Mapping Binary (On/Off) Gene Expression for Pathway & Tissue Analysis

A THESIS
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Raymond M Moore

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Jean-Pierre Kocher, Claudia Neuhauser

August 2012

Acknowledgements

Jean-Pierre Kocher
Chen Wang
Claudia Neuhauser
Hugues Sicotte
Yuji Zhang

Special Thanks to:
Matthew McCall

Dedication

This thesis is dedicated to Caitlin & my Parents.

Abstract

Microarray data contain information about the level of expression of genes that can be informative of changes taking place in cells. This information has been widely used to study the changes in gene expression between normal and cancer cells. Gene expression has been used as a biomarker predictive of the progression of a disease and to identify drug targets specifically expressed in cancer tissue. Although the level of expression of a gene can change by multiple folds across tissues, the simplest information (on/off status) about a gene is whether it is or is not expressed in a given tissue.

In this study we propose to take advantage of a large set of tissue specific gene expression data to study the profile of gene expression in pathways and tissues. To perform this analysis, we leverage and improve a method that computes the on/off status of a gene from their level of expression.

The percent of genes with on status in a given tissue was selected to summarize across bio specimen. The gene state method was applied to sets of tissue specific expression microarray extracted from the GEO database. We then studied the profile of on/off state of genes in KEGG pathways across several tissues. The data were then used to calculate a distance between gene sets. Using all genes, a distance could be calculated between normal and cancer tissue, as well as pairwise comparisons between each tissue type. The gene sets were then narrowed and selected based on pathway annotation from KEGG. This demonstrated an ability to identify known cancer pathways based on their gene signature distances. The results affirm known cancer pathways by calculating relative distances.

Table of Contents

List of Tables

List of Figures

Introduction

Goals of the Project

Gene expression has been used as a biomarker in various studies to predict the progression of a disease. It has also been used to identify drug targets in cancer tissue. (Parmigiani, 2002) Expression data is usually acquired through microarray, which provides information about the level of expression by measuring the amount of messenger ribonucleic acid (mRNA). RNA translates to protein, which is the functional mechanism in a cell. Measuring the level of RNA can provide insight into the changes taking place in a cell. The experiment entailed finding a method to calculate the on/off gene state from gene expression data. Population studies, which investigate cellular changes across a population, could utilize this approach which can provide data by viewing all pathways or tissues simultaneously

The level of expression of a gene varies greatly across tissues and even within tissues. From a functional point of view, the most basic information about a gene is whether or not it is expressed in a given tissue, which can be summarized as two states: on or off. Reducing gene expression into on and off is important because it simplifies the data. The Boolean description of the on or off state of a gene yields a straightforward interpretation. The gene states reflect the biological mechanism of a gene being on and used or not and unused. The modeling of biological networks into Boolean "control circuits" was originally developed by Kauffman (1969). Despite possible oversimplification, the information obtained from this Boolean approach could reveal insights into relationships between genes, corresponding pathways and across tissue types that would be otherwise hard to interpret (Kauffman, 1973).

We discuss the implementation of a specific method that enabled us to model Boolean networks. The primary aim of this work was the assembly of a workflow to process gene expression data from microarrays (Affymetrix) and convert the level of expression into the on/off state of a gene. This workflow included several applications that had been previously developed. However these applications could not be readily assembled without significant modifications.

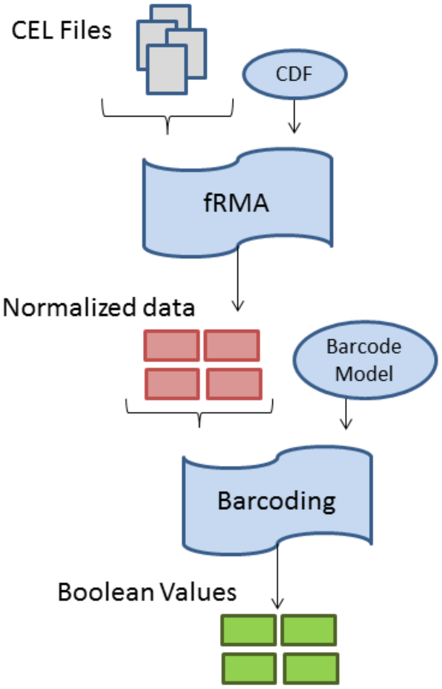


Figure 1: Workflow Overview. Each sample is normalized by fRMA, then applied to the barcoding algorithm, which utilizes the previously created model to calculate Boolean values.

This workflow was designed to process a large number of microarrays from different tissues and different experiments. Therefore, the critical first step was data normalization to adjust the distribution of gene expression to be comparable. This step was accomplished by fRMA (McCall, Irizarry 2011), a recently developed method that allowed for the sequential normalization (preferably by batch) of microarray data. The normalized data are then processed by an application called “Barcode” (McCall, Uppal, Jaffee, Zilliox, Irizarry, 2011) that converts level of expression into on/off state of a gene. By default, these two applications use a standard CDF (cel definition file) annotation file provided by

Affymetrix that relates each probe to a probe set. (A probe set is a collection of probes intended to target specific sequences in a gene.) Since the knowledge of the human genome is constantly expanding, the CDF provided by Affymetrix has become outdated. A newly updated CDF was proposed by several research groups, to address the changes from the older version. (Wang, Verhaak, Purdom, Spellman, Speed, 2011)

The work performed for this thesis includes the 3 following specific aims:

1. Enable fRMA and the “Barcode” application to process data with the new CDF provided by Brain Array lab. (Thompson, Fan, 2011)
2. Build these applications into a workflow and interface to the BORA repository system developed by the Bioinformatics Core at Mayo Clinic.
3. Describe the behavior of the on/off state of gene in pathways across different tissue types.

Introduction to Microarray Technology and its Application in Gene Expression

Microarray is a technique that captures the level of expression of genes on a genome wide scale by measuring the quantity of RNA molecules bound to their complement. The technique enables fast processing of ribonucleic acid (RNA) molecules, indicating the level of gene expression. Microarray technology creates a snapshot of the level of expression for thousands of genes in a biospecimen. The process of microarray can be summarized into four steps: obtain a selection of RNA molecules from a sample, select a manufactured chip and perform the experiment, preprocess the raw intensity data, then analyze the data according to the experimental type.

Microarray Preparation

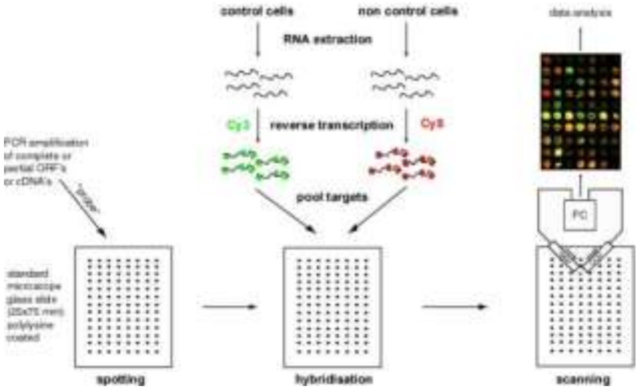


Figure 2: The process of creating a microarray chip can really be simplified down to three basis steps. Spotting, where the probes are arranged on the chip. Hybridization, where the sample RNA is applied to the chip and allowed to bind to its compliment. Scanning, where the fluorescence is measured to identify the amount of RNA bound to the chip. (Eichinger, 2009)

There are two methods for manufacturing microarray chips: DNA fragment deposition and *in situ* synthesis (Zhang, 2006). One of the best known manufacturers is Affymetrix. In the manufacture of Affymetrix chips two techniques are required: photolithography & solid-phase DNA synthesis. The process begins with photolithographic masks that allow light to react only over specific cells on the chip. The open spots in the mask allow light to convert the protective group on the terminal nucleotide into a hydroxyl group. The hydroxyl group is reactive enough to create a new bond to the next nucleotide in the sequence. The mask changes each round and is designed to correspond to all the cells having the same nucleotide added. It takes many rounds of masking and adding the nucleotides to create a chip with thousands of individual sequences called “probes.” Microarray preparation begins with the extraction of RNA. Then the collections of RNA molecules are applied to the microarray chip. The microarray chip containing RNA molecules are inserted into a machine to be read.

Design of an Affymetrix Chip

Two types of probes exist on an Affymetrix chip; reference probes (or perfect match) and mismatch probes. During the hybridization step the DNA probes on the microarray will form heteroduplexes with their target RNA due to the Watson-Crick base pairing rule. (Zhang, 2006) Fluorophores are added to measure the level of fluorescence, which is proportional to the amount of heteroduplexes. The reference probes (PM) are specifically targeting a known sequence in the genome within an expected gene. The mismatch probes (MM) have a single nucleotide that is altered intentionally. The altered nucleotide is meant to detect sequences that have cross-hybridized. Cross-hybridization happens when a DNA/RNA sequence binds to a probe that is similar but not its perfect complement, even with a single nucleotide difference (Zhang, 2006). RNA hybridization can be influenced by many conditions including temperature, humidity, salt, target solution volume and the sequence itself (Zhang, 2006). Including mismatch probes on a chip is intended to be a strategy for detecting noise generated by the non-specific binding. The concept was that low expressed probes may be entirely made up of the noise or baseline and thus, unreliable. Additionally, some base line is expected for every probe. However, the usefulness of the mismatch probe is degraded by the fact that it is not necessarily removing noise, but potentially, viable information as well. This being understood, few normalization techniques continue to utilize the mismatch probes.

Affymetrix Standard File to Store Gene Expression Results

The dense intensity data is stored in files that have the extension cel and adhere to a standard format (".CEL File Extension" 2011). The cel file stores the calculated numeric interpretation from the digital image of the features on the microarray chip. The core of the file is made up of the calculated feature intensity, which is obtained from pixel values of the image file created by digitally reading the light signatures in the microarray reading machine. In addition to the feature data, there are various other values to describe the microarray including flags, outliers, masked features, sub-grids and records of the algorithms used to calculate the data (".CEL File Extension" 2011). Every value stored in a cel file comes from a single feature, which represents one probe. The information about a single probe is not informative. The microarray chip is designed to contain multiple features to statistically derive useful information about a gene. The cel file is loaded into a normalization algorithm.

Preprocessing of Gene Expression Data Obtained from Microarray

Background on Normalization Methods

Normalization, also known as data transformation or preprocessing, is the method of converting the microarray raw intensities into biologically interpretable signals. Normalization ensures constant variability at all intensity levels and eliminates experimental bias (Zhang). In addition, Microarray can produce values that contain noise specific to the facility and version, which also requires consideration of normalization procedures. The most widely used techniques are robust multi-array analysis (RMA), GeneChip RMA (gcRMA), Model Based Expression Index. (MBEI), and Probe Logarithmic Intensity Error (PLIER), all of which are referred to as multi-array analysis (McCall, Irizarry 2011).

The Multiple Array Normalization Method

Robust multi-array average (RMA) employs log-transformed PM values, and adjusts those values based on a global background reference computed across all the included arrays (Irizarry, Hobbs, Collin, Beazer-Barclay, Antonellis, Scherf, et al. 2003). It is a well-established methodology that has been shown to be more effective than techniques that include MM values (Irizarry, 2003). The RMA method has an alteration to allow for iterative processing without the need for re-analyzing every array when a new one is added, known as frozen Robust Multi-array Analysis (fRMA). (McCall, Irizarry 2011). The premise of fRMA is to calculate a set of parameters that are essentially frozen, from a very large curated database of microarrays.

Background on Application of On/Off State of a Gene

Initally, the developers of Barcode came up with two applications; identifying orthologous genes via phylogenetic tissue analysis and determining an unknown tissue or cell type through gene expression. The first application consisted of choosing tissues from mouse and human samples, and looking at clustering of expression values and barcoded values. The results showed that the expression values clustered according to species, while the barcoded values successfully identified orthologous tissues. (McCall, 2011) The other application employed by the Barcode group was to use barcoded values to identify tissue or cell type of an unknown sample. Searching through a library of cell types previously barcoded and matching the unknown sample to the most likely cell type. The result of the application was very successful in matching cell or tissue types by their gene expression barcoded values (McCall, 2011).

The Argument for Informative Pathway Level Expression Analysis

The Barcode group demonstrated a set of applications for Boolean values, however, there is another group investigating pathway analysis techniques that could benefit from gene state information. A literature search revealed a group using normalized gene intensities that could be investigated further with Boolean data. The technique employed was built on the concept of existing pathway networks, a predefined set of genes selected from other methods. The method used was a three step strategy of obtaining a predefined set of genes, calculating the activity of that pathway and then identifying the differentially expressed pathways. The results of the research were made available via a webserver. The differential analysis includes an evaluation score based on randomly permutated sample labels. (Tomfhor, 2005)

Affymetrix Present/Absent Call

There exist many approaches of categorizing probe intensities; the quality function provided by the microarray company Affymetrix is the present/absent score. The quality measure was specifically designed as a quality metric and was not meant to contain biologically relevant information. This score was also provided by our automated workflows, meaning that it was readily available along with the initial data. It was proposed that perhaps this easily obtained score had some usefulness in identifying the gene expressions that are noise (absent) or actual (present). A simple way to determine if the score was a viable candidate was to demonstrate a separation of absent and present across expression values. Our analysis indicates that there was no distinct separation between samples identified as absent and those identified as present, thus confirming that it was indeed a poor candidate to identify on/off.

Background

Origin and Retrieval of Data

Gene expression microarrays can add up to a massive amount of data. There is a data management system available at Mayo Clinic called BORA. BORA stands for Biological Oriented Repository Architecture which is a system designed to integrate multiple genomic data to enable integrative analysis. It is intended to be used by researchers to integrate multiple data types to maximize information cross-analysis. The data already loaded into BORA includes normal, cancer and various disease conditions across more than a hundred tissue descriptions. The tissue terminology was collected by experts within the Mayo Clinic group as well as my own text mining scripts.

The Need for Better Summarization, a New CDF

The Original Affymetrix CDF Annotation File

The information used to aggregate which probes belong together in a probe set is located in the Affymetrix chip definition file (CDF). This file is critical to the normalization step, because it defines which features on the chip belonged to which probe set. A probe set is intended to target a gene within the genome. ("CDF FILE" 2009) The target is determined by a selection based on a fully sequenced and annotated genome. However, the understanding of where genes are located is continually updated. As knowledge of the genome advances, it exposes the fact that some probes target the wrong gene.

New CDF: Probe Set Equals a Gene

Gene level summarization was initially chosen due to publications that suggested gene expression biomarker panels validated better with an updated definition of probe sets. The decision was based on the popularity in the research community. A validation was performed by looking at 246 Glioblastoma (GBM) & 175 Ovarian samples from the Cancer Genome Atlas (TCGA). (Wang, 2011) They used the digital sequence data as a basis of comparison against the three platforms available; Affymatrix U133, Exon, and Agilent. An updated definition removed ambiguity by rearranging the probes to represent the appropriate gene. Genes with long sequences would receive the greatest benefit. The better selectivity should also increase the ratio between signal and noise. Most often a probe set was selected to represent a gene from an array, whereas gene level summarization removed the process of selection and enforce the same microarray analysis. In addition to rearranging probes to their appropriate gene, some probes were simply removed from analysis based on poor performance. The conclusion was that gene level analysis outperformed the simple, probe set analysis for all three platforms. (Wang, 2011)

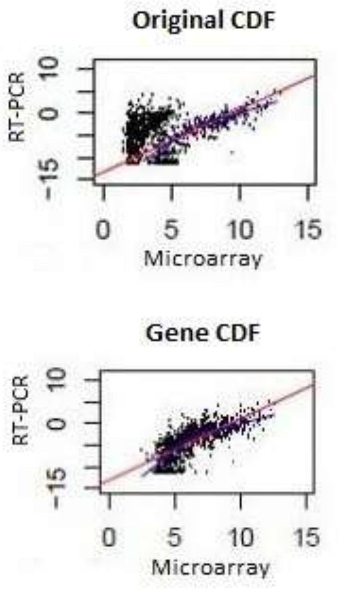


Figure 3: The better validation was what we were trying to show in the plots. Using a custom gene CDF also did significantly better than averaging multiple probe sets. The probes with high hybridization efficiency “rescued” the probes with low affinity based on the robust averaging method.

The data retrieved was already preprocessed through fRMA with an alternative CDF, mapped directly to genes. The two main advantages of the new Brain Array CDF were higher probe sensitivity and probes mapped based on the latest sequence knowledge. (Thompson, Fan 2011). Most of the remapping occurs where previous probes were incorrectly identified in the original CDF created nearly a decade ago. An additional motivation for gene level summarization, which evolved during the analysis, was into wanting to integrate gene expression data across platforms, since that is currently impossible.

Introduction to Barcode

The Equation to Creating the Model:

Barcode is a technique developed by McCall, Zilliox & Izzari (2011) to categorize a single gene expression as being 0 or 1 (akin to on or off) and thus creating a barcode effect in a given sample. The strategy entails a mixed model system, per gene (or probe set) specific to a microarray platform. The estimation process employs a variation of the Probability of Expression (POE) algorithm (Parmigiani, Garrett, Anbazhaghan, Gabrielson, 2002). The modified algorithm attempted to determine the proportion (1-p_g) of silenced (noise) normal N(μ_g, τ^2_g) distribution and expressed uniform

$$(y_{ijg} \mid \Theta_{jg}) \sim N(\Theta_{jg}, \sigma_g^2)$$
$$\Theta jg \mid \mu_g \sim (1 - p_g) \times N(\mu_g, \tau_g^2) + p_g \times U(\mu_g, S_g)$$
$$\mu_g \sim N(\xi, \lambda^2)$$
$$\tau_g^2 \sim IG(\alpha, \beta)$$

Equation 1: y_{ijg} is the observed $\log_2$ intensity value for the g^{th} gene in the i^{th} sample of j^{th} tissue type. The Θ_{jg} is the resulting mixed model of some proportion of a normal distribution N(μ_g, τ^2_g), considered the “silent” distribution of noise and some proportion of the uniform distribution of true expression U(μ_g, S_g).

$U(\mu_g, S_g)$ distribution from the silenced mean to the saturation point S_g, per gene. All gene expression intensity values, per tissue per sample are represented by y_{ijg}, the $\log_2$ intensity from fRMA, on gene (g) from sample (i) and tissue type (j) (Zilliox, Irizarry 2007). Thus the collection of Θ_{jg} is the average gene intensities per gene in a given tissue. It is assumed that the mean μ_g comes from a normal distribution $N(\xi, \lambda^2)$ on a global scale external to a specific gene, but rather based on biological constraints. Additionally, $\tau^2{}_g$ is proposed to exist from an inverse gamma distribution, in the same global sense to account for gene intensity outliers.

Estimation of the Noise/Off Distribution

Highly diverse microarray data were necessary to obtain good results. The EM algorithm employed estimates the off (noise) distribution. In general, most genes are not expressed in every tissue. There are some genes, often referred to as house-keeping genes that are always expressed. Accounting for these genes can be achieved by using a microarray study performed as a cross-hybridization experiment, to obtain expression values that are specifically background noise. Since it was not possible to estimate all of the parameters simultaneously, the programmatic structure consists of three parts. First, estimate a placeholder threshold via EM, leaving out higher level allowing the model to fit gene-by-gene. Second, estimate the highest/global level parameters (μ & τ) minus the off (noise) distribution. Third, re-estimate the parameters in the first step, using the hierarchical model including the highest/global parameters from the second step. This process then creates a model of parameters, which model the normal distributions that represent the noise otherwise known as off.

The Decision Boundary

It is from the resulting model, that the barcode value 0 or 1 (on or off respectively) could be calculated for a given probe set in a sample. The 1 or 0 value was obtained in multiple steps. The first step was to determine where the observed $\log_2$ intensity values lied in the noise distribution. Every probe set is partnered to an off (noise) distribution in the model. A p-value was calculated on the corresponding model's off (noise) distribution for that probe set. The p-value was converted into a Boolean value, by being either above or below a decision boundary. The decision boundary is a tunable parameter. The current decision boundary was calculated to be the number that provided the best performance in a leave-one-out categorization of an unknown tissue sample. The decision value was selected by the value which was able to correctly identify a random sample into its correct tissue/cell type.

$$z_g = \frac{y_g - \hat{\mu}_g}{\hat{\tau}_g}$$

$$b_g = \left\{ \frac{1\ if \quad z_g > C}{0 \quad otherwise} \right\}$$

Equation 2:
Calculating the Barcode value is performed as a logic bound to a constant (C). The value y_g is the observed gene intensity, μ_g & τ_g are the model mean and variance, respectively.

Methods

Barcode Methods & Script Descriptions

The available Barcode models within BioConductor (Gentleman, Carey, Bates, Bolstad, Dettling, Dudoit, 2004) were calculated from fRMA normalization with the CDF provided by Affymetrix, thus incompatible with our alternative CDF normalization. Thus it became necessary to contact the original developer for the ability to create new Barcode models that use an alternative CDF. Matt McCall, the original script developer, shared various pieces of Rscript code. The collection of scripts needed to be reviewed, connected and edited to fit into a more generalized format that utilized the parallel computational and archival resources available. The entire process was to understand and order each piece, then connect them and ensure the entire pipeline was error free.

Input Data Prep

The first three scripts rearranged and formatted the input data. These three scripts were provided by Matt McCall, and were specific to the original development system. Thus it was necessary to write new scripts to replace the original scripts. The data existed primarily as flat files. The first script copied the fRMA normalized expression data into a

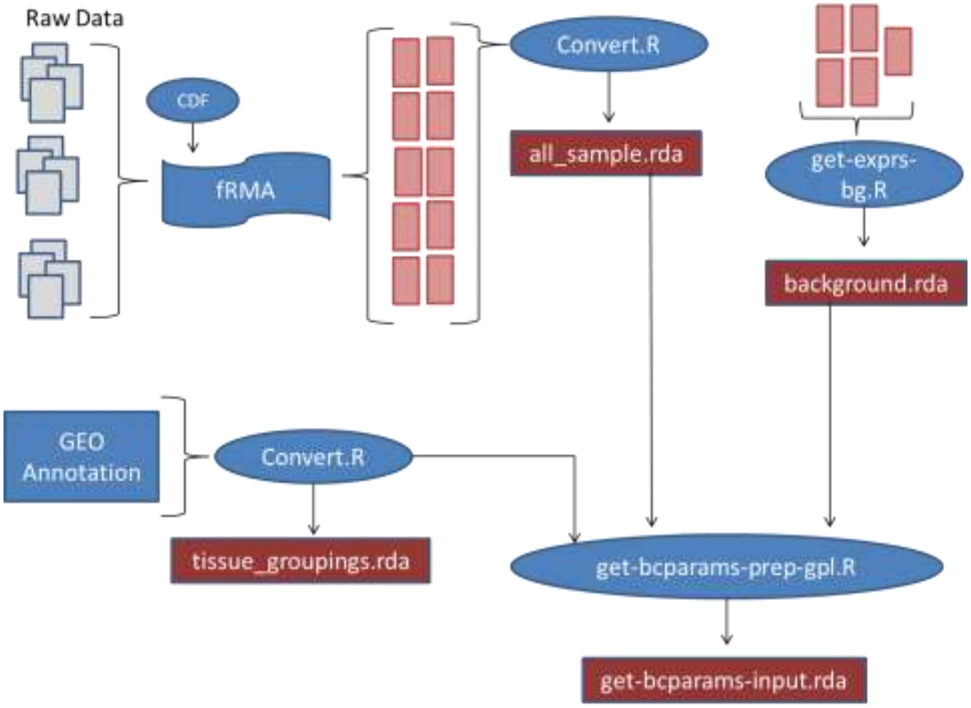


Figure 4: Some formatting is necessary to prepare the various inputs. All of the microarray samples must be normalized in the same way. They are they loaded into a R data structure. Additionally the background samples and the sample annotation are loaded into similar R data structures. The annotation contains the information about tissue and disease for each sample. The various data structures are then merged into a single input file.

single R structure, where the rows were the genes, labeled with Entrez Id, and the columns were samples, denoted by a GSM number. The script created a matrix style data frame contained within a single file named e.rda, which was the format necessary for the original algorithm. Since the algorithm is dependent on tissue type, a description of the sample's origin was required. The text based data was stored in a single flat file consisting of columns of information about the tissue type for each sample. The second Rscript was converted the flat file of text into a matrix data structure named tab-GPL.rda. The third and final piece of the input was the background expression. A Rscript nearly identical to the first compiled all of the background expression files into a single R data structure called ebg.rda.

The three input files previously described were the input for a single initiation script. Unlike the previous steps, this preparation script and most of the subsequent scripts were unaltered from the original. This script creates multiple R structures that contain the intersection between the input samples and the tissue descriptions. The results were R structures that ordered values of expression and tissue information, within a file labeled get-bcparams-input.rda.

Round 1: Estimate Distribution without Global Effect

The first round of estimation employed the EM algorithm. The EM algorithm is computer code for the Probability of Expression algorithm, designed by Parmigiani. Round 1 estimation was designed to make the per gene estimation without the global factors. This step involved an iterative optimization function used to obtain the most likely normal mean and variance values for the probe set $\log_2$ intensity. Like most optimization functions it is susceptible to local minima in the stochastic descent. From a computational perspective, any gene estimate could fail optimization. The code was originally written to detect failed optimization, in those cases the initial parameters were adjusted and rerun. The probability of the estimation failing multiple times was very rare. The resulting data structure from this preliminary estimation was a data structure containing each gene's mean and variance of the off (noise) distribution, which is a normal distribution.

Computational Speed Up

Parallelization of the EM algorithm could be achieved in per-gene calculations. The original scripts were adapted to our multi-node computational resource, which was achieved by leveraging our SGE installment. The script accepts two integers that indicate

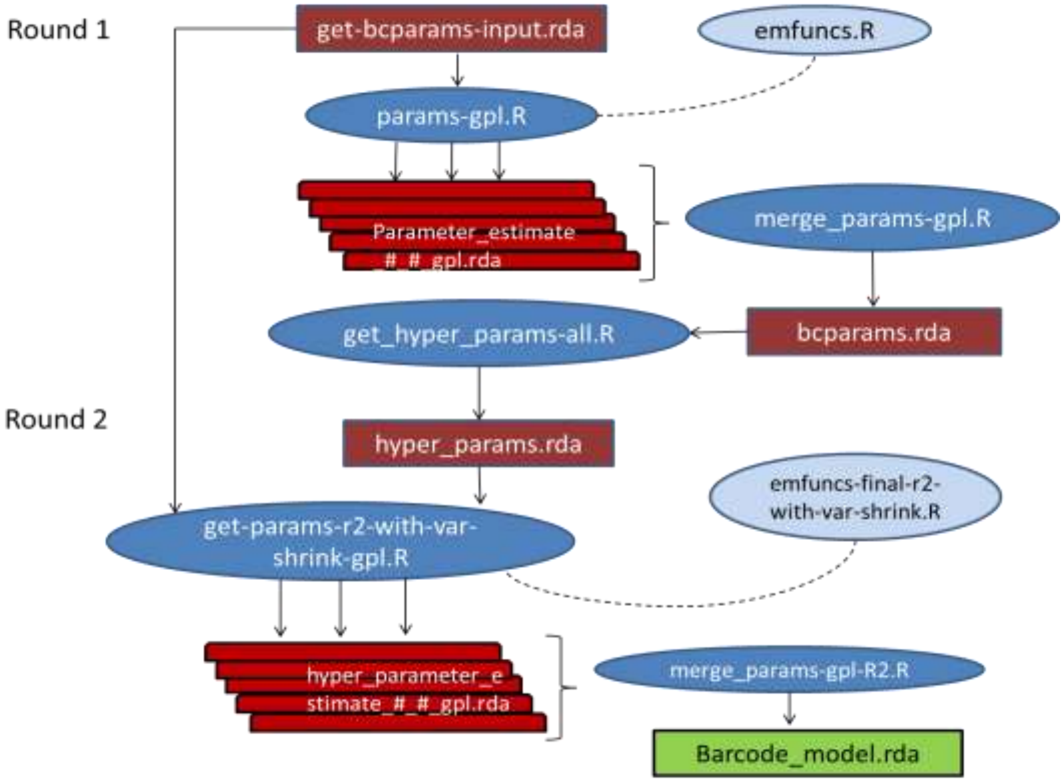


Figure 5: Model Estimation Workflow.

the slice of rows, the genes, which the script computes. In a parallel context this allowed the script to be called and run on individual nodes, each working on a separate slice of the input data. Reading from a single file may act as a limiting factor however, it was not the computationally heavy part of the algorithm, compared to the optimization component, and thus the slowdown created by reading from a single file is minimal handled within SGE optimally. The result of the parallelization was a collection of files. Merging happened in the next script, merge-params-gpl.R, recombine multiple files into a single file, bcparams.rda, in preparation for Round 2.

Round 2: Re-estimation taking Global Effects into Account

As noted, the basis of the approach was to get a rough estimate, and then obtain the global variables, referred to as hyper parameters, to describe the overall trends within the dataset. The script get-hyper-params-all.R simply followed the equation in obtaining the normal distribution and inverse gamma distribution of the entire dataset. The data were stored in the file hyper_params.rda, which is half of the input for the subsequence script. Nearly identical to the first round script, get-params-r2-with-var-shrink.R accepted both the original input space get-bcparams-input.rda and the hyper parameters to perform estimation under the EM algorithm, with the addition of the hyper parameters. This script follows the same parallel strategy as the first round estimation and results in a similar output, as well as a separately designated directory. The same merge script compiles the multiple files into a single file and the result became the barcode model from which all of the results can be computed.

Validation of Barcode Modifications

In adapting an already established algorithm it was important to show that no significant changes were made that would invalidate the algorithm. The entire data set was run, normalized by fRMA under the old CDF, to inspect the precision of values outputted, with a new input. A minimal amount of difference was expected, for genes close to the decision boundary. The validation model was done by computing data points that represent probe sets rather than previously being summarized into genes. The validation focused on the consistency of the algorithm. The comparison was a measure of exact point-by-point comparison, therefore if the original algorithm produced a 0 and the modified algorithm produced a 0, it was considered an agreement. The same agreement exists with 1's matching in both cases. However, if the original produced a 0 and the modified model produced a 1 or vice versa, it was considered a disagreement. A total of ~6,000 microarray files were barcoded by both models (original & modified). The results from these two models where then compared point-by-point comparison resulted in 89% identical results. That meant that 11% of the data had switched from one result to the other. Determining if the 11% mis-match, is significant is not trivial. It

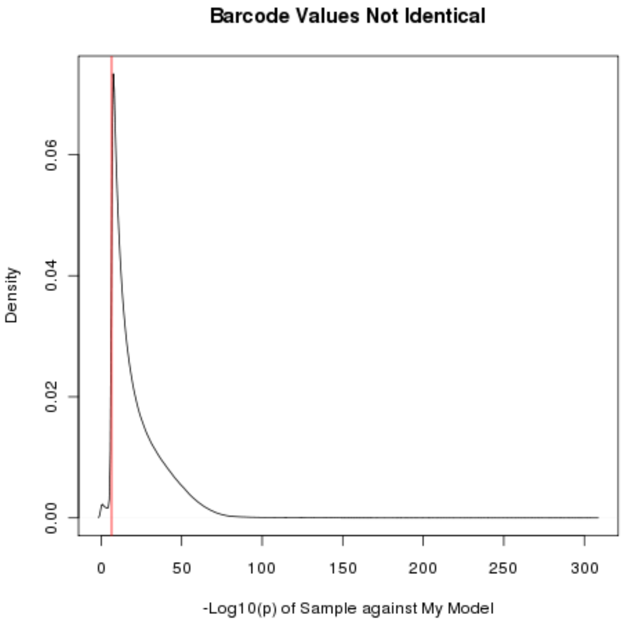


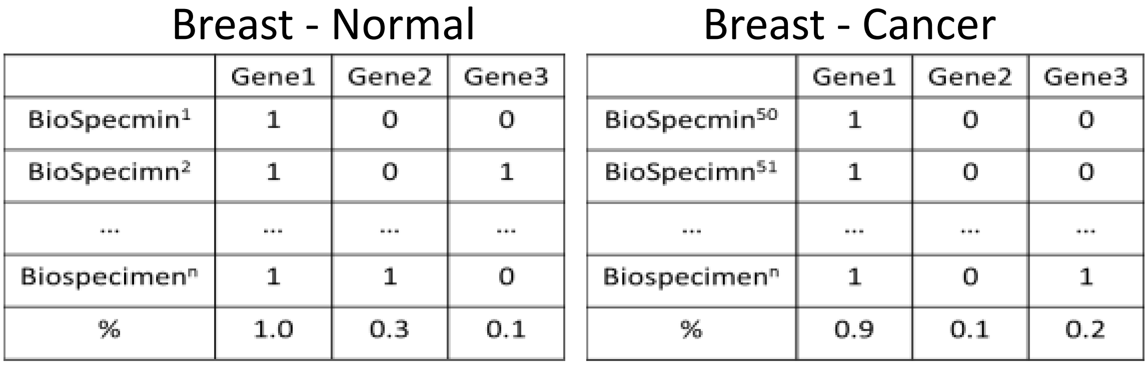
Figure 6: Validation with original barcode.

would require having a known truth to compare both models against, which does not currently exist. However it is possible to identify the location of the mis-matches in relation to the decision boundary, which both models share. As a result, there is a peak which can be interpreted as the values “Off originally, On in the modified model”. Our claim for this result is the difference in microarrays used in the initial creation of the barcode model. Genes that were originally located near the decision boundary would be highly sensitive to slight changes, such as additional microarrays added to the original collection. The important point was that there is a subset of genes that are influenced by the input space and our modification, which could represent genes with high variability.

Results & Discussion

Gene Expression Analysis in Tissues

The summarization of gene expression into Boolean values is a means to simplify the complex intensity data into more interpretable values with regards to the biology of a cell. The simpler can then be used in additional layers of analysis easily. (Kauffman, 1974)The complexity quickly increases by combining tissue descriptions, curated pathway networks and Boolean gene expression, but this more comprehensive cross-section can provide additional perspectives into the behavior of gene expression. The analysis was narrowed to 14 tissues; Breast, Cervix, Colon, Fallopian Tube, Gastric Tissue, Kidney, Liver, Lung, Ovary, Pancreas, Prostate, Sigmoid Colon Mucosa, Skin, Tongue Squamous Cells. Since the output values for each gene is a Boolean value, either 0 or 1, summarization of the state of a gene in a tissue is done by computing percentage of time the gene is ‘on’ in the set of bio specimens in this tissue. This strategy intends to provide a single value that captures the empirical variability of a gene within a tissue. The on /off state of gene can vary within bio specimens a given tissue type. Thus any gene at 0.5 or 50% would reflect the highest variability. When comparing tissues between themselves or pathways state between tissues, the percentage of gene in ‘on’ state is used as its ‘coordinate’ in the state space. For instance the state of a pathway in a tissue that includes n genes is defined as a n-dimensional vector of this percentage value. By representing the data in this manner the Euclidean distance equation can be applied, to determine the distance between those two pathways. Note that this distance can be similarly computed between all genes in two tissues to compare them.



Breast - Normal

	Gene1	Gene2	Gene3
BioSpecmin[1]	1	0	0
BioSpecimn[2]	1	0	1
...	...	...	...
Biospecimen[n]	1	1	0
%	1.0	0.3	0.1

Breast - Cancer

	Gene1	Gene2	Gene3
BioSpecmin[50]	1	0	0
BioSpecimn[51]	1	0	0
...	...	...	...
Biospecimen[n]	1	0	1
%	0.9	0.1	0.2

Figure 7: Example of Euclidean distance determination. This approach uses the tissue signal as a coordinate in an nth dimensional configuration, where n is the number of genes.

Comparison between Normal and Cancer Tissue

We computed the difference between normal tissue and the cancer tissue. Calculating the Euclidean distance between the two signatures of normal and cancer provides a metric of difference. The result of the distances between normal and cancer are shown in the table 1. According to this method the tissue with the shortest distance between normal and cancer is Gastric tissue, suggesting that the change from an on/off perspective is minimal. From a biological perspective, this indicates that the on/off variance between genes in the two tissues quite similar. The tissue with the largest distance is Lung cancer, suggesting that the disease affects the state of a larger number of genes or impact more dramatically the state of some genes. The tissue that appears to have the least amount of gene expression in common is lung normal from lung cancer.

Table 1: Tissue Normal Vs. Cancer

Tissue	Cancer
Gastric Tissue	18.28577
Tongue Squamous Cells	31.62825
Breast	37.11043
Pancreas	38.12178
Colon	40.73352
Prostate	46.14632
Fallopian Tube Epithelium	52.39634
Kidney	55.13081
Sigmoid Colon Mucosa	56.02789
Skin	65.77019
Cervix	69.26119
Ovary	70.78509
Liver	70.8988
Lung	78.3623

Comparison between Normal Tissues

The same analysis was performed across tissue types by computing the distance between two normal tissues. For instance, as expected, the distance between Breast and Ovarian displays the closest distance, suggesting a high degree of similarity between these two tissues. Interestingly it is well know that the cancer state of these two tissues is very similar. On the other hand, Breast is more distant from Cervix and Fallopian Tube Epithelium, than Ovarian, which suggests that our metric does not capture a signature specific to female organs. We also observe that Colon and Lung being the farthest away, among all the tissues listed.

Table 2 Normal Vs. Normal Tissue Comparison

	Breast	Cervix	Colon	FTE	Gastric	Kidney	Liver	Lung	Ovary	Pancs	Proste	SCM	Skin	TSC

Breast	0.00													
Cervix	61.37	0.00												
Colon	49.11	75.24	0.00											
Fallopian Tube	56.82	55.83	71.21	0.00										
Gastric Tissue	63.83	32.32	76.28	58.09	0.00									
Kidney	67.23	37.51	79.85	60.88	34.42	0.00								
Liver	64.83	32.72	76.78	59.05	30.51	33.03	0.00							
Lung	72.77	44.51	84.50	66.73	45.25	40.71	43.08	0.00						
Ovary	28.90	62.99	48.20	60.83	66.89	69.87	66.63	74.46	0.00					
Pancreas	69.21	42.82	81.34	61.14	40.30	34.48	39.80	30.49	71.71	0.00				
Prostate	48.81	74.63	40.06	69.39	76.08	79.33	76.62	83.91	48.82	80.67	0.00			
Colon Mucosa	57.41	43.48	70.16	51.17	41.36	48.91	46.27	57.23	61.68	51.94	71.21	0.00		
Skin	42.98	72.21	32.77	67.62	74.10	77.19	74.45	81.96	44.37	78.62	34.76	68.6	0.00	
Tongue	41.99	68.93	49.80	63.51	70.68	74.02	71.06	78.63	41.14	75.57	47.34	65.94	45.71	0.00

Pathway Focused Comparison of Normal versus Cancer Tissue

We computed the pathway centric distance between normal and cancer of the same tissue type Pathways were obtained from KEGG (Kanehisa, 2012). All the distances were computed between normal and cancer for each tissue type and for all pathways. This analysis was performed on 185 KEGG pathways across the 14 tissue and 2 disease states resulting in 2,590 distances.

How the Mean is Computed

This metric provides a means to rank pathways in a similar way to differential analysis. The expected pathways that have the greatest distance would be known cancer pathways. As such, the most commonly first ranked pathway is the KEGG “Pathways in

Cancer”, which provides a good affirmation. Additionally, the p53, WNT and MapK signaling pathways are ranked highly as well. These three pathways are commonly known to be associated with cancer. In addition to detecting pathways in cancer that are farthest from normal, the pathways closest to normal are also revealed. We would expect that pathways that act very similar between normal and cancer are fundamental for survival, most likely house-keeping pathways. The pathways with the least amount of distance were typically protein and amino acid metabolism, and sugar degradation.

Table 3 Pathways Normal Vs. Cancer

Mean Rank Score	*Pathway Name*
1.6	PATHWAYS_IN_CANCER
2.9	CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
5.6	MAPK_SIGNALING_PATHWAY
6.0	NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION
6.5	FOCAL_ADHESION
6.6	ENDOCYTOSIS
8.9	CHEMOKINE_SIGNALING_PATHWAY
9.4	REGULATION_OF_ACTIN_CYTOSKELETON
9.5	CELL_ADHESION_MOLECULES_CAMS
11.1	PURINE_METABOLISM
13.1	TIGHT_JUNCTION
19.1	WNT_SIGNALING_PATHWAY
19.9	SPLICEOSOME
20.6	PYRIMIDINE_METABOLISM

Future Work

Application of Other Microarray Platforms

This analysis contains a significant amount of statistics based work. This project only entailed a single Affymetrix array; U133-Plus2. This single array was selected due to it being the largest represented array in the GEO database, with 66,894 samples publicly available. Of that collection only samples with Tissue and disease details available were used. The first step would be to obtain the rest of the samples and tissue annotations (only possible through literature searching and human selection). That would be followed by applying the whole platform of that data to Barcode model training. Once that is completed, the second desire would be to obtain other platforms, Affymetrix or otherwise. By obtaining multiple platforms, it should be feasible to make gene expression comparisons across microarray platforms.

Utilization of More Robust Differential Calculation

There is a potential pitfall that biases the Euclidean distance calculations for describing the difference between normal and cancer. The summarization over bio specimen of the same tissue type fails to account for the variance of that percentage. Hence a more complicated, more robust method is proposed by determining the difference through permutation probability. The permutation probability approach would be used to calculate a more representative value for gene behavior across bio specimen.

Integration with Archive Resources

The Barcode workflow was created to initially from data within BORA and built with the intention to loop back to deposit the Boolean results back into BORA. Currently the data used to create the modified Barcode model is being loaded into the repository, but the resulting barcoded values has yet to be stored.

Expansion to Open Source Software

One of the most widely used open source software used for pathway analysis is Cystoscape (Shannon, Markiel, Ozier, Baliga, Wang, Ramage, et. al., 2003). Cytoscape is a client software program that contains tools to integrate datasets and interact with graphical representations. A direct link between Cystoscape and BORA has been developed, to leverage the stored data in a databank. Once the Barcoded values of gene states are loaded into BORA, this link could be used to perform more targeted, deeper analysis.

References

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., et al. (2011). NCBI GEO: Archive for functional genomics data sets--10 years on. *Nucleic Acids Research, 39*(Database issue), D1005-10. doi:10.1093/nar/gkq1184

CDF FILE. (2009, March). In *Affymetrix® CDF Data File Format*. Retrieved April 10, 2012, from http://www.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cdf.html

.CEL File Extension. (2011, May 18). In *FileInfo.com*. Retrieved February 22, 2012, from http://www.fileinfo.com/extension/cel

Eichinger, L. (2009, January 7). Figure 1: Principle of DNA microarray analysis. In *Dictyostelium discoideum Microarray analysis*. Retrieved April 5, 2012, from http://www.uni-koeln.de/med-fak/biochemie/transcriptomics/07_analysis.shtml

Glass, L. (1975). Classification of biological networks by their qualitative dynamics. *Journal of Theoretical Biology, 54*(1), 85-107.

Glass, L., & Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology, 39*(1), 103-129.

Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A.J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J.Y. and Zhang J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5(10): R80.

Hellwig, B., Hengstler, J. G., Schmidt, M., Gehrmann, M. C., Schormann, W., & Rahnenfuhrer, J. (2010). Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics, 11*, 276. doi:10.1186/1471-2105-11-276

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Research, 31*(4), e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England), 4*(2), 249-264. doi:10.1093/biostatistics/4.2.249

Jia, P., Zheng, S., Long, J., Zheng, W., & Zhao, Z. (2011). dmGWAS: Dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics (Oxford, England), 27*(1), 95-102. doi:10.1093/bioinformatics/btq615

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M.; KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40, D109-D114 (2012).

Kauffman, S. (1969). Homeostasis and differentiation in random genetic control networks. *Nature, 224*(5215), 177-178.

Kauffman, S. (1974). The large scale structure and dynamics of gene control circuits: An ensemble approach. *Journal of Theoretical Biology, 44*(1), 167-190.

Kim, J., Patel, K., Jung, H., Kuo, W. P., & Ohno-Machado, L. (2011). AnyExpress: Integrated toolkit for analysis of cross-platform gene expression data using a fast interval matching algorithm. *BMC Bioinformatics, 12*(75), 1. Retrieved April 5, 2012

Lage, K., Hansen, N. T., Karlberg, E. O., Eklund, A. C., Roque, F. S., Donahoe, P. K., et al. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America, 105*(52), 20870-20875. doi:10.1073/pnas.0810772105

McCall, M. N., & Irizarry, R. A. (2011). Thawing frozen robust multi-array analysis (fRMA). *BMC Bioinformatics, 12*, 369. doi:10.1186/1471-2105-12-369

McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., & Irizarry, R. A. (2011). The gene expression barcode: Leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research, 39*(Database issue), D1011-5. doi:10.1093/nar/gkq1259

Ochab-Marcinek, A., & Tabaka, M. (2010). Bimodal gene expression in noncooperative regulatory systems. *Proceedings of the National Academy of Sciences of the United States of America, 107*(51), 22096-22101. doi:10.1073/pnas.1008965107

Parmigiani,G., Garrett,E.S., Anbazhaghan,R. and Gabrielson,E. (2002) A statistical framework for expression-based molecular classification in cancer. *J. Roy. Stat. Soc. B* (Stat. Meth.), 64, 20.

Ponten, F., Gry, M., Fagerberg, L., Lundberg, E., Asplund, A., Berglund, L., et al. (2009). A global view of protein expression in human cells, tissues, and organs. *Molecular Systems Biology, 5*, 337. doi:10.1038/msb.2009.93

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research 2003 Nov; 13(11):2498-504*

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550. doi:10.1073/pnas.0506580102

Thompson, Robert, and Fan Meng. "Brain Array." *MBNI Microarray Lab*. University of Michigan - Ann Arbor, Aug. 2011. Web. 7 Feb. 2012. http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/15.0.0/entrezg.asp

Wang, X. V., Verhaak, R. G., Purdom, E., Spellman, P. T., & Speed, T. P. (2011). Unifying Gene Expression Measures from Multiple Platforms Using Factor Analysis. *Analysis. PLoS ONE*, *6*(3).

Zhang, Aidong. *Advanced Analysis of Gene Expression Microarray Data*. Singapore: World Scientific Publishing, 2006. N. pag. Print.

Zilliox, M. J., & Irizarry, R. A. (2007). A gene expression bar code for microarray data. *Nature Methods, 4*(11), 911-913. doi:10.1038/nmeth1102