Quantifying Genome Sequence Compatibility Between Viruses and Their Hosts

Patrick Schreiner

Dinucleotide CpG under-representation is well known to occur in mammalian genomes. Previous studies have concluded that the leading hypothesis for this under-representation of CpG dinucleotide content can be explained by methylation of cytosine residues and the corresponding deamination that occurs in 5-metycytosine. Viruses that have infected humans for an extended period of time exhibit these same dinucleotide CpG deficiencies, although CpG dinucleotide content does not result in methylation throughout the viral genome. It is suggested that genomic signatures of viral genomes have a direct correlation with the content of the host DNA in which they infect, with few exceptions. This correlation is expressed significantly at both the dinucleotide and tetranucleotide level. The similarity of the viral genome to its specific host's genome allows the virus to more efficiently escape immune response. We have developed software for the analysis of di-, tri-, and tetra-nucleotide usage correspondences between a wide variety of viruses and their hosts. Examination was performed not only at overall genome compositional biases but using a sliding window approach.