

A Probabilistic Approach to Study Yeast's Gene Regulatory Network

Pinto F.R

Centro de Química e Bioquímica, Departamento de
Química e Bioquímica, Faculdade de Ciências da Universidade
de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal

Corresponding author: Pinto F.R, Centro de Química e Bioquímica,
Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade de Lisboa,
Campo Grande, 1749-016 Lisboa, Portugal, E-mail: frpinto@fc.ul.pt; Tel: +351 217500891; Fax: +351 217500088

Received December 14, 2008; Accepted February 25, 2009; Published February 27, 2009

Citation: Pinto F.R (2009) A Probabilistic Approach to Study Yeast's Gene Regulatory Network. J Comput Sci Syst Biol 2: 043-050.

Copyright: © 2009 Pinto F.R. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Using only the transcription network structure information, a probabilistic model was developed that computes the probabilities with which a pair of genes responds simultaneously (*SR*) or differentially (*DR*) to a random network perturbation. Study of yeast's transcription regulatory network in association with gene expression profiles shows that *SR* and *DR* probabilities are significantly associated with the distribution of strong co-expression. It is 100 fold more probable to observe co-expression when $P(SR) \approx 0.5$ for a random perturbation of 3 transcription factors (TFs), allowing for perturbation spread until a depth of 3 connections in the regulatory network. The model also predicts that positive co-expression enhancement is related with the proportion of common TFs (number of TFs that regulate both genes in a pair divided by the total number of TFs that regulate at least one gene in the pair), and not to the absolute number. The relationship between the model derived probabilities and other graph-theoretic measures used to analyse biological networks is discussed.

Key words: gene regulatory networks; network perturbation; response probability; co-regulation; co-expression; and graph-based measures

Abbreviations

GTOM – generalized topological overlap measure

$P(DR)$ – probability of differential response to network perturbation

$P(NR)$ – probability of neutral response to network perturbation

$P(SR)$ – probability of simultaneous response to network perturbation

TF – transcription factor

TFS – transcription factor similarity

TOM – topological overlap measure

Introduction

Systems biology has recently re-emerged (Westerhoff and Palsson, 2004), after a long quiescent period since its first steps (Bertalanfy, 1928), due to the acknowledgement of the components' interactions importance for an enhanced understanding of living organisms. Analysis of these interactions can be achieved through the identification of biological networks. Typically, they are characterized by their connectivity distributions (Barabasi and Oltvai, 2004) or by the enrichment in small network motifs, as compared with randomly connected networks (Milo, et al., 2002). The present work studies transcription regulatory networks, in particular for the yeast *Saccharomyces cerevisiae*. The most abundant information available about these networks

corresponds to a topological model or wiring diagram (Schlitt and Brazma, 2005). In other words, the network consists on a directed graph, where each node corresponds to a gene and respective gene product, and a directed interaction between A and B means that A gene product is a transcription factor (TF) that regulates the expression of gene B. For *S. cerevisiae* these interactions have been gathered from the literature (Guelzim, et al., 2002) and detected with high throughput experimental techniques like chromatin immunoprecipitation combined with DNA chip technology (ChIP-chip) (Lee, et al., 2002). Additionally, this model organism is relatively rich in microarray gene expression studies. Soon after the availability of topological models of yeast's transcriptional regulatory network, several researchers evaluated the agreement between network structure information and its dynamical behaviour. Some studies only categorized pairs of genes into two or three classes: no common TFs, one or more common TF and two or more common TFs (Allocco, et al., 2004; Yeung, et al., 2004; Yu, et al., 2003). Others looked at blocks of target genes modulated by the same set of TFs, measuring the impact of the number of TFs on block co-expression (Herrgard, et al., 2003). These studies used the same general approach to detect co-expression (applying a threshold to an expression correlation measure), but relied on stringent criteria to recognize co-regulation. In the present report the association patterns that emerge under more generic definitions are explored. With this aim, a probabilistic framework was developed to predict when two genes respond simultaneously to a random perturbation.

Methods

Gene Expression Data

The gene expression data used in this work (Gasch, et al., 2000) consists of 173 cDNA array experiments, involving around 30 different environmental perturbations of yeast cultures. Some perturbations are monitored for several time points. The actual dataset, normalization procedure and other pre-processing methods are described in the paper supplementary website: http://www-genome.stanford.edu/yeast_stress. Pearson's linear correlation coefficient was calculated between the expression profiles of every pair of genes. For a given gene pair, if the number of missing values was greater than 10% of the total number of arrays, the corresponding correlation coefficient was not used in further analysis.

Regulatory Network Information

The yeast's transcription network topology was obtained from the public dataset of Lee and colleagues. The promoter

binding sites of 106 TFs were detected through a chromosome immuno-precipitation followed by hybridization in a DNA chip. As proposed by the authors, a cut-off of $p < 0.001$ was used to define that a promoter region is bound by a given TF (Lee, et al., 2002).

Co-expression Thresholds and Detection of Associations

The association between co-expression and network perturbation responses was assessed by computing the probability of finding strong co-expressions among pairs of gene with similar response probabilities. Having a similar response probability means it is within an interval centred on a value of interest. The range of the interval was always selected to be 1/10 of the total observed range of the respective response probability. The strong co-expression probabilities were computed for 100 equally spaced values across the total range of observed response probabilities. Positive and negative co-expressions were analysed separately. Positively co-expressed gene pairs were defined as the 0.5% gene pairs with the highest expression correlation coefficients. This was equivalent to the application of a correlation threshold $r_{i+} = 0.82$. Analogously, negatively co-expressed gene pairs were defined as the 0.5% gene pairs with the lowest expression correlation coefficients, corresponding to a threshold of $r_{i-} = -0.73$.

Randomization Procedure

To estimate the range of strong co-expression probabilities in the null case of independence between co-expression and perturbation responses, 0.5% of gene pairs were randomly assigned as strongly co-expressed and all the strong co-expression probabilities were re-computed. This procedure was repeated 2000 times, retaining for each value of the perturbation response probability the minimum and maximum limit values of the strong co-expression probability. In a probability profile composed of 100 points, the null probability of finding at least one point out of the random range will be 0.05, according to a Bonferroni correction for multiple testing (Quinn, 2002). All the procedures, graphs and calculations were implemented in Matlab (Release 14).

Results

Most of the common expression datasets are collections of expression values, relating two gene expression states, one before and the other after a specific environmental perturbation. The correlation coefficient between expression profiles is dependent on the number of times the two genes respond simultaneously and the number of times only one

of the genes responds. In this section estimates are derived for the probabilities of observing each kind of response from available network data. This is done in the absence of information about strengths and signs of each regulatory interaction.

In the following derivation, a random perturbation is applied, where a limited number of TFs e are perturbed, and around them the perturbation propagates through all the possible pathways until a maximum depth d .

The response of a given pair of genes is in one of three classes: a) both genes respond (simultaneous response, SR); b) only one of the genes responds (differential response, DR); c) none of the genes respond (neutral response, NR).

NR s should only randomly affect the pair-wise correlation coefficients of expression profiles. SR s effect will depend on interaction signs and DR s should contribute to lower correlation coefficients.

For different d values and for every pair of genes A and B, it is possible to count the transcription factors (TFs) in each of the following four classes: a) number of TFs that regulate both genes – X_d ; b) number of TFs that regulate only gene A – Y_d ; c) number of TFs that regulate only gene B – W_d and d) number of TFs that do not regulate any of the genes – Z_d . Every transcription factor among the N present in the network can be classified in this way. In a random perturbation, e TFs are perturbed. For a given pair of genes, x_d , y_d , w_d and z_d will represent the numbers of perturbed TFs in each of the four possible classes.

Perturbing at least one common TF ($x_d > 0$), irrespective of whatever other TFs are perturbed, is considered to be sufficient to elicit an SR . If $x_d = 0$, it can still be possible to observe an SR if y_d and w_d are both greater than zero. This means that, by chance, the two target genes are regulated after a perturbation by TFs that exclusively regulate each one of the target genes separately. If none of the perturbed TFs regulates any of the target genes ($z_d = e$), then the pair of genes will show a NR s. All the other cases correspond to DR s.

In this deduction two main assumptions are made: a) when more than one TF acts on a given promoter, even if some are activating and others repressing, there is always a net response from the target gene, and b) every TF has an equal probability of being perturbed. The latter is not necessarily reasonable. It could be that TFs are involved in perturbation

responses more frequently because their activity is modulated by a higher number of signal transduction cascades. However, in the absence of information about the activity frequency of different signalling pathways, the least biased hypothesis is the equal-probability one.

Knowing the six parameters (e , d , X_d , Y_d , W_d and Z_d) it is possible to compute the probability of observing any of the possible responses (SR , DR or NR) after a random network perturbation:

$$P(SR | e, d) = P(x_d > 0) + P((x_d = 0) \wedge (y_d > 0) \wedge (w_d > 0)) \tag{1}$$

$$P(DR | e, d) = P((y_d > 0) \wedge (w_d = 0) \wedge (x_d = 0)) + P((w_d > 0) \wedge (y_d = 0) \wedge (x_d = 0)) \tag{2}$$

$$P(NR | e, d) = P((z_d > 0) \wedge (x_d = y_d = w_d = 0)) \tag{3}$$

Applying probability calculus, expressions (1), (2) and (3) can be converted to algebraic functions. The simplest case to calculate is:

$$P(NR | e, d) = \frac{Z_d C_e}{N C_e} \tag{4}$$

Where ${}^n C_p$ is number of combinations of n elements taken p at a time.

$$P(DR | e, d) = \frac{Z_d + Y_d C_e + Z_d + W_d C_e - 2 \cdot Z_d C_e}{N C_e} \tag{5}$$

$$P(SR | e, d) = \frac{N C_e - Z_d + Y_d C_e - Z_d + W_d C_e + Z_d C_e}{N C_e} \tag{6}$$

For $e=1$:

$$P(NR | e = 1, d) = \frac{Z_d}{N} \tag{7}$$

$$P(DR | e = 1, d) = \frac{Y_d + W_d}{N} \tag{8}$$

$$P(SR | e = 1, d) = \frac{X_d}{N} \tag{9}$$

As only one factor is perturbed, the factors in Y_d and W_d

always contribute to DR and not to SR . It is also noticeable that $P(DR)$ only depends on the sum $(Y_d + W_d)$ and not on the relative proportion of each one alone.

The values of X_d , Y_d , W_d and Z_d at the various depths d , are completely defined by the regulatory network architecture. Consequently, for a perturbation with depth d , the network architecture also defines $P(SR)$, $P(DR)$ and $P(NR)$ values for a given pair of genes. I surveyed the relations between the probability of observing strong correlations in gene expression and the values of $P(SR)$ and $P(DR)$ for different values of e (1 to 4) and d (1 to 9, the diameter of the regulatory network used in this study). A change in the parameters e and d induces smooth modifications in the strong correlation probability profiles for $P(SR)$ and $P(DR)$. Strong positive expression correlations are more clearly associated with high $P(SR)$ for $e=3$ and $d=3$. The intensity of the strong co-expression enrichment grows from $d=1$ until $d=3$, and decreases for higher depths. $P(DR)$ for those parameter values is also a characteristic example of the remaining $P(DR)$ profiles. These particular results are shown in Figure 1. The two most remarkable observations made from Figure 1 are that the greatest

enrichments of positive strong expression correlations happen near the maximum values of $P(SR)$ and at the minimum possible values of $P(DR)$. Additionally, strong positive correlations are less frequent for higher values of $P(DR)$. Near a $P(SR)$ of 0.5, half of the gene pairs have a correlation coefficient higher than the top 0.5% threshold, meaning that strong positive correlations are concentrated 100 times. Strong negative expression correlations only show some significant deviation from the random null model when they are stratified by $P(DR)$ values. Although much less intense, they have an inverted profile relatively to the positive correlations. That is, they are less frequent for low values of $P(DR)$ and slightly enriched for some intervals of greater $P(DR)$ values.

Combining expressions (8) and (9):

$$\frac{X_d}{X_d + Y_d + W_d} = \frac{P(SR | e = 1, d)}{P(SR | e = 1, d) + P(DR | e = 1, d)} = P(SR | z_e = 0, e = 1, d) \tag{10}$$

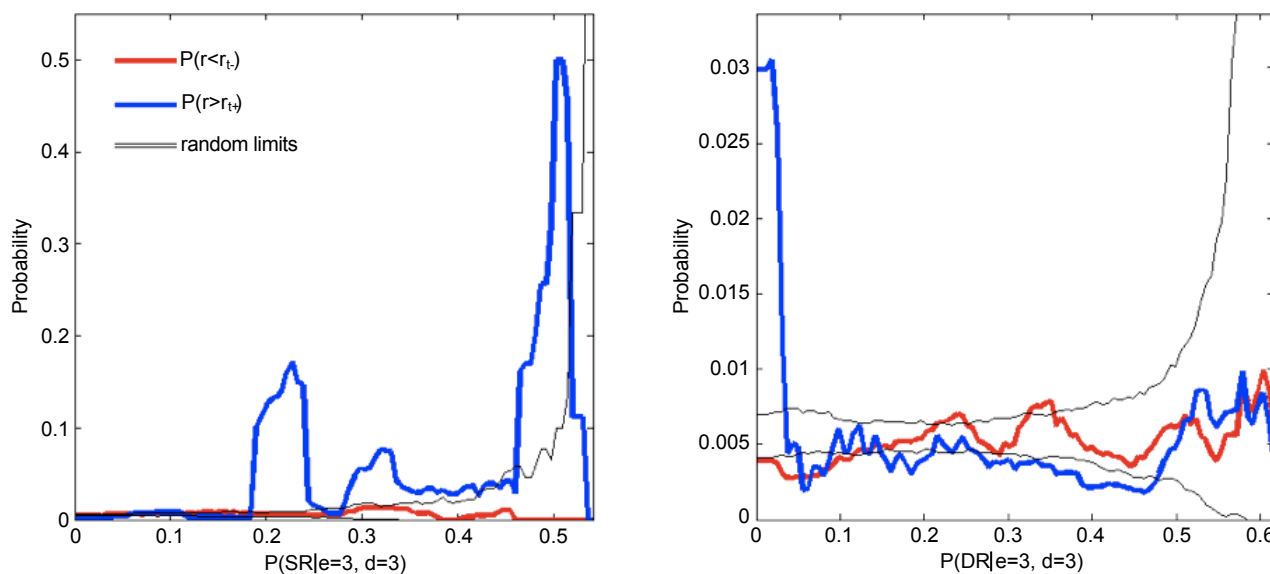


Figure 1: Plots of the relation between strong co-expression and the probability of observing simultaneous (SR) or differential (DR) responses of a target gene pair after network perturbations initially affecting $e=3$ TFs and spreading until a maximum depth of $d=3$. Blue lines represent the probability of finding strong positive co-expression ($P(r > r_+)$, with a positive threshold $r_+ = 0.82$ and where r is the linear correlation coefficient between two gene expression profiles) among pairs of genes with similar response probabilities. Red lines represent the probability of finding strong negative co-expression ($P(r < r_-)$, with a negative threshold $r_- = -0.73$) among pairs of genes with similar response probabilities. Black lines represent the maximum and minimum randomized probabilities of finding strong co-expression among pairs of genes with similar co-regulatory similarity measure, after 2000 random permutations of the gene expression correlation matrix.

According to expression (10), the proportion of common direct TFs is the probability of observing a simultaneous response knowing that it is not neutral - $P(SR|z_d=0, e=1, d=1)$. If we recall that neutral responses should not affect correlation coefficients between expression profiles, the proportion in expression (10) should be more related with co-expression than the number of common TFs alone. After observing the gene expression profile of a sufficiently high number of perturbation experiments or distinct cellular states it would be expectable that genes sharing more common TFs are more frequently co-regulated. On the other hand, if two genes are regulated by very distinct TF sets, it would be expectable to observe perturbations where only one of the genes is regulated. Measuring the proportion of common TFs does effectively account for both the common TFs and the exclusive TFs of a given target gene pair.

In fact, Notebaart and colleagues (Notebaart, et al., 2008) successfully used a transcription factor similarity (*TFS*) measure to explain the coupling of metabolic flux between two enzyme-coding genes. This success was not attained when graph distance between the genes in the transcription regulatory network was used instead of the *TFS*. They defined *TFS* as the total number of shared TFs between two genes divided by the total number of unique TFs regulating the two genes, which is equivalent to expression (10).

Other well-known graph-theoretic measure used to analyse biological networks is the topological overlap measure (*TOM*) (Ravasz, et al., 2002). Interestingly, *TOM* is also closely related to the perturbation response probabilities derived in our model:

$$\begin{aligned}
 TOM &= \frac{P(SR | e = 1, d = 1)}{\min(P(AR | e = 1, d = 1), P(BR | e = 1, d = 1))} \\
 &= \frac{X_1}{\min((X_1 + Y_1), (X_1 + W_1))} \tag{11}
 \end{aligned}$$

Where, for the pair of genes A and B, $P(AR)$ and $P(BR)$ are the probabilities that genes A and B, respectively, respond to the perturbation. Yip and Horvath expanded *TOM* to a generalized form (*GTOM*) that presents associations with gene function that are more robust to uncertainties in network topology data (Yip and Horvath, 2007). In this probabilistic model language, the generalization corresponds

to the possibility to consider perturbations with higher depths:

$$\begin{aligned}
 GTOM(d) &= \frac{P(SR | e = 1, d)}{\min(P(AR | e = 1, d), P(BR | e = 1, d))} \\
 &= \frac{X_d}{\min((X_d + Y_d), (X_d + W_d))} \tag{12}
 \end{aligned}$$

It is readily apparent from expression (12) that *GTOM* can be further generalized by allowing perturbations involving more than one TF:

$$GTOM(e, d) = \frac{P(SR | e, d)}{\min(P(AR | e, d), P(BR | e, d))} \tag{13}$$

These relationships and equivalences between our probabilistic approach and common graph based measures further suggests that $P(SR)$ and $P(DR)$ estimates can provide valuable information about the correlation between regulatory network topology and function. On the other hand, measures like *TSF*, *TOM* or *GTOM* are also enriched with the use of the presented response probabilities. They gain a more concrete biological meaning, which can potentiate the interpretation of the associations between their values and network functional properties. *TSF* may be read as the probability of observing a simultaneous response to network perturbations, knowing that the response is not neutral. *TOM* and *GTOM* are also proportional to the probability of simultaneous response, but their values are normalized by the response probability of the gene that responds less frequently to network perturbations. Thus, when *TOM* or *GTOM* are 1, it does not mean that both genes respond exactly to the same network perturbations, but that one gene is sensible to a set of perturbations that is contained in the set of perturbations that have an effect in the other gene.

Discussion

A main result of this work consisted in proposing a simple rational connection between the architecture of the regulatory network topology and its functional behaviour. Using a minimal probabilistic model it was possible to justify the differences between computing the proportion versus the absolute number of common TFs. In addition to the effect

of common TFs, the use of the proportion has also accounts for the TFs exclusively regulating the expression of one gene in a target pair. As a consequence, it reflects both the probability of observing a simultaneous response, $P(SR)$ and a differential response, $P(DR)$, after a system perturbation. The absolute count is more directly related with the $P(SR)$ and does not include the impact of $P(DR)$ in the experimental correlation coefficients between expression profiles.

It is also shown that integrating information about indirect factors may be important for the analysis of gene expression data. The results obtained with the $P(SR)$ indicate that including factor information until 3 regulatory levels up of the target genes provides the most strong association between the probability of simultaneous response and the probability of observing strong co-expression. This observation does not necessarily imply that on average, the perturbations associated with the used dataset had a depth of three connections. It could alternatively mean that the past history of network activations or inhibitions could be relevant for the response to newer perturbations.

Besides the associations found between response probabilities and strong correlation of expression profiles, this approach was further validated by the relationship with other methodologies (*TFS*, *TOM*, *GTOM*), which on their own have previously demonstrated their utility in providing valuable insights into the functional organization of biological regulatory networks. Perturbation response probabilities enhance the interpretation of graph-based measures by attributing them a biological meaning related to the network capacity to respond to and propagate random perturbations. In fact, the use of $P(SR)$ and $P(DR)$ allows an extra generalization of the *GTOM* measure. It is plausible that the latter can have an increased performance since the most strong association between $P(SR)$ and expression profiles occurred for perturbations of 3 TFs propagating through 3 network connection levels ($e=3$ and $d=3$).

Conclusions and Perspectives

The presented approach uses minimal information about transcriptional regulatory network. I believe this approach can be useful as it can be easily validated with most common microarray experiments. More detailed models of transcriptional regulatory network dynamics, using Boolean, Bayesian, stochastic or differential equations frameworks would be optimally validated against expression profiles with long and well-sampled time series not so commonly

produced. As the network information grows and becomes more complete, the probabilistic model shown here can provide better predictions. The simplicity of the input information and model assumptions may allow the analysis of integrated networks including simultaneously several mechanisms of gene transcription or protein activity regulation.

I finally propose that coarse level analysis like the one presented here can provide a useful bridge between large gene expression datasets and more fine-detailed models of biological network dynamics.

References

1. Allocco DJ, Kohane IS, Butte AJ (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* 5: 18.
2. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5: 101-113.
3. Bertalanfy LV (1928) *Kritische Theorie der Formbildung*. Borntraeger Berlin.
4. Gasch AP, Spellman PT, Kao CM, Carmel HO, Eisen MB, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11: 4241-4257.
5. Guelzim N, Bottani S, Bourguin P, Kepes F (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 31: 60-63.
6. Herrgard MJ, Covert MW, Palsson BO (2003) Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res* 13: 2423-2434.
7. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar JZ, et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799-804.
8. Milo R, Shen OS, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: simple building blocks of complex networks. *Science* 298: 824-827.
9. Notebaart RA, Teusink B, Siezen RJ, Papp B (2008) Co-regulation of metabolic genes is better explained by

- flux coupling than by network distance. PLoS Comput Biol 4: e26.
10. Quinn GP, Keough MJ (2002) *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge.
 11. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297: 1551-1555.
 12. Schlitt T, Brazma A (2005) Modelling gene networks at different organisational levels. *FEBS Lett* 579: 1859-1866.
 13. Westerhoff HV, Palsson BO (2004) The evolution of molecular biology into systems biology. *Nat Biotechnol* 22: 1249-1252.
 14. Yeung KY, Medvedovic M, Bumgarner RE (2004) From co-expression to co-regulation: how many microarray experiments do we need?. *Genome Biol* 5: R48.
 15. Yip A, Horvath S (2007) Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 8: 22.
 16. Yu H, Luscombe NM, Qian J, Gerstein M (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet* 19: 422-427.