

Reconstruction of Ancient Operons From Complete Microbial Genome Sequences

Yuhong Wang¹, John P. Rose², Bi-Cheng Wang², Dawei Lin²

1. Department of Molecular Biology, Jilin University, Changchun 130023, PRC

2. Southeast Collaboratory Biochemistry and Molecular Biology Department, University of Georgia, Athens, GA 30602, USA
dlin@uga.edu

Abstract

Completed genomes not only provide DNA sequence information, but also reveal the relative locations of genes. In this paper, we propose a new method for reconstruction of “ancient operons” by taking advantages of the evolutionary information in both orthologous genes and their locations in a genome. The basic assumption is that the closer two genes were in an ancient genome, the more likely they will stay close in the current genome. An assembly of non-random neighboring pairs of genes in current genomes should be able to reconstruct the gene groups that were together at a certain point of time during evolution. Given the fact that genes that are close neighbors are more likely functionally related, the gene groups generated by this assembly process are named “ancient operons”.

The assembly is only meaningful when enough non-random pairs can be found. This was made possible by over 100 microbial genomes available in recent years. For proof of concept, we chose 63 non-redundant complete microbial

genomes from RefSeq database [May 2003 release] at NCBI. In order to normalize the effect of protein sequence mutations and other changes due to evolution, we only consider assembly of COGs (Cluster of Orthologous Group) in these genomes. There are total 4901 COGs from NCBI COG database are used.

The assembly process is similar to the one that assembles DNA sequences into contigs. In our case, the neighbor COG pairs are used as basic assembly units. A target function is defined based on neighbor frequency of pair-wise link among all 4901 COGs after analysis for all 63 genomes. We used random cost algorithm, a global optimization algorithm to minimize the target function and assembled COGs into contigs. The significance of these contigs are then assessed by statistical methods. The results suggest that the assembled contigs are statistically and biologically significant. This method and the assembled ancient operons provides a new way for studying microbial genomes, their evolution and for annotating proteins of unknown functions.