International Research Journal of Biochemistry and Bioinformatics (ISSN-2250-9941) Vol. 2(10) pp. 208-215, October, 2012
Available online http://www.interesjournals.org/IRJBB
Copyright © 2012 International Research Journals

Full Length Research Paper

Effects of expression profile data variation on boolean gene regulatory network predictions

Kanda Euatham[1], Natee Tongsiri*[2,3]

[1]Bioinformatics Research Laboratory, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand 50200
[2]The Centre of Excellence in Mathematics, Commission on Higher Education, Sri Ayutthaya Rd., Bangkok, Thailand 10400
[3]Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand 50200

Accepted October 10, 2012

Reconstructing the regulatory relationships between genes using multiple time point expression profile data (EPD) is a powerful computational method to gain insight into gene networks. One such method uses binary on/off relationships to characterize the under- and over-expression of genes acting in unison. This approach uses only the relative expression levels of the genes of interest at multiple time points. One aspect of the EPD these methods often fail to account for is the inherent variability in the measurements of the gene expression levels. We characterize the variability in expression levels for a single time point to measure the inherent variability in that dataset. We then generate multiple new expression profile data samples from the original data and measured variability. These new datasets are then binarized to test whether the gene network relationships change due to the random sampling. This also allows us to test different variation magnitudes to set limits on how large the inherent variability should be to yield reproducible results for the binary gene network method. We find that the current variabilities in EPD are too large to yield reproducible gene regulatory networks, but that the data for some particular genes are sufficient to generate reproducible binarizations.

Keywords: Stochastic modeling, binarization, *S. cerevisiae*, gene chip.

INTRODUCTION

One of the main outstanding problems in molecular biology is to determine the functional relationships between differentially regulated genes. This field is complicated by the fact that all of the genes in an organism share a common genome but protein expression can vary dramatically depending on the position of the expressed gene within the organism, external factors such as heat and light and temporal variation. Since the advent of the genomic era a growing body of biological data has become available which can begin to describe these differences in gene expression. In particular, DNA microarray technology is now able to provide a snapshot of the expression levels of all of the genes in a particular organism at one time point for a given tissue sample (Bassett DE *et al*., 1999; Eisen MB *et al*., 1998). Various computational techniques have been developed to help analyze these data sets including Bayesian Networks (Zou M and Conzen SD, 2005; Yu J *et al*. 2004; Werhli AV *et al*., 2006; Friedman N *et al*., 2000), Differential Equations (Gardner TS *et al*., 2003; Gebert J *et al*., 2007; Quach M *et al*., 2007), Fuzzy Logic (Ressom H *et al*., 2003) and Information Theoretic Models (Stuart J.M *et al*., 1003; Margolin AA *et al*., 2006; Rao A *et al*., 2007). One limitation of these methods is that they require large amounts of data to function effectively and often cannot be applied to the types of sparse time point datasets which are publically available (Baldi P and Long AD, 2001). In particular, most of these methods require that the measured expression levels are accurate, which is often not a valid assumption due to inherent measurement inaccuracies in expression profile data.

Another set of methods which can infer regulatory relationships between expressed genes are called Boolean Regulatory Networks. These methods have been extensively studied because they simplify the

*Corresponding Author E-mail: nongpimmy@hotmail.com

functional analysis to the point that current data sets are biologically meaningful (Martin S *et al.*, 2007; Faure A *et al.,* 2006; Lähdesmäki H *et al.*, 2003; Liang S *et al.*, 1998). Boolean methods make the simplifying assumption that genes are either expressed (1) or unexpressed (0). This allows relationships between genes that are being coregulated to be highlighted and reduces much of the difficulty with inaccurately measured expression profiles since all that is required for the Boolean analysis is a determination of whether a gene is being expressed or not expressed at a given time point. With these simplifying assumptions, Boolean modeling can be used to reconstruct gene regulatory relationships including whether sets of genes are coexpressed or have a regulatory effect on each other. For these methods to function properly an effective binarization technique is required which can categorize the genes into on/off binary data that represents the biological meaning of the original raw data. The package called 'BoolNet'(Müssel C *et al.*, 2010) in the R programming language (R Development Core Team, 2005) is statistical program that can construct such binarizations from a given expression profile dataset and then further generate a gene regulatory network from these binarizations. It is often assumed though that the when dealing with Boolean networks the fact that the methods to binarize time point data are full developed implies that the binarizations are accurate. In this research we test this underlying assumption by doing a statistical analysis of the variation inherent in a given expression profile dataset using the repeated gene measurements. We then generate new datasets with the same average expression levels and variability as in the original dataset and using these new randomly generated datasets test the consistency of the binarization methods to produce biologically meaningful on/off expression profiles. This analysis allows a determination of how large the variation in the expression profile data should be to yield reproducible results for the binary gene network method.

MATERIALS AND METHODS

Data Set

The expression profile dataset used in this work consisted of 7 distinct time points taken for the test species *Saccharomyces cerevisiae* (Yeast) under which the Glucose levels changed to study the metabolic effect of glucose deprivation in yeast (DeRisi JL *et al.*, 1997). The initial 5 time points had non-zero glucose levels whereas the glucose for the final two time points was completely exhausted. The raw data was downloaded from the Metabolic Time course website (http://cmgm.stanford.edu/pbrown/explore/additional.htm) *S. cerevisiae* has been widely used to investigate gene expression since this genome has been extensively characterized, the genetic regulatory mechanisms for most genes in yeast are already known, and gene expression profile chips are cheaply available to further study this test organism.

Data Set Preprocessing

The original raw expression profile data was preprocessed in two ways to make the data from the different time points as comparable as possible in the standard fashion. First a background intensity correction was performed to normalize the spot intensities against the ambient background of the overall expression data. This normalization was performed on both the red and green data channels which then yielded the standard $\log_2(R/G)$ data values. These intensity values were then Lowess normalized to set the average log offset to zero as expected since most genes are assumed to not dramatically change from one timepoint to another. The Lowess normalization was performed using the online MIDAW (Romualdi C *et al.*, 2005) normalization server.

Measurement of Variability in the Expression Profile Data

Using the Yeast Dataset described above, all genes in the gene expression profile data with multiple measurements for a single time point were extracted. There were 52 distinct genes with two distinct measurements for each of the time points.

Stochastic Sampling of Paired Measurements

To determine the expected profile distribution from the paired samples a randomly generated dataset consisting of normally distributed data with a mean of 0 and a variance of 1 were generated using the standard PERL algorithm from the Perl Cookbook (Christiansen T and Torkington N, 1998). The expected mean and variance of the paired samples measurements for the 52 sample size was determined from these randomly generated datapoints.

Binarization of EPD

The package BoolNet which was developed in the R programming language was used to perform the binarizations of the seven time point datasets. Two different parametric methods of binarization (k-means and edge detector) were tested by setting the appropriate options for BoolNet.

Table 1. Average magnitude between the paired expression profile values for the 52 duplicated genes at each time point in *S. cerevisiae*.

Time point	1	2	3	4	5	6	7
Average Magnitude deviation	0.209	0.226	0.217	0.184	0.243	0.433	0.389

Table 2. The standard deviation, and upper and lower bounds on the variability in the measurement of the gene expression values for each of the seven time points.

Time point	1	2	3	4	5	6	7
95% Lower Bound EPD	0.153	0.166	0.159	0.135	0.178	0.317	0.285
Average Expected EPD	0.185	0.201	0.193	0.163	0.216	0.384	0.346
95% Upper Bound EPD	0.234	0.254	0.243	0.206	0.273	0.486	0.437

RESULTS AND DISCUSSION

Measurement of Variability in the Expression Profile Data

Time series expression profile data records the concentrations of selected genes at several distinct time points to determine how changes in the gene expression levels correlate with specified treatment effects. For a single time point the same DNA sample is used to determine the expression levels of all of the genes on the microarray chip. For many such microarray chips there are repeated measurements for some of the genes on the chip. The expression levels for these repeated genes provide a measure of the variability in the expression profile data. For example in the microarray chip for *S. cerevisiae* the gene YAL001C is measured two times. When the variability in the measured expression levels is low, these two values provide an accurate measure of the true expression level and the difference between the two measured values is small. As the variation in the measured expression levels increases the average difference between the two random measurements should increase as well. To test this we extracted all of the genes with repeated measurements from the expression profile data of the seven time point series data for *S. cerevisiae* and determined the magnitude of the differences between the paired data. For this data set there were 52 genes with repeated measurements and the average magnitude of the differences for this data for each time point are shown in Table 1. These values quantify how much variability is present in the expression profile data, but what is required is the relationship between these deviations and the variation in the original gene expression values.

Stochastic Sampling of Paired Measurements

To determine how large the average magnitudes between paired measurements sampled from the same distribution, we used a sampling function with a mean of zero and standard deviation of one to experimentally determine these values. Using the Gaussian normal sampling function from the Perl Cookbook[21], we generated 10,000 datasets of fifty random samples each from a distribution with a mean of zero and a standard deviation of one. To verify that this sampling function was generating random samples with the correct properties, we measured the averages and standard deviations within each of these 10,000 datasets and found that the overall average of these 500,000 random samples was 0.0005 with a standard deviation of 0.9988. Although we used the standard Gaussian sampling function, it is still important to verify that the routine is working correctly and that the generated properties are as expected. Next we paired the data from these 10,000 datasets to generate 5,000 pairs of sample values in groups of fifty. We then measured the magnitudes between these pairs and determined the average and standard deviation for each group of fifty data points. The overall average of the magnitudes for the groups of fifty was 1.126 ± 0.12 and the average standard deviation within these samples was 0.849 ± 0.10. This directly relates the measured average magnitudes between the pairs of samples and the standard deviation (STD) of the original distribution. Due to the small group sizes of fifty there is a reasonable amount of variability about this value. Therefore it is most accurate to say that the standard deviation of the distribution from which the paired data was drawn should 95% of the time lie in the range of 0.891 to 1.366 of the average magnitude of the paired data. This provides a

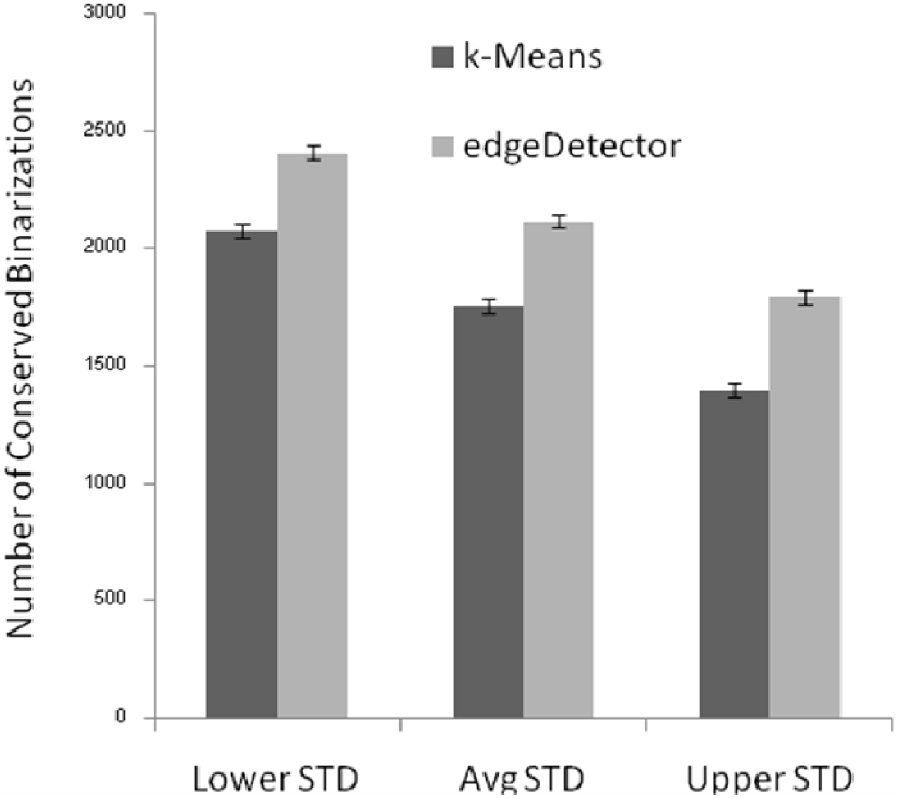

Figure 1. The total number of genes with the same binarization as in the original data for the two BoolNet methods. The Lower and Upper STD values show the number of genes with the same binarization as in the original dataset for the 95% bounds on the predicted variation in the expression values.

direct relationship between the measured deviations shown in Table 1 and the variability in the original distribution. With this measure of the standard deviation of the gene expression data, it is possible to accurately model the original data using the actual expression values for each gene.

Binarization of EPD

To see which of the two binarization methods used in the BoolNet package produces the most robust model for the inputs to the gene regulatory analysis, we first tested the two methods on the original expression dataset. The k-Means method analyzed the 6,153 genes in approximately 34 seconds and the edgeDetector method analyzed the same data in just under 4seconds. The binarization results from these two methods were dramatically different with only 39.7% of the genes being binarized the same way. For expression profile data although the edgeDetector method is almost 10 times faster than the k-Means method, it is probably not an appropriate choice in general since this method focuses on the largest difference in expression between two time points. If one time point is highly over- or under-expressed this would bias the binarization to segregate only that time point even if other timepoints had high or low expression levels. Most likely this is one reason the k-Means method is set as the default binarization method.

Then, we generated randomly sampled datasets with the means of the original data and variation based on the variation analysis described above. For each of the following analyses we generated 1,000 distinct datasets, with the standard deviations for each of the time points shown in Table 1. For each of the three measures of the amount of variation in the gene expression values (i.e. average, upper- and lower-bounds on the variation in the system) the 1,000 datasets were stored for later use and then analyzed using the BoolNet package. The two distinct R BoolNet binarization methods were run on each of the three groups of 1,000 datasets to determine the binarization for each gene. This analysis took on average over eight hours per 1,000 sample dataset for the k-Means method and approximately an hour for the edgeDetector method. Figure 1 shows the number of genes which had the same predicted binarization as in the original data set. Ideally for a method to be robust, all of the genes should have the same binarization in the original dataset and for the generated datasets. As can be seen in Figure 1, out of the 6,153 genes in the *S. cerevisiae* expression profile data, roughly 30% were

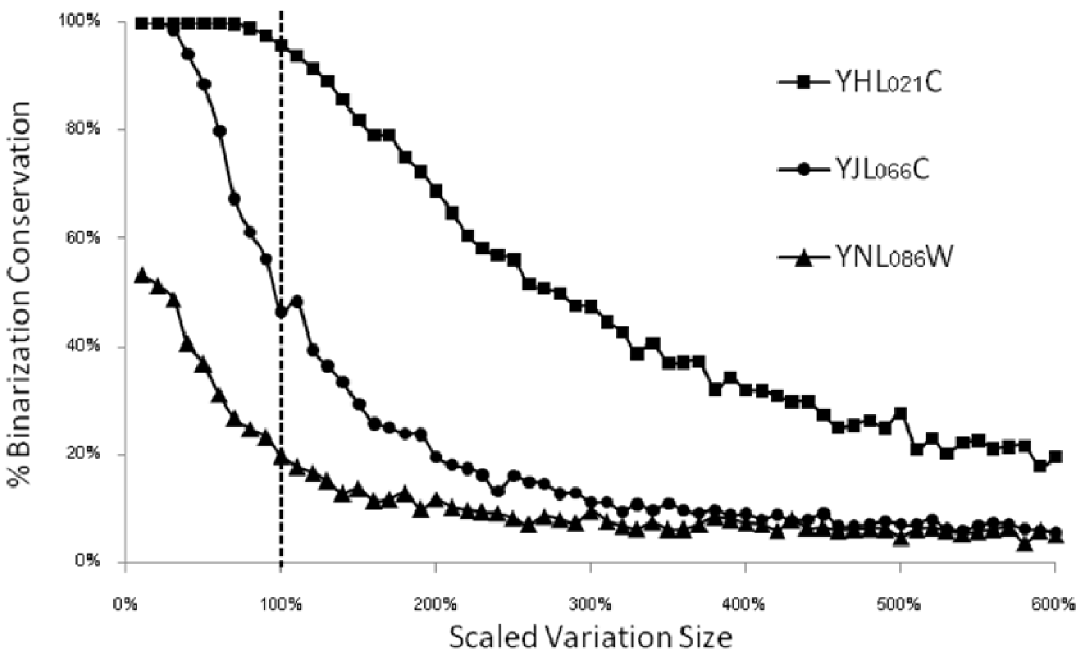

Figure 2. The %conserved binarization for three genes with different sizes of variation in the sampled distributions ranging from 0 to 600%.

found to have conserved binarizations for the randomly sampled datasets. The edge Detector method performed better over the entire tested range with the predicted %conserved binarization falling between 30 -40% with an average of 34%. The k-Means method fell in the same general range with the expected values lying between 23 -34% with the average value lying slightly above 28%. In all of these cases the binarization methods are clearly not robust to variability in the original expression data, but the predicted values were remarkably consistent having only 30 genes standard deviations for the conserved binarization within each of the 1,000 generated data sets. Ideally the conserved binarization should share 95% or 99% conservation to the original dataset so that the predicted regulatory relationships could be based on binarization values which are not highly dependent on the variability inherent in the expression profile data.

Single Gene Binarizations

Although the total number of genes with conserved binarizations was remarkably consistent over a range of STD values, the amount of conservation for particular genes was much less conserved. For example, for the average predicted STD values using the k-Means method, 28.5% of the genes had conserved binarizations. Looking at individual genes across the 1000 generated samples, there was a huge amount of variation in the binarization conservation ranging from only 7 conserved cases for the gene YPL082C all the way up to greater than 95% conservation for 80 different genes including gene YHL021C.

To get a better grasp on the changes in conserved binarization for different variation amounts, we chose three genes which fell in the high (YHL021C), medium (YJL066C) and low (YNL086W) conservation regimes to test how the %conservation changed with different variability sizes. For each gene, 1,000 randomly sampled datasets for the 7 time points were generated for different variability sizes. To exhaustively test these relationships sample data sets were generated ranging from 10% to 600% the size of the measured variability in 10% variability steps. This produced a total of 60 datasets with increasing variability sizes for each gene with 1000 estimates for the gene expression values. For each of these datasets, the BoolNet k-means binarization method was then run to determine whether the generated expression profile values yielded the same binarization as was found for the original data. This result is shown in Figure 2. where the dashedline for 100% shows the %binarization conservation for the actual measured amount of variation from the gene expression values.

As the amount of variation approaches 0 the %binarization conservation should approach 100% as would be expected since a zero percent variation is exactly identical to the original data. As the size of the variation grows larger the %binarization conservation decreases as expected since for large variations the relationships between the time point data are completely randomized. Figure 2 clearly shows that the %binarization conservation for these three genes can vary widely while following the same general pattern. The level of %binarization conservation is dependent on the

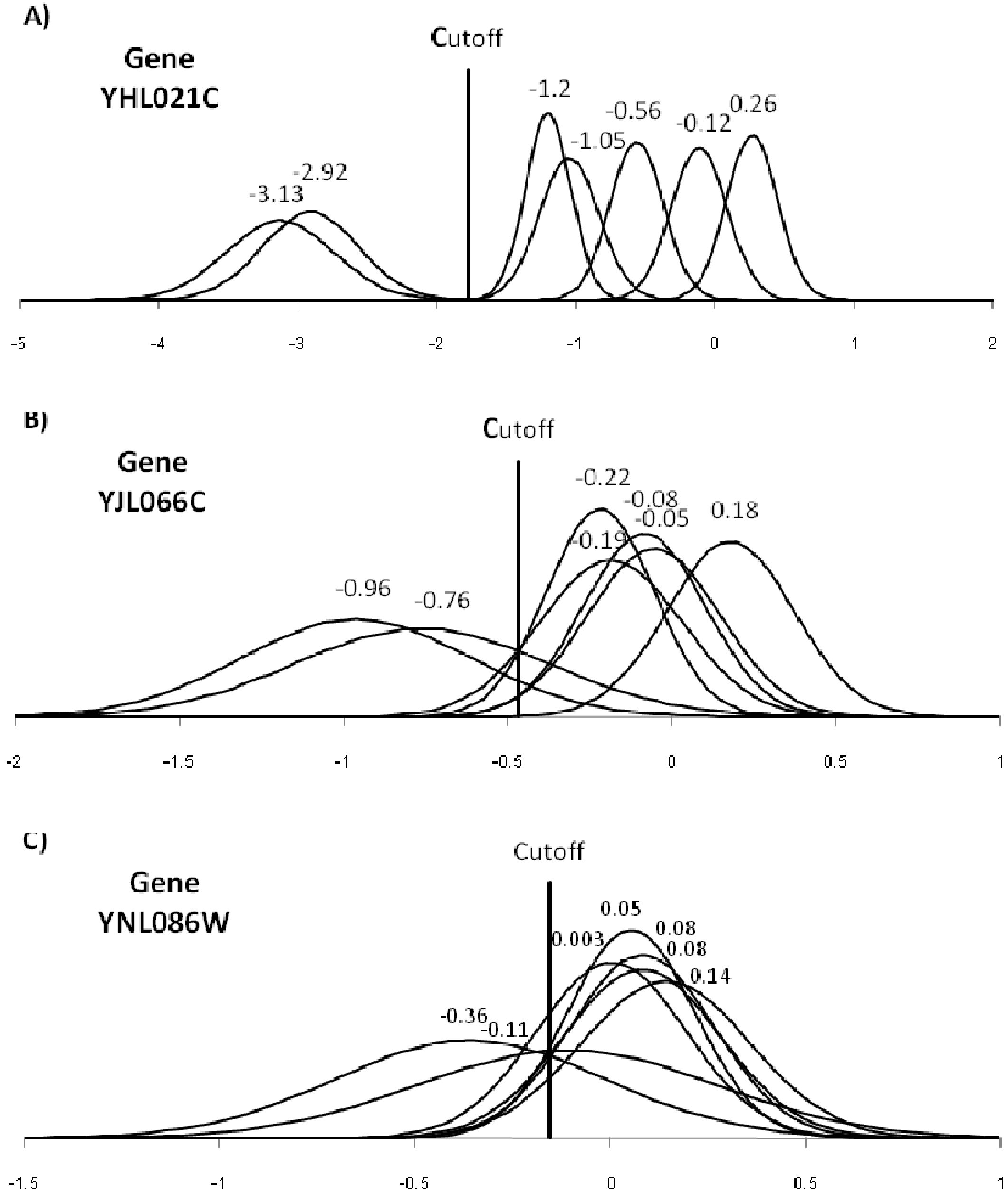

Figure 3. The gene expression values and measured variation for the three genes A) YHL021C, B) YJL066C and C) YNL086W. The normal curves for each time point have the average expected standard deviations shown in Table 2 and unit area.

actual pattern of gene expression at the different time points for that gene. Figures 3A) - C) show the pattern of gene expression for the same three genes examined in Figure 2. For each time point, the measured amount of variation around the expression value is shown and these figures are a visual depiction of the distributions the random samples are being drawn from. The cutoff generated by the BoolNet binarization program is also show in these figures to clearly show the two groups of expression values being binarized to 0 and 1 respectively. The pattern of expression values for gene YHL021C in Figure 3A) shows that the positions of the two time points on the left are far enough away from the five time points on the right that the amount of variation in the gene expression values does not often affect the binarizations for the sampled datasets. In Figures 3B) and 3C) the separation between the different time points are smaller leading to a greater change due to the variation in the data when sampling from these distributions. If the different time points are sufficiently far apart as is the case in Figure 3A) the variability in the expression data does not affect the ability of BoolNet to

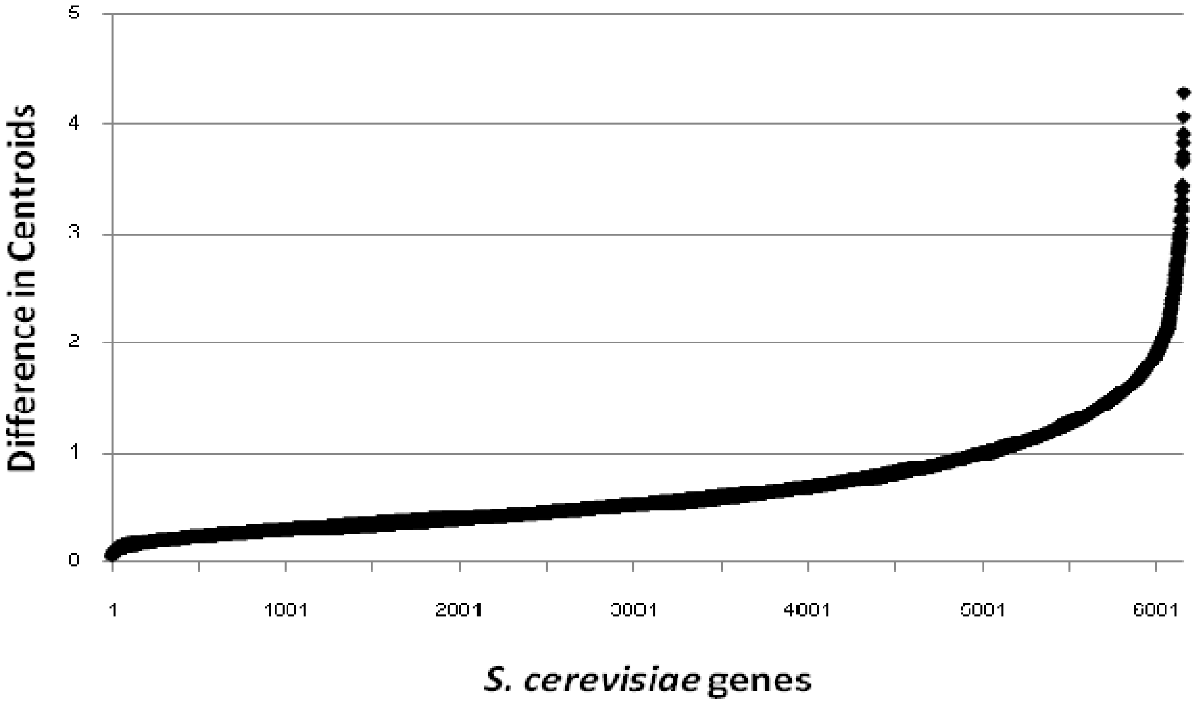

Figure 4. The differences in centroids for the k-Mean binarization of the two groups for every gene in the *S. cerevisiae* sorted by magnitude, 4988 genes had centroids closer than 1 showing there was no significant differential expression.

perform an accurate binarization.

Biological Considerations

The binarization algorithms used by BoolNet assume that there exists at least one of each state in the expression data. This means that if a particular gene is not differentially expressed the BoolNet package will still assign it a binarization with two distinct expression classes. From a mathematical standpoint the BoolNet k-Means binarization routine correctly selects the optimal cutoff to produce the largest difference in centroids for the two groups of expression values, but whether the differences between these groups is biologically meaningful is a separate question. In general the minimum biologically meaningful difference in expression profile results is a 2-fold change in expression which is equivalent to a 1 difference in the log base 2 expression profile values. To have biologically meaningful binarizations the centroids for the two groups should be separated by at least 1 to ensure there is a possible change in expression between the different timepoints. Figure 4 shows the differences between the centroid values for the binarizations of the original 6153 genes.

Using a standard 1 value cutoff as the minimum requirement for the two centroids of the expressed and non-expressed groups, 4988 of the 6153 genes were found to not be differentially expressed. When using the BoolNet package with expression profile data this cut on the gene expression data should always be performed before any analysis of the binarization relationships are undertaken. The 1 cutoff as shown in Figure 4 is a safe cutoff to remove most of the non-differentially expressed genes. Biologically most genes are not expected to be differentially expressed in general, but only under distinct treatment conditions which stress the system such as heat, low glucose levels, high or low light levels, the presence of heavy metals, etc.

Taking into account the almost 5000 genes which are not differentially expressed, the percent conserved binarizations for the differentially expressed genes improves dramatically from 28.5% for the average STD values to 65.7% which is over a 2-fold improvement. Ideally this value would be over 95% so that the predicted binarizations are conserved the majority of the time, but this shows nicely the fact that trying to perform a binarization on data for which that value does not have a well defined meaning leads to poor results. Once the non-differentially expressed genes are accounted for, the BoolNet binarization routines do begin to better represent the data.

There are two main points that can be taken away from this: 1) For accurate enough gene expression measurements, the binarization methods used in BoolNet will produce consistently robust gene regulatory predictions. 2) The variability in the gene expression measurements are currently too large to produce consistently robust predictions. Although for some genes, the expression pattern is distinct enough to overcome the inherent variability in the gene expression measurements.

All in all, using a binarization method to classify expressed genes into expressed and unexpressed classes is a powerful technique to reconstruct gene regulatory relationships. It is important though to quantify the uncertainties in the system including the inherent variability in the expression profile measurements. Depending on the quality of the expression profile data and the pattern of the time series data, it is currently possible to reconstruct the gene regulatory relationships using BoolNet at least for some of the more highly differentially expressed genes.

ACKNOWLEDGEMENT

The authors would like to acknowledge support from the Research Professional Development Project under the Science Achievement Scholarship of Thailand (SAST).

REFERENCES

Baldi P, Long AD (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics,* 17: 509 –519.

Bassett DE, Eisen MB, Boguski MS (1999). Gene expression informatics—it's all in your mine. *Nature Genetics,* 21: 51-55.

Christiansen T, Torkington N (1998). Generating Biased Random Numbers. In: *Perl Cookbook Tips and Tricks for Perl Programmers,* O'reilly Publishing Company, Ch. 2.10.

DeRisi JL, Iyer VR, Brown PO (1997). Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science,* 278: 680-686.

Development CR, Team R (2005). A language and environment for statistical computing, reference index version 2.14.1. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA,* 95: 14863 –14868.

Faure A, Naldi AL, Chaouiya C, Thieffry D (2006). Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics,* 22: 124–131.

Friedman N, Linial M, Nachman I, Pe'er D (2000). Using Bayesian Networks to Analyze Expression Data. *J. Comp. Bio.,* 7: 601–620.

Gardner TS, diBernardo D, Lorenz D, Collins JJ (2003). Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. *Science Reports,* 301: 102-105.

Gebert J, Radde N, Weber GW (2007). Modeling gene regulatory networks with piecewise linear differential equations. *Euro. J. Oper. Res.,* 181: 1148-1165.

Lähdesmäki H, Shmulevich I, Yli-Harja O (2003). On learning gene regulatory networks under the Boolean network model. *Machine Learning,* 52: 147-167.

Liang S, Fuhrman S, and Somogyi R (1998). Reveal, A General Reverse Engineering Algorithm For Inference of Genetic Network Architectures. *Pacific Symposium on Biocomputing,* 3: 18-29.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics,* 7: 1-15.

Martin S, Zhang Z, Martino A, Faulo J-L (2007). Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics,* 23: 866–874.

Müssel C, Hopfensitz M, and Kestler HA (2010). BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics.,* 26: 1378 -1380.

Quach M, Brunel N, d'Alche´-Buc F (2007). Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference. *Bioinformatics,* 23: 3209–3216.

Rao A, Hero AO, States DJ, Engel JD (2007). Using directed information to build biologically relevant influence networks. *Comput. Syst. Bioinfo. Conf.,* 6: 145-156.

Ressom H, Wang D, Varghese RS, Reynolds R (2003). Fuzzy logic-based gene regulatory network. *The 12th IEEE International Conference on Fuzzy Systems 2003 FUZZ,* 03: 1210-1215.

Romualdi C, Vitulo N, Del Favero M, Lanfranchi G (2005). MIDAW: a web tool for statistical analysis of microarray data. Nucl. Acids. Res., 33: 644-649.

Stuart JM, Segal E, Koller D, Kim SK (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science,* 302: 249-255.

Werhli AV, Grzegorczyk M, Husmeier D (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics,* 22: 2523–2531.

Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics,* 20: 3594–3603.

Zou M, Conzen SD (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics,* 21: 71–79.