Inference of Transcriptional Regulation Relationships from Gene Expression Data

Andrew T. Kwon
Dept. of Computer Science
University of British Columbia
Vancouver, BC, Canada
tjkwon@cs.ubc.ca

Holger H. Hoos*
Dept. of Computer Science
University of British Columbia
Vancouver, BC, Canada
hoos@cs.ubc.ca

Raymond Ng
Dept. of Computer Science
University of British Columbia
Vancouver, BC, Canada
rng@cs.ubc.ca

ABSTRACT

We propose a new method for finding potential regulatory relationships between pairs of genes from microarray time series data and apply it to expression data for cell-cycle related genes in yeast. We compare our algorithm, dubbed the event method, with the earlier correlation method and the edge detection method by Filkov *et al.* When tested on known transcriptional regulation genes, all three methods are able to find similar numbers of true positives. The results indicate that our algorithm is able to identify true positive pairs that are different from those found by the two other methods. We also compare the correlation and the event methods using synthetic data and find that typically, the event method obtains better results.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences

1. INTRODUCTION

A genetic regulatory network is a system in which proteins and genes bind to each other and act as complex input-output system for controlling cellular functions. For a normal cell life cycle to take place, a cell needs to have in place a correctly working regulatory network for control. Many of the known regulators that control mRNA levels work at the level of transcription (other control mechanisms, not considered here, are based on post-transcriptional modifications). Many of these regulators are components of protein complexes that regulate the transcription of other genes. Insights into the nature and function of various pathways in the network are of interest to many researchers, as these are the key to a better understanding of many important biological problems.

In order to study the regulatory network, it is necessary to have a means to measure the gene expression at different time points, so that one can observe and infer which genes are being regulated by looking at their expression levels. Until the development of cDNA microarrays researchers could perform experiments only on a limited number of genes at a time, even though these genes are part of a large network. Microarray technology allows researchers to study gene expression on a large scale; but it also poses new challenges, as one must now find ways to sort through and extract useful information from the massive amounts of data.

Before one can determine the overall regulatory network structure, it is important to identify genes that have direct regulatory relationships. Due to the complex nature of the network, even this is not an easy task. In fact, there are many different variables associated with protein expression besides the mRNA levels, which means that cDNA microarray data alone does not present the researchers with a complete picture. However, although it may be incomplete, this data still contains a significant amount of information pertaining to the cellular protein levels, and can thus provide researchers with useful and interesting information that can help them focus their research efforts.

The problem addressed in this paper is that of determinating which pairs of genes have direct regulatory interactions given a large number of gene expression profiles obtained from microarray data. We propose a new method called the *Event Method* for finding potential regulatory pairs from gene expression data and evaluate it against previous methods using real and synthetic data sets.

The remainder of this paper is organized as follows. In Section 2, we give a brief overview of the existing algorithms for solving the problem of finding gene regulation pairs. In Section 3, we present a detailed explanation of our algorithm. In Section 4, we discuss the experimental results of our algorithm using real and synthetic data sets. Finally, in the concluding Section 5 we provide a brief summary of our results and indicate some directions for future work.

2. PRIOR WORK

There are a number of previous approaches for extracting regulation information from microarray data. These include methods ranging from simple correlation analysis and clustering to the application of Bayesian networks.

The first is the *correlation method*, which tests whether two variables share a significant linear relationship with each other by their Pearson correlation coefficient. The correlation method has been used heavily as the basis of many clus-

*To whom correspondence should be addressed

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SAC 2003, Melbourne, Florida, USA © 2003 ACM 1-58113-624-2/03/03 ...$5.00.

tering analyses, as it is very useful in determining whether two variables have a strong global similarity. With microarray time series data, one would expect those genes with regulatory relationships to exhibit globally similar gene expression profiles, which could be tested using the correlation method. However, while this method is good at determining genes that share global similarity, it does not take into account the fact that it often takes time for the regulator gene product to exert its influence on its target gene. Moreover, the correlation method strongly favours global similarity over more localized similarities arising from conditional regulatory relationships.

The second method is the *edge detection method* by Filkov *et al.* [3], who focused on improving the local edge detection ability compared to other methods, such as correlation. The edge detection method scans through each gene expression curve to determine where major changes in expression level (edges) occur, and removes spurious edges from consideration. To produce a score, the edge detection method sums up the number of edges in two gene expression curves that share the same direction and are within reasonable distances of each other. Gene pairs that are likely to have an activation relationship are given high scores. In calculating the score, the method also makes sure that those edges that are farther apart would get lower scores. The edge detection method in its current form can only determine potential activation relationships.

The third approach is the usage of *Bayesian networks* [9, 4, 1]. A Bayesian network is a graphical representation of conditional independence in a multivariate probability distribution. For gene regulatory network inference, the directed acyclic graph of a Bayesian network represents the structure of the gene regulatory network, while the set of parameters for the graph represents the statistical hypothesis behind the network. Gene X regulates gene Y if and only if there is a direct edge from X to Y in the graph. In order to construct the Bayesian network, one needs to learn the network using the observed data. However, this can be computationally hard, especially if the temporal aspects of the gene expression data are taken into account.

3. THE EVENT METHOD

The correlation and edge detection methods can be described as the opposite ends of a spectrum: the correlation method focuses on the global match of two profiles, while the edge detection method focuses on strong local matches. Also, the correlation method cannot make use of the temporal evidence in the data, and neither method takes into account the directionality of regulation. While one may be able to accommodate these factors with Bayesian networks, the associated computational costs can be very high. Thus, we feel the need to develop a more balanced method that can detect both global and local similarity features and take temporal issues into account. We also want our method to accomplish this in a computationally efficient way. The event method that we describe in this section is designed to meet these criteria.

There are two types of regulation at the level of transcription—activation and inhibition. In the activation process, the product of gene A affects the transcription process of gene B such that the production rate for gene B increases. Conversely, the inhibition process involves gene A's product decreasing the production of gene B. Activation or inhibition can take place through the regulator directly binding to the targeted gene or by binding another regulator and thus controlling it indirectly.

If one is hypothesizing that gene A activates gene B, one would expect to see in their data a rise in A followed by a corresponding rise in B, and a fall in A followed by a fall in B. The expectation would be reversed for inhibition. One would also expect to observe a certain amount of time delay between two corresponding events. The algorithm tracks these directional changes, dubbed "events," by calculating the slope of the expression profile at each time interval. These events signify the state of the gene expression at an instant—whether there is an increase in the expression, or decrease, or neither. Thus, depending on the slope value, the algorithm marks each event as rising (R), constant (C), or falling (F), resulting in a string of events. Under ideal circumstances, if the hypothesis of A activating B were correct, each event in A should be matched with a corresponding event in B.

To perform the matching of corresponding events while taking noise and temporal issues into account, we perform a sequence alignment of the event strings and obtain a numerical score that reflects the likelihood of A and B having a regulatory relationship. Also, since we do not know beforehand whether A or B should be hypothesized as the regulator, we evaluate both hypotheses by performing the algorithm in both directions and choosing the higher-scoring result. For inhibitory regulation relationships, we first complement the event string of the gene hypothesized to be inhibited by changing each R to F, and vice versa, while C remains unchanged. Then, we perform the alignment and scoring steps as explained above. Because the inhibitor gene is exerting its influence on the inhibited gene, the time delay relationship between the two remains unchanged.

To illustrate the main stages of the algorithm, Figure 1 outlines the process for YGL207W and YDR224C, two genes from the yeast gene expression data obtained from the alpha factor arrest experiment [8]. These genes are known to have an activation relationship in transcription regulation.

3.1 Conversion of Data into Events

In order to compare two gene expression curves, we first convert the raw data to a string of events. An event at a specific time interval represents the directional change of the gene expression curve at that instant. The conversion process involves the following steps.

1. Before calculating the slope at each time point, we first perform smoothing and filtering on the raw data to remove any significant noise present in the curve that may lead to erroneous interpretations. Smoothing is performed by a sliding window method in which at every time point, the data points within the specified window are averaged, so that small irregularities may be removed. The algorithm increases smoothing window sizes, ω, until it finds one where the event strings obtained with ω and $\omega + 1$ are equal to each other, or until $\omega = \omega_{max}$, where ω_{max} is the maximum allowed window size.

2. Next, we calculate the slope at each time point based on the smoothed data.

3. Finally, the slope values are converted into events. The

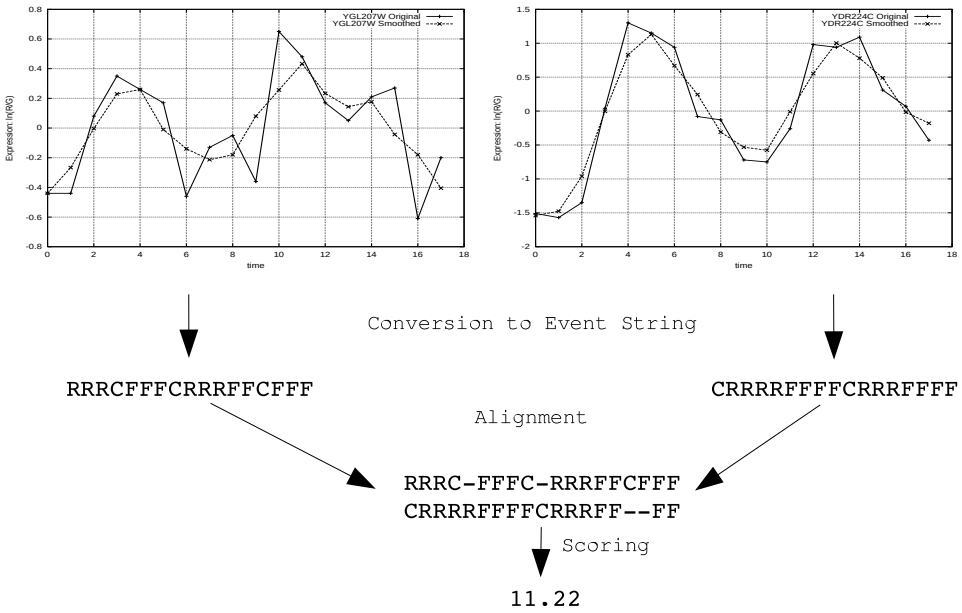


Figure 1: Outline of the Event Method. Expression curves for two yeast genes YGL207W and YDR224C (both before and after smoothing shown) are converted into event strings, which are then aligned and scored.

slope values to be classified as constant (C) are determined based on a threshold parameter θ that specifies the percentage of the data points for the gene that are to be classified as constant. The boundary slope values that result in this percentage of constant events are then used to classify all segments of the given expression profile: segments with slope greater than the higher boundary are classified as R, those with slope between the boundaries as C, and the rest as F.

For each gene expression curve, this process results in a string of event characters. For our experiments, we chose to perform smoothing with $\omega = 3$ and $\theta = 0.2$. These two values were chosen empirically, as they gave the highest number of true positive results when tested on biological data sets. The event strings for YGL207W and YDR224C are shown in Figure 1.

3.2 Alignment of Event Strings

Now that we have the event strings, we need to determine whether the order of the events indicate a possible regulatory relationship by finding the best match between the two strings while taking the noise and time delays into account. This problem can be approached similarly to the problem of biological sequence alignment; given the two event strings, we can efficiently determine their best alignment according to a suitably defined scoring function (see below) using a modified version of the Needleman-Wunsch algorithm for global sequence alignment [7] that takes into account the time delays between aligned event characters. We also need to ensure that there is no negative time delay—if the working hypothesis is that gene A activates gene B, events in A must always occur before their counterparts in B. The alignment result of YGL207W and YDR224C, our running example, is shown in Figure 1.

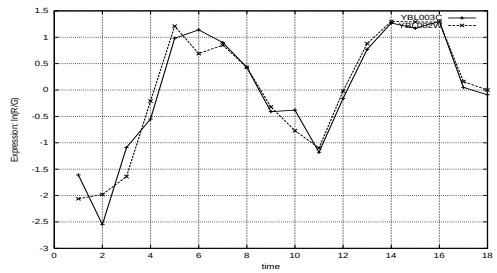
3.3 Scoring Matrix

Our scoring function is based on a scoring scheme for individual event characters. We capture this scheme in a scoring matrix that takes into account the time delay between the two events being compared, as shown in Table 1. This matrix is a form of similarity matrix used to evaluate how well two gene expression profiles match our working hypothesis. The weights of the matches, $S(dT)$, are functions of the time delay dT. As dT between the two events increases, their score is decreased in a linear fashion. This is to emphasize the fact that if two events are too far apart from each other, it is unlikely that they reflect a regulatory relationship. A linear time delay penalty was chosen over an exponential penalty after evaluating both schemes empirically.

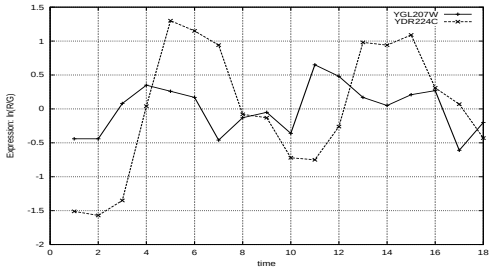
Table 1: Scoring Matrix for the Event Method ($0 < S(dT) < 1$, $N = 0$, $0 < \alpha < 1$, $0 < \beta < 1$, dT = time delay between two events. If dT is negative, the match is assigned ∞ as penalty, since such matching is not allowed.)

	R	C	F
R	$S(dT)$	0	$-\beta(dT)$
C	0	0	0
F	$-\beta S(dT)$	0	$\alpha S(dT)$

Just as in protein sequence alignments, different event matches have different weights. In our algorithm, the R-R matches are assigned higher weights than the F-F event matches. This is signified by the constant α ($0 < \alpha < 1$), that is multiplied to $S(dT)$ for F-F matches. Our data comes from cellular mRNA levels. While corresponding increases in mRNA levels of two genes are good indicators of a potential regulatory relationship, this may not be true for decreases in mRNA levels, because the latter may reflect other factors, such as different half-lives of different mRNAs. Thus, it appears reasonable to assign more weight to R-R matches. Any event that is matched with C is assigned the neutral score of 0. Constant events define regions of uncertainty. They could be due to any number of reasons, from simple noise to saturation of cellular mRNA or other factors. Thus, as they cannot aid in determining the potential



(a) YBL003C and YBL002W



(b) YGL207W and YDR224C

Figure 2: Two pairs of gene expression profiles that score differently in the correlation and the event methods. (a)YBL003C and YBL002W score high in both methods, as they are almost identical. (b)YGL207W and YDR224C score high in the event method, but relatively low in the correlation method because of the temporal lag between the corresponding events.

regulatory relationship of the genes being compared, they are assigned neutral scores. Finally, if there is a R-F mismatch, a penalty specified by $-\beta$ $(0 < \beta < 1)$ is multiplied to $S(dT)$.

This scoring matrix allows us to control the behaviour of our algorithm in a detailed and meaningful way. The parameters associated with the scoring matrix can be changed as necessary according to the details of the data that are being analyzed by the algorithm. For our experiments, values of $\alpha = 0.7$ and $\beta = 0.3$ were used.

3.4 Comparison with Existing Methods

While the correlation method is good at detecting the global similarity between two sequences, time delays can reduce its effectiveness for finding gene regulatory relationships. An example for this limitation is shown in Figure 2.

The edge detection method strongly focuses on the local matches between two gene expression curves. While it is important to identify genes with a high degree of local similarity in the respective expression profiles, this bias has the undesirable effect of ignoring weaker but still significant profile similarities. Consider the expression profiles for genes YGL207W and YER111C shown in Figure 3; these genes are known to have an activation relationship with each other. While both correlation and the event method assign high scores to the pair, the edge detection scores it rather low because of an insufficient number of edges matched.

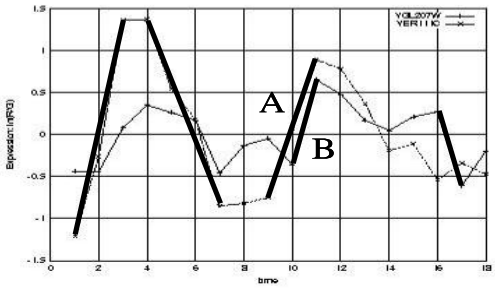


Figure 3: The edges found by the edge detection method in the profiles of genes YGL207W and YER111C (bold lines). Only the edges marked A and B can be matched with each other.

4. EXPERIMENTAL RESULTS

In order to assess our algorithm against existing methods, we conducted an empirical comparative performance analysis of the three methods—the correlation method, the edge detection method, and the event method, on various sets of real-world and synthetic data.

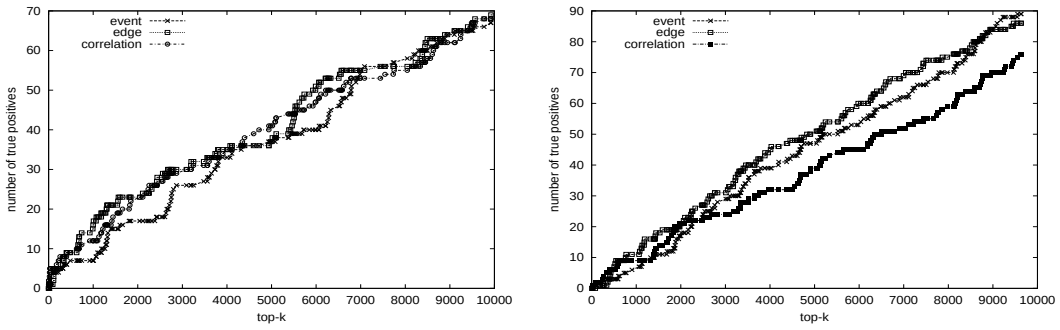
4.1 Spellman's Data Sets

Spellman *et al.* [8] sought to build a comprehensive catalogue of cell cycle-regulated genes in the yeast *Saccharomyces cerevisiae*. They performed a series of microarray experiments in which they took mRNA level measurements for all yeast genes at regular time intervals. They then combined their results with those by Cho *et al.* [2] to produce a more comprehensive collection of data. The test samples were synchronized so that all the cells would be at the same stage in their cell cycle. Three different methods were employed to arrest the cells at the same stage: alpha-factor arrest, elutriation, and arrest of CDC15 and CDC28 temperature-sensitive mutants. The reported expression levels are the log ratios of the test sample expression by control sample expression level measurements.

Because Spellman's data sets contained expression profiles for all open reading frames in the yeast genome, numbering over 6000, it was necessary to find a subset of these genes in order to reduce the search space. Filkov *et al.* [3] created a subset of 888 known transcriptional regulation pairs, including 647 activations and 241 inhibitions. We used the alpha-factor and CDC28 data sets for our experiment. After filtering out all genes with a significant number of missing data points from these data sets, we analyzed the known regulation subsets using the three algorithms.

From the 888 known regulation pairs, the number of distinct genes that could be analyzed was 348 for alpha data set and 458 for CDC28 data set. This means that there are over 120,000 possible pairs that can be formed in the alpha data set, and over 200,000 pairs in the CDC28 data set. In the context of our evaluation, every pair of genes that occurs in the list of 888 known pairs is a true positive. We should note that because only the event method takes directionality into account, we had to compensate for this when comparing with the other two methods, lowering the number of possible pairs by half. Figures 4 shows how many true positives were found by the three methods in their top-k candidates, where k varies from 0 to 10,000. All three methods performed comparably.

Table 2 shows the degree of overlap between the pairs



(a) Alpha Data Set

(b) CDC28 Data Set

Figure 4: True Positive Distributions for Top-k ($0 < k < 10,000$)

Table 2: Overlapping Results Among Three Methods (All Results / True Positive Results)

Methods	Alpha	CDC28
Event + Correlation	3367 / 11	2916 / 9
Event + Edge	2081 / 0	3362 / 0
Correlation + Edge	1989 / 0	2252 / 0

returned by the three algorithms, when looking at the top-10,000 rankings by each method. No more than 1/3 of the results returned by any two methods overlap with each other, indicating that the event method finds significantly different pairs from the other two methods. Table 2 also shows the number of overlapping true positive pairs when looking at the top-10,000 rankings by each method. It is evident that there is very little overlap, if at all.

The event method produced the list of potential *inhibitory* regulations as well, but as the other two methods are not designed to find inhibitions, they could not be compared to each other. Using the same parameters as above, the event method found 27 true positive inhibitory relationships in the alpha data set, and 17 in the CDC28 data set within the top-10,000 ranking pairs. Using different parameters produced better results, but lowered the number of true positive activations it found.

There were a few significant problems associated with the obtained results that made the analysis inconclusive. First, because of the poor resolution of the data, many pairs of genes showed almost identical expression curves, making a more detailed analysis difficult. If most of these pairs with strong similarities were in fact regulatory pairs or at least be involved in similar pathways, one could assign strong confidence to using these methods for analyzing the data. However, when the high-ranking pairs returned by the three methods were compared against the yeast gene databases, this was not always the case. While many pairs shared common promoters or were components of the same protein complex, others were unrelated genes with no obvious relationship.

The edge detection method posed another serious problem. The program, kindly provided by its original authors, had difficulty in finding significant edges in many gene profiles, thereby giving zero score to a high portion of the genes. Even for those genes that received nonzero scores, the number of edges that were found per gene was extremely low, which casts doubt on the significance of the scores.

4.2 Synthetic Data Sets

Because of the limitations of Spellman's data sets, we used additional synthetic data sets for a more detailed evaluation of the algorithms. Four data sets were designed to test specific features of the three algorithms. In our experiment, each data set consisted of curves designed to show regulatory relationship and randomly produced curves. The regulatory curves, named $gene_i$, where $0 <= i <= 10$, were produced in such a way that $gene_i$ and $gene_{i+1}$ would differ with respect to the factors listed below. For example, with curves produced to test for constant time delay, $gene_{i+1}$ would be a time-shifted version of $gene_i$. Our synthetic data sets take the following factors into consideration:

1. *Constant Time Delay:* The time delays between corresponding changes in two curves are always constant. A signal curve $gene_{i+1}$ is generated by time shifting $gene_i$ by a fixed amount.

2. *Irregular Time Delay:* The time delays between corresponding changes in two curves are variable. A signal curve $gene_{i+1}$ is generated by time shifting $gene_i$ by a randomly chosen amount.

3. *Partial Matching:* Only a section of the two curves can be matched, and the rest of the curves are filled with random edges. A signal curve $gene_{i+1}$ is generated by randomly changing $gene_i$ within a specified range.

4. *Differential Weighting of Events:* The two curves $gene_i$ and $gene_{i+1}$ share major, matching rising edges, but the rest of the curves are filled with random edges. This is to test for differential weighting, where rising edges are weighed more than falling edges.

To evaluate how the algorithms perform with the synthetic data sets, we counted the number of true positive pairs that they found. In this case, we need to account for the fact that pairing $gene_x$ with $gene_{x+r}$, where r is relatively small, should be considered as better matches than pairing $gene_x$ with a random curve. Thus, we specified a range parameter r so that for $gene_x$, a match with any of $gene_{x-r}$ through $gene_{x+r}$ would qualify as a true positive pairing. We generated and tested five sets of data for each category, then averaged the total number of true positives found. Each data set contained 11 signal curves and 11 random curves. The range for true positive classification was

Table 3: Average Number of True Positives from Synthetic Data Sets ($r = 2$)

Data Set	Correlation	Event
Constant Time Delay	31.6	39.8
Irregular Time Delay	27.2	33.8
Partial Matching	44.6	40.6
Differential Weighting	36.2	45.0

set at 2. Unfortunately, we could not compare the performance of the event method to that of the edge detection method, as the implementation of the method that we had available was unable to produce non-zero scores in most of the synthetic data pairs that we tested.

The results listed in Table 4 show that except for the Partial Matching sets, the event method was superior to the correlation method. We should note that the advantage enjoyed by the correlation method was diminished when time delay was introduced to the Partial Matching data sets.

5. CONCLUSIONS AND FUTURE WORK

We presented a new algorithm, called the *event method*, that can find potential activation and inhibition pairs from gene expression data. The event method is based on some key features of gene expression, such as time delays and asymmetry between rising and falling edges. It is computationally efficient. The method is shown to perform comparably to the correlation and edge detection methods in finding true positive regulation pairs from Spellman's yeast data sets, and outperforms correlation on our synthetic data sets. More results from this study can be found in [6].

In light of the limitations of the data used in our experiments, it would be interesting to consider other types of higher-quality time-series data. Also, integrating the microarray data with other types of a priori knowledge would help narrowing down the search space. Creating a more realistic synthetic data set for testing the algorithms should also prove to be interesting. An ideal synthetic data set would come from an artificial regulatory network that incorporates as many features of the real one as possible so that it would be a reliable indicator of how the algorithms would perform with good, high-quality data. Such a network would consider the reaction kinetics of various pathways, the effect of polymerization necessary for proteins to become active, formation of protein complexes etc. [5, 10].

It would be desirable to study the effects of changing the number of event types on the performance of the event method. We used three event types in our experiments, but one could increase this number so that the event strings would represent the gene expression profiles in more detail. The effects of the algorithm's parameters on its performance should be further investigated. While we did some empirical tests in choosing the parameters we used, the quality of the data we were working with here may have prevented us from gaining more insight into the behaviour of the algorithm as the parameters change. Finally, it is possible that by using a local alignment algorithm instead of the global alignment method used here, the performance of the event method could be improved by focussing it more on local changes in the gene expression profiles that may reflect complex, conditional regulatory relationships.

Once we are left with potential regulatory pairs, it would be prudent to remove any spurious pairings from the ranked list. One possible approach is to perform transitive closure removal: If there is a regulatory relationship between genes A and B, and B and C, any high scores between genes A and C may be due to the fact that they are related through gene B only. Removing such pairs would allow more potential candidates to be placed in the rankings. Ultimately, methods for combining the ranked pairings into hypotheses on larger fragments of the underlying regulatory network should be studied. Transitive closure removal may help in this context, as it would result in clusters of genes that are connected by regulatory relationships.

6. ACKNOWLEDGMENTS

We would like to sincerely thank Vladmir Filkov and Jizu Zhi for making available their code and data, as well as for valuable additional information on their results. We would also like to thank the anonymous reviewers of this paper for providing us with valuable suggestions.

7. REFERENCES

[1] Y. Barash and N. Friedman. Context-specific bayesian clustering for gene expression data. In *Proceedings of RECOMB 2001.*

[2] R. J. Cho *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, 2:65–73, 1998.

[3] V. Filkov, S. Skiena, and J. Zhi. Identifying gene regulatory networks from experimental data. In *Proceedings of RECOMB 2001.*

[4] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Comp. Biol.*, 7(3-4), August 2000.

[5] T. Knight and G. Sussman. Cellular gate technology. http://www.ai.mit.edu/people/tk/ce/cellgates.ps, 1997.

[6] A. Kwon. Inference of transcriptional regulation relationships with gene expression data. Master's thesis, submitted to Dept. of Computer Science, Univ. of British Columbia, 2002.

[7] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Mol. Biol.*, 48:443–453, 1970.

[8] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* microarray hybridization. *Mol. Biol. of the Cell*, 9:3273–3297, December 1998.

[9] P. Spirtes, C. Glymour, and R. Scheines. Constructing bayesian network models of gene expression networks from microarray data. In *Proceedings of the Atlantic Symposium on Computational Biology.* Genome Information Systems & Technology, 2000.

[10] R. Weiss. *Cellular Computation and Communications Using Engineered Genetic Regulatory Networks.* PhD thesis, Dept. of E. Eng. and Computer Science, MIT., September 2001.