ABSTRACT

Genome annotation came into existence when the first human genome sequencing project got completed in the year 2003. Next generation sequencing technology has now made it more important. The genome annotation is useful to make a database of the information present in the DNA. Genome annotation can be used to reduce the time and cost involved in the drug discovery process, to custom design drugs, and to develop personalized medicine. The DNA sequences contain huge amount of hidden information which is associated with the functionality of the organism. The DNA regions of interest include exons, introns, genic regions, inter genic regions, untranslated regions, translational start sites, promoter regions, splice sites, microsatellite and minisatellite repeats in the DNA sequences. To identify these features various computational methods have been developed and these are broadly grouped into model dependent and model independent. The model dependent methods need a priori knowledge about the datasets and offer more accuracy. However, due to the model dependency these methods cannot be generalized. On the contrary model independent methods do not require any prior knowledge and are independent of training requirement. The model independent methods generally exploit the presence of certain periodicities specific to a particular region in DNA sequence. Signal processing based methods are highly helpful in identifying the presence of these periodicities for annotating various genomic regions. Analysis of the genomic data using Signal Processing tools is called Genomic Signal Processing. In this research work, Signal Processing based methods have been developed for the identification of short exons, acceptor splice sites, microsatellites, and minisatellites in the DNA sequences. The simulation of these algorithms has been done in MATLAB (R2010) environment and performance has been evaluated using standard data sets and performance metrics.