Operon Prediction in Cyanobacteria using Comparative Genomics

Natalia Khuri
Carnegie Institution, DPB
260 Panama Street
Stanford, CA 94305
nkhuri@stanford.edu

Nikhila S Rao
College of Engineering,
San Jose State University
San José, CA 95192
nrao@email.sjsu.edu

Abstract

In this work, we performed a comprehensive analysis of the genome organization of eight publicly available cyanobacterial genomes and of the extensively studied Escherichia coli K12. Our comparison revealed striking similarities in organization of different microbial genomes. We examined the orientation of putative CDS, and distance distribution patterns between adjacent putative CDS in the whole genome as well as between genes within operons and across operon borders. Our results enabled us to calculate distance log-likehoods for gene pairs to be part of the same operon and to apply this metric to predict operons in two newly sequenced cyanobacterial genomes, Synechococcus OS-A and OS-B'.

1. Introduction

Operons are the basic organizational genomic units that are characteristic of a prokaryotic cell. They are a group of adjacent, co-expressed and co-regulated genes that have the same orientation and encode functionally linked proteins in prokaryote genomes. Genes in the same operon are generally part of the same metabolic pathway. Prediction of the operon structure in prokaryotic genomes is becoming increasingly important not only for identifying co-regulated genes but also for automated finding of regulatory elements.

Operon prediction methods can be classified into two groups. The first makes use of the conserved operon structures from genomes of related organisms and the second, of the fact that the genes within the operon are separated by shorter intergenic distances than those across operon borders. The rationale of the first approach is that functionally important segments of genomes co-evolve and are, therefore, conserved across related species and sometimes across species. The second method uses the intergenic distances within the operons and across the operon borders to compute the probability of two adjacent genes to be part of the same operon. The advantage of both methods is that they do not require experimental data for the newly sequenced genomes.

2. Operon Prediction Approach

We combined both comparative genomics and distance analysis to infer operon intergenic distance model. The comparative genomics method was adopted from [2] and the distance analysis method from [1]. The overview of our approach is shown in Figure 1. Our method assumes that gene pairs are more conserved on the same DNA strand when compared to gene pairs on the opposite strands. Thus, we assume that the genes in an operon will be located on the same strand (i.e. have the same orientation). We confirmed this assumption using eight publicly available cyanobacterial genomes and the genome of *Escherichia coli K12*. We refer to these genomes as reference genomes. We have also compiled an operon data set consisting of 256 polycistronic transcription units from *Escherichia coli* K-12, 13 operons from *Nostoc* sp. PCC7120, 21 from *Synechocystis* sp. PCC6803, 20 from *Synechococcus* sp. WH8102, and one from *Prochlorococcus marinus* SS120 and also from *Prochlorococcus marinus* MED4. We examined the orientation of putative CDS, and distance distribution patterns between adjacent putative CDS in each of the reference genome as well as between genes within operons and across operon borders. In the **Orientation Analysis** phase, we confirm that the characteristics of the operon data set carry over to the whole genome organization of the reference and target genomes. This is done by comparing various metrics of singletons and the directons over each data set. Directons are defined to be consecutive genes transcribed in the same direction with no intervening gene on the complementary strand. The directon size is the number of consecutive genes that makes up the

directon. Directons consisting of a single gene are termed singletons.

In the **Intergenic Distance Analysis** step, we took advantage of the observation that intergenic distances are smaller than inter-operon distances. We calculated the intergenic distances between genes in the operon data set and between all genes in the reference and target genomes. These results were used to calculate the distance log-likehoods for gene pairs to be part of the same operon and apply this metric to predict operons in the two target genomes. Log-likelihood is a quantitative estimate of the correlation of two events – a higher log-likelihood score indicates a higher probability of a gene pair being part of an operon. This score is calculated using the following formula:

$$LL[dist] = \log\left(\frac{\#\ of\ Genepairs WO[dist] \,/\, Total\ \#\ of\ Genepairs WO[dist]}{\#\ of\ Genepairs at TUB[dist] \,/\, Total\ \#\ of\ Genepairs at TUB[dist]}\right)$$

where WO = Within Operon and TUB = Transcription Unit Border.

In the final step, we used log-likelihood scores to predict operons in the target genomes.

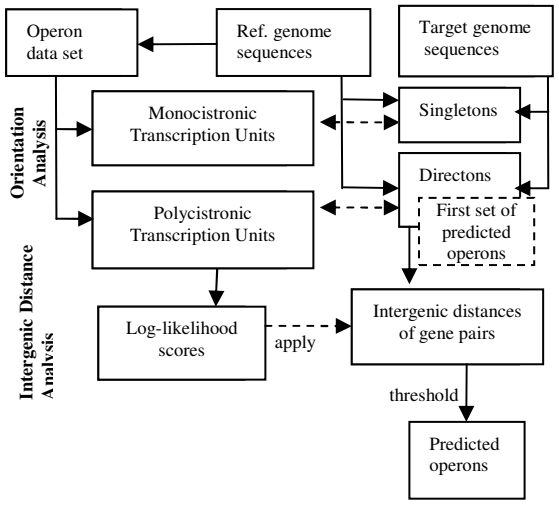

Figure 1. Overview of operon prediction method.

3. Results

Our analysis of the reference genomes revealed striking similarities between cyanobacterial genome organization and that of the *E. coli*. All genomes have approximately the same number of singletons (~10%-17% of all CDSs). Though the directon sizes in reference and target genomes range from 2 to 46, the frequency of directons with more than six genes is very small as is the frequency of operon sizes greater than six. The average operon size for the operon data set is around 3.5. The average directon size is 4.0. The two target genomes have 620 and 635 directons with 2418 and 2426 genes, respectively.

The intergenic distance (IGD) between any two adjacent genes, g_1 and g_2 is defined as: IGD = start of g_2 – (end of g_1 + 1). Based on this, the IGD can either be positive or negative (overlapping genes). The minimum intergenic distance in the operon data set (considering those genomes which have a non-negligible number of operons) is an overlap. The same is observed in directons of reference and target genomes. The most frequent intergenic distance within operons in E.coli operons is -4, which is also the most frequent distance for IGDs in all reference and target genomes. As seen in Figure 2, the log-likelihood score for smaller IGDs is positive. As the IGD increases, the log-likelihood becomes negative. This indicates that the gene pairs with smaller IGDs have a higher probability of being part of an operon.

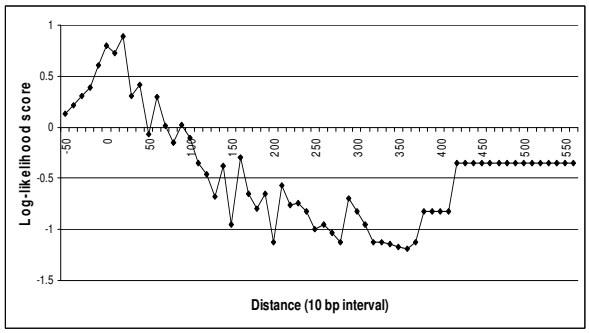

Figure 2. Log-likelihood score for 10 bp intervals

Using a threshold of zero, there are 595 predicted operons with 896 gene pairs in *Synechococcus* OS-A and 563 predicted operons with 856 gene pairs in *Synechococcus* OS-B′. The average operon size is around 2.5 in both genomes.

4. References

[1] Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y. and Jiang, T. Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. *Nucleic Acids Research, Volume 32, Number 7, pp. 2147-2157*, 2004.

[2] Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides, J. Operons in Escherichia coli: Genomic analyses and predictions. *Proc. Natl. Acad. Sci., Volume 97, No. 12, pp 6652-6657,* 2000.

5. Acknowledgements

The research was funded by the FIBR program at NSF and an NSF-ROA grant to Natalia Khuri.