Transcriptional regulatory network topology from statistics of DNA binding sites

A. Kabakçıoğlu*

Department of Physics, Koç University, 34450 Sarıyer, İstanbul, Turkey

Available online 20 July 2007

Abstract

We show that the out-degree distribution of the gene regulation network of the budding yeast, *Saccharomyces cerevisiae*, can be reproduced to high accuracy from the statistics of TF binding sequences. Our observation suggests a particular microscopic mechanism for the observed universal global topology in these networks. The numerical data and analytical solution of our model disagree with a simple power-law for the experimentally obtained degree distribution in the case of yeast.

© 2007 Elsevier B.V. All rights reserved.

PACS: 87.17.Aa; 89.75.Fb; 89.75.Da

Keywords: Transcriptional regulatory network; Regulatory element; Consensus sequence; Degree distribution

1. Introduction

Regulation of gene expression is a central concept in cell biology. It helps one to understand how a single fertilized egg develops into a multicellular organism with a variety of cells, how each of them can manufacture and use different sets of proteins, and how they adopt their mode of operation to changes in the environment, when each individual cell hosts the very same genome. The DNA microarray technique developed in the late 1990s [1] generated a flood of gene expression data, which made the large-scale analysis of gene regulation at a chosen instance of the cell’s life-cycle accessible. Microarray experiments deliver a direct and simultaneous measure of the expression levels for thousands of genes (possibly the whole genome, as in the case of the budding yeast—*Saccharomyces cerevisiae* [2]). This information may then be used to identify genes which regulate each other’s transcription [3,4].

Functional organization of the cell is based on such communicating genes and is summarized in a transcriptional regulatory network. An edge in this network reflects the presence of a regulatory mechanism between the two genes represented by the two nodes terminating the edge. The dominant mechanism of regulation relies on particular proteins called transcription factors (TF), that bind to short DNA segments of TF-specific sequences [5] at the regulatory regions of the controlled gene. Hence, the regulatory network’s
topology can be seen as a by-product of the matching statistics between the regulatory regions of genes and the DNA sequences that TFs exclusively bind. For simplicity, we will assume here that these two sequence sets come from the same distribution, which, as far as the out-degree distribution is concerned, allows us to consider one regulatory sequence (RS) per gene instead of two. Model details and possible improvements are discussed later in the text.

We propose that the statistics of this generic matching rule and the amount (but not the content) of information in these sequences determine the topology of genetic regulatory networks [4,6,7]. In this article, we demonstrate our point by focusing on a single topological signature, the out-degree distribution. A thorough demonstration of the suggested connection examining a variety of global structural features has been published elsewhere [11]. To this end, we reconsider a model recently introduced [8] for RNA interference and later reinterpreted to include TF-based regulation in the cell [9]. In this model, one represents genes by short random RSs and estimates analytically the out-degree distribution for an unrealistic but analytically tractable RS length distribution. We find a qualitatively different behavior for low- and high-degree regimes, which disagrees with a unique power-law suggested earlier [6,10]. Next, we consider a more realistic scenario and numerically show that the results can be tuned to quantitatively agree with the available data for the yeast, while the qualitative features of the analytical solution are still preserved. We finally argue that the model not only reproduces the out-degree distribution but the optimal RS length distribution thus obtained also matches that found for the yeast experimentally.

2. Description of the model

We start with defining a “reduced genome” as a pool of gene regulatory segments. For simplicity we take one RS per gene and represent each as a random binary sequence with \( \ell \) bits, where \( \ell \) is chosen from a generic length distribution \( P(\ell) \). We denote the RS of the \( i \)-th gene by \( G_i \) as shown in Fig. 1 and refer to \( \ell \) as “length”, which in reality is closer to the information content [12] of the sequence than its actual length on the DNA. A more biologically accurate model should consider two sequences per gene (see Ref. [11]). Yet, the present model, with its simplicity, better serves the purpose of developing an intuition for the causal connection between the RS statistics and the network topology.

Another important feature of an RS length distribution is that it is peaked at an intermediate value, since selective binding onto the DNA implies that the recognized sequences typically contain an amount of information large enough to be encountered sufficiently seldom. This, in fact, is the case for the yeast, as shown below. Here, as a generic choice, we will assume a Gaussian length distribution for the RS lengths. On the other hand, if a Poisson process is used for generating RSs of arbitrary length (as in Ref. [9]) one ends up with an exponentially decaying length distribution which, although unrealistic, has the merit of being fully analytically solvable. Our strategy will be to use the Gaussian model for reproducing the experimental results and to interpret them under the light of the analytical solution which provides insight to the features observed.

Once a pool of RSs is generated, the gene network (where each RS corresponds to a vertex) is constructed through the adjacency matrix (the edges) defined by the matching condition

\[
\begin{align*}
  w_{ij} = \begin{cases} 
    1, & G_i \subseteq G_j, \\
    0, & \text{otherwise.}
  \end{cases}
\end{align*}
\]

By \( G_i \subseteq G_j \) we mean the sequence \( G_i \) appears as a subsequence of \( G_j \), \( w_{ij} = 1 \) indicates a directed link from \( G_i \) to \( G_j \). Note once again that by doing so, we coalesce in a single object \( G_i \), two different entities related to a gene: the RS used for regulating the gene itself and the nucleotide sequence which its product binds to for

\[
\begin{align*}
  G_i & \quad G_{i+1} & \quad G_{i+2} \\
  \cdots & \quad 11101000 & \quad 010011100 & \quad 11111 \cdots \\
  \text{Reduced Genome}
\end{align*}
\]

Fig. 1. The binary representation of the genome used to simulate the transcriptional regulatory network. Each gene is represented by the information content of its regulatory segment, which in our model is a random binary string with length chosen from a Gaussian distribution.
regulating other genes. We assume here that the two are picked from the same distribution. In this case, going one step further and assigning the same sequence to both entities for a gene does not change the expected out-degree distribution, although higher-order correlations are altered. Since we are only interested in the out-degree distribution, we will adopt this one-string-per-gene approach for which we can make use of earlier numerical and analytical results [8,9]. For example, it was shown in Ref. [9] that, certain obvious generalizations such as allowing less than perfect matches, considering an alphabet of four letters, or matching complementary (rather than identical) sequences do not qualitatively alter the conclusions derived from this simpler model.

3. Out-degree distribution

In order to properly interpret the resultant out-degree distribution, we provide below a qualitative summary for the analytical solution of the model [9]. To this end, it is convenient to group the $G_i$ into subsets according to their lengths. So let

$$G_l = \{G_i | l_i = l\}.$$ 

A central quantity of interest is the probability that a member of $G_l$ has an outgoing edge terminating in $G_k$ which we denote by $p(l, k)$. A good approximation for $p(l, k)$ is given by

$$p(l, k) = \begin{cases} \frac{(k - l + 1)}{2^l}, & l \leq k, \\ 0, & l > k, \end{cases}$$

(2)

where the exponential factor is the probability of match for a substring of length $l$ and $k - l + 1$ is the number of substrings of length $l$ in an RS of length $k$. A more accurate analytical expression for $p(l, k)$ can be found in Ref. [9].

Using Eq. (2) and for a large number of genes one obtains the out-degree for an RS of length $l$ to be a Gaussian distribution with a calculable mean $d_l$ and variance $\sigma_l^2$. The total out-degree distribution follows from summing up the contributions from each $G_l$. With increasing RS length $l$, the spacing between the Gaussian peaks for successive $l$'s approaches zero faster than the peak widths $\sigma_l \sim \sqrt{d_l}$. Two regimes immediately surface:

(1) $l < l_c$, where the inter-peak separation is larger than the widths so that the out-degree distributions for successive $G_l$ are well separated, giving a series of peaks. The probability of a degree falling in this regime is determined dominantly by a single value of $l$.

(2) $l > l_c$, where distributions for different $G_l$ strongly overlap so that the out-degree probability is a cumulative quantity with contributions from many RS lengths.

The crossover length $l_c$ is given by the condition

$$d_l - d_{l+1} \simeq \sigma_l$$

(3)

and corresponds roughly to $l_c \simeq 10$.

When $l < l_c$, degree distributions for each $G_l$ are well separated. The full probability distribution exhibits distinct peaks, with the area under each peak centered around $d_l$ being equal to $P(l) \simeq \langle n_l \rangle / \sum n_l$, the probability for having an RS of length $l$. Clearly, further properties of this regime depend on the particular length distribution $P(l)$ one considers. E.g., when the lengths are distributed exponentially one obtains a series of peaks enveloped by a power-law decaying with the exponent $\gamma_2 \simeq 0.5$ [9]. In contrast, a Gaussian function which more accurately represents the RS length distribution in the yeast suppresses sequences that are both too short and too long. As a result, distinct peaks for each $G_l$ are less prominent (with a possibly larger $\gamma_2$). But more importantly, the peaks become invisible due to insufficient sampling in a realistically long model genome (see Fig. 2a). This could be the case for the yeast network, as well.

For $l > l_c$, the analysis presented above ceases to be valid, since the peaks start overlapping. In this regime, the dominant contribution to the out-degree distribution comes from rare matches between relatively long strings. Consequently, the probability for an RS of length $l$ having degree $k$ is given by a Poisson distribution.
The out-degree distribution for small degree \( d \) is then the expectation value of a Poissonian random variable in an ensemble of varying RS lengths:

\[
p(d) = \sum_{l=l_c}^{\infty} p(l) \left( \frac{dl}{d!} \right) e^{-d_l},
\]

(4)

where \( l_c \) is the crossover length obtained from Eq. (3). A saddle-point treatment of Eq. (4) gives the scaling behaviour in this regime with small out-degrees as

\[
p(d) \approx d^{-\gamma_1},
\]

(5)

with scaling exponent \( \gamma_1 = \frac{1}{2} + \gamma_2 \gtrsim 1.0 \). Both \( \gamma_1 \) and \( \gamma_2 \) are in perfect agreement with earlier numerical results [8] and a \( \gamma_1 > 1.0 \) is in good agreement with Ref. [7].

The distribution obtained from above considerations is shown in Fig. 2a. The first scaling regime exhibits a short-lived power-law decay. It is robust to sample to sample fluctuations due to self-averaging, since a lot of “long” RSs contribute to the probability of the out-degrees. However, the proximity of the crossover point limits validity of Eq. (5) to less than a decade wide region. In contrast, the tail hosts much smaller probability per degree. The oscillation characteristics to this regime are suppressed in the case of a Gaussian length distribution which filters out short RSs dominant here (to be compared with Refs. [8,9]). The number of genes in the yeast (~6000) allow only a poor sampling of the tail of the degree distribution, rendering such features invisible.

4. Model vs yeast

4.1. Out-degree distribution

For comparison, we refer to an independent compilation of the same data from the experiments on the yeast’s genome [2]. Microarray experiments yield mRNA levels of all 6178 genes of the yeast, which later are subjected to further analysis for extracting the pairs of genes whose expression levels are correlated in time. We picked an extensive analysis available to us [4], although several other data sets exists with the same qualitative degree distribution reported by several groups earlier [6,10]. Based on quite similar experimental data, Ref. [4] concludes that there is no power-law behavior, whereas Refs. [6,7,10] suggest a fit to a single power-law in the full degree spectrum.

In order to compare our model’s prediction with the experiments it is crucial to realize that the experimental data probes a single realization, i.e., that of yeast. Therefore we should not average over many genomes, but pick a single realization with a number of genes as in yeast. The parameters of \( P(l) \), the Gaussian RS length distribution are then fixed by requiring the best match with the yeast out-degree distribution in Ref. [4]. We set the average and the standard deviation of \( P(l) \) be 14 and 1.5, respectively. These values yield the same
percentage of genes (~45%) with a nonzero out-degree and the same amount of probability in the two scaling regimes.

Fig. 2 displays the out-degree distribution of the Gaussian model and the yeast’s regulatory network [4] side by side. The model degree distribution shown is a typical one chosen out of a handful of samples we generated with the above parameter values. The low-degree end of the theoretical distribution shows the same kind of power-law decay as obtained from the experiment, practically independent of the model parameters and the realization. This is a self-averaging regime where the power-law is a generic feature resulting from the cumulative contribution of many RS lengths. On the other hand, the expected well-separated peaks of the tail are not visible due to the insufficient number of yeast genes that sample this regime. Exact positions of the data points here differ from sample to sample for the same reason.

4.2. A cross-check: true RS lengths of yeast

Ultimately, the conjectured connection between the specific binding sequences for the regulatory proteins and the out-degree distribution of the network can be cross-checked. A yet small database of “consensus sequences” that some DNA binding yeast proteins exclusively attach onto is available on several databases on the internet. We converted the available data from one such database (http://rulai.cshl.edu/SCPD) to a distribution in bitwise information content. Consensus sequences have positions where the nucleotide type is completely determined (2 bits), or can be one of the two alternatives (e.g., a pyrimidine or a purine: 1 bit), or a non specific gap-filler (0 bit). Fig. 3 compares the information content of the regulatory elements thus obtained with that predicted by our Gaussian model based on the out-degree distribution alone. The overlap of the two supports the conjectured connection between the out-degrees and the RS lengths. Note that the model outcome, which was optimised solely with respect to the out-degree distribution derived from experiments on yeast, is not biased in any way to agree with this consensus sequence data.

5. Discussion

Certainly, regulation of gene expression in the living cell is much more complex than depicted here. Nevertheless, the agreement between the presented generic model and the yeast’s transcriptional regulatory network suggests that the shape of the out-degree distribution is predominantly a statistical effect based on sequence-matching, presumably involving the short DNA segments recognized by various transcriptional regulatory proteins. One corollary of this statistical connection is the robustness of our conclusions to the experimental data sets chosen for comparison, since the relevant features are the probability distributions (which do not differ from one database to another) rather than the number of genes, TFs, or interactions. Our model should benefit from including further aspects of gene expression such as, including further information on promoter regions, combinatorial regulation or the possibility of multiple RSs for a single gene, especially if

Fig. 3. The bit-length distribution of the specific DNA segments which a set of regulatory proteins bind, compared with the optimal model Gaussian RS length distribution that reproduces the experimental out-degree distribution.
further features of the network are of interest. Further investigation on yeast and other organisms is planned as future work.

Acknowledgments

I am grateful to M. Hörnquist for making the data available, M. Baiesi for beneficial discussions, A. Erzan, M. Mungan and D. Balcan for continuing collaboration, and the organizers of the PASI-2006 School on Disorder and Complexity for an exceptionally inspiring meeting.

References


