Available online at www.sciencedirect.com

Physica A 386 (2007) 764–769

www.elsevier.com/locate/physa

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.

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

*Tel.: +90 2123381830; fax: +90 2123381559.
E-mail address: akabakcioglu@ku.edu.tr

0378-4371/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2007.07.013

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 ℓ bits, where ℓ is chosen from a generic length distribution $P(\ell)$. We denote the RS of the ith gene by G_i as shown in Fig. 1 and refer to ℓ 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

$$w_{ij} = \begin{cases} 1, & G_i \subset G_j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

By $G_i \subset G_j$ 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

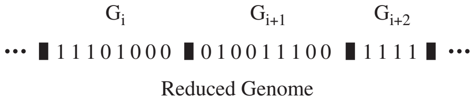

Fig. 1. The binary representation of the genome used to simulate the transcriptional regulatory network. Each gene is represented by the information content of its regulatory segment, which in our model is a random binary string with length chosen from a Gaussian distribution.