Name of Scholar: Charu Kapil

Name of Supervisor: Dr. Mohamad Aman Jairajpuri

Name of Co-Supervisor: Dr. Amit Sharma (SCBG, ICGEB, New Delhi)

Department: **PCEL, Department of Bioscience, JMI, New Delhi**

Title of thesis: ***Bioinformatics-based genomic and proteomic analyses of malaria parasite Plasmodium falciparum***

Malaria pathogenesis is cause by *Plamodium* species resulting into millions of clinical cases and deaths annually. *Plasmodium falciparum* is extremely AT-rich genome (80% AT content) of genome size nearly 23 Mbp wrapped in 14 chromosomes and with approximately 5500 proteins. Genomic and proteomic comparison study of this parasite with others will give the sequence differences. These *Pf* sequences containing unique insertions can act as potential targets against which drugs can be prepared.

We compared 85 eukaryotic genomes including *P. falciparum* belonging to different genomic compositions: AT-rich, GC-rich and neutrals class and analysed spatial distribution of nucleotides in a codon in these genomes. Genomes were converted to tri-nucleotide codons followed by probability of occurrence of each nucleotides (ATGC) calculation at all the three positions of a codon. Based on similar probabilities of nucleotide occurring at the first position 85 eukaryotic organisms were clustered in four A-, T-, G- and C- groups. We studied each group for the deviation of all four nucleotide and pair of nucleotides (AT/GC) at the end two positions in a codon and four nucleotides trend at the first position itself. *Pf* belongs to A- and C- group, nucleotides Thymine and Cytosine introduces more variation at the second and third position and the first position has more variation of Guanine and Thymine nucleotide. Likewise, di-nucleotide AT content is more at the third position than second position. Further we carried forward the analysis at the proteome level by picking 12 (out of 85) representative organisms from AT-rich, GC-rich and neutral class including 6 pathogens.

Abstract

Whole proteome of 12 organisms were compared and based on sequence similarity 871 common proteins were catalogued. These proteins were analysed for the unique insertions, length of insertions and their residue preferences. Our key observations include >40 residue length insertions were abundant in pathogens - *P. falciparum* followed by *T. gondii* and *L. major*; the causative agents of malaria, toxoplasmosis and leishmaniasis. Residues like Glu and Asp are over-represented in proteomes irrespective of genomic compositions or pathogenicity where as Asn is exclusive to *Pf. P. falciparum* proteome contains highest number of proteins with unique insertions (124) with >40 residue insertions. Also >40 residue insertions and unique insertions were found predominantly in helicases, kinases and aminoacyl tRNA synthetases from *Pf, Tg & Lm,* whereas 62 common proteins contained no insertions in 12 studied proteomes.

The proteome study provides a comprehensive knowledge of insertions among common proteins of 12 eukaryotic organisms. Among all studied proteome *Pf* shows more number of specific protein families (helicases, kinases and t-RNA synthetases) with >40 residue, unique insertions and several proteins with no insertions. Presence of large sized insertions in pathogen might indicate their crucial role in regulating the parasite's function.

Out of the three major protein families: helicases, kinases and tRNA-synthetases, we analysed tRNA-synthetase in detail for structural mapping of these unique insertions. Database searches and Hidden Markov Model based method were used for classification of 37 tRNA synthetases in two classes. Phylogenetic analysis listed few aaRS enzymes closer to plant and of bacterial origin. Also we modeled few aaRSs for sequence-structural mapping of these unique inserts. This analysis therefore opens up a new study area that may unravel specific roles played by insertion sequences. Structural mapping of such large sized and unique insertions containing proteins from aaRSs in parasitic organisms raise questions about their critical involvement in modulating their function.