Journal of Biomedical Informatics 42 (2009) 287–295

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Use of Bayesian networks to probabilistically model and improve the likelihood of validation of microarray findings by RT-PCR

Sangeeta B. English [a,1,*], Shou-Ching Shih [b,1], Marco F. Ramoni [d], Lois E. Smith [c,1], Atul J. Butte [a,1]

[a] *Stanford Center for Biomedical Informatics Research (BMIR), Stanford University School of Medicine, 251 Campus Drive, Stanford CA, 94305, USA*
[b] *Pathology Department, Beth Israel Deaconess Medical Center Research North, 99 Brookline Avenue, Boston, MA 02215, USA*
[c] *Department of Ophthalmology, Children's Hospital Boston, and Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115, USA*
[d] *Children's Hospital Informatics Program, Harvard-MIT Division of Health Sciences and Technology and Harvard-Partners Center for Genetics and Genomics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston MA 02115, USA*

ARTICLE INFO

Article history:
Received 7 January 2008
Available online 26 August 2008

Keywords:
Bioinformatics
Bayesian network
Microarray
RT-PCR
Microarray data

ABSTRACT

Though genome-wide technologies, such as microarrays, are widely used, data from these methods are considered noisy; there is still varied success in downstream biological validation. We report a method that increases the likelihood of successfully validating microarray findings using real time RT-PCR, including genes at low expression levels and with small differences. We use a Bayesian network to identify the most relevant sources of noise based on the successes and failures in validation for an initial set of selected genes, and then improve our subsequent selection of genes for validation based on eliminating these sources of noise. The network displays the significant sources of noise in an experiment, and scores the likelihood of validation for every gene. We show how the method can significantly increase validation success rates. In conclusion, in this study, we have successfully added a new automated step to determine the contributory sources of noise that determine successful or unsuccessful downstream biological validation.

1. Background

Genome-wide technologies such as gene expression microarrays offer the possibility of large-scale screening to find new genes and pathways involved in complex biological processes. However, potentially, a lot of interesting findings are lost because of issues of noise. A technique that identifies the sources of noise to guide subsequent choices of genes to validate would improve genomic technology. We describe such a general technique in this study.

The success of validating genes determined to be significantly involved in a process by using genomic technologies, such as gene expression microarrays, is quite variable. Though there are dozens of published methods to determine the most statistically significantly differentially expressed genes given two sets of microarray data, the amplitude of difference does not necessarily correspond with the likelihood of successful validation using a more sensitive measurement technique such as quantitative real time reverse transcriptase-polymerase chain reaction (RT-PCR). It is more commonly assumed that higher fold changes (greater than 2) are more likely to validate with gold standard methods such as RT-PCR. This assumption is based on previous work showing genes with this level of fold change are most likely to validate on another microarray technology [1]. It is also assumed that genes with small mRNA expression changes (under 2-fold) or low expression levels as measured by microarrays validate less often, because of the increased noise in these measurements [2]. Without a new approach, we cannot predict which of these small fold changes are reproducible without validating all genes, which is not feasible. This is unfortunate because interesting changes in biological systems may be governed by genes showing small differences in expression, especially when measured in the context of complex tissues consisting of many different cell types. In addition, many of the most biologically interesting genes are expressed at the lowest levels, such as transcription factors.

Our goal here is to consider the process of validation by RT-PCR, and to model the success and failure of validation of a gene as a variable independent from the degree of likelihood of a significant change in that gene. Our hypothesis was that a machine-learning method can be used to learn which sources of noise weighted most importantly in successful gene validation.

Failure of validation has been attributed to the multiple sources of biological and technological noise present in these high-throughput measurement systems. Potential sources of biological noise include circadian and other influences present at the time samples are acquired [3,4] and tissue complexity [5]. Sources of technical noise include variability in scanning and high back-

[*] Corresponding author. Fax: +1 650 723 7070.
E-mail address: english1@stanford.edu (S.B. English).
[1] These authors contributed equally to this work.

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.08.009

ground signals [6], and dye-specific biases in experiments using cDNA arrays [7]. As microarray generations advance, more genes and transcripts are probed, but the specific probes designed to detect the RNA may change, and this can introduce irreproducibility of measurement [8]. Noise is also introduced when samples being compared are measured on different platforms; previous studies have shown varying degrees of reproducibility for samples measured on spotted cDNA and oligonucleotide microarrays [9].

In this work, we show how we tested our hypothesis by using Bayesian networks, a method to model variables and the conditional probabilities between them, to probabilistically model the likelihood of successful validation, given input sources of noise [10]. Bayesian approaches have been previously used at many different levels of microarray analysis. At the raw image level, Bayesian networks built using pixel data have been used to model and improve the quality of microarray measurements [11]. At the processed measurement level, Baldi and Long showed how a Bayesian *t*-test could be used on estimates of distributions of gene expression measurements, and showed how this approach better compensates for lack of replicates [12]. Long, et al., then showed the success of this approach in an *E. coli* microarray experiment [13]. Ibrahim, et al., used a similar approach, with the crucial difference that a correlational structure between genes was modeled [14]. Broet, et al., used a hierarchical model taking into account multiple discrete levels of gene expression change [15]. At the multi-gene level, Bayesian and other probabilistic networks have been inferred from microarray data [16–20].

Our method is different than the many previous uses of Bayesian approaches, in that we are modeling the success and failure of validation as a variable dependent on multiple estimators of measurement noise. It is important to note that we are not modeling the likelihood of a significant change in a gene given multiple measurements, nor are we building a Bayesian network of genes.

The model we describe in this paper was built and tested on time-series measurements in the domain of diabetic retinopathy. At the onset of the analysis of the microarray data in this biological system, most of the observed differences in gene expression between mice under normal oxygen and hyperoxia conditions were small, possibly because the retina is a complex organ with different cell types. Very few genes were found to be significantly different using established methods such as Significance Analysis of Microarrays (SAM) [21]. Of the genes chosen by significance using conventional *t*-tests, only 33% initially validated by RT-PCR (examples shown in Fig. 3). Here, we show how we first created and trained a Bayesian network on the success and failures of these genes. We then validated a new set of genes guided by the network. The success rate of validation for the tested set of genes improved to 92%. Interestingly, the genes in the test set showed very small expression differences for the Affymetrix data; in most cases less than 1.5-fold. In conclusion, we increased the likelihood of selecting genes that are successfully validated by the introduction of an automated step after the traditional bioinformatics step in microarray analyses.

2. Methods

2.1. Microarray data

Mice were exposed to air containing a normal concentration of oxygen, and hyperoxia, a process that parallels early stages of

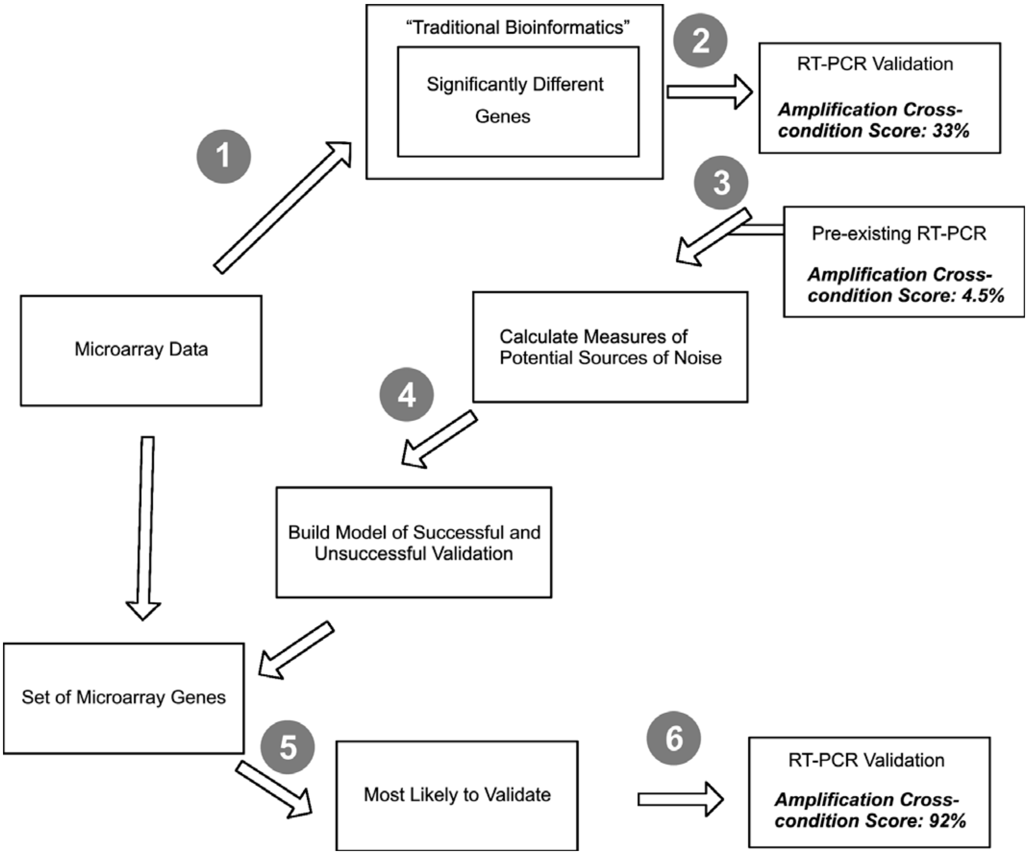
"Traditional Bioinformatics"
Significantly Different Genes
2
RT-PCR Validation
Amplification Cross-condition Score: 33%
1
3
Pre-existing RT-PCR
Amplification Cross-condition Score: 4.5%
Microarray Data
Calculate Measures of Potential Sources of Noise
4
Build Model of Successful and Unsuccessful Validation
Set of Microarray Genes
5
Most Likely to Validate
6
RT-PCR Validation
Amplification Cross-condition Score: 92%

Fig. 1. Schema used for the analysis. Microarray data was used to generate a list of significant gene expression changes using pair-wise *t*-test comparisons (1). These genes were validated by RT-PCR (2). The successes and failures for these genes as well as 22 additional genes for which there was pre-existing RT-PCR data (3) were used to train a Bayesian network. This became our predictive model for future successful validation (4). Based on the predicted likelihood of validation, we performed RT-PCR validation on 12 newly selected genes (5) and our success rate improved to 92%, despite these genes actually performing worse on the specific *t*-test comparisons than the original set (6).

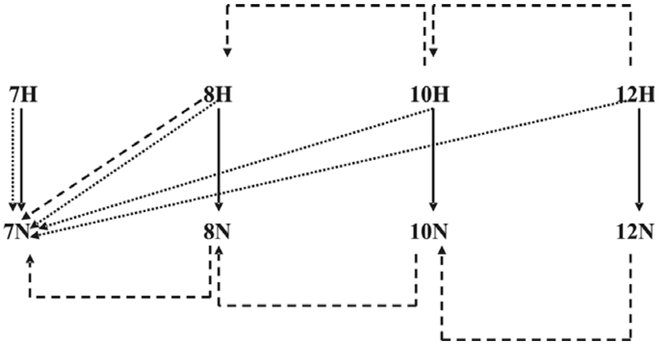

Fig. 2. Three sets of comparisons of biological interest.(1) *Consecutive comparisons* (dashed lines). (2) *Cross-condition comparisons* (solid lines). Note: The 7H →7N comparison exists for Affymetrix data alone. (3) *Baseline comparisons* (dotted lines).

diabetic retinopathy with vascular loss [22]. Litters of postnatal day-7 (P7) C57Bl/6 mice with nursing mothers were exposed to 75% ± 2% oxygen for 12-h (P7 + 12-h), 24-h (P7–P8), 72-h (P7–P10), or 120-h (P7–P12) periods in a sealed incubator. Age-matched room-air mice (P7, P8, P10, and P12) were used as controls. At each time point for both control and oxygen-treated mice, we pooled retinas from 8 different mice of 8 different litters to reduce biologic variability. The entire RNA preparation process was repeated three times at each of the four time points for both control and oxygen-treated groups, so that 24 RNA samples of 8 pooled retinas each were collected. RNA was hybridized to Affymetrix MOE3430A microarrays using established protocols. The image files were analyzed with Microarray Suite 5.0 (MAS 5.0) software. For each condition, we measured RNA at the time-points P7, P8, P10 and P12 with biological triplicates for each time-point. Data is available under accession GSE1816 at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus.

2.2. RT-PCR

RNA was prepared for real time RT-PCR from different pools of mice using the same methods as were used for microarray measurements. The same time-points were used for both techniques except for the P7 time-point under hyperoxia, which was not measured by RT-PCR. Both cDNA preparation and the quantitative real-time RT-PCR were performed as described previously [23]. Briefly, 100 ng of purified total RNA was reverse transcribed into cDNA using murine leukemia virus reverse transcriptase and random primed hexamer (Invitrogen, Bethesda, MD). The ABI Prism 7700 Sequence Detection System (Applied Biosystems) and the SYBR Green master mix kit (Qiagen) were used for detecting real-time RT-PCR products from 0.25–2.5 ng reverse transcribed cDNA samples. Cyclophilin-A, which exhibits a constant level in comparison with 18S rRNA in the retina samples, was used as the normalizer. PCR reactions for each sample were done in duplicate for target genes and triplicate for precisely quantified and 10-fold serially diluted cDNA templates. Copy numbers were determined from the cDNA standard curves. The level of target gene expression was calculated after normalizing against the 10^6 cyclophilin-A copies in each sample and data are presented as mRNA copies per 10^6 cyclophilin-A copies.

2.3. Algorithm

We first used a conventional technique to determine lists of significantly differentially expressed genes for validation by real time RT-PCR (Fig. 1). Many of these genes did not validate. A Bayesian network was trained on the success and failure of RT-PCR validation of these and other genes, given the input of multiple characteristics of the genes. After the creation of the Bayesian network, the network was then applied to a set of genes from the microarray, and a sample of genes were then chosen for RT-PCR based on having the highest predicted likelihood of validation. Each of these steps is described below.

2.4. Selection of candidate genes for the Bayesian network analysis

First, three sets of pair-wise comparisons of biological interest were defined (Fig. 2): (A) *Consecutive comparisons*, where consecutive time-points are compared with each other in both the normoxia and hyperoxia time series. (B) *Cross-condition comparisons*, where each hyperoxia time-point is compared to its corresponding normoxia time-point. (C) *Baseline comparisons*, where each of the hyperoxia time-points are compared to the day 7 normoxia time-point.

A list of genes significantly different from each pair-wise comparison in each of these sets of comparisons was generated using a two-tailed Student's *t*-test with unpaired values with Welch correction for unequal variance [24], and selecting genes with $p < 0.01$, similar to many previous studies [25,26]. Methods that compensated for multiple-hypothesis testing, such as SAM, resulted in no significantly expressed genes in most comparisons [21]. Genes were selected to be validated by RT-PCR because they were positive in at least one of the cross-condition and baseline comparisons as shown in Table 2. Eighteen genes were initially validated by RT-PCR. To this set, another 22 genes were added, for which there was pre-existing RT-PCR data measured before the microarrays were measured, bringing the total number to 40 genes used to train the network.

2.5. Database used for induction of Bayesian network

Six input variables and six output variables were calculated for each gene in the training set and were used to create the database (Table 1). The six input variables were estimates of noise based on the Affymetrix measurements and probe set characteristics. We also included two output success (outcome) variables and four output variables. The output success variables incorporate both Affymetrix and RT-PCR data and define a successful RT-PCR validation of the Affymetrix data. The network is queried to predict these two success variables. The other four output variables incorporate only RT-PCR data. These four output variables were not predicted using the Bayesian network and do not define a successful RT-PCR validation. Though we call these variables output variables, we actually have them as input during training, but they were missing when the network was actually used in the test set. While they are not crucial to the development of the network or the prediction of the output success variables, we designed them into the network to secondarily learn predictors for these RT-PCR variables from the original microarray data. We also wanted to learn the relative input of the RT-PCR and Affymetrix data in the output success variables.

The input variable *Corr-Replic-Affy* is the mean of three Pearson correlation coefficients calculated from the three possible pair-wise pairings of triplicate vectors. These are from eight measurements, four measurements in the normoxia time-series and four measurements in the hyperoxia time-series. This variable is a measure of the strength of the linear relationship within replicates.

The input variable *Unique-Affy-Probe* assumes a "yes" value if the probes in the Affymetrix probe set are unique for the measured transcript (i.e. probe set identifier ends with "_at") and a 'no' value if the probes are potentially shared between multiple transcripts of the same or different genes (i.e. probe set identifier ends with "_a_at" or "_s_at") or probes for which the rules for cross-hybridization were dropped (i.e. ends with "_x_at"). We observed differ-

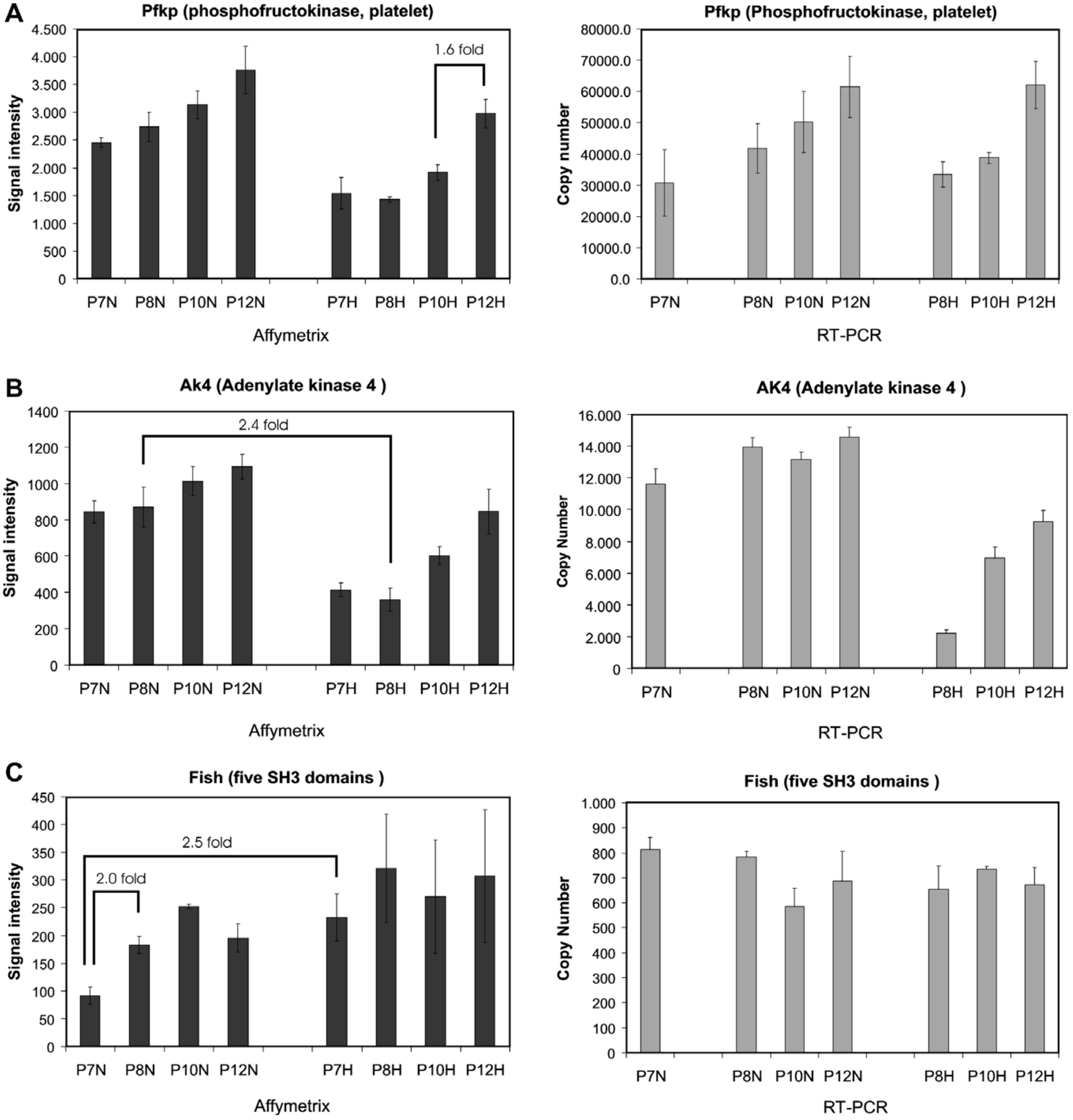

Fig. 3. Example of genes with successful and unsuccessful validation. (A) *Pfkp* is an example of a gene with high *Consec-Dir-Match* score (0.83) because the Affymetrix and RT-PCR data have similar trends across both the normoxia and hyperoxia time series, even though by microarray the highest Consecutive comparison fold difference is only 1.6-fold. (B) *Ak4* is an example of a gene with high *Ampli-Cross-Condn-Match* score (1.00), indicating that at corresponding time points, the difference between hyperoxia and normoxia seen in the RT-PCR measurements was at least as high as that seen in the Affymetrix measurements; the highest fold difference in a Cross-condition comparison is only 2.4-fold for the Affymetrix measurements. (C) *Fish* is an example of a gene that validated poorly and showed low scores for both output variables, despite having a higher maximum *Consecutive* comparison fold difference (2.0-fold) and higher maximum *Cross-condition* comparison fold difference (2.5-fold) than the other two genes.

ent patterns for the different types of probe sets for the same gene, and hypothesized that the type of probe set might influence the success of validation and the output variables. The input variable *Mean-Affy* was defined as the mean expression level for a gene across all time-points and replicates. The input variable *Pcalls* reflects the count of the Present Detection calls as reported by the Affymetrix MAS 5.0 software for each gene across all 24 microarrays [27].

The input variable *Cross-Condn-Ttest-Affy* reflects the count of significant *t*-tests in Cross-condition comparisons (as defined earlier), ranging from 0 to 4. The input variable *Consec-Baseln-Ttest-Affy* indicates the count of significant *t*-tests in Consecutive and Baseline comparisons, ranging from 0 to 9. *Corr-Replic-Rt*, *Mean-Rt*, *Cross-Condn-Ttest-Rt,* and *Consec-Baseln-Ttest-Rt* were exactly similar to their corresponding input variables, except calculated using the RT-PCR measurements.

Table 1
Variables used to train the Bayesian network

Variables (for each gene)	Definition
Input Variables	
Pcalls	Number of “Present” detection calls
Unique-Affy-Probe	Type of microarray probe
Corr-Replic-Affy	Mean pair-wise triplicate correlation
Consec-Baseln-Ttest-Affy	Number of significant Consecutive and Baseline comparison t-tests
Cross-Condn-Ttest-Affy	Number of significant Cross-condition *t*-tests
Mean-Affy	Mean signal across all 24 microarrays
Output Variables	
Corr-Replic-Rt	Within duplicate correlation
Consec-Baseln-Ttest-Rt	Number of significant Consecutive and Baseline comparison t-tests
Cross-Condn-Ttest-Rt	Number of significant Cross-condition t-tests
Mean-Rt	Mean expression level
Output Success Variables	
Consec-Dir-Match	RT-PCR measurements successfully validated the direction (increased or decreased) of change
Ampli-Cross-Condn-Match	RT-PCR measurements successfully validated the fold changes across conditions

Six input variables, four output variables, and two output success variables were used to train the Bayesian network. Details of these variables are explained in the text.

Table 2
Number of genes with significant *t*-tests for one or more of the defined biologically interesting comparisons

	Number of genes with a positive *t*-test for one or more of the Cross-condition comparisons	Number of genes with a positive *t*-test for one or more of the Baseline comparisons
Before Network	15/18 (83%)	3/18 (17%)
After Network	8/12 (67%)	None

Two strict definitions of a successful RT-PCR validation were used as output success variables on which the Bayesian network was queried to predict. These were as follows:

2.5.1. Consec-Dir-Match

In order to define whether the Affymetrix and RT-PCR data showed similar patterns for the normoxia and hyperoxia time series, we considered the set of Consecutive comparisons in Fig. 2. Biologically, these comparisons are an indicator of change in expression level for a gene between time points in the normal and hyperoxia conditions. Each of the above comparisons was assigned a 1 if the Affymetrix and RT-PCR fold changes were in the same direction (increased or decreased expression) and zero other-wise. The average of the six was called the *Consec-Dir-Match* variable and ranged between 0 and 1. A *Consec-Dir-Match* value of greater than 0.5 was defined as a success.

2.5.2. Ampli-Cross-Condn-Match

Because we were primarily interested in the effects of hyperoxia on each gene at each time-point, Cross-condition comparisons were also considered informative. A fold ratio for a gene was computed by taking the arithmetic mean of replicate expression measurements for that gene under hyperoxia and normoxia at a given time point and dividing the means. Each of the 3 Cross-condition comparisons was assigned a 1 if the fold ratio calculated using the RT-PCR measurements was equal or higher in magnitude than the fold ratio calculated using the Affymetrix measurements, and greater than a minimum relevance threshold of 1.4-fold; otherwise a zero was assigned. The average over the three fold changes is a factor between 0 and 1 and defines the *Ampli-Cross-Condn-Match* variable. An *Ampli-Cross-Condn-Match* variable value of greater than 0.5 was defined as a success. The training data set is available as Supplemental Table 1: english supplemental table 1.xls.

Regarding the threshold, it does not make sense to employ a threshold of less than 0.5 for either of the success variables, as that would mean that less than half of the comparisons met the success criteria. For both success variables, if the threshold (ranging from 0 to a maximum value of 1) is plotted against performance, the trend of performance versus threshold is increasing. As a result, a threshold of 0.5 is a reasonable and conservative pick. The performance is defined as the improvement in the success rate of the test set compared to the initial set of 18 genes picked on bioinformatics criteria that formed part of the training set.

2.6. Statistical modeling of the Bayesian network

This database of variables (Supplemental Table 1) was then used to create the structure of and ascertain the distributions for the Bayesian network. This was done using *Bayesware Discoverer*, which explores a working subset of models as defined by the user by identifying an order with which the variables in the database will be evaluated [28]. The higher the assigned rank of a variable, the greater the number of other variables that will be tested as potential precedent variables; thus, output success variables were placed at the highest rank. The threshold Bayes factor was set at 3, a conservative value that reduces false positives. No limit was put on the maximum number of parents (precedent variables) that a variable can have. The Prior Precision, encoding the confidence of prior distributions, was set at the default value of 1. As the construction of Bayesian networks is markedly difficult when handling continuous variables, all variables were discretized. In most cases, the continuous variables were discretized into two bins of equal length by taking the range of values, and creating two bins using the midpoint of the range. In the case of *Cross-Condn-Ttest-Affy*, three bins were used.

After the Bayesian network was created using the training data, a set of the genes measured on the microarrays was evaluated using the network. We identified a set of Affymetrix probe-sets whose expression values across time-points correlated to one or more of five markers of endothelial cells with a Pearson correlation coefficient of greater than 0.8. The 5 genes used as markers were ICAM-1, PECAM-1, Tie-1, Tie-2 and VE-cadherin. The number of probe-sets correlating to each of these markers ranged from 220 to 450. The network was used to predict which of these genes were likely to successfully validate by RT-PCR. Based on the predicted high likelihood of validation in the two output success variables *Consec-Dir-Match* and *Ampli-Cross-Condn-Match*, a sample of 12 genes was selected to test the network. Significance in the difference in validation success rates was determined using a χ^2 test in Microsoft Excel 2002 (Redmond, Washington). The test data set is available as Supplemental Table 2: english supplemental table 2.xls.

3. Results

3.1. Training and testing the network

Our goal was to develop an automated method to predict which genes implicated in an experiment using microarray measurements would be most likely to successfully validate by RT-PCR, trained using the experience from a starting set of genes that successfully and unsuccessfully passed validation. To improve our ability to generalize findings, the training set consisted of two sets of genes. We initially attempted validation of 18 genes chosen on the basis of the statistically significant differences between comparison groups, as defined by *t*-tests. This set of 18 genes became the training set. To add data on genes failing validation, we then added another 22 genes to the training set for which there was pre-existing RT-PCR data performed under the same biological conditions, along with the microarray data acquired after the RT-PCR measurements. We gauged the success and failure of this validation and calculated estimates representing potential sources of noise. These inputs were used to create and train a Bayesian network. Two output success variables, *Consec-Dir-Match* and *Ampli-Cross-Condn-Match* were defined for the successful validation of Affymetrix microarray data by RT-PCR, using strict criteria. Success using the *Consec-Dir-Match* variable identifies genes that show similar patterns of expression across the time series for microarray and RT-PCR data. Success using *Ampli-Cross-Condn-Match*, which identifies genes with expression differences between corresponding normoxia/hyperoxia time-points that can be validated, was the most relevant to the biological question being asked. Of the initial set of 18 genes, selected using *t*-tests, the rate of genes with successful *Ampli-Cross-Condn-Match* was only 33% and the rate of genes with successful *Consec-Dir-Match* was 50%. Of the additional 22 genes, only one had a successful *Ampli-Cross-Condn-Match* (4.5%). In the full training set, there were 17% of genes with successful *Ampli-Cross-Condn-Match* and 57% of genes with successful *Consec-Dir-Match*. Fig. 3 shows examples from the training set of genes that served as successfully and unsuccessfully validated genes.

3.2. Results from the Bayesian network

After extracting the network from the data (Fig. 4), we applied it to a set of genes on the microarray. The retina is a complex organ made up of different kinds of cells, of which endothelial cells are a small proportion. Under conditions of hyperoxia, the endothelial cells undergo vaso-obliteration and thus are likely to show genes changes between normoxia and hyperoxia. These gene changes are likely very small as the whole retina was used in the experiment. It was of interest to identify potential novel endothelial cell genes that showed expression differences between normoxia and hyperoxia conditions that could be validated by RT-PCR. Towards this end, we applied the network to a set of genes whose mean expression values correlated to one or more of five known markers of endothelial cells. Running the genes through the network identifies genes that are more likely to meet our criteria for successful validation by RT-PCR. And as genes from other cell types would also show gene changes between normoxia and hyperoxia conditions, restricting to genes that show a correlation with known markers of endothelial cells may be more likely to identify genes potentially expressed in endothelial cells.

We picked a set of 12 genes to validate out of the genes most likely to have a successful *Consec-Dir-Match* and *Ampli-Cross-Condn-Match*, our strict definitions of biological validation. The distributions of microarray measurements for the entire set of genes,

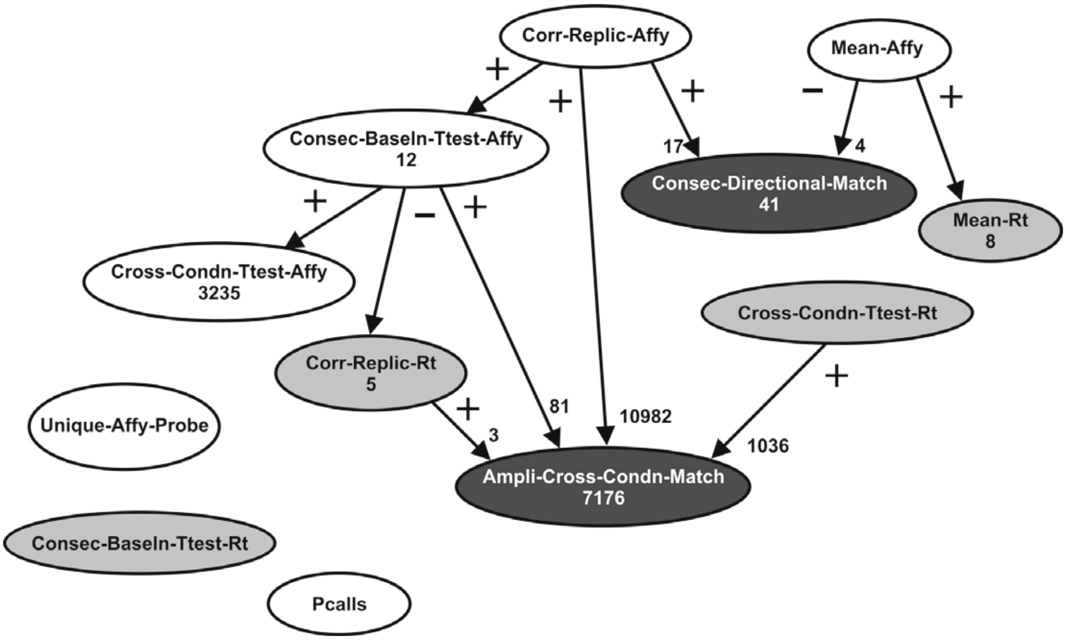

Fig. 4. Topology of the identified Bayesian network. The identified Bayesian network consists of nodes (variables) connected by significant conditional dependencies between variables. The six input variables (black text on white) reflect calculated indications of biological or technical noise in microarray measurements. The four output variables (black text on gray) indicate similar calculations made using the RT-PCR measurements. The two output success variables (white text on black) reflect the event of successful biological validation (as strictly defined in the text). Dependencies can be positively (plus sign) or negatively (minus sign) correlated. Numbers within nodes indicate the Bayes factor between the given set of parent dependencies (i.e. the most probable) and no parent dependencies. For example, as a model explaining the given training data, the set of four parent dependencies into the *Ampli-Cross-Condn-Match* variable is 7176 times more likely than no parent dependency. Numbers on arrows indicate the Bayes factor between the given set of parent dependencies and the set without the annotated dependency. For example, adding the *Corr-Replic-Affy* parent dependency to the other three parent dependencies into *Ampli-Cross-Condn-Match* improves the likelihood of explaining the output variable by 10,982-fold. The interpretation of this network is that *Corr-Replic-Affy*, which measures the intra-replicate correlation between gene measurements by microarray, is the most informative input variable in the network in that it is strongest in conditioning both output success variables.

Table 3
RT-PCR validation rates of genes selected without and with the network

	Number of genes with a *Consec-Dir-Match* > 0.5	Number of genes with an *Ampli-Cross-Condn-Match* > 0.5
Before Network	9/18 (50%)	6/18 (33%)
After Network	8/12 (67%)	11/12 (92%)

The success rates without the network are for the set of 18 genes which later formed part of the training data for the network.

the training set, and the test set are similar as determined by ANOVA ($p = 0.09$). We found that performing RT-PCR on genes where the Bayesian network predicted successful validation resulted in a significant improvement in the rate of successful validation (Table 3). Compared to the results from our initial set of genes selected on bioinformatics criteria, the rate of genes with *Consec-Dir-Match* improved from 50% to 67%, and for *Ampli-Cross-Condn-Match*, the rate improved significantly from 33% to 92% (χ^2 p = 0.000005). Of note, the genes chosen for the test set actually score more poorly than the original set of genes based on the traditional bioinformatics criteria (Table 2) suggesting that in this experiment, successful validation is governed more by choosing genes with less noisy measurements than by choosing genes based on statistical significance.

The goal of Bayesian network extraction is to find the model with the maximum marginal log-likelihood compared to the next most probable model given the data. The marginal log-likelihood of the entire network is the sum of the marginal log-likelihood of each dependency. In order to do this, for each variable, we find the most probable parent dependencies given the data. The marginal log-likelihood of each variable is used to compute the Bayes factor. For a given variable, the Bayes factor is computed as a ratio between the most probable dependency and other sets of dependencies, including the null hypothesis of no dependence. The Bayes factor serves as a measure of how many times more probable is one dependency compared to another.

Our resultant Bayesian network (Fig. 4) identified a single most informative variable, *Corr-Replic-Affy*, which measures the strength of the linear relationship within replicates using the Pearson correlation coefficient. Both output success variables conditionally depend on *Corr-Replic-Affy*. The likelihood of *Consec-Dir-Match* and *Ampli-Cross-Condn-Match* explaining the output variable increased 17- and 10,982-fold, respectively, after the addition of *Corr-Replic-Affy* as a parent dependency. In addition, these were positive correlations, in that a higher value for this variable, indicative of stronger correlation between the triplicate measurements by microarray, was associated with greater likelihood of validation by RT-PCR. We used this variable as the primary basis for selecting genes for the test set, for further validation.

Besides *Corr-Replic-Affy*, the only other input variable conditioning *Ampli-Cross-Condn-Match* is *Consec-Baseln-Ttest-Affy*. The relationship is positive, in that the greater the number of significant comparisons for a gene, the greater the likelihood of a positive match for that gene, but weaker than *Corr-Replic-Affy*. The output variables *Corr-Replic-Rt* and *Cross-Condn-Ttest-Rt* also condition *Ampli-Cross-Condn-Match.*

Other than *Corr-Replic-Affy*, the other variable that conditions *Consec-Dir-Match* is *Mean-Affy*, which represents the mean expression level across all time-points as measured by the microarrays. Specifically, for genes with high *Corr-Replic-Affy,* if the mean expression measurements are high (in this experiment, above 5600 arbitrary Affymetrix units), the likelihood of *Consec-Dir-Match* is 90.0%, compared to 99.3% if the mean expression levels are lower. For genes with low *Corr-Replic-Affy*, the likelihood of *Consec-Dir-Match* is 80.6% if the mean expression level is low, which drops to only 5.6% if the mean expression level is high. This is a surprising result, as genes at higher expression levels are commonly considered as being measured with less noise. But *Mean-Affy* is a weaker parent variable for *Consec-Dir-Match* than *Corr-Replic-Affy*, as indicated by the fact that adding the *Mean-Affy* dependency improves the likelihood of explaining *Consec-Dir-Match* only by 4 as compared to 17-fold for *Corr-Replic-Affy*.

Another unexpected relation was seen in that *Consec-Baseln-Ttest-Affy* conditions *Corr-Replic-Rt*, but this relationship is negative. In other words, the greater the number of significant comparisons for a gene, the lower the likelihood of a high correlation between replicates, as measured by RT-PCR.

As expected, the input variable *Mean-Affy* conditions the output variable *Mean-Rt*. *Mean-Affy* and *Mean-Rt* are the mean expression measurements across all time-points for a single gene as measured by Affymetrix microarray and RT-PCR respectively. It is encouraging that there is a direct relationship between the two variables, and this can serve as a positive control.

The four non-success output variables also provide useful information to the experimentalist. For example, it is informative that correlation between RT-PCR replicates as well as significant *t*-tests for RT-PCR both contribute to a successful RT-PCR validation of Affymetrix data. It is also informative that none of the Affymetrix nodes points to *Cross-Condn-Ttest-Rt,* while this is not the case for *Corr-Replic-Rt.*

Finally, the orphaned input and output variables, whose distributions neither condition nor are conditioned by any other variables, also provide information. These variables include *Pcalls*, *Unique-Affy-Probe*, and *Consec-Baseln-Ttest-Rt*. In other words, the type of probe set and the number of Detection calls as reported by the Affymetrix MAS 5.0 software did not significantly influence the likelihood of a successful validation.

4. Discussion

Using Bayesian networks, we were successful in our goal of creating an automated method that can determine which factors contribute most significantly to the successful RT-PCR validation of genes implicated from microarray measurements. We trained a Bayesian network on the successes and failures of an initial set of genes using input sources of noise. This helped us identify predictors of successful validation that we did not know of a priori. Significantly, the Bayesian network can allow for the successful validation of genes that show small fold changes, such as one shown in Fig. 3A; many biologically interesting genes are in this category. Of the successfully validated genes in our test set, most showed microarray data fold changes of less than 1.5-fold. The network can also allow for the validation of genes that show low expression levels.

With the Bayesian network created on our preliminary experimental data, we determined that optimizing the selection of genes based on "within gene" correlation of microarray measurements would most greatly improve our rate of biological validation, as defined by equal or greater fold-changes by RT-PCR between hyperoxia and normoxia at each time point than by microarray. We have shown that optimizing "within gene" correlation in the subsequent selection of genes allowed us to significantly increase rates of validation, as the model predicted.

When used to boost validation rates in this experiment, the network provided few results that were counter-intuitive. The inverse relationship between a gene's mean expression level across all microarrays (*Mean-Affy*) and the likelihood of validation (*Consec-Dir-Match*) is unexpected, as genes at higher expression levels are thought to have measures that are less influenced by technical noise. In actuality, gene expression differences between time points in both conditions were less likely to be reproducible when mean gene expression was higher, even when only the direction of change was evaluated. Similarly, the degree of correlation of a gene within replicate RT-PCR measurements (*Corr-Replic-Rt*) drops slightly as the number of significant Base-line and Consecutive comparisons for that gene increases (*Consec-Baseln-Ttest-Affy*), though this relation is relatively weak (evidenced by the Bayes Factor of only 5 as compared to no relationship existing). Further interpretation of these unexpected relations is limited by the size of the training set of genes.

Use of the Bayesian network successfully allowed us to find those variables, representing sources of noise, that when optimized would allow a greater rate of validation. However, the lack of connections between variables was also illuminating. For example, the likelihood of validating a difference between hyperoxia and normoxia and finding an increased fold-change (*Ampli-Cross-Condn-Match*) was not conditioned on the number of "Present" Detection calls (*Pcalls*), or the mean expression level (*Mean-Affy*). Intuitively, without the network, we would have considered choosing genes with higher expression levels or with many "Present" calls, as recommended by others [29]. With the network, we instead found an alternate variable (*Corr-Replic-Affy*) to optimize that would yield a higher rate of validation.

Although a high correlation between replicates was the best predictor of success for this particular experiment, we are not making the case that correlation within replicates will be the best predictor for another experiment. We are making the case that it is possible to statistically learn from failures in validation and improve choices midstream within an experiment.

The approach that we used has some limitations. We recognize that there are many more potential sources of noise those discussed here, but for the purposes of the present paper we have only discussed a few. For example, our approach could be improved by adding variables representing the distribution of raw intensities over each probe pair of every probe set. Another useful variable might be the number of probes that match sequences in the NCBI Reference Sequence (RefSeq) database, as this can implicate the validity of expression measurements [30]. We acknowledge that over-fitting may be a serious issue as not every gene was validated using RT-PCR, and only a small sample of genes was used to train and test the network. It is also true that we cannot completely discount the possibility of an introduced bias based on the selection of the test set. However, two separate issues are involved (1) identifying gene changes that are relevant to the experimentalist and (2) identifying reproducible (verifiable) gene changes. The methods used to select genes for the training and the test sets both target relevant gene changes. The second issue, the reproducibility of gene changes, is dependent on noise factors. Another factor is that straight discretization by range may oversimplify the complex distributions of each variable. In addition, our initial bioinformatics criteria of selecting genes by count of significant *t*-tests may appear naïve when alternate methods are available, including methods involving permutation testing, such as SAM [21]. However, using SAM we obtained no significantly changed genes at the earlier time-points, possibly because of the smaller number of replicates. As our goal was to increase the number of true positives that successfully validate by a gold-standard method such as RT-PCR, the technique used to identify significantly changed genes is less critical. Finally, we also acknowledge that our results may not be generalizable to other platforms and analysis methods other than differential expression. However, whichever the platform used for a microarray study, there will be noise factors that negatively affect the reproducibility of the data. And learning these sources of noise and using the knowledge to guide further choices for validation might be expected to improve the reproducibility.

In spite of the above limitations, we suggest that our method can improve the optimization of countless other parameters which are not presently used or weighted highly by existing bioinformatics methods.

5. Conclusions

This approach represents an improvement in the standard methodology of genomic exploration. Based on our results, we suggest the addition of a new automated step to determine the contributory sources of noise that determine a successful or unsuccessful RT-PCR validation, and we suggest taking advantage of this step midway through the validation of any experiment guided by microarray data. This process can be tailored to different experiments and conditions in other biological systems. Though we used Bayesian networks as our automated prediction step, other supervised learning methods can be used as well, such as support vector machines or decision trees [31,32]. If output variables are carefully crafted to match ones own definition of a successful validation, the resultant network may significantly improve validation efforts.

Acknowledgments

The work was supported by grants from the V. Kann Rasmussen Foundation, National Eye Institute, National Institute of Diabetes and Digestive and Kidney Diseases, National Cancer Institute, National Institute of Allergy and Infectious Diseases, the Harvard-MIT Division of Health Sciences and Technology, and the Lawson Wilkins Pediatric Endocrine Society.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2008.08.009.

References

[1] Tan PK, Downey TJ, Spitznagel Jr EL, Xu P, Fu D, Dimitrov DS, et al. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 2003;31:5676–84.

[2] Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 2002;12:111–39.

[3] Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, et al. Coordinated transcription of key pathways in the mouse by the circadian clock. Cell 2002;109:307–20.

[4] Storch KF, Lipan O, Leykin I, Viswanathan N, Davis FC, Wong WH, et al. Extensive and divergent circadian gene expression in liver and heart. Nature 2002;417:78–83.

[5] Huminiecki L, Lloyd AT, Wolfe KH. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. BMC Genomics 2003;4:31.

[6] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30:e15.

[7] Fang Y, Brass A, Hoyle DC, Hayes A, Bashein A, Oliver SG, et al. A model-based analysis of microarray experimental error and normalisation. Nucleic Acids Res 2003;31:e96.

[8] Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, et al. Reproducibility of gene expression across generations of Affymetrix microarrays. BMC Bioinformatics 2003;4:27.

[9] Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 2002;18:405–12.

[10] Heckerman D. A tutorial on learning with bayesian networks. In: Redmond WA, Microsoft Research. 1995.

[11] Hautaniemi S, Edgren H, Vesanen P, Wolf M, Jarvinen AK, Yli-Harja O, et al. A novel strategy for microarray quality control using Bayesian networks. Bioinformatics 2003;19:2031–8.

[12] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t* -test and statistical inferences of gene changes. Bioinformatics 2001;17:509–19.

[13] Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. J Biol Chem 2001;276:19937–44.

[14] Ibrahim JG, Chen M-H, Gray RJ. Bayesian models for gene expression With DNA microarray data. J Am Stat Assoc 2002;97:88–99.

[15] Broet P, Richardson S, Radvanyi F. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. J Comput Biol 2002;9:671–83.

[16] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000;7:601–20.

[17] Husmeier D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 2003;19:2271–82.

[18] Kim S, Imoto S, Miyano S. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. Biosystems 2004;75:57–65.

[19] Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet 2004;36:1090–8.

[20] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003;34:166–76.

[21] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116–21.

[22] Smith LE, Wesolowski E, McLellan A, Kostyk SK, D'Amato R, Sullivan R, et al. Oxygen-induced retinopathy in the mouse. Invest Ophthalmol Vis Sci 1994;35:101–11.

[23] Shih SC, Robinson GS, Perruzzi CA, Calvo A, Desai K, Green JE, et al. Molecular profiling of angiogenesis markers. Am J Pathol 2002;161:35–41.

[24] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes in C: the art of scientific computing. Cambridge: Cambridge University Press; 1993.

[25] Costigan M, Befort K, Karchewski L, Griffin RS, D'Urso D, Allchorne A, et al. Replicate high-density rat genome oligonucleotide microarrays reveal hundreds of regulated genes in the dorsal root ganglion after peripheral nerve injury. BMC Neurosci 2002;3:16.

[26] Hakak Y, Walker JR, Li C, Wong WH, Davis KL, Buxbaum JD, et al. Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. Proc Natl Acad Sci USA 2001;98:4746–51.

[27] Affymetrix. Statistical Algorithms Description Document. In; 2002.

[28] Sebastiani P, Ramoni M, Crea A. Profiling your customers using Bayesian networks. SIGKDD Explorations 2000;1:1.

[29] Shippy R, Sendera TJ, Lockner R, Palaniappan C, Kaysser-Kranich T, Watts G, et al. Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. BMC Genomics 2004;5:61.

[30] Mecham BH, Wetmore DZ, Szallasi Z, Sadovsky Y, Kohane I, Mariani TJ. Increased measurement accuracy for sequence-verified microarray probes. Physiol Genomics 2004;18:308–15.

[31] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;97:262–7.

[32] Quinlan J. C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, Calif.; 1992.