Analysis Techniques for Microarray Time-Series Data

Steven Skiena
Department of Computer Science
State University of New York
Stony Brook, NY 11794-4400
http://www.cs.sunysb.edu/~skiena

November 7, 2000

Gene Expression Analysis

With extensive DNA sequence data emerging, determining the structure of gene regulatory networks is the next great challenge in biology.

Time series microarray expression data of all 6601 yeast genes over multiple cell division cycles (Spellman/Cho) facilitates identifying regulatory elements.

Given a set of time series data, we seek to *suggest* possible inhibition/activation relations between the genes.

Our methodology: using signal processing techniques, reduce the time series data set to a graph of possible pairwise gene relationships.

Cho/Spellman Data Sets

Four sets of time series, each using a different cell synchronization method.

Not all of the 6600 orfs yielded data for each time point in each series.

Data	Period obs.	Period det.	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

Each series (except for elu) clearly ran for more than one cell cycle.

Previous Work

Several teams have attempted to extract gene regulatory data from the Cho/Spellman data:

- Eisen (PNAS '98) / Spellman (Mol. Bio. '98) using clustering / promoter analysis.
- Chen, Skiena, Filkov (RECOMB '99) using signal processing and combinatorial optimization.
- Friedman, et.al (RECOMB '00) using Bayesian networks.
- Differential equation modeling (CHC-99), wavelets (KD-2000), and singular value decomposition (HMMCBF-2000).

Not much has been said about the accuracy of predictions from these systems. . .

Outline of Talk

Describe our previous system (RECOMB '99) for proposing regulatory relations.

Assessment of the potential of inducing regulatory relations from Cho/Spellman data.

Improved edge detection algorithms for detecting regulatory relations.

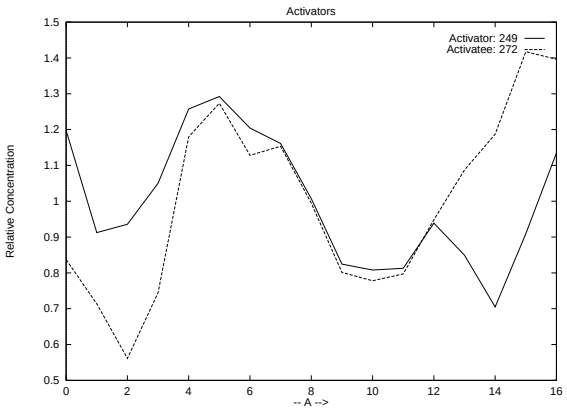
Periodicity and phase shift analysis for time-series integration.

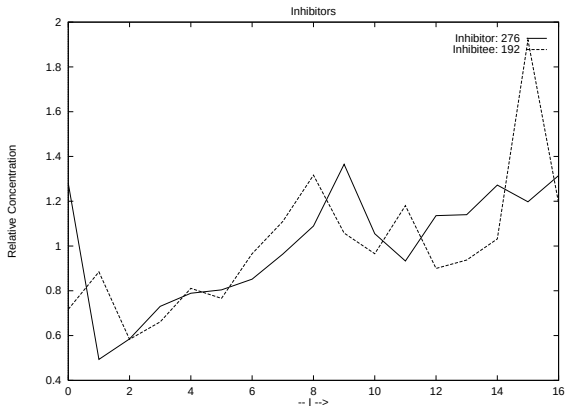
Comparing correlations of distinct length sequences

Correlation significance of small alphabet sequences.

Phases of Regulatory Candidate Identification

- Pre-Filtering unreliable expression data.
- Clustering
- Curve Smoothing.
- Calculating regulation scores.
- Optimizing regulation assignments.

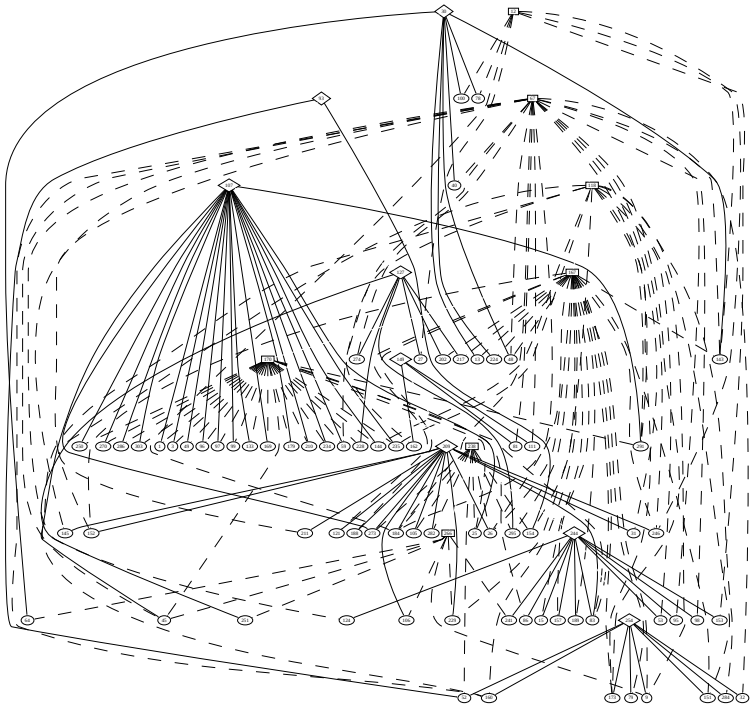




Candidate activator (L) / inhibitor (R) pairs.

Experimental Results

This network contains 7 proposed activators and 8 proposed inhibitors.



Prof. James Konopka, a yeast specialist at Stony Brook, observed several potentially interesting features in our network, including genes involved with cell division cycle, DNA replication, and amino acid synthesis.

Correlation and Similarity

The good things about using correlation as a similarity measure for gene expression data are (1) it is scale invariant,and (2) it seems to work well.

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

However, there are at least two potential problems:

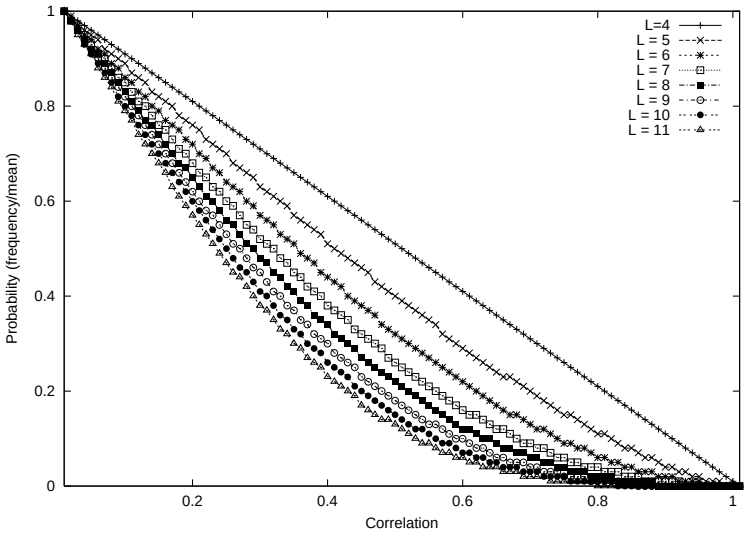
- Correlation is not a metric, and hence does not satisfy the triangle inequality.

- It is not meaningful to compare correlations of sequence pairs of different lengths.

Correcting for Sequences of Different Lengths

We stress that the Cho/Spellman data sets are *short* time series, which invalidates certain standard assumptions.

In particular, short sequences are far more likely to have chance high correlations than long sequences.

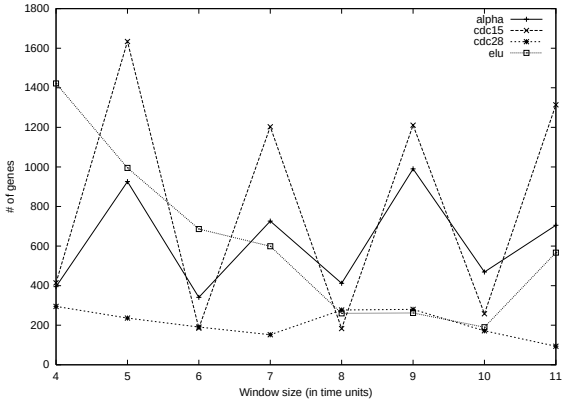


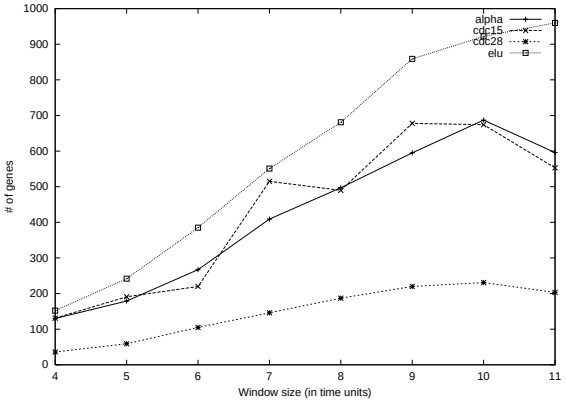
By plotting the cumulative distribution function of positive correlation coefficients, we could normalize significance by length.

Inferring Cycle Lengths

Inferring cell cycle lengths involves measuring the correlation between a prefix of a time-series and its suffix.

Selecting the shift period which maximized the highest correlation didn't work well (see left figure) because high correlations of short sequences were given too much weight.





After normalizing to account for sequence length, we obtain a peak at the proper place.

Inferring Cycle Offsets

Different synchronization methods leave cells in different states of the cell cycle.

Interleaving the different time-series requires determining the relative phase shift of each pair of experiments.

We seek the time shift which maximizes the number of orfs whose correlation across the series pair is maximized.

Data set 1	Data set 2	Shift
alpha	cdc28	$0-2$ samples
cdc28	cdc15	$1-2$ samples
cdc15	alpha	0 samples

Our computed shifts appear basically on target which known results about the length of different phases of the cell cycle.

Correlation of Small Alphabet Sequences

One approach to dealing with the high error rates associated with gene expression data is *bucketing* the observed values into a small number of bins.

In the limiting case, such data can be quantized to 0/1, i.e. *binary* sequences.

The *Hamming distance*, or number of bit mismatches, provides a natural distance metric between pairs of binary strings.

How well do the correlation coefficient and Hamming distance agree in scoring binary sequence similarity?

Analysis

Two n-bit binary sequences can be viewed as a single sequence of length n over the alphabet $\Delta = \{\binom{1}{1}, \binom{1}{0}, \binom{0}{1}, \binom{0}{0}\}$.

Let a, b, c, d denote, respectively, the counts of character from Δ in the pair (X, Y).

Clearly $b + c$ is the Hamming distance of the sequence.

Analysis of the correlation coefficient reduces to:

$$\text{corr}(X, Y) = \frac{a - (a+b)(a+c)/n}{\sqrt{((a+b) - (a+b)^2/n)((a+c) - (a+c)^2/n)}}$$

Exhaustive search over all a, b, c, and d is tractable for large n.

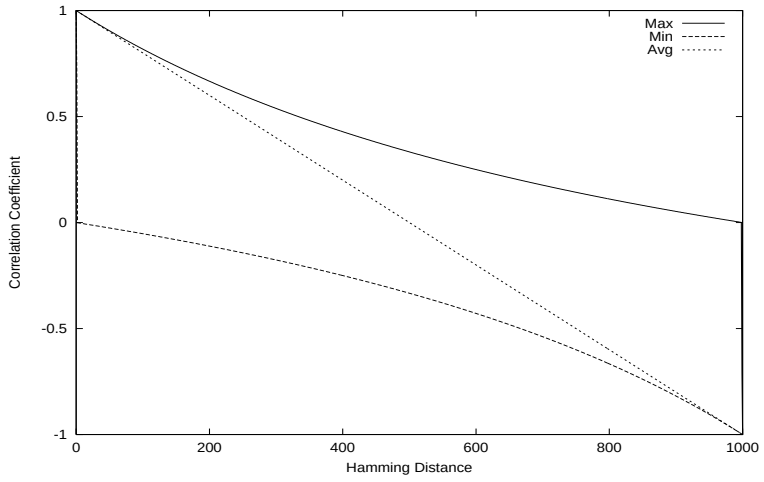
Average and Extremal Values

The potential gap between Hamming distance and correlation *can be* distressingly large:

Two sequences of length n which differ in only one position ($a = 1$, $b = 1$, $c = 0$, and $d = n - 2$) have correlation of

$$\sqrt{1/2 - 1/2(n-1)} \approx 1/\sqrt{2} \approx 0.7067$$

Two sequences of length n can differ in two positions ($a = 0$, $b = 1$, $c = 1$, and $d = n - 2$) and have a correlation of $-1/(n-1) \approx 0$.



Still, the average bounds fall exactly on the desired line.

Evaluating Known Regulatory Pairs

To analyze the potential for determining regulatory pairs from the Cho/Spellman data, we constructed a data base of all *known* regulations from the Yeast Protein Database YPD.

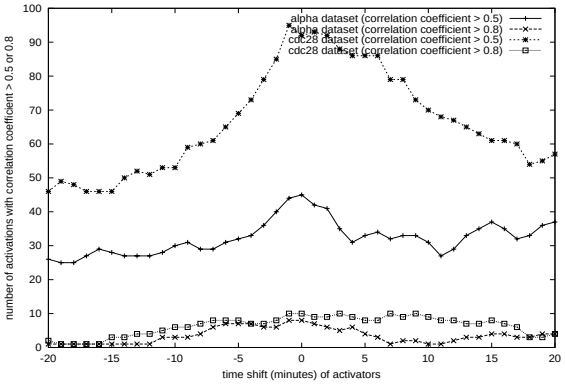
After mapping most of these names to the Spellman/Cho data:

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

Less than 20% of these known regulatory pairs scored a correlation above 0.5.

Time Shift Required?

This fraction did not increase when we relatively shifted the time series:



A Ph.D biologist (Zhi) examined all known pairs by eyeball and saw no common pattern to indicate possible transcription regulation in more than 20% of these cases.

Conclusion: It seems futile to hope to induce large networks from this data.

However, it is still reasonable to identify interesting pairs of expressed genes.

Improved Edge Detection

Based on our study of known regulatory pairs, we developed an improved edge detection function.

It seeks to eliminate narrow peaks and troughs as likely experimental error.

It ignores variation of 10% in gene expression level as likely experimental error in defining local minima and maxima.

It measures the time difference between peaks relative to the longest biologically plausible delay, about 15 minutes.

Edge Detection Algorithm

- *Primary edges* link neighboring local maxima and minima.
- *Secondary edges* line all primary edges whose height

$$height = \frac{high\_point\_expression - low\_point\_expression}{average\_expression\_of\_the\_gene}$$

  is greater than a threshold, typically 30%. 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.

Pairs of genes are *scored* solely based on their quadrary edges.

The similarity score S_g between G_a and G_b is given as:

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

where $d \in \{-1, 1\}$ denotes the agreement of the slopes of e_a and e_b.

Results

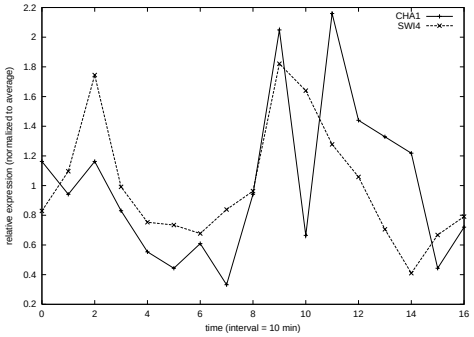
We scored how well both measures did at identifying clearly interesting / uninteresting series pairs, as judged by our biologist (Zhi).

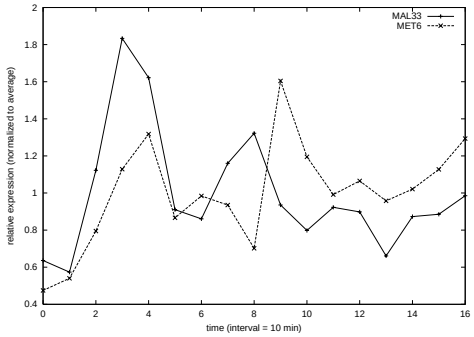
Alpha Data							
correlation coefficient				edge function			
thresh	total	good	bad	thresh	total	good	bad
> 0.85	107	5	0	> 0.6	96	5	0
> 0.8	192	5	2	> 0.5	223	5	0
> 0.7	703	5	7	> 0.4	557	7	6
> 0.6	1852	9	13	> 0.3	1581	11	15

CDC 28 Data							
correlation coefficient				edge function			
thresh	total	good	bad	thresh	total	good	bad
> 0.85	289	2	2	> 0.6	146	1	0
> 0.8	628	5	4	> 0.5	398	3	0
> 0.7	1826	22	15	> 0.4	1236	11	3
> 0.6	3903	31	19	> 0.3	3401	19	20

We are making fewer mistakes at high thresholds than the correlation coefficient.

Interesting Pairs

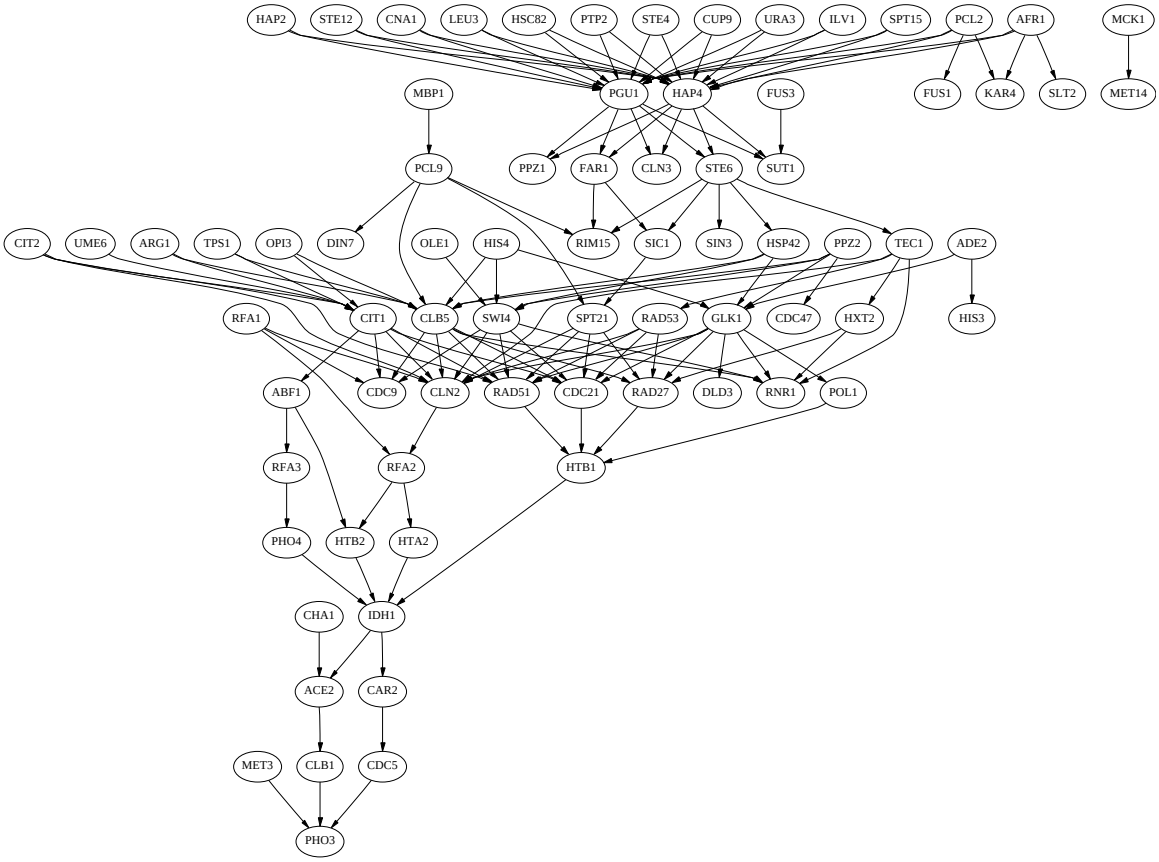




Interesting regulatory pairs detected by our edge function but not by correlation.

See http://www.cs.sunysb.edu/∼skiena/gene

Network of 140 activations in Alpha with score > 0.4



Thanks

This has included joint work with:

- Ting Chen, Dept. of Mathematics, Univ. of Southern California.

- Vladimir Filkov, Dept. of Computer Science, SUNY Stony Brook.

- Jizu Zhi, Center for Biotechnology, SUNY Stony Brook.

Funding from NSF, ONR, SB Center for Biotechnology, and industrial sources.