CONCLUSIONS

Current genomic revolution has unlocked the potential to understand the gene regulation at molecular, cellular and physiological levels. The first step in this process is to identify the genes present in a genome and study the expression patterns of the gene influenced by regulatory signals. Several programs are available in the public domain that can identify genes from the DNA sequence using 'signals' and 'contents' of the DNA sequence. Gene prediction programs using 'contents' information are limited from identifying only protein coding genes in the genome. So in order to derive an *ab initio* gene prediction system purely based on signals, in this project I attempted to create models for gene regulatory elements.

The start and end of any gene is marked by their promoter and termination signals where from RNA polymerases begin and terminate transcription. The transcription start site was initially identified by Down and Hubbard using generalised linear model based probabilistic algorithm called Eponine (Down and Hubbard, 2004; Down and Hubbard, 2002). In this project, I attempted a number of methods including Eponine to identify transcription termination signals responsible for RNA polymerase stop and release from DNA.

Termination of polymerase does not happen at the cleavage site and the RNA polymerase transcribes DNA even 2 kb downstream before releasing from the DNA. Recent experiments confirm the presence of a pause site downstream of the cleavage site required for transcription termination. In chapter 1, I have detailed the mechanism of transcription termination and compared with other systems known to occur *in vivo*. Attempts to identify the pause elements have so far not been successful in deriving a consensus sequence. However experimental and computational analyses indicate the sequence might be A-rich and G-rich and bind MAZ and Sp1 protein to stop transcription from running-over to the neighbouring genes.

In this project, I first used base compositional analysis to study any significant changes in the nucleotide distribution in the sequences around cleavage site. The differences in the composition were found concentrated within 100 and 50 bases upstream and downstream of cleavage site and these are linked to the poly(A) signal and GT rich region known earlier.

No significant changes were found in the sequences where polymerase is likely to pause. Then, I investigated for the presence of any secondary structures that can potentially stop polymerase as a similar mechanism is found in prokaryotes and histone genes. So to analyse this, I used Nussinov and Zuker algorithms. Base pair maximisation principle-based Nussinov algorithm did not find any stem-loop structures. Free energy minimization based Zuker algorithm, however, predicted the possibility of RNA secondary structure in the sequences 100 to 650 bases downstream of cleavage site. Correlation with GC and GT percentage showed they are unlikely to be caused by sequence artefacts. Confirming these structures using biochemical experiments will help us to understand the mechanism of transcription termination of protein coding genes and correlate them with histone and prokaryotic gene transcription termination.

After analysing for secondary structures in DNA, I used the probabilistic machine learning algorithm based on Bayes theorem and Generalized Linear Models, Eponine, for scanning motifs responsible for transcription termination. The model captured poly(A) signal and auxiliary sequence motifs along with a few multiplex signals that might be responsible for polymerase II pause and termination. An evaluation of this termination model against annotated human chromosomes shows that the model performs better than existing methods. However a significant number of predictions also appear near the annotated start site and first intron of genes. In chapter 3, I have tried to explain these biases and false positives at this region using hypothesis derived from previous knowledge. I propose that a significant number of predictions made by the model that are not correlated with available annotations are not really false predictions and they are likely to have biological functions. It would be interesting to test these hypotheses by devising appropriate molecular and biochemical experiments.

Apart from the bias towards transcription start site and first intron, I found approximately 10% of predictions lie within genes and their density is correlated with gene length and intron size. Interestingly shorter introns were found to have higher prediction density and most of them are likely to be alternative termination or polyadenylation site of the gene. Early experiments show this is possible as at least 22% of mRNAs was recorded to undergo alternative polyadenylation often in a tissue- and time-specific manner (Legendre and Gautheret, 2003). Previous programs developed to find the end of the gene and alternative

polyadenylation site are mainly dependent on poly(A) signals and Eponine differs from them by using other downstream signals. A comparison with one such previous program, ERPIN (Legendre and Gautheret, 2003), showed Eponine performed better in identifying transcription termination sites.

I then extended the application of Eponine to develop splice site and translation models to meet the objective of creating an *ab initio* gene prediction system. These models are explained in chapters 4 and 5. Donor and acceptor site models were trained from sequences from chromosome 22 along with appropriate negative datasets. Positional variations in splice site models were captured using a Delta distribution rather than the usual Gaussian distribution. The models picked the known signals near donor and acceptor sites. Acceptor sites, as expected, were difficult to predict relative to the donor site as acceptor sites show variation in the regulatory elements (Lund *et al.,* 2000). Moreover, the Eponine acceptor site model did not capture branch point signal where lariat formation occurs. A comparison of the models with annotated sites of chromosome 20 showed the models have good positional accuracy and performed comparably with GeneSplicer (Pertea *et al.,* 2001) and StrataSplice (Levine, 2001a). I also noticed that there is a scope for improvement of performance of Eponine splice site models by using local GC variation as employed by StrataSplice.

Likewise, I attempted to identify translation start and stop codons and regulatory elements near by that determine translation initiation and termination by the ribosomal machinery. Translation start model learnt the famous Kozak sequence and performed better than NetStart (Pedersen and Nielsen, 1997b) although less well than ATGpr (Salamov *et al.,* 1998a).

After training all the Eponine models, I combined them using the dynamic programming framework based GAZE (Howe *et al.,* 2002) to develop a gene prediction system called GenePred. Various versions of GenePred developed by tweaking the input features and score values showed all the models are comparable with GENSCAN (Burge and Karlin, 1997) in identifying genes from the genomic sequence. In cases of Novel_transcripts and Putative genes, GenePred was found to be better than GENSCAN in identifying these genes. However, GenePred had difficulty in determining the annotated exon-intron structure of the

genes. This is expected as the GenePred uses only signal information in predicting the candidate genes.

Thus in this project, I developed various models that influence gene regulatory elements and linked them together to derive an *ab initio* gene prediction system that uses only these gene regulatory signals and not dependent on protein coding information. During this attempt, I found interesting observations like distribution of termination sites near transcription start site, first intron and short introns. Results from experiments confirming these observations will help us to discern the transcriptional machinery and reconsider the current concepts of gene regulation in the eukaryotic genome.