Analysis Techniques for Microarray Time-Series Data

Vladimir Filkov Steven Skiena Jizu Zhi
Dept. of Computer Science and Center for Biotechnology
State University of New York
Stony Brook, NY 11794-4400
{vlfilkov—skiena—zjizu}@cs.sunysb.edu

Abstract

We introduce new methods for the analysis of short-term time-series data, and apply them to gene expression data in yeast. These include (1) methods for automated period detection in a predominately cycling data set and (2) phase detection between phase-shifted cyclic data sets. We show how to properly correct for the problem of comparing correlation coefficents between pairs of sequences of different lengths and small alphabets. In particular, we show that the correlation coefficient of sequences over alphabets of size two can exhibit very counter-intuitive behavior when compared with the Hamming distance. Finally, we address the predictability of known regulators via time-series analysis, and show that less than 20% of known regulatory pairs exhibit strong correlations in the Cho/Spellman data sets. By analyzing known regulatory relationships, we designed an edge detection function which identified candidate regulations with greater fidelity than standard correlation methods.

1. INTRODUCTION

New experimental technologies in molecular biology (particularly oligonucleotide and and cDNA arrays) now make it possible to quickly obtain vast amounts of time-series data on gene expression in a particular organism under various conditions. Extensive time-series data on gene expression in yeast (Saccharomyces cerevisiae) have been obtained by Cho [3] and Spellman [10] using microarrays, greatly expanding our knowledge of which genes involved in cell cycle regulation.

The importance of the Cho and Spellman data sets is perhaps best revealed by the variety of methodologies being applied to analyze it. Clustering studies and promoter analysis [4, 10] have been used to classify genes according to where they are active in the cell cycle. We [1] have developed a system for proposing putative gene regulatory networks by identifying activators and inhibitors using signal processing and combinatorial optimization. Friedman et al [6] have built a system similar in spirit analyzing the same data, but based instead on Bayesian networks. Techniques for analyzing these data sets using differential equation modeling [2], wavelets [8], and singular value decomposition (SVD) [7] have also been explored.

In general, this analysis has succeeded in revealing certain gross periodicities in the data corresponding to the cell and other cycles, and generated untested predictions of putative regulators. Microarray technology is limited in sensitivity when comparing the relative concentrations of different genes in a single experiment. Further, the costs associated with current microarray technologies, as well as the nature of these experiments do not allow for experiments to be repeated. This means that microarray data cannot be analyzed for accuracy or precision, since there can be no statistical study of a single measurement. Therefore, the available gene expression data is intrinsically noisy and difficult to analyze.

In this paper, we address several fundamental questions concerning possible limitations to which informed predictions can be generated using these data sets:

- *Predictability of Known Regulations via Time-Series Analysis*

  Previous systems for inferring regulatory networks from these microarray data sets [1, 6] have bravely assumed that the Cho/Spellman data sets contain sufficient information to identify a substantial fraction of all regulators. To test this hypothesis, we analyzed the literature to compiled a database of known regulatory relations in yeast. Less than 20% of these regulatory pairs exhibited strong correlations in the Cho/Spellman data set, clearly insufficient to infer large regulatory pathways.

- *Improved Edge Detection for Regulatory Relations*

  By analyzing the known regulatory relations which were actively expressed in the Cho/Spellman data, we were able to design and analyze an edge detection function which identifies these relations with greater fidelity than standard correlation methods. We believe that our edge detector is now significantly better at identifying interesting regulatory candidates than previous work [1].

- *Periodicity and Phase Shift Analysis for Time-Series Integration*

  The Cho/Spellman data sets comprise four distinct time-series data sets, each measuring gene expression

Data set	Period observed	Period detected	δt	# samples	# full orfs
alpha	66 ± 11 min.	70 ± 7 min.	7	18	3361
cdc28	90 ± 10 min.	100 ± 10 min.	10	17	1188
cdc15	70 ± 10 min.	90 ± 10 min.	10/20	24	3453
elu	—	—	30	14	4753

Table 1: Cho/Spellman data sets with observed and detected cell-cycle times.

for all 6178 ORFs in yeast. Each data set uses cell cultures synchronized by a particular method (cdc15, cdc28, alpha, and elutriation), which starts each in distinct phases of the cell cycle, with varying cell cycle lengths and sampling frequency. Integrating these distinct time-series into a common reference frame requires methods for inferring periodicity and phase shifts on noisy data. Our techniques yield period length and shift values consistent with those observed by the experimenters, and are of independent interest.

- *Comparing Correlations of Distinct Length Sequences*

  Our techniques for periodicity and phase shift analysis require comparing the significance of correlations of short sequences of different lengths. Properly interpreting these correlations is a subtle problem. For example, two random sequences of length five will have a correlation of 0.8 or higher roughly 60% of the time, a likelihood which reduces sharply with increased sequence length. We show how to correct for such anomalies, and apply it successfully to identify periodicity / phase shift in the Cho/Spellman data.

- *Correlation Significance of Small Alphabet Sequences*

  The limited resolution inherent in measuring noisy experimental data can be modeled by grouping data values into a small number of bins for analysis. This yields data streams over a small alphabet, and alphabet of size two in the limiting case. We explore the significance of the standard correlation coefficient measure on sequences over alphabets of length two, and show that the results can be inherently misleading. For example, we show that two sequences of any length can have a correlation coefficient near 0 (i.e. essentially uncorrelated) even though they are identical in all but 2 positions! We introduce analytical methods to bound the regions where allowable Hamming distance/correlation coefficient pairs exist, which are of independent interest.

This paper is organized as follows. In Section 2, we describe the particulars of the Cho/Spellman data sets which we analyze throughout this paper. In Section 3, we explore techniques to extract cycle period length and phase shifts by appropriately weighting correlation coefficients by the length of the associated sequences. In Section 4, we compare the correlation coefficient with the Hamming distance metric as a similarity measure for sequences of size two alphabets, with surprising results. In Section 5, we demonstrate that only a small fraction of all known regulatory relations are exhibited in these data sets, and use these results to guide the design of an effective edge detector for identifying interesting pairs of expressed genes.

2. MICROARRAY DATA SETS

Our experiments were conducted on the data of Cho [3] and Spellman [10], which is available at *http://genome-www.stanford.edu/cellcycle/*. It is comprised of four time-series, where each data set contains temporal concentration measurements for all 6178 ORFs in yeast. Each of the experiments starts with a cell population which has been synchronized with distinct methods (cdc15, cdc28, alpha factor, and elutriation Elu), which arrest the cells in the same state by introducing external substances, changing environmental conditions, or selecting cells of the same size and hence are likely in the same state. The time-series courses have been repeated through one+ periods for Elu, two+ periods for alpha and cdc28, and three+ periods for cdc15.

Table 1 gives the cycle-times for each data set, as reported by the original researchers [10], the cycle-times for each data set that we detected, the time interval between sampling points, the number of sampling points in each experiment, and the number of curves (out of 6178) that had no missing points in them. This last column gives the number of sequences from each data set with complete enough data for analysis.

Out of the four data sets, we found the cdc28 (Cho) and alpha data sets to be the most reliable (i.e. without discrepancies and permitting analysis) because the cdc15 data set had a lot of missing points and the elu data set had been sampled for one period, and only coarsely at that.

3. INTEGRATING DATA SETS

We sought to interleave all four of the Cho/Spellman cell cycle data sets, with the goal that the integrated data set would gain precision and accuracy through smaller gaps between sampling times. In addition, the integrated time series would clearly reveal genes whose expression grossly differs in one data set from the rest. This would be indicative of either experimental error or genes whose expression is effected by one of the synchronization factors.

Interleaving the data sets requires identifying the cell-cycle periods of each data sets as well as detecting the cell cycle phase each experiment started in after synchronization. Since the experimenters grossly observed the cell-cycle length of each time course, we can use these lengths to assess to performance of our algorithms. Further, Spellman [10] determined the shifts as a side effect of their cycling gene detection algorithm, which roughly performed a Fourier analysis over all genes, by maximizing the sum of all the transforms over different phase shifts.

The problems of determining the cell cycle period and detecting phase shift of a data set are intimately connected, since it is impossible to identify the position in the cell cycle without knowing the length of the cell cycle.

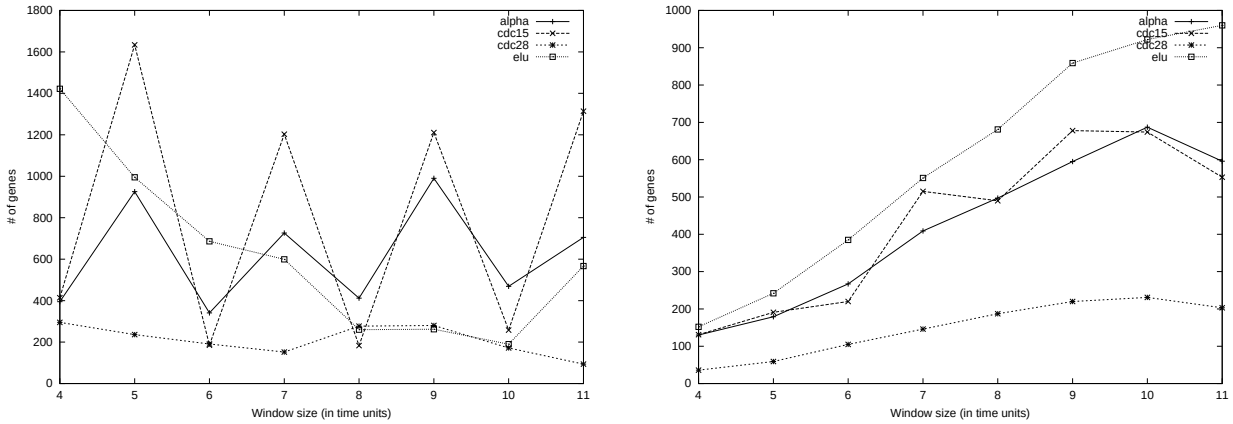

Figure 1: Inferred periods for all data sets using (left) standard and (right) length-corrected correlation coefficients.

3.1 Cycle detection in Microarray Data

Fourier analysis is the standard method of identifying cycles in periodic data sets when the number of sampling points is large, but this is not the case with the gene expression data sets. Instead, we explored how well discrete similarity methods can work, which we used for analyzing short sequences in [1]. We have previously explored geometric approaches to cycle detection [5].

Our main idea is as follows. Consider a periodic time-series curve with more than one period. The similarity of a "window" (i.e. a sub-range of the time-series function) with the same-length window immediately following it should be greatest when their length equals the period of the whole curve . Thus by varying the length of the window we can find the optimal period for each gene.

Since each yeast gene's expression follows the general cell cycle if it exhibits any cycling behavior, we expect that this method will detect the periods of the cycling genes while extracting random periods for the non-cycling genes. Thus we expect the period of the data set to become apparent when averaging over all genes.

3.1.1 Methods and Results

We chose the standard correlation coefficient to be our similarity measure of two sequences x_i, y_i of length n,

$$r(X,Y) = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sqrt{(\sum {x_i}^2 - (\sum x_i)^2/n)(\sum {y_i}^2 - (\sum y_i)^2/n)}} \quad (1)$$

The correlation coefficient is known to give good results when employed in clustering and analyzing time-series [4, 10]. Curves that are visually very similar typically score as being highly correlated.

More formally, let $C_{i,1}, C_{i,2}, \ldots, C_{i,k}$ denote the k values of the expression profile C_i associated with orf i. We correlated the time series $C_{i,1}, C_{i,2}, \ldots, C_{i,t}$ with $C_{i,t}, C_{i,t+1}, \ldots, C_{i,2t}$ for all possible inferred periods t, where $5 \leq t \leq k/2$. This constant 5 denotes the shortest length sequence where we deemed the correlation coefficient as having any predictive power. Thus, for each orf we obtained an array of correlations over t. Summing up these correlations over all expression curves on each of the four data sets yielded the results in Figure 1(a).

As is apparent from Figure 1(a), three of the four data sets exhibit several local maxima. Only the elu data set does not show any significant number of oscillating genes. Alternating peaks and troffs are particularly apparent for the alpha and cdc15 data sets. We observe the following from Figure 1(a):

- There are very strong indications of an underlying oscillatory phenomenon for the cdc15 and alpha data sets, with a period smaller than the cell cycle. This period is on the order of 3 sampling points or approximately 20 minutes for alpha.
- There is also an apparent periodic signal present which has a smaller period, somewhere around 40 minutes. This phenomenon was also reported in [8]. These frequencies are apparently undetectable by Fourier analysis, but can be found with methods suitable for analysis of short, coarsely sampled sequences, like ours or wavelet analysis.
- The actual cell-cycle periods are in fact strongly expressed in all the graphs, but they are not necessarily the most strongly expressed, indicating that uncorrected correlations are not sufficient to detect the periods.
- There is an overall tendency towards smaller periods, i.e. all graphs are somewhat skewed towards left.

3.1.2 Correcting for Correlations of Sequences of Different Length

Our cycle detection algorithm required comparing correlations of sequences as short as 5 (4 time units) with correlations of sequences as long as 12 (11 time units) in analyzing the Cho/Spellman data sets. However, high correlations have a greater probability of occurring by chance on short sequences than long sequences. The phenomenon caused curves of Figure 1(left) to be skewed towards smaller shifts. Here we propose a statistical method to correct for this phenomenon.

We can model pairs of sequences without meaningful correlations to be random strings over finite alphabets. Under such a model, shorter sequences correlate higher than longer sequences do, primarily because the event space is much

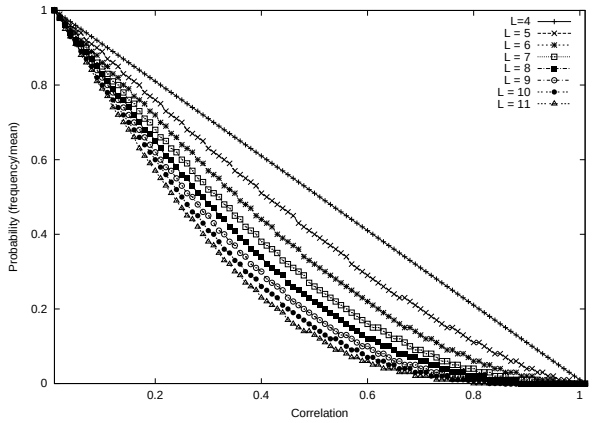

Figure 2: Cumulative distribution function of positive correlation coefficients for random sequences of length 4 to 11.

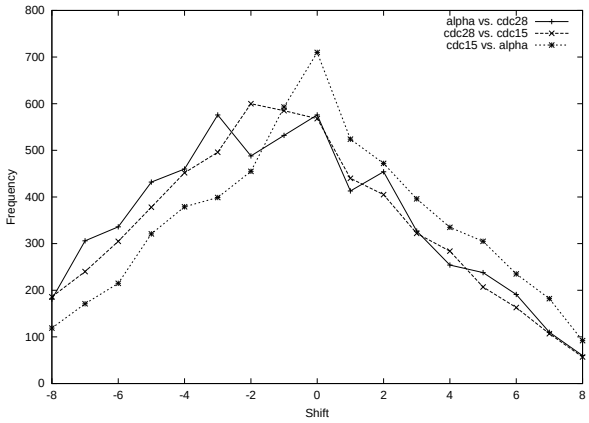

Figure 3: Phase shifts maximizing correlation for all pairs of alpha, cdc15, and cd28 data sets.

smaller for shorter strings over a fixed size alphabet. Thus in order to compare correlations of sequences of different sizes, we needed to account for this probabilistic effect.

We ran a Monte Carlo experimental study where we generated twenty million random sequences of different lengths to determine the likelihood of each correlation value occurring by chance. Each of the random numbers came from a standard normal distribution over the alphabet $\{1, \ldots, 1000\}$. The results are reported in Figure 2. We see that the probability that two random sequences achieve a correlation of $\geq x$ decrease sharply with increasing sequence length. Given any two sequences of length bounded by our study, Figure 2 gives the likelihood of their correlation occurring by chance. Such likelihoods can be meaningfully compared for sequences of different lengths.

Statistically, Figure 2 gives the distributions of the correlation coefficient for various lengths. Finding analytical forms for these distributions is known to be a difficult problem [9].

The results of Figure 2 can be used to correct the plots of Figure 1(a) in comparing correlations of different length cycle windows. Corrected plots are provided in Figure 1(right). Clearly, the oscillations in the raw data have been dampened and even elliminated. Moreover, the corrections have revealed peaks around a single time point for all data sets except the Elu data, which in fact did not cycle. Further, after converting the observed cycle length from number of points to time for each experiment using the sampling frequencies in Table 1, we deduce that the periods are: 70 minutes for alpha, 100 minutes for cdc28, and 90 minutes for cdc15. As shown in Table 1, these computational results match the observed cycle times very well.

3.2 Phase Shift detection

In addition to changing period lengths, cell synchronization procedures leave cells in different stages of the cell cycle. These phase shifts must also be recovered to interleave the data.

We used a similar method to detect shift as cell-cycle length, but this time correlated across data sets. Interleaving these data sets required finding all pairs of relative phase shifts, from which absolute phase shifts could be extracted. To do this, we shifted corresponding curves from different data sets along each other.

Using the periods of the time series computed in Section 3.1, we normalized the periods of the cdc28 and cdc15 time series curves with respect to the alpha data set so that all resulting time series had the same period length. Then we correlated the corresponding time-series curves from different data sets across all possible shift lengths, obtaining for each ORF pair a shift vector of correlations. By summing up these shift vectors for all curves, we obtain frequency distributions and can establish which shifts maximized the correlation for most curve pairs. This distribution is provided in Figure 3.

These results demonstrate that alpha and cdc15 are unshifted relative to each other, while cdc28 shifts 1 to 2 time periods from cdc15. The results are less clear comparing alpha to cdc28, likely representing a shift of from 0-2 units. In fact, the alpha and cdc28 time-series both started in $G1$ phase of the cell cycle, and cdc15 in M phase [10]. Our observed shifts are basically on target. For the transformed periods of roughly 70 minutes each, the computed shift offset of 14 minutes lies within the M - $G1$ time difference, since the M phase occupies roughly 50% of the cell cycle time and $G1$ roughly 15%.

4. CORRELATION COEFFICIENT AND SMALL ALPHABETS

The correlation coefficient is highly regarded as a measure of similarity between pairs of sequences, and has been widely used to analyze gene expression data [4]. However, over the course of our experiments indications arose that the correlation coefficient may perform very badly in comparing sequences drawn over small, finite alphabets. Such data arises naturally when attempting to smooth noisy data signals by quantizing them into a small number of bins. Therefore, in this section we take a closer look at the behavior of the correlation coefficient on small alphabet sequences.

We will restrict our attention to the limiting case of sequences over two letter alphabets, i.e. sequences drawn from the alphabet $\Sigma = \{m, l\}$. For said sequences, the Hamming distance, defined as the number of positions in which the sequences agree, offers a natural measure of similarity, in fact a well-defined distance metric. To get an objective measure

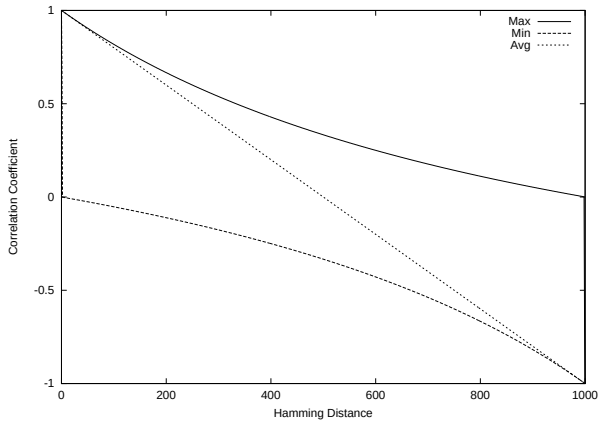

Figure 4: Correlation coefficient bounds and average for Hamming distance, for sequences of length 1000.

of how the correlation coefficient performs on two letter alphabet sequences, we can compare its performance to the Hamming distance metric.

Observe that any sequence pair $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ over $\Sigma = \{m, l\}$ can be viewed as a sequence of length n over the alphabet $\Delta = \{\binom{m}{m}, \binom{m}{l}, \binom{l}{m}, \binom{l}{l}\}$. Let a, b, c, d denote, respectively, the counts of character from Δ in the pair (X, Y). Thus a denotes the number of $\binom{m}{m}$'s, b the number of $\binom{m}{l}$'s, c the number of $\binom{l}{m}$'s and d the number of $\binom{l}{l}$'s. The sequence of symbols can be permuted without changing the correlation coefficient of (X, Y):

$$\underbrace{\binom{m}{m}, \ldots, \binom{m}{m}}_{a}, \underbrace{\binom{m}{l}, \ldots, \binom{m}{l}}_{b}, \underbrace{\binom{l}{m}, \ldots, \binom{l}{m}}_{c}, \underbrace{\binom{l}{l}, \ldots, \binom{l}{l}}_{d}$$

Observe that

$$\begin{array}{lcl} \sum x_i y_i & = & am^2 + (b+c)ml + dl^2 \\ \sum x_i & = & m(a+b) + l(c+d) \\ \sum y_i & = & m(a+c) + l(b+d) \\ \sum {x_i}^2 & = & m^2(a+b) + l^2(c+d) \\ \sum {y_i}^2 & = & m^2(a+c) + l^2(b+d) \\ n & = & a+b+c+d \end{array}$$

so the correlation coefficient formula (Equation 1) reduces (after moderate algebra) to:

$$r(X,Y) = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \quad (2)$$

Hence, any two sequence pairs with the same counts a, b, c, d will have the same correlation coefficient, since eq. 2 does not depend on m or l. Further, any sequence pairs yielding distinct coefficients must have combinatorially different a, b, c, d counts. Observe that the Hamming distance of X and Y, $h = b + c$.

These observations make tractable the problem of exhaustively enumerating all possible correlation values for two sequences of length n, over Σ as a function of h, since the correlation is now a function of a, b, and h.

Using these ideas we ran a study over all combinatorially "different" sequence pairs over Σ, using an exhaustive search algorithm running in $O(n^3)$ time. We computed the maximal and minimal correlation coefficient values for each possible Hamming distance. We also computed the expected correlation for each h, by weighing each correlation coefficient as follows:

$$E_h = \frac{\sum r\, N(a,b,c,d)}{\sum N(a,b,c,d)}; N(a,b,c,d) = \frac{n!}{a!b!c!d!}$$

where the summations are over all different a, b, c, d, while $h = b + c$ is kept constant. $N(a, b, c, d)$ denotes the number of different ways to column-permute a sequence pair with counts a, b, c, d, and $\sum N(a, b, c, d) = 4^n$, where the sum is over all different a, b, c, d.

The results, shown in Figure 4 define tight upper and lower bounds to possible correlation coefficient as a function of the Hamming distance, and show the expected value of the correlation for each h.

Even though the expected correlation graph in Figure 4 indicates proper behavior of the correlation coefficient on the average, the upper and lower bounds demonstrate that even for very small Hamming distances, the possible range of correlations is distressingly large. For example, consider the following sequence pair of length n, which differs in only 1 of n positions: $a = 1$, $b = 1$, $c = 0$, and $d = n - 2$. These sequences have correlation of $\sqrt{1/2 - 1/2(n-1)}$, which approaches $1/\sqrt{2} \approx 0.7067$ for large n. Further, the sequence pair with the parameters: $a = 0$, $b = 1$, $c = 1$, and $d = n-2$, which differ in 2 of the n positions, correlate to $-1/(n-1)$ which approaches 0 for large n.

We must therefore conclude that the correlation coefficient is inadequate to compare two-letter-alphabet sequences, and must be viewed with caution in measuring the similarity of quantized or small alphabet sequences.

Note that for alphabets of size 3 or more, the correlation coefficient is no longer independent of the letters (numbers) of the alphabet, e.g. $(1, 2, 4)$ and $(1, 4, 2)$ correlate to $1/7$, whereas $(1, 2, 3)$ and $(1, 3, 2)$ correlate to $1/2$.

5. ASSESSING AND IMPROVING REGULATORY PAIR DETECTION

At least three software systems [1, 2, 6] have been developed to predict gene regulations using the Spellman/Cho time-series data. However, none of these systems was rigorously evaluated as to the quality of the resulting predictions. In this section, we demonstrate that the Spellman/Cho is inherently inadequate to identify the vast majority of known regulatory relations, but use insights from this analysis to construct an edge detector which appears better at selecting biologically interesting pairs of genes.

5.1 Evaluating Known Regulatory Pairs

To analyze the potential for determining regulatory pairs from the Spellman/Cho data sets, we began by constructing a database of known regulatory relations in yeast. Specifically, a keyword search on the Yeast Protein Database YPD (*http://www.proteome.com*) performed in February 2000 yielded 1007 regulated genes in yeast. By reviewing the published literature on these 1007 genes, we collected 888 transcriptional regulations of which 647 were activations and 241 were inhibitions. Altogether, 486 genes were involved in these transcriptional regulations.

We then mapped these 486 genes to the two highest-quality time series data sets, cdc28 and alpha. The results are described in Table 2.

Only 366 genes were successfully mapped to the cdc28 data set, with the balance of 120 genes failed primarily be-

	time points	time intervals	genes mapped	activations	inhibitions
cdc28	17	10 min.	366	469	155
alpha	18	7 min.	335	343	96

Table 2: Mapping known regulations to public data sets

cause of differing naming conventions between the literature and the data set. Of original 888 regulations, 469 activations and 155 inhibitions mapped correctly to the cdc28 data set. For the alpha data set, we selected only the 335 genes for which all 18 time-series points were available. The remaining genes include 343 known activations and 96 known inhibitions.

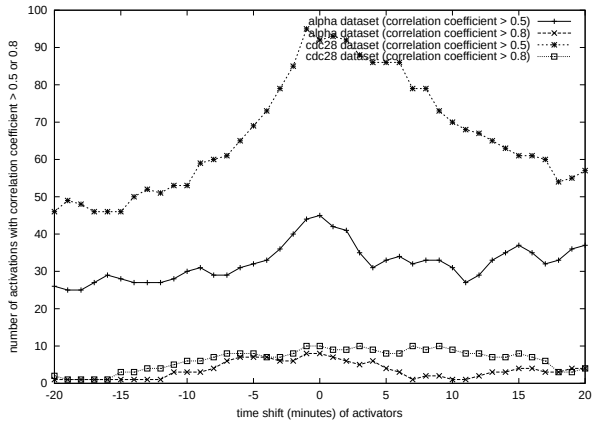

Figure 5: Effect of shift on the correlation coefficient

We then performed correlation coefficient analysis on each pair of activations in the remaining 469 and 343 known activating relations. The result shown in Figure 5 demonstrates a poor correlation between the activators and their corresponding activatees. Less than 20% of known regulations (93/469 for the cdc28 data set, 44/343 for the alpha data set) scored > 0.5 correlation coefficient between the activator and activatee. Shifting the curves of every activator (to compensate for the time-lag in activation) did not improve the number of correlations for any shift between ± 20 minutes, as shown in Figure 5. These curve shifts were performed on re-sampled interpolated data to facilitate short shifts. Finally, eyeball examination by a Ph.D. biologist (Zhi) of all plotted pairs of known regulations failed to extract a common pattern which could indicate a possible transcription regulation in more than 20% of these cases. Together, these results confirm the complexity of biological systems, and demonstrate that gene regulation is much more complicated than this model and data set can deal with.

5.2 An Improved Edge Detection Function

Although two genes within a regulated pair do not necessarily have similar expression patterns, expressionally correlated genes usually suggest a functional correlation. Thus correlation coefficient-based clustering has been widely used in analyzing time course expression data [4]. Here we propose an edge function, which was aimed to more accurately predict functional correlated genes. The correlation coefficient method fails to pick up strong local signals, as opposed to global similarity, particularly in noisy data. Our locally based edge detector algorithm is as follows:

1. Assign each time point of every gene a label depending upon whether it is a local minima or maxima. While comparing the expression value of neighboring points, we allow a given expression error level, typically 10%. Thus whenever the difference of expression value between two points is less than this expression error level, neither point can define a local optima.

2. For each gene, the labeled minima and maxima are connected using a four-step edge construction process:

   - *Primary edges* link neighboring local maxima and minima.
   - *Secondary edges* link all primary edges whose normalized expression, $(max - min)/average$, is greater than a threshold, typically 30%. Here, max, min, and $average$ pertain to the expression level. This accounts for the minimal biologically significant expression level change. Any changes below this level are probably due to experimental error.
   - *Tertiary edges* result from merging adjacent secondary edges of similar direction.
   - *Quadrary edges* result from eliminating narrow peaks or troughs. For example, in Figure 6(a), the narrow trough of CHA1 from time point 9 to 11 is probably due to an error at time point 10. Thus the edges from time points 9 to 10 and from 10 to 11 are eliminated.

   Finally, each gene is represented by an array of quadrary edges which we consider as biologically significant and reliable expression level changes.

3. Pairs of genes are *scored* solely based on their quadrary edges. To score genes G_a and G_b, we score each quadrary edge (e_a) of G_a against each quadrary edge (e_b) of G_b. provided the time difference between the two edges is $\leq \delta_{max}$. The similarity score S_g between G_a and G_b is given as:

$$S_g = \sum_{all\ e} d(1 - \frac{\delta}{\delta_{max}})/\sqrt{n_a n_b}$$

   where d denotes the agreement of the slopes of e_a and e_b. Specifically, $d = 1$ if the signs of the slopes agree, otherwise $d = -1$. The parameter δ_{max} defines the maximum allowable time difference (in minutes) between the middle of e_a and e_b. A typical setting is $\delta_{max} = 15$ minutes for yeast expression data sets. The parameter δ denotes the observed time difference (minutes) between the middle of e_a and e_b, and $0 \leq \delta \leq \delta_{max}$. Edge pairs with time difference $> \delta_{max}$ are considered biologically meaningless and are simply ignored. Larger values of δ imply less similarity between the edges. The counts n_a and n_b denote the total number of quadrary edges in G_a and G_b, respectively.

   Observe that, like the correlation coefficient, the range of S_g is between -1 and 1, with 1 denoting the highest positive score, and -1 the highest negative score.

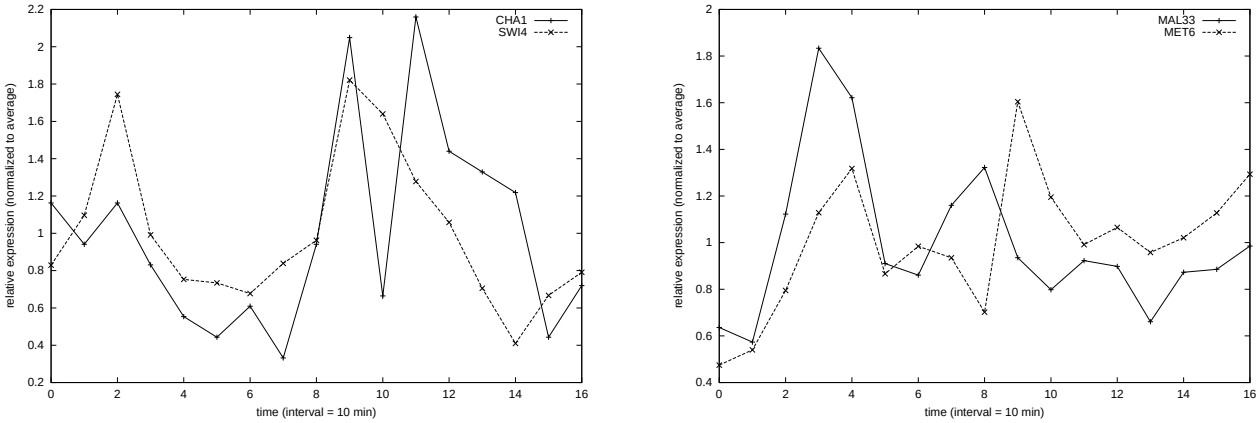

Figure 6: Interesting regulatory pairs detected by our edge function but not by correlation.

Alpha Data								CDC28 Data							
correlation coefficient				edge function				correlation coefficient				edge function			
	putative regs.				putative regs.				putative regs.				putative regs.		
thresh	total	good	bad	thresh	total	good	bad	thresh	total	good	bad	thresh	total	good	bad
> 0.85	107	5	0	> 0.6	96	5	0	> 0.85	289	2	2	> 0.6	146	1	0
> 0.8	192	5	2	> 0.5	223	5	0	> 0.8	628	5	4	> 0.5	398	3	0
> 0.7	703	5	7	> 0.4	557	7	6	> 0.7	1826	22	15	> 0.4	1236	11	3
> 0.6	1852	9	13	> 0.3	1581	11	15	> 0.6	3903	31	19	> 0.3	3401	19	20

Table 3: Putative regulatory relations for alpha and cdc28 data by threshold, with scoring of favorable and unfavorable pairs.

5.3 Results

Our results showed that this algorithm identifies certain interesting putative pairs missed by correlation coefficient methods, because we focus on local rather than global features.

We used the following methodology to evaluate the performance of our edge detector. A Ph.D level biologist (Zhi) manually compared all known regulatory relations in the alpha and cdc28 data sets, classifying them into sets which (a) clearly revealed a possible regulatory relation, (b) clearly *did not* reveal a regulatory relations or were inherently ambiguous. Only 27 of the 343 activations in the alpha and 63 of the 469 activations in the cdc28 data sets were classified in the positive group a, with the remainder classified in the negative group b.

The result of scoring all gene pairs from the 335 alpha genes and 366 cdc28 genes are Table 3. A perfect classifier / threshold pair would score only positive pairs above the threshold. Table 3 demonstrates that our proposed edge scoring function misclassifies no pair above the 0.5 level.

The established efficacy of our edge detection function makes it interesting to study other highly scoring pairs of genes, namely the 192 alpha pairs which correlated > 0.8 and 223 alpha pairs which scored > 0.5 using our edge function. For the cdc28 data set, 628 pairs correlated > 0.8 and 398 pairs scored > 0.5 using our edge function. There is some agreement between the two scoring measures for 88 alpha pairs and 127 cdc28 pairs occurred in both sets. Eyeball examination showed that all these common pairs carried a characteristic of strong and smooth oscillation pattern. However, we observed certain interesting pairs of genes using edge function correlated poorly. For example, the pairs in Figure 6, MAL33/MET6 correlated 0.32 and CHA1/SWI4 correlated 0.47. For MAL33/MET6, shifting the curve of MAL33 5 min, 10 min and 15 min to the right still brings the correlation coefficient score to only 0.41, 0.42, 0.27, respectively. In the case of CHA1/SWI4, the low expression of CHA1 at time point 10 is the main reason why correlation coefficient scored low. Our edge function, which decided this time point was probably experimental error, and ignored it, scored it well.

In conclusion, we believe our edge function provides an interesting method to analyze time course gene expression data. More results of our experiments are available at *http://www.cs.sunysb.edu/ ~skiena/ gene*. Figure 7 shows that the collection of highest scoring regulatory pairs in alpha are well distributed among many genes. While we do not propose such a diagram yields any semblance of the complete regulatory network, we do believe these pairs warrant further study.

As more accurate experimental data emerges, perhaps with error bars, such edge detection functions should become even more reliable and predictive.

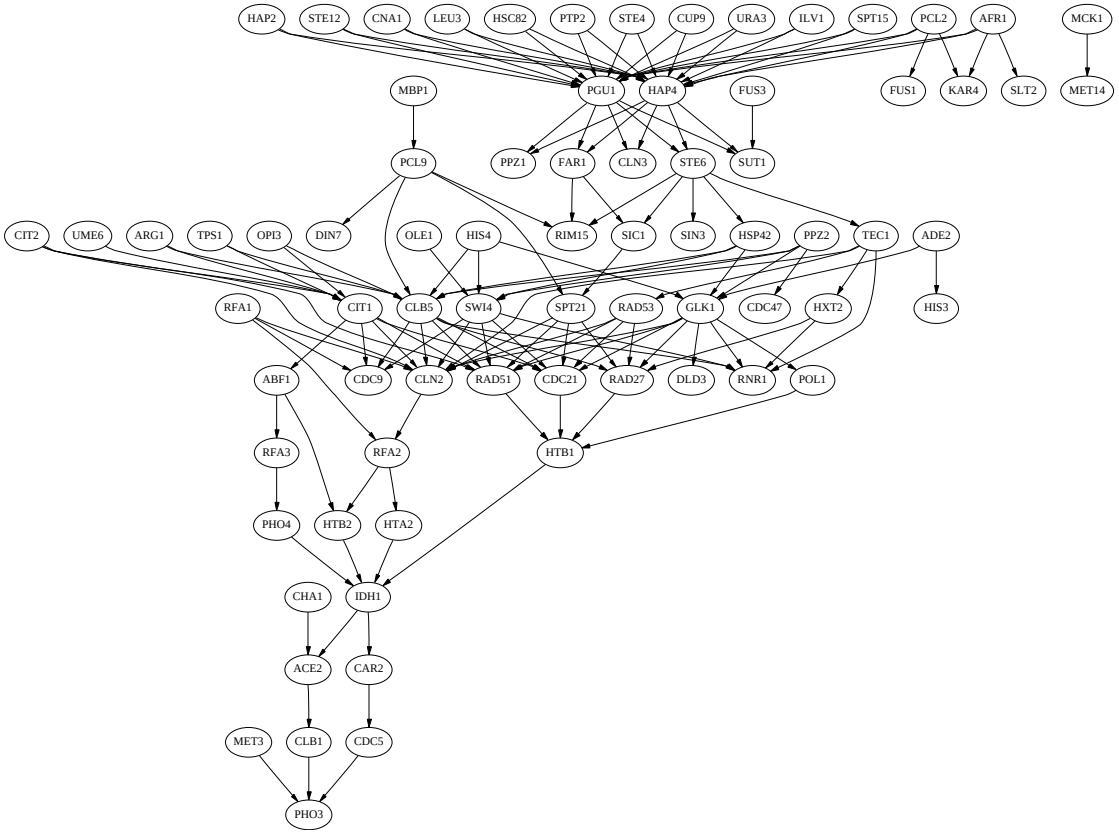

Figure 7: Network of the 140 activations in Alpha data set with score > 0.4

6. REFERENCES

[1] T. Chen, V. Filkov, and S. Skiena. Identifying gene regulatory networks from experimental data. In S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, pages 94–103. ACM Press, 1999.

[2] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. In *Pacific Symposium Biocomputing '99*, pages 29–40, 1999.

[3] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[4] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science*, 85:14863–14868, 1998.

[5] V. Filkov. Covering points on a circle with circular arcs. In *Proc. 12th Canadian Conf. Computational Geometry*, August 2000.

[6] A Friedman, M. Linial, I. Nachman, and D. Péer. Using bayesian networks to analyze expression data. In Ron Shamir, Satoru Miyano, Sorin Istrail, Pavel Pevzner, and Michael Waterman, editors, *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB-00)*, pages 127–135, N.Y., April 8–11 2000. ACM Press.

[7] N. Holter, M. Mitra, A. Maritan, M. Cieplak, J. Banavar, and N. Fedoroff. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Nat. Acad. Sci.*, 97:8409–8414, July 18, 2000.

[8] R. R. Klevecz and H. B. Dowse. Tuning in the transcriptome: Basins of attraction in the yeast cell cycle. *Cell Proliferation*, in Press, 2000.

[9] G. Snedecor and G. Cochran. *Statistical Methods.* Iowa State University Press, Ames, IA, 7th edition edition, 1980.

[10] 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 by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.