

Phylogenetic Detection of Conserved Gene Clusters in Microbial Genomes

Brian P. Anton¹, Yu Zheng², Richard J. Roberts³, and Simon Kasif⁴

Keywords: conserved gene cluster, phylogenetic tree, operon prediction

1 Introduction.

Genes that reside in conserved proximity in a number of microbial genomes are often functionally related [1]. Thus, identification of such "conserved gene clusters" provides an effective channel for functional gene annotation, microarray screening, and pathway reconstruction. The problem of devising a robust method of identifying these conserved clusters and evaluating their significance has a number of implications for comparative, evolutionary, and functional genomics. Here we describe a new method for detecting conserved gene clusters by taking advantage of information in the phylogenetic tree of the organisms in which they occur. We show that our method can overcome the common problem of overestimation of significance due to bias in the genome database and thereby achieve better accuracy than more naïve approaches. This general approach may serve as a platform for many other comparative genomic analyses such as operon and regulatory site prediction.

2 Method.

Since a gene cluster can be thought of as a series of one or more overlapping gene pairs, we focus on the simpler problem of identifying and evaluating conserved gene pairs. The significance of the conservation of a gene pair is measured by a conservation score as described below. Since each gene belongs to two gene pairs, it is assigned conservation scores C_u and C_d with its upstream and downstream partner, respectively. A gene with an insignificant C_u but significant C_d value marks the start of a conserved cluster, and a gene with a significant C_u but insignificant C_d value marks the end.

We model evolution as a stochastic process acting on a rooted phylogenetic tree. The tree is constructed from only those genomes that contain the gene pair of interest, assuming that the ancestors of genomes containing the pair also contained the pair. Every branch of the tree connecting node X with its immediate ancestor Y is associated with a probability $P(X|Y)$ that X contains the gene pair given that Y also did. We show that the logarithm of the probability of observing a gene pair in a set of genomes related by phylogenetic tree T is the sum of $\log[P(X|Y)]$ for all X and Y in T .

We make the assumption that $\log[P(X|Y)]$ is proportional to the length of the branch connecting X and Y . Therefore, the sum of all branch lengths in T is proportional to the log of the probability of

¹ Bioinformatics Graduate Program, Boston University, Boston, MA, USA. E-mail: anton@bu.edu

² New England Biolabs, 32 Tozer Road, Beverly, MA, USA. E-mail: zhengy@neb.com

³ New England Biolabs, 32 Tozer Road, Beverly, MA, USA. E-mail: roberts@neb.com

⁴ Dept. of Biomedical Engineering, Boston University, Boston, MA, USA. E-mail: kasif@bu.edu

observing the pattern of gene pair conservation represented by T . This sum forms the basis of our conservation score C .

3 Results.

Genome sequence data in public repositories does not represent an even sampling across evolutionary space, but rather tends to cluster around model organisms. As a result, treatment of gene pair observations as independent events is misleading, failing to distinguish selection from vertically inherited synteny among close relatives. Ad hoc sampling methods have been devised to overcome this [3], but our method uses all available data and automatically downweights observations among close relatives, where recombination events have not had sufficient time to act.

We identified gene clusters in 127 sequenced microbial genomes, results of which are available at <http://genomics10.bu.edu/optimus>. Figure 1 shows the C_u profile for the genomic region around the *entCEBA* operon as we progressively include data from more genomes in the reference set. The scores of genes in the operon increase more rapidly than those in the surrounding region as more data is added.

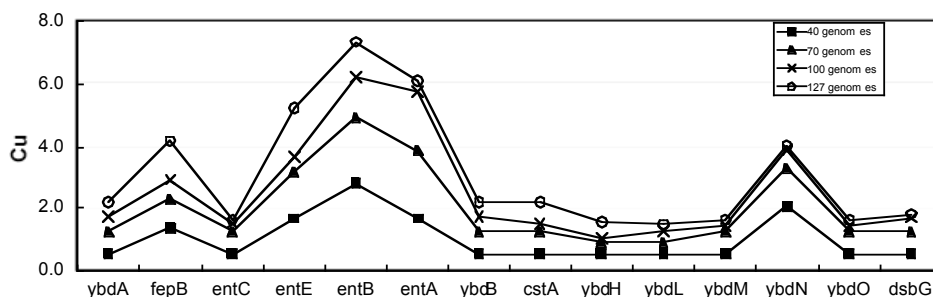


Figure 1: Effect of increasing reference genomes on C_u score profile around the *entCEBA* genomic region.

We tested the predictive power of our method for identifying known *E. coli* operons [2]. Using a cutoff C value of 5.0, our method achieves 71% sensitivity and 80% specificity against this data set. In addition, our data suggests that as many as 10-40% of genes in microbial genomes are contained in conserved gene clusters, and this is independent of genome size. Finally, our method makes possible the easy identification of gene clusters found in very few, yet highly divergent, species. These represent potential cases of horizontal gene transfer or other biologically interesting phenomena.

4 References.

- [1] Dandekar, T., Snel, B., Huynen, M., and Bork, P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23:324-8.
- [3] Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol* 16:332-46.
- [2] Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A., and Koonin, E.V. 2002. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* 30:2212-23.