Differential Evolution with Application to Operon Prediction

Li-Yeh Chuang, Yi-Cheng Chiang, and Cheng-Hong Yang, *Member, IAENG*

***Abstract*—An operon is the basic unit of transcription. The structural gene in the operon is co-transcribed into a single-stranded mRNA sequence, allowing operons to contribute to the understanding of transcription rules. However, experimental methods for detecting operons are extremely difficult and time-consuming to execute, thus using operon prediction as pre-treatment can greatly reduce the cost of performing an experimental assay. Previous studies have used different algorithms) with biological properties to predict genome operons distributions. This study uses a differential evolution (DE) algorithm with biological properties to predict the operons of bacterial genomes. The biological properties include the intergenic distance, the metabolic pathway, the cluster of orthologous groups (COG), gene length ratio and operon length. The *Escherichia coli* genome is used to train the evaluation standards of each property. The present study proposes DE for operon prediction, and also compares the effectiveness of the five properties as presented by ROC curves. Results indicate that intergenic distance, metabolic pathway and COG provide better operon prediction results. The respective accuracy values for the *B. subtilis, P. aeruginosa PA01, S. aureus* and *M. tuberculosis* genomes were 0.923, 0.954, 0.963 and 0.963. A comparison with other methods in the other literature demonstrates that the proposed method can effectively be used for operon prediction.**

***Index Terms*—operon prediction, differential evolution, intergenic distance, metabolic pathway, cluster of orthologous groups.**

I. Introduction

In prokaryotic organisms, operons of bacterial genomes contain valuable information regarding protein functions that can be used in drug design. An operon contains a promoter, an operator, one or more continuously-structural genes, and a terminator. The structural gene is co-transcribed into a single strand of mRNA. This provides information that is translated into proteins. However, experimental methods for detecting operons are extremely difficult and time-consuming [1] and effective prediction methods are urgently needed. This research focuses on using machine learning and biological properties for operon prediction. Since the co-transcribed genes have the same biological properties, machine learning can be applied to these biological properties for operon prediction. The prediction results of an assay can be used as reference data, thus greatly reducing costs and improving the effectiveness of experimental detection.

In recent years, studies have proposed several properties for use in inferring prokaryote operon structures, namely intergenic distance, conserved gene clusters, functional relations, genome sequence-based, and experimental evidence [2]. Genome sequence-based promoters and terminators are most commonly used for operon prediction for these five properties [3], with intergenic distance being the simplest to predict. It is widely used in operon prediction because the distance between operon pairs (i.e., adjacent genes within a single operon) is significantly smaller than the distance between non-operon pairs (i.e., adjacent genes within different operons), thus intergenic distance on its own can yield good operon prediction results [2]. Since genes in the same operon often show similar functional relations, this property also provides good prediction results. Metabolic pathways [4], clusters of orthologous groups [5], and gene ontologies [3] are also often used for operon prediction.

Operon prediction methods proposed in recent years include hidden Markov models [6], support vector machines [7], probabilistic learning [8], Bayesian networks [9], fuzzy guided genetic algorithms [1], genetic algorithms [10] and differential evolution [11]. This study uses the differential evolution of an optimization algorithm to predict operons. The *Escherichia coli* (NC_000913) genome was used to train the fitness value of a gene pair, and accuracy testing was conducted using four test data sets. The fitness function evaluation standard was based on the intergenic distance, the metabolic pathway, the cluster of orthologous groups (COG), gene length ratio and operon length of the *E. coli* genome. The log-likelihood [12] was used to assess the scores of biological properties.

We propose a simple and highly accurate computational method for operon prediction. The direction and distance between adjacent genes was used to encode chromosomes during the initialization process, and subsequent iterations were conducted with consideration of the relationship of adjacent and nearby genes to produce an operon combination. The proposed method was tested on the *B. subtilis* (NC_000964), *P. aeruginosa PA01* (NC_002516), *S. aureus* (NC_002952) and *M. tuberculosis* (NC_000962) genomes. Experimental results on the four test data sets indicate that the proposed method obtained higher levels of accuracy, sensitivity, and specificity than can be obtained from other methods from the literature.

L. Y. Chuang is with the Chemical Engineering Department, I-Shou University, 84001, Kaohsiung, Taiwan. (e-mail: chuang@isu.edu.tw).

Y. C. Chiang is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 80778, Kaohsiung, Taiwan. (e-mail: a09210917@yahoo.com.tw).

C. H. Yang is also with the Electronic Engineering Department, National Kaohsiung University of Applied Sciences, 80778, Kaohsiung, Taiwan. (corresponding author to provide phone: 886-7-3814526#5639; e-mail: chyang@cc.kuas.edu.tw).

II. Methodology

A. Training score based on biological properties

In this study, the *E. coli* genome is used to train various property scores, followed by accuracy tests on the testing data genomes. Predictors are easier to build for large data sets like the *E. coli* genome. We applied five biological properties for operon prediction: the intergenic distance, the metabolic pathway, the cluster of orthologous groups (COG), gene length ratio and operon length. These five properties for the *E. coli* genome were used to assess the possibility of an assumed operon, with assessment scores calculated by the log-likelihood method. The properties and score assessment method are introduced below.

1) Intergenic Distance

Adjacent genes within the same operon are usually characterized by short distances, and adjacent genes may sometimes even overlap. Hence a short intergenic distance indicates that genes are more likely to be located in the same operon. [10]. Yan and Moult [13] further proposed that the distance distribution frequency of non-operon pairs increases with distance, and gradually becomes exceeds the frequency of operon pairs. We chose this feature as an evaluation criterion. The log-likelihood method for the scores is given in Eq.1:

$$\mathrm{LL}_{\mathrm{Property}}\left(gene_i, gene_j\right)=\ln\left(\frac{N_{WO}\left(property\right)/TN_{WO}}{N_{TUB}\left(property\right)/TN_{TUB}}\right) \qquad (1)$$

where property can be distance, pathway or COG. In intergenic distance, $N_{WO}(property)$ 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. Table I shows the score of each interval of the *E. coli* genome based on 10bps [14]. The table shows that, if the distance between a gene pair is -4 bps, the score of the gene pair is 2.22656. It also shows that shorter distances between gene pairs often obtain higher scores.

2) Metabolic Pathway

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, this property can also be used to predict whether a gene pair is located in the same operon. Using Eq.1 to calculate the gene pair score of metabolic pathways based on the *E. coli* genome shows that, if the adjacent gene has the same metabolic pathway, the gene pair has a score of 2.671; otherwise the score is 0.

3) Cluster of Orthologous Groups

The cluster of orthologous Groups (COG) contains three levels biological functions; each level can be subdivided into several functional categories. The first level is divided into four main categories, namely (1) information storage and processing, (2) cellular processing and signaling, (3) metabolism and (4) different COG categories. We use Eq.1 to 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 using Eq.2. Table 2 shows the training scores of this property.

$$\mathrm{LL}_{\mathrm{COG}}\left(gene_i, gene_j\right)=\ln\left(\frac{1-N_{WO}\left(COG\right)/TN_{WO}}{1-N_{TUB}\left(COG\right)/TN_{TUB}}\right) \qquad (2)$$

4) Gene length ratio

WO pairs are often associated with big values of the natural logarithm of the length ratio when the $\log_n$ of the length ratio is examined. The length ratio property is best able to predict whether the gene pair is located in an operon [15]. The pair-score of the gene length ratio is calculated as the natural logarithm of the length ratio of the upstream and downstream genes [15]. It is defined by the following equation:

$$\mathrm{LL}_{\mathrm{glr}}\left(gene_i, gene_j\right)=\ln\left(\frac{length_i}{length_j}\right) \qquad (3)$$

where $length_i$ and $length_j$ are the length of the upstream and downstream gene, respectively.

TABLE I
Intervals of intergenic distance using the logarithmic likelihood method for E. coli genome

Interval	Score	Interval	Score	Interval	Score
[-∞, -99]	-0.82457	[30, 39]	0.568643	[170, 179]	-1.83357
[-100, -91]	0.00000	[40, 49]	-0.67375	[180, 189]	-1.98772
[-90, -81]	1.478014	[50, 59]	-0.52852	[190, 199]	-1.51772
[-80, -71]	0.00000	[60, 69]	-0.43437	[200, 209]	-2.35497
[-70, -61]	-0.31375	[70, 79]	-0.6435	[210, 219]	-1.98772
[-60, -51]	0.00000	[80, 89]	-0.6322	[220, 229]	-3.4918
[-50, -41]	0.533552	[90, 99]	-0.55887	[230, 239]	-2.23556
[-40, -31]	-0.22673	[100, 109]	-1.48787	[240, 249]	-2.25966
[-30, -21]	0.379401	[110, 119]	-1.15683	[250, 259]	-2.79865
[-20, -11]	2.019145	[120, 129]	-1.43768	[260, 269]	0.00000
[-10, -1]	2.22656	[130, 139]	-1.84221	[270, 279]	-3.33417
[0, 9]	2.2105	[140, 149]	-2.66512	[280, 289]	-2.1329
[10, 19]	2.340637	[150, 159]	-1.80384	[290, 299]	-2.83947
[20, 29]	1.564274	[160, 169]	-1.78965	[300, ∞]	-2.96611

TABLE II
FREQUENCIES OF ADJACENT PAIRS FOR DIFFERENT COG FUNCTIONAL CATEGORIES AND THEIR SCORES IN THE E. COLI GENOME

COG main categories of the first level	OP pairs frequency	NOP pairs frequency	Score
Information storage and processing	0.046	0.018	0.9360
Cellular processing and signaling	0.105	0.023	1.4996
Metabolism	0.271	0.085	1.1543
Different COG categories	0.579	0.873	-0.4112

5) Operon length

The operon length is given by the number of genes in an operon [9]. De Hoon et al. [16] calculated a prior probability of adjacent gene pairs within the same operon based on a list of 635 experimentally verified operons. In the study, the assessment of the prior probability is based on the experimentally verified operons of the *E. coli* genome. If an operon consists of multiple genes, the probability of operon appearance decreases [17]. The probability *P*, i.e. the pair-score of the operon length, is calculated by the following equation:

$$P_i = \frac{\bar{n} - 1}{\bar{n}} \tag{4}$$

where $\bar{n}$ is the average operon length given by the total number of genes in all operons divided by the total number of operons in the genome. P_i represents the probability of the next gene being located in the same operon. We infer that the gene pair is located in the same operon if a random number is smaller than P_i.

B. Differential Evolution

The differential evolution algorithm (DE) was proposed by Storn and Price in 1995 [18] and has been shown to have superior solving ability. The DE algorithm is similar to the genetic algorithm (GA) and particle swarm optimization (PSO) in that they are all optimized algorithms. The differential evolution algorithm includes three steps: mutation, recombination and selection. In selection, DE uses a one-to-one elimination mechanism to update the chromosome, which is similar to the recording of the best experience in PSO. DE considers the correlation between multiple variables, and this coupling provides an advantage over PSO. DE also has superior random search performance and simple parameter settings, causing it to be widely used in various fields including data mining, electronic engineering and decision support. Several DE processes are introduced below, including (1) Chromosome encoding, (2) Initialization, (3) Fitness evaluation, (4) Mutation, (5) Recombination and (6) Selection.

1) Chromosome encoding

To evaluate prediction accuracy, we must first define the adjacent gene pair for operon prediction. Adjacent genes in the same operon are called operon pairs (OP) and are positive. If an operon contains only a single gene or if it contains an adjacent gene within a different operon it is called a non-operon pair (NOP) and the gene pair is negative. If we assume an adjacent gene within the same operon, then the upstream gene of the adjacent gene will be coded 1. On the other hand, if the gene is coded 0, the gene and downstream gene are assumed to be NOP. For example, coding the chromosome $x_i = (1, 1, 0, 0, 1, 0)$ indicates the assumption that Gene$_1$, Gene$_2$ and Gene$_3$ are located in the same operon; Gene$_4$ is a single-gene operon; and Gene$_5$ and Gene$_6$ belong to a single operon.

2) Initialization

The initialization process is divided into two steps. In the first step obtains the preferred initial solution while the second step facilitates the execution of the DE algorithm. As shown in Fig. 1, in the first step we use the direction and distance of adjacent genes to generate a binary coding and randomly generate a threshold for each chromosome between from 0-600bps [17]. The distance is calculated by Eq. 5 [16]. If the distance of the adjacent gene is greater than the random threshold value and the adjacent gene has the same direction, the upstream gene is encoded as 1 (e.g., Gene$_1$); Gene$_2$ is encoded as 0 because the distance between Gene$_2$ and Gene$_3$ exceeds the threshold. Gene$_n$ is encoded as 0 because it is the last gene in the genome. In the second step, we give a random value for each gene. If the gene is encoded as 1 in the first step, the random value is assigned from 5 to 10; if the gene is encoded as 0 in the first step, the random value is assigned from 0 to 4, so as to complete the encoding of the chromosome's decimal sequence.

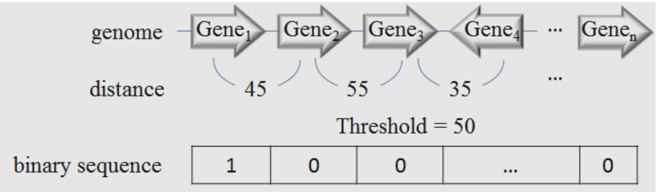

Fig 1. Diagram of binary sequence

$$\text{distance} = Gene_2\_start - (Gene_1\_finish + 1) \tag{5}$$

where $Gene_1\_finish$ is the base end position of the upstream gene, and $Gene_2\_start$ is the base start position of the downstream gene.

3) Fitness evaluation

In this study, we converted the decimal chromosome encoding of DE into binary encoding for assessment, and used the intergenic distance, the metabolic pathway, the cluster of orthologous groups (COG), gene length ratio and operon length properties to calculate the fitness value. Using the training scores of the *E. coli* genome to obtain the overall pair-score of the adjacent genes, Eq. 6 is then used to calculate the fitness value of the c^{th} putative operon.

$$\begin{aligned} fitness\,(operon_{th}) = & \sum_{i=1}^{m\text{-}1}(d_i + P_i + \mathrm{LL}_{\mathrm{glr}}(gene_i, gene_{i+1})) \\ & -(d_m + P_m + \mathrm{LL}_{\mathrm{glr}}(gene_m, gene_{m+1}) \\ & + \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(\mathrm{S}_{path}(gene_i, gene_j)}{n} \times m \\ & + \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(\mathrm{S}_{COG}(gene_i, gene_j))}{n} \times m \end{aligned} \quad (6)$$

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

$$fitness_{c^{th}} = \sum_{i=1}^{c} fitness\,(operon_i) \quad (7)$$

where *c* is the number of operons in the particle.

4) Mutation

In DE, each chromosome (Target vector, $X_{i,G}$) randomly selects three variable vectors ($X_{r1,G}$, $X_{r2,G}$ and $X_{r3,G}$) from the chromosome group, and uses Eq. 8 to combine the three variable vectors into a donor vector ($V_{i,G+1}$). In Eq.8, *F* is a scale factor which controls the length of the exploration vector ($X_{r2,G} - X_{r3,G}$).

$$V_{i,G+1} = X_{r1,G} + F(X_{r2,G} - X_{r3,G}) \quad (8)$$

where *i* is the target chromosome and *G* is the number of generations.

5) Recombination

Once the donor vector is generated by mutation, the target vector ($X_{j,i,G}$) and donor vector ($V_{j,i,G}$) are exchanged by the crossover rate (*CR*), and thus generate $u_{i,G+1}$ (trial vector or final offspring) by Eq.9.

$$u_{j,i,G+1} = \begin{cases} V_{j,i,G} & ,if\ rand \le CR \\ X_{j,i,G} & ,if\ rand > CR \end{cases} \quad (9)$$

where *rand* is a random number between 0 and 1; *j* is the dimension of the chromosome *i* under examination.

6) Selection

The resulting $u_{i,G+1}$ is evaluated following a one-by-one spawning strategy, such as Eq. 10. $u_{i,G+1}$ replaces x_i when $f(u_{i,G+1}) \le f(X_{i,G})$; otherwise, replacement does not occur.

$$X_{i,G+1} = \begin{cases} u_{i,G} & ,if\ F(u_{i,G}) \ge F(X_{i,G}) \\ X_{i,G} & ,otherwise \end{cases} \quad (10)$$

C. Parameter settings

In this study, the parameter value for the population number *P* is 20, the iteration number *G* is 100, the scale factor (*F*) is 0.5, the crossover rate (*CR*) is 0.5, and the initialization thresholds are between 0 and 600 bps.

III. EXPERIMENTALRESULTSANDDISCUSSIONS

A. Data sets

In the study, experimental data sets consisted of the *E. coli*, *B. subtilis*, *P. aeruginosa PA01, S. aureus* and *M. tuberculosis* genomes with 4430, 4160, 5566, 2656 and 3988 genes, respectively. All experimental data and annotated genes can be downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov/). The data records the definition, name, number, start position, end position, direction, and product names of each gene. We obtained the experimental operon data of the *E. coli* and *B. subtilis* genome from the OperonDB [19] and DBTBS (http://dbtbs.hgc.jp/) [20] databases; and the operon data of the *P. aeruginosa PA01* genome, *S. aureus* and *M. tuberculosis* genome from the ODB (http://odb.kuicr.kyoto-u.ac.jp/) [21]. 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/).

B. Performance measurement

Tables III and IV show the medical diagnostic assessment methods. TP and FP represent true and false positives, and TN and FN represent true and false negatives. Table III is used to calculate sensitivity (SN), specificity (SP) and accuracy (ACC) [17]. For example, a gene sequence is encoded as 111010, our prediction result is 110110. Gene$_1$, Gene$_2$ and Gene$_5$ are TP, Gene$_3$ is FN, Gene$_4$ is FP, and Gene$_6$ is TN. Finally, sensitivity, specificity and accuracy are calculated using the equations in Table IV and are compared with results obtained by the other methods. It should be noted that the proposed method achieved a good balance between sensitivity and specificity.

TABLE III
THE POSITIVE AND NEGATIVE EVALUATION

True Prediction	Positive	Negative
Positive	TP	FP
Negative	FN	TN

TABLE IV
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)

C. Prediction results

We use the DE algorithm to identify the highest probability of operon combinations in a gene sequence, and compare the result with the experimentally verified operons to calculate TP, FN, TN, and FP and to evaluate accuracy, sensitivity, and specificity. The results, shown in Table V, are compared to those of the other methods. As explained in the discussion section, the intergenic distance, metabolic pathway and COG are used to predict operons. The proposed method obtains accuracy values of 0.923, 0.954, 0.963 and 0.963, respectively, for the *B. subtilis*, *P. aeruginosa PA01*, *S. aureus* and *M. tuberculosis* genomes. Although we only used

TABLE V
ACCURACY, SENSITIVITY, SPECIFICITY OF THREE GENOMES

Genome	Methodology	Accuracy	Sensitivity	Specificity
B. subtilis (NC_000964)	DE	**0.923**	**0.910**	0.934
	BPSO [17]	0.921	0.887	0.945
	UNIPOP [22]	0.792	0.782	0.821
	GA [10]	0.883	0.873	0.897
	Using both genome-specific and general genomic information [15]	0.902	N/A	N/A
	SVM [7]	0.889	0.900	0.860
	ODB [21]	0.632	0.499	**0.992**
	FGA [1]	0.882	N/A	N/A
	JPOP [23]	0.746	0.720	0.900
P. aeruginosa PA01 (NC_002516)	DE	**0.954**	**0.967**	0.935
	BPSO [17]	0.933	0.930	**0.939**
	GA [10]	0.813	0.870	0.763
S. aureus (NC_002952)	DE	**0.963**	**0.972**	0.945
	BPSO[17]	0.959	0.959	**0.959**
	Genome-wide operon prediction in *Staphylococcus aureus* [24]	0.920	N/A	N/A
M. tuberculosis (NC_000962)	DE	**0.963**	**0.963**	**0.963**
	BPSO [17]	0.951	0.944	**0.963**
	A Predicted Operon map for Mycobacterium tuberculosis [25]	0.908	N/A	N/A

three features for prediction (fewer than are used in other operon prediction methods), our method achieved a good balance between sensitivity and specificity. Since the resulting prediction accuracy compares well with that achieved by other methods, the proposed method can be used to solve operon prediction problems.

D.Discussion

The DE algorithm is similar to a genetic algorithm and particle swarm optimization, but it also considers the multivariate correlation, and hence has an advantage over PSO in solving problems where variables are coupled. DE uses a one-on-one elimination mechanism to update the population, making it easier for DE to find the global optima.

Since the genome contains many genes (i.e., the solution space is very large), the initialization step is very important for operon prediction. To enhance DE prediction performance, we use the direction and distance to generate the initial population with random values produced. This improves the fitness value of the chromosome population in the initialization step, and updating the population effectively improves operon prediction accuracy through multiple iterations. The direction of the adjacent gene is important for operon prediction because adjacent genes with different directions must belong to different operons, and can thus effectively predict NOP to enhance prediction accuracy and specificity. And the threshold point of intergenic distance, adjusting the initial threshold to 600 bps raises the sensitivity and specificity of the gap [17]. Therefore, we used these two conditions for initialization.

Most methods use the properties of adjacent genes to determine whether a gene pair is OP or NOP, while ignoring the importance of the relationship between a gene and its neighbors. To increase the likelihood of finding an optimal solution, the DE fitness functions must consider the properties of nearby genes. The log-likelihood method is used to design the fitness function and to assess the scores of each property. In this study, we selected the *E. coli* genome as the training data since the *E. coli* genome has been extensively studied in experiments, and the majority of its operons have been experimentally confirmed, thus increasing the credibility of *E. coli* as a training data set. Theoretically, the use of additional properties for operon prediction should yield prediction results with a higher degree confidence.

In operon prediction, biological property selection and fitness function design both directly affect the prediction results. Even though adjacent genes have related features, they could possibly belong to different operons, and hence the two factors above are the key to successful operon prediction. In theory, the more features used in prediction, it higher the resulting prediction accuracy. However, some features require considerable time investment without providing commensurate improvement. We selected the metabolic pathway and cluster of orthologous groups to predict operons because DVDA [26] only used homologous genes for prediction, yielding unsatisfactory results. ODB [21] used the intergenic distance, metabolic pathway, microarray and gene order conservation as properties, but failed to achieve a good balance between sensitivity and specificity. Therefore, we chose properties based on property utilization and prediction results.

Due to the *M. tuberculosis* genome has the scarcity of experimentally verified operon data; hence we don't provided ROC curves of *M. tuberculosis*. In Figs. 2 to 4, the ROC curves of operon prediction show two messages. The first message indicates that the gene length ratio property is not suitable for use with the distance pathway and COG for operon prediction. The second message represents that the result of operon prediction did not significantly improve when the operon length property is added to calculate the fitness value. It can thus be assumed the gene length ratio and operon length is not suitable for operon prediction with distance, pathway and COG. In Figs. 4, the scarcity of experimentally verified operon data results in the roughness of the ROC curve of *S. aureus* [22]. Since the operon length is a severely biased method of prediction, since the probability is directly dependent on the number of WO pairs and TUB pairs (25). Therefore, we choose intergenic distance, metabolic pathway and the cluster of orthologous groups as the basis of the evaluation fitness. The three features used in this study are the same in those used in the GA study. However, even though GA also used microarray expression data, the proposed method achieved a higher accuracy level, indicating that the three features used in DE are effective for operon prediction.

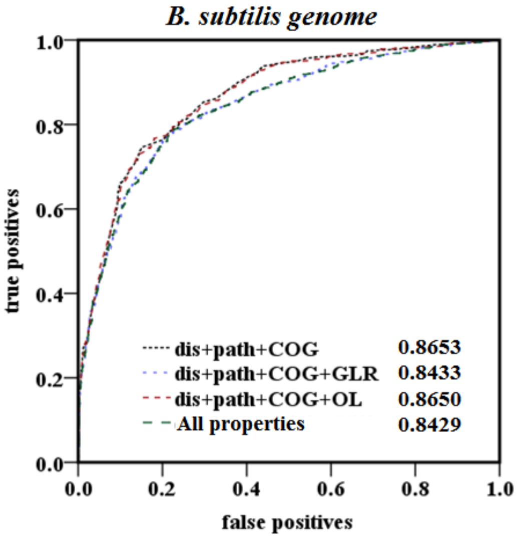

Fig 2. ROC curves of operon prediction of *B. subtilis* genome

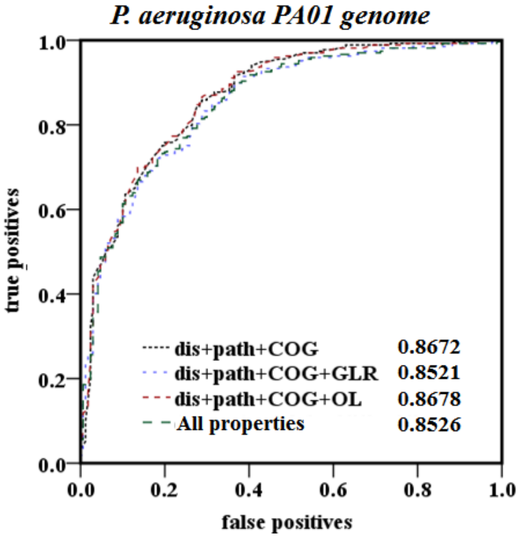

Fig 3. ROC curves of operon prediction of *P. aeruginosa* genome

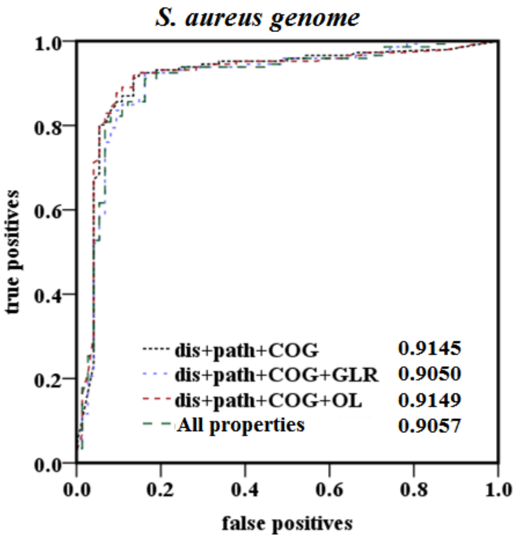

Fig 4. ROC curves of operon prediction of *S. aureus* genome

IV. Conclusions

An effective operon prediction method with improved differential evolution is proposed. The initialization step considers the direction and intergenic distance of adjacent genes, and the log-likelihood method is used to design the fitness function to further improve evaluation accuracy. Experimental results show that DE, using only three kinds of biological properties, can obtain excellent prediction results. Future research will use a greater variety of biological properties to predict operons and provide related prediction results to provide a better understanding of the impact of other features on the operon prediction problem.

References

[1] E. Jacob, R. Sasikumar, and K. N. Nair, "A fuzzy guided genetic algorithm for operon prediction," *Bioinformatics,* vol. 21, pp. 1403-7, 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. 367-75, Sep 2008.

[3] 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. 25-9, May 2000.

[4] 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. 1221-30, Aug 2002.

[5] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," *Science,* vol. 278, pp. 631-7, Oct 24 1997.

[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. 233-40, Jun 2006.

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

[9] J. Bockhorst, M. Craven, D. Page, J. Shavlik, and J. Glasner, "A Bayesian network approach to operon prediction," *Bioinformatics,* vol. 19, pp. 1227-35, 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. 151-9, Oct 2007.

[11] L.-Y. Chuang, Y.-C. Chiang, and C.-H. Yang, "A Differential Evolution for Operon Prediction," *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2013*, 13-15 March, 2013, Hong Kong, pp. 79-84.

[12] 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. 6652-7, Jun 6 2000.

[13] Y. Yan and J. Moult, "Detection of operons," *Proteins,* vol. 64, pp. 615-28, Aug 15 2006.

[14] P. Romero and P. 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.

[15] 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.

[16] M. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano, "Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information," in *Pac. Symp. Biocomput*, 2004, pp. 276-287.

[17] L. Y. Chuang, J. H. Tsai, and C. H. Yang, "Binary particle swarm optimization for operon prediction," *Nucleic Acids Res,* vol. 38, p. e128, Jul 2010.

[18] R. Storn and K. Price, "Differential evolution-a simple and efficient adaptive scheme for global optimisation over continuous spaces,"

International Computer Science Institute, Berkley, CA, Tech. Rep. TR, pp. 95-012, 1995.

[19] 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. D479-82, Jan 2009.

[20] 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. D93-6, Jan 2008.

[21] 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. D358-62, Jan 1 2006.

[22] G. Li, D. Che, and Y. Xu, "A universal operon predictor for prokaryotic genomes," *Journal of bioinformatics and computational biology,* vol. 7, pp. 19-38, 2009.

[23] 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.

[24] 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.

[25] P. Roback, J. Beard, D. Baumann, C. Gille, K. Henry, S. Krohn, H. Wiste, M. Voskuil, C. Rainville, and R. Rutherford, "A predicted operon map for Mycobacterium tuberculosis," *Nucleic Acids Res,* vol. 35, pp. 5085-5095, 2007.

[26] M. T. Edwards, S. C. G. Rison, N. G. Stoker, and L. Wernisch, "A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context," *Nucleic Acids Res,* vol. 33, pp. 3253-3262, 2005.