

A Rapid peptide based Alignment –free method to construct Genome trees Independent of sequence annotation

Vikash Kumar, Sunil Kumar* sunil20051@rediffmail.com
Institute of Life Sciences, Nalco Square, Bhubaneswar-751023, India

Abstract

We present here a rapid, peptide-based alignment-free method for constructing genome trees in bacteria. The protein sequence information from a given complete genome was used to generate a pool of peptides of varying length from tetramers (4 residues) to pentadecamers (15 residues). A distance measure between two genomes was developed by computing the number of common peptides between a given pair of species. Peptides of 10 amino acid residues were found to be optimal in satisfying the condition of zero background peptide matches due to chance alone and in maximizing the number of genes covered. Both informational and operational gene sets (as defined by Fitz et al. 1999) contribute to the distance measure. Confidence on the tree was assessed using the jumble option that varies the input order of species. Most of the nodes were supported at 100% in 1000 jumbles. The genome tree constructed using this approach is in agreement with those reported in the literature using traditional methods of sequence alignment procedures. This peptide based method is inherently simple, rapid and provides useful information for molecular phylogenetics studies using complete genome sequence information. We call this method as peptide based phylogenetic reconstruction (PBPR). It can be used even without prior annotation information.

Materials and Methods

Peptide distances between a given species/strain pair was computed using Jacquard's Coefficient. We define the Peptide distance (PD)

D_{ij} = Distance between any two species i & j .

$D_{ij} = 100 - \{((P_k / (P_i + P_j + P_k))) * 100\}$

P_i = Total Number of non-redundant peptides in species 'i'

P_j = Total Number of non-redundant peptides in species 'j'

P_k = Total Number of common peptides between species 'i' and species 'j'

A 55 X 55 matrix between the species was created.

Results and discussion

The topology of the tree strongly supports the monophyly of the two domains of life Archaea & Bacteria. The results are remarkably similar to results from phylogenetic analysis of the SSU r-RNA gene, increasing confidence in both methods and suggesting that a tree of life can indeed be constructed and used to understand early microbial evolution.

We found that in constructing the genome tree, peptides from operational genes are contributing more than informational genes. In traditional method all the molecules, which were used for making phylogeny, were from informational sets.

The distribution of peptides mapped back to genes in an all versus all comparison of peptide libraries of *Escherichia coli*, *Mycoplasma genitalium* and

Mycobacterium tuberculosis at different peptide lengths is given in Table 1.

Species Pair 1 st organism Vs 2 nd organism	1st Species		2nd Species	
	Informational	Operational	Informational	Operational
<i>Escherichia coli</i> Vs <i>Mycobacterium tuberculosis</i>	19%	76%	18%	63%
<i>Mycobacterium tuberculosis</i> Vs <i>Mycoplasma genitalium</i>	35%	51%	43%	54%
<i>Mycoplasma genitalium</i> Vs <i>Escherichia coli</i>	35%	62%	27%	71%

References:

1. Felsenstein, J. *PHYLIP*, version 3.5. University of Washington, Seattle, 1993.
2. Fitz-Gibbon S & House C.H. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nuc. Acid. Res.* 27 (21):4218 -4222, 1999
3. Page, M and D, Roderic A program for drawing phylogenies on MacOS and MS Windows, 2001

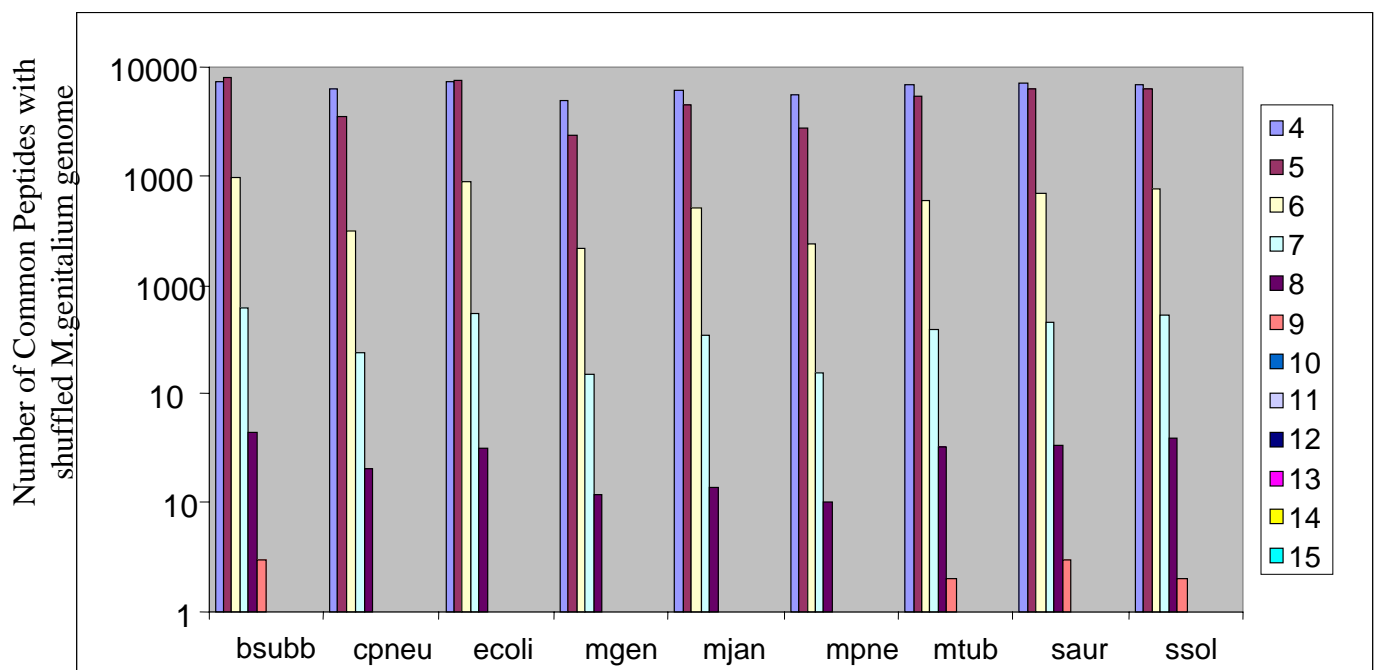


Figure 1