A Microorganism Transcriptional Regulation Algorithm Based on Generalized Regression Neural Network

Hui Li

Software Engineering College
Zhengzhou University of Light Industry
Zhengzhou 450002, China
E-mail: lihui@zzuli.edu.cn

Received: November 06, 2018 **Accepted: April 25, 2019**

Published: June 30, 2019

Abstract: *Considering the importance of operon in microorganism transcriptional regulation, this paper sets up a new operon prediction model based on artificial neural network (ANN). Specifically, multiple genome information, ranging from intergenic distance (IGD), orthologous protein cluster (OPC), conserved gene pair (CGP) to system evolution spectrum (SES), were preprocessed by log-likelihood fraction and wavelet transform, and then inputted to the GRNN for operon prediction. The experimental results in E. coli K-12 and B. subtilis 168 show that our model is a valid and feasible way to predict operon. The research findings shed new light on the prediction of operon information of new species.*

Keywords: *Microorganism transcriptional regulation, Operon prediction, Generalized regression neural network.*

Introduction

The term “operon” first appeared in the research of protein regulation mechanism, referring to a cluster of genes co-regulated by operators, promoters and terminators [3]. This cluster controls the gene manipulation in the transcription of mRNA, an intermediate product in protein synthesis, against a DNA template. As shown in Fig. 1, an operon is a transcriptional unit consisting of manipulating genes, public promoters, terminators, as well as other regulatory elements and structural genes.

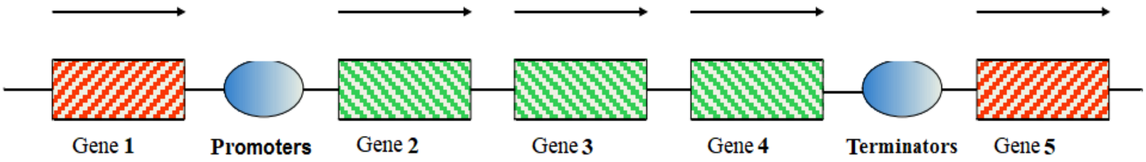

Fig. 1 Structure of operon

Considering the importance of the operon to the regulatory network, this paper explores the operon structure in the entire genome, with the aim to guide the reconstruction of biochemical and metabolic networks, and promote the research of microorganism transcriptional regulation [4, 8, 11-12].

Preliminaries

Definition of operon prediction

Currently, there are mainly two ways to define operation prediction, namely, the neighboring gene pair prediction (NGPP) method and the gene cluster prediction (GCP) method. The former is implemented in two steps: (1) Computing the attribute relationships of neighboring genes, and judging whether each and all neighboring genes belong to the same operon; (2) Counting the number of genes in each and all operons in the entire genome.

On the upside, the attribute calculation is convenient, uncomplex, and the relationship between neighboring genes can be identified accurately. On the downside, this method only tackles the neighboring genes, failing to apply to single-gene operons.

The GCP method divides the putative operon gene clusters considering the distance between the operon genes, and then analyze the genes in each cluster to obtain the final predicted operon. This method is quite advantageous in that it can predict single-gene operons and disclose the genetic relationships within the entire operon in a robust manner. However, the GCP method faces high complexity and inconvenience in the computation of some properties. By this method, all subsequent calculations need to be performed separately on the two strands of the genome, because operons only exist on the same strand of the genome.

In the NGPP model, the neighboring genes in the same operon are defined as a pair of neighboring operators (NO pair), and those in different operons as a pair of transcription unit boundaries (TUB pair) [5]. In the GCP model, two genes in the same operon of a gene cluster are defined as an operon pair (OP), while those in different operons of a gene cluster as a non-manipulated pair (NMP). In the cluster of neighboring genes shown in Fig. 2, Operon 1 and Operon 2 are two neighboring operons. For the NGPP model, b-c, c-d, d-e and e-f are NO pairs, and b-a, d-e and f-g are TUB pairs; For the GCP model, b-c, c-d, b-d and e-f are OPs and other gene pairs in the cluster are NMPs.

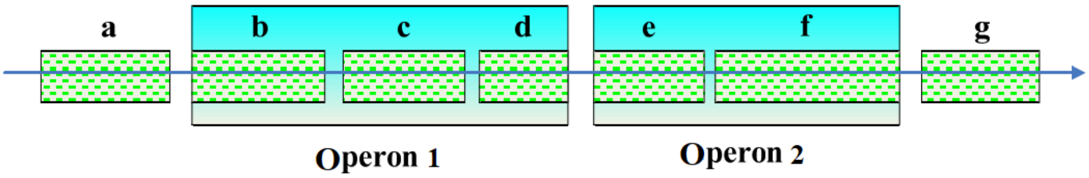

Fig. 2 Neighboring gene cluster

Data preprocessing

The operon prediction involves a dazzling array of attribute information, which differs in range, unit, significant and even form of expression. For example, the intergenic distance (IGD) is described digitally, while the gene ontology is represented by strings. This calls for preprocessing of the data used for prediction. Otherwise, it will be impossible to achieve robust prediction, not to mention outputting accurate results. In the existing studies, the related data are mainly preprocessed by log-likelihood fraction and wavelet transform.

Log-likelihood fraction, in log-likelihood fraction [9], different types of attributes are normalized by computing the relationship between two genes concerning a certain attribute value, and that between two genes in the same operon, in light of two prior probabilities:

$$LL\left(C_1|f\left(g_a,g_b\right)\right)=\log\frac{f\left(g_a,g_b\right)C_2}{f\left(g_a,g_b\right)|C_2}, \tag{1}$$

where $P\left(f\left(g_a,g_b\right)|C_1\right)$ and $P\left(f\left(g_a,g_b\right)|C_2\right)$ are priori probabilities for conditions C_1 and C_2 to satisfy the attribute $f\left(g_a,g_b\right)$, respectively; g_a and g_b are the attribute values of genes *a* and *b* in the gene pair *a-b*, respectively, used for operon prediction; $LL\left(C_1 \mid f\left(g_a,g_b\right)\right)$ is the probability that gene pair *a-b* belongs to C_1 under $f\left(g_a,g_b\right)$.

For the NGPP model, *a* and *b*, as a pair of neighboring genes, is an NO pair and a TUB pair under C_1 and C_2, respectively; for the GCP model, *a* and *b*, as any pair of genes in the gene cluster, is an OP pair and an NMP pair under C_1 and C_2, respectively.

With different metrics of information, some attributes can measure non-neighboring gene pairs, such as conserved gene pairs (CGPs) and gene ontology similarities. Some can only measure neighboring attribute information, namely, the distance between genes and the minimum free energy. For gene cluster prediction, the original log-likelihood fraction formula should be extended as:

$$L\left(g_a,g_b\right) \pm \sqrt{\left|l\left(g_a,g_{c_1}\right)\right| * \left|l\left(g_{c_1},g_{c_2}\right)\right| * \cdots * \left|l\left(g_{c_n},g_b\right)\right|}, \tag{2}$$

where $l\left(x,y\right)$ is a simple form expression of $LL\left(OP \mid f\left(x,y\right)\right)$; $c_1, c_2, \cdots, c_n$ are the genes between gene a and gene b. If all the log-likelihood scores to the right of the equation are positive, the sign of the equation is positive. Otherwise, the sign of the equation is negative. For intergenic distance and minimal free energy, Eq. (1) should be used if the gene pairs in the gene cluster are neighbors, and Eq. (2) should be used if otherwise.

Wavelet transform

Wavelet transform is a novel mathematical approach of data processing like Fourier transform, except for its ability to localize the signal in time and frequency. Since its introduction in 1996, wavelet transform has been widely used in many fields related to bioinformatics [1, 10]. This approach can be adopted to optimize and denoise unwanted attribute information, laying the basis for accurate operon prediction.

The wavelet transform techniques like denoising and compression can be employed to process the attribute information processed by log-likelihood fraction, which may contain much noise information due to the use of probability statistics. Through the processing, the attribute information will become less volatile and more realistic.

The effect of wavelet transform is illustrated in Fig. 3 below, where the dashed line describes the attribute information processed by log-likelihood fraction, and the solid line represents that processed by wavelet transform. Obviously, the attribute information curve becomes smoother, less volatile and more realistic after wavelet transform.

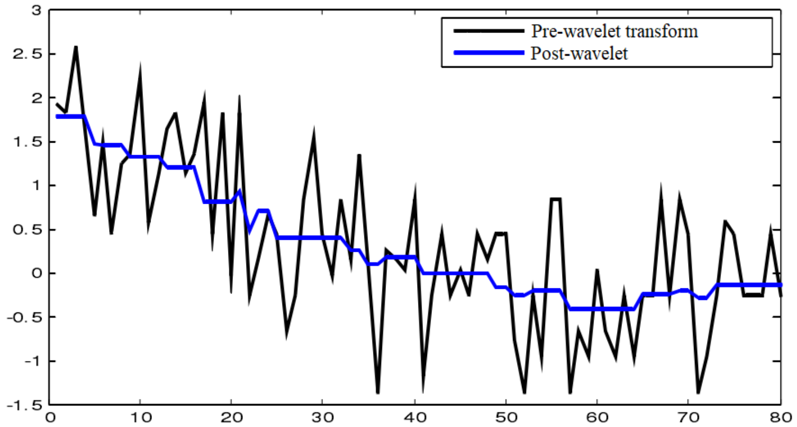

Fig. 3 The effect of wavelet transform

Evaluation of prediction effect

The evaluation of prediction effect is the key to determining the predictive power of the operon prediction algorithm. The most commonly used evaluation indices and tools include sensitivity, specificity, accuracy and receiver operating characteristic (ROC) curve. The first three indices can be defined as follows [6]:

$$\text{Sensitivity} = \frac{TP}{OP}, \tag{3}$$

$$\text{Specificity} = \frac{TP}{TP+FP}, \tag{4}$$

$$\text{Accuracy} = \frac{TP+TN}{OP+TUB}, \tag{5}$$

where *TP* and *TN* are the number of corrected predicted NO pairs and TUB pairs, respectively; *FP* is the number of TUB pairs that are predicted incorrectly as NO pairs; *FN* is the number of NO pairs that are predicted incorrectly as TUB pairs; *OP* and *TUB* are the number of NO pairs and TUB pairs in the genome, respectively.

The ROC curve is a 2D graph with the *FP* value on the abscissa and the *TP* value on the ordinate. The area above the curve is negatively correlated with the prediction effect.

Operon prediction model based on generalized regression neural network

Generalized regression neural network (GRNN)

The artificial neural network (ANN) is a mathematical computational model inspired by biological neural networks. Each ANN consists of a set of interrelated artificial neurons, and completes the computing process by mimicking the interaction of the biological nervous system. The structure of the ANN changes adaptively through input adjustment in the learning phase.

One of the most popular ANNs is the radial basis function (RBF network), which uses the RBF as activation functions. The GRNN is an improved version of the RBF network [2]. As shown in Fig. 4, the GRNN retains most of the features of the RBF network, while removing the weight connection between the hidden layer and the output layer. With strong nonlinear mapping and fast training speed, the GRNN has been a popular tool in function approximation and other research areas.

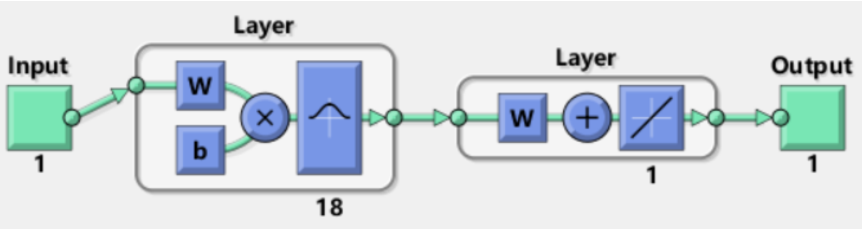

Fig. 4 The structure of GRNN

Description of operon prediction model

Taking neural networks as classifiers, this paper proposes an operon prediction model capable of fusing a variety of attribute information. In this model, several genetic attributes of the genome are selected, the classifier is replaced with the clustering algorithm for operator prediction, and four attributes are subjected to clustering operation, including IGD, CGP, similarity of gene ontology and minimum free energy of gene sequence. The proposed model only uses the calculated attribute information, rather than that derived from experimental data. It can be easily applied to newly-sequencing species, providing biologists with new species of operon information.

The model is implemented in the following steps. To begin with, the position and orthologous cluster function of each gene were obtained from the entire genome that has been sequenced and annotated. Then, the CGP and the phylogenetic information of the gene were calculated based on the existing information. After that, the log-likelihood fraction and wavelet transform were used to process various inputs. Finally, multiple attributes were fused by a GRNN for operon prediction.

To verify its prediction effect, the proposed model was applied to test two kinds of microorganisms: *Escherichia coli* (strain K-12) (*E. coli* K-12) and *Bacillus subtilis* (strain 168) (*B. subtilis* 168). The experimental process is as follows: Firstly, the IGDs of all genes in the two species were calculated using the genome data downloaded from GenBank, and the functional classification information of the orthologous clusters were extracted from the annotation file. Meanwhile, the information on CGP and system evolution spectrum (SES) of the two species were calculated using the 360 genome-wide data of sequenced complete microorganisms. Afterwards, experimentally validated operon data were extracted from RegulonDB and ODB, optimized and denoised by wavelet transform, and introduced to calculate the log-likelihood fractions of the four attributes of both species. Finally, the log-likelihood fractions were taken as the training sample for the proposed operon prediction model.

Specific flow of the operon prediction model

(1) Preprocessing of the input data

On intergenic distance, the distances between all neighboring genes in the predicted genome were calculated by Eq. (1); next, the prior probabilities of NO pair and TUB pair were computed at different IGDs, respectively; afterwards, the log-likelihood fractions of neighboring genes were obtained by Eq. (2) at different IGDs; finally, the log-likelihood fractions were optimized and denoised through wavelet transform.

On orthologous protein cluster (OPC), the functional relationships of the OPCs of neighboring gene pairs were classified by the degree of similarity, using the annotation information of the predicted genome; next, the prior probabilities of NO pair and TUB pair were computed under different similarity classifications; afterwards, the log-likelihood fractions of neighboring genes were obtained by Eq. (2) under different similarity classifications; finally, the log-likelihood fractions were optimized and denoised through wavelet transform.

On CGP, the number of conserved genes in the comparative genomic set was calculated for all pairs of neighboring genes in the predicted genome; next, the prior probabilities of NO pair and TUB pair were computed at different numbers of conserved genes; afterwards, the log-likelihood fractions of neighboring genes were obtained by Eq. (2) at different numbers of conserved genes; finally, the log-likelihood fractions were optimized and denoised through wavelet transform.

On system evolution spectrum, the phylogenetic distances of neighboring gene pairs were calculated after calculating the SES of all the genes in the predicted genome; next, the prior probabilities of NO pair and TUB pair were computed at different phylogenetic spectral distances; afterwards, the log-likelihood fractions of neighboring genes were obtained by Eq. (2) at different phylogenetic spectral distances; finally, the log-likelihood fractions were optimized and denoised through wavelet transform.

(2) GRNN-based operon prediction

The main difficulty in operon prediction lies in the prediction of complex, unknown biometric problems involving various biological attributes. To realize effective prediction, the multiple attributes should be fused correctly and efficiently [7]. With strong nonlinear mapping ability and fast training speed, the GRNN provides an ideal solution to the fusion of various attributes, laying a solid basis for operon prediction. Compared with the RBFNN, the GRNN has an additional linear output layer beyond the output layer. Here, an operon prediction model is set up based on the GRNN, and implemented on the Matlab. Fig. 5 shows the simplified structure of the GRNN adopted in our model.

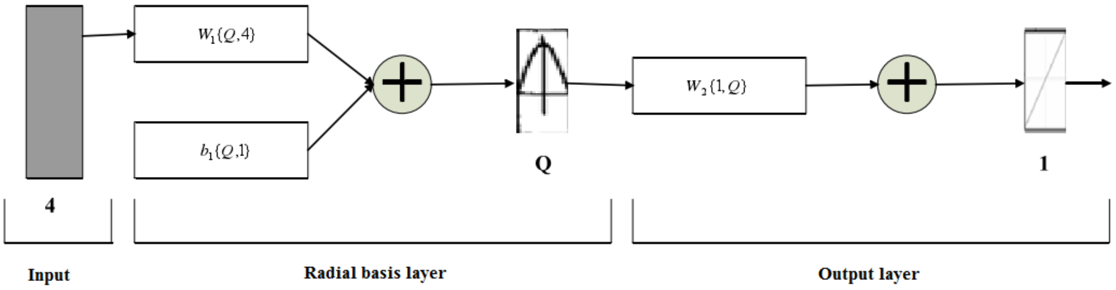

Fig. 5 The structure of the GRNN in our model

There are four nodes on the input layer of the GRNN, which correspond to the four genome attributes obtained through data preprocessing. Meanwhile, the output layer of the network has only one node, i.e. a probability within [0, 1]. This output indicates whether the neighboring gene pair belongs to the same person. In addition, the GRNN contains only one hidden layer. The number of hidden layer nodes is usually set the same as that of input vectors. However, this setting is not feasible if there are so many input vectors as to dampen the network performance. Thus, the number of hidden layer nodes in our model was determined iteratively: gradually increasing the number of hidden layer nodes from one to the number of input vectors, finding the number that minimizes the output error and taking this number as the number of hidden layer nodes.

Experimental results

The proposed GRNN operon prediction model was verified through an experiment on *E. coli* K-12 and *B. subtilis* 168. The 1/2 operon of the two species was selected as the training set, and the other 1/2 operon as the test set for cross-checking. The mean sensitivity, specificity and accuracy of the prediction model on *E. coli* were 88.6%, 89.2% and 88.9%, respectively, and 87.4%, 85.5% and 86.3% on *B. subtilis*, respectively. For comparison, joint prediction of operons (JPOP) and operon finding software (OFS) were also tested on the same dataset of *E. coli* K-12 and *B. subtilis* 168.

The predicted results are shown in Tables 1 and 2 below. Clearly, the GRNN model outperformed both JPOP and OFS in the sensitivity, specificity, and accuracy of the operon prediction results.

Table 1. The prediction results of three methods on *E. coli* K-12

Prediction method	**Sensitivity**	**Specificity**	**Accuracy**
JPOP	84.5%	83.8%	86.1%
OFS	85.7%	84.9%	86.5%
The proposed model	88.1%	89.5%	88.3%

Table 2. The prediction results of three methods on *B. subtilis* 168

Prediction method	**Sensitivity**	**Specificity**	**Accuracy**
JPOP	83.2%	80.9%	82.7%
OFS	85.1%	81.3%	82.6%
The proposed model	87.6%	85.2%	86.6%

One of the four attributes was removed in turns and compared with the previous prediction results, aiming to test the role of each attribute in operon prediction. The results in Table 3 show that the best sensitivity, specificity, and accuracy of the prediction appeared when all four attributes were considered. In addition, the prediction effect was the worst when the IGD information was deleted, revealing that IGD has an important impact on the operon. Thus, the coefficient of IGD should be increased in the prediction model. The prediction effect did not change much at the removal of the SES. A possible reason lies in the fact that the SES is partially covered by the system gene profile and the conserved gene. Furthermore, the SES distance used in this section is a simple Hamming distance, which can be further improved to enhance the prediction effect of our model.

Table 3. Comparison between four-attribute prediction and three-attribute predictions

Information	**Sensitivity**	**Specificity**	**Accuracy**
Excluding intergenic distance	75.4%	85.7%	83.1%
Excluding OPC	84.2%	87.3%	86.1%
Excluding CGP	85.6%	87.9%	86.3%
Excluding SES	86.1%	88.1%	87.3%
All four attributes	88.5%	89.7%	88.8%

The prediction effect of our model was further evaluated by comparing its ROC curve with that of the OFS and that of the JPOP (Fig. 6).

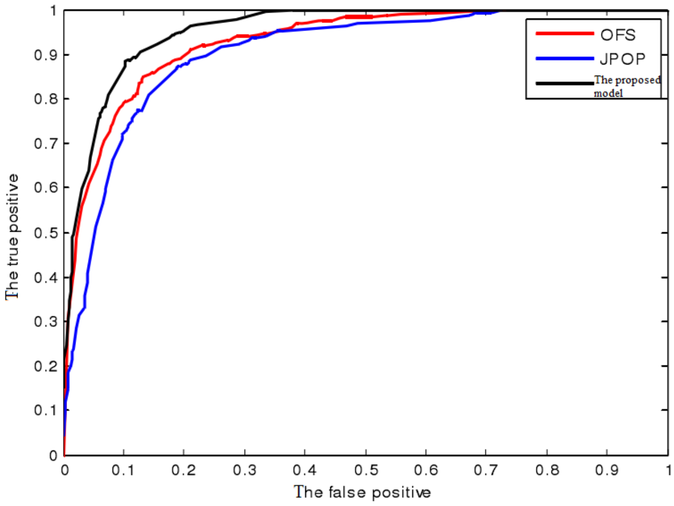

Fig. 6 ROC curves of three different methods

As shown in Fig. 6, our model had a smaller area above the ROC curve than the other two methods, an evidence of good mean sensitivity and specificity. Figs. 7 and 8 respectively show the ROC curves on the two species when our model is adopted for prediction using all four attributes and one of the four attributes. It can be seen that the prediction results in the case of all four attributes were much better than the single-attribute cases.

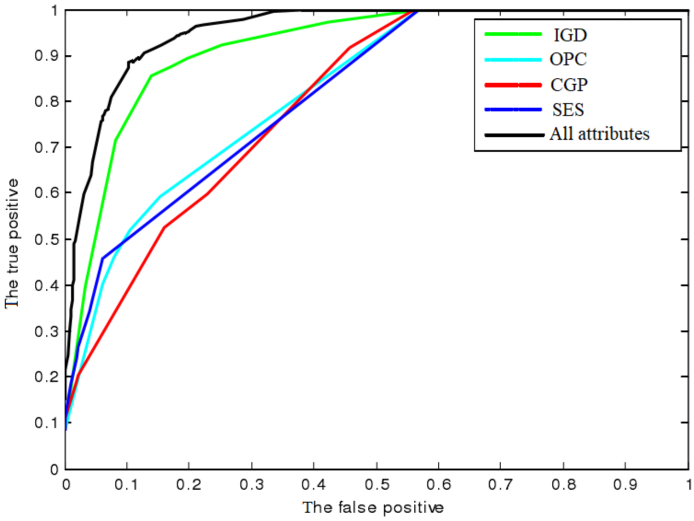

Fig. 7 ROC curves on *E. coli* K-12 under different attribute combinations

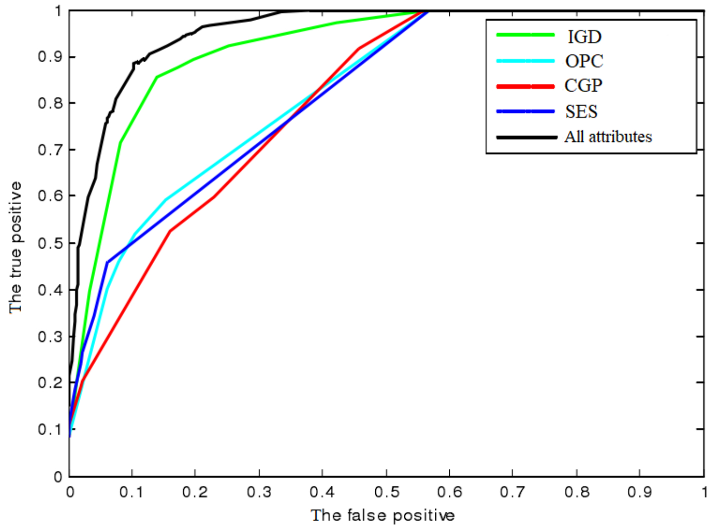

Fig. 8 ROC curves on *B. subtilis* 168 under different attribute combinations

Conclusion

Operon is a basic transcription unit in complex biological processes of microorganisms. It provides many valuable information in such field as biopharmaceutics, protein function, and biological regulation mechanism. In view of these, the author put forward an ANN operon prediction model, which relies on the GRNN and four attributes (i.e. IGD, OPC, CGP and SES) to realize operon prediction. The information of the four attributes were preprocessed by log-likelihood fraction and wavelet transform, and then inputted to the GRNN for operon prediction. The experimental results in *E. coli* K-12 and *B. subtilis* 168 show that our model is a valid and feasible way to predict operon.

References

1. Arneodo A., B. Audit, E. Bacry (2017). Thermodynamics of Fractal Signals Based on Wavelet Analysis: Application to Fully Developed Turbulence Data and DNA Sequences, Physica A Statistical Mechanics & Its Applications, 254(1), 24-45.
2. Cigizoglu H. K., M. Alp (2006). Generalized Regression Neural Network in Modelling River Sediment Yield, Advances in Engineering Software, 37(2), 63-68.
3. Klappenbach J. A., P. R. Saxman, J. R. Cole (2001). RRNDB: The Ribosomal RNA Operon Copy Number Database, Nucleic Acids Research, 29(1), 181-184.
4. Kubota T., Y. Tanaka, N. Takemoto (2014). Chorismate-dependent Transcriptional Regulation of Quinate/shikimate Utilization Genes by LysR-type Transcriptional Regulator QsuR in *Corynebacterium glutamicum*: Carbon Flow Control at Metabolic Branch Point, Molecular Microbiology, 92(2), 356-368.
5. Mao X., V. Olman, R. Stuart (2010). Computational Prediction of the Osmoregulation Network in *Synechococcus* sp. WH8102, BMC Genomics, 72(5), 291-291.
6. Petrovska-Delacretaz D., S. Lelandais, J. Colineau (2007). Operon Prediction Using Both Genome-specific and General Genome Information, Nucleic Acids Research, 35(1), 288-298.
7. Ranjan S., R. K. Gundu, A. Ranjan (2006). MycoperonDB: A Database of Computationally Identified Operons and Transcriptional Units in *Mycobacteria*, BMC Bioinformatics, 7(5), 9-16.
8. Sáenz-Mata J., F. B. Salazar-Badillo, J. F. Jiménez-Bremont (2014). Transcriptional Regulation of *Arabidopsis thaliana*, WRKY, Genes Under Interaction with Beneficial

Fungus *Trichoderma atroviride*, Acta Physiologiae Plantarum, 36(5), 1085-1093.

9. Salgado H., G. Moreno-Hagelsieb, T. F. Smith (2000). Operons in *Escherichia coli*: Genomic Analyses and Predictions, Proceedings of the National Academy of Sciences of the United States of America, 97(12), 6652-6657.
10. Tsonis A. A., P. Kumar, J. B. Elsner (1996). Wavelet Analysis of DNA Sequences, Physical Review E Statistical Physics Plasmas Fluids & Related Interdisciplinary Topics, 53(2), 1828-1839.
11. Vemuri G. N., E. Altman, D. P. Sangurdekar (2006). Overflow Metabolism in *Escherichia coli* during Steady-state Growth: Transcriptional Regulation and Effect of the Redox Ratio, Applied & Environmental Microbiology, 72(5), 3653-3661.
12. Zhang Z., Y. Liang (2019). Operon Prediction Model Based on Markov Clustering Algorithm, International Journal Bioautomation, 23(1), 105-116.

Hui Li, M.Sc.
E-mail: lihui@zzuli.edu.cn

Hui Li received her Master’s degree in Pattern Recognition and Intelligent System from Zhengzhou University, Zhengzhou, China, in 2007. She is a lecturer at Zhengzhou University of Light Industry. Her research interests include intelligent control, data mining and data processing.

© 2019 by the authors. Licensee Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).