Synopsis

Promoter region is the key regulatory region, which enables the gene to be transcribed or repressed by anchoring RNA polymerase and other transcription factors, but it is difficult to determine experimentally. Hence an *in silico* identification of promoters is crucial in order to guide experimental work and to pin point the key region that controls the transcription initiation of a gene. Analysis of various genome sequences in the vicinity of experimentally identified transcription start sites (TSSs) in prokaryotic as well as eukaryotic genomes had earlier indicated that they have several structural features in common, such as lower stability, higher curvature and less bendability, when compared with their neighboring regions. In this thesis work, the variation observed for these DNA sequence dependent structural properties have been used to identify and delineate promoter regions from other genomic regions. Since the number of bacterial genomes being sequenced is increasing very rapidly, it is crucial to have procedures for rapid and reliable annotation of their functional elements such as promoter regions, which control the expression of each gene or each transcription unit of the genome. The thesis work addresses this requirement and presents step by step protocols followed to get a generic method for promoter prediction that can be applicable across organisms. The each paragraph below gives an overall idea about the thesis organization into chapters.

An overview of prokaryotic transcriptional regulation, structural polymorphism adapted by DNA molecule and its impact on transcriptional regulation has been discussed in introduction chapter of this thesis (chapter 1).

Standardization of promoter prediction methodology - Part I

Based on the difference in stability between neighboring upstream and downstream regions in the vicinity of experimentally determined transcription start sites, a promoter prediction algorithm has been developed to identify prokaryotic promoter

sequences in whole genomes. The average free energy (E) over known promoter sequences and the difference (D) between E and the average free energy over the random sequence generated using the downstream region of known TSS (RE_{av}) are used to search for promoters in the genomic sequences. Using these cutoff values to predict promoter regions across entire *E. coli* genome, a reliability of 70% has been achieved, when the predicted promoters were cross verified against the 960 transcription start sites (TSSs) listed in the Ecocyc database. Reliable promoter prediction is obtained when these genome specific threshold values were used to search for promoters in the whole *E. coli* genome sequence. Annotation of the whole *E. coli* genome for promoter region has been carried out with 49% accuracy.

Reference

Rangannan, V. and Bansal, M. (2007) **Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability.** *J Biosci*, **32**, 851-862.

Standardization of promoter prediction methodology - Part II

In this chapter, it has been demonstrated that while the promoter regions are in general less stable than the flanking regions, their average free energy varies depending on the GC composition of the flanking genomic sequence. Therefore, a set of free energy threshold values (TSS based threshold values), from the genomic DNA with varying GC content in the vicinity of experimentally identified TSSs have been obtained. These threshold values have been used as generic criteria for predicting promoter regions in *E. coli* and *B. subtilis* and *M. tuberculosis* genomes, using an *in-house* developed tool 'PromPredict'. On applying it to predict promoter regions corresponding to the 1144 and 612 experimentally validated TSSs in *E. coli* (genome %GC : 50.8) and *B. subtilis* (genome %GC : 43.5) sensitivity of 99% and 95% and precision values of 58% and 60%, respectively, were achieved. For the limited data set of 81 TSSs available for *M. tuberculosis* (65.6% GC) a sensitivity of 100% and precision of 49% was obtained.

Reference

Rangannan, V. and Bansal, M. (2009) **Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition.** *Mol Biosyst*, **5**, 1758 - 1769.

Standardization of promoter prediction methodology - Part III

In this chapter, the promoter prediction algorithm and the threshold values have been improved to predict promoter regions on a large scale over 913 microbial genome sequences. The average free energy (AFE) values for the promoter regions as well as their downstream regions are found to differ, depending on their GC content even with respect to translation start sites (TLSs) from 913 microbial genomes. The TSS based cut-off values derived in chapter 3 do not have cut-off values for both extremes of GC-bins at 5% interval. Hence, threshold values have been derived from a subset of translation start sites (TLSs) from all microbial genomes which were categorized based on their GC-content. Interestingly the cut-off values derived with respect to TSS data set (chapter 3) and TLS data set are very similar for the in-between GC-bins. Therefore, TSS based cut-off values derived in chapter 2 with the TLS based cut-off values have been combined (denoted as TSS-TLS based cutoff values) to predict promoters over the complete genome sequences. An average recall value of 72% (which indicates the percentage of protein and RNA coding genes with predicted promoter regions assigned to them) and precision of 56% is achieved over the 913 microbial genome dataset. These predicted promoter regions have been given a reliability level (low, medium, high, very high and highest) based on the difference in its relative average free energy, which can help the users design their experiments with more confidence by using the predictions with higher reliability levels.

Reference
Rangannan, V. and Bansal, M. (2010) **High Quality Annotation of Promoter Regions for 913 Bacterial Genomes.** *Bioinformatics*, **26,** 3043-3050.

Web applications

PromBase : The predicted promoter regions for 913 microbial genomes were deposited into a public domain database called, PromBase which can serve as a valuable resource for comparative genomics study for their general genomic features and also help the experimentalist to rapidly access the annotation of the promoter regions in any given genome. This database is freely accessible for the users via the World Wide Web `http://nucleix.mbu.iisc.ernet.in/prombase/`.
EcoProm : EcoProm is a database that can identify and display the potential promoter regions corresponding to EcoCyc annotated TSS and genes. Also displays predictions for whole genomic sequence of *E. coli* and EcoProm is available at `http://nucleix.mbu.iisc.ernet.in/ecoprom/index.htm`.
PromPredict : The generic promoter prediction methodology described in previous chapters has been implemented in to an algorithm 'PromPredict' and available at `http://nucleix.mbu.iisc.ernet.in/prompredict/prompredict.html`.

Analysing the DNA structural characteristic of prokaryotic promoter sequences for their predominance

Sequence dependent structural properties and their variation in genomic DNA are important in controlling several crucial processes such as transcription, replication, recombination and chromatin compaction. In this chapter 6, quantitative analysis of sequences motifs as well as sequence dependent structural properties, such as curvature, bendability and stability in the upstream region of TSS and TLS from *E. coli*, *B. subtilis* and *M. tuberculosis* has been carried out in order to assess their predictive power for promoter regions. Also the correlation between these

structural properties and GC-content has been investigated. Our results have shown that AFE values (stability) gives finer discrimination rather than %GC in identifying promoter regions and stability have shown to be the better structural property in delineating promoter regions from non-promoter regions.

Analysis of these DNA structural properties has been carried out in human promoter sequences and observed to be correlating with the inactivation status of the X-linked genes in human genome. Since, it is deviating from the theme of main thesis; this chapter has been included as appendix A to the main thesis.

General conclusion

Stability is the ubiquitous DNA structural property seen in promoter regions. Stability shows finer discrimination for promoter prediction rather than directly using %GC-content. Based on relative stability of DNA, a generic promoter prediction algorithm has been developed and implemented to predict promoter regions on a large scale over 913 microbial genome sequences. The analysis of the predicted regions across organisms showed highly reliable predictive performance of the algorithm.