Improving Operon Prediction in *E. coli*

Phuongan. Dam[1,2] , Victor Olman[1] and Ying Xu[1,2]
[1] *Department of Biochemistry and Molecular Biology, University of Georgia at Athens,*
[2] *Computational Biology, Institute, Oak Ridge National Laboratory*
phd@csbl.bmb.uga.edu, olman@csbl.bmb.uga.edu and xyn@csbl.bmb.uga.edu

Abstract

In bacterium, genes working in the same pathway or interacting with each other are often organized into operons. Currently, the prediction accuracy for operon/boundary gene pairs is fairly good in *Escherichia coli*, however, such a high level of success in recognizing a gene pair as a boundary or operon pair does not automatically transcribe into a high level of accuracy in predicting the boundary of operons. We found that for several operon prediction programs, the prediction accuracy is often less accurate when the intergenic region of a gene pair is between 40 to 250 base pairs. In our approach, multiple features of the intergenic region, gene length and available microarray data in *E. coli* were used to improve the accuracy of the operon prediction programs in general and of gene pairs in the above intergenic region in particular. These features were scored according to a log likelihood formula, and the result suggests that we can gain up to 8% increase in the accuracy level for gene pairs with the intergenic distance between 40-250 base pairs. For other regions, the newly added features also give a moderate improvement in prediction accuracy. Furthermore, the accuracy in predicting transcript boundary is also improved, comparing to methods using the intergenic distance and functional annotation alone. We are currently fine-tuning our program to predict all operons in *E. coli*, and applying this method to predict operons in other organisms.

1. Introduction

Biologically, the presence of operons serves as a mechanism for transcriptional regulation of gene expression. Computationally, successful prediction of operons will improve our capability in functional annotation of conserved or hypothetical genes predicted in newly sequenced genomes. Currently, the overall prediction accuracy for operon/boundary gene pairs is fairly good, and operon prediction programs can achieve a specificity level of 85-90% in *E. coli*. Although there is a high level of success in recognizing a gene pair as a boundary or operon pair, the accuracy of predicting the boundary of experimentally-verified transcripts is not very good [4].

Previously, it has been shown that many features of the gene pairs can be used to predict whether the pair is an operon or boundary pair. The attributes include the intergenic region, the functional annotation, the phylogenetic profile, the conservation of a gene pair across multiple genomes, codon usage, the present of a Rho-independent terminator signal and gene expression profile. Furthermore, using multiple features often improves the accuracy of predicting methods.

In this paper, several features of experimentally-verified gene pairs in *E. coli* including the ratio of gene lengths, the frequency of G and TT in the intergenic regions, the phylogenetic profiles, the conservation of gene neighborhood across multiple genomes, the correlation of gene expression profiles, and the functional relationship between genes in a pair were examined. Then, the attributes were comprehensively used to predict gene pairs as well as transcription unit boundary. Unlike previously described methods, we divided the gene pairs according to the lengths of the intergenic regions. Our approach accompanying with the use of multiple features improve the accuracy of gene-pair classification by as much as 7-8%, as well as improving the accuracy of predicting transcription unit boundaries.

2. Methods and Results

Previously, it has been shown that in *E. coli*, the intergenic distance between gene-pairs can be used as a predicting parameter. Around 90% of observed gene-pairs in *E. coli* that are separated by less than 40 base pairs are operon pairs, whereas around 88% of observed pairs having the intergenic region of 250 base pairs or more are boundary pairs (data not shown). However, in the region from 40 to 250 base pairs, it is less certain to classify a pair of genes based only on the length of the intergenic region. When analyzing the accuracy of several currently available operon prediction programs in differentiating operon pairs vesus boundary pairs, we found that the prediction accuracy is often less when the intergenic region of a gene pair is between 40 to 250 base pairs (data not shown). When evaluating the predicting results from three methods including OFS,

VIMSS and JPOP[1, 3, 5], we found that all methods did very well when the intergenic length of a gene pairs is less than 40 or more than 250 base pairs. However, in the 40-250 region, the prediction accuracy drops below the 80% (Table 1).

Table 1. Accuracy level of several prediction methods.

Intergenic region	VIMSS	OFS	J-POP
Under 40 bp	0.92	0.93	0.80
From 40 - 100 bp	0.67	0.72	0.61
From 110-250 bp	0.77	0.73	0.78
Over 250 bp	0.83	0.76	0.81

We characterized several characteristics of experimentally-verified gene pairs including the ratio of the lengths and the base composition of the intergenic regions. In conjunction, we also used established features that were shown to differentiate operon pairs and boundary pairs. Figure 1 shows the distribution of ratio of lengths of known gene pairs. In this step, the distribution of natural log of ratio of lengths of gene pairs were calculated, and the results suggested that the distribution of operon pairs is significantly different from the distribution of boundary pair (result not shown). Similarly, we found that the distribution of G and TT frequencies are significantly different between operon pairs and boundary pairs (data not shown). Similar to previously described methods, we found that the phylogenetic profile, conservation of gene pair across multiple genomes, gene expression profile and annotated functions of gene pairs are important features in classifying a pair of genes as operon or border pairs.

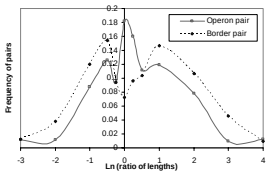

Figure 1. Distribution of ratio of lengths of known gene pairs.

For each feature, we calculate the log likelihood (LLH) score according to previously described formula [2]. Then, the total score was calculated as followed:

Total score = $\sum_{i=1}^{n} LLH_i$, n is the number of attributes.

Table 2 shows the result of our method of prediction, using multiple features of gene pairs. In this table, the accuracy of prediction was calculated as the total number of correctly-predicted operon and boundary pairs divides to the total number of known pairs. In each column, the features used include all features used in previous columns and listed features in the present column. Phylogenetic distances, LLH of Riley's functional annotation and protein complex were the same as previously described [1, 4]. We also observed that small peptides of length 33 amino acids or less are all leader peptides of operons. Therefore, we assigned a large score (10) for a gene pair if the first gene is smaller than 34 amino acids. Our result shows an overall improvement when multiple features were used, especially the accuracy in predicting gene pairs having intergenic region between 40-250 base pairs. Using all features, 69% of known, long transcripts that contains 2 or more genes were accurately predicted. When including single-gene transcription units, our accuracy rate is 79%. When omitting Riley's functional classification and protein complex information, we got 67% and 77% in predicting long transcripts and all transcripts.

Table 2. Contribution of various attributes toward the overall accuracy of gene-pair prediction.

IG length	# pairs	core1[a]	core2[b]	p-dist1[c]	p-dist2[d]	corr[e]	func[f]	k2[g]	IG dist[h]
Under 40 bp	691	0.91	0.93	0.93	0.93	0.93	0.93	0.93	0.93
From 40-100 bp	244	0.61	0.72	0.72	0.74	0.75	0.77	0.78	0.79
From 110-250 bp	326	0.76	0.83	0.84	0.85	0.85	0.85	0.85	0.85
Over 250 bp	169	0.88	0.92	0.93	0.93	0.93	0.93	0.93	0.93

[a] core1 includes 4 features: length ratio, leader peptide, G and TT composition.
[b] 3 features: conserved neighbor , entropy and Hamilton distances of the phylogenetic profiles with cut-off 10^{-15}.
[c] Hamilton distance of the phylogenetic profiles with cut-off 10^{-6}.
[d] Entropy distance of the phylogenetic profiles with cut-off 10^{-6}.
[e] Spearman correlation calculated from microarray data.
[f] Riley functional annotation and protein complex.
[g] Conserved gene-pair where 2 genes are less than 3 genes apart.
[h] Intergenic distance.

3. Discussion

When using all features described above, we achieved the same accuracy level as another method that required multiple annotated functions such as Riley's functional classification, protein complex formation, biological pathway information of the gene pairs and upstream, downstream genes [4]. Because such detailed annotation of functions for all genes in a genome is difficult to obtain, we expect that our method will perform better in genomes that are not well-studied previously, but have fair amount of microarray data available.

References

1. Chen, X., et al., *Computational Prediction of Operons in Synechococcus sp. WH8102.* Genome Inform Ser Workshop Genome Inform, 2004. **15**(2): p. 211-22.
2. Moreno-Hagelsieb, G. and J. Collado-Vides, *A powerful non-homology method for the prediction of operons in prokaryotes.* Bioinformatics, 2002. **18 Suppl 1**: p. S329-36.
3. Price, M.N., et al., *A novel method for accurate operon predictions in all sequenced prokaryotes.* Nucleic Acids Res, 2005. **33**(3): p. 880-92.
4. Romero, P.R. and P.D. Karp, *Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases.* Bioinformatics, 2004. **20**(5): p. 709-17.
5. Westover, B.P., et al., *Operon prediction without a training set.* Bioinformatics, 2005. **21**(7): p. 880-8.