Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases

David T. Pride,[1,2,6] Richard J. Meinersmann,[4] Trudy M. Wassenaar,[5] and Martin J. Blaser[2,3]

[1]Department of Microbiology and Immunology, Vanderbilt University, Nashville, Tennessee 37235 USA; [2]Departments of Medicine and Microbiology, New York University School of Medicine, and [3]VA Medical Center, New York, New York 10016 USA; [4]USDA Agricultural Research Service, Athens 30604 Georgia, USA; [5]Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany

We compared nucleotide usage pattern conservation for related prokaryotes by examining the representation of DNA tetranucleotide combinations in 27 representative microbial genomes. For each of the organisms studied, tetranucleotide usage departures from expectations (TUD) were shared between related organisms using both Markov chain analysis and a zero-order Markov method. Individual strains, multiple chromosomes, plasmids, and bacteriophages share TUDs within a species. TUDs varied between coding and noncoding DNA. Grouping prokaryotes based on TUD profiles resulted in relationships with important differences from those based on 16S rRNA phylogenies, which may reflect unequal rates of evolution of nucleotide usage patterns following divergence of particular organisms from a common ancestor. By both symmetrical tree distance and likelihood analysis, phylogenetic trees based on TUD profiles demonstrate a level of congruence with 16S rRNA trees similar to that of both RpoA and RecA trees. Congruence of these trees indicates that there exists phylogenetic signal in TUD patterns, most prominent in coding region DNA. Because relationships demonstrated in TUD-based analyses utilize whole genomes, they should be considered complementary to phylogenies based on single genetic elements, such as 16S rRNA.

Biases in nucleotide composition and organization in prokaryotic genomes have long been recognized (Muto and Osawa 1987), with the representation of short oligonucleotide combinations as a focus of analysis (Henaut et al. 1996; Gelfand and Koonin 1997; Rocha et al. 1998). Dinucleotide frequencies within organisms represent genomic signatures, which may result from selective pressures as a result of dinucleotide stacking, DNA conformational tendencies, DNA replication and repair mechanisms, or selection by restriction endonucleases (Karlin et al. 1998), and codon usage also may influence nucleotide usage because it affects translational efficiency (Grantham et al. 1981; Grosjean and Freirs 1982; Sharp et al. 1993). However, constraints beyond dinucleotide frequencies and codon usage preferences can be identified only through analysis of longer oligonucleotide words (Pride and Blaser 2002). Methods available for determining the significance of oligonucleotide word frequencies include Markov chain analysis (Schbath et al. 1995; Cardon and Karlin 1994), which involves determining word frequencies by removing biases in their constituent oligonucleotides; however, the evolutionary significance of oligonucleotide word frequencies in prokaryotes has not been fully addressed.

Evolutionary inferences based on gene sequences, such as 16S rRNA (Woese and Fox 1977; Woese et al. 1990) are considered reliable indicators of prokaryotic ancestry; however, because evolutionary constraints are multidimensional (Koonin et al. 2000), analysis of a single gene is insufficient to fully understand the divergence between related life forms. The universally conserved 16S rRNA, with conservative rates of nucleotide substitution, is generally accepted as the standard for assessing microbial evolution; however, analysis of other gene loci often may not be phylogenetically congruent (Doolittle 1999). Such incongruencies often result from horizontal gene transfer, which obscures evidence of recent common ancestry (Holmes et al. 1999). With an increasing number of complete genomic sequences available, it now can be determined whether the relationships revealed from phylogenies based on 16S rRNA are reflected in the nucleotide usage patterns of individual organisms. Analysis of complete genomes can identify the extent to which nucleotide usage has evolved after divergence from recent common ancestors and can provide insight into selective pressures on usage not addressed by 16S rRNA sequences nor fully revealed in codon usage preference analyses.

Because analysis of tetranucleotide frequencies provides insights beyond those inferred from analysis of codon usage biases, we sought to develop an analytical method to examine their conservation across and between prokaryotic genomes. Our goals were to compare alternative models for determining tetranucleotide frequency divergences to understand the extent to which tetranucleotide usage is shared for multiple genomes and their plasmids and bacteriophages, and to determine whether tetranucleotide usage divergences exhibit phylogenetic signal compared with phylogenies based on 16S rRNA.

RESULTS

Representation of Tetranucleotide Combinations in Microbial Genomes

For the studied microbial genomes, we analyzed the tetranucleotide usage deviations from expectations (TUD) to de-

[6]Corresponding author.
E-MAIL Prided01@med.nyu.edu; FAX (212) 252–7164.
DOIFTNArticle and publication are at http://www.genome.org/cgi/doi/10.1101/gr.335003.