SUMMARY

In the post genomic era, annotating the increasing number of genomes coming out from the genome sequencing projects is becoming a huge challenge to researchers. The challenge is especially difficult when it comes to annotating the eukaryotic genomes due to the size of sequence data and the complexity of genes and other structural and functional components therein. This necessitates the development and use of computational methods for automated genome annotation. The type of features that can be detected and described in any genomic sequence include the location of the protein-coding genes; the structure of the detected genes; the probable translations of every transcript into a protein product; the location of repetitive sequences and their nature; and the location of genes encoding noncoding RNAs. The main purpose of this work was to explore new methods of genome annotation by focusing on two of the major and primary components that are detected and described in any eukaryotic genomic sequence. This includes the location and characterizing the structure of protein-coding genes and identifying repetitive DNA sequences. Six stages were decided upon to achieve this goal. These include the compilation of gene prediction data sets and development of a new method for generation of non-homologous data sets; evaluation of the existing eukaryotic gene prediction programs; development of an integrative method for prediction of protein-coding genes; development of a new method for identification and characterization of repetitive DNA; demonstration of the new methods in gene prediction and repetitive DNA identification on large eukaryotic sequences; and development of Web-based servers implementing the new methods for easy public access.

In the first stage, all the existing data sets available on gene prediction were compiled. This database of existing sequence data sets is maintained on GeneBench Web server. While the compilation of existing data sets is useful, researchers have their sequences with which they evaluate the gene finding programs. However, a good

evaluation always requires the use of a non-homologous data set. GeneBench server also allows for the creation of non-homologous gene data set based on similarity among the protein product. The server uses the PROSET program to compare the input sequences and produces a filtered set of non-homologous sequences. The method uses the available CDS information to generate and compare the protein products of the query genomic DNA sequences. In addition, the server also has a compilation of all the standard and new predictive accuracy measures that has been used till date. The Web implementation of GeneBench server is available at http://www.imtech.res.in/raghava/genebench/.

Gene identification methods are categorized into two based on whether they learn from training set to create a gene structure model on which the query sequences are modeled (ab initio methods) or they employ similarity to known genes and proteins to locate protein coding genes in DNA (similarity based methods). One of the important criteria to be fulfilled before developing any new method is to establish the predictive accuracy achieved by existing algorithms in the field. Evaluation of the ab initio based methods on HMR195 & Burset/Guigo data sets suggested a sensitivity and specificity of ~90% at nucleotide level and sensitivity and specificity of ~~75% at exon level with a comparatively high false positive rate. The BLASTX based similarity search method has a slightly lower accuracy at nucleotide level and near zero prediction at exon level. The advantage of using similarity-based method is the identification of the gene in almost all sequences. However, this is not enough to accurately predict genes in every single instance. This necessitates the development of new methods that utilize the consensual properties of the predictions from different programs, and therefore will improve prediction accuracy.

In this thesis, a combination based gene prediction method is developed through the integration of evidences from similarity searches against protein sequence databases with the prediction from the ab initio gene finders. The method improves upon the prediction accuracy achieved by both approaches through a novel strategy of including evidences from similarity to available intron sequences is attempted to filter the high number of false positives. Two ab initio methods—Genscan and HMMgene, and three ab initio based combination methods—EUI, EUI-Frame and GI methods are used for integration of BLASTX based similarity search against RefSeq protein sequence

database. Additional information from splice site predictions is also included for improving the accuracy of exon prediction. Based on the study on inclusion of different evidences, a new method EGPred is developed. Comparison of the new method, EGPred, against existing programs on two different sequence data sets suggests an increase in gene prediction accuracy of 4%-10% at the exon level.

Repetitive DNA sequences constitute a major component of eukaryotic genomes. Though relatively common in genomes and implicated in various pathologies, repeat sequences are the most understudied of all genomic components. It is therefore important to locate and analyze the repetitive sequences. Fast Fourier transformation (FFT) is a powerful mathematical tool that is used to identify periodicity in biological sequences. Though surprisingly simple and very effective, FFT has not been extensively used for prediction of repeats. In this technique a DNA sequence is converted in a series of binary digital signals, which are then Fourier transformed to obtain a peak at frequency that is inverse of the periodicity observed in the sequence, i.e. $f = 1/3$ would give a peak for periodicity of 3 bp while 1/4 is for periodicity of 4 bp. The new method, Spectral Repeat Finder (SRF), is used to predict all types of direct and dispersed repetitive sequences in microsatellites, minisatellites and other longer repeat sequences. Initially, FFT is performed on query sequence to identify the length of various repeat units in any genomic sequence. In the next step, a second window-based FFT analysis is performed on DNA sequence to identify the protein coding regions that has these repeat units. After identifying the repeat regions, an exact method is used to characterize the repeat unit and to derive a consensus sequence of the repeat unit. The method accurately applied on known and previously annotated repeat regions.

The new gene prediction method, EGPred, is used to annotate the human chromosome 13. Predictions obtained from this experiment show a large number of unaccounted and possibly novel genes in the genomic sequence that are potentially protein coding based on similarity to protein sequences. Results from the evaluation also emphasize that gene prediction in eukaryotes is not yet completely solved. On the other hand, SRF method is used to characterize the repetitive sequences in yeast genome. Results suggest that the SRF method produces more accurate results and more numbers of repeats in comparison to other repeat finding method due to the use of motif finding exact

method. The method is highly suitable to locate and identify repeat units of 2 bp to even more than 300 bp long.

The methods developed in this thesis for identification and of protein-coding genes and repetitive sequences in eukaryotic genomic sequences are implemented as Web based servers and are freely available to public for use through Internet. The EGPred method is available at http://www.imtech.res.in/raghava/egpred while the SRF method is available at http://www.imtech.res.in/raghava/srf. The purpose of this thesis was consequently realized.