

# INFORMATION THEORY ANALYSIS OF WHOLE GENOME SEQUENCES

**Dr. S. Krishnaswamy**

Centre of Excellence in Bioinformatics

School of Biotechnology

Madurai Kamaraj University, Madurai 625 021

krishna@mrna.tn.nic.in , mkukrishna@gmail.com

Enormous sequence data has become available with the advent of high throughput sequencing and database availability. Understanding the mathematical laws and operational principles governing the storage and processing of information in these one-dimensional sequence will help unravel the complexity of the system and explore the systematics of the biological system. The talk looks at genomes from an information theoretic point of view. An information theoretic analysis of 142 prokaryotic genomes and 155 eukaryotic chromosomes is presented. Our results suggest that information theory provides a framework for understanding messaging strategies and looking at the evolutionary aspects of genomes. Despite size and compositional variations, both prokaryotic and eukaryotic genomes do not deviate significantly from an equiprobable and random situation. Chromosomes in eukaryotic organisms maintain similar information densities ( $I_d$ ) suggestive of common informational restraints. Interestingly, in *A. thaliana* and human chromosomes the  $I_d$  values are similar also for the two arms of the chromosomes. Inter and intra-strand  $A=T$  and  $G=C$  rules of Chargaff are broadly adhered to in all genomes. An inverse correlation is seen between the contribution of compositional redundancy (RD1) and Shannon redundancy, arising from dinucleotide (RD2) and trinucleotide (RD3) frequency distributions. This suggests a balance between strategies to combat error involving variation in nucleotide composition and variation in the order of occurrence of nucleotides.