# A MUTUAL INFORMATION ANALYSIS OF tRNA SEQUENCE AND MODIFICATION PATTERNS DISTINCTIVE OF SPECIES AND PHYLOGENETIC DOMAIN

FRANCISCO M. DE LA VEGA, CARLOS CERPA, GABRIEL GUARNEROS

*Department of Genetics and Molecular Biology, CINVESTAV-IPN, A.P. 14-740, Mexico City 07000, Mexico*

ROBERT M. FARBER

*Complex Systems Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM and Santa Fe Institute, Santa Fe, NM, USA*

Transfer RNA is a central molecule in the transmission of information from DNA to proteins as depicted in the central dogma of molecular biology. It is the first RNA molecule for which tertiary structure was determined and is the most extensively modified nucleic acid in the cell; soon after its discovery it was shown to contain modified nucleosides. After the first nucleotide sequence of a tRNA was reported in 1964, vast amount of gene and RNA sequences of tRNA molecules has accumulated, and they were recently collected by M. Sprintze in a specialized sequence databank. There has been considerable effort in finding sequence and structural features within tRNA isoacceptors that may define their identity to aminoacyltRNA synthases. Conversely, little is known about specific features of tRNAs that are exclusive to a species, taxa or phylogenetic domain level. Upon simple examination of the tRNA sequence data available, universally conserved features, such as the CCA 3' terminus in the acceptor stem, and several modified nucleosides, indicated a common evolutionary origin. However, it can be demonstrated that some sequence features, including a subset of the possible modified nucleosides, are specific for Archea, Bacteria, or Eucarya. This suggests that these have evolved after the three phylogenetic domains divided. Here we have used an information theoretic quantity, called mutual information, a measure of covariation, to further pursue the search for specific features of tRNAs to particular taxa. Our data included the complete set of structurally aligned tRNA sequences from *E.coli*, *B. subtilis*, *H. volcani*, *M. capricoleum*, *H. influenzae*, *S. cerevisiae* (cytoplasmic) and *S. cereviseae* (mitochondrial), taken from the Sprintze database or GenBank and removing redundancies. Our preliminary results suggest that certain pairs of nucleotides along the tRNA sequences of a particular species, do not have independent mutations or modication patterns as measured by mutual information. After excluding from the analysis the obvious covariations arising from the tRNA helical stems, we found pairs of sites with high mutual information that were highly predictive of each dataset. For some instances a statistical argument can not be sustained, mainly because some datasets are composed by a reduced set of examples, like in *M. capricoleum* or yeast mitochondrial tRNAs.

These results are compared with the behavior of a neural network trained with the same data and that is able to classify tRNAs sequences according to species. The species-specific interdependence among some sequence sites may have implications for previously undescribed structural or functional relationships between the tRNA molecules and the maturation and/or translational machineries unique for each taxa. Complex interactions between tRNA nucleotides and components of the translational apparatus exist. Therefore, a detailed taxonomic study of these features as started here, may reveal the interplay that ensued during the evolutionary story of such an intricate and essential piece of the cell's biosynthetic machinery.

This research has benefited from grant 4826N from the National Council of Science and Technology (CONACyT, Mexico) to FMV and a Howard Huges Medical Institute grant to GG. Part of this work was developed during the *Patterns in Biological Sequences* Workshop held on June 1995 at the Aspen Center for Physics.