Journal of Theoretical Biology 263 (2010) 281–294

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Information content based model for the topological properties of the gene regulatory network of *Escherichia coli*

Berkin Malkoç [a], Duygu Balcan [b], Ayşe Erzan [c,*,1]

[a] Department of Physics, Faculty of Sciences and Letters, Istanbul Technical University, Maslak 34469, Istanbul, Turkey
[b] Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University, Bloomington, IN 47408, USA
[c] Department of Physics, Faculty of Sciences and Letters, Akdeniz University, Dumlupınar Caddesi, 07058 Antalya, Turkey

ARTICLE INFO

Article history:
Received 3 June 2009
Received in revised form
21 November 2009
Accepted 23 November 2009
Available online 3 December 2009

Keywords:
Systems biology
Stochastic modeling
Combinatorics
Emergence

ABSTRACT

Gene regulatory networks (GRN) are being studied with increasingly precise quantitative tools and can provide a testing ground for ideas regarding the emergence and evolution of complex biological networks. We analyze the global statistical properties of the transcriptional regulatory network of the prokaryote *Escherichia coli*, identifying each operon with a node of the network. We propose a null model for this network using the content-based approach applied earlier to the eukaryote *Saccharomyces cerevisiae* (Balcan et al., 2007). Random sequences that represent promoter regions and binding sequences are associated with the nodes. The length distributions of these sequences are extracted from the relevant databases. The network is constructed by testing for the occurrence of binding sequences within the promoter regions. The ensemble of emergent networks yields an exponentially decaying in-degree distribution and a putative power law dependence for the out-degree distribution with a flat tail, in agreement with the data. The clustering coefficient, degree–degree correlation, rich club coefficient and *k*-core visualization all agree qualitatively with the empirical network to an extent not yet achieved by any other computational model, to our knowledge. The significant statistical differences can point the way to further research into non-adaptive and adaptive processes in the evolution of the *E. coli* GRN.

1. Introduction

Complex biological systems such as transcriptional regulatory networks (Ma et al., 2004; Samal and Jain, 2008), protein–protein interaction networks (Spirin and Mirny, 2003) or metabolic networks (Jeong et al., 2000) all require the satisfaction of physical and chemical constraints between pairs of molecules. There is a growing body of knowledge regarding specific protein–DNA or protein–protein interactions and metabolic pathways. Nevertheless, it would be hopeless at this stage to try to predict the universal (Bergmann et al., 2004) features or the global architecture of the gene regulatory network or the proteomic network on this basis. On the other hand, rudimentary forms of many complex structures we observe in biology can arise spontaneously, given the combinatoric profusion of possible ways in which the simple building blocks, such as nucleotides, amino acids, etc., can be associated with each other (Dawkins, 1986, 2006; Kauffman, 1993). In particular, given a sufficiently long sequence (such as the genome), and a set of complex biological molecules (such as the proteins), it is very likely that some of them will have affinities for certain subsequences of the genome and bind them, giving rise to an interaction network.

We propose that important insights could be gained into the emergence of biological networks by employing null models making use of the combinatoric properties of random sequences (Kim et al., 2003). Comparison with real biological data could enable us to distinguish between: (i) generic properties of such networks, (ii) features that could spontaneously evolve under the kinetics of duplication and divergence (Wagner, 1994; Sengun and Erzan, 2006; Lynch, 2007) and (iii) those features that must have clearly evolved in response to specific selection pressures (see, e.g., Kashtan et al., 2009 and references therein). Such a null model for the global statistics of a biological network would not aim to model it on a node to node basis, but would provide a much more appropriate starting point for further elaboration than would a "classical" random network (Erdös and Rényi, 1959, 1960).

In previous work (Balcan and Erzan, 2004; Mungan et al., 2005) we have demonstrated that any collection of random sequences of varying lengths with a matching rule between them naturally gives rise to a complex network. In a recent paper

* Corresponding author. Tel.: +90 242 310 2279.
E-mail address: erzan@itu.edu.tr (A. Erzan).
[1] Permanent address: Department of Physics, Faculty of Sciences and Letters, Istanbul Technical University, Maslak 34469, Istanbul, Turkey.

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.11.017

(Balcan et al., 2007) it was shown that, the statistical properties of the transcriptional gene regulatory network (GRN) of the eukaryote *Saccharomyces cereviciae* (yeast) (Teixeira et al., 2006) could be predicted by such a null model, with random sequences representing the binding sites of the transcription factors and the promoter regions. The information content of the binding sequences, extracted from their probability matrices (Teixeira et al., 2006), was used to determine their effective length distribution. The only free parameter employed in this model was the exponent of the long tailed power law distribution (Almirantis and Provata, 1999) exhibited by the promoter region lengths (Harbison et al., 2004). This exponent was chosen to obtain the best fit of the model networks to the empirical GRN. In fact, the model was not very sensitive to the precise value of this exponent.

It should be mentioned that similar ideas of sequence matching have been employed earlier in constructing model genomes and interaction networks (Banzhaf and Kuo, 2004; Geard and Wiles, 2003; Reil, 1999; van Noort et al., 2004; Wagner, 1994; Watson et al., 2004). However, in these models no attempt was made to take into account the variability in the specificity of the interactions and the results were not very realistic.

Taking the content-based network approach of Balcan et al. (2007) we here propose a model which is able to predict many of the global statistical properties of the GRN of *Escherichia coli*, on the basis of the distribution of the degree of specificity of the binding sequences/sites (Sengupta et al., 2002). If the high degree of agreement between the GRN of yeast and the content-based model is not due to pure chance, we should be able to demonstrate that the model also captures the characteristics of a prokaryotic GRN that is known to be organized somewhat differently.

In prokaryotes, in contradistinction to eukaryotes, one encounters a more complex, hierarchical organization of promoter regions and groups of genes which they regulate (Alberts et al., 2002; Almirantis and Provata, 1999; Browning and Busby, 2004; Ma et al., 2004; Okuda et al., 2007; Salgado et al., 2006b; Warren and Wolde, 2003). Genes which are to be co-regulated typically occur in tandem. A group of one or more genes which are very often (but not necessarily always) transcribed together into one mRNA is called a "Transcription Unit" (TU). A maximal series of genes that can be transcribed into an mRNA, organized into one or more tandem or overlapping TUs, are termed an "operon" (see Fig. 1). The operon, and the TUs that it may contain, are (generally) preceded by promoter regions (PRs). As in the case of eukaryotes, special proteins called transcription factors (TFs) have high affinities for certain sequences, called "binding sequences", or somewhat misleadingly, "binding sites" (bs) within these PRs.[2] The binding of a bs by a TF may facilitate or suppress the transcription of all the genes in the associated TU, thereby regulating their expression.

In the next section we present our information-content based null-model for the *E. coli* GRN. In Section 3, we describe the biological information which determines the model parameters. In Section 4 we outline our simulation results which we obtain by generating an ensemble of realizations of the model GRN, compute the topological properties of these different realizations and compare them with the empirical network which we extract from the RegulonDB v6.0 (Gama-Castro et al., 2007). Section 5 will be devoted to a discussion of our findings.

[2] The terminology here is far from uniform. Certain authors, e.g., Alberts et al. (2002), refer to the DNA sequence bound by the RNA polymerase as a "promoter", and the region in which the binding sequences of the transcription factors regulating a given TU are to be found, is termed a "gene control region" or a "regulatory region". On the other hand, Harbison et al. (2004), Berg et al. (2004), and Almirantis and Provata (1999) among many others use the term in the way we have defined it here.

Fig. 1. Hierarchical organization of the *E. coli* genome. Shown are an operon (big box), which constitutes a transcription unit (TU) in itself, and a smaller box embedded in it, which is another TU. Promoter regions (PRs), shown here as striped horizontal bars, are attached to both the whole operon and the TU embedded in it. The vertical bars indicate different genes within the TUs. A blob on the left hand side represents a transcription factor which may bind a binding site (bs) within one of the PRs, initiating the transcription of the TU associated with that PR. The drawing is not to scale.

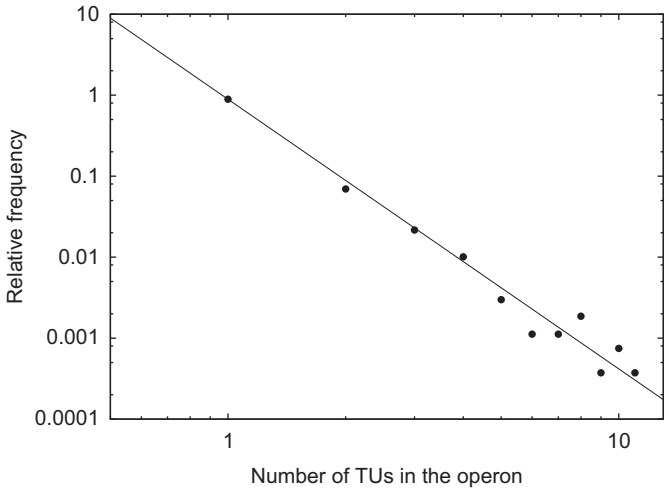


Fig. 2. Transcription units per operon. Distribution of the number of transcription units (TUs) per operon for the *E. coli* genome, extracted from the RegulonDB v6.0 (Gama-Castro et al., 2007). The straight line is the power law $x^{-\nu}$, with $\nu = 3.333 \pm 0.045$.

2. The model

To build the model network, we choose a certain number of nodes, each representing an operon. For each node, we pick from an empirically determined distribution (see Fig. 2), the number of TUs to be associated with that operon. This will determine the number of random PR sequences which we will assign to the node. The lengths of these PR sequences will be chosen from the empirical PR length distribution (see Section 3 and Fig. 3).

We then randomly choose an empirically determined percentage (see Table 1) of operons which will incorporate TF coding genes. Each TF-coding node will be assigned one or more binding sequences its TF will recognize. The number of bss per TF obeys the empirical distribution in Table 1. The length of each distinct bs will be independently chosen from the bs length distribution (see Fig. 4) determined (see next section and Appendix A) from the information content of the known binding sequences.

Finally, we consider each TF-coding operon (with at least one bs, by definition), and ask if any of these sequences are contained as a subsequence within any of the PRs assigned to any of the operons. A directed edge is then drawn from the operon coding the TF to each of the operons whose PRs contain the bs associated with this TF. Self-interactions are included in this scheme. Nodes connected by directed edges going both ways are considered to be

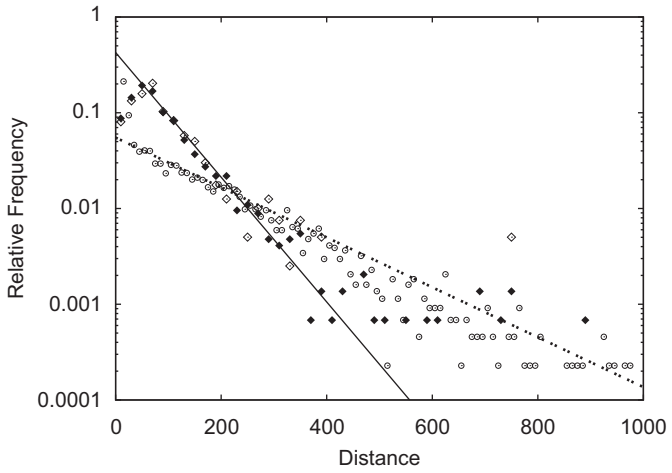


Fig. 3. Where do the binding sites occur? The distribution of intergenic distances (Rudd, 2000) in base pairs (circles) and the distances from the centers of the binding sequences (bsc) to the start codons of the operons or transcription units (TUs) which they regulate (from the RegulonDB v6.0). The latter distances are shown as filled diamonds for operons, empty diamonds for TUs within an operon. The absolute values of the distances, which may be negative (upstream) or positive (downstream) were taken. The distributions were fitted to exponential functions $\sim \exp(-bl)$, omitting the first three points and roughly ten outlying data points (out of about 100) with distances up to $l \simeq 2500$. The b values are given in Table 1.

Table 1
Network parameters and genome data extracted from the RegulonDB (Gama-Castro et al., 2007) and Ecogene (Rudd, 2000).

Number of nodes (operons)				2684			
Number of known TFs				159			
Percentage of TF coding operons				5.9			
Candidate PR lengths				b			
Intergenic				$0.00648 \pm 9 \times 10^{-5}$			
bsc to start codon (operon)				0.0156 ± 0.0003			
bsc to start codon (TU in op.)				0.014 ± 0.001			
bsc to start codon (op.s and TUs)				0.0152 ± 0.0007			
Number n of bs per TF	1	2	3	4	5	7	9
Percentage of TFs with n bs	76.6	11.7	1.7	3.3	3.3	1.7	1.7

The percentage of transcription factors (TFs) recognizing n distinct binding sequences has been calculated on the basis of the 60 TFs for which this information is available. The parameter b refers to fits to distributions of the form $\exp(-bl)$ for the intergenic and for binding site center to start codon distances l (see Fig.3).

connected by one bi-directional edge. The resulting directed network is one realization of our model GRN.

2.1. Information content and specificity of the connections

Since we require an exact match for each connection, the number of characters in the bs determines the specificity of the interaction. In a real genome, the consensus bs represents a number of slightly different sequences to which the same TF binds. Therefore, we identify the length of a model binding sequence with the effective length computed (Balcan et al., 2007) from the probability matrices of the set of similar sequences recognized by a given TF (Gama-Castro et al., 2007; Gershenzon et al., 2005; Li et al., 2002; Pachkov et al., 2007; Stormo, 2000) (see Appendix A).

We define (Balcan et al., 2007) the effective information content of a consensus bs as its Shannon information (Shannon, 1949; Avery, 2003) *relative* to a random sequence,

$$I_m = \sum_{i=1}^{l_m} \sum_{j=1}^{4} p_{ij}^{(m)} \ln p_{ij}^{(m)} + l_m \ln 4. \quad (1)$$

Here l_m is the length of the mth bs and the elements of the probability matrix, $p_{ij}^{(m)}$, are the probabilities for encountering the jth nucleotide (from among A,T,C,G) at the ith site, computed over the different instances of this bs within the different PRs. The last term comes from subtracting off the information content of a random sequence of length l_m, with equal probabilities for the four letters. Note that any other choice of the background distribution would have just shifted the lengths by a constant amount per position (see Appendix A).

The binding sequences and the PRs will be represented in Boolean characters (see Appendix A). We define the effective lengths of the binding sequences represented in Boolean characters to be

$$l_m^{\text{eff}} = [I_m / \ln 2], \quad (2)$$

where the square brackets indicate integer part. The probability matrices are frequently reported in the databases on the assumption that they factorize site-wise, which is most probably not the case (Berg et al., 2004; Benos et al., 2002; Bilu and Barkai, 2005; O'Flanagan et al., 2005; Okuda et al., 2007), and the effective bitwise lengths are probably slightly overestimated as a result (see Section 3).

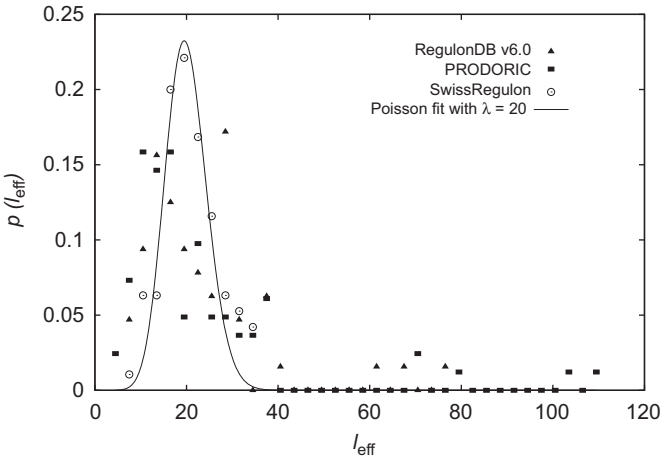


Fig. 4. Information content of the binding sequences. Effective binary length distribution for the binding sequences of *E. coli* extracted from three different databases: RegulonDB v6.0 (Gama-Castro et al., 2007) (triangles), PRODORIC (Münch et al., 2003) (squares), SwissRegulon (Pachkov et al., 2007) (circles). The solid line is the truncated Poisson distribution with mean = 20, and normalized over the finite range of the data points represented by circles, as obtained from the SwissRegulon data. The data points fall right on the curve.

2.2. Connection probabilities

Within our model, the probability of connection between two nodes can be estimated analytically. The probability of encountering a Boolean bs sequence of length l, associated with one node, within the PR sequence (of length $k \geq l$) of a second node, is given to a very good approximation (see Mungan et al., 2005, Fig. 2) by

$$p(l,k) \simeq 1-(1-2^{-l})^{k-l+1}. \quad (3)$$

For $l \gg 1$, this can be further approximated by $p(l,k) \sim (k-l+1)2^{-l}$. Thus the effective length distribution of the bs is highly relevant to the topological properties of the resulting interaction matrix,

with the connection probabilities depending exponentially on the bs lengths, and only linearly on the length of the PR.

The resulting network can be considered as a superposition of random networks, with connection probabilities which depend on the properties of different classes of nodes, here labeled by the lengths of the bs (if any) and PR sequences associated with them. For analytical results on how such a superposition can actually give power law or exponential distributions, the reader should consult Mungan et al. (2005) and Balcan and Erzan (2007). It should be noted that different distributions of sequence lengths give rise to markedly different behavior. The interaction network of a set of random sequences with an exponential length distribution was solved analytically by Mungan et al. (2005), and shown to exhibit an out-degree distribution with two different scaling regimes and a non-monotonic in-degree distribution with a Gaussian tail.

3. Extracting the model parameters from biological data

3.1. The GRN at the operon level

As stated above, to analyze and model the GRN at the operon level, we identify the nodes of the network with the operons. A TF coded by any one of the genes belonging to any TU within an operon contributes one or more out-going edges to the node associated with that operon. Conversely, any TF which binds a bs contained within a promoter region (PR) associated with a node will contribute an in-coming edge to that node, regardless of whether it regulates the operon itself, or a TU within the operon.

In the RegulonDB v.6.0 the number of TF coding operons are 159 out of a total of 2684, or about 5.9% of all operons. The *E. coli* genome is reported to have only one operon which codes two TFs, all the rest code one or zero. There are a couple of TFs which are complexes formed out of two proteins coded by genes in two different operons but we have neglected this fine detail. We picked 5.9% of the nodes on the average and assigned one binding sequence to each, indicating that they code candidate TFs. The distribution of the number of PRs (effectively, the number of TUs) associated with each operon is determined from the RegulonDB and is shown in Fig. 2.

3.2. Determining the PR lengths

Prokaryotes have a very high proportion of "coding" to intergenic material in their genomes, in comparison to eukaryotes. In *E. coli* coding material constitutes about 89% of the whole genome, while in *S. cereviciae* the ratio is just about inverted. The number of genes, on the other hand, are comparable in the two organisms; as a consequence, the average intergenic distance is much smaller on the prokaryotic genome, and distributed exponentially (Almirantis and Provata, 1999).

There is no clear cut prescription for the determination of the lengths of the PRs. We decided to focus on the distances (in number of base pairs, with the absolute value taken) between the start codons of the TUs and the binding site centers (bsc) recognized by the TFs regulating them (see Fig. 3). We believe this quantity is most clearly indicative of the length of the region in which the bs could possibly occur. The PR length distribution found in this way is not conditional on whether the regulated TU is buried inside an operon, or is located right at the beginning. By contrast, the intergenic distances found from the EcoGene database (Rudd, 2000), also displayed in Fig. 3, are larger for pairs of inter-operonic pairs of genes than they are for intra-operonic pairs (Okuda et al., 2007). The continuous line in Fig. 3, fitted to the relative frequency of bsc-to-start-codon distances taken from the RegulonDB, corresponds to an exponential distribution $p_{\mathrm{PR}}(l) \sim \exp(-bl)$ with $b = 0.0152 \pm 0.0007$ (see Table 1). In performing the actual simulations, the PR lengths were randomly selected from the empirical distribution of all the bsc-to-start-codon distances shown as diamonds in Fig. 3, uniformly shifted upwards by 9 base pairs to allow the shortest PRs to accommodate binding sequences of typical length. The mean of this shifted distribution is 91 bp, with a few datum points at distances as large as 2500 bp. The absolute range of the distribution is comparable with that for *S. cereviciae*, although there the distribution decays only as a power law (Almirantis and Provata, 1999; Balcan et al., 2007).

3.3. Probability matrices of the binding sequences

The most important problem was in determining the effective lengths (see Eqs. (1) and (2)) of the binding sequences. We have analyzed the *E. coli* data starting from version 5.6 of the RegulonDB and subsequently updated our data with versions 5.7, 5.8 and 6.0. In the successive updates of the data base, the most telling difference was in the small but extremely important (see Eq. (3)) upward shift of the minimum effective bs length appearing in RegulonDB v6.0. Moreover, inspection of the sets of sequences used for the generation of certain probability matrices in the RegulonDB revealed that, if the sequences were clustered into several distinct sets (rather than being considered as variants of the same bs) they could be better aligned. This would result in several probability matrices with fewer columns but with larger matrix elements, leading to larger relative information content (Eq. (1)) for several *distinct* binding sequences (Fu and Weng, 2004).

The literature search for an alternative source for weight matrices yielded the SwissRegulon (Pachkov et al., 2007) and PRODORIC (v2.0) (Münch et al., 2003). The effective bs length distributions obtained via Eqs. (1) and (2) from these three databases are displayed in Fig. 4. Data was binned into intervals of size three.

The weight matrices (see Appendix A) quoted in the SwissRegulon database were obtained by Pachkov et al. (2007) by re-clustering and re-aligning the binding site data from RegulonDB, using the clustering algorithm "PROCSE" of van Nimwegen et al. (2002). In Fig. 4, the computed length distribution in bits is indicated by the open circles. This distribution is clearly much smoother than the others. The lower limit of its range agrees with that of the RegulonDB v6.0 and it has a mean of 20 bits. We have superposed on this set of points a truncated Poisson distribution with the same mean, normalized over their finite range. It can be seen that most of the points fall right on the curve, which interpolates over the gaps in the data. In our simulations we have randomly chosen our bs lengths from this smoothed, Poissonian distribution.

We have found from the SwissRegulon (Pachkov et al., 2007) data that the number of binding sequences per TF obeys the distribution given in Table 1. Each bs contributes a single length to the empirical length distribution, regardless of the number of other binding sites a TF may have. In constructing the model genome each bs length is drawn independently from the truncated Poisson distribution shown in Fig. 4.

The parameters for the GRN network of *E. coli* are shown in Table 1.

4. Simulation results

We chose the size of the model and empirical networks to be comparable, even though the absolute size of the network can be

normalized away for all of the statistical graph properties discussed below, except for the k-core analysis. The empirical network has 683 nodes (out of 2684) that have at least one edge connecting them to some other node. For the model networks we started with 2684 nodes as in the empirical network; the number of connected nodes range between 575 and 1372, with a mean of 982 and standard deviation 162. The model networks have on the order of 2000 edges, whereas the empirical network has about 1300.

In our simulations we randomly pick 5.9% of the nodes to be candidates for TF coding nodes; however, only about half of them actually connect (see Table 1) and we end up with about half the number of TFs as the empirical network.[3]

In order to study the statistical properties of our model, we have performed two sets of simulations. In the first, we have computed the topological properties of 100 realizations of the model. Below, in Figs. 5–11, we display the scatter plots we thus obtained. The properties computed for the empirical network are superposed on these simulation results. We have also performed k-core analysis of the empirical and model networks. Finally, we have extracted statistics of the motifs (Milo et al., 2002, 2004) encountered in both the empirical and model networks.

In the second set of simulations, we have randomized the empirical and model networks while keeping the in- and out-degrees of each node fixed. Comparisons of the topological properties of randomized versions of the empirical and model networks are available in the Supplementary Material. Not surprisingly, the empirical network moves closer to the null model under randomization, while the 100 randomized versions of one randomly picked model network generate another, statistically identical realization of the original ensemble. As remarked below, the clustering coefficient and motif statistics of the empirical graph are most strongly affected by the randomization, while the degree–degree correlation function is almost left invariant.

The simulation code for generating the adjacency matrices and computing the graph properties was written in C^{++}. A reasonably annotated version is available upon request. The random number generator we used was a C^{++} implementation by Richard Wagner[4] of the Mersenne Twister (Matsumoto and Nishimura, 1998).

4.1. Degree distributions

A basic tool of graph analysis is the degree distribution of the nodes (Albert and Barabasi, 2002; Dorogovtsev and Mendes, 2002). The degree of a node is the total number of nodes to which it is connected, by one or more directed or undirected edges (see Appendix B).

In Fig. 5 we present the results for the degree distribution $p(k)$, on a log–log plot. The emerging picture is rather similar to that of the GRN of yeast (Balcan et al., 2007). It is very gratifying that here too, the empirical data points (red disks) fall right on top of the scatter of points from 100 independent realizations of the model. For each realization we have picked the relevant sequence lengths, TF numbers and bs numbers, from the appropriate distributions and generated the random sequences independently. In Fig. 6 we have plotted the averages over the 100 realizations. The error bars correspond to one standard deviation in each case. In this semi-log plot, one may discern that the initial part of the $p(k)$ curve is exponential in both the model and empirical networks; the difference between the slopes of the fitted curves is somewhat in excess of the error bars (see Table 2).

By plotting the in- and out-degree distributions separately one can see that the small degree part of the distribution in Fig. 5 comes essentially from the in-degrees, while the larger degrees are contributed by the out-degrees. In Fig. 7, the semi-log plot for the in-degree, and in Fig. 8, the log–log plot for the out-degree distribution, one has strong qualitative agreement between the

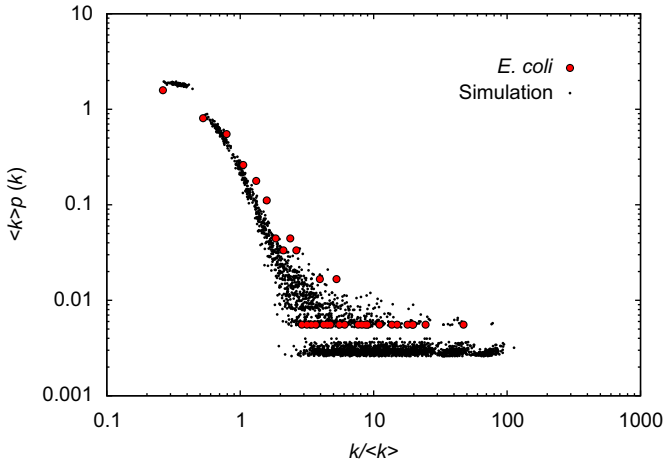


Fig. 5. Degree distribution. Plot of the degree distribution of the transcriptional gene regulatory network (GRN) of *E. coli* extracted from the RegulonDB v6.0 (red disks), compared with the scatter plot (black points) of the degree distribution of 100 independent realizations of the model network, in which we have included only those nodes with degree greater than zero. To account for the fluctuations in the network size, the horizontal axis has been scaled with the average degree per node $\langle k\rangle$, an extensive quantity. The probability $p(k)$ (the vertical axis) has been multiplied by the average degree in anticipation of an exponential fit to the distribution, shown in Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

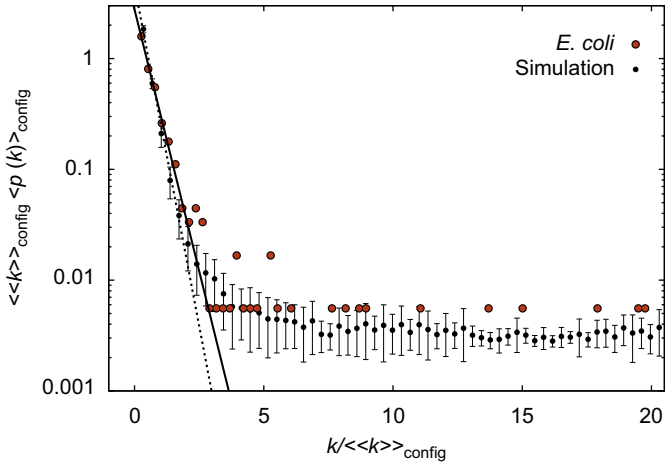


Fig. 6. Exponential fit to the degree distribution. Semi-logarithmic plot of the degree distribution of the transcriptional gene regulatory network (GRN) of *E. coli* from the RegulonDB v6.0, compared with the degree distribution of the model network, averaged over 100 realizations (black discs). The configurational average is taken over the set of independent realizations of the model network. Error bars stand for one standard deviation. Numerical results for the fit to $\sim\exp(-k/\xi)$ of the initial range of the empirical and model distributions are given in Table 2. The vertical axis has been multiplied by the degree averaged over the nodes (and for the model networks, also the realizations) in order to scale away the fit parameters ξ.

[3] We have repeated the calculations with twice the number of nodes that are TF candidates, ending up with approximately the empirical number of TFs which actually bind other nodes. The statistical distributions characterizing the network, normalized as they are by the size of the network in each case, remain the same. There is a slight reduction in the scatter due to the larger network. Data is available upon request.

[4] http://www-personal.umich.edu/~wagnerr/MersenneTwister.html

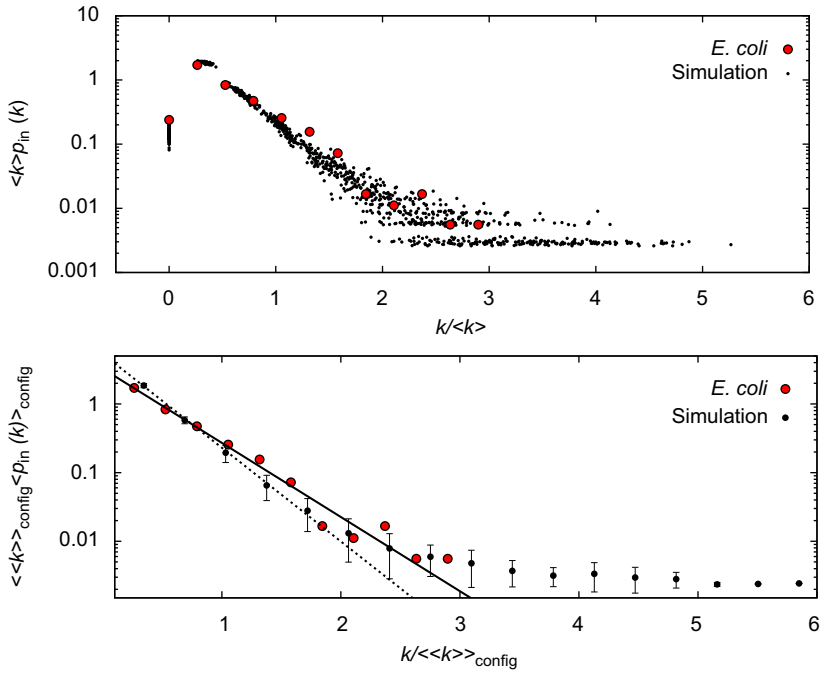


Fig. 7. The in-degree distribution. Semi-logarithmic plot of the in-degree distribution of the transcriptional gene regulatory network (GRN) of *E. coli*, extracted from the RegulonDB v6.0, compared with the in-degree distribution of the model network. The lower panel has been averaged over 100 realizations of the model. Error bars stand for one standard deviation. Numerical results for the characteristic in-degree, ξ_{in}, found from fitting $\langle p(k_{in})\rangle_{config} \sim \exp(-k/\xi_{in})$ to the initial ranges of the distributions, are given in Table 2. The datum point with zero in-degree and the flat tail is truncated in the lower panel.

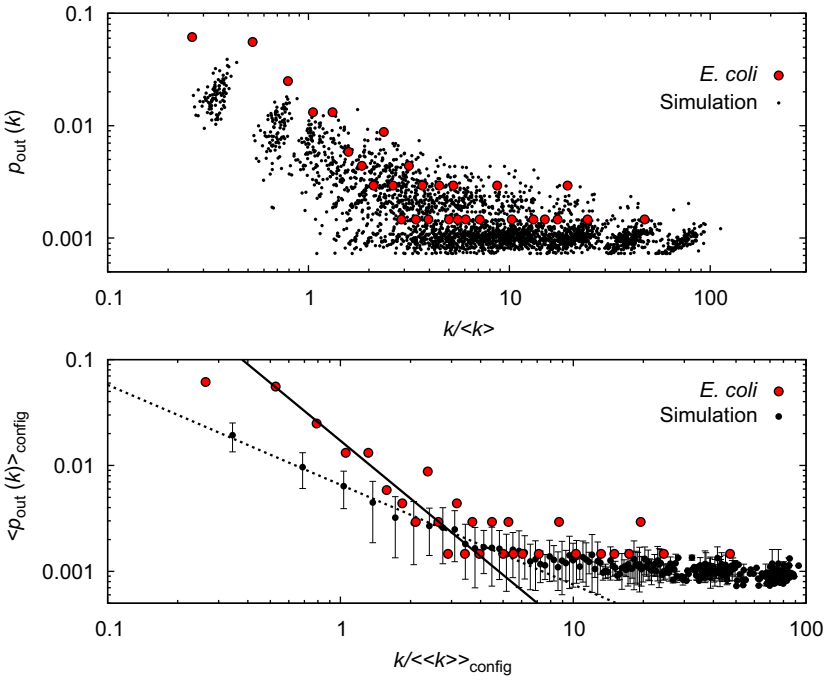


Fig. 8. The out-degree distribution. Both the scatter plot and the averages over 100 realizations of the simulation (black discs) are shown, compared with empirical distribution (red discs). Error bars in the lower panel stand for one standard deviation. Values of the exponents for the fits to $\langle p(k)\rangle_{config} \sim k^{-\gamma}$ are given in Table 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

model and empirical networks, with, however, a quantitative difference in both the average in-degree per node and the power in the initial, scaling range of the out-degree (also see Table 2). In the clustering of the model points on the far right side of the out-degree distribution, we see a faint remnant of the discrete peaks which would be there for a much larger network (Mungan et al., 2005). Within the model, these peaks arise from the shortest bs sequences which are most frequently to be encountered in the longer PRs. As one goes to smaller degrees, the peaks merge and give rise to a continuous distribution.

We find that the degree distribution of the empirical gene regulatory network of *E. coli*, as well as that of yeast (Balcan et al.,

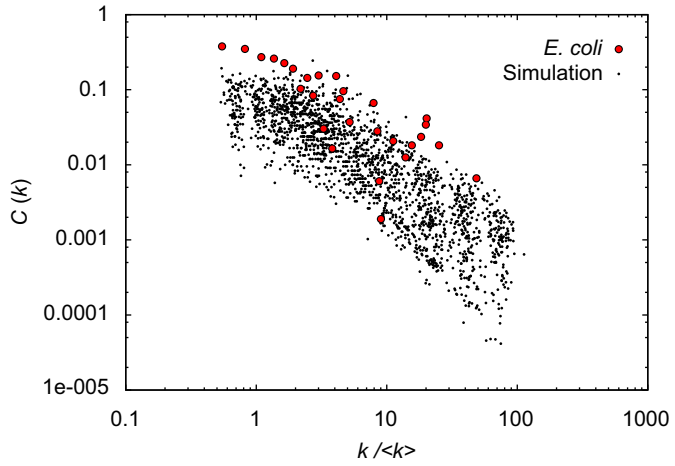


Fig. 9. Clustering coefficient $C(k)$. The *E. coli* data from the RegulonDB v6.0 is shown as red discs. The scatter of black points corresponds to 100 realizations of the model network. All self-interactions have been removed from the network before calculating the clustering coefficient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

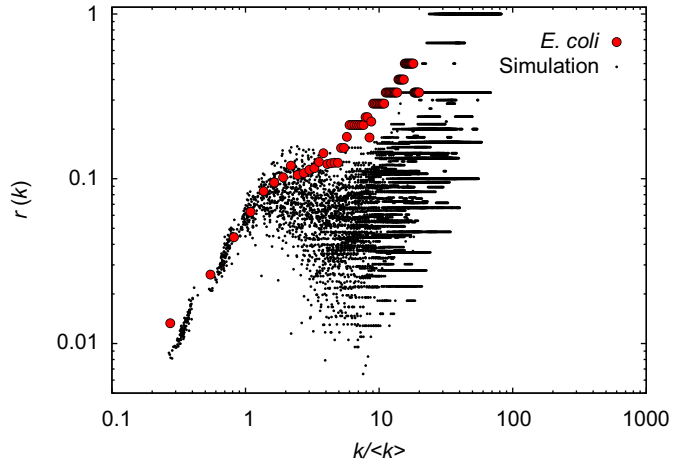


Fig. 10. The rich-club coefficient $r(k)$. The *E. coli* data from the RegulonDB v6.0 are shown as red discs. The scatter of black points corresponds to 100 realizations of the model network. All self-interactions have been removed from the network before calculating $r(k)$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

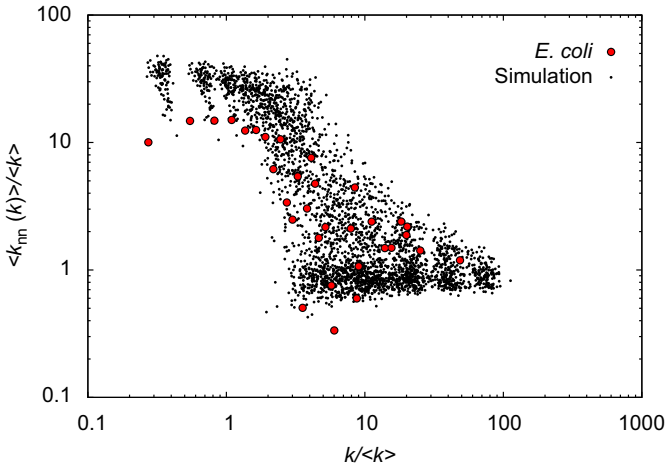


Fig. 11. The degree–degree correlation function. The expected degree of nodes neighboring a degree-k-node is denoted by $k_{nn}(k)$. All self-interactions have been removed from the network before calculatingthe correlation function. Since this correlation function is an extensive quantity, both the vertical and horizontal axis have been normalized by the average degree. The *E. coli* data (red disks) is from the RegulonDB v6.0. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Comparison of the average degrees and degree distributions of the empirical and model networks.

	E. coli	**Model**
Average degree $\langle k \rangle$	3.796	2.906
ξ	0.46 ± 0.01	0.35 ± 0.03
Average in-degree $\langle k_{in} \rangle$	1.977	1.454
ξ_{in}	1.81 ± 0.09	0.94 ± 0.03
Average out-degree $\langle k_{out} \rangle$	1.977	1.454
γ	0.40 ± 0.01	0.32 ± 0.01

Only those nodes (operons) with non-zero degree have been included in the network statistics. The parameters ξ, ξ_{in} refer to the exponential fits to the degree and the in-degree distributions (see Figs. 6 and 7). The out-degree distribution has a putative power law behavior $k_{out}^{-\gamma}$ for relatively small degrees (see Fig. 8). These numbers are only provided for ease of quantitative comparison of the model and empirical networks. The power law fits are valid only within a range of about a decade and do not represent any claims that the respective networks are scale free. See text, Section 4.1.

2007), are much richer than so far suspected. There were early claims (Barabasi and Oltvai, 2004; Bergmann et al., 2004; Dobrin et al., 2004; Vazquez et al., 2004) that gene regulatory networks were scale free, with the degree distribution decaying as a universal power law, $p(k) \sim k^{-\gamma}$, where $\gamma \simeq 2$, with, perhaps an exponential cutoff. For this operon-level analysis we find no evidence for the power law with an exponential cutoff claimed for the overall degree distribution, and we are not aware of convincing arguments indicating that a process of preferential attachment (Barabasi and Albert, 1999) is in operation. On the other hand, our model reproduces the exponential decay of the in-degree distribution, also noted by Guelzim et al. (2002) for yeast. For ease of comparison, it may be useful to supply some numbers. The putative power law behavior of the out-degree distribution has an exponent of $\gamma \simeq 0.3$, over a very limited range of about one decade. For the empirical network, a comparable fit within an interval of again about a decade yields 0.4 (see Table 2). Note that both these numbers are much smaller than 3, expected for the "preferential attachment" model.

4.2. Higher order correlations

Other quantities of interest, which reflect higher order correlations between the nodes than just pair-wise connectivity, are the clustering coefficient $C(k)$ as a function of the degree (Bollobás, 1998; Albert and Barabasi, 2002; Dorogovtsev and Mendes, 2002), the correlations between the degrees of neighboring nodes (Colizza et al., 2005) and the rich-club coefficient (Zhou and Mondragon, 2003; Colizza et al., 2006). These quantities are defined in Appendix B. Before calculating these three quantities, we have removed the self-interactions from both the empirical and model networks. It should be noted that the empirical network has 93 self-interactions, while the model networks have between 0 or 1 self interaction per realization. Since the bs and PR sequences associated with the nodes have been generated independently, this null model does not incorporate the abundance of self-regulatory interactions in the prokaryotic genome (Lynch, 2007).

The clustering coefficient $C(k)$ is shown in Fig. 9. The data points for the *E. coli* network follow the same qualitative trend as those for the model network; however, they are systematically

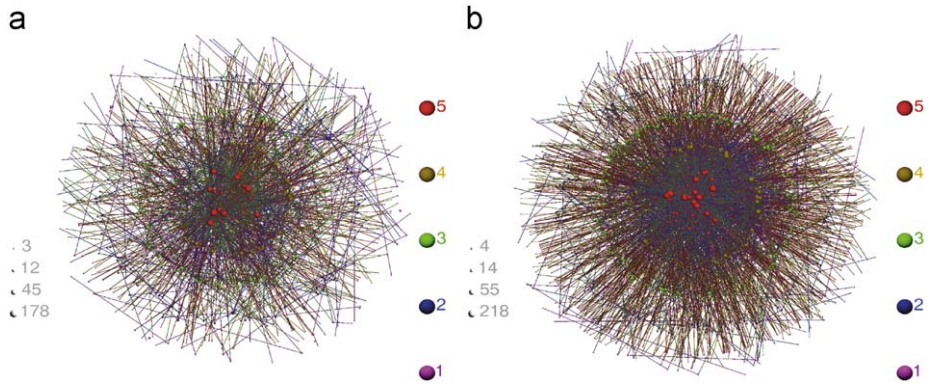


Fig. 12. The *k*-core analysis. (a) The empirical *E. coli* gene regulatory network from the RegulonDB, and (b) a model network (a typical realization, number 48) from our ensemble of model networks. The *k*-cores have been visualized using the visualization tool developed by Alvarez-Hamelin et al. (2005), LaNet-vi, which is available online at (http://xavier.informatics.indiana.edu/lanet-vi/). The color code indicated on the right corresponds to the shell number (coreness), while the size of each ball is proportional to the degree of the corresponding node. A sample of values are given on the left, the last one being the largest degree on the network. The thickness of the shells corresponds to the spread in the coreness of the nodes to which members of a given shell are connected.

shifted to higher values, more markedly so for small k values. It is clear on the log–log plot that the curves followed by both the empirical and the model data points deviate downwards from a straight line and therefore the decay is faster than a power law, contrary to previous claims to this effect by Vazquez et al. (2004).

The simulation results for the rich-club coefficient $r(k)$ (see Fig. 10) show a more pronounced non-monotonicity where the empirical network displays a shoulder. There is a shift to higher values in the high-degree end, indicating a greater incidence of inter-connections between high-degree nodes than expected on the basis of uncorrelated binding sequences and PRs. (Both of these effects also show up in the motif statistics (Milo et al., 2002, 2004), and it will be further discussed below.) Colizza et al. (2006) find that the rich club coefficient displays a monotonic increase with the degree for real-world networks such as the internet, air transportation networks, and scientific collaborations, as well as random graphs and the scale free networks yielded by the preferential attachment model of Barabasi and Albert (1999). It is interesting that the only departure from this behavior is a small non-monotonicity, or shoulder, for the protein–protein interaction network, which is a constraint-satisfaction type network, like the gene regulatory network we are considering here.

The average degree of nodes that are nearest neighbor to degree-k nodes (the so-called $k-k$ correlation, or $k_{nn}(k)$), is plotted in Fig. 11. Here again, one has close qualitative agreement between the GRN of *E. coli* and the set of model networks. However, in this case the data points are shifted downwards by almost a factor of four in the small degree region, indicating that the average degree of neighbors of low-degree nodes is four times smaller than what one would expect on the basis of the model. This fact is also reflected in the k-core analysis of the network; see next subsection. In the *E.coli* GRN, the low in-degree nodes are generally those with high out-degrees, regulating a large number of TUs, which are not themselves regulating. Thus their neighbors will have degrees that are below the average. In the model, for any TF-coding node the PR lengths and the length of the bs associated with the TF are chosen independently. Therefore, there is no correlation between the in-degree and out-degree of a node.

4.3. Hierarchical structure

A different way of analyzing the graph properties is the k-core analysis (Bollobás, 1998). The iterative method for determining the different layers, or shells, is described in Appendix B. The visualization (Alvarez-Hamelin et al., 2005) of the different k-shells is a very concise way to display the hierarchical organization of the graph. In Fig. 12 we show the k-core analysis of the GRN of *E. coli* at the operon level and a representative realization of our model network. Both have five shells. For this application we chose the overall fraction of potential regulatory nodes (i.e., those coding transcription factors) to be such that, the number of actually connected regulatory nodes was equal to the empirical number, 159 (see Table 1, main text, and Supplement, Table 1).

The similarity between the k-core visualizations for the empirical and model networks is very close, with both showing a very marked hierarchical organization. All the nodes of highest degree (hubs) reside in the innermost core of the graph and are highly connected amongst each other. Nodes of different coreness (residing in different shells) are preferentially connected directly to the innermost core, with this tendency being more pronounced in the model network. This structure has also been found in the *E. coli* GRN at the gene level. See Ma et al. (2004).

In the Supplement we also provide plots of the shell populations and the connectivity between nodes belonging to different shells. It is instructive to contrast the k-core analysis of the rather similar yeast GRN with that of Barabasi-Albert scale free graphs of the same size. The latter yield much fewer shells (only three compared to nine for the empirical and model networks) and no hierarchical organization, with the nodes in different shells being connected to each other seemingly at random (Fig.1, Supporting Text 2, Balcan et al., 2007).

4.4. Motif statistics

Finally let us consider the motif statistics, reported in Figs. 13 and 14. We have used the motif finder program "FANMOD" (freely available online at http://www.minet.uni-jena.de/~wernicke/motifs/) developed by Wernicke and Rasche (2006). We see that in the model network bi-directional edges are totally absent, so that a number of motifs present in the *E. coli* GRN are simply ruled out. However, we may note that although the absolute values of the Z-scores for the motifs in a randomly selected realization of the model network are smaller than the values encountered in the real network, they do consistently have the same sign, i.e., they depart from the randomized versions in the same direction as the empirical network. These results may be compared with the motif statistics at the gene level reported by Ma et al. (2004).

Size-3 motifs

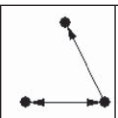
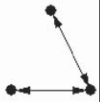
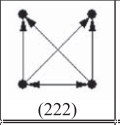
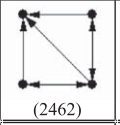
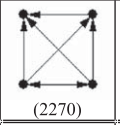
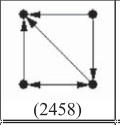
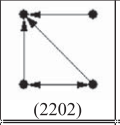
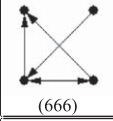
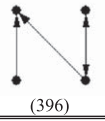
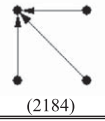
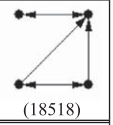
Motif Type / Source	 (46)	 (14)	 (6)	 (36)	 (38)	 (12)	 (166)	 (164)	 (78)
RegulonDB	0.36732 (16.301)	1.0329 (-15.981)	92.568 (15.664)	2.7342 (-12.201)	0.63246 (6.7199)	2.6099 (-6.2338)	0.0055236 (1.5481)	0.046951 (-0.84301)	0.0027618 (0.28815)
Random	0.053932 (0.019225)	1.6352 (0.037689)	91.722 (0.054002)	3.3329 (0.049074)	0.29975 (0.049511)	2.8976 (0.046151)	0.0020076 (0.0022712)	0.051191 (0.0050301)	0.0025905 (5.946e-004)
Model	*Not found*	*Not found*	95.499 (1.4397)	1.8346 (-1.4494)	0.14321 (1.4492)	2.5236 (-1.4336)	*Not found*	*Not found*	*Not found*
Random			95.462 (0.02512)	1.8732 (0.026608)	0.10307 (0.027699)	2.5612 (0.026234)			

Fig. 13. Network motifs. Percentages of size-3 motifs found in the *E. coli* network generated using data from RegulonDB v6.0 and in one realization of the model. Corresponding values obtained from 1000 randomized networks are also given. Numbers in the parentheses for the randomized networks stand for the standard deviations whereas the ones for the original network are the *Z*-scores for that motif. The *Z*-score is defined as the difference between the value for the original network and the mean value over 1000 randomizations divided by the standard deviation. Motifs are ordered in decreasing value of the *Z*-score for the RegulonDB network from left to right. The numbers below the graphs identify the motifs.

Size-4 motifs (only the ones with a *Z*-score above three in the RegulonDB network are listed)

Motif Type / Source	 (222)	 (2462)	 (2270)	 (2458)	 (2202)	 (666)	 (396)	 (2184)	 (18518)
RegulonDB	0.21002 (38.85)	0.00020496 (21.559)	0.0012981 (20.103)	0.0025962 (12.31)	0.015509 (10.935)	0.00936 (9.0363)	0.026645 (-8.518)	0.047961 (-8.0081)	0.00027328 (6.7408)
Random	0.0057148 (0.005259)	1.3486e-006 (9.445e-006)	3.2579e-005 (6.295e-005)	0.00028541 (1.877e-004)	0.0030946 (0.001135)	0.0020388 (8.102e-004)	0.062474 (0.004206)	0.074015 (0.003253)	1.5618e-005 (3.823e-005)
Model	*Not found*	*Not found*	*Not found*	*Not found*	*Not found*	*Not found*	*Not found*	0.0398 (-1.2375)	*Not found*
Random								0.042489 (0.002173)	

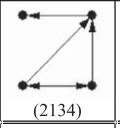
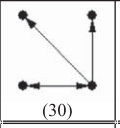
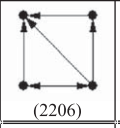
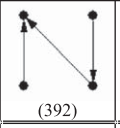
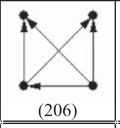
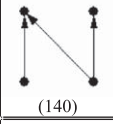
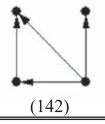
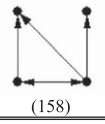
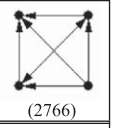
Motif Type / Source	 (2134)	 (30)	 (2206)	 (392)	 (206)	 (140)	 (142)	 (158)	 (2766)
RegulonDB	0.0084718 (6.5921)	1.1248 (6.5352)	0.001708 (4.7971)	0.074743 (-4.6899)	0.058756 (4.3935)	6.0642 (-3.8253)	1.4287 (3.6374)	0.47545 (3.1671)	0.00013664 (3.0439)
Random	0.0011244 (0.001115)	1.8018 (0.1036)	0.00039949 (2.728e-004)	0.11711 (0.009033)	0.019001 (0.009049)	7.1049 (0.27204)	0.77002 (0.18107)	0.18013 (0.093247)	1.2782e-005 (4.069e-005)
Model	*Not found*	*Not found*	*Not found*	0.0597 (-1.6417)	0.022514 (4.6942)	5.5056 (-1.9132)	0.39466 (1.8919)	*Not found*	*Not found*
Random				0.071442 (0.007152)	0.0055961 (3.604e-003)	5.9085 (0.21059)	0.24436 (0.079445)		

Fig. 14. The statistics for motifs of size 4.

5. Discussion

In this paper we have presented a null model for a complex biological system. We have provided a detailed comparison of the model with the actual biological network, the transcriptional gene regulatory network of *E. coli*. We believe this study contributes to an understanding of how and to what extent such structures might emerge from combinatoric considerations alone.

Our analysis of the most up to date data on the transcriptional gene regulatory network of *E. coli* shows that the somewhat simplified picture of scale free graphs (Barabasi and Oltvai, 2004; Dobrin et al., 2004; Vazquez et al., 2004; Bergmann et al., 2004), with exponents $2<\gamma<3$, and modeled by a "preferential attachment" growth rule (Barabasi and Albert, 1999), is not applicable for the statistical features of the *E. coli* GRN at the operon level. The success of our combinatoric model lies in its detailed reproduction of "non-universal" details and trends in the statistical features of the empirical network.

In Section 4 we have shown that the degree distribution of the *E. coli* network is dominated by the in-degree for small degrees, and has an exponential decay in this region. This behavior is faithfully reproduced by the model network. The model out-degree distribution exhibits a putative power law behavior over a small interval of about a decade, as do the *E. coli* data, with a small power comparable to that of the empirical network (see numerical values in Table 2). The long flat tail of the out-degree extending beyond the initial scaling region is present in both the model and the empirical graphs and is not just scatter due to the

paucity of data, as can be seen from analytic computations (Mungan et al., 2005; Balcan and Erzan, 2007), albeit for different sequence length distributions. We find that the clustering coefficient does not follow a power law as claimed by Vazquez et al. (2004), while our model is able to reproduce the form of the variation with the degree. The rich club coefficient displays the same overall increase, as well as a marked non-monotonicity as a function of $k/\langle k\rangle$, at the same values where the graph for the empirical network displays a shoulder. The k-core analysis and the k-shell population distribution as a function of the coreness (see Supplementary material) are in agreement with the main features of the empirical network.

It should be noted that the data that goes into our model network is of two kinds. (i) The number of operons, TFs, and the number of different bs bound by the same TF. These determine the total size of the network and set lower and upper bounds on the total number of edges, but cannot in any way lead to even a qualitative prediction regarding the degree distributions or the other topological properties of the network. (ii) The distribution of the information content of the connections, an attribute superficially having nothing to do with the topology. Our model provides a theoretical framework within which the second type of data is used to predict the specificity of the connections, and thereby the statistics of the network topology. It should be noted that the range of possibilities given just the first kind of data are nearly infinite in the absence of a model, and therefore even a qualitative agreement between the empirical and model networks is an important achievement.

5.1. Quality of the data and the predictive power of the model

The second point we would like to make is that the quantitative agreement between the empirical GRN and our model networks improved steadily with the discovery of larger and larger numbers of regulatory interactions in the *E. coli* genome. It can be seen from Fig. 4 that there is rather poor agreement between different data bases regarding the effective binary length distribution of the binding sequences. The crucial increase in the minimum effective information content of the binding sequences reported in the successive versions of RegulonDB (starting from v5.6 (Salgado et al., 2006a)), and the improved distribution we derived from the SwissRegulon database (Pachkov et al., 2007), resulted in a radical improvement in the agreement between the model networks and the *E. coli* GRN. Thus, we may say that given the correct bs and PR length distributions, the model is able not just to mimic but to virtually *predict* qualitative features of the *E. coli* GRN as reported in the RegulonDB v6.0.

5.2. Effects that have been neglected

A number of possible reasons can be cited for the small but persistent difference between the distribution of empirical in-degrees and those estimated from the model network.

A high degree of overlap is found between consensus sequences in the relatively short promoter regions of *E. coli*, leading us to conjecture that even if more than one interaction is allowed in principle, only one of them will be realized at any given time. "Transcriptional interference" (Shearwin et al., 2005), where interference between RNA polymerase binding two close-by sites inhibits transcription of one or both of the TUs, has recently been studied and modeled (Dodd et al., 2007; Sneppen et al., 2005). Such effects can have further consequences for the reduction of the actual regulatory interactions from those that are possible purely on the basis of combinatoric arguments.

Several workers (Buldyrev et al., 1995; Kugiumtzis and Provata, 2004) have claimed that correlations within intergenic regions lead to reduced information content (and effective bitwise length) of PRs, by several percent. This would reduce the real connectivity of the actual networks to below what we conjecture on the basis of random PR sequences.

The binding sequences we obtained from the RegulonDB were slightly anti-correlated on average. We define the average distance per nucleotide between pairs of binding sequences of the same length l as

$$h = 1-(1/4)\sum_{l}[|S_l|l]^{-1}\sum_{\mu,\nu\in S_l}\sum_{i=1}^{l}\sum_{j=1}^{4}p_{ij}^{(\mu)}p_{ij}^{(\nu)}, \tag{4}$$

where S_l is the set of binding sequences of length l, and $|S_l|$ is the size of this set; μ and ν indicate different binding sequences within such a set, and $p_{ij}^{(\mu)}$ is the probability matrix for this binding sequence. Had the binding sequences been totally random, with the probabilities for the bases A,T,C,G given by $\frac{1}{3},\frac{1}{3},\frac{1}{6},\frac{1}{6}$, we would have gotten 0.728 for h, whereas from the RegulonDB(v5.7) we found 0.737, i.e., the binding sequences are farther from each other on the average than random sequences, by 1%. Thus, the probability for encountering overlapping binding sequences within a random PR is actually lower than had the former been random, but this is a very small effect which is below noise level in the present discussion. (Note that, in the absence of joint probabilities for the occurrence of different nucleotides at given sites of distinct binding sequences, the mutual information between them, constructed from just the probability matrix, is identically zero.)

Bilu and Barkai (2005) report that in those cases where more than one TF is binding a PR region, a lower specificity is tolerated, i.e., the binding sequences in this region are "fuzzier". This means that the effective lengths of the binding sequences sought out by the same TF may actually vary between different regions of the genome, an effect we have not taken into account in this model. By keeping the effective length of the consensus sequences fixed, independently of the length of the PR in which they are to be sought, we slightly disadvantage the binding probabilities at the shorter PRs compared to what seems to be observed. This effect, however, is of the same order (and opposite sign) as that which would be induced by the anti-correlation between the binding sequences, an effect which we ignore. We believe these two small corrections effectively cancel each other out and that we are not in error in neglecting both of them.

5.3. Evolution of correlations

Above all, it is necessary to understand that certain features of the empirical network could never be reproduced by such a naive null-model as the one we propose. We have already mentioned the absence of self-interactions in the model networks, where the protein product of the gene binds its PR and regulates its own transcription. Besides these, there are highly conserved, very special regulatory sub-graphs which, say, include regulatory nodes with extremely large number of connections, even though the binding sequence which they recognize is highly specific, requiring the satisfaction of a very large number of constraints. The correlations between connections, embodied especially in the clustering coefficient and the motif statistics (Milo et al., 2002, 2004) of the network, are other features which are not included in our model, where the assignments of all the sequences associated with the nodes are made independently. It is quite possible, that in the course of the evolution of the GRN, certain nodes with a high out-degree, regulating a relatively large number of TUs, were selected from among those having small in-degrees, introducing a

negative correlation between these quantities and leading to the observed discrepancies.

We believe that instead of comparing the empirical motif statistics with those of purely random networks, it is more meaningful to compare them with the present null model. The most striking feature of the motif statistics of the *E. coli* GRN is the high incidence of bi-directional interactions, giving rise to motifs with the highest *Z*-scores that can be seen in Figs. 13 and 14. Such bi-directional interactions are in fact present in our model but to a much smaller extent than in the empirical network. Note, in Table 2, that the average total degree is very slightly less than the sum of the in- and out-degrees for the model networks, while it is markedly different from this sum for the empirical network. The simplest, and most likely (Berg et al., 2004; Babu et al., 2006) mechanism to give rise to such interactions is the duplication of TF-coding genes and their promoter regions, a feature which is not present in this model, but which can easily be built in and has already been considered by Sengun and Erzan (2006). Another feature which could very easily be incorporated into the model is homologies between the TFs leading to similarities between binding sequences. The high incidence of the feed-forward loop (motif number 36, a high *Z*-score motif) and the large rich-club coefficient would be accounted for if there were a high overlap between the binding sequences of TFs which regulate each other in a cascade. Further work on the evolution of content-based model genetic networks by non-adaptive processes (Lynch, 2007) is in progress.

5.4. Comparison with the yeast GRN

It is instructive to compare our findings for *E. coli* with those for *S. cereviciae* (yeast) (Balcan et al., 2007). The two genomes differ most markedly in the distribution of the lengths of the promoter regions, with, however, a rather similar distribution of effective lengths for the binding sites. Comparing the GRN of *E. coli* with that of *S. cereviciae* one finds great qualitative similarities between the two, with an essentially exponential distribution of the in-degree, a rather scattered out-degree distribution suggesting a power-law distribution, and clustering coefficients, degree–degree correlations and rich-club coefficients that qualitatively look very similar.

Comparing the *E. coli* and yeast (Balcan et al., 2007) networks with respect to their *k*-core decomposition is also interesting. A sharp difference between the *E. coli* and yeast GRNs shows up in the shell population distribution as a function of the coreness: In the case of yeast, the shell population decreases linearly with coreness, whereas for *E. coli* the decrease is exponential. The model networks mimic these respective behaviors perfectly in both cases. For *E. coli*, the very tightly hierarchical connectivity of the model network, with edges going almost strictly up and down the coreness hierarchy, is disrupted in the empirical network of *E. coli* to a greater extent than is the case for yeast. Although the exponential growth trend in the connections to the high coreness nodes is common to both these organisms, a greater incidence of in-shell (transverse) connections are visible in Fig. 12 than in the corresponding Fig. 2 of Balcan et al. (2007). This behavior is graphically illustrated in the Supplementary Fig. 8, where the empirical graph deviates from the exponential growth of the connectivity to higher coreness nodes, and shows an excess of connections to nodes of low coreness. This agrees with a well known feature of the prokaryotic genome where there is an abundance of small regulatory loops and self-regulatory interactions (Lynch, 2007).

The quantitative agreement between the *E. coli* genome and our model networks is overall less than the corresponding agreement found for yeast (Balcan et al., 2007). This could be ascribed to the absence of any fitting parameters in the present study, while in the case of yeast, the (unknown) exponent of the length distributions of the promoter regions was optimized to get the best fits. However, the qualitative dependence of various network features on this number was very weak. We conjecture that selective pressures on the very compact prokaryotic genome might have caused greater departures from purely combinatoric features in the *E. coli*, than is the case for the yeast genome.

5.5. Generic features coming from a large number of independent constraints

In this paper we have constructed a model of the prokaryotic GRN. We should recall that we do not intend to model the GRN on a node to node basis, but only with respect to its global statistical properties. We have checked whether the number of transcription units (TUs) which a TF regulates is correlated with the information content of its binding sequence, and found no correlation at all, for any of the data bases used. (No such one-to-one correspondence was found in the case of *S. cereviciae* either (Balcan et al., 2007).) Thus the high degree to which our model predictions are borne out points to a phenomenon of a more fundamental nature (Gerland et al., 2002). It seems to imply that the distribution of the specificity of the connections seems to arise independently of the actual lengths of the binding sequences recognized by the TFs, but nevertheless has essentially the same truncated Poisson distribution as the latter.

The clue to this convergence lies in considering an arbitrary number m of independent conditions for, say, a genomic interaction to be established. The probability $\mathcal{P}_m$ for the satisfaction of all m of these conditions will be a product of the individual probabilities, viz.,

$$\mathcal{P}_m = \prod_{i=1}^{m} p_i. \tag{5}$$

Each of the probabilities p_i can be expressed as $p_i = 2^{-\alpha_i}$, where $\alpha_i = -\ln_2(p_i)$. Thus, $\mathcal{P}_m = 2^{-\lambda}$ where $\lambda = \sum_i^m \alpha_i$. Even if the α_i are not identically distributed, as long as the mean and variance exists for each α_i, and, for example, Lyapunov's condition (see any standard text on probability theory, e.g., Koralov and Sinai, 2007) is fulfilled, we can avail ourselves of the central limit theorem, to claim that for m sufficiently large, λ is Gaussian distributed around $\sum_i^m \langle \alpha_i \rangle$ with variance $\sum_i^m \sigma_i^2$, summed over the individual variances. At the level of precision of the fit in Fig. 4, this distribution would be indistinguishable from a Poissonian, especially for $\langle \lambda \rangle$ as large as 20, as found here. This argument is in fact very general and need not apply only to the establishment of genomic interactions. It has to do with the well-known fact that the distribution of probabilities for the satisfaction of a large number of independent conditions is log-normally distributed.

Acknowledgments

It is a pleasure to thank Volkan Sevim for a critical reading of our manuscript. We thank Meltem Sevgi for having obtained an early version of the PR-length distribution in the initial stages of this work. Berkin Malkoç acknowledges support from the Scientific and Technological Research Institute of Turkey (TÜBİTAK) National Scholarship Program for PhD Students. Ayşe Erzan would like to acknowledge partial support from the Turkish Academy of Sciences.

Appendix A. Information content of sequences

The methodology here closely follows that presented in Balcan et al. (2007). We define the information content of an ensemble of sequences of N letters drawn from an alphabet of r letters as (Shannon, 1949; Avery, 2003),

$$I \equiv \sum_{i=1}^{N}\sum_{j=1}^{r} p_{ij}\ln p_{ij}, \tag{A.1}$$

where p_{ij} is the probability of encountering the jth letter in the ith position. Note that this is a negative quantity, and therefore is sometimes defined with an overall $(-)$ sign out front, in analogy with the thermodynamic entropy. It makes sense to subtract from this expression the information content of a sequence of the same length but with letters drawn at random, namely $I^{(0)} \equiv N\sum_{j=1}^{r} p_j^{(0)}\ln p_j^{(0)}$, where $p_j^{(0)}$ are the "background" probabilities of the different letters. This gives an information content that is relative to the random case. We have taken the background probabilities to be uniformly equal to $1/r$ in Eq. (1).

A "bit" is a binary variable having the values of 0 and 1, and is universally accepted as the basic unit of coded information. We note that any symbol within an alphabet of r letters can be uniquely assigned a binary (Boolean) code of length n, where n is the first integer greater than or equal to $\ln r/\ln 2$. Any sequence coded in an alphabet of r letters can therefore be recoded in 0 and 1s. Thus the nucleic acids of the genomic code, of which there are four, can be uniquely represented by 00, 01, 10 and 11, i.e., by four sequences of two bits. The definition we have chosen for the effective lengths of the binary sequences, Eq. (2), is nothing but the number of bits necessary to code a binary sequence with information content equal to that represented by the probability matrix for the consensus sequence.

The terminology regarding the probability matrices is not at all uniform. Some authors prefer to quote frequencies rather than probabilities, as in the "alignment matrices" defined by Li et al. (2002). "Weight matrices" (Benos et al., 2002), sometimes called Position Specific Weight Matrices (Gershenzon et al., 2005; Pachkov et al., 2007) may be used instead of probability matrices, and they are defined as $w_{ij}^{(m)} = \ln p_{ij}^{(m)} - \ln p_j^{(0)}$, where the $p_j^{(0)}$ are the background probabilities for the nucleotides j over the whole genome; m indexes a particular consensus bs. Within this convention the (relative) information content of a sequence of length l_m is defined as

$$I_m = \sum_{i=1}^{l_m}\sum_{j=1}^{4} p_{ij}^{(m)} w_{ij}^{(m)}. \tag{A.2}$$

Note that this differs from our definition, in that the subtracted quantity is not the information content of the random series but $\sum_{i=1}^{l_m}\sum_{j=1}^{4} p_{ij}^{(m)}\ln p_j^{(0)}$.

Appendix B. Topological characterization of complex networks

The in-degree of a node is defined as the number of directed edges incident upon that node. The out-degree is, conversely, the number of directed edges leading out of a node. The (total) degree is the number of distinct neighbors of a node, with the neighborhood being established either with in- or out-edges, or both.

The clustering coefficient (Bollobás, 1998; Albert and Barabasi, 2002; Dorogovtsev and Mendes, 2002) as a function of the degree k is defined as

$$C(k) = |G_k|^{-1}\sum_{i\in G_k}\sum_{\mu<\nu;\mu,\nu\in\Omega_i} 2e_{i,\mu\nu}/[k(k-1)], \tag{A.3}$$

where G_k is the set of nodes of degree k, i ranges over the elements of this set, $|G_k|$ is the size (the number of elements) of this set, Ω_i is the set of neighbors of the ith node, μ and ν range over this neighborhood and $e_{i,\mu\nu}$ is either zero or one depending upon whether the μth and νth members of the neighborhood of i are disconnected or connected.

The degree–degree correlation function (Colizza et al., 2005) is defined as

$$k_{nn}(k) = \sum_{k'} k' p(k|k'), \tag{A.4}$$

where $p(k|k')$ is the conditional probability that a node with degree k has a neighbor of degree k'.

The rich-club coefficient (Zhou and Mondragon, 2003; Colizza et al., 2006) is the total number of edges connecting nodes with degree greater than k, normalized by the maximum possible number of such connections,

$$r(k) = 2e_{>k}/[N_{>k}(N_{>k}-1)], \tag{A.5}$$

where $N_{>k}$ is the total number of nodes with degree greater than k and $e_{>k}$ is the total number of edges between such nodes.

The k-core analysis (Bollobás, 1998) of the network into different layers, or "shells" is performed via the following iterative procedure: All nodes that are at least of degree 1 will be called the 1-core of the graph. The graph may consist of more than one connected component. To start the iteration, all nodes which are connected with one edge only are eliminated by severing that edge. The process is repeated until no nodes remain which are singly connected to the graph. What remains of the graph is the 2-core, and all nodes outside it are termed the 1st shell (although some of them might have had degree greater than unity). At the second stage, one searches for nodes that are doubly connected to the rest of the graph and removes them together with their edges, and the process is repeated until none such are left. This yields the 3-core, and those nodes which have been removed at this stage make up the 2-shell. In each k-core, the nodes are of degree $\geq k$, and the k-shell consists of those nodes that belong to the k-core but not to the $k+1$st core. The "coreness" of a node is defined as the k value of the shell to which it belongs (Alvarez-Hamelin et al., 2005). One proceeds as outlined above until all nodes are exhausted. This means that once $k_{\max}$ has been reached, iteratively removing all nodes with degree $k_{\max}$ leaves an empty set of nodes.

Appendix C. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2009.11.017.

References

Albert, R., Barabasi, A.L., 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74 (1), 47–97.
Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. Molecular Biology of the Cell. Garland, New York.
Almirantis, Y., Provata, A., 1999. Long- and short-range correlations in genome organization. J. Stat. Phys. 97 (1–2), 233–262.

Alvarez-Hamelin, I., Dall'Asta, L., Barrat, A., Vespignani, A., 2005. k-core decomposition: a tool for the visualization of large scale networks. arXiv:cs/0504107 〈http://arxiv.org/abs/cs.NI/0504107〉.

Avery, J., 2003. Information Theory and Evolution. World Scientific, Singapore.

Babu, M.M., Teichmann, S.A., Aravind, L., 2006. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J. Mol. Biol. 358 (2), 614–633.

Balcan, D., Erzan, A., 2004. Random model for RNA interference yields scale free network. Eur. Phys. J. B 38 (2), 253–260.

Balcan, D., Erzan, A., 2007. Content-based networks: a pedagogical overview. Chaos 17 (2), 026108-1–026108-14.

Balcan, D., Kabakçıoglu, A., Mungan, M., Erzan, A., 2007. The information coded in the yeast response elements accounts for most of the topological properties of its transcriptional regulation network. PLoS ONE 2 (6), e501.

Banzhaf, W., Kuo, D.P., 2004. Network motifs in natural and artificial transcriptional regulatory networks. J. Biol. Phys. Chem. 4 (2), 85–92.

Barabasi, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Barabasi, A.-L., Oltvai, Z., 2004. Network biology: understanding the cell's functional organization. Nat. Genet. 5, 101–113.

Benos, P.V., Bulyk, M.L., Stormo, G.D., 2002. Additivity in protein–DNA interactions: how good an approximation is it? Nucleic Acids Res. 30 (20), 4442–4451.

Berg, J., Willmann, S., Lässig, M., 2004. Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. 4 (1), 42.

Bergmann, S., Ihmels, J., Barkai, N., 2004. Similarities and differences in genome-wide expression data of six organisms. PLoS Biol. 2, 85–93.

Bilu, Y., Barkai, N., 2005. The design of transcription-factor binding sites is affected by combinatorial regulation. Genome Biol. 6 (12), R103.

Bollobás, B., 1998. Modern Graph Theory. Springer, New York.

Browning, D.F., Busby, S.J., 2004. The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. 2 (1), 57–65.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.K., Simons, M., Stanley, H.E., 1995. Long-range correlation properties of coding and noncoding DNA sequences: Genbank analysis. Phys. Rev. E 51 (5), 5084–5091.

Colizza, V., Flammini, A., Maritan, A., Vespignani, A., 2005. Characterization and modeling of protein–protein interaction networks. Physica A 352 (1), 1–27.

Colizza, V., Flammini, A., Serrano, M.A., Vespignani, A., 2006. Detecting rich-club ordering in complex networks. Nat. Phys. 2 (2), 110–115.

Dawkins, R., 1986. The Blind Watchmaker. Norton and Co. Inc, New York.

Dawkins, R., 2006. Climbing Mount Improbable. Penguin, London.

Dobrin, R., Beg, Q.K., Barabasi, A.-L., Oltvai, Z.N., 2004. Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. BMC Bioinf. 5, 10.

Dodd, I.B.B., Shearwin, K.E.E., Sneppen, K., 2007. Modelling transcriptional interference and DNA looping in gene regulation. J. Mol. Biol. 369 (5), 1200–1213.

Dorogovtsev, S.N., Mendes, J.F.F., 2002. Evolution of networks. Adv. Phys. 51 (4), 1079–1187.

Erdös, P., Rényi, A., 1959. On random graphs. Publ. Math. 6, 290.

Erdös, P., Rényi, A., 1960. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. 5, 17.

Fu, Y., Weng, Z., 2004. Improvement of transfac matrices using multiple local alignment of transcription factor binding site sequences. In: IEMBS '04, vol. 2, pp. 2856–2859.

Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza Spinola, M.I.I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A.M.M., Treviño Quintanilla, L., Collado-Vides, J., 2007. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. Nucleic Acids Res. 36 (Database Issue), D120–D124.

Geard, N., Wiles, J., 2003. Structure and dynamics of a gene network model incorporating small RNAs. In: The 2003 Congress on Evolutionary Computation, pp. 199–206.

Gerland, U., Moroz, J.D., Hwa, T., 2002. Physical constraints and functional characteristics of transcription factor–DNA interaction. Proc. Natl. Acad. Sci. USA 99 (19), 12015–12020.

Gershenzon, N.I., Stormo, G.D., Ioshikhes, I.P., 2005. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. Nucleic Acids Res. 33 (7), 2290–2301.

Guelzim, N., Bottani, S., Bourgine, P., Kepes, F., 2002. Topological and causal structure of the yeast transcriptional regulatory network. Nat. Genet. 31, 60–63.

Harbison, C.T., Gordon, B.D., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.-B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, A.P., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E., Young, R.A., 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431 (7004), 99–104.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.-L., 2000. The large-scale organization of metabolic networks. Nature 407, 651–654.

Kashtan, N., Mayo, A., Kalisky, T., Alon, U., 2009. An analytically solvable model for rapid evolution of modular structure. PLoS Compt. Bio. 5, e1000355.

Kauffman, S.A., 1993. The Origins of Order, Self-Organization and Selection in Evolution. Oxford University Press, New York.

Kim, J.T., Martinetz, T., Polani, D., 2003. Bioinformatic principles underlying the information content of transcription factor binding sites. J. Theor. Biol. 220 (4), 529–544.

Koralov, L.B., Sinai, Y.G., 2007. Theory of Probability and Random Processes. Springer, Berlin, New York, p. 134.

Kugiumtzis, D., Provata, A., 2004. Statistical analysis of gene and intergenic DNA sequences. Physica A 342 (3–4), 623–638.

Lynch, M., 2007. The evolution of genetic networks by non-adaptive processes. Nat. Rev. Genet. 8, 803–813.

Li, H., Rhodius, V., Gross, C., Siggia, E.D., 2002. Identification of the binding sites of regulatory proteins in bacterial genomes. Proc. Natl. Acad. Sci. USA 99 (18), 11772–11777.

Ma, H.-W., Kumar, B., Ditges, U., Gunzer, F., Buer, J., Zeng, A.-P., 2004. An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. 32 (22), 6643–6649.

Matsumoto, M., Nishimura, T., 1998. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. (TOMACS) 8 (1), 3–30 (Special issue on uniform random number generation).

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U., 2004. Superfamilies of evolved and designed networks. Science 303 (5663), 1538–1542.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., 2002. Network motifs: simple building blocks of complex networks. Science 298 (5594), 824–827.

Münch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E., Jahn, D., 2003. Prodoric: prokaryotic database of gene regulation. Nucleic Acids Res. 31 (1), 266–269.

Mungan, M., Kabakçıoglu, A., Balcan, D., Erzan, A., 2005. Analytical solution of a stochastic content-based network model. J. Phys. A Math. Gen. 38 (44), 9599–9620.

O'Flanagan, R.A., Paillard, G., Lavery, R., Sengupta, A.M., 2005. Non-additivity in protein–DNA binding. Bioinformatics 21, 2254–2263.

Okuda, S., Kawashima, S., Kobayashi, K., Ogasawara, N., Kanehisa, M., Goto, S., 2007. Characterization of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*. BMC Genom. 8 (1), 48.

Pachkov, M., Erb, I., Molina, N., van Nimwegen, E., 2007. Swissregulon: a database of genome-wide annotations of regulatory sites. Nucleic Acids Res. 35 (Database Issue), D127–D131.

Reil, T., 1999. Dynamics of gene expression in an artificial genome—implications for biological and artificial ontogeny. Lecture Notes in Computer Science, vol. 1674. Springer, Berlin, pp. 457–466.

Rudd, K.E., 2000. Ecogene: a genome sequence database for *Escherichia coli* K-12. Nucleic Acids Res. 28 (1), 60–64.

Spirin, V., Mirny, L.A., 2003. Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. 100, 12123.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Díaz-Peredo, E., Sánchez-Solano, F., Santos-Zavaleta, A., Martínez-Flores, I., Jiménez-Jacinto, V., Bonavides-Martínez, C., Segura-Salazar, J., Martínez-Antonio, A., Collado-Vides, J., 2006a. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. 34 (Database Issue), D394–D397.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Peralta-Gil, M., Penaloza-Spinola, M.I., Martinez-Antonio, A., Karp, P.D., Collado-Vides, J., 2006b. The comprehensive updated regulatory network of *Escherichia coli* K-12. BMC Bioinf. 7 (1), 5.

Samal, A., Jain, S., 2008. The regulatory network of *E. coli* metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. BMC Syst. Biol. 2 (21 doi:10.1186/1752-0509-2-21.

Sengun, Y., Erzan, A., 2006. Content-based network model with duplication and divergence. Physica A 365 (2), 446–462.

Sengupta, A.M., Djordjevic, M., Shraiman, B.I., 2002. Specificity and robustness in transcription control networks. Proc. Natl. Acad. Sci. 99 (4), 2072–2077.

Shannon, C.E., 1949. Communication in the presence of noise. Proc. IRE 37, 10–21.

Shearwin, K.E., Callen, B.P., Egan, J.B., 2005. Transcriptional interference—a crash course. Trends Genet. 21 (6), 339–345.

Sneppen, K., Dodd, I.B., Shearwin, K.E., Palmer, A.C., Schubert, R.A., Callen, B.P., Egan, B.J., 2005. A mathematical model for transcriptional interference by RNA polymerase traffic in *Escherichia coli*. J. Mol. Biol. 346 (2), 399–409.

Stormo, G.D., 2000. DNA binding sites: representation and discovery. Bioinformatics 16 (1), 16–23.

Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A.R., Mira, N.P., Alenquer, M., Freitas, A.T., Oliveira, A.L., Sá-Correia, I., 2006. The yeastract database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. Nucleic Acids Res. 34 (Database Issue), D446–D551.

van Nimwegen, E., Zavolan, M., Rajewsky, N., Siggia, E.D., 2002. Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics. Proc. Natl. Acad. Sci. USA 99, 7323–7328.

van Noort, V., Snel, B., Huynen, M.A., 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep. 5 (3), 280–284.

Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J.-P., Oltvai, Z.N., Barabasi, A.-L., 2004. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proc. Natl. Acad. Sci. 101, 17940–17945.

Wagner, A., 1994. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. Proc. Natl. Acad. Sci. USA 91 (10), 4387–4391.

Warren, P.B., Wolde, T.P.R., 2003. Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. J. Mol. Biol. 342 (5), 1379–1390.

Watson, J., Geard, N., Wiles, J., 2004. Towards more biological mutation operators in gene regulation studies. Biosystems 76 (1–3), 239–248.

Wernicke, S., Rasche, F., 2006. FANMOD: a tool for fast network motif detection. Bioinf. Appl. Note 22 (9), 1152–1153.

Zhou, S., Mondragon, 2003. The rich-club phenomenon in the internet topology. IEEE Commun. Lett. 8 (3), 180–182.