Available online at www.sciencedirect.com
ScienceDirect
Genomics 89 (2007) 460–470
GENOMICS
www.elsevier.com/locate/ygeno

Empirical comparison of tests for differential expression on time-series microarray experiments

Ernest A. Fischer, Michael A. Friedman, Mia K. Markey *

Department of Biomedical Engineering, University of Texas at Austin, Campus Code C0800, 1 University Station, Austin, TX 78712, USA

Received 13 March 2006; accepted 30 October 2006
Available online 22 December 2006

Abstract

Methods for identifying differentially expressed genes were compared on time-series microarray data simulated from artificial gene networks. Select methods were further analyzed on existing immune response data of Boldrick et al. (2002, Proc. Natl. Acad. Sci. USA 99, 972–977). Based on the simulations, we recommend the ANOVA variants of Cui and Churchill. Efron and Tibshirani's empirical Bayes Wilcoxon rank sum test is recommended when the background cannot be effectively corrected. Our proposed GSVD-based differential expression method was shown to detect subtle changes. ANOVA combined with GSVD was consistent on background-normalized simulation data. GSVD with empirical Bayes was consistent without background correction. Based on the Boldrick et al. data, ANOVA is best suited to detect changes in temporal data, while GSVD and empirical Bayes effectively detect individual spikes or overall shifts, respectively. For methods tested on simulation data, lowess after background correction improved results. On simulation data without background correction, lowess decreased performance compared to median centering.

Keywords: Microarray analysis of gene expression; DNA microarray; Area under curve; ROC curve; Rank-sum tests; Analysis of variance

Microarray technology has allowed for cell- or organism-wide observation of transcriptional activity via relative measurements of most of the cell's or organism's mRNA. With such large-scale measurements comes the potential for spurious conclusions, particularly considering the numerous sources of experimental error and the typical lack of experimental replication.

Many challenges in microarray analysis have been investigated. For example, the normalization of microarray data, while by no means a closed issue, has been approached in a systematic fashion, as with the comparisons of transformations for noise types by Cui et al. [1]. Likewise, methods for class comparison of static microarray data have been shown to account for lack of replication and/or multiple simultaneous comparisons [2–9]. Furthermore, microarray data normalization is essential for effective determination of differentially expressed genes. Hoffmann et al. demonstrated the profound effect the choice of normalization technique had on analysis results [10]. Qin et al. observed a similar disparity in results depending on measured background signal subtraction [11]. Therefore, in this study we considered the impact of normalization and whether the background is removed.

Methods are available to identify differentially expressed genes in static microarray data (e.g., normal versus cancer). For discussion and comparison of some of these statistical tests refer to [5,11–15]. On the other hand, many microarray experiments consider not only different experimental conditions, but also the effect of the treatment over time, such as in the immune response of an animal after infection. Analyzing time series microarray expression data has its own challenges, such as the strong autocorrelation between successive time points [16]. Additionally, in time-series microarray experiments changes in the timing or pattern of expression are as important as changes in concentration magnitude but may not be as readily revealed with existing methods. Current methods find differences based, in general, on deviations from mean normal expression. This approach is reasonable for static microarray data but is not necessarily appropriate for time-course microarray data.

Some methods have been proposed specifically for identifying differentially expressed genes in time-course data. However,

* Corresponding author. Fax: +1 512 471 0616.
E-mail address: mia.markey@mail.utexas.edu (M.K. Markey).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.10.008

most are not generally applicable [17], including cluster analysis [18,19], decomposition [20–22], custom-tailored models dependent on the specific experiment [23], quadratic regression [24], and B-spline-based approaches [17,25]. These methods are a step in the right direction for analysis of time-series microarrays because they aim to make comparisons based on pattern differences. Yet, for example, B-splines are appropriate only for relatively long (>10 time points) experiments [16] and quadratic regression requires multiple replicates per time point to be effective. Differential expression through cluster analysis requires the definition of potential candidate profiles of interest. These candidate profiles must be defined a priori or possibly deduced from the data via decomposition. Even then, determining differential expression is not straightforward. Both B-splines and cluster analysis/decomposition methods work adequately on time-series data such as the Spellman et al. yeast cell cycle [26], which has many time points and only a few expected biologically meaningful patterns of expression. In general, though, they are not appropriate for comparing individual genes [16,17].

Alter et al. showed that GSVD provides a mathematical framework for the comparison of two expression datasets and illustrated GSVD's ability to discover similar regulatory programs across species [20]. In essence, their application of GSVD allows for the classification of genes across two microarray datasets to the common set of expression patterns resulting from the decomposition. We have applied GSVD with an added means of quantifying the relationship between the same gene in two paired expression datasets to provide a metric for class comparison (over/underexpression versus no change). Whether representative of the underlying biology or not, the decomposition results in a set of characteristic patterns or components for the datasets, which, we believe, makes this method well suited to comparing temporal microarray datasets. However, as we demonstrate, GSVD is not most appropriate in all situations. Rather, GSVD is best suited to detect only those changes in expression characterized by subtle variations or small deviations in magnitude from the normal expression levels.

To characterize the conditions, if any, under which currently available tests for differential expression are capable of detecting changes in time-series microarray experiments, we compared each method's performance on data simulated from the artificial gene networks of Mendes et al. [27]. Typically, class comparison methods are tested on random data, with a given number of "genes" offset, providing the pool of changed genes. Generating data in this fashion does not represent the true complexity of a real expression dataset and, more importantly, assumes that the actual differences are those with a statistically measurable discrepancy in mean expression between samples. Real data would be the ideal choice for testing the effectiveness of the comparison methods; however, evaluation is difficult without knowledge of the expected response. Gene network simulations provide an attractive alternative as they include the regulatory relationships that are an important influence on gene transcription levels, creating data with realistic responses to experimental conditions.

To supplement the simulation studies, tests for differential expression were also compared on a dataset described by Boldrick et al. measuring the response of human peripheral blood mononuclear cells (PBMCs) to the gram-negative bacterial pathogen *Escherichia coli* [28]. Briefly, gram-negative bacteria are known to trigger a response through the toll-like receptor (TLR) signaling pathway, specifically through TLR4 [29]. The downstream signaling triggered by the activation of TLR4 activates the transcription factors (TFs) NF-κB (composed of the NFKB1 and RELA subunits) and AP-1 (composed of the FOS and JUN subunits) [30–32], among others. We evaluated the downstream targets of NF-κB and AP-1 in a manner similar to that of the artificial networks.

Results

Justification of gene network "truth"

We took a number of approaches to justify the choice of truly changed genes as those genes directly influenced by the

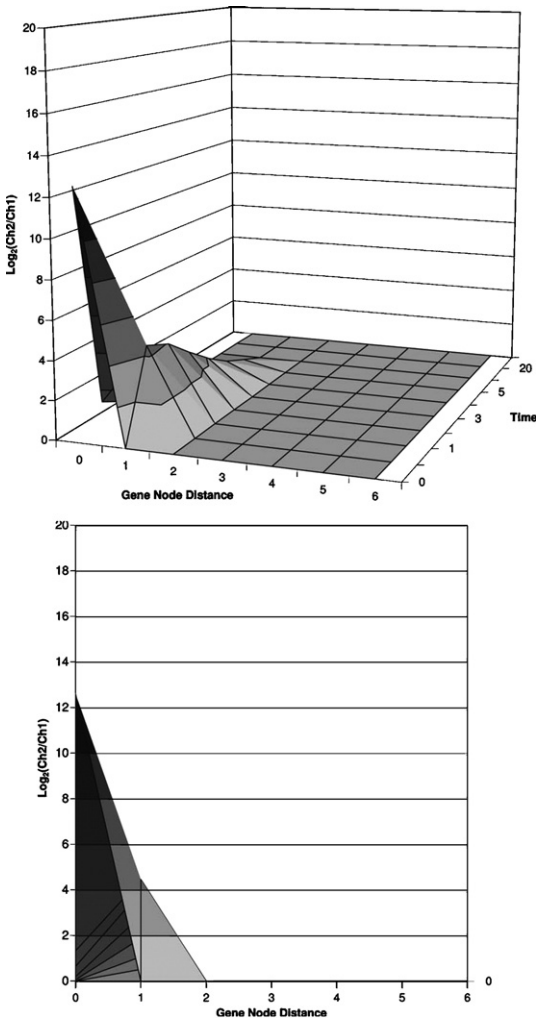

Fig. 1. Log_2 of the noise-free, simulated Ch2/Ch1 data versus the distance from the changed gene for the scale-free network Century SF-007 was averaged across all simulations in the low initial concentration, perturbation (pt–lo) simulations. Shown are all time points (top) and time 0 only (bottom).

explicitly altered gene, including the altered gene itself. First we examined the data before the noise had been applied. Example graphs of the averaged $\log_2$ ratios of the raw, noise-free data versus distance from the changed gene are shown in Figs. 1 and 2. The raw data for all networks tended to have the largest average ratio at one or two nodes away from the changed gene. Based on this analysis, the truth could reasonably include genes one and two nodes away. Therefore, all comparisons were carried out defining truth either way. For all methods on all networks, the results were higher when the truth was defined as one node distance from the changed gene. In most cases, the same methods had the best relative performance with either definition of truth for the respective network and condition. In the few that were different, the method that had been highest using one node distance as the truth was not statistically different that the best performing method when using the two-node distance truth. Thus, the conclusions of this study are not dependent on defining truly changed genes as those within one node rather than two nodes from the altered gene. The results with truth defined as one node away are presented here and those for two node distances away are available in the supplementary data.

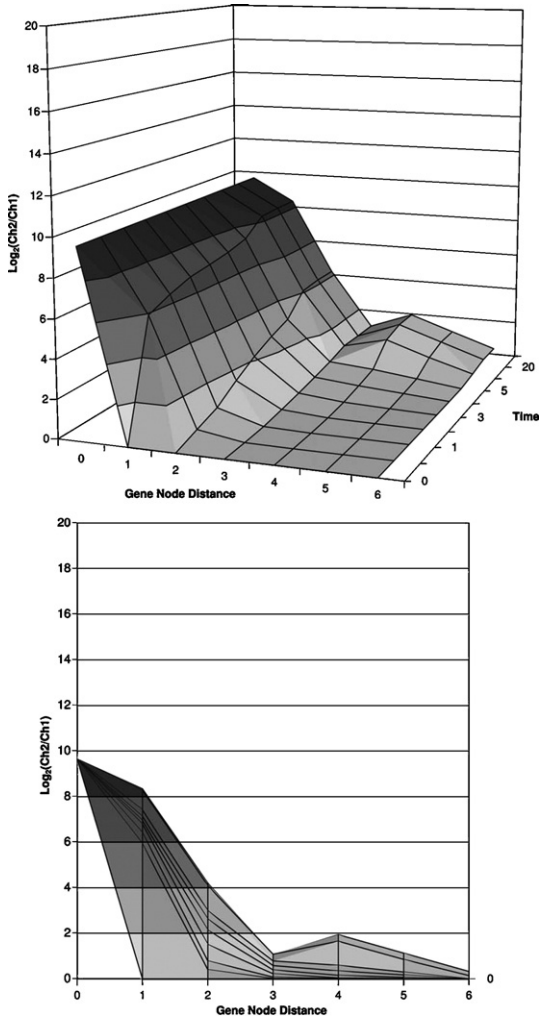

Fig. 2. Log_2 of the noise-free, simulated Ch2/Ch1 data versus the distance from the changed gene for the Erdös and Rényi network Century RND-005 was averaged across all simulations in the low initial concentration, mutation (mt–lo) simulations. Shown are all time points (top) and time 0 only (bottom).

Table 1
Comparison of normalizations for the SAM *t* statistic

Normalization	Scale-free			
	pt–lo	pt–hi	mt–lo	mt–hi
Bck subt/lowess (0.3)	0.74	0.58	0.93	0.72
Bck subt/med center	0.66	0.63	0.92	0.76
Lowess only (0.3)	0.60	0.48	0.84	0.58
Median center only	0.56	0.58	0.76	0.67

Analysis of simulated data

Similar to the findings of Hoffmann et al. [10] and Qin et al. [11], results for all methods were dependent on the choice of normalization. Since we superimposed uniform noise onto the simulated data, only global median centering or lowess normalizations were implemented. In addition, comparisons were performed on data with and without global background subtraction to analyze the effects of background on time-series microarray analysis. Table 1 shows an example of the dependence on normalization.

Owing to the normalization dependency, each comparison method was carried out using each normalization, on either background-subtracted or non-background-subtracted data. The highest average areas under the curve (AUC) are reported in Table 2 (global background-subtracted data) and Table 3 (data without background subtraction). Figs. 3 and 4 display examples of the resulting receiver operating characteristic (ROC) curves. In particular, the GSVD, F3, and empirical Bayes Wilcoxon rank sum (EB WRS) results are shown in Fig. 3 for a single, perturbation low initial concentration (pt–lo) network simulation. Fig. 4 has the results for a single, knockout (mt–lo) network simulation using GSVD, F2, and EB WRS. Refer to the available online supplement for the AUC by network and simulation for each class comparison method.

Regardless of normalization, we observed a decrease in detection for all comparison methods on all networks within perturbation or mutant conditions under which the gene was changed to a significantly low value compared to when the gene was initially overexpressed (pt–lo vs pt–hi and mt–lo vs mt–hi). Also, results did seem to change depending on the network topology, as scale-free network results were higher than those for small-world and Erdös and Rényi networks.

For the scale-free networks simulations after background subtraction (Table 2), the GSVD comparison, F2, and F3 outperformed all other methods under the pt–lo condition ($p<0.001$ for GSVD over significance analysis of microarrays (SAM)). GSVD outperformed all other methods under the pt–hi condition ($p<0.001$ for GSVD over F3). ANOVA variants F2 and F3 had the best performance under the knockout condition (mt–lo; $p=0.001$ for F2 over B) and the GSVD, F2, and F3 performed best in the constitutive overexpression simulations (mt–hi; $p<0.001$ for F2 over EB). On the small-world network, background-subtracted data, F2 and F3 performed significantly

Table 2
Average AUC for all simulations per network type and experimental condition, by comparison method on background-subtracted (baseline normalized) data

Comparison method (normalization)	Network type											
	Erdös and Rényi				Scale-free				Small-world			
	pt–lo	pt–hi	mt–lo	mt–hi	pt–lo	pt–hi	mt–lo	mt–hi	pt–lo	pt–hi	mt–lo	mt–hi
EBayes WRS (none)	0.54	0.50	0.84	0.66	0.59	0.55	0.90	0.72	0.62	0.59	0.84	0.80
t test (lowess)	0.64	0.54	0.88	0.65	0.68	0.54	0.91	0.68	0.67	0.60	0.87	0.80
SAM (lowess)	0.67	0.54	0.88	0.66	0.74	0.58	0.93	0.72	0.71	0.63	0.89	0.83
MLE-T (lowess)	0.65	0.53	0.89	0.65	0.66	0.51	0.91	0.66	0.64	0.60	0.87	0.79
Cyber-T (lowess)	0.62	0.53	0.84	0.58	0.65	0.56	0.91	0.66	0.62	0.59	0.84	0.70
B (lowess)	0.67	0.55	0.89	0.67	0.74	0.57	0.93	0.71	0.70	0.64	0.90	0.83
ANOVA–F1(lowess)	0.63	0.52	0.89	0.66	0.67	0.51	0.89	0.67	0.66	0.58	0.87	0.79
ANOVA–F2 (lowess)	**0.70**	0.54	**0.89**	**0.68**	0.83	0.62	0.96	0.78	0.76	0.65	**0.91**	**0.85**
ANOVA–F3 (lowess)	0.69	0.55	0.89	0.68	**0.84**	0.64	**0.97**	0.79	**0.76**	**0.67**	0.90	0.85
GSVD (none)	0.61	**0.56**	0.55	0.55	0.83	**0.78**	0.73	**0.81**	0.70	0.64	0.61	0.64

The highest result per network type and condition is shown in bold.

best in the perturbation low simulations (pt–lo; $p<0.001$ for F3 over SAM). F2 and F3 had the highest average AUC in the pt–hi, mt–lo, and mt–hi simulations, although many methods were statistically indistinguishable. GSVD performed worse than the other methods in both mutant simulations of scale-free and small-world networks. Of the *t* test and modifications on background-subtracted data SAM and B performed best, albeit not as good as methods mentioned above.

The AUC obtained on background-subtracted data (Table 2) were higher than for data without background subtraction (Table 3). On the scale-free simulations, GSVD significantly outperformed other methods under all conditions except for mt–lo (pt–lo, $p<0.001$ for GSVD over *t;* pt–hi, $p<0.001$ for GSVD over F3; mt–hi, $p<0.001$ for GSVD over EB WRS). EB WRS performed significantly best on the mt–lo scale-free network simulations ($p=0.001$ for EB WRS over *t*). EB WRS also performed best on the mt–lo and mt–hi small-world simulations, although not significantly so.

Without the background subtraction, the ANOVA variants no longer outperform other methods, and in fact, have results similar to those of the *t*-test variant methods (in contrast to ANOVA's superior performance on the background normalized data). Also of note is that lowess normalization is detrimental to most methods when the background is not corrected. Only on the *t* test, SAM, and F1 did lowess on non-background-normalized data improve results compared to median centering.

Finally, we attempted to combine some of the highest performing methods by averaging the outputs. For methods in which the output is not a *p* value, the output was rescaled to the range (0, 1). We averaged the ANOVA variants together (F); the variants with GSVD (F_G), the *B* statistic (F_B), and EB WRS (F_EB); and GSVD with the *B* statistic (GSVD_B) and EB WRS (GSVD_EB). Results are shown in Tables 4 and 5 for background-subtracted and non-background-subtracted data, respectively. The combinations did not provide a statistically significant increase in performance for any particular condition or network type. However, in the case of the F_G on background-subtracted data and GSVD_EB on non-background-subtracted data, the average statistic did demonstrate a good overall performance regardless of network type and experimental condition.

EB WRS and to a lesser extent GSVD were more robust with regard to background subtraction compared to other methods. For both EB WRS and GSVD, uncorrected data gave the best

Table 3
Average AUC for all simulations per network type and experimental condition, by comparison method on data *without* background subtraction

Comparison method (normalization)	Network type											
	Erdös and Rényi				Scale-free				Small-world			
	pt–lo	pt–hi	mt–lo	mt–hi	pt–lo	pt–hi	mt–lo	mt–hi	pt–lo	pt–hi	mt–lo	mt–hi
EBayes WRS (none)	0.49	0.50	0.79	0.56	0.59	0.61	**0.90**	0.68	0.57	0.59	**0.79**	**0.76**
t test (lowess)	0.61	0.53	0.84	0.61	0.64	0.53	0.86	0.64	**0.62**	0.57	0.78	0.73
SAM (lowess)	**0.62**	0.53	**0.85**	0.60	0.60	0.48	0.84	0.58	0.60	0.55	0.76	0.70
MLE-T (med ctr)	0.50	0.47	0.80	0.56	0.52	0.55	0.75	0.62	0.56	0.53	0.73	0.66
Cyber-T (med ctr)	0.52	0.49	0.80	0.56	0.56	0.53	0.76	0.62	0.59	0.52	0.73	0.66
B (med ctr)	0.62	0.53	0.85	0.61	0.56	0.43	0.83	0.52	0.61	0.55	0.77	0.71
ANOVA–F1 (lowess)	0.55	0.53	0.80	0.61	0.57	0.49	0.85	0.58	0.53	0.53	0.73	0.67
ANOVA–F2 (med ctr)	0.57	0.55	0.79	0.60	0.57	0.62	0.84	0.67	0.52	0.54	0.74	0.66
ANOVA–F3 (med ctr)	0.59	**0.59**	0.79	**0.62**	0.59	0.64	0.85	0.67	0.52	0.53	0.73	0.65
GSVD (none)	0.59	0.57	0.55	0.56	**0.76**	**0.75**	0.65	**0.76**	0.59	**0.63**	0.57	0.61

The highest result per network type and condition is shown in bold.

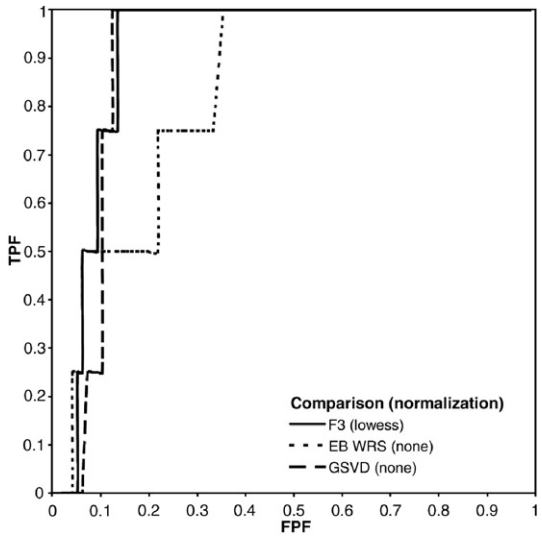

Fig. 3. Representative ROC curves for three example comparison methods. Results are shown for scale-free network Century SF-032, low initial concentration, perturbation (pt–lo) simulation 10.

results although median centering gave results that were slightly lower if not similar. The *t* test and variants and ANOVA and variants all had decreased performance when median centering was used compared to lowess on the background-subtracted data. The reverse was true for the *t* test variants (except the *t* test itself and SAM) and ANOVA and variants on the non-background-subtracted data, for which median centering gave the best performance. The recommended normalization for each comparison method is summarized in Supplemental Table S3.

Analysis of Boldrick et al. innate immune response data [28]

For the analysis of the Boldrick et al. data we considered only EB WRS, F, and GSVD since, based on the simulations, these three methods seemed to perform the best depending on experimental conditions. Fig. 5 shows raster displays of subsets of the Boldrick et al. *E. coli* and untreated data. The genes are sorted by the *p* values from each comparison method. The left three raster displays are of the 1000 *least* extreme genes based on the *p* value from the EB WRS, F, and GSVD comparison methods, and the right three are of the 1000 *most* extreme genes for EB WRS, F, and GSVD.

The extremity of each method's *p* value is the experimenter's indication of the level of the gene production or activity in the samples. The actual gene activity is represented in the measurements themselves, bright red or green. From the three displays of the least extreme genes in Fig. 5 (left) it appears that the least extreme genes determined by GSVD actually demonstrate less activity compared to the least extreme genes from the EB WRS and F methods, which both have some low *p* values corresponding to spikes of activity in the sample (some moderately green or red time points). It is from the most extreme genes as measured by each method (Fig. 5, right) that we can infer the types of actual activity that each method will deem as changed. EB WRS apparently measures the most activity from genes with changes spanning most if not all experiments (nearly solid bands of red or green). The F comparison method is similar but also measures cycling between overexpression and underexpression (green to red bands of multiple consecutive experiments) as extreme activity. And finally, GSVD measures the most activity from genes with multiple spikes that do not necessarily span multiple experiments (multiple, individual green or red time points).

For further evaluation of the comparison methods, we identified targets for the TFs FOS and JUN (AP-1) and NFKB1 and RELA (NF-κB) using the literature and compiled resources TRANSPATH (http://www.biobase.de), the Kyoto Encyclopedia of Genes and Genomes (KEGG) [31], and the Ingenuity Pathway Analysis knowledge base (http://www.ingenuity.com). The identified targets in each module are available in Supplemental Table S4. Then, similar to the manner in which Tavazoie et al. used combinatorics to measure category enrichment [41], we then use the Fisher exact test to measure the enrichment of a TF module. Please refer to the supplementary methods for more details. The Fisher's exact test *p* value can then be interpreted as an indication of the activity of the each TF and can be compared to the expected activity. FOS, JUN, NFKB1, and RELA are all expected to be active in response to *E. coli*. Raster displays of the genes in each module are shown in Fig. 6. The results of the measurement of module enrichment using each different comparison are shown in Table 6.

Using an $\alpha < 0.05$ cutoff on the Fisher's exact test *p* value to indicate activity, the FOS, JUN, and NFKB1 modules were shown to be active using all three comparison methods. The RELA module was not found active by GSVD, while EB WRS and F did lead to an indication of activity. From the raster displays in Fig. 6 it appears that there is significant activity in all modules. The RELA a module contains a number of genes with

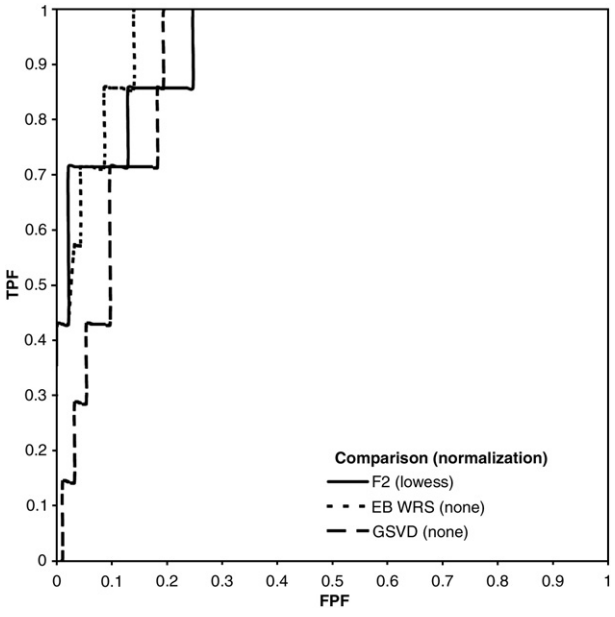

Fig. 4. Representative ROC curves for three example comparison methods. Results are shown for scale-free network Century SF-020, low initial concentration, mutation (mt–lo) simulation 6.

Table 4
Average AUC for all simulations per network type and experimental condition, using combined comparison methods on data *with* background subtraction

Comparison method (normalization)	Network type											
	Erdös and Rényi				Scale-free				Small-world			
	pt–lo	pt–hi	mt–lo	mt–hi	pt–lo	pt–hi	mt–lo	mt–hi	pt–lo	pt–hi	mt–lo	mt–hi
F (lowess)	0.69	0.55	0.89	**0.68**	0.83	0.63	0.96	0.79	**0.76**	0.66	0.91	0.85
F_B (lowess)	0.70	0.55	**0.90**	**0.69**	0.83	0.63	0.96	0.79	**0.77**	**0.67**	**0.91**	**0.87**
F_EB (lowess/none)	0.65	0.52	0.87	**0.68**	0.78	0.63	0.95	**0.82**	0.74	0.66	0.90	**0.86**
GSVD_F (none/lowess)	0.67	**0.57**	0.79	0.63	**0.88**	0.77	0.92	**0.85**	**0.77**	**0.68**	0.84	0.83
GSVD_B (none)	0.62	**0.58**	0.83	0.65	**0.85**	0.78	0.95	**0.87**	0.74	**0.67**	0.86	**0.86**
GSVD_EB (none)	0.59	0.56	0.76	0.65	0.79	0.74	0.91	**0.86**	0.72	0.66	0.84	0.85

Results higher than those for the highest individual methods (Table 2) are shown in bold.

significant, constant overexpression over the time period, precisely the type of change that GSVD is not expected to detect well.

Discussion

Several methods were found to detect differentially expressed genes adequately in the mutant simulations in which the absence or overexpression of a gene was persistent. This was expected since a mutation alters the entire temporal expression pattern of the mutated gene and all the genes that it influences. The comparison of the mean of those expression patterns is more likely to differ from the mean of the corresponding normal expression patterns. Thus, identifying differentially expressed genes in the time-series mutant persistent case is similar to that in a static experiment.

While the current methods were found to detect changing genes adequately in the mutation simulations, there was a difference between the two mutation subtypes. All methods for identifying differentially expressed genes performed better in the knockout simulations (mt–lo) than in the persistent overexpression simulations (mt–hi). This is likely an artifact of the gene network simulations. Mendes et al. model the gene relationships in the artificial networks using kinetic behavior similar to Hill kinetics [27]. The kinetic curves of target genes have a sigmoidal shape and exhibit saturation at large parent-gene concentrations. Hence, increasing a parent gene's concentration in the simulations, which is the case in the persistent overexpression simulations, may not have as marked effect on the gene's children. As the networks are modeling transcriptional kinetics, it may be that too much of a transcriptional regulatory gene in a real biological system is not as easily discriminated compared to the absence of that gene.

The comparison methods studied were less able to identify differentially expressed genes in the perturbation simulations. An outside stress is imposed on the models in the perturbation simulations, but since the relationships between the genes are not altered, the system is able to correct for the stress in a manner that affects only a portion of the temporal expression pattern. Therefore, identifying differentially expressed genes in the time-series perturbation experiments differs from that in static experiments. It is not surprising that some methods developed for the static comparison performed poorly when applied to perturbation time-series data. It is also likely that these transient changes are not always biologically significant.

ANOVA, particularly F2 or F3, gave the best results on background-subtracted data for most conditions and network topologies. ANOVA is similar to the other statistical methods in that it finds differences based on variance in mean expression among groups. Its increased power compared to the other methods is attributed to the additional term in the linear model that accounts for the time variable. This advantage was reduced on the data without background subtraction. Although ANOVA can include a time variable, as presented here, ANOVA does not include the temporal ordering. A variant of ANOVA, which extends the ANOVA variants (F1, F2, and F3) to incorporate the chronological element of the experiments, was proposed by Park et al. [42]. It is likely that this additional information can further improve the ANOVA results.

Table 5
Average AUC for all simulations per network type and experimental condition, using combined comparison methods on data *without* background subtraction

Comparison method (normalization)	Network type											
	Erdös and Rényi				Scale-free				Small-world			
	pt–lo	pt–hi	mt–lo	mt–hi	pt–lo	pt–hi	mt–lo	mt–hi	pt–lo	pt–hi	mt–lo	mt–hi
F (lowess)	0.58	0.58	0.79	0.61	0.58	0.63	0.84	0.67	0.52	0.54	0.74	0.66
F_B (lowess)	0.62	0.56	0.83	0.61	0.59	0.57	0.86	0.62	0.56	0.53	0.76	0.68
F_EB (lowess/none)	0.54	0.54	0.79	0.59	0.62	0.65	0.89	0.70	0.57	0.60	0.78	0.75
GSVD_F (none/lowess)	0.60	**0.60**	0.74	**0.63**	0.73	0.73	0.81	**0.78**	0.61	0.63	0.74	0.71
GSVD_B (none)	**0.64**	**0.62**	0.80	**0.66**	0.67	0.64	0.81	0.65	0.55	0.51	0.68	0.66
GSVD_EB (none)	0.54	0.54	0.74	0.58	0.73	0.73	0.84	**0.78**	0.61	0.62	0.77	0.75

Results higher than those for the highest individual methods (Table 3) are shown in bold.

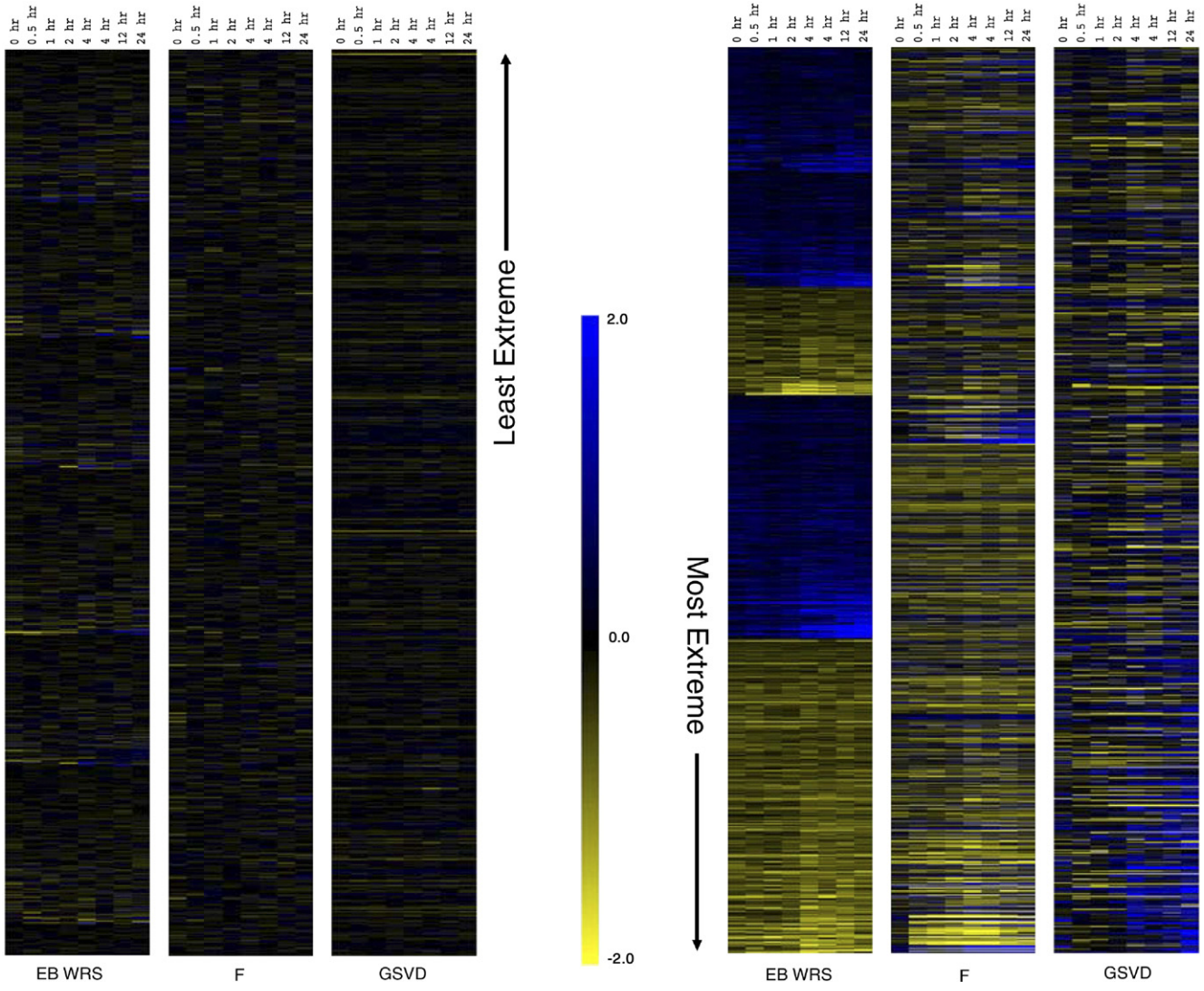

Fig. 5. Raster displays of the $\log_2(R/G)$ values from the Boldrick et al. immune response experiments [28]. The red, channel 2 data are from the measurements of the PBMC response to *E. coli* and the green, channel 1 data are the measurement of untreated PBMCs. The data are sorted by *p*-value output of the EB WRS, F, and GSVD comparison methods. The left three raster displays show the 1000 least extreme genes, as measured by the comparison methods. The right three show the 1000 most extreme genes.

The GSVD method appears to provide some improvement over most methods for the detection of more subtle changes in a gene's expression (when a stress is imposed on a normal biological system). This is likely because it is a pattern-based comparison, rather than a comparison of mean expression differences. On the other hand, GSVD performed poorly in the knockout experiments for the same reason. The persistent change may only offset the expression rather than change the pattern of expression. Additionally, as Peddada et al. note, a strong correlation does not necessarily correspond to matching patterns, and likewise a low correlation does not always result from differing patterns [18]. Therefore, an alternative metric for comparing the GSVD projections may improve the GSVD differential expression method.

From the analysis of the Boldrick et al. [28] data (Fig. 5) we can better understand the types of changes each method detects most effectively. The changes that the EB WRS method detects appear to be those a classical statistical test would easily detect, i.e., overall changes in mean between groups. ANOVA appears to also detect those changes and others in which the gene activity is fluctuating but not necessarily over the entire time frame. GSVD, relative to EB WRS and ANOVA, poorly detects the overall changes in mean, but seems to best detect those changes that are spikes of sudden gene activity in one or a few time points.

The results were more reliable on the background-subtracted data; therefore the background should be removed whenever possible. In most instances, however, the background cannot be properly removed. This should factor in the choice of analysis method. Based on this study, and the fact that true biological networks have both scale-free and small-world characteristics, we recommend the ANOVA variant combination for analyzing background-subtracted data. Although the performances of F2 and F3 were similar, we recommend averaging ANOVA results. It is likely that the noise we superimposed on the simulated data is more uniform than an actual array. The F3 method estimates variance from the entire array; hence, the uniformity of our

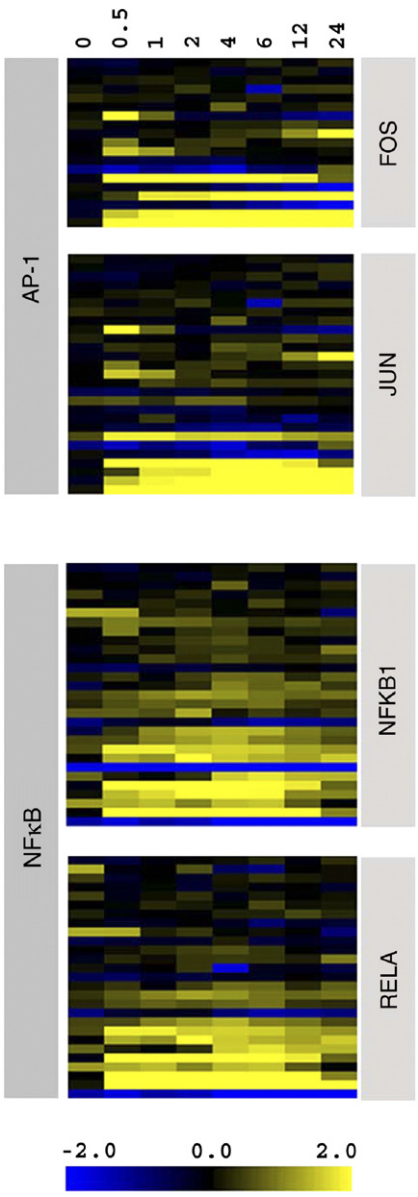

Fig. 6. Raster displays of the $\log_2(R/G)$ values from the Boldrick et al. immune response experiments [28] for genes in the FOS, JUN, NFKB1, and RELA TF modules. The red, channel 2 data are from the measurements of the PBMC response to *E. coli* and the green, channel 1 data are from the measurements of untreated PBMCs.

simulations may have favorably biased the results of the F3 method.

When the background cannot be removed, EB WRS is recommended. But EB WRS will likely detect only changes in gene expression that are present over the entire time scale. GSVD can instead be used for background-subtracted or non-background-subtracted data. But GSVD is most suitable in the special cases in which only short-term or subtle changes are expected, for example, when comparing normal responses of different genotypes. Combining either EB WRS or ANOVA with GSVD would probably provide the best results on time-series data since changes of multiple types could potentially be detected. In the simulations, a combination of the ANOVA variants with GSVD achieved the most consistent performance on background subtracted data, while a combination of EB WRS with GSVD was most consistent when the background was not removed. Furthermore, EB WRS and GSVD were most robust to non-background-subtraction and worked best without any normalization. While this could again be a consequence of the uniformity of the data, median centering only slightly affected the results of EB WRS and GSVD; therefore, complex normalization can be avoided. In general, we recommend EB WRS, or the EB WRS–GSVD combination, when the background cannot be removed.

In addition to determining which genes are differentially expressed in a time-series microarray experiment, experimenters may be interested in the time point at which the gene becomes changed. The methods considered in this study are not designed for that purpose. It is possible to use the recommended methods to isolate the changed genes and then evaluate the expression patterns of only those genes for the specific point(s) of change. This evaluation can be performed using methods for single arrays, such as sliding window z scores or a modified one-sample t test (preferably a test with stable variance estimates, such as a penalized t test) using the reference sample(s) as the null and each experiment measurement as the test point. But the analysis should be completed first on all arrays, rather than individually, so as not to lose the temporal context of the experiments.

Materials and methods

Gene network simulations

We simulated microarray data using the artificial gene networks described by Mendes et al. [27]. We compared networks from all three available topology categories: Erdös and Rényi [33], "small-world" [34], and "scale-free" [35]. Most networks within the cell have been shown to approximate a scale-free topology and have features of the so-called small-world effect [36], although the exact topology has not been firmly established. Additionally, it is not known whether analysis of the corresponding microarray data depends on the type of topology. The GEPASI 3 Biochemical Simulator software was utilized to simulate all data [37].

Simulation conditions

We selected five networks from each topology and performed 10 simulations for each under the following experimental conditions:

- Normal: Genes in each network were initialized to concentration values near steady-state levels.
- Perturbation—high: A single gene was selected from each network and initialized to a concentration considerably *higher* than the steady-state level.
- Perturbation—low: A single gene was selected from each network and initialized to a concentration considerably *lower* than the steady-state level.

Table 6
Results of TF module activation using the Fischer exact test

	EB WRS	F	GSVD
FOS	0.031	0.027	>0.001
JUN	0.010	0.007	>0.001
NFKB1	>0.001	>0.001	>0.001
RELA	0.001	>0.001	0.156

Activity of the TF modules FOS, JUN, NFKB1, and RELA as measured by the Fischer's exact test using the three microarray comparison methods EB WRS, F, and GSVD as the activity level of individual genes in the TF modules.

- Mutation—high: A single gene was selected from each network and initialized to a concentration considerably *higher* than the steady-state level. The kinetics of the selected gene were altered such that it was no longer under regulatory control.
- Mutation—low: A single gene was selected from each network and initialized to a concentration considerably *lower* than the steady-state level. The kinetics of the selected gene were altered such that it was no longer under regulatory control.

Only one gene per network was explicitly altered. All other genes were initialized to values near their normal steady-state levels. The normal simulations correspond to the reference experiments and are handled as the Cy3/channel 1 data. All the other simulations are handled as the Cy5/channel 2 data. Perturbations represent conditions under which a gene's concentration is changed from the normal state, but the system is capable of correcting for the change. The situation is analogous to a temporary stress purposefully imposed on a normal cell by the experimenters. The mutations, on the other hand, are most analogous to either knockout studies (low) or the unregulated over-expression of a gene (high), such as in a cancerous cell.

Each simulation was run from time 0 until steady state was reached. For the chosen networks, this corresponded to a time of approximately 40 min. The time to reach steady state is representative of some biological processes, while other biological systems can take hours or more to respond. For our purposes, however, the absolute times are irrelevant; only relative changes on a time scale are necessary. Concentration values were sampled every 0.5 min, but the comparisons were evaluated with samples from times 0, 0.5, 1, 2, 3, 4, 5, 10, 20, and 40 min to give the relative distribution of "arrays" likely in a time-course microarray experiment (assuming noncyclic conditions).

Noise

The GEPASI simulator provides exact gene concentrations, as defined by the model parameters, at specified time points. To evaluate the class comparison methods in a realistic manner, characteristic microarray noise was superimposed onto the data. The noise implemented follows the Rocke and Durbin two-component noise model [38].

Using the software available from the model's authors, we obtained estimates for the parameters for both the Cy3 and the Cy5 intensity (to introduce a dye bias) from the classic Spellman et al. yeast cell cycle α-factor microarray dataset [26]. The parameter estimates from this dataset were then used to generate similar variants for each raw "array" output from the GEPASI simulations.

Differential expression

The following methods were compared on the simulated arrays: EB WRS [4], *t* test, the Bayesian or regularized *t* test (Cyber-T) [2], the SAM [8] test statistic (without the permutation analysis), the *t* statistic with maximum-likelihood standard error estimate (MLE-T) [7], the *B* statistic [5], the "three flavors" of *F* test (F1, F2, and F3) [15], and our implementation of GSVD. Cui and Churchill provide a detailed discussion of several of these methods [15]. We discuss GSVD below and the other methods in the supplementary material.

GSVD for class comparison

For a variety of reasons, the methods that have been proposed specifically for identifying differentially expressed genes in time-course data are not generally applicable [17]. Instead, methods developed for static microarray comparisons are also used for time-course data because they typically require calculation of only a mean and variance. However, this strategy may not be best suited to all the types of changes that could be expected in a time-course experiment. Fig. 7 illustrates two such changes, a phase shift (Fig. 7A) and inverse expression patterns (Fig. 7B), although there are certainly others. For this reason we aimed to find a method that did not make comparisons using the mean and variance of the gene expression values but rather compared the time-course expression patterns.

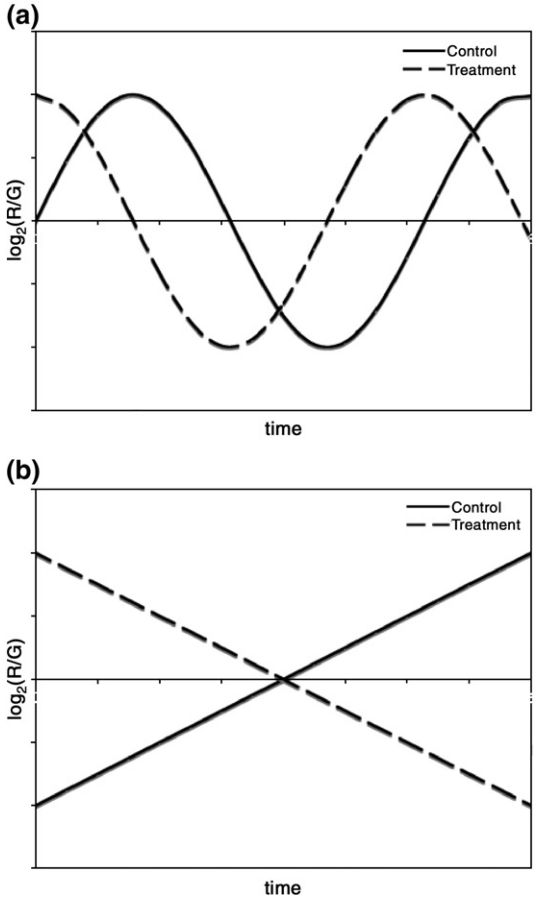

Fig. 7. Examples of control and treatment microarray measurements that would not be uncovered by traditional statistical methods used for determining differential expression. Examples shown are a phase shift (a) and inverse expression patterns (b).

GSVD is the transform of two microarray datasets, e_1 and e_2, with m genes and n experiments each, into "eigengenes" × "eigenarrays" space,

$$e_1 = u_1 \varepsilon_1 v^T, \tag{1}$$

$$e_2 = u_2 \varepsilon_2 v^T, \tag{2}$$

where each dataset e_i is represented by a common set of eigengenes, v^T. Refer to Alter et al. [20] and Golub and Loan [39] for algorithmic details.

Alter et al. used the common set of eigengenes, or expression patterns, to classify genes in the two decomposed datasets. The authors determined that the individual patterns represented regulatory programs or independent cellular processes. This allowed for genes to be grouped into various processes based on which pattern the gene most closely matched [20]. Our application of GSVD allows for the quantification of the similarity or difference in a gene-by-gene manner by comparing the control and treatment microarray measurements (for a cDNA array, channel 1 reference is compared to channel 2 treatment).

If there has been no change in the expression of a gene between the two sets of experimental treatments, the common set of expression patterns should be equally representative of that gene in one dataset, e_1, compared to the other dataset, e_2. To evaluate the deviation from this assumption the difference between the projections of each dataset, e_1 and e_2, onto the common pattern set of eigengenes ($e_1 \cdot v^T$) is quantified. In practice, a given pattern will be overrepresented in one dataset versus

the other. Therefore, to compare the projections we use the weighted Pearson's correlation, r_{xy},

$$r_{xy}(i) = \frac{\sum_{j=1}^{n} w_j (x_{ij} - \bar{x}_w(i))(y_{ij} - \bar{y}_w(i))}{\sqrt{\sum_{j=1}^{n} w_j (x_{ij} - \bar{x}_w(i))^2 \sum_{j=1}^{n} w_j (y_{ij} - \bar{y}_w(i))^2}} \quad (3)$$

where

$$\bar{x}_w(i) = \frac{\sum_{j=1}^{n} w_j x_{ij}}{\sum_{j=1}^{n} w_j} \text{ and } \bar{y}_w(i) = \frac{\sum_{j=1}^{n} w_j y_{ij}}{\sum_{j=1}^{n} w_j}. \quad (4)$$

The weights, w_j, are just $\pi/4$ added to the normalized, absolute value of the antisymmetric angular distances between the datasets as shown in Eq. (5):

$$\bar{x}_w(i) = \frac{\sum_{j=1}^{n} w_j x_{ij}}{\sum_{j=1}^{n} w_j} \text{ and } \bar{y}_w(i) = \frac{\sum_{j=1}^{n} w_j y_{ij}}{\sum_{j=1}^{n} w_j} \quad (5)$$

Each GSVD comparison carried out in this fashion results in a series of weighted Pearson correlations, one metric for each gene pair, i, of the m gene pairs. The metrics are then bootstrapped to generate a null distribution. Namely, the m pairs of genes are sampled 1000 times with replacement, maintaining the pairing. The result is 1000 datasets of pseudo-GSVD correlations, r^*_{xy}, that taken together should approach the underlying distribution of the data. The significance of the ith correlation from the actual array data is quantified as the proportion of bootstrap correlations that are lower than the ith correlation (Eq. (6)):

$$P\{r_{xy}(i)\} = \frac{\#\{r^*_{xy} < r_{xy}(i)\}}{1000^* m}, \quad (6)$$

Normalization

For cDNA microarray data, the Cyber-T and SAM documentation recommends median centering the $\log_2$ ratios. Lönnstedt and Speed [5] also recommend this type of global normalization for the B statistic, in addition to lowess correction (in this case performed using the Bioconductor package Limma [40]). For EB WRS Efron and Tibshirani recommend standardized median centered ratios [4].

For ANOVA, Cui and Churchill [15] recommend normalization based on the curvature. Since the noise we introduced was consistent for all "arrays" we simply applied the MAANOVA lowess correction. Of the data input schemes considered for GSVD, we recommend using single-channel nonnormalized data with GSVD, rather than the log transformed data, as it gave better relative results.

Data normalization across arrays can considerably influence the results of the comparisons. For this reason, and due to the nature of the simulations and noise, we normalized the data in the following ways: median centering or lowess correction with multiple smoothing parameters.

Given that the superimposed noise did not incorporate any per-spot or print-tip anomalies, the background in the simulations was estimated as the minimum measurement for a particular array. This approach is more of a baseline normalization rather than background subtraction. However, since it is increasingly common to forgo background subtraction, particularly when background measurements are unreliable, this approach is useful as it allows us to evaluate the effects of background noise on the analysis.

Evaluation

We evaluated all comparison methods by means of the area under the receiver operating characteristic curve. The AUC for each curve was calculated using the trapezoid area approximation (equivalent to the Wilcoxon–Mann–Whitney statistic).

To generate the curves, all comparison results were compared to the "truth" for that network. The networks are represented as directed graphs, with genes as nodes and arrows as regulatory relationships. An arrow from one gene, or node, to another indicates a regulatory influence of the prior on the latter. We defined the truth as those genes influenced directly by the gene we explicitly altered in that network, including the altered gene itself. Thus, those genes with arrows directly from the altered gene were assumed to have changed, as illustrated in Fig. 8. In other words, the altered gene and its children comprised the "truth."

The AUC for the 10 simulations from each of the five networks from each topology category were averaged to come up with a metric to assess the different differential expression methods. Significance of this metric was evaluated using a two-tailed paired t test at the 0.05 level on the AUC results for each simulation of a given network type/experimental condition between the two methods to be compared, in other words, the highest average AUC for a given condition compared to the next highest.

Boldrick et al. innate immune response data [28]

We chose the Boldrick et al. [28] data because the study was designed to measure transcriptional activity associated with the innate immune response. Innate immunity is perhaps the best understood portion of the immune response, and the corresponding TFs that become activated have been studied in considerable depth. A subset of the Boldrick et al. arrays was selected from the "Diversity" block of experiments. Specifically, we used the untreated and *E. coli* time-course array sets. In each set, samples were taken at 0, 0.5, 2, 4, 6, 12, and 24 h.

In all of the arrays, channel 1 was a universal control. We used the universal control to normalize all arrays. Moderate intensity-dependent curvature was evident, an example of which is shown in the supplement. We corrected the curvature with lowess using a smoothing parameter of 0.25, which, through visual observation, provided optimal removal of curvature. The other smoothing parameters used, 0.2, 0.3, and 0.4, did not affect the conclusions of this study. After normalization, replicate gene measurements were averaged together.

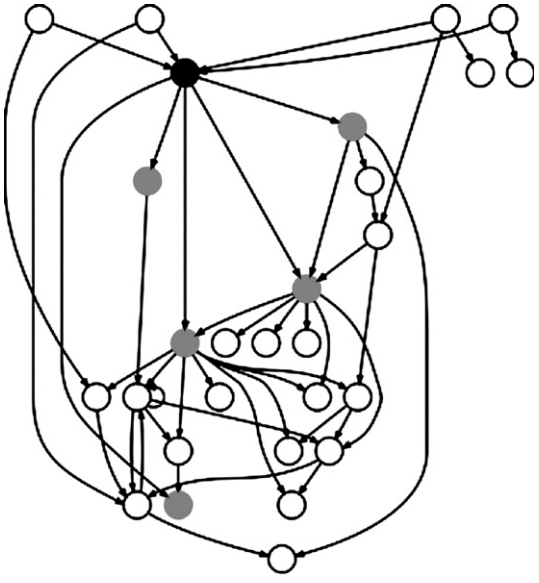

Fig. 8. Example portion of a "scale-free" artificial gene network: CenturySF-046. The explicitly changed gene is highlighted in black, and genes expected to change as a result are highlighted in gray.

Acknowledgments

We thank Zack Mahdavi and Chris Kite for technical support and Dan Bozinov for helpful discussions on microarray normalization.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.10.008.

References

[1] X. Cui, M.K. Kerr, G.A. Churchill, Transformations for cDNA microarray data, Stat. Appl. Genet. Mol. Biol. 2 (2003).
[2] P. Baldi, A.D. Long, A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes, Bioinformatics 17 (2001) 509–519.
[3] M.K. Kerr, M. Martin, G.A. Churchill, Analysis of variance for gene expression microarray data, J. Comput. Biol. 7 (2000) 819–837.
[4] B. Efron, R. Tibshirani, Empirical Bayes methods and false discovery rates for microarrays, Genet. Epidemiol. 23 (2002) 70–86.
[5] I. Lonnstedt, T. Speed, Replicated microarray data, Stat. Sin. 12 (2002) 31–46.
[6] M. Newton, C. Kendziorski, C. Richmond, F. Blattner, K. Tsui, On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data, J. Comput. Biol. 8 (2001) 37–52.
[7] J.G. Thomas, J.M. Olson, S.J. Tapscott, L.P. Zhao, An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles, Genome Res. 11 (2001) 1227–1236.
[8] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. USA 98 (2001) 5116–5121.
[9] R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, R.S. Paules, Assessing gene significance from cDNA microarray expression data via mixed models, J. Comput. Biol. 8 (2001) 625–637.
[10] R. Hoffmann, T. Seidl, M. Dugas, Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis, Genome Biol. 3 (2002) research0033.1–0033.11.
[11] L.-X. Qin, K.F. Kerr, Contributing Members of the Toxicogenomics Research Consortium, Empirical evaluation of data transformations and ranking statistics for microarray analysis, Nucleic Acids Res. 32 (2004) 5471–5479.
[12] T. Aittokallio, M. Kurk, O. Nevalainen, T. Nikula, A. West, R. Lahesmaa, Computational strategies for analyzing data in gene expression microarray experiments, J. Bioinform. Comput. Biol. 1 (2003) 541–586.
[13] P. Broberg, Statistical methods for ranking differentially expressed genes, Genome Biol. 4 (2003).
[14] W. Pan, A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments, Bioinformatics 18 (2002) 546–554.
[15] X. Cui, G.A. Churchil, Statistical tests for differential expression in cDNA microarray experiments, Genome Biol. 4 (2003) 210–220.
[16] Z. Bar-Joseph, Analyzing time series gene expression data, Bioinformatics 20 (2004) 2493–2503.
[17] Z. Bar-Joseph, G. Gerber, I. Simon, D.K. Gifford, T.S. Jaakkola, Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes, Proc. Natl. Acad. Sci. USA 100 (2003) 10146–10151.
[18] S.D. Peddada, E.K. Lobenhofer, L. Li, C.A. Afshari, C.R. Weinberg, D.M. Umbach, Gene selection and clustering for time-course and dose–response microarray experiments using order-restricted inference, Bioinformatics 19 (2003) 834–841.
[19] G. Zhu, P.T. Spellman, T. Volpe, P.O. Brown, D. Botstein, T.N. Davis, B. Futcher, Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth, Nature 406 (2000) 90–94.
[20] O. Alter, P.O. Brown, D. Botstein, Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, Proc. Natl. Acad. Sci. USA 100 (2003) 3351–3356.
[21] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Natl. Acad. Sci. USA 97 (2000) 10101–10106.
[22] T.D. Moloshok, R.R. Klevecz, J.D. Grant, F.J. Manion, W.F. Speier IV, M.F. Ochs, Application of Bayesian decomposition for analysing microarray data, Bioinformatics 18 (2002) 566–575.
[23] X.L. Xu, J.M. Olson, L.P. Zhao, A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model, Hum. Mol. Genet. 11 (2002) 1977–1985.
[24] H. Liu, S. Tarima, A.S. Borders, T.V. Getchell, M.L. Getchell, A.J. Stromberg, Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments, BMC Bioinformatics 6 (2005).
[25] Y. Luan, H. Li, Clustering of time-course gene expression data using a mixed-effects model with B-splines, Bioinformatics 19 (2003) 474–482.
[26] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Mol. Biol. Cell 9 (1998) 3273–3297.
[27] P. Mendes, W. Sha, K. Ye, Artificial gene networks for objective comparison of analysis algorithms, Bioinformatics 19 (2003) 122ii–129ii.
[28] J.C. Boldrick, A.A. Alizadeh, M. Diehn, S. Dudoit, C.L. Liu, C.E. Belcher, D. Botstein, L.M. Staudt, P.O. Brown, D.A. Relman, Stereotyped and specific gene expression programs in human innate immune responses to bacteria, Proc. Natl. Acad. Sci. USA 99 (2002) 972–977.
[29] E. Lien, T.K. Means, H. Heine, A. Yoshimura, S. Kusumoto, K. Fukase, M.J. Fenton, M. Oikawa, N. Qureshi, B. Monks, R.W. Finberg, R.R. Ingalls, D.T. Golenbock, Toll-like receptor 4 imparts ligand-specific recognition of bacterial lipopolysaccharide, J. Clin. Invest. 105 (2000) 497–504.
[30] K. Takeda, S. Akira, Toll-like receptors in innate immunity, Int. Immunol. 17 (2005) 1–14.
[31] M. Kanehisa, The KEGG database, Novartis Found. Symp. 247 (2002) 91–101.
[32] T. Kawai, S. Akira, Toll-like receptor downstream signaling, Arthritis Res. Ther. 7 (2004) 12–19.
[33] P. Erdos, A. Renyi, On random graphs, Publ. Math. Debrecen. 6 (1959) 290–297.
[34] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440–442.
[35] A.L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[36] A.L. Barabasi, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nat. Rev. Genet. 5 (2004) 101–113.
[37] P. Mendes, GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems, Comput. Appl. Biosci. 9 (1993) 563–571.
[38] D.M. Rocke, B. Durbin, A model for measurement error for gene expression arrays, J. Comput. Biol. 8 (2001) 557–569.
[39] G.H. Golub, C.F. Van Loan, Matrix Computation, Johns Hopkins Univ. Press, Baltimore, 1996.
[40] G.K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004).
[41] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, Systematic determination of genetic network architecture, Nat. Genet. 22 (1999) 281–285.
[42] T. Park, S.-G. Yi, S. Lee, S.Y. Lee, D.-H. Yoo, J.-I. Ahn, Y.-S. Lee, Statistical tests for identifying differentially expressed genes in time-course microarray experiments, Bioinformatics 19 (2003) 694–703.