Asian Journal of Engineering and Technology (ISSN: 2321 – 2462)
Volume 02 – Issue 05, October 2014

A Teaching-Learning-Based Optimization for Operon Prediction

Mei-Lee Hwang
Dept. of Chemical Eng., I-Shou University, Kaohsiung, Taiwan

Yi-Cheng Chiang
Dept. of Electronic Eng., National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

Yu-Da Lin
Dept. of Electronic Eng., National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

Li-Yeh Chuang
Dept. of Chemical Eng., I-Shou University, Kaohsiung, Taiwan

Cheng-Hong Yang
Dept. of Electronic Eng., National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
Email: chyang {at} cc.kuas.edu.tw

ABSTRACT--- ***Operons are the basic unit of transcription and can be used to understand the transcription regulation in a given prokaryotic genome. Currently, the sequence and gene coordinates of organisms can be rapidly identified, but their operons remain unknown. Moreover, the experimental methods detecting operons are extremely difficult and time-consuming to execute. Operon prediction as pretreatment can greatly reduce the cost of performing an experimental assay. Many algorithms and biological properties have been proposed but the resulting predictions still require improvement in terms of sensitivity, specificity, and accuracy. This study uses a teaching-learning-based optimization (TLBO) algorithm with three biological properties for operon prediction: the intergenic distance, the metabolic pathway, and the cluster of orthologous groups (COG). These properties for the Escherichia coli genome are used to train the evaluation standards of fitness function of gene pairs. The experimental results use the accuracy (ACC), sensitivity (SN) and specificity (SP) to evaluate our prediction method and as a basis for comparison with other methods to validate that the proposed method can effectively solve operon prediction problems.***

Keyword--- operon prediction, teaching-learning-based optimization, intergenic distance, metabolic pathway, cluster of orthologous groups

---

1. INTRODUCTION

Operons contain one or more structural genes and are transcribed into multi-gene mRNAs. In the case of a multi-gene transcript, the set of genes found in the transcript is arranged in tandem in the chromosome. Thus, operons can be used to understand gene transcription rules. Operons contain valuable information for drug design and protein functions. However, experimental methods for detecting operons are extremely difficult and time-consuming [1], raising the importance of developing an effective prediction method. Currently, many biological properties are already used to infer prokaryotic operons, including the distance between genes, conserved gene clusters, functional relationships based on genomic sequences, and experimental evidence [2]. Intergenic distance is a simplest prediction property because the distance between operon pairs is significantly less than the distance between non-operon pairs. Thus, intergenic distance on its own can yield good operon prediction results [2]. Since genes in the same operon often show similar functional relations, the intergenic distance can provide good prediction results. Metabolic pathways [3], clusters of orthologous groups [4], and gene ontologies [5] are also often used to predict operons.

In recent years, many algorithms have been used to predict operons, such as hidden Markov models [6], support vector machines [7], probabilistic learning [8], Bayesian networks [9], fuzzy guided genetic algorithms [1], and genetic algorithms [10]. FGA uses the intergenic distance, metabolic pathway, conservation across multiple genomes, and the similarity of protein functions to design a fitness function assessment method. GA uses the intergenic distance, metabolic pathway, cluster of orthologous groups gene function (COG), and microarray expression data to assess the putative operon, and the algorithm

sums all scores as a basis for assessing the merits of chromosomes. However, these methods do not consider the importance of direction in operon prediction. Therefore, algorithms cannot determine the better parent chromosome at initialization, thus limiting solution quality.

Since the co-transcribed genes have the same biological properties, machine learning can be applied to these operon properties for operon prediction. The predicted results of an assay can be used as reference data. Thus, costs can be greatly reduced and the effectiveness of experimental detection can be improved. We propose an effective teaching-learning-based optimization algorithm to predict operons, and use the direction and distance between adjacent genes to encode chromosomes during the initialization process, thereby obtaining powerful initial population. To calculate fitness, we use three biological properties of the *Escherichia coli* genome to train the evaluation standards of fitness function of gene pairs. Finally, we tested our method on the four genomes. Experimental results indicate that our method obtained higher levels of accuracy, sensitivity, and specificity than other methods from the literature.

2. METHODOLOGY

2.1 Teaching–Learning-Based Optimization (TLBO)

Rao et al. (2011) proposed teaching–learning-based optimization (TLBO) [11] as a clustering intelligent optimization algorithm and analog for human teaching and learning of philosophy. It is based on the effect of the influence of a teacher on learner output. Teacher quality affects learning outcomes; hence, a highly learned person is generally considered to be a teacher who then shares knowledge with learners to improve the population quality. A good teacher trains learners such that they can have better results in terms of their marks or grades. Moreover, learners also learn from interaction between themselves. TLBO is explained in detail as follows:

1) Initialization: In TLBO, each learner is a feasible solution. To obtain an outstanding initial population, learners are produced based on the distance and direction of the adjacent genes. Each learner generates a random number from 0 to 600 as a threshold, and three situations are considered into coding learner. (1) When the adjacent gene has the same direction and a distance below the threshold value, the gene will be encoded as *rand*(5-10). (2) If adjacent gene has the same direction but the adjacent gene distance is greater than the threshold value, the upstream gene will be encoded as *rand*(0-5). (3) The adjacent gene has a different direction or comes last in the gene sequence, thus the upstream gene will be encoded as 0. To evaluate prediction accuracy, the learner's decimal encoding must be converted to binary.

2) Fitness evaluation: Adjacent genes within the same operon are usually characterized by short distances, and may sometimes even overlap. Hence, a short intergenic distance indicates that genes are more likely to be located in the same operon [10] [12]. Genes within an operon often participate in the same biological process [7], and co-transcribed genes often share the same properties and functional relations. Therefore, we use intergenic distance, metabolic pathway, and COG gene properties to calculate the fitness value, and the training scores are obtained by the overall pair-score of the adjacent genes of the *E. coli* genome. Intergenic distance, metabolic pathway and COG gene properties are described below:

a) Intergenic distance: This feature is used as an evaluation criterion, along with intervals of intergenic distance using the logarithmic likelihood (LL) method for E. coli genome, and the score of each interval is assessed at 10bps [13].

$$LL_{Property}(gene_i, gene_j) = \ln\left(\frac{N_{WO}(property)/TN_{WO}}{N_{TUB}(property)/TN_{TUB}}\right) \quad (1)$$

where $N_{WO}(property)$ and $N_{TUB}(property)$ respectively correspond to genes with the same characteristics on the number of *WO* and *TUB* pairs. TN_{WO} and TN_{TUB} are the total pair numbers of *WO* and *TUB*, respectively.

b) Metabolic Pathway: This property is used to predict whether or not a gene pair is located in the same operon. Eq.1 is used to calculate the gene pair score of the metabolic pathways based on the *E. coli* genome.

c) Cluster of Orthologous Groups (COG): COG contains three levels of biological functions; each level can be subdivided into several functional categories. The first level is divided into four main categories, including (1) information storage and processing, (2) cellular processing and signaling, (3) metabolism, and (4) different COG categories. Eq.1 can calculate the scores of categories (1), (2), and (3) of the first level. Gene pairs have a score for one of these three categories when the gene pair shares the same categories. If the gene pair belongs to different COG categories, the score of this category is calculated with Eq.2.

$$LL_{Property}(gene_i, gene_j) = \ln\left(\frac{1-N_{WO}(COG)/TN_{WO}}{1-N_{TUB}(COG)/TN_{TUB}}\right) \quad (2)$$

Eq. 3 is used to calculate the fitness value of the c^{th} putative operon.

$$fitness\ (operon_{th}) = \sum_{i=1}^{m-1}(d_i) - d_m + \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(S_{path}(gene_i, gene_j) + S_{COG}(gene_i, gene_j))}{n} \times m \quad (3)$$

where *m* and *n* are respectively the total number of genes and gene pairs in the $operon_{th}$, respectively. Finally, the fitness value of a learner is calculated as the sum of the fitness values from all putative operons in the learner as follows:

$$fitness_{c^{th}} = \sum_{i=1}^{c} fitness(operon_i) \quad (4)$$

3) Teacher Phase: In a population, learner improvement depends largely on the presence of a highly learned person (*teacher*) and the mean level of population (*M*), which are defined as follows:

$$M_g = \frac{1}{n}\sum_{i=1}^{n} X_{g,i} \quad (5)$$

$$teacher_g = X_g\_best \quad (6)$$

where *g* represents the number of generations; *n* is the number of learners; and *i* is the target learner. In addition, the amount of learner learning ($Difference\_Mean_i$) is updated according to the difference between the existing and the new mean given by following:

$$Difference\_Mean_i = r_i(teacher_g - T_f M_i) \quad (7)$$

where T_f is a teaching factor that decides the value of the mean to be changed, and r_i is a random number in the range [0, 1]. The value of T_f can be either 1 or 2 which is again a heuristic step and decided randomly with equal probability as in Eq.8.

$$T_f = round[1 + rand(0,1)] \quad (8)$$

The above difference modifies the existing learner according to the following expression.

$$X_{new,i} = X_i + Difference\_Mean_i \quad (9)$$

4) Learner Phase: Learners increase their knowledge by two different means. In this step a learner randomly learns with other learners (randomly selecting another learner X_j, such that $i \neq j$). A learner learns something new if the new learner has relatively more knowledge. The learner modification step is expressed as:

$$X_{new,i} = \begin{cases} X_{old,i} + r_i(X_i - X_j), f(X_i) > f(X_j) \\ X_{old,i} + r_i(X_j - X_i), f(X_i) \leq f(X_j) \end{cases} \quad (10)$$

5) Parameter settings: In this study, the parameter value for the population (learner) number *P* is 20, the iteration number *G* is 100, and the initializatial threshold is between 0 and 600 bps.

3. EXPERIMENTAL RESULT AND DISCUSSIONS

3.1 Data sets

The study considers experimental data sets from the *E. coli*, *B. subtilis*, *P. aeruginosa PA01*, *S. aureus*, and *M. tuberculosis* genomes, respectively containing 4430, 4160, 5566, 2656, and 3988 genes. All experimental data and annotated genes can be downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov/). We obtained the experimental operon data of the *E. coli* and *B. subtilis* genome from the OperonDB [14] and DBTBS (http://dbtbs.hgc.jp/) [15] databases, respectively. The operon data of the *P. aeruginosa PA01*, *S. aureus*, and *M. tuberculosis* genome are obtained from the ODB (http://odb.kuicr.kyoto-u.ac.jp/) [16]. The genome's metabolic pathway and COG were respectively obtained from KEGG (http://www.genome.ad.jp/kegg/pathway.html) and NCBI (http://www.ncbi. nlm.nih.gov/COG/).

3.2 Prediction results and Discussion

Table I is used to calculate sensitivity, specificity, and accuracy [17]. In this table, TP and FP represent true and false positives, and TN and FN represent true and false negatives. In this study, we use three biological properties to design the fitness function. The results show the function can obtain a good balance between sensitivity and specificity. The TLBO algorithm is proposed to predict operons, and its search function can identify the highest probability of operon combinations in a gene sequence. Best the solution compared with experimentally-verified operons is used to calculate TP, FN, TN, and FP to evaluate accuracy, sensitivity, and specificity. Experimental results are shown in Table II. For some data sets, BPSO obtains a higher specificity than TLBO, but TLBO not only obtains a better balance between sensitivity and specificity, but also obtains better accuracy. In addition, TLBO provides better accuracy, sensitivity, and specificity than the other methods and obtains superior overall results.

Given the importance of the initialization step for operon prediction, we use the direction and distance of biological properties to obtain a powerful initial population. Thus, the updated population can effectively improve the accuracy of operon prediction through multiple iterations. In operon prediction, a short intergenic distance indicates that genes are more likely to be located in the same operon, and the direction of the adjacent gene is important for operon prediction because genes in the same operon can share a direction, while adjacent genes in different directions must belong to different operons. Therefore, this study uses two biological characteristics as the initial basis: threshold of intergenic distance (adjusting the initial threshold to 600 bps raises the sensitivity and specificity of the gap, and improves prediction accuracy) and the direction of the adjacent gene (which effectively enhance prediction accuracy and specificity).

Generally, the use of more functions increases prediction accuracy along with computation time. However, not all features are applicable to all genomes, and they must be chosen carefully. On the other hand, adjacent genes have related properties, but still have a high probability of falling into different operons, and forecasting results are directly influenced by the choice of biological properties and the design of the fitness function. Hence, in the proposed method, we select biological characteristics for high usage rates as the basis for fitness, and the log-likelihood [18] statistical method is used to establish a reliable fitness function. In calculating fitness terms, although the proposed method only uses three features for prediction (fewer than are used in other operon prediction methods), the prediction results achieve a better balance between sensitivity and specificity. The TLBO can thus greatly reduce the cost of performing an experimental assay.

4. REFERENCES

[1] E. Jacob, R. Sasikumar, and K. N. Nair, "A fuzzy guided genetic algorithm for operon prediction," *Bioinformatics*, vol. 21, pp. 1403-1407, Apr 15 2005.

[2] R. W. Brouwer, O. P. Kuipers, and S. A. van Hijum, "The relative value of operon predictions," *Brief Bioinform*, vol. 9, pp. 367-375, Sep 2008.

[3] Y. Zheng, J. D. Szustakowski, L. Fortnow, R. J. Roberts, and S. Kasif, "Computational identification of operons in microbial genomes," *Genome Res*, vol. 12, pp. 1221-1230, Aug 2002.

[4] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," *Science*, vol. 278, pp. 631-637, Oct 24 1997.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-29, May 2000.

[6] T. Yada, M. Nakao, Y. Totoki, and K. Nakai, "Modeling and predicting transcriptional units of Escherichia coligenes using hidden Markov models," *Bioinformatics*, vol. 15, pp. 987-993, 1999.

[7] G. Q. Zhang, Z. W. Cao, Q. M. Luo, Y. D. Cai, and Y. X. Li, "Operon prediction based on SVM," Comput Biol Chem, vol. 30, pp. 233-240, Jun 2006.

[8] M. Craven, D. Page, J. Shavlik, J. Bockhorst, and J. Glasner, "A probabilistic learning approach to whole-genome operon prediction," Proc Int Conf Intell Syst *Mol Biol*, vol. 8, pp. 116-127, 2000.

[9] J. Bockhorst, M. Craven, D. Page, J. Shavlik, and J. Glasner, "A Bayesian network approach to operon prediction," *Bioinformatics*, vol. 19, pp. 1227-1235, Jul 1 2003.

[10] S. Wang, Y. Wang, W. Du, F. Sun, X. Wang, C. Zhou, and Y. Liang, "A multi-approaches-guided genetic algorithm with application to operon prediction," *Artif Intell Med*, vol. 41, pp. 151-159, Oct 2007.

[11] R. Rao, V. Savsani, and D. Vakharia, "Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems," *Computer-Aided Design*, vol. 43, pp. 303-315, 2011.

[12] Y. Yan and J. Moult, "Detection of operons," *Proteins*, vol. 64, pp. 615-628, Aug 15 2006.

[13] P. Romero and P. D. Karp, "Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases," *Bioinformatics*, vol. 20, pp. 709-717, 2004.

[14] M. Pertea, K. Ayanbule, M. Smedinghoff, and S. L. Salzberg, "OperonDB: a comprehensive database of predicted operons in microbial genomes," *Nucleic Acids Res*, vol. 37, pp. D479-D482, Jan 2009.

[15] N. Sierro, Y. Makita, M. de Hoon, and K. Nakai, "DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information," *Nucleic Acids Res*, vol. 36, pp. D93-D96, Jan 2008.

[16] S. Okuda, T. Katayama, S. Kawashima, S. Goto, and M. Kanehisa, "ODB: a database of operons accumulating known operons across multiple genomes," *Nucleic Acids Res*, vol. 34, pp. D358-D362, Jan 1 2006.

[17] P. Dam, V. Olman, K. Harris, Z. Su, and Y. Xu, "Operon prediction using both genome-specific and general genomic information," *Nucleic Acids Res*, vol. 35, pp. 288-298, 2007.

[18] H. Salgado, G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides, "Operons in Escherichia coli: genomic analyses and predictions," *Proc Natl Acad Sci U S A*, vol. 97, pp. 6652-6657, Jun 6 2000.

[19] L.-Y. Chuang, J.-H. Tsai, and C.-H. Yang, "Binary particle swarm optimization for operon prediction," *Nucleic Acids Res*, vol. 38, pp. e128-e128, 2010.
[20] G. Li, D. Che, and Y. Xu, "A universal operon predictor for prokaryotic genomes," *J Bioinform Comput Biol*, vol. 7, pp. 19-38, Feb 2009.
[21] P. Dam, V. Olman, K. Harris, Z. Su, and Y. Xu, "Operon prediction using both genome-specific and general genomic information," *Nucleic Acids Res*, vol. 35, pp. 288-298, 2007.
[22] X. Chen, Z. Su, P. Dam, B. Palenik, Y. Xu, and T. Jiang, "Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome," *Nucleic Acids Res*, vol. 32, pp. 2147-2157, 2004.
[23] L. Wang, J. D. Trawick, R. Yamamoto, and C. Zamudio, "Genome-wide operon prediction in Staphylococcus aureus," *Nucleic Acids Res*, vol. 32, pp. 3689-3702, 2004.
[24] P. Roback, J. Beard, D. Baumann, C. Gille, K. Henry, S. Krohn, H. Wiste, M. I. Voskuil, C. Rainville, and R. Rutherford, "A predicted operon map for Mycobacterium tuberculosis," *Nucleic Acids Res*, vol. 35, pp. 5085-5095, 2007.

Table I. Evaluation method for operon prediction

Value to be estimated	**Equation for estimation**
Sensitivity	TP/(TP+FN)
Specificity	TN/(FP+TN)
Accuracy	(TP+TN)/(TP+FP+TN+FN)

Table II. Accuracy, sensitivity, specificity of four genomes

Genome	**Methodology**	**Accuracy**	**Sensitivity**	**Specificity**
B. subtilis (NC_000964)	TLBO	0.916	0.893	0.953
	BPSO [19]	**0.921**	0.887	0.945
	UNIPOP [20]	0.792	0.782	0.821
	GA [10]	0.883	0.873	0.897
	Using both genome-specific and general genomic information [21]	0.902	N/A	N/A
	SVM [7]	0.889	**0.900**	0.860
	ODB [16]	0.632	0.499	**0.992**
	FGA [1]	0.882	N/A	N/A
	JPOP [22]	0.746	0.720	0.900
P. aeruginosa PA01 (NC_002516)	TLBO	**0.948**	**0.952**	**0.941**
	BPSO [19]	0.933	0.930	0.939
	GA [10]	0.813	0.870	0.763
S. aureus (NC_002952)	TLBO	**0.973**	**0.993**	0.931
	BPSO [19]	0.959	0.959	**0.959**
	Genome-wide operon prediction in *Staphylococcus aureus* [23]	0.920	N/A	N/A
M. tuberculosis (NC_000962)	TLBO	**0.963**	**0.963**	**0.963**
	BPSO [19]	0.951	0.944	**0.963**
	A Predicted Operon map for Mycobacterium tuberculosis [24]	0.908	N/A	N/A