Cellular and Molecular Biology
E-ISSN : 1165-158X / P-ISSN : 0145-5680
www.cellmolbiol.org

Original Research

The link between intergenic distances and controls exerted on the transcriptional regulation; an inferential approach

M. Ahmadi[1], B. Salehi[2*]

[1] Young Researchers and Elites Club, North Tehran Branch, Islamic Azad University, Tehran, Iran
[2] Mycobacteriology Research Center, National Research Institute of Tuberculosis and Lung Diseases (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran

Correspondence to: bahar.salehi007@gmail.com

Received February 8, 2017; **Accepted** February 20, 2017; **Published** February 28, 2017

Doi: http://dx.doi.org/10.14715/cmb/2017.63.2.3

Abstract: The position of genes along the genome is an important evolutionary factor for organizing gene regulation. Hence, transcriptional regulatory network have been studied much more extensively from gene distributions rather than other viewpoints. The systematics of intergenic distances, therefore, should be taken into account as an important source of information on the controls exerted on gene expression by various biological mechanisms. Here we study a collection of features including, intergenic and interoperonic distances, distances between isolated genes, distances between regulatory genes and distances between isolated and regulatory genes/operons in order to provide a more informative picture of gene distributions via firstly discovering the statistical model of these features. We find that all these features significantly follow the lognormal distribution. Then we test a few important biological hypotheses particularly in relation to controls exerted in the transcriptional regulatory network in a completely inferential approach using generalized p-value based on the discovered model. We find that mean distances of isolated genes significantly is less than that of regulatory genes. These findings are consistent with the previous evidences that many biological variables, especially in relation to systems biology, follow lognormal distribution. furthermore, our results inferentially support the crucial hypothesis on the two distinct logical types of control, namely digital control (i.e. control mediated by specific transcription factors) and analog control (i.e. control mediated by distribution of supercoiling energy and based on gene neighborhood) previously proposed by studying expression patterns.

Key words: Lognormal distribution; Distances between genes; Transcriptional regulatory network; Generalized P-value; Digital and analog control.

Introduction

Scientific researchers are still interested in studying the well-studied circular genome of *E. coli* which is used extensively for organizing gene regulation, investigating different types of controls used on gene expression, analyzing 3D structure of DNA (i.e. DNA topology).

In 80's De Martelaere and Van Gool (1) and Jurka and Savageau (2) discussed the gene density of the circular chromosome of *E. coli* as an important source of information on the gradually shaping of the system during the evolution and particularly as a means of using genome 3D structure for regulatory aims for the first time.

Two other studied focus on intergenic and interoperonic distances (3, 4). In both works, genes or operons regulating each other or pairs of genes or operons co-regulated by other genes are defined as regulatory pairs in order to study specific patterns in the distributions of them. Warren and ten Wolde (4) find that operons in such regulatory pairs have a generally reduced distance, suggesting an evolutionary pressure to reduce interoperonic distances in such regulatory pairs for efficient regulation. It should be noted that their method comprises of the partial pair correlation function and nearest neighbor distance probability density function as two summary characteristics of point process statistics.

Hermsen et al. (5) found that divergent gene pair, i.e., genes with opposite orientation, tend to have larger distances when oriented away from each other compared to convergent gene pair, i.e., those oriented towards each other. They discussed that this bias is because of the larger size of the upstream control region compared to the downstream one.

From systems viewpoint, the role of transcription factors in controlling gene regulation via direct binding to target genes has been mainly investigated (6). The transcriptional regulatory network (TRN) of *E. coli* has been built by collecting information of the mentioned interactions in to a database called RegulonDB (7, 8). This perspective provides thorough topological perceptions into the hierarchical organization of TRNs (9, 10) and their compositions made of specific network motifs (8). The use the TRN for interpreting expression patterns (11, 12) has been disclosed both the abilities and the limitations of this viewpoint. In evaluation process of this perspective, it became recently apparent that this view is not representative alone and very different regulatory mechanisms have to be considered as well such as changes in the DNA structure on a small scale (13, 14) and its alterations on larger (15) scale. Thus, to understand the gene regulation organization, a clear distinction of the different control types is essentially needed as the first step of investigating their effect on regulation.

This observation that closeness of genes elucidate some aspects of observed gene expression patterns (16-18) confirms the relationship between these two research areas, gene distribution and TRN. Particularly, the interaction between two types of control in gene expression profiles in *E. coli*, network-mediated and that mediated by DNA 3D structure, has been analyzed by (18).

The study (18) name these two control types as digital (due to the fact that the TRN provides discrete variable quantity on the transcriptional connections between discontinuous components, e.g. a particular pair of regulator and regulated gene) and analog (due to continuous information of the specific genes expression control provided by distributions of supercoiling energy in the genome), respectively. The biological hypothesis underlying this categorization is that each type of gene regulation is of specific length scale. Hence, intergenic distance has been studied extensively in order to help researchers offer the best model of gene distributions. Obviously, investigations, in this context, differ from two points of view; 1) categorization of genes 2) the way of testing the hypothesis such as using descriptive or inferential statistics.

Sonnenschein et al. (19) studied a new set of categories, which had not been analyzed before that time. They distinguished between "regulatory" genes (i.e. either being regulated by or genes producing a transcription factor regulates other genes) and "isolated" genes (i.e. genes not involved in regulation mediated by transcription factors). They found that these types of genes show a clear statistical repulsion and have different ranges of correlations. In particular, they found that isolated genes have a tendency for shorter intergenic distances. It should be added that they have employed a non-classical correlation function to test the hypotheses about the pre-defined categorized genes.

In terms of gene classifications, similar to (19), we analyze the above mentioned gene categories including regulatory and isolated genes using a inferentially statistical approach (i.e. generalized P-value). Needless to say, the biological hypothesis which motivates this classification is the different length scales of gene regulation mechanisms.

The novel feature of our work lies in two points: (1) statistical modeling of gene distributions in the transcriptional regulatory network (TRN) via modeling of the most complete collection of variables in relation to TRN has been already considered. The model of these variables has not been discovered before. Our finding here, lognormal distribution is the best model for mentioned variables, fits to the hypothesis that many crucial biological data follow lognormal distribution (2) our approach is completely inferential, using generalized p-value and generalized confidence intervals, to compare digital and analog control from their length scales viewpoint to elucidate the importance of obtaining the data model. Our hypothesis, based on the findings from (18, 19) is that two distinct types of controls, digital and analog, have different length scales and particularly isolated genes have a tendency to shorter intergenic distances as they tend to be co-regulated by spatial neighborhood via topology of the genome.

In this study, in order to investigate gene distributions, some critical gene spacing variables including intergenic and interoperonic shortest distances, shortest distances between two regulatory genes named by RV, two isolated genes named by IV and an isolated gene together with a regulatory one named by RIV are considered. In all these variables we analyze the shortest distance along the circular genome to the nearest neighbor of the respective type. We, similar to (19), do not consider the orientation of genes/ operons along the genome or their sizes. In fact, we consider every gene just by a single point, i.e., its center. We checked that our results remain unchanging when we consider other definitions of the "distance" between two genes. Results in the following section are presented both on the gene level and on the operon level.

Materials and Methods

Generalized p-value

In the statistical analysis of this paper we focus on a collection of critical variables which are representative features of gene distributions. Besides LPP-Plot (lognormal probability plot), the statistical models of these variables are discovered using the nonparametric method of Anderson Darling. This method tests if a given sample of data is drawn from a particular probability distribution function. More technically, it compares the empirical cumulative distribution function of the sample data with the distribution expected if the data were normal. If this observed difference is sufficiently large, the test will reject the null hypothesis of population normality. Needless to say, for the log normality test the hypotheses are H_0: data follow lognormal distribution vs. H_1: data do not follow lognormal distribution.

After discovering the lognormal distribution as the model of the afore-mentioned critical variables, we use generalized p-value in order to inferential analysis of some crucial biological hypotheses. Thus, our main reference is (20). In the following, we describe this method briefly.

Let X_1 and X_2 be two independent lognormal random variables, and let μ_1 and σ_1^2, respectively, represent the mean and variance of variable $Y_1 = \ln(X_1)$. Obviously, $Y_1 = \ln(X_1) \sim N(\mu_1, \sigma_1^2)$. Similarly, let μ_2 and σ_2^2, respectively, represent the mean and variance of variable $Y_2 = \ln(X_2)$ where $Y_2 = \ln(X_2) \sim N(\mu_2, \sigma_2^2)$. Many of the parameters of interest concerning the lognormal distribution (e.g. the mean of X_1) are functions of both μ_1 and σ_1^2 and it appears difficult to conduct exact and/or optimum tests and obtain confidence intervals. In particular, this is the case for the mean of the lognormal distribution given by

$$E(X_i) = E(\exp(Y_i)) = \exp(\eta_i), i = 1,2 \qquad (1)$$

where

$$\eta_i = \mu_i + \sigma_i^2/_2 , i = 1,2. \qquad (2)$$

Obviously, the problem of testing

$$H_0: \eta_1 \leq \eta_2 \quad \text{Vs.} \quad H_1: \eta_1 > \eta_2 \qquad (3)$$

is equivalent to test the following hypotheses:

$$H_0: \eta_1 - \eta_2 \leq 0 \quad \text{Vs.} \quad H_1: \eta_1 - \eta_2 > 0. \tag{4}$$

To address this problem, let $X_{1i}, i = 1,2,3,\ldots,n_1$ and $X_{2i}, i = 1,2,3,\ldots,n_2$ denote independent random samples from the lognormal distributions of X_1 and X_2, respectively. Suppose that $Y_{1i} = \ln(X_{1i}), i = 1,2,3,\ldots,n_1$, and $Y_{2i} = \ln(X_{2i}), i = 1,2,3,\ldots,n_2$. Define

$$\bar{Y}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} Y_{ij} \tag{5}$$

and

$$S_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2, i = 1,2. \tag{6}$$

Furthermore, let $\bar{y}_1, \bar{y}_2$, s_1^2 and s_2^2 stand for the observed values of their respective variables.

Let

$$T_i = \bar{y}_i - \frac{\bar{Y}_i - \mu_i}{S_i/\sqrt{n_i}} s_i/\sqrt{n_i} + \frac{1}{2}\frac{\sigma_i^2}{S_i^2}s_i^2 \tag{7}$$

$= \bar{y}_i - \dfrac{Z_i}{U_i/\sqrt{n_i - 1}} s_i/\sqrt{n_i} + \dfrac{1}{2}\dfrac{s_i^2}{U_i^2/(n_i - 1)}, i = 1,2$; are also independent. Define the generalized test variable as

$Z_i = \sqrt{n_i}\ (\bar{Y}_i - \mu_i)/\sigma_i \sim N(0,1)$ $Z_i = \sqrt{n_i}\ (\bar{Y}_i - \mu_i)/\sigma_i \sim N(0,1)$

$$T = i = 1,2_2 - (\eta_1 - \eta_2). \tag{8}$$

It can be easily shown that T satisfies the three following requirements any generalized test variable should meet. Hence, the generalized p-value for testing the hypotheses of (3) or (4) is given by

$$P(T \leq 0 | \eta_1 - \eta_2). \tag{9}$$

The three above mentioned requirements are as follows:

(a) The distribution of T is stochastically monotone with respect to $\eta_1 - \eta_2$.

(b) The observed value of T is free of any unknown parameters.

(c) For $\eta_1 - \eta_2 =$, then the distribution of T is free of any unknown parameters.

Transcriptional regulatory network and distribution of genes

We obtained the data from RegulonDB (version 6.2), (12) which is a database specifically dedicated to the transcriptional regulation of *E. coli*. A total number of 4600 genes, 3035 isolated genes, 1565 regulatory genes, 2660 operon, 1906 isolated operon, and 754 regulatory operons are included in this database.

Results and Discussion

First we present the model of the gene and operon distributions and compare their distributions in a completely inferential approach. Then, we present the model of isolated genes and also that of regulatory ones and compare their distributions in order to investigate the two distinct logical types of control (digital and analog) from length scale view point using generalized p-value which is firstly applied in biological sciences (to the best of our knowledge).

Figure 1 shows the distribution of shortest distances on the gene level (Figure 1a) and that on the operon level (Figure 1b). This Figure shows also the LPP-Plot of shortest intergenic and interoperonic distance variables (Figure 1c and 1d).

Figure1a vs. Figure 1b illustrates that distances between operons seems to be larger than distances between genes, confirming the hypothesis of the evolutionarily systematic omission of intra-operon distances, when passing from genes to operons. To test this hypothesis in an inferential way we need first to discover the model of the intergenic and interoperonic distance variables.

Figure 1a and Figure1b also reveal that intergenic and interoperonic distances are inherently positive variables both gets less quantities with grater frequencies but greater quantities with less frequencies. Hence, it seems that these variables, particularly the interoperonic distance variable, follow lognormal distribution. We tested this hypothesis descriptively using LPP-Plot (lognormal probability plot) and inferentially using nonparametric tests such as Anderson-Darling and Rayan-Joiner tests. Figure 1c reveals that the intergenenic distance variable roughly follows the lognormal distribution, supporting the observation shown by the histogram in figure 1a. The Anderson darling test shows that the lognormal significantly is the model of the gene distributions ($P<0.005$) as well. The small amount of p-value may be due to our incomplete current knowledge about the unknown genes in the transcriptional regulatory network. Considering such a discussion, the lognormal distribution is the best model that represents the gene distributions. Figure 1d reveals that the interoperonic distance variable follows the lognormal distribution as expected by its histogram in figure 1b. In addition, The Anderson darling

Test shows that the lognormal significantly is the model of the operon distributions ($P<0.05$).

Since we have already discovered that the intergenic and interoperonic distance variables significantly follow the lognormal model, then we now report the result of testing the following hypotheses:

H_0: "means of intergenic and interoperonic distance variables are equal"

H_1: mean of intergenic distance variable is less than that of interoperonic one".

Testing the recent hypotheses was performed in a completely inferential approach via generalized p-value method. After 1000 iteration, the obtained generalized p-value ($P<0.001$) suggests that the null hypothesis is significantly rejected, confirming the result descriptively shown by histograms in (19). This result also supports the hypothesis of the systematic omission of intra-operon distances, when passing from genes to operons.

Figure 2 shows the distribution of shortest distances between (regulatorily) isolated genes (IV variable) by Figure 2a and distribution of shortest distances between regulatory genes (RV variable) by figure 2b. This figure also shows the LPP-Plot of IV variable (Figure 2c) and that of RV variable (Figure 2d).

Comparison of Figure2a and Figure 2b indicates that

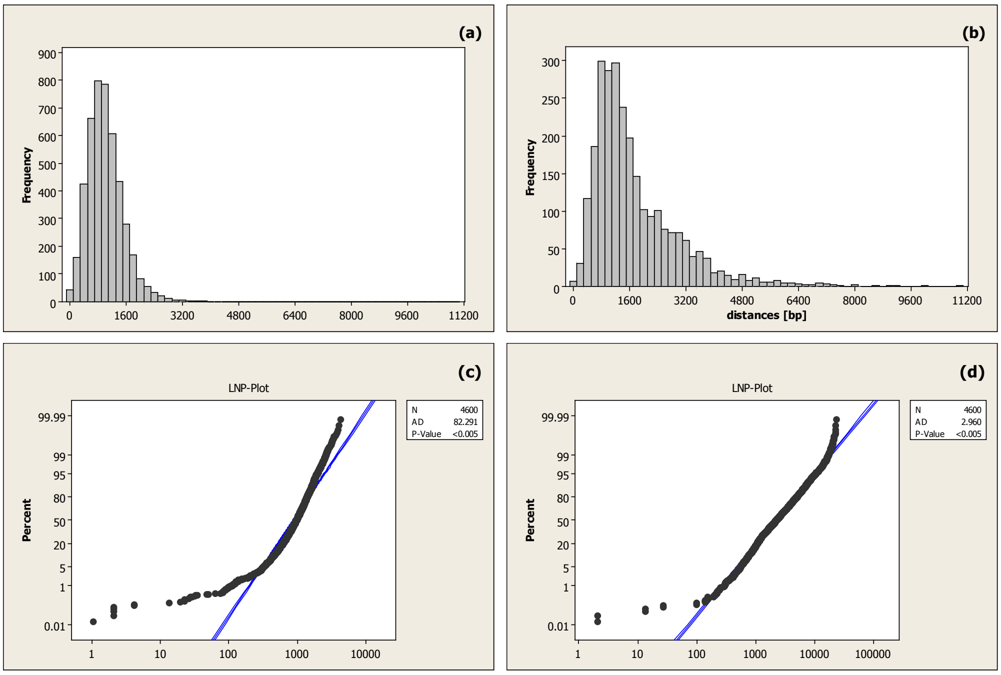

Figure 1. Distribution of shortest intergenic and interoperonic distances and their LPP-Plot. (a) Histogram of shortest intergenic distance, (b) histogram of shortest interoperonic distance, (c) LPP-Plot of shortest intergenic distance, (d) LPP-Plot of shortest operonic distance.

distances between isolated genes seems to be less than distances between regulatory ones. This hypothesis is due to the difference in mechanisms of two distinct logical types of controls including analog control where genes predominately regulated by 3D structure of the genome and digital control where genes mainly regulated by direct binding of the transcription factors to the regulatory region upstream of the genes. Additionally, Figure 2a and Figure 2b reveals that the IV and RV variable appear to follow lognormal distribution. The same as the approach taken to verify similar intuitive observation of figure1, we investigate this hypothesis descriptively using LPP-Plot and inferentially using nonparametric tests such as Anderson-Darling and Rayan-Joiner tests. Figure 2c and 2d indicate that the IV and RV variables both roughly follow the lognormal distribution, supporting the observation shown by the histogram in figure 2a and 2b. The Anderson darling test ensures that the log normal significantly is the model of the isolated and regulatory gene distributions (P<0.05).

Since we have already discovered that the IV and RV variables significantly follow the lognormal model, we now are able to report the result of testing the following critical hypotheses.

H_0: The means of IV and RV variables are equal
H_1: The mean of IV is less than that of RV

Testing the recent hypotheses was performed in a completely inferential approach via generalized p-value method. After 1000 iteration, the obtained the p-value (P<0.001) shows that the null hypothesis is significantly rejected, ensuring that the distances between isolated genes seem to be less than those between regulatory

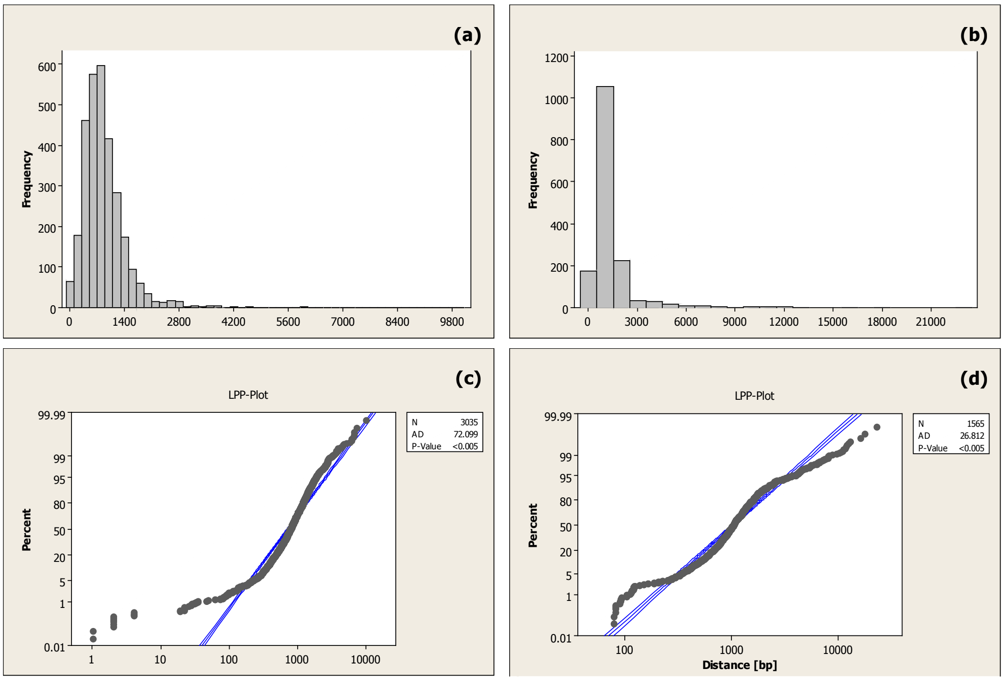

Figure 2. Distribution of IV and RV variables and their LPP-Plot. (a) Histogram of IV variable, (b) Histogram of RV variable, (c) LPP-Plot of IV variable, (d) LPP-Plot of RV variable.

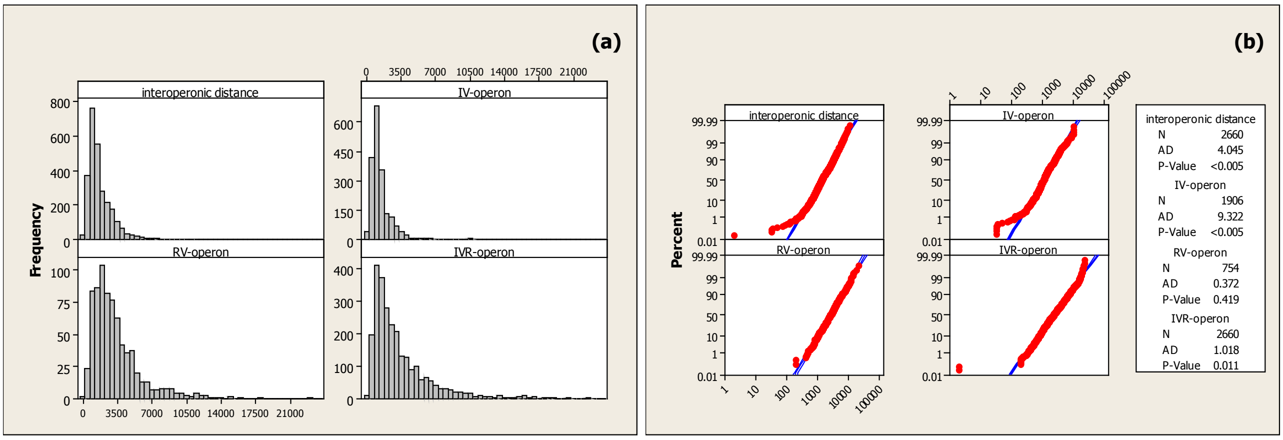

Figure 3. Distribution of interoperonic distance variable, IV, RV and IRV and their LPP-Plot. (a) Histogram of interoperonic distance variable, IV, RV and IRV (b) LPP-Plot of interoperonic distance variable, IV, RV and IRV.

ones. This result supports that reported by (19) and also confirms that digital control exerts relatively in larger scale in length perspective than analog control.

In the level of genes we also found that the shortest distance between regulatory and isolated elements, RIV, significantly follow the lognormal distribution. Thus, the monte carlo estimation of generalized p-value after 1000 iteration showed that the mean of RIV is greater than that of IV. That is, the following null hypothesis, is significantly rejected (P<0.001).

H_0: the means of RIV and IV are equal

H_1: the mean of RIV is greater than that of IV

Similarly, we find that RIV values averagely are larger than RV values. These two recent results suggest that genes of the same type (i.e. isolated or regulatory) prefer to be closer to each other while repulse genes of the different type. This may result from real pushing away in addition to relative "attraction" of the genes of the same type toward each other. This repulsion is interpreted as an unmixing of genes mainly regulated by transcription factors (digital control) and genes largely regulated by DNA topology (analog control).

Most of the operons in E. coli consist of only a single gene, some operons, however, contain as many as 15 genes. Henceforth, results of the 2nd phase of this work, i.e. repeat of above analyses at the operon level, will be presented.

Figure 3 shows histograms of interoperonic distance variable, IV, RV, and IRV and LPP-Plot of each one at the

Operon level. All these variables, like in the case of gene level while more closely, follow the lognormal distribution (see figure 3a and 3b). The results of hypotheses testing regarding the means of these variables provided the same conclusions obtained in the gene level.

References

1. De Martelaere D, Van Gool A. The density distribution of gene loci over the genetic map of *Escherichia coli*: Its structural, functional and evolutionary implications. Journal of molecular evolution. 1981;17(6):354-60.
2. Jurka J, Savageau MA. Gene density over the chromosome of *Escherichia coli*: frequency distribution, spatial clustering, and symmetry. Journal of bacteriology. 1985;163(2):806-11.
3. Képes F. Periodic transcriptional organization of the *E. coli* genome. Journal of molecular biology. 2004;340(5):957-64.
4. Warren P, Ten Wolde P. Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. Journal of molecular biology. 2004;342(5):1379-90.
5. Hermsen R, Ten Wolde PR, Teichmann S. Chance and necessity in chromosomal gene distributions. Trends in Genetics. 2008;24(5):216-9.
6. Palsson B, Palsson BØ. Systems biology: Cambridge university press; 2015.
7. Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spinola MI, Contreras-Moreira B, et al. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic acids research. 2008;36(suppl 1):D120-D4.
8. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nature genetics. 2002;31(1):64-8.
9. Ma H-W, Kumar B, Ditges U, Gunzer F, Buer J, Zeng A-P. An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic acids research. 2004;32(22):6643-9.
10. Yu H, Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. Proceedings of the National Academy of Sciences. 2006;103(40):14724-31.
11. Gutierrez-Rios RM, Rosenblueth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR, et al. Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. Genome research. 2003;13(11):2435-43.
12. Herrgård MJ, Covert MW. Reconciling gene expression data with known genome-scale regulatory network structures. Genome research. 2003;13(11):2423-34.
13. Travers A, Muskhelishvili G. DNA supercoiling—a global transcriptional regulator for enterobacterial growth? Nature Reviews Microbiology. 2005;3(2):157-69.
14. Travers A, Muskhelishvili G. Bacterial chromatin. Current opinion in genetics & development. 2005;15(5):507-14.
15. Wright MA, Kharchenko P, Church GM, Segrè D. Chromosomal periodicity of evolutionarily conserved gene pairs. Proceedings of the National Academy of Sciences. 2007;104(25):10559-64.
16. Blot N, Mavathur R, Geertz M, Travers A, Muskhelishvili G. Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome. EMBO reports. 2006;7(7):710-5.
17. Li S, Liberman LM, Mukherjee N, Benfey PN, Ohler U. Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. Genome research. 2013;23(10):1730-9.
18. Marr C, Geertz M, Hütt M-T, Muskhelishvili G. Dissecting the

logical types of network control in gene expression profiles. BMC systems biology. 2008;2(1):18.
19. Sonnenschein N, Hütt M-T, Stoyan H, Stoyan D. Ranges of control in the transcriptional regulation of Escherichia coli. BMC systems biology. 2009;3(1):1.
20. Krishnamoorthy K, Mathew T. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. Journal of statistical planning and inference. 2003;115(1):103-21.