would be reclassified as a weak halo. This correction was performed because more often than not, if a spot had a mean upper correlation value of less than -.9, then it was a halo or weak halo. We chose to reclassify them as weak halos because the model was better at predicting halos.

2. If the spot was classified as a weak halo and had two or more comets located in a range of 7 spots below and above it, it would be reclassified as a comet. This correction was performed since halos and weak halos almost never appeared within a comet strip.

3. If the mean upper correlation was greater than -.6 and the spot was classified as a weak halo, it would be reclassified as a regular spot. This correction was performed to increase the precision of our model since most weak halos had mean upper correlation values less than -.6 .

For the pixel intensity set, since it was not classified using any correlation attributes, only the second correction could be performed.

3. RESULTS

3.1 Performance of Statistical Attribute Model

Tables 1 and 2 show the results of the confusion matrix of the five-fold cross validation of the training data and the confusion matrix for the results of the classification and correction on the test set. Misclassifications of weak halos have the highest relative frequency in both validations. Precision and recall of the test validation were calculated based on correct classification of halos (both types) versus non-halos. For the cross-validation, Weka determined precision and recall by computing the weighted averages of these performance metrics for all classes. Precision loss in the test set validation is due to high frequency of misclassifications of non-halos as weak halos. The time it takes to generate the attributes and classify and correct the spots using this method is 1.65 seconds per image. Table 3 shows the performance statistics for both validations.

Table 1. Confusion Matrix for 5-fold Cross Validation (Stats Set)

Actual Class↓	Pred. Val.→	Blank	Comet	Halo	Reg.	Weak
Blank		21	0	0	1	0
Comet		0	132	0	8	2
Halo		0	2	84	0	15
Regular		1	8	0	136	0
Weak Halo		0	15	5	3	60

Table 2. Confusion Matrix for Test Results (Stats Set)

Actual Class↓	Pred. Val.→	Halo	Weak Halo	Non-Halo
Halo		230	41	16
Weak Halo		9	139	77
Non-Halo		26	195	140390

Table 3. Performance Metrics of Statistical Model

Model Eval.	Recall or TPR	FPR	Precision	F-Measure
Cross Validation	.878	.04	.88	.877
Test Validation	.818	.0016	.655	.727

3.2 Performance of Pixel Intensity Attribute Model

Tables 4 and 5 show the results of the confusion matrix of the five-fold cross validation of the training data and the confusion matrix for the results of the classification and correction on the test set. As with the statistical attribute model, misclassifications of weak halos have the highest relative frequency in both validations. Precision loss in the test set validation is due to high frequency of both types of misclassifications of non-halos, though misclassifications as weak halos is still the higher of the two. The time it takes to generate the attributes and classify and correct the spots using this method is about 7 minutes per image. Table 6 shows the performance statistics for both validations.

4. DISCUSSION

Although cross-validation on the training data sets suggested that the pixel intensity model would be the superior of the two, evaluation of the two models' performance on the test data sets shows that the statistical model outperforms the pixel intensity model on almost every level. While both models demonstrate relatively high recall, the precision of the pixel intensity model is much too low to serve the desired function, as reflected in its low

Table 4. Confusion Matrix for 5-fold Cross Validation (Pixel Intensity Set)

Actual Class↓	Pred. Val.→	Blank	Comet	Halo	Reg.	Weak
Blank		21	0	0	1	0
Comet		0	134	0	5	2
Halo		0	3	89	0	9
Regular		0	5	1	131	0
Weak Halo		0	8	12	4	66

Table 5. Confusion Matrix for Test Results (Pixel Intensity Set)

Actual Class↓	Pred. Val.→	Halo	Weak Halo	Non-Halo
Halo		246	23	18
Weak Halo		19	149	57
Non-Halo		177	771	139663

Table 6. Performance Metrics of Pixel Intensity Model

Model Eval.	Recall or TPR	FPR	Precision	F-Measure
Cross Validation	.898	.033	.897	.896
Test Validation	.854	.007	.316	.461

f-measure. On the other hand, the f-measure for the statistical model is acceptable given the task of classification amongst thousands of spots. Dividing the number of non-halos classified as halos by the number of images in the test set suggests that there will only be 4 spots examined unnecessarily per image. Furthermore, since it takes less than 2 seconds to generate the attributes and classify the spots, this method would save the users at least 10 minutes per image in time spent trying to find the halos and classify them themselves.

Misclassification of weak halos as halos and vice versa is not of great concern because we anticipate that statistical analyses will be based on the presence or absence of a halo. Misclassification of halos and weak halos as comets and vice versa is substantially more important. However, as comets are technical artifacts, we anticipate that changes in NAPPA protocols will reduce the prevalence of comets on the arrays, which will in turn enable improved classification.

5. CONCLUSION

It can be concluded that the halo classification task can be successfully automated using our software with acceptable recall and precision using the statistical attributes.

In the future, we look to improve the performance of the model specifically in the area of classifying weak halos. We believe that the difficulty that the model has in classifying weak halos can be explained by the similarities of the space surrounding weak halos with those of comets. The use of divergent field gradient response and inclusion of new attributes such as S/N ratio may be possible solutions for enhancing our process of halo detection. In addition, we would like to develop metrics that can quantify the size and intensity of each halo and include an evaluation of the inter-rater variability of our halo analysts.

6. ACKNOWLEDGMENTS

This research was supported by Early Detection Research Network (NIH/NCI 7U01CA117374), the Virginia G. Piper Foundation, and start-up funds from the Department of Biomedical Informatics, Arizona State University.

7. REFERENCES

[1] Abràmoff, Michael D., Paulo J. Magalhães, and Sunanda J. Ram. "Image processing with ImageJ." *Biophotonics international* 11.7 (2004): 36-42.

[2] Anderson, Karen S., et al. "Protein microarray signature of autoantibody biomarkers for the early detection of breast cancer." *Journal of proteome research* 10.1 (2010): 85-96.

[3] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[4] Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD Explorations Newsletter* 11.1 (2009): 10-18.

[5] Ramachandran, Niroshan, et al. "Next-generation high-density self-assembling functional protein arrays." *Nature methods* 5.6 (2008): 535-538.