Bacterial Whole Genome Phylogeny Using Proteome Comparison and Optimal Reversal Distance

Noppadon Khiripet
National Electronics and Computer Technology Center, Thailand
khirin@nectec.or.th

Abstract

Traditional phylogenetic tree reconstruction is based on point mutations of a single gene. This approach is hardly suitable for genomes whose genes are almost identical and hardly captures evolutionary scenarios. To reconstruct a more conclusive phylogenetic tree of bacterial genome, all currently available complete bacterial genomic sequences were downloaded from the National Center for Biotechnology Information (NCBI). Each individual proteome was blasted against the collection and provided a number of homologous genes shared with others. Moreover, the syntenies of each two genomes can be considered as two signed permutations. One permutation can be rearranged into another in finite steps, called reversal distance. These two measures were combined and yield a phylogenetic tree that is highly consistent with the bacterial taxonomy.

1. Introduction

The goal of a phylogenetic tree reconstruction is to analyze the evolutionary relationships among a group of organisms. Typically, the relationships have been revealed through examinations of rRNA sequences and other essential genes that can be aligned into a multiple sequence alignment. However, the selection of sequences may generate conflicting results for the evolutionary pathways of organisms.

The advances in sequencing technologies have produced a vast amount of sequence data, which give rise to the opportunity to analyze the evolution of the organisms on the genome scale. Meanwhile, the huge amount of data poses challenges for information processing, visualization and computational complexity.

Recently, there are many efforts contributed to the whole genome phylogeny. These efforts employ either complete gene sets of DNA sequences, or complete protein sequence sets [1], [2], [3] or information from gene order [4], [5]. Little efforts have been done on capturing the evolution footprints by utilizing both measures.

In this study, a new phylogenetic tree reconstruction method is proposed. A new evolutionary distance measurement based on both proteome comparison and reversal distance was developed and applied to a dataset of 225 microbes.

2. Method

The dataset was downloaded from the National Center for biotechnology Information (NCBI) and had been processed according the following steps. Firstly, the *blastp* program was used to compare protein sequences from each pair of genomes with expectation score threshold set to 1e-5. This computational process had been carried out over a few months on a cluster of computers with 64 Itanium CPUs, 1.5Ghz, 4GB RAM each. Secondly, the similarity scores were used to evaluate the distance between each two genomes by considering the gene content measure and gene order measure as described bellows. Then, an adjacency matrix of the combined distance measure was constructed and fed to the *neighbor* program of the *phylip* suite to reconstruct the phylogenetic tree.

2.1. Gene content measure

The gene content measure is described by the ratio of protein ortholog found between each pair of genomes to the number of all proteins of an organism. In other words, the fraction of shared homologous proteins in organism *A* compared to *B* is calculated as

$$C_{AB} = \frac{|G_A \cap G_B|}{\min(|G_A|, |G_B|)} \quad (1)$$

2.2. Gene order measure

When comparing two genomes and considering the orthologous genes, the order of these genes of one genome can be rearranged into the order of genes of

another genome by repeatly reversing fragments of its DNA, e.g.

```
D0: 1,-7,6,-10,9,-8,2,-11,[-3,5],4
D1: 1,-7,6,-10,9,-8,2,-11,-5,[3,4]
D2: 1,-7,6,-10,9,-8,2,[-11,-5,-4,-3]
D3: 1,-7,6,[-10,9,-8,2,3,4,5],11
D4: 1,[-7,6,-5,-4,-3,-2],8,-9,10,11
D5: 1,2,3,4,5,[-6],7,8,-9,10,11
D6: 1,2,3,4,5,6,7,8,[-9],10,11
D7: 1,2,3,4,5,6,7,8,9,10,11
```

The minimum number of reversing steps is call optimal reversal distance [6].

2.3. Combined measure

Both gene content measure and gene order measure are taken into account by forming the combined score

$$D_{AB} = C_{AB} - \frac{R_{AB}}{\min(|G_A|, |G_B|)} \qquad (2)$$

where R_{AB} is the reversal distance between genome *A* and *B*.

3. Results

Applying the Neighbor-Joining method to the 225x255 adjacency matrix of 225 baterial genomes yielded the resulting whole genome phylogenetic tree. Due to space limitation, only a subset of the tree is shown as in figure 1.

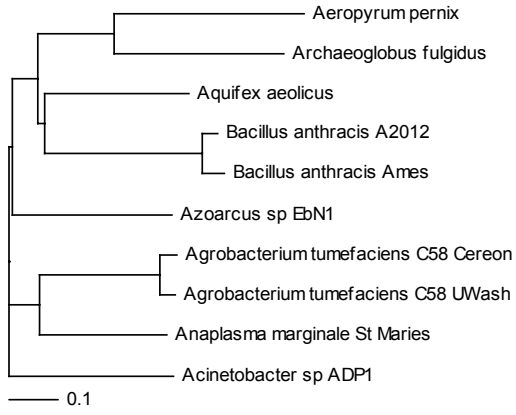

Figure 1: Whole genome phylogeny constructed by Neighbor-Joining using the combined distance matrix.

4. Discussion

The study presents a new pairwise evolutionary distance measure based on both the gene content (proteome comparisons) and the reversal distance. The resulting phylogenetic tree is expected to capture the rich evolutionary information in the whole genomic sequences. This method was applied to a dataset of 225 microbes and most of the phylogenetic results are similar to the standard taxonomy tree (*http://ncbi.nlm.nih.gov/Taxonomy*). This result demonstrated that the new combined measure is effective for whole genome phylogeny construction.

There is a major problem viewing the resulting phylogenetic tree. Common existing phylogenetic tree visualization tools are not able to display readable tree with a large number of nodes. A novel approach such as visualizing the phylogenetic tree in three dimension hyperbolic space [7] is being explored.

Acknowledgements

This work is supported by the National Electronics and Computer Technology Center, Thailand, Bioinformatics Project No C24803.

References

[1] S.R. Henz, D.H. Huson, A.F. Auch, K. Nieselt-Struwe, S.C. Schuster, "Whole-genome prokaryotic phylogeny", *Bioinformatics*, 2005 May 15, 21(10) pp 2329-2335

[2] B. Snel, P. Bork, and M. A. Huynen. "Genome phylogeny based on gene content". 1999, *National Genetics*, 21, pp108–110.

[3] C. House and S. Fitz-Gibbon. "Using homolog groups to create a whole-genomic tree of freeliving organisms: An update", *Molecular Evolution*, 2002, 54:pp539–547

[4] D. Sankoff, "Analytical approaches to genomic evolution", *Biochimie*, 75, 1993, pp409-413

[5] J.O. Korbel, B. Snel, M.A. Huynen and P. Bork, SHOT: a web server for the construction of genome phylogenies", *Trends in Genetics*, 2002 March 1, 18(3), pp158-162

[6] P. Pevzner, and G. Tesler, "Genome rearrangements in mammalian evolution: lessons from human and mouse genomes", *Genome Res*, 2003 Jan; 13(1), pp. 37-45.

[7] T. Hughes, Y. Hyun and Jones, C.D., A.B. Smith, and D.a. Libertes, "Visualising very large phylogenetic trees in three dimensional hyperbolic space" *BMC Bioinformatics 2004 Apr 29;5(1):48*