Operon Prediction Model Based on Markov Clustering Algorithm

Zhenmei Zhang[1,2], Yongquan Liang[1*]

[1]*College of Computer Science and Engineering*
Shandong University of Science and Technology
Qingdao 266590, China
E-mail: zhangzhenmei@sdufe.edu.cn

[2]*College of Management Science and Engineering*
Shandong University of Finance and Economics
Jinan 250014, China
E-mail: lyq@sdust.edu.cn

[*]*Corresponding author*

Received: April 10, 2018 **Accepted: March 05, 2019**

Published: March 31, 2019

Abstract: *There are many operon prediction models, but few methods can be applied to the operon prediction of new sequencing species effectively. In this paper, an operon prediction model based on Markov clustering algorithm is proposed. The model uses some generic attribute information of genomes and graph clustering algorithm instead of classifier to predict operon. Similarly to most operon prediction models, E. coli K12 and B. subtilis str. 168 were used to assess the prediction capability of the proposed model, the experiment results show that the proposed model has better capability of operon prediction than some other classical operon prediction methods.*

Keywords: *Operon Prediction, Markov Clustering Algorithm, Generic Attribute, Graph Clustering.*

Introduction

With the booming of the human genome project, bioinformatics is developing rapidly. In this process, people have realized that the construction of gene expression regulatory network is the key to reveal the mystery of life. Prediction of operon as a predictive gene expression regulatory network has attracted wide attention from researchers [2, 5, 8].

The concept of operon comes from the theory of protein synthesis regulatory mechanism proposed by Yaniv [16]. The theory indicates that operon is a basic unit related to transcription. Its characteristics mainly include:

1) One operator contains one or more genes, and its transcription direction is consistent.
2) The distance between adjacent genes in operon is less than that between different operon genes.
3) There is a promoter in the upstream of an operon, and there is a terminator in downstream, and there is usually no promoter or terminator inside it.
4) The genes in the operon are functionally related and belong to the same functional classification.

The prediction of the operon can provide reference for people to recognize and construct biochemical and metabolic networks [10], and provide information for the study of biopharmaceuticals, protein functions and regulatory mechanisms.

The bioinformatics methods for operon prediction can be classified into two categories: training set method and independent training set method. Training set method mainly includes neural network method and machine learning method. Neural network has nonlinear transformation function. In the prediction of operon, the required data source can be extracted from the known operon. Each data source is used as input sample to train the neural network after processing. The trained network was applied to the genome studied and the operon of the genome was predicted. Radakovits et al. [14] use neural network to train each data source, and the prediction results are more accurate than the single data sources. But the neural network of the single hidden layer is less capable of dealing with the nonlinear separable classification problem. In general, the combination of multiple sources makes the problem not linearly separable, so it is necessary to further improve the network structure to get better results. Machine learning algorithm also plays an important role in operon prediction. Zaidi and Zhang [17] used naive Bayes network to predict operon, and combined with dynamic programming algorithm to improve prediction accuracy and sensitivity. Du et al. [4] apply Markov model to predict operon. Neural network algorithm and machine learning algorithm are trained by training set, combining multiple data sources instead of simple linear superposition, so the sensitivity of prediction results is improved. But they all rely too much on data sets, it makes the algorithm less universal.

Comparative genomic approach is a common independent training set method. The method is based on the following hypothesis: if a series of sequences and functions can be preserved in the genome, the reorganized gene may be in the same operon. Radakovits et al. [14] find the conservative gene pairs by comparison of the genome in the process of prediction, and divide the results according to certain rules. The forecast results are arranged in order according to the score. The comparative genome approach is less dependent on the experimental data and the training set, but it loses the unique information on the genome, which affects the sensitivity of the algorithm. Genetic algorithm can simulate the process of biological evolution and find the optimal solution or quasi optimal solution of the problem domain [18, 19]. In the prediction of operon, the combination of genetic algorithm and other intelligent computing methods can improve the prediction accuracy. Jacob et al. [7] used the data sources such as inter-genic distance, metabolic network, conservative gene pair and gene function annotation, combining genetic algorithm with fuzzy rules, to cluster all the genes according to certain rules. The operon prediction using genetic algorithm is not dependent on the training set, and it is intelligent to cluster the characteristics of the gene cluster operon, and the accuracy and sensitivity are better, but the definition of the fitness function is more complex.

Operon prediction related attribute information

We can use one or more genome information to predict operon. Generally speaking, the more types of function and attribute information are used, the better prediction results can be achieved. But more attribute information does not necessarily have a better prediction effect. Therefore, the right choice of attribute information has an important influence on the prediction effect of operon. The common attribute information which used to predict operon is as follows.

Inter-genic distance

Inter-genic distance is one of the common attributes of operon prediction [6, 9, 12]. The inter-genic distance is the distance between two genes in a genome sequence, usually is the number of base pairs with two genes. The inter-genic distance can be accurately calculated based on the location and termination position of genes in genome annotation information. The inter-genic distance between neighborhood genes g_a and g_b can be defined as follows:

$$d(g_a, g_b) = |s(g_b) - e(g_a) + 1|,$$

where g_a and g_b are two neighborhood genes, $s(g_b)$ and $e(g_a)$ denote the starting position and ending position of gene g respectively. $d(g_a, g_b)$ represents the distance between gene g_a and gene g_b.

Conserved genes clusters

Previous studies have shown that genes belonging to the same operon are conserved among multiple genomes [1, 3]. Conserved gene pairs refer to two neighborhood genes x_1 and y_1 on the same chain of biological genomes. If there is another two neighborhood genes x_2 and y_2 on the same chain as the comparative genome, and the sequence similarity between gene x_1 and x_2, gene y_1 and y_2 meets the requirements. At the same time, the similarity between them is higher than that between x_1 and y_1, x_2 and y_2, so x_1 and y_1 have conserved gene pairs in this comparative genome. Fig. 1 is an example of a conserved gene cluster between Genome X and Genome Y_1 to Y_n.

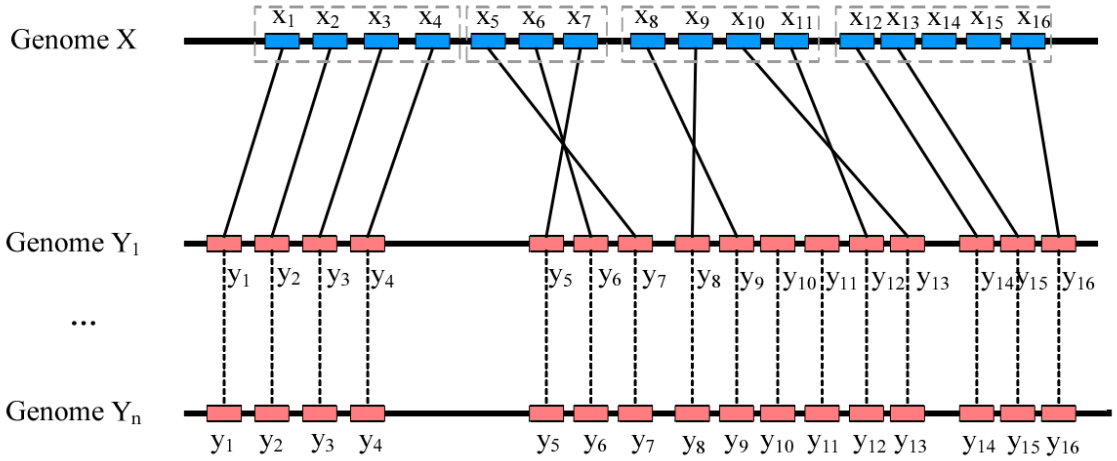

Fig. 1 Schematic diagram of conserved genes clusters

In this example, the Genome X has 16 genes. At the same time, the Genome Y_1 to Y_n also contains 16 genes. There are 14 orthologous genes between Genome X and Y_1. These orthologous genes belong to 4 conserved gene clusters. Gene cluster $x_1x_2x_3x_4$ and $y_1y_2y_3y_4$ constitute the identical conservative gene cluster.

Phylogenetic spectrum

The phylogenetic spectrum of a gene is a binary string, and the phylogenetic spectrum of the genome is a binary matrix. Each location of the matrix indicates whether the gene of the row has similar genes in the genome of the column. If there is a similar gene, the location is 1. If there is no similar gene, the location is 0.

Phylogenetic spectrum can provide information of homologous genes in evolution, it can reflect the similarity of gene function classification and metabolic pathway. Fig. 2 shows the phylogenetic spectrum of species. Species II, III and IV are genomes for comparative analysis. As shown in Fig. 2, the phylogenetic relationship of gene g_c and g_d in species I is exactly the

same, so their functional classification and metabolic pathways may also have high correlation, and they are most likely to belong to the same operon.

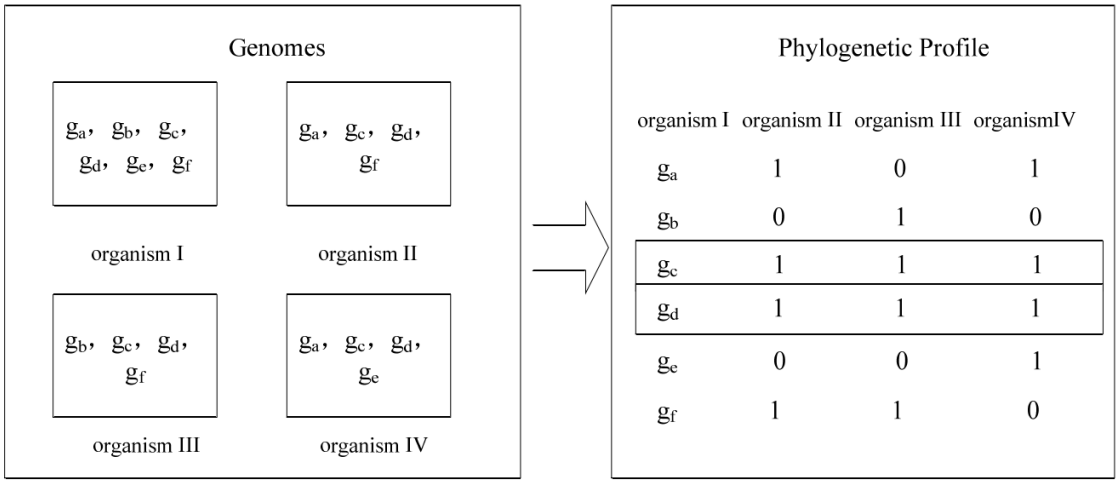

Fig. 2 Simple phylogenetic spectrum

Operon prediction model based on graph clustering method

Model description

Using data from existing databases, operons of newly sequenced species can be effectively predicted. However, only a few researchers have proposed methods that can be effectively applied to the operon prediction of newly sequenced species. Due to the use of inter-genomic specific attribute information and over-fitting classifiers, most current methods do not have good generalization capabilities for operon prediction of new species. In the paper, we propose a graph clustering model using Markov clustering algorithm for operon prediction. The model uses some generic attribute information of the genome while using a clustering algorithm instead of a classifier. The model performs clustering operations based on four kinds of attribute information: inter-genic distances, conserved gene clusters, gene ontology similarities, and minimum free energy of inter-genic sequences. The model differs from the existing operon prediction models and methods in that gene clusters are used to perform operon predictions in place of existing neighboring gene pairs.

Genes belonging to the same operon have the same transcription direction and shorter inter-genic distances. Therefore, the genes on the same strand are firstly processed and operon candidate gene clusters are generated according to the inter-genic distance. Then the four kinds of attribute information of gene pair distances, conserved gene clusters, gene ontology similarities, and minimum free energy of inter-genic sequences in each gene cluster were calculated. Afterwards, four kinds of attribute information values of gene pairs in each gene cluster are processed using log-likelihood scores. Finally, the final operon information is obtained from the candidate operon cluster using the Markov clustering algorithm. As with most operon prediction methods, the microbial *E. coli* K12 and *B. subtilis* str. 168 were used evaluations to propose the model's ability to predict in a single genome. The results showed that the average sensitivity, specificity, and accuracy were 92.9%, 90.2%, 91.7%, and 89.9%, 88.4%, and 89.1%, respectively. Experimental results show that the proposed model can effectively predict operons, and the prediction ability is better than other common operon prediction programs such as JPOP [14], OFS [15], MA-GA [13]. Although the predicted results are slightly different from the single-genome test on *E. coli* and *B. subtilis* by using specific and global genomic information, the method has better generalization ability in multi-species cross-

validation. For example, using a constructed model to perform tests on *P. furiosus* is better than other existing methods. The test results show that operon prediction model not only has a good effect in single species testing, but also can obtain effective results in new species.

Model specific process

The operon can be considered as a special gene cluster, so the genome is divided into candidate operon clusters using special rules. In the previous related studies, a gene clustering method based on positive and negative strands was proposed. The method generates candidate operon gene clusters by dividing adjacent genes on the same strand into the same gene cluster. The advantages of this method are simple and fast. However, the DBTBS database shows that the inter-genic sequences within one operon of *B. subtilis* may contain other operons on the corresponding genomic strand. For example, sigA, dnaG, antE, and yqxD are the four genes in the genome. Their positions in the genome are - - + -, where "+" denotes the positive strand and "-"denotes the negative strand. In *B. subtilis*, sigA, dnaG, and yqxD belong to the same operon. Obviously, it can be seen that the use of positive and negative strands for gene clustering does not correctly predict this type of operon.

In the paper, a novel method for generating operon candidate gene clusters using inter-genic distances is proposed. Since the operon is a transcription unit, genes within one operon are expected to be closer to each other. From existing operon databases, such as RegulonDB, DBTBS and DOOR, it can be found that the inter-genic distances of neighboring gene pairs within all operons are less than 600 bp. Obviously, the attribute is very important for operon prediction.

After obtaining the log-likelihood scores of the four kind of attribute information in all candidate operon gene clusters, the weight of each attribute is set to 1, and then the four attributes are input into the Markov clustering algorithm to predict the final operators. In experiment, the expansion rate and expansion rate of the Markov clustering algorithm were set to 2 and 1.2, respectively. Fig. 3 is a schematic diagram of operon prediction using the Markov clustering algorithm.

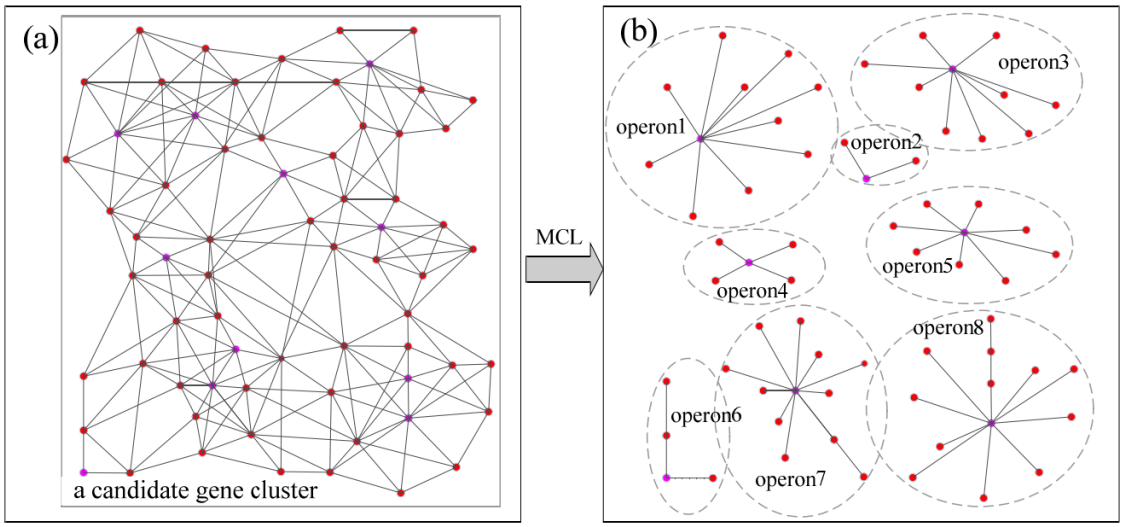

Fig. 3 A schematic diagram of the operon prediction in Markov clustering algorithm

The vertices in Fig. 3 represent genes, and the edges represent the relationships between genes. Fig. 3(a) shows the relationship between genes in a hypothetical candidate operon cluster.

Fig. 3(b) shows predicted operon results using the Markov clustering algorithm for this putative gene cluster.

Experimental results and analysis

Matlab is used to implement and simulate the proposed model. The effectiveness of the operon prediction verification algorithm was tested by using predictive models on three species, *E. coli* K12, *B. subtilis*, and *P. furiosus*. It verifies the performance of the operon method proposed in this paper. The model uses a graph clustering algorithm instead of the commonly used classifier for operon prediction. However, log-likelihood fractions are based on species, so single-species experiments and cross-species experiments were performed to verify the validity of the proposed algorithm. The corresponding operon information for *E. coli* is obtained from the RegulonDB database. The corresponding operon information for *B. subtilis* was obtained from the DBTBS database. The data of micrococcus assays for *Pyrococcu*s were derived from published studies [11]. Therefore, *E. coli* and *B. subtilis* were used as predictive species in a single species experiment, and the above three species were used as predicted species in a cross experiment.

The role of four kinds of attribute information

In order to evaluate the effects of the four genomes attribute information, the frequency distributions of operon pairs and transcription unit boundaries on different inter-genic distances, different numbers of conserved gene clusters, different gene ontologies similarity scores, and different minimum free energies were counted. It is estimated that operon pairs will have smaller inter-genic distances, greater numbers of conserved gene clusters, higher oncogene functional similarities, and greater inter-gene sequence minimum free energy than transcriptional unit boundary pairs.

Fig. 4 shows the distribution frequency of the operon pairs and transcription unit boundary pairs under different gene distances.

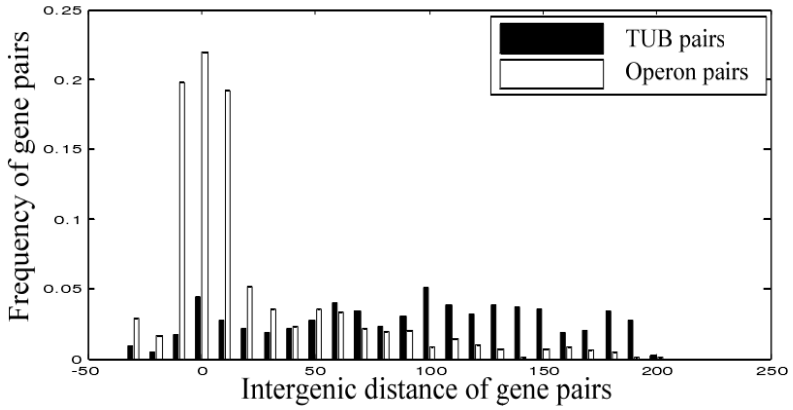

Fig. 4 Frequency distribution under different gene distances

From Fig. 4, it can be found that the inter-genic distances of most operon pairs are between -10 bp and 20 bp, while at the same time the distance between most transcription unit boundaries is greater than 50 bp.

Fig. 5 shows the distribution frequency of operon pairs and transcription unit boundary pairs under different conserved gene cluster numbers.

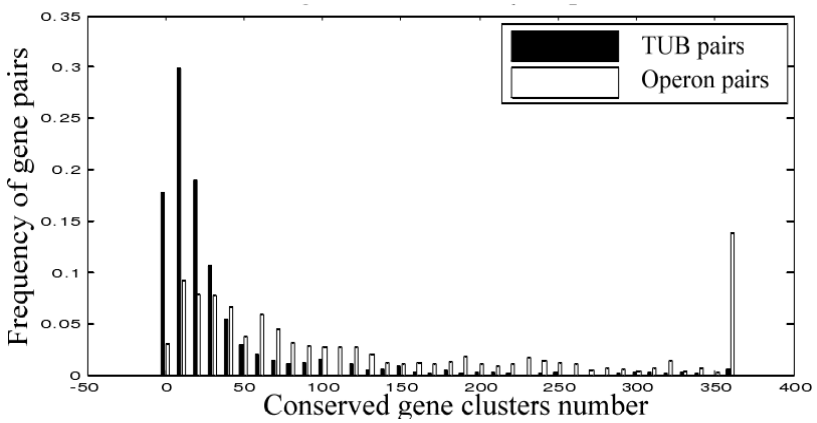

Fig. 5 Frequency distribution under different conserved gene cluster numbers

From Fig. 5, it can be found that the operon pair has a very high number of conserved gene clusters and the number of conserved gene clusters at the transcriptional cell boundary pair is small. Therefore, the number of conserved gene clusters can correctly determine whether two gene pairs belong to one operon.

Fig. 6 shows the distribution frequency of operon pairs and transcription unit boundary pairs under different gene ontology similarity scores.

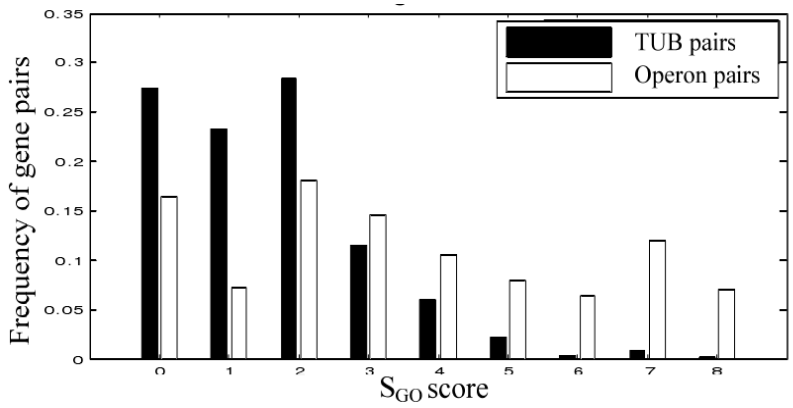

Fig. 6 Frequency distribution under different gene ontology similarity scores

From Fig. 6, it can be found that the gene ontology similarity score of most operon pairs is greater than 3, but the gene ontology similarity score of most transcription unit boundary pairs is less than 3.

Fig. 7 shows the distribution frequency of operon pairs and transcription unit boundary pairs under different inter-gene sequence minimum free energies. From Fig. 7, it can be found that the minimum free energy of the inter-genic sequences of most operon pairs is greater than -4, but the minimum free energy of the inter-genic sequences of most transcription unit boundaries is less than -4.

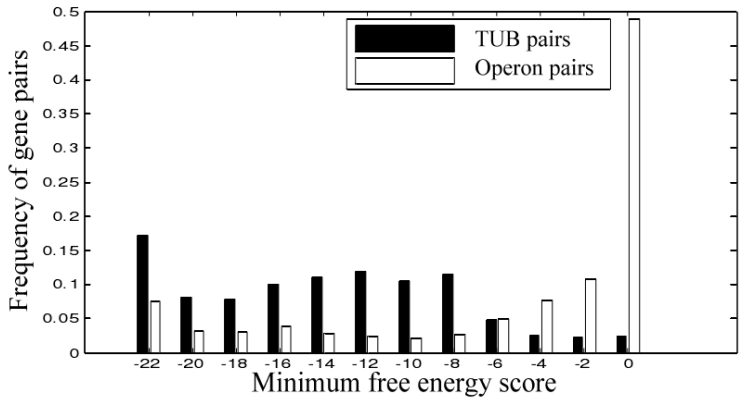

Fig. 7 Frequency distribution under different inter-gene sequence minimum free energies

Single species validation

In the process of validation, the prior probability of the log-likelihood score is calculated by the predicted species itself. Half of the species' operators were used to calculate log-likelihood scores, while the other half of the operons was used to evaluate the effects of operon predictions. The predicted average sensitivity, specificity and accuracy in *E. coli* and *B. subtilis* were 91.3%, 90.5%, 92.3% and 88.7%, 87.8%, and 89.6%, respectively. In order to compare with existing algorithms, the two species were predicted using the existing JPOP, OFS, MA-GA and specific and global genomic information methods. The prediction results are shown in Table 1 and Table 2.

Table 1. Five methods for validation of single species on *E. coli*

Prediction method	Sensitivity	Specificity	Accuracy
JPOP	84.8%	83.5%	84.7%
OFS	85.9%	84.6%	85.5%
MA-GA	88.7%	83.2%	86.2%
UGSGG	92.6%	91.7%	92.6%
The proposed model	91.3%	90.5%	92.3%

Table 2. Five methods for validation of single species on *B. subtilis*

Prediction method	Sensitivity	Specificity	Accuracy
JPOP	84.2%	80.7%	83.1%
OFS	85.1%	80.6%	83.7%
MA-GA	87.2%	87.8%	87.3%
UGSGG	90.1%	88.5%	90.3%
The proposed model	88.7%	87.8%	89.6%

It can be seen from Table 1 and Table 2, the average sensitivity, specificity, and accuracy of the proposed model are better than JPOP, OFS, and MA-GA. The results of the prediction are slightly different than the effects of single genome tests predicted using *E. coli* and *B. subtilis* using specific and global genomic information methods. It may be due to the fact that our proposed model doesn't use a classifier. In the paper, log-likelihood scores are used to train operon predictions using genomic information from the same species as a training set.

Multi-species validation

In the process of validation, the prior probability of the log-likelihood score was calculated from species different from the predicted species. The prior probability used to predict the log-likelihood fraction of the *E. coli* operon was calculated from *B. subtilis*, whereas the calculation method was the same. The average sensitivity, specificity, and accuracy of species crossover prediction in *E. coli* and *B. subtilis* were 90.6%, 91.1%, 91.3%, and 83.8%, 88.6%, and 86.8%, respectively. The operon prediction algorithm was also validated on the new genome. However, in addition to *E. coli* and *B. subtilis*, due to the lack of information on known operons from other species, gene chip-validated *Pyrococcus* operon information was used to validate the proposed method's ability to predict new species. The average sensitivity, specificity, and accuracy of the three methods for *E. coli*, *B. subtilis*, and *P. aureus* are shown in Table 3, Table 4, and Table 5.

Table 3. Prediction results of species cross validation operon on *E. coli*

Prediction method	Sensitivity	Specificity	Accuracy
OFS	86.2%	91.3%	88.2%
UGSGG	88.5%	91.8%	90.2%
The proposed model	90.6%	91.1%	91.3%

Table 4. Prediction results of species cross validation operon on *B. subtilis*

Prediction method	Sensitivity	Specificity	Accuracy
OFS	90.8%	83.1%	81.1%
UGSGG	80.2%	86.3%	82.3%
The proposed model	83.8%	88.6%	86.8%

Table 5. Prediction results of species cross validation operon on *P. aureus*

Prediction method	Sensitivity	Specificity	Accuracy
OFS	75.3%	77.2%	76.2%
UGSGG	87.5%	73.2%	82.5%
The proposed model[a]	87.9%	81.5%	82.9%
The proposed model[b]	88.7%	82.6%	85.9%

[a] log-likelihood fraction calculation from *E. coli*
[b] log-likelihood fraction calculation from *B. subtilis*

From Tables 3-5, it can be seen that the average sensitivity, specificity, and accuracy of the proposed model prediction results are better than OFS and the use of specific and global genomic information methods. The experimental results prove that our proposed operon prediction model has strong capabilities for the operon prediction of new species. The effect of the prediction is better than the existing OFS methods and the use of specific and global genomic information methods.

In order to further evaluate the effectiveness of the proposed operon prediction model, the results of operon prediction using single attribute information and using all four information about attribute were compared. Fig. 8 shows the ROC curve for operon prediction using the single attribute and all four information about attribute in three species.

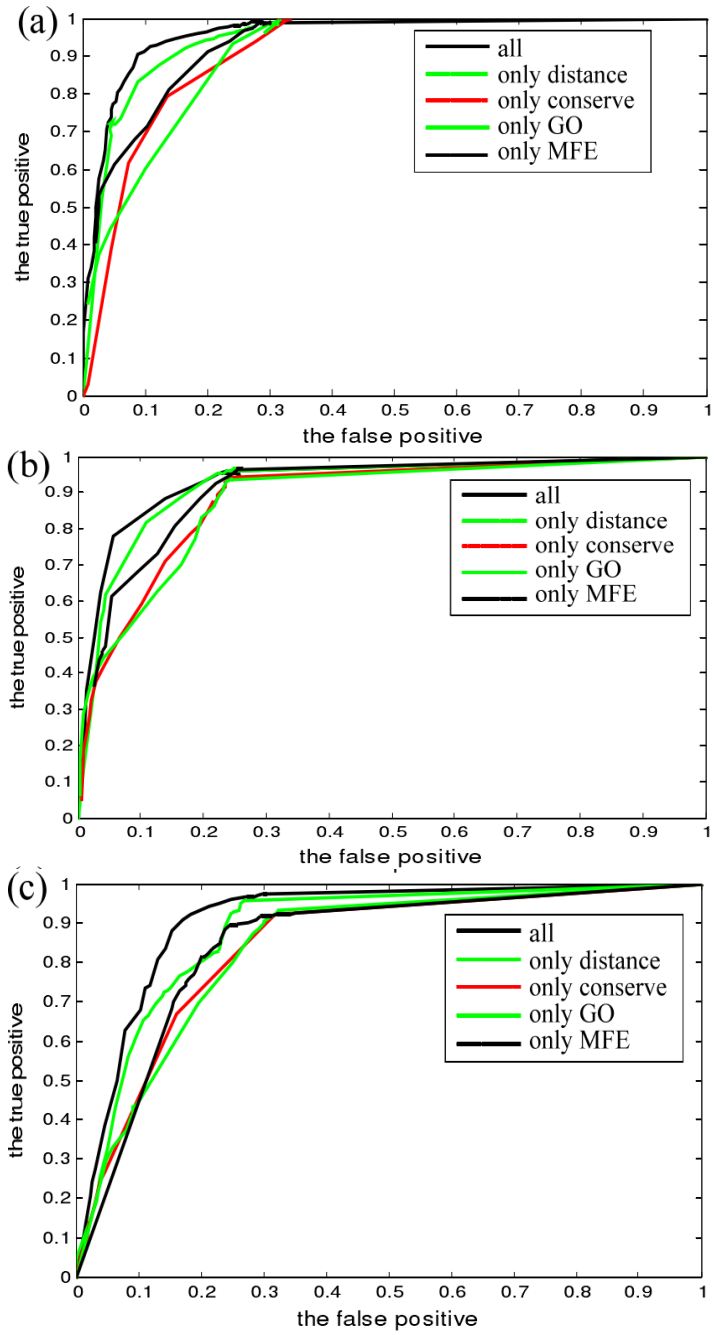

Fig. 8 ROC curve of using a single attribute and all attributes
(a) *E. coli*; (b) *B. subtilis*; (c) *P. aureus*

It can be seen from the Fig. 8, the information for operon prediction results using of all attribute information much better than using a single attribute information operator to predict the results. Using the inter-genic distance alone as the attribute information to perform operon predictions is better than using other properties alone. Therefore, two genes are closely related to the distance between genes and whether they belong to one operon. In later studies, the coefficient can be increased appropriately.

Conclusion

The operon is the basic transcription unit in the microbial complex biological process. It provides much valuable information on biopharmaceuticals, protein functions, and biological

regulation mechanisms. In this paper an operon prediction model based on graph clustering algorithm is proposed. The model is based on the Markov clustering algorithm and uses the inter-gene distance, conserved gene clusters, gene ontology similarity, and minimum free energy information of inter-gene sequences for operon prediction. The model differs from the existing operon prediction model and method in that gene clusters are used instead of existing neighboring gene pairs, and graph clustering algorithms are used in place of currently used classifiers for operon prediction. Experimental results show that the model can effectively predict operons, and the prediction ability is better than other common operon prediction methods such as JPOP, OFS, and MA-GA.

Acknowledgements

Fund Project: The State Key Research Development Program of China under Grant 2016YFC0801403, Shandong Provincial Natural Science Foundation of China under Grant ZR2018MF009 and ZR2015FM013, the Special Funds of Taishan Scholars Construction Project, and Leading Talent Project of Shandong University of Science and Technology.

References

1. Bratlie M. S., J. Johansen, F. Drabls (2010). Relationship between Operon Preference and Functional Properties of Persistent Genes in Bacterial Genomes, BMC Genomics, 11(1), 71-80.
2. Chitsaz H., J. L. Yeegreenbaum, G. Tesler (2011). De Novo Assembly of Bacterial Genomes from Single Cells, Nature Biotechnology, 29(10), 915-923.
3. Ding Y., X. Liu, F. Chen (2014). Metabolic Sensor Governing Bacterial Virulence in *Staphylococcus aureus*, Proceedings of the National Academy of Sciences of the United States of America, 111(46), E4981.
4. Du W., Z. Cao, Y. Wang (2014). Operon Prediction by Markov Clustering, International Journal of Data Mining & Bioinformatics, 9(4), 424-436.
5. Giovannini D., G. Cappelli, L. Jiang (2012). A New *Mycobacterium tuberculosis*, Smooth Colony Reduces Growth inside Human Macrophages and Represses PDIM Operon Gene Expression. Does an Heterogeneous Population Exist in Intracellular Mycobacteria?, Microbial Pathogenesis, 53(3), 135-146.
6. Gostick D. O., H. G. Griffin, C. A. Shearman (2010). Two Operons that Encode FNR-like Proteins in *Lactococcus lactis*, Molecular Microbiology, 31(5), 1532-1535.
7. Jacob E., R. Sasikumar, K. N. R. Nair (2005). A Fuzzy Guided Genetic Algorithm for Operon Prediction, Bioinformatics, 21(8), 1403-1407.
8. Krushkal J., R. M. Adkins, Y. Qu (2010). Bioinformatic Analysis of Gene Regulation in the Metal-reducing Bacterial Family *Geobacteraceae*, BMC Bioinformatics, 11(Suppl 4):P11.
9. Luo Y., D. Mcshan, M. Matuszak (2016). We-ab-207b-02: A Bayesian Network Approach for Joint Prediction of Tumor Control and Radiation Pneumonitis (rp) in Non-small-cell Lung Cancer (NSCLC), Medical Physics, 43(6), 3804-3804.
10. Maclean R. C. (2010). Predicting Epistasis: An Experimental Test of Metabolic Control Theory with Bacterial Transcription and Translation, Journal of Evolutionary Biology, 23(3), 488-493.
11. Manjasetty B. A., M. R. Chance, S. K. Burley (2014). Crystal Structure of *Clostridium acetobutylicum* Aspartate Kinase (CaAk): An Important Allosteric Enzyme for Amino Acids Production, Biotechnology Reports, 3, 73-85.
12. Matsutani M., M. Ogawa, N. Takaoka (2013). Complete Genomic DNA Sequence of the East Asian Spotted Fever Disease Agent Rickettsia Japonica, PloS ONE, 8(9), e71861.

13. Moharana K. C., M. R. Dikhit, B. R. Sahoo (2015). GAOPP: Operon Prediction in Prokaryotes Using Genetic Algorithm, Current Bioinformatics, 10(3), 299-305.
14. Radakovits R., R. E. Jinkerson, S. I. Fuerstenberg (2013). Draft Genome Sequence and Genetic Transformation of the Oleaginous Alga *Nannochloropsis gaditana*, Nat Commun, 4:2356.
15. Westover B. P., J. D. Buhler, J. L. Sonnenburg (2005). Operon Prediction without a Training Set, Bioinformatics, 21(7), 880-888.
16. Yaniv M. (2011). The 50th Anniversary of the Publication of the Operon Theory in the Journal of Molecular Biology: Past, Present and Future, Journal of Molecular Biology, 409(1), 1-6.
17. Zaidi S. S. A., X. Zhang (2016). Computational Operon Prediction in Whole-genomes and Metagenomes, Briefings in Functional Genomics, 16(4), 181-193.
18. Zeng J., L. Yong (2017). The Use of Adaptive Genetic Algorithm for Detecting Kiwifruit's Variant Subculture Seedling, Int J Bioautomation, 21(4), 349-356.
19. Zhang M. (2016). Snake Model Based on Improved Genetic Algorithm in Fingerprint Image Segmentation, Int J Bioautomation, 20(4), 431-440.

Zhenmei Zhang, Ph.D.
E-mail: zhangzhenmei@sdufe.edu.cn

Zhenmei Zhang has received Ph.D. degree from the College of Computer Science and Engineering, Shandong University of Science and Technology, China. She is currently working in the Shandong University of Finance and Economics. Her research interests include data mining and machine learning.

Yongquan Liang
E-mail: lyq@sdust.edu.cn

Yongquan Liang is currently working in the College of Computer Science and Engineering, Shandong University of Science and Technology. His current research interests include data mining and intelligent information processing. Up to now, he has published over 100 papers in national and international journals or conferences.