Molecular Evolution of a Tree of Life and the Last Universal Common Ancestor.

W.L. Duax, R. Huether, D. Diziak, Dept of Structural Biology. SUNY Buffalo, NY 14203. duax@hwi.buffalo.edu

Early attempts to use genomic analysis to create an evolutionary tree of all species produced contradictory results. Horizontal gene transfer was considered the source of the disparity. Because it is unlikely that the ribosome or its integral proteins arose from horizontal transfer we are exploring the potential to trace evolution with each of the 50 ribosomal proteins present in bacterial ribosomes and to compare the 50 independent trees for consistency. In the course of doing so we have discovered the power of having all members of a protein family perfectly aligned. The three dimensional folds of each of the fifty ribosomal proteins (RibP) are fully conserved in all bacterial species for which complete genomes have been reported (2800 as of 8/1/11). Over 98% of all members of each bacterial RibP (bRibP) can be accurately aligned with retention of a small number of fully conserved identities commonly including Alanines (Ala), Prolines (Pro), Arginines (Arg), and immutable Glycines (Gly) and precisely located insertion or deletion sites (Indels) of restricted size. Accurate full length alignment permits identification of sequence positions where a single residue difference can separate Gram positive from Gram negative bacteria. Co-evolution of different combinations of amino acids in other positions in the aligned sequence separates bRibP sequences of different phylum, classes, orders and genera from one another. Accurate alignment makes it possible to follow sequence divergence through 3 billion years of evolution, reveals that the entire sequences of the ribosomal proteins of individual genera have remained over 95% identical throughout their evolution, detects Gene bank errors in gene length and annotation, and reveals that codon bias and use in ribosomal proteins is not determined by tRNA population. We detect significant sequence homology between all bRibPs and RibPs in the same two dozen chloroplasts (clRibP) and find that the greatest sequence homology is between RibPs of cyanobacteria and these clRibPs. Of the RibPs analyzed thus far, the sequence and three dimensional structure of ribosomal protein S19 from the small subunit appears to be the most highly conserved in all species. When we align 3987 sequences of S19 (2353 bacteria, 1528 eukaryotes, and 106 archaea) only two residues are fully conserved (a Gly and an Arg). One or more Gly residues in each RibP family are immutable because the Gly conformations (phi and psi) are in regions of the Ramachandran plot where other amino acids are rarely tolerated. Arg conservation is associated with maintenance of charge balance and direct interaction with specific sites on rRNA and tRNA involved in ribosomal function. Sixteen other sequence positions have 90% conservation of amino acid identities and one sequence position, immediately adjacent to the fully conserved Gly, is occupied by only two amino acids.

An Asp in this position isolates 95% of all gram-positive (G+) bacteria (866). An Asn in this position isolates 1487 bacteria, 1528 eukaryotes and 106 archaea. This [Asp,Asn] Gly sequence forms a tight turn with an internally hydrogen-bonded network and directly interacts with ribosomal RNA via five additional hydrogen bonds [figure 1]. The amino acid composition of the S19 sequences is found to diverge in the order G+ bacteria, to G- bacteria, to eukaryotes. Optimizing GARP content of the minimum universal fingerprint of S19 isolates actinobacteria. Twelve species of Actinobacteria are found to have the highest GARP composition. All of the RibPs of the large and small subunits in these species have full length sense/antisense open reading frames, a hallmark of the DNA of genes of ancient species. Examination of codon use in the proteins of the 30S subunits of six of these species reveals that 24 codons in which the third nucleotide is A or T are never or rarely found in their DNA. No tRNAs cognate to these "unused" codons are found in the genomes of those six species. These findings reveal that not all species use a 64 letter code, that the earliest peptides and proteins had a bias in amino acid composition favoring amino acids of tRNA synthetase II family, that immutable Gly residues are the key to aligning all members of major protein families, that ribosomal protein S19 is a rosetta stone for accurate determination of a rooted evolutionary tree of all living species and that the last universal common ancestor at the root of the tree is closely related to either *K. radiotolerans* or *C. flavogenans*.

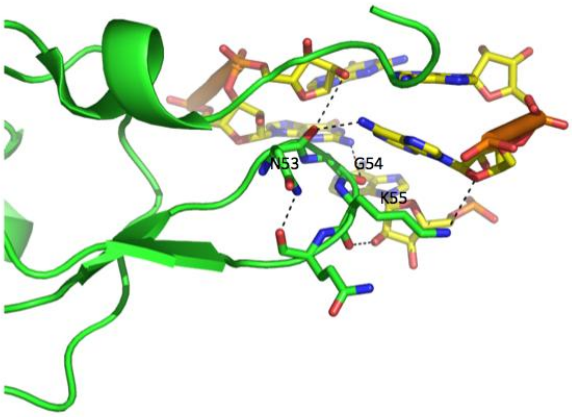

Figure 1: Residues N53G54K55 in the gram negative bacterium *T. thermophilus* creates a hydrogen bonding network with the rRNA and intera residues in S19.

Support in part by: Mr Roy Carver, Stafford Graduate Fellowship, Caerus Forum Fund, The East Hill Foundation and the generous help from a number of High School students from the Buffalo NY area.