

# Constructing Networks with Correlation Maximization Methods

Joseph C. Mellor<sup>1</sup>  
mellor@bu.edu

Jie Wu<sup>2</sup>  
jiewu@bu.edu

Charles DeLisi<sup>2</sup>  
delisi@bu.edu

<sup>1</sup> Program in Bioinformatics, Boston University

<sup>2</sup> Department of Biomedical Engineering, Boston University

## Abstract

Problems of inference in systems biology are ideally reduced to formulations which can efficiently represent the features of interest. In the case of predicting gene regulation and pathway networks, an important feature which describes connected genes and proteins is the relationship between active and inactive forms, i.e. between the “on” and “off” states of the components. While not optimal at the limits of resolution, these logical relationships between discrete states can often yield good approximations of the behavior in larger complex systems, where exact representation of measurement relationships may be intractable. We explore techniques for extracting binary state variables from measurement of gene expression, and go on to describe robust measures for statistical significance and information that can be applied to many such types of data. We show how statistical strength and information are equivalent criteria in limiting cases, and demonstrate the application of these measures to simple systems of gene regulation.

**Keywords:** regulatory networks, gene networks

## 1 Introduction

The rapid increase in DNA sequencing has enormously amplified our ability to assay the expression levels of genes by massively parallel microarray assays, thus paving the way for systems biology. It has also forced the development of non-homology based computational methods for inferring the function of unannotated genes. These new methods, unlike those based solely on sequence homology, are contextual; i.e. they place genes within a network of other genes having correlated functions, and they contribute directly to the delineation of the cellular circuitry that controls, amplifies and otherwise modulates changes in phenotypic expression.

Broadly speaking there are two classes of methods, those based on evolutionary principles, and those based on high throughput biochemical assays. The former include correlations in the pattern of occurrences of a pair of genes across a set of phylogenetically distinct genomes; the tendency to preserve proximity across a set of genomes [27], and the tendency of functionally related distinct genes to fuse as genomes evolve [28]. Among the latter are DNA microarrays which measure the expression of thousands of genes over hundreds of conditions, and high throughput assays for identifying protein-protein [11, 17, 26] and protein-DNA [13, 16] interactions.

We find that the problem of drawing systems-level inferences from phylogenetic profiles and expression array analysis can be formulated in a strikingly similar manner. Here we outline a mathematical approach to the analysis of this data, and briefly illustrate how it can be applied to pathway elucidation. The abstract similarity is that the observations to be analyzed can be represented by pairs of binary strings. A phylogenetic profile of a gene across a set of  $N$  genomes is a binary vector, with component 0 when the gene is absent and 1 when it is present. Two genes are assumed to be functionally related when their profiles are correlated above some prespecified threshold.

Similarly, the expression of a gene across a set of  $N$  experiments can, to first approximation, be represented as a string of zeroes and ones, a zero being when the gene is down relative to its average in those experiments and 1 if it is up. How strongly down- or up-regulated a gene needs to be to be assigned a 0 or 1 is itself an important question which will be addressed below. Here too, if the profiles are such that the similarity between them could only have arisen by chance with a small probability less than some threshold, they are assumed to be functionally related. In general, the question as such is structured in similar ways as previous studies of probabilistic or deterministic gene networks [2, 4, 5, 6, 8, 10, 18, 22, 24, 25]. The information-based method we employ possesses the qualitative clarity of Boolean models while also being more robust to noise and sampling effects. Connections in these networks are representative of significant probabilistic correlation between binary states of two components, and imply hypothetical functional relationships. We demonstrate our approach as it is applied to a small system of regulation in the yeast *Saccharomyces cerevisiae*.

## 2 Methods and Results

### 2.1 Information and Correlation in Binary String Representations

The relation between any two random (i.e. no within string correlations) binary strings,  $X$  and  $Y$ , of the same length,  $N$ , can be characterized by four variables: one for the length, two to specify the number of ones in each string ( $x$  and  $y$ ), and one to specify a relation between the two strings, for example, the number of positions ( $z$ ) at which each string is equal to one (1). We are interested in the probability that the observed similarity between two random binary strings is a chance event, and the use of this probability measure in determining new relationships between biological variables.

The correlation problem is stated generally: Given a string  $N$  long with  $y$  ones, what is the probability that a string  $X$  will be picked at random from strings  $N$  long, such that it has  $x$  ones and mismatches the profile of  $Y$  at  $(z_0 + z_1)$  positions,  $z_0$  of these being mismatched zeroes?

Mismatched zeroes in  $Y$  can be opposite any of  $x$  ones in  $X$ . The first zero can be placed in  $x$  ways, the second in  $x - 1$  ways ... the last in  $x - z_1 - 1$  ways. Overall, they can be placed in  $x(x - 1)(x - 2) \dots (x - z_1 - 1)$  ways. These ways remain identical even when the  $z_1$  zeroes are permuted; therefore, the total number of ways to place all mismatched zeroes in  $Y$  is

$$\binom{x}{z_1}$$

The number of ways to distribute the remaining  $(N - y - z_1)$  matched zeroes over  $N - x$  positions is

$$\binom{N - x}{N - y - z_1}$$

Now only the ones remain in  $Y$  to be distributed, and there are  $y$  of them. There are also  $y$  remaining positions in  $X$  in which to distribute them. But since the number of ones and the number of positions to which they can be assigned is the same (all ones necessarily being identical), there is only one way to place the  $y$  ones. There is thus no additional probability term necessary for these. The total number of distributions meeting the combined constraints is the product of the above two expressions. The required probability is this product divided by the number of ways of distributing the sequence  $Y$  in an unconstrained manner; *viz*,

$$\binom{N}{y}$$

Next, we define

$$\begin{aligned}
T &\equiv x!y!(N-x)!(N-y)! \\
&\text{and} \\
B &\equiv (x-z_1)!z_1!(y+z_1-x)!(N-y-z_1)!N! = (x-z_0)!z_0!(N-y-z_1)!N! \\
P &= T/B
\end{aligned} \tag{1a}$$

$P$  can also be written in terms of  $z$ , which could be, for example, the number of experiments in which both genes are over expressed, or when two genes are present in the same genome,

$$P(z|N, x, y) = \frac{x!y!(N-x)!(N-y)!}{(z!)(N!)(N+z-x-y)!(x-z)!(y-z)!} \tag{1b}$$

We assume that two genes will be functionally related if the probability evaluated by eq 1 is below some small value  $p^*$ , usually taken to be  $10^{-4}$ .

Mutual information,  $I(X, Y)$  also provides a measure of correlation between binary strings. In fact there is an exact relation between mutual information and the asymptotic form of eq 1. By definition, mutual information is the reduction in uncertainty about a string given information about a related string:

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

Here  $H(X)$  and  $H(X|Y)$  are the entropy and conditional entropy of  $X$ . The joint probabilities,  $p(i, j)$ , are easily written in terms of  $x/N$ ,  $y/N$ ,  $z/N$ . In particular

$$\begin{aligned}
p(1, 1) &= \frac{x - z_0 - z_1}{N} = p_x - p_{z_0} \\
p(0, 0) &= \frac{N - y - z_1}{N} = 1 - p_y - p_{z_0} = 1 - p_x - p_{z_1} \\
p(1, 0) &= \frac{z_0}{N} = p_{z_0}; \quad p(0, 1) = \frac{z_1}{N} = p_{z_1}
\end{aligned}$$

where

$$p_x \equiv p(X = 1) = \frac{x}{N}; \quad p(X = 0) \equiv 1 - p_x = 1 - \frac{x}{N}, \quad p_y = \frac{y}{N}, \quad p_{z_0} = \frac{z_0}{N}, \quad p_{z_1} = \frac{p_{z_1}}{N},$$

For completion we further note

$$\begin{aligned}
p(Y = 1|X = 1) &= \frac{y - z_0}{x} = \frac{p_y - p_{z_0}}{p_x}; \quad p(0|0) = \frac{N - x - z_0}{N - x} = \frac{1 - p_x - p_{z_0}}{1 - p_x} \\
p(1|0) &= \frac{z_0}{N - x} = \frac{p_{z_0}}{1 - p_x}; \quad p(0|1) = \frac{z_1}{x} = \frac{p_{z_1}}{p_x}
\end{aligned}$$

The limiting form of eq. 1 is obtained using Sterling's approximation:

$$w! \approx \sqrt{2\pi w} w^w e^{-w}$$

If all the variables  $N$ ,  $x$ ,  $y$ ,  $z$ , as well as differences entering the expression for  $P$ , are sufficiently large ( $> 10$ ) this approximation can be applied with essentially no error. When applied to eq 1, the exponential terms will cancel. In addition, for the limit of large  $N$ , terms involving the reciprocal square root of  $N$  will go to zero, and can be ignored. For the remaining terms, we then have

$$\begin{aligned}
T &\equiv x^x y^y (N-x)^{(N-x)} (N-y)^{N-y} \\
&= x^x y^y N^{2N} N^{-x} N^{-y} (1-p_x)^{N-x} (1-p_y)^{N-y} \\
&= p_x^{N p_x} p_y^{N p_y} N^{2N} (1-p_x)^{N(1-p_x)} (1-p_y)^{N(1-p_y)}
\end{aligned}$$

and

$$B \equiv N^{2N} p_{z_1}^{N p_{z_1}} p_{z_0}^{N p_{z_0}} (p_x - p_{z_1})^{N(p_x - p_{z_1})} (1 - p_y - p_{z_1})^{N(1 - p_y - p_{z_1})}$$

Forming  $P$ , taking the log, and dividing by  $N$ , we obtain

$$\begin{aligned} - \lim_{n \rightarrow \infty} \frac{1}{N} \log \frac{T}{B} &= (-p_x) \log p_x - (1 - p_x) \log(1 - p_x) \\ &= (-p_x) \log p_x - (1 - p_x) \log(1 - p_x) + p_{z_1} \log p_{z_1} + p_{z_0} \log p_{z_0} \\ &+ (p_x - p_{z_1}) \log(p_x - p_{z_1}) + (1 - p_y - p_{z_1}) \log(1 - p_y - p_{z_1}) \\ &= H(X) + H(Y) - H(X, Y) = I(X, Y) \end{aligned}$$

Therefore the limiting form of  $P$  is

$$P = 2^{-NI(X, Y)}$$

When  $I = 1$ ,  $P = 2^{-N}$  as expected; i.e. the probability that two random strings  $N$  long will match exactly is  $2^{-N}$ , the probability being  $\frac{1}{2}$  at each position. When  $I = 0$ , there is no relation, which means that the similarity is certainly the result of chance ( $P = 1$ ).

We apply this formalism to the relationship between two nonhomologous genes in *Saccharomyces cerevisiae*, using their phylogenetic profiles across 40 microbial species (Figure 1). The yeast gene *FAS1* has verified function, encoding a subunit of the fatty acid synthetase. The other gene, *CEM1*, encodes a yeast gene with unknown function, but with homology to known  $\beta$ -keto-synthases. The probability  $P$  that the profiles between the two genes is what is observed in Figure 1 is approximately  $5 * 10^{-12}$ .

	Species																																										
	Alu	Hbs	Mac	Mth	Mja	Tac	Two	Pto	Pab	Pyu	Sso	Ape	Scs	Spo	Ecu	Aab	Tma	Syn	Nos	Nme	NmA	Rso	Hpy	Jhp	Qje	Alu	Sme	Bme	Mlo	Cor	Rpr	Rco	Cr	Cpn	Tpa	Bbu	Lur	Mpu	Mpn	Mge			
<i>CEM1</i> (YER061C)	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	
<i>FAS1</i> (YKL182W)	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0

Figure 1: Binary correlation using phylogenetic profiles of two genes in *Saccharomyces cerevisiae*. Two yeast genes (*CEM1*, *FAS1*) have an inferred functional relationship in acyl-carbon synthesis based on the profiles of the genes in 40 species. A one (1) denotes the presence of an ortholog of the yeast gene in the genome of the corresponding species.

## 2.2 Uncertainty of Regulation between Gene Expression Profiles

The same method described above can be applied to the expression states of genes in a collection of transcription profiles. Only instead of judging the “presence” or “absence” of a gene in the phylogenetic context, we look at the relative increase or decrease of the gene’s transcription level. Consider the summary of expression profiles of the yeast genes *STE12*, *MCM1* and *STE2*, shown in Table 1. *Ste12p* and *Mcm1p* are transcription factors known to bind upstream of *STE2*, and the proteins have been shown to interact physically [15, 21]. In this example, a change in expression for a gene is significant when it deviates by more than one standard deviation from mean, taking into account measurement of the gene in all experiments. For data, we collected microarray experiments in yeast representing 550 diverse conditions [7, 9, 23]. We then normalized this data with standard techniques [20] such that each gene has zero mean and unit variance across all conditions. A measurement in this normalized set then represents the number of standard deviations that the gene falls relative to its mean. We assign a state of zero to a gene when it is significantly under-expressed, and a state of 1 when it is over-expressed. The null model assumes equal probabilities for the gene being under- and over-expressed. Given real data, we employ our method to determine whether for certain combinations of input states this assumption doesn’t hold.



When *STE12* and *MCM1* are down ( $STE12, MCM1 = (0, 0)$ ), *STE2* has a split output. *FAR1* correlates with the two expression possibilities of that output, being 1 in 7/8 instances when *STE2* is 1, and being 0 in 8 of the 11 instances when *STE2* is 0. As a result, the average entropy associated with the output of the  $(STE12, MCM1) = (0, 0)$  state is reduced by a factor of 2 when *FAR1* is also down. While we can say that *FAR1* is a potential additional regulator of *STE2*, the mechanism linking *FAR1* to the regulation scheme is unclear at best. It is possible that *FAR1* is upstream or downstream of the signal causing regulation of *STE2*, or that the situation involves feedback. As such, the method only identifies potentially indirect relationships between regulators and targets. Our next section shows how improvements can be made that give better resolution of the general causality problem. Figure 2 shows a subnetwork of genes involved in the regulating response to mating signal and the cell cycle in yeast.

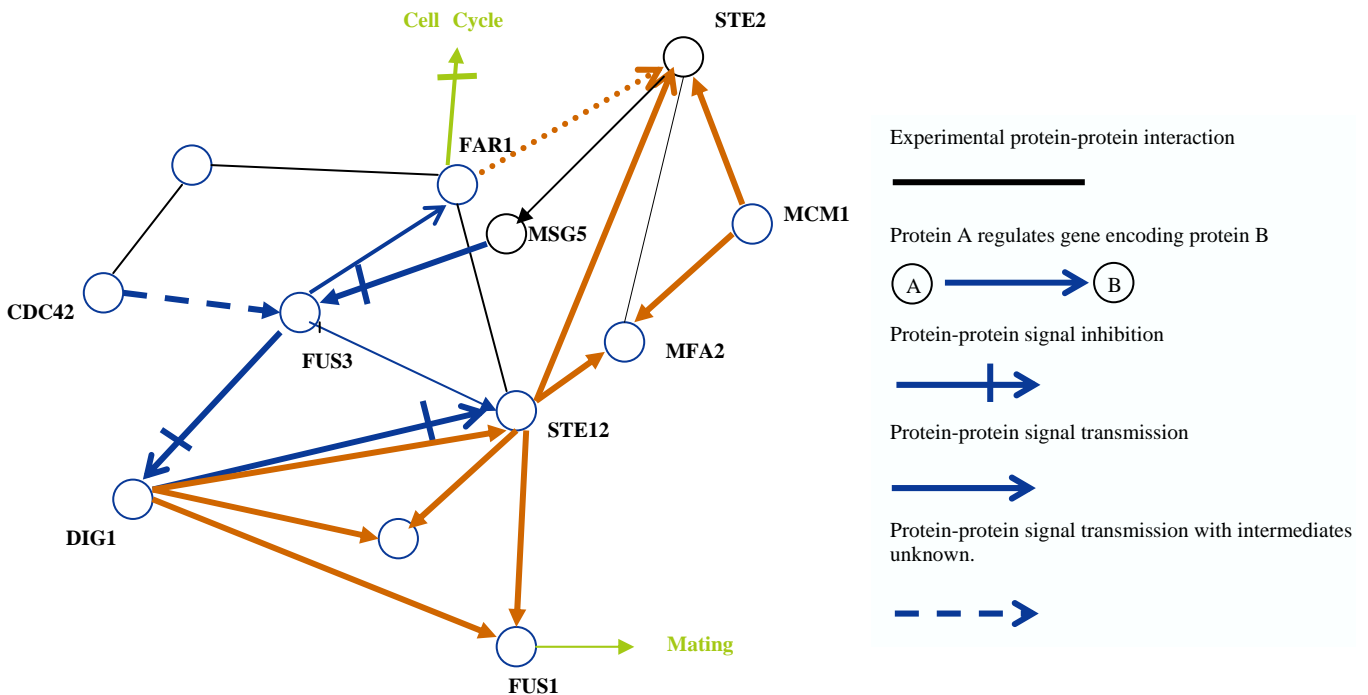


Figure 2: A yeast subnetwork of genes involved in mating response and cell cycle control. Combining above results with KEGG [12] pathway information shows that *FAR1* is potentially involved in regulating mating-response genes such as *STE2*.

### 2.3 Biological Changes in Gene Expression: Avoiding Arbitrary Thresholds

One of the main difficulties with microarray and other high throughput data is low specificity. To guard against this problem, the requirement for classifying a gene as significantly changed is usually very stringent. This reduces false positives, but at the expense of potential loss of considerable information. Here we outline the concept behind an objective and promising criterion that extends our binary correlation methods to extract maximum information from continuous measurements of gene expression data. Detailed tests and biological predictions produced by the method are presented elsewhere (*JCM and CD, submitted*).

We premise that gene regulation can be effectively described both (a) as a discrete process in which biological alteration of expression is partitioned into two or more states of over- or under-expression and (b) as a continuous process in which the biological threshold for ascribing a state may change as a function of context. The idea then is to allow expression thresholds to adjust to fit context, choosing

those which maximize the mutual information between a target gene and its regulators(s). This allows us to draw conclusions related to (i) whether the regulator affects the target (ii) if so, which state of the regulator results in which state of the target and (iii) the experimental conditions under which relations (ii) hold. We delineate the proof of principle in this approach for the case of two expression profiles - one for a regulator and one for a target gene.

Define values  $E_i^X$  and  $E_i^Y$  as expression levels of genes  $X$  (the regulator) and  $Y$  (the target) in the  $i^{th}$  experiment. A partitioning function  $\delta(E_i)$  maps these expression levels to states “up” and “down”, subject to a set of threshold parameters. The expression thresholds for the regulator and target ( $E^X$ ,  $E_{down}^Y$ ,  $E_{up}^Y$ ) are the parameters that vary during the optimization procedure. The instances of states of  $X$  and  $Y$  in the  $i^{th}$  experiment are defined by

$$\delta(E_i^X) = \begin{cases} 0, & |E_i^X| < |E^X| \\ 1, & |E_i^X| \geq |E^X| \end{cases} \quad \delta(E_i^Y) = \begin{cases} -1, & E_i^Y \leq E_{down}^Y \leq 0 \\ 1, & E_i^Y \geq E_{up}^Y \geq 0 \end{cases}$$

Summing values of  $\delta(E_i)$  over the set of experiments will enumerate those times each gene is in different states. We wish to consider from  $n_{exp}$  total experiments the number of events when different cases of genes  $X$  and  $Y$  are observed independently and jointly. For example, the numbers of experiments in which the target gene  $Y$  is up and down, respectively, are given by

$$N(E_{up}^Y) = \frac{1}{2} \sum_{i=1}^{n_{exp}} (\delta(E_i^Y) + 1) \quad \text{and} \quad N(E_{down}^Y) = \frac{1}{2} \sum_{i=1}^{n_{exp}} (\delta(E_i^Y) - 1)$$

And, the union (sum) of these observations is

$$N(E_{down}^Y, E_{up}^Y) = \sum_{i=1}^{n_{exp}} |\delta(E_i^Y)|$$

Also, for the regulator gene  $X$ , the number of experiments where  $X$  exceeds the threshold  $E^X$  is given by

$$N(E^X) = \sum_{i=1}^{n_{exp}} \delta(E_i^X)$$

The number of joint observations – that is, experiments where both target  $Y$  and regulator  $X$  exceed all thresholds (either up or down) is

$$N(E^X, E_{down}^Y, E_{up}^Y) = \sum_{i=1}^{n_{exp}} |\delta(E_i^X)| \cdot |\delta(E_i^Y)|$$

Here we note that across a set of experiments given by  $N(E^X, E_{down}^Y, E_{up}^Y)$ , the target gene can variably be up ( $E_i^Y \geq E_{up}^Y$ ) or down ( $E_i^Y \leq E_{down}^Y$ ), but across this same set the regulator is exclusively either up ( $E^X > 0$ ) or down ( $E^X < 0$ ), depending on the choice of model. The choice, i.e., the *a priori* regulator state, is itself a fixed feature of the relationship, and we are interested primarily in how that state correlates with a change in the target state.

The goal is to minimize a target function, call it  $K(X, Y, P)$ , which depends on the thresholds through the above eqs for  $X$ ,  $Y$  and the function  $P$ . We denote  $P$  as the significance criteria, which we fix at an acceptable level in order to optimize the target function  $K$ . For two binary strings, a suitable significance function is

$$P_{sig} = \begin{cases} P(p|q) = \sum_{k=0}^{n_{down}} (n_{total} C k) (q_{down})^k (q_{up})^{n_{total}-k}, & p_{up} < q_{up} \\ P(p|q) = \sum_{k=0}^{n_{up}} (n_{total} C k) (q_{up})^k (q_{down})^{n_{total}-k}, & p_{down} < q_{down} \end{cases}$$

where

$$\begin{aligned} n_{down} &= N(E^X, E_{down}^Y) \\ n_{up} &= N(E^X, E_{up}^Y) \\ n_{total} &= N(E^X, E_{down}^Y, E_{up}^Y) \end{aligned}$$

An appropriate choice of target functions is the Kullback-Leibler entropy between posterior and prior states of the target gene:

$$K(p|q) = p_{down} \lg \left( \frac{p_{down}}{q_{down}} \right) + p_{up} \lg \left( \frac{p_{up}}{q_{up}} \right)$$

where

$$\begin{aligned} p_{down} &= \frac{N(E^X, E_{down}^Y)}{N(E^X, E_{down}^Y, E_{up}^Y)}, & p_{up} &= \frac{N(E^X, E_{up}^Y)}{N(E^X, E_{down}^Y, E_{up}^Y)} \\ q_{down} &= \frac{N(E_{down}^Y)}{N(E_{down}^Y, E_{up}^Y)}, & q_{up} &= \frac{N(E_{up}^Y)}{N(E_{down}^Y, E_{up}^Y)} \end{aligned}$$

The maximization of this function optimizes the mutual information at large  $N$ , by incrementally adjusting the threshold parameters. This expresses the determinacy of regulation of gene  $y$  at a defined threshold  $E^Y$ , given information on the transcription state of gene  $x$  exceeding some expression threshold  $E^X$ .

Now we elaborate the earlier example of the gene *STE2* being regulated by *STE12* and *MCM1*. For gene-specific expression thresholds  $E^{STE2}$ ,  $E^{STE12}$  and  $qE^{MCM1}$ , we apply the optimization criteria outlined above, and the results of this maximization using the conditional entropy are gene-specific expression cutoffs are shown in Figure 3. We are clearly able to describe the regulation of *STE2* just as well as with strict cutoffs shown in the earlier example. The reason for this is that the partitioning of the expression of *STE2* into different states converges to an optimum which derives from the biological context - in this case, the states of its input regulators.

By not imposing an ‘‘arbitrary’’ threshold for over- or under-expression, we are able to find the location of the maximum signal in the data set. Here, the optimum thresholds  $E^{STE12} = 0$ ,  $E^{MCM1} = 0$ ,  $E^{STE2} = -1.6$  and  ${}^1E^Y = 0.7$  give the maximum regulation effect in *STE2*, in which it is down-regulated with *STE12* up and *MCM1* down. The significance (calculated as the cumulative binomial probability above) corresponding to this configuration is highly significant, at  $3 \times 10^{-6}$ . The inset of Figure 3 shows how the expression-state distribution of *STE2* changes markedly when observed with expression of its two regulators.

### 3 Discussion

The use of binary correlation methods shows significant promise in extracting functional information from otherwise noisy biological data sources and problems that aren’t immediately transparent to standard statistical techniques. Our method may be particularly useful for microarray data, where the ratio of signal to noise is difficult to quantify, yet where meaningful associations can occur even at relatively low signal level. By using an information-based approach, we are able to effectively determine the latent uncertainty of simple mechanisms of regulation by comparing the expression profiles of paired genes and their regulators, and also quantify the amount of information gained by the association of yet other genes with this pair. Another variation on this technique partitions the expression of genes into optimized windows of over- and under-expression that correspond to, depending on exact choice of criteria, the significant or informative modes of regulation. The relationships between genes derived

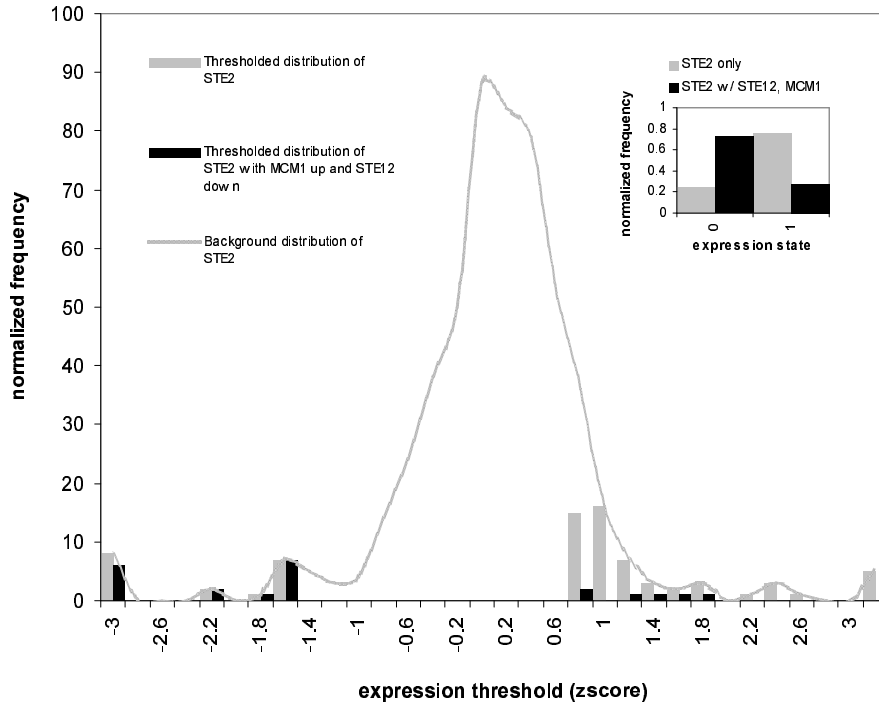


Figure 3: Expression-threshold optimization finding maximum correlation between binary states of two regulators ( $MCM1$ ,  $STE12$ ) and a target gene ( $STE12$ ).

in this method bear some similarity to the class of representations known as Boolean models [1, 14]. The logic of regulation itself and combinatorial effects, has been discussed in recent literature as a Boolean process [3, 29], and we hope our representation of the problem may be useful for exploring these complex features of eukaryotic transcription. We have yet to explore whether the method will have the predictive capabilities of robust Boolean models; we have chosen thus far to focus only on the statistical features and formalism of the procedure. We envision the general applicability of the methods we outline to the analysis of gene regulation from microarray experiments, and the discovery of potentially important features in mechanisms controlling transcription and signal transduction.

## References

- [1] Akutsu, T., Miyano, S., and Kuhara, S., Algorithms for inferring qualitative models of biological networks, *Pac. Symp. Biocomput.*, 293–304, 2000.
- [2] Akutsu, T., Miyano, S., and Kuhara, S., Inferring qualitative relations in genetic networks and metabolic pathways, *Bioinformatics*, 16(8):727–734, 2000.
- [3] Buchler, N.E., Gerland, U., and Hwa, T., On schemes of combinatorial transcription logic, *Proc. Natl. Acad. Sci. USA*, 100(9):5136–5141, 2003.
- [4] Chen, T., He, H.L., and Church, G.M., Modeling gene expression with differential equations, *Pac. Symp. Biocomput.*, 29–40, 1999.
- [5] Friedman, N., Linial, M., Nachman, I., and Pe’er, D., Using Bayesian networks to analyze expression data, *J. Comput. Biol.*, 7(3-4):601–620, 2000.

- [6] Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J., Inferring genetic networks and identifying compound mode of action via expression profiling, *Science*, 301(5629):102–105, 2003.
- [7] Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O., Genomic expression programs in the response of yeast cells to environmental changes, *Mol. Biol. Cell.*, 11(12):4241–4257, 2000.
- [8] Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A., Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, *Pac. Symp. Biocomput.*, 422–433, 2001.
- [9] Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburttty, K., Simon, J., Bard, M., and Friend, S.H., Functional discovery via a compendium of expression profiles, *Cell*, 102(1):109–126, 2000.
- [10] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F., Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics*, 18 Suppl 1:S233–s240, 2002.
- [11] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. USA*, 98(8):4569–4574, 2001.
- [12] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, 32 Database issue:D277–D280, 2004.
- [13] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A., Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 298(5594):799–804, 2002.
- [14] Liang, S., Fuhrman, S., and Somogyi, R., Reveal, a general reverse engineering algorithm for inference of genetic network architectures, *Pac. Symp. Biocomput.*, 18–29, 1998.
- [15] Madhani, H.D. and Fink, G.R., Combinatorial control required for the specificity of yeast MAPK signaling, *Science*, 275(5304):1314–1317, 1997.
- [16] Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E., TRANSFAC: Transcriptional regulation, from patterns to profiles, *Nucleic Acids Res.*, 31(1):374–378, 2003.
- [17] Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B., MIPS: A database for genomes and protein sequences, *Nucleic Acids Res.*, 30(1):31–34, 2002.
- [18] Pe’er, D., Regev, A., Elidan, G., and Friedman, N., Inferring subnetworks from perturbed expression profiles, *Bioinformatics*, 17 Suppl 1:S215–224, 2001.
- [19] Peter, M., and Herskowitz, I., Direct inhibition of the yeast cyclin-dependent kinase Cdc28-Cln by Far1, *Science*, 265(5176):1228–1231, 1994.
- [20] Quackenbush, J., Microarray data normalization and transformation, *Nat. Genet.*, 32 Suppl:496–501, 2002.

- [21] Roberts, R.L. and Fink, G.R., Elements of a single MAP kinase cascade in *Saccharomyces cerevisiae* mediate two developmental programs in the same cell type: Mating and invasive growth, *Genes. Dev.*, 8(24):2974–2985, 1994.
- [22] Segal, E., Taskar, B., Gasch, A., Friedman, N., and Koller, D., Rich probabilistic models for gene expression, *Bioinformatics*, 17 Suppl 1:S243–S252, 2001.
- [23] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell*, 9(12):3273–3297, 1998.
- [24] Steffen, M., Petti, A., Aach, J., D’Haeseleer, P., and Church, G., Automated modelling of signal transduction networks, *BMC Bioinformatics*, 3(1):34, 2002.
- [25] Tegner, J., Yeung, M.K., Hastay, J., and Collins, J.J., Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling, *Proc. Natl. Acad. Sci. USA*, 100(10):5944–5949, 2003.
- [26] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, 403(6770):623–627, 2000.
- [27] Wu, J., Kasif, S., and DeLisi, C., Identification of functional links between genes using phylogenetic profiles, *Bioinformatics*, 19(12):1524–1530, 2003.
- [28] Yanai, I., Derti, A., and DeLisi, C., Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes, *Proc. Natl. Acad. Sci. USA*, 98(14):7940–7945, 2001.
- [29] Yuh, C.H., Bolouri, H., and Davidson, E.H., Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene, *Science*, 279(5358):1896–1902, 1998.