Analysis of a Proposed Universal Fingerprint Microarray

Michael Doran, Raffaella Settimi, Daniela Raicu, Jacob Furst
School of CTI, DePaul University, Chicago, IL

Mathew Schipma, Darrell Chandler
Bio-detection Technologies Center, Argonne National Laboratory, Chicago, IL

Abstract

Microarray technology allows biologists to test for a large number of DNA sequences with a single test. Each microarray may consist of hundreds or thousands of probes to test for specific sequences. These devices are typically developed to target a specific DNA type. A proposed universal microarray that uses a few hundred randomly selected probes is evaluated on a collection of micro-organism DNA; the collection consists of 25 closely related isolates that are used to test the limits of the device. A test based on statistical confidence intervals is used to explore the hypothesis that the microarray contains enough information to discriminate between isolates. Multiple replicates are combined to improve results, which leads to a potential method of raising the accuracy of the test to a desired target level. Work still in progress using machine learning algorithms demonstrates that closely related isolates can be identified based on information in the microarray data.

1 Introduction

The goal of the universal microarray is to use a standard set of probes to allow for fine grained classification of microbial isolates. Prior work has shown that this device is able to distinguish between families of isolates but has limited ability when applied to strains that are closely related [1]. *Bacillus anthracis* strains are of particular interest in the context of demonstrating the ability of a universal microarray since it is known as one of the most genetically homogenous bacterial species. Since this microbe causes anthrax, the ability to distinguish between strains from different origins is interesting because of the potential forensic applications related to the identification of the source of a bio-terror threat. To improve the device's ability to detect small differences, such as those between *B. anthracis* strains, a different gel based microarray technology was adopted to provide more accurate readings.

The task of processing the microarray data can be broken down into several basic steps. First, the images are processed and summarized as probe level data. Then, a statistical procedure is needed for determining the equivalence of a new sample to a fingerprint. An equivalence test should provide information at a specified level of confidence about how many probes may be significantly different. Ideally this would give an exact answer to whether or not the isolates are identical. Because of the amount of variance in the data and the relatively small number of replicates available in relation to the number of probes we are not able to answer this question with a high confidence

level. Using confidence interval based technique proposed in this paper we are able to demonstrate that there is enough information in the images of the microarrays to correctly identify isolates more than 70% of the time based on the assumption that the sample data is a representative subset of the readings that would be obtained from other isolates from additional repetitions and strains.

The obtained classification results are well below a perfect classification rate. However, these results are still of interest because the isolates studied are known to be very hard to distinguish, so this test is pushing the limits of what is possible today. Additional replicates are shown to improve the classification accuracy so an experimenter is able improve results by adding data. Since this is one of the hardest tests set to work with the device will probably provide much better results with other isolates.

2 Data Set Description and Preprocessing

The microarray consists of 4 blocks each of size 10 by 10 probes providing 400 readings per test; 390 of the 400 probes contain actual probes, the others are control spots used to register the microarray images and provide extra information for intensity normalization across slides. The Automated Microarray Image Analysis Toolbox for Matlab [2] is used to determine the average foreground and background intensity for each probe. A single value computed as the log of the ratio of the foreground to background is used as the intensity of the probe following a standard practice in the industry [3][4] and prior work with this chip [1].

Image normalization is usually required to adjust for systemic differences between microarray images [5]. Normalization algorithms that transform the data by fitting a curve through the points are not appropriate for small data sets since there are not enough data points to minimize the effect of outliers, especially at the high end of the intensity spectrum (where the variance in the data increases greatly even after the log transform of the ratio between foreground (probes) and background. Therefore, quantile normalization [6] has been chosen to transform each microarray's intensity distribution to a template (Figure 1).

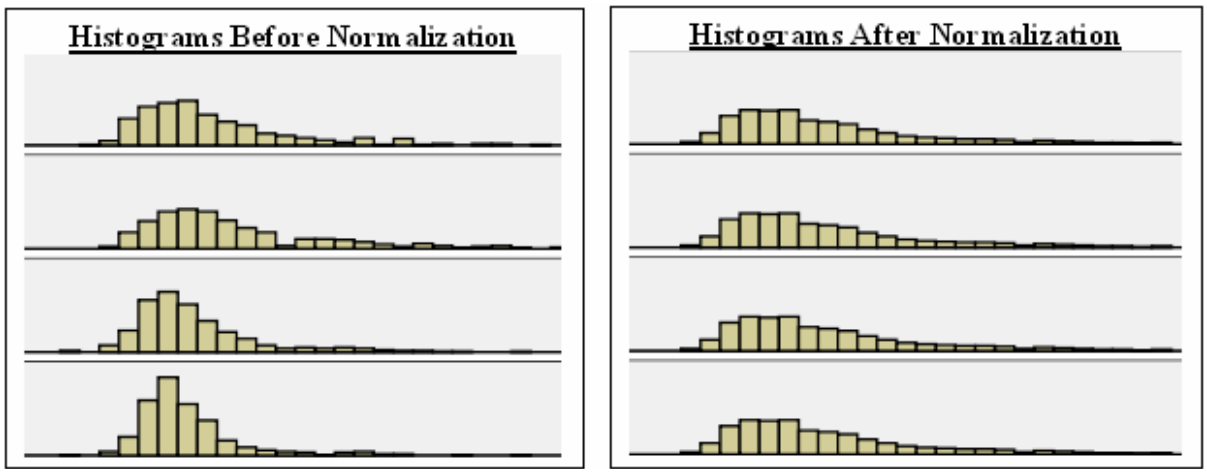

Figure 1: Quantile normalization for one of the probes

If very different isolates are used this may not be the best approach, but it is a reasonable assumption when all of the isolates are extremely closely related. For the n^{th} quantile of each set of probes, each intensity value is set to the median value of all the intensities and thus, this approach does not require the selection of a base image to normalize all other data to.. As an example of the effect experimental factors may have on the data, at a 95% confidence level using the Bonferonni correction for multiple hypotheses testing, the number of significant probes between isolates 2 and 3 decreased from 52 to 5 after normalization. While it is not known how many probes should actually hybridize at significantly different levels for these isolates, the anticipated numbers is relatively small. Prior work suggests that there may be approximately 3 significantly different probes [1]. The new chips use improved materials, which may provide the ability to detect more differences.

3 Statistical Test of Profile Equivalence

The task of comparing a test isolate to a library fingerprint can be approached statistically as a series of parallel hypotheses tests. This procedure follows the approach from Allan Willse [1]. The statistical test for each probe is the decision between:

H_0: the probe intensities for the test and library fingerprints are the same
H_a: the probe intensities are different

A significance level α is selected that specifies the confidence level of the test result. Selection of an appropriate α depends on knowledge of the target application. For a single test a probe where $p < \alpha$ would be selected as significantly different. For $\alpha = 0.05$ and a test that includes 390 probes, the expected number of probes where $p < \alpha$ when the intensities are the same is around 19.5 (390 * 0.05). This is problematic when working with a data set where the goal is to distinguish between isolates with fewer than 10 significant differences. A procedure for working with multiple hypothesis testing is needed to correct for Type I errors (the likelihood of rejecting H_0 when H_0 is in fact true [7]; type II errors (the possibility of not rejecting H_0 when H_a is true) also clearly hurt the effectiveness of the test. Benjamini and Hochberg's False Discovery Rate (FDR) method provides a technique that allows for the regulation of the proportion of Type I errors among the rejected hypothesis [8]. The FDR method is a good choice for the microarray application because it is more likely to avoid Type II errors while still providing control of Type I errors compared to approaches like the Bonferonni method. The complete resulting profile equivalence between a sample and a library finger print will consist of the confidence level α, the multiple hypothesis control method, the number of significantly different probes and a list of probes with an associated p values.

4 Confidence Interval Based Classification

A simple classification algorithm was developed to categorize test cases and provide a base line for evaluating several machine learning algorithms used for microarray data classification.

Our classification procedure defines *an isolate fingerprint as a set of confidence intervals (C.I.) for ratio intensity values for each spot in the micro-array.*

Let $\{Y_{1jk},\ldots,Y_{Njk}\}$ be the average spot intensities for N spots on a micro-array where DNA fragments from a microorganism j, for $j=1,\ldots,J$, are hybridized. We assume that the hybridization experiment is replicated K times for each bug, and the index $k=1,\ldots,K$ identifies the k-th replicate..

Let μ_{ij} be the average spot intensity level for isolate $j=1,\ldots J$ at spot $i=1,\ldots N$. For any given microorganism j, we define its fingerprint for the probe hybridization levels as the set of N confidence intervals for the average spot intensity levels μ_{ij}, $i=1,\ldots, N$.

The $(1-\alpha)$% confidence interval for μ_{ij} is defined as

$$\left(\bar{y}_{ij} - t_{(1-\alpha/2,K-1)} s_{ij}/(K-1), \bar{y}_{ij} + t_{(1-\alpha/2,K-1)} s_{ij}/(K-1)\right)$$

where K is the number of replicates, $\bar{y}_{ij}$ is the sample average intensity and s_{ij} is the sample standard deviation of the intensity at spot i for isolate j computed from the K intensity levels. Thus for each isolate, the $(1-\alpha)$% confidence intervals for the N probes are used to define the isolate fingerprint.

Whenever DNA fragments from a certain isolate are hybridized on an array, the observed spot intensity levels are expected to lie within the confidence interval based fingerprint of that specific isolate. The C.I. fingerprint classifier is implemented by comparing the observed spot intensity levels of a tested microorganism with the fingerprint confidence intervals for a certain isolate. We will say that there is a *positive sub-match* between the tested microorganism and an isolate j, if the observed spot intensity level of a certain probe lies within the fingerprint C.I. corresponding to that probe for isolate j. The classifier counts the number of positive sub-matches between a tested microorganism and the isolate j.

We can therefore define a similarity metric between the tested microorganism and an isolate fingerprint, as the number of positive sub-matches. A very high similarity score with bug j fingerprint suggests that the tested microorganism belongs to the bug j strain. Thus, for two isolates A and B, the classification producing the highest number of sub-matches will identify whether a bug is of type A or B.

Using a data set with a diverse set of isolates (26 isolates, each replicated 9 times), this technique was able to correctly classify 73% of the samples. Closely related samples where classified poorly, on average 55% of the time, but the incorrectly classified samples where almost always confused with the related sample.

Averaging test samples together was shown to have a large positive impact on the classification rate. Intuitively this makes sense because averaging generates a test case with very few outliers so most probe intensities should fall in the expected range. This is important because it means that results can almost always be improved by adding additional replicates.

5 Conclusions

At this point we are able to demonstrate that there is enough information in the data to classify extremely closely related isolates using the universal microarray. Additional work is needed to verify that classification results are a side effect of

information in the data as opposed to side effects of systemic experimental factors. Finally, we are working on implementing the classification and profiling algorithms so that they are directly accessibly by anyone interested in examining the data. This implementation can also serve as a starting point for analyzing future results from similar microarrays.

References

[1] Alan Willse, Darrell P. Chandler, Amanda White, Miroslava Protic, Don S. Daly, and Sharon Wunschel (2005) "Comparing Bacterial DNA Microarray Fingerprints", *Statistical Applications in Genetics and Molecular Biology*: Vol. 4: No. 1, Article 19.

[2] Amanda M. White, Don S. Daly, Alan R. Willse, Miroslava Protic, Darrell P. Chandler: Automated Microarray Image Analysis Toolbox for MATLAB. Bioinformatics 21(17): 3578-3579 (2005)

[3] Allison, Page, Beasley and Edwards. "DNA Microarrays and Related Genomic Techniques: Design, Analysis and Interpretation of Experiments". Chapman & Hall CRC. 2006.

[4] Eisen, M.B, Spellman, P.T., Brown, P.O. and Botstein, D., *Cluster analysis and display of genome-wide expression patterns,* Proceedings of the National Academy of Sciences USA 95, 14863-14868 (1998).

[5] Lee MLT, Kuo FC, Whitmore GA, Sklar J, *Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations*. Proceedings of the National Academy of Sciences USA 97: 9834-9839 (2000).

[6] Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance*. Bioinformatics 19(2):185-193

[7] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick, "Multiple Hypothesis Testing in Microarray Experiments" (August 2002). *U.C. Berkeley Division of Biostatistics Working Paper Series.*

[8] Benjamini, Y.and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289-300.