Bio-Medical Materials and Engineering 24 (2014) 3719–3727
DOI 10.3233/BME-141200
IOS Press

A link partition approach for finding overlapping functional modules in the transcriptional regulatory network

Qingyu Zou, Fu Liu[*], Tao Hou, Yihan Jiang and Reifeng Mo
College of Communication Engineering, Jilin University, Changchun 132000, People's Republic of China

Abstract. The transcriptional regulation of cellular functions is carried out by the overlapping functional modules of a complex network. In this paper, a statistical approach for detecting functional modules in the transcriptional regulatory networks (TRNs) is studied. The proposed method defines modules as groups of links rather than nodes since nodes naturally belong to more than one module. Furthermore, the proposed algorithm is evaluated on the Escherichia coli TRN. The experimental results demonstrate that it detected a suitable number of overlapping modules that were biologically meaningful without any prior knowledge about the modules.

Keywords: Transcriptional regulation, network, functional module

1. Introduction

Transcriptional regulation is one of the most important mechanisms in gene expression. Using a network to represent the complicated process of transcriptional regulation has produced a powerful approach that is capable of effectively illustrating the dynamic interplay among the components involved in this process [1]. One gene's expression can be controlled by another gene's gene product. Thus, a directed graph can be used to model the transcriptional regulation process. In transcriptional regulatory networks (TRNs), the transcriptional regulator coding genes and target genes are represented as nodes, and the control interactions between them are indicated as directed links [2–4].

In order to understand how the mechanism of genes in the TRNs relate to one another as well as the influence of the topologies of TRNs on the biological process, there is an extremely useful analytical approach that involves extracting mesoscale structures, known as modules, from a network, which are defined as a group of co-regulated genes that likely share a common biological function. Module structure is one of the most important features of TRNs [5]. Therefore, detecting modules in TRNs are a critical component of understanding the relationship between the topology structure and represent function [6].

[*]Corresponding author: Fu Liu, College of Communication Engineering, Jilin University, Changchun 132000, People's Republic of China. Tel.:13610708679; Fax: 0431-85095775; E-mail: liufu@jlu.edu.cn.

Thus far, a large number of algorithmic approaches have been proposed to detect modules in different types of networks [7,8]. Optimization and clustering algorithms are two of the primary kinds of algorithms for detecting modules in complex networks. The underlying idea of the optimization methods is to define a quantity that is high for `good' network divisions and low for 'bad' ones and then to search through the possible divisions to find the one with the highest score. Numerous different measures for assigning scores have been proposed, such as the likelihood-based measures [9], fluid-flow [10], information theoretic [11], and others [12], but the most widely used approach is the modularity [13]. On the other hand, the clustering algorithms first estimate the strength of the link between each pair of nodes based on different methods, such as the link betweenness [14], link clustering coefficient [15], information centrality [16], similarity based on random walks [17], clustering centrality [18], and so on. Then, the partition results of the networks are obtained by either merging the two nodes with the highest link strength repeatedly (the agglomerative method) or by removing the link with the lowest strength repeatedly (the divisive methods).

Whereas nearly all of these methods are focused on the module of nodes, Yong-Yeol Ahn et al. [19] and T. S. Evans et al. [20] have recently conducted research that was focused on the cluster of links in undirected networks with the purpose of uncovering overlapping modules. However, TRNs are a directed network. The most common approach for detecting modules in directed networks has been to simply ignore the link directions and apply algorithms designed for undirected networks [21,22]. However, by discarding the direction of the links, it is evident that important information about the network's structure is simply being neglected, and with this information, a more accurate determination of the modules could be constructed.

In this paper, a new algorithm based on links similarity is proposed to detect the overlapping functional modules in the TRN of Escherichia coli. Based upon links similarity, the original TRN has been transformed into a weighted, undirected link network whose nodes are the original network's links and the link weight is the links similarity of the original network. Then, we used a hierarchical clustering algorithm in the transformed network to identify module structure. Moreover, in order to measure the strength of the module structure and to obtain the most relevant modules, an improved Newman-Girvan modularity Q [23] was used.

We compared the performance of our algorithm with two successful methods, with one designed by Resendis et al [21] and the other by Ahn et al [19]. Resendis et al identified eight functional modules from the TRN of Escherichia coli based upon the shortest path between the nodes. Ahn et al reinvented modules as groups of links rather than nodes and then show that this approach naturally incorporates overlap while revealing hierarchical organization.

2. Materials and methods

2.1. Link similarity

The link similarity is a measure of the closeness between a pair of links. It is clear that in the same network module the node-node connections are denser, and the shortest paths between pairs of nodes are shorter than in different modules. According to this principle, the similarity $S(e_{il},e_{jk})$ between links e_{il} and e_{jk} that is shown in Figure 1 is:

$$S(e_{il},e_{jk}) = \alpha NS(e_{il},e_{jk}) + (1-\alpha) DS(e_{il},e_{jk}) \tag{1}$$

where $NS(e_{il},e_{jk})$ represents the interlinkage closeness degree between links e_{il} and e_{jk}. $DS(e_{il},e_{jk})$ measures the distance of links e_{il} and e_{jk}, and α (between zero and one) is a parameter to adjust the weight of $NS(e_{il},e_{jk})$ and $DS(e_{il},e_{jk})$, as shown in Figure 1.

$$NS(e_{il},e_{jk}) = \frac{\left(n_{-}(i)\cap n_{-}(j)\right)+\left(n_{+}(l)\cap n_{+}(k)\right)}{\left(n_{-}(i)\cup n_{-}(j)\right)+\left(n_{+}(l)\cup n_{+}(k)\right)} \tag{2}$$

where $n_{+}(i)$ is the number of neighbors of a node i that direct it, and $n_{-}(i)$ is the number of neighbors of a node i that it directs.

$$DS(e_{il},e_{jk}) = \left(\frac{sp_{lj}+sp_{ki}}{2dia}\right)\delta\left(sp_{lj},sp_{ki}\right) \tag{3}$$

where sp_{lj} is the number of nodes in the shortest path [24] between nodes l and j. dia is the length of the longest shortest path in the TRN. The $\delta(sp_{lj,}sp_{ki})$ function is 1 if neither sp_{lj} nor sp_{ki} are zero; otherwise, it is 0.

2.2. Hierarchical clustering

After calculating the similarities for all the link-pairs in the TRN, a new weighted network was constructed, named link-net, in which the nodes are links of the TRN, and the links express the likenesses of the TRN links. Then, using hierarchical clustering, nodes in the link-net were clustered based upon their degree [25]. The clustering processes are described in the following three steps:

– Calculate the degree for each node in link-net.
– To initialize, assign each node to a cluster; then, merge the clusters iteratively using the single linkage function according to the nodes' degrees.
– Stop merging when all nodes belong to a unique cluster.

The trace of the clustering process is then stored in a dendrogram, which contains all the information of the hierarchical module organization. The similarity value at which the two clusters merge is considered to be the strength of the merged module and is encoded as the height of the relevant dendrogram branch to provide additional information.

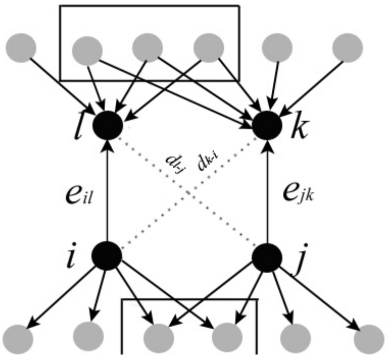

Fig. 1. Link similarity measure between links e_{ik} and e_{jk}.

2.3. Dendrogram partition

Hierarchical clustering methods repeatedly merge groups until all the elements are members of a single cluster. This eventually forces highly disparate regions of the network into single clusters. In order to find meaningful modules rather than just the hierarchical organization pattern of modules, it is important to know where to partition the dendrogram. Modularity Q [13] has been widely used for similar purposes,

$$Q=\sum_{i=1}^{k}\left(e_{ii}-a_i^2\right) \tag{4}$$

where e_{ii} is the fraction of links belonging to module i in the total weight of all the links, and a_i is the fraction of links connecting module i with other modules. In order to apply Q to the weighted network, the number of links is replaced by the sum of the weights. Then, the new modularity Q_{od} is given by:

$$Q_{od}=\sum_{i=1}^{k}\left(\left(\frac{\omega_i}{\sum_{i=1}^{k}\omega_i}\right)-\left(\sum_{j=1}^{k}\left(\frac{\psi_{ij}}{\sum_{i=1}^{k}\omega_i}\right)\right)^2\right) \tag{5}$$

where ω_i is the sum of the link weights in module i. ψ_{ij} is the sum of the link weights between module i and j.

3. Results and discussion

To clearly explain how the proposed method works, a small-scale example directed network consisting of six nodes and nine links is presented in Figure 2A, where node 2 is shared by two modules. First, an 6×6 adjacency matrix A (Figure 2B) was constructed in which A_{ij}=1 if i and j are connected; otherwise, it equals 0. Then, the links similarity matrix setting α=0.4 (Figure 2C) was calculated and the original network was transformed to link-net (Figure 2D). Finally, the nodes in link-net were clustered, and the modularity Q_{od} for each partition was calculated to find the meaningful one (Figure 2E). As shown in Figure 2F, when the Q_{od} is at its maximum, two overlapping modules have been explicitly recovered: one contains the number 1, 2, 5, and 6 nodes, and the other contains the number 2, 3, and 4 nodes.

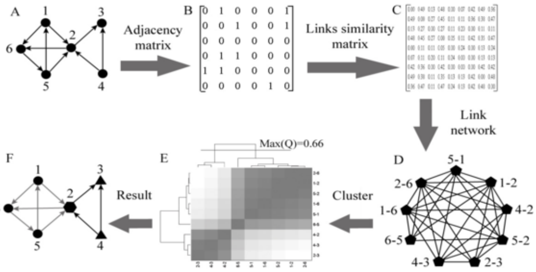

Fig. 2. Detecting module structure in the toy example network using the proposed method. (A) original network, (B) adjacency matrix, (C) links similarity matrix, (D) transformed network (link-net), (E) hierarchical clustering, and (F) partition result.

Escherichia coli is considered the most complete available prokaryotic, and therefore, it was selected for this study. The gene–gene transcriptional regulatory relationships were selected from the RegulonDB database [26,27]. A TRN model represents the molecular regulation process of transcription. A gene X directly regulates a gene Y if the protein that is encoded by X is a transcriptional factor for Y. Transcriptional regulator coding genes and target genes are represented as nodes, and the interactions between them are marked as links in the TRN. To construct an integrated *Escherichia coli* TRN, the interactions of A-E and A-F are replaced by B-E, B-F, and C-E, and C-F if a regulatory gene A, which is encoded by gene B and C, regulates gene E and F. The resulting *E. coli* TRN includes 1680 nodes and 4150 interactions with 186 regulatory genes controlling the expression of 1491 genes. The basic properties [24] of the resulting *E. coli* TRN are shown in Table 1.

The functional modules in the Escherichia coli TRN were detected using the proposed method. α is set at 0.5 in order to equalize the interlinkage closeness degree and distance of links. As shown in Figure 3, picture A is the hierarchical clustering dendrogram of the regulating nodes in the TRN link-net, and picture B is a plot of the fitted curve of modularity Q_{od}. The modularity graph is aligned with the dendrogram so that the modularity values for different divisions of the network can be directly read. The dendrogram is divided into eight clusters when the modularity reaches its maximum value. It is evident that the peak in the modularity (the dotted line) corresponds to a perfect identification of them.

4930 gene functional annotations of 1498 TRN nodes were extracted from the GeneProtEC [28,29] database and 1272 nodes distributed into 5228 annotations from the Gene Ontology [30,31] database. Merged two sets of data 6788 functional annotations of 1640 TRN nodes were obtained. In order to measure the effectiveness of the proposed module detecting algorithm, a direct validation was used by comparing identified clusters with a list of Escherichia coli functional modules, which is annotated with genes corresponding to functional class according to Monica Riley's MultiFun system [32,33] and obtained from the GeneProtEC and Gene Ontology database.

Table 1

Properties of the TRN of Escherichia Coli

The number of nodes and links						Average of properties		
TF	GENE	ConTF	Link	TFlink	Trace	Degree	Spl	Cluster
189	1491	135	4150	267	121	4.796	2.715	0.313

Note: TF is the number of regulatory factors, GENE is the number of regulated genes, ConTF is the number of interconnecting regulatory factors, Link is the number of links in TRN, TFlink is the number of links among regulatory factors, and Trace is the number of self-citations. Degree, Spl and Cluster represent average of degree, shortest path length, and clustering coefficient of TRN, respectively.

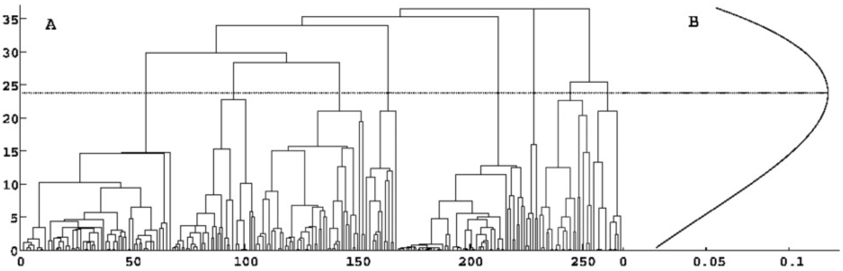

Fig. 3. Detecting modules of TRN from the link dendrogram. A is the link dendrogram of the regulatory genes network. B is the modularity Q_{od}.

The comparison results of our algorithm are shown in Table 2. The functions *consistency*$_i$, *intersection*$_i$ and *overlap*$_i$ were defined to measure how well the algorithm grouped nodes into a functionally correlated module *i*.

Consistency was defined as:

$$consistency_i = \frac{|inde_{ii} \cap reale_{ii}|}{|inde_{ii}|} \times 100\% \tag{6}$$

Intersection was defined as:

$$intersection_i = \frac{|inde_{ii} \cap reale_{ii}|}{|reale_{ii}|} \times 100\% \tag{7}$$

Overlap was defined as:

$$overlap_i = \left[\frac{1}{n-1} \sum_{j \neq i} \left| \frac{|inde_{ij} \cap reale_{ij}|}{|inde_{ij} \cup reale_{ij}|} \right| \right] \times 100\% \tag{8}$$

where $inde_{ii}$ is the nodes of an identified module *i*; $reale_{ii}$ is the nodes of a practical module *i*; $inde_{ij}$ and $reale_{ij}$ are the nodes shared between an identified and practical module *i* and *j*, respectively; n is the number of modules; and the absolute value sign represents the number of nodes.

Consistency is the fraction of the number of accurate identifications out of the total identifications. Furthermore, intersection is the fraction of the number of accurate identifications out of the total practical predictions, and finally, overlap is the fraction of the number of identified overlapping predictions out of the total overlapping predictions from the identified and practical predictions. They are the measurement of algorithm accuracy. The maximum for consistency, intersection, and overlap is 81.3%, 61%, and 75.5%, respectively, which indicates that a majority of the genes in the same identified module have consistently functional annotation.

Table 2

Accuracy rating for module partition results

Module	Biological function	Consistency	Intersection	Overlap
M1	Carbon compound utilization	80%	51.2%	66.3%
M2	Macromolecule degradation	55.6%	14.7%	57.1%
M3	Energy metabolism	51.3%	46.5%	42.5%
M4	Energy production/transport	52.2%	42.9%	50.9%
M5	Building block biosynthesis	62.5%	8.8%	57.6%
M6	Macromolecules biosynthesis	46.2%	61%	44%
M7	Central intermediary metabolism	81.3%	20%	75.5%
M8	Phosphorous Sulfur Nitrogen Metabolism	39.5%	46.9%	49.4%
	Average	58.5%	36.5%	55.4%

Table 3

The same degree of overlapping nodes

Modules	M1	M2	M3	M4	M5	M6	M7	M8
M1	None	66.7%	60.0%	50.0%	100.0%	45.0%	78.6%	63.6%
M2	66.7%	None	42.9%	40.0%	50.0%	66.7%	100.0%	33.3%
M3	60.0%	42.9%	None	54.5%	20.0%	58.3%	20.0%	42.1%
M4	50.0%	40.0%	54.5%	None	60.0%	30.0%	80.0%	41.7%
M5	100.0%	50.0%	20.0%	60.0%	None	33.3%	100.0%	40.0%
M6	45.0%	66.7%	58.3%	30.0%	33.3%	None	50.0%	25.0%
M7	78.6%	100.0%	20.0%	80.0%	100.0%	50.0%	None	100.0%
M8	63.6%	33.3%	42.1%	41.7%	40.0%	25.0%	100.0%	None
Mean	66.3%	57.1%	42.5%	50.9%	57.6%	44.0%	75.5%	49.4%

M1: Carbon compound utilization; M2: Macromolecule degradation; M3: Energy metabolism; M4: Energy production/transport; M5: Building block biosynthesis; M6: Macromolecules biosynthesis; M7: Central intermediary metabolism; and M8: Phosphorous Sulfur Nitrogen Metabolism.

Algorithms based on link similarity are more suitable for finding modules in the TRN because it is common that the functional modules of the TRN are overlapping. Table 3 shows the same degree of functional modules overlapping nodes between identified and practical partitions. From Table 3, it can be seen that the accuracy rate of 64.3% overlaps among functional modules is greater than 50% of the total, and 14.3% overlaps is equal to 100% of the total.

Furthermore, the Ahn's approach and Resendis's algorithm were both ran on the *Escherichia coli* TRN, and the resulting data was compared to the results obtained from the algorithm proposed in this study. The modules obtained by each algorithm were compared with practical functional modules. The cutoff parameter was set to obtain eight clusters for each algorithm. For each identified module, the consistency, intersection and overlap values were calculated. As shown in Figure 4, the accuracy of the proposed algorithm is higher than that of the other two algorithms. There are more significant functional modules generated by the proposed method than by the other two algorithms.

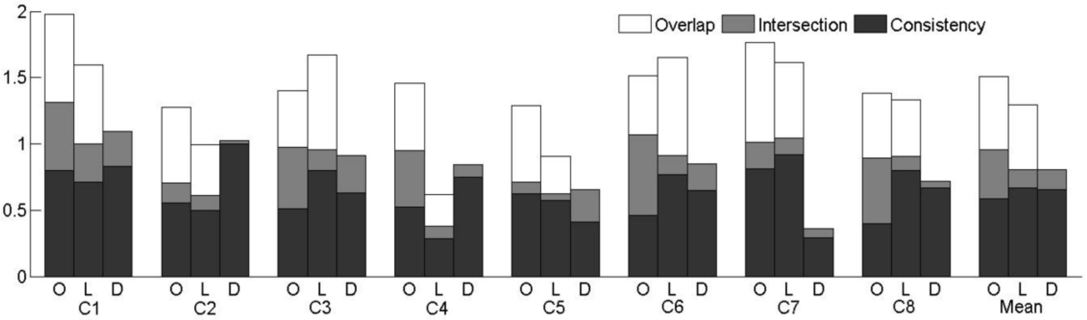

Fig. 4. Consistency, intersection, and overlap of each module calculated by three module detecting algorithms. O: Our method; L: Ahn's approach based on link similarity; and D: Resendis's algorithm based on the shortest path. C1: Carbon compound utilization; C2: Macromolecule degradation; C3: Energy metabolism; C4: Energy production/transport; C5: Building block biosynthesis; C6: Macromolecules biosynthesis; C7: Central intermediary metabolism; and C8: Phosphorous Sulfur Nitrogen Metabolism.

4. Conclusion

In the postgenome era of animalcule, a major research area is the discovery of the functional modules from the network level. In addition, the TRN is a very important and specific biological network. Links in TRN are directive, and functional modules are commonly overlapping. Previous module detecting methods in directed networks generally either ignore the link directions or divide a node into only one module. A rising challenge is how to discover the overlapping functional modules in a directed network. To overcome this challenge, in this study, a new algorithm based on the link similarity was developed, which can be used in TRN. A new measure of link similarity and a new modularity Q_{od} were introduced. The proposed algorithm was applied to the *Escherichia coli* TRN. The identified modules were confirmed by the function annotations of genes. The experimental results show that the identified modules approximately correspond to practical modules in terms of function annotations.

Additionally, the performances of this study's method were compared with two previous classic algorithms. The quantitative comparison of consistency, intersection, and overlap revealed that this study's method outperforms the other previous competing algorithms. The results show that the proposed method is efficient for discovering the overlapping function modules of large-scale directed networks. However, its full potential remains unexplored. The accuracy rating of the identified results was a bit lower than expected. In this study, the work primarily focused on the highly overlapping module structure of complex networks, but the hierarchy that organizes these overlapping modules holds great promise for further study.

References

[1] M.M. Babu, Structure, evolution and dynamics of transcriptional regulatory networks, Biochemical Society Transactions **38** (2010), 1155–1178.

[2] S.B.T. de-Leon and E.H. Davidson, Modeling the dynamics of transcriptional gene regulatory networks for animal development, Developmental Biology **325** (2009), 317–328.

[3] M.M. Babu, N.M. Luscombe, L. Aravind et al., Structure and evolution of transcriptional regulatory networks, Current Opinion in Structural Biology **14** (2004), 283–291.

[4] E.H. Davidson, J.P. Rast, P. Oliveri, et al., A genomic regulatory network for development, Science **295** (2002), 1669–1678.

[5] A. Martinez-Antonio, S.C. Janga and D. Thieffry, Functional organisation of Escherichia coli transcriptional regulatory network, Journal of Molecular Biology **381** (2008), 238–247.

[6] P.J. Mucha, T. Richardson, K. Macon et al., Community structure in time-dependent, multiscale, and multiplex networks, Science **328** (2010), 876–878.

[7] G. Palla, I. Derenyi, I. Farkas et al., Uncovering the overlapping community structure of complex networks in nature and society, Nature **435** (2005), 814–818.

[8] M.E.J. Newman, Community detection and graph partitioning, Epl. **103** (2013), 6.

[9] B. Karrer and M.E. Newman, Stochastic blockmodels and community structure in networks, Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. **83** (2011), 016107.

[10] A.A.H. Zanjani and A.H. Darooneh, Finding communities in linear time by developing the seeds, Physical Review E **84** (2011), 036109.

[11] M. Rosvall and C.T. Bergstrom, An information-theoretic framework for resolving community structure in complex networks, Proc. Natl. Acad. Sci. USA **104** (2007), 7327–7331.

[12] Z. Li, S. Zhang, R.-S. Wang et al., Quantitative function for community detection, Physical Review E **77** (2008), 036109.

[13] M.E.J. Newman and M. Girvan, Finding and evaluating community structure in networks, Physical Review E **69** (2004), 1–16.

[14] M. Girvan and M.E.J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences of the United States of America **99** (2002), 7821–7826.

[15] F. Radicchi, C. Castellano, F. Cecconi et al., Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA **101** (2004), 2658–2663.

[16] S. Fortunato, V. Latora and M. Marchiori, Method to find community structures based on information centrality, Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. **70** (2004), 056104.

[17] P. Pons and M. Latapy, Computing communities in large networks using random walks, Computer and Information Sicences-Iscis 2005 Proceedings **3733** (2005), 284–293.

[18] B. Yang and J.M. Liu, Discovering global network communities based on local centralities, Acm. Transactions on the Web **2** (2008), 1–32.

[19] Y.Y. Ahn, J.P. Bagrow and S. Lehmann, Link communities reveal multiscale complexity in networks, Nature **466** (2010), 761–U711.

[20] T.S. Evans and R. Lambiotte, Line graphs, link partitions, and overlapping communities, Physical Review E **80** (2009), 016105.

[21] O. Resendis-Antonio, J.A. Freyre-Gonzalez, R. Menchaca-Mendez et al., Modular analysis of the transcriptional regulatory network of E-coli, Trends in Genetics **21** (2005), 16–20.

[22] H.W. Ma, J. Buer and A.P. Zeng, Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach, Bmc Bioinformatics **5** (2004), 199.

[23] M.E.J. Newman, Detecting community structure in networks, European Physical Journal B **38** (2004), 321–330.

[24] F. Emmert-Streib and M. Dehmer, Networks for systems biology: conceptual connection of data and function, Iet Systems Biology **5** (2011), 185–207.

[25] R. Xu and D. Wunsch, Survey of clustering algorithms, IEEE Transactions on Neural Networks **16** (2005), 645–678.

[26] S. Gama-Castro, H. Salgado, M. Peralta-Gil et al., RegulonDB version 7.0: transcriptional regulation of escherichia coli K-12 integrated within genetic sensory response units (Gensor Units), Nucleic Acids Research **39** (2011), D98–D105.

[27] H. Salgado, I. Martinez-Flores, A. Lopez-Fuentes et al., Extracting regulatory networks of escherichia coli from RegulonDB, Methods Mol. Biol. **804** (2012), 179–195.

[28] M. Riley, Genes and proteins of escherichia coli K-12 (GenProtEC), Nucleic Acids Res. **25** (1997), 51–52.

[29] M.H. Serres, S. Goswami and M. Riley, GenProtEC: an updated and improved analysis of functions of escherichia coli K-12 proteins, Nucleic Acids Research **32** (2004), D300–D302.

[30] G.O. Consortium, The gene ontology project in 2008, Nucleic Acids Research **36** (2008), D440–D444.

[31] T.G.O. Consortium, The gene ontology: enhancements for 2011, Nucleic Acids Res. **40** (2012), D559–564.

[32] M.H. Serres and M. Riley, MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products, Microb. Comp. Genomics **5** (2000), 205–222.

[33] E. Becker, B. Robisson, C.E. Chapple et al., Multifunctional proteins revealed by overlapping clustering in protein interaction network, Bioinformatics **28** (2012), 84–90.