Synopsis

Understanding the structural organisation of genome and insightful information of gene has helped in exploring the regulation and expression of the gene at a molecular and cellular level. In eukaryotes, the gene regulation is a complex process compared to prokaryotes regarding the presence of distant promoter region and noncoding elements inside the gene. After the report of noncoding DNA, previously known as junk DNA, the functionality of these DNA has been enlightened. Advancement of technologies in DNA sequencing, genetic engineering has helped in the repertoire of sequence information of the functional region of coding and noncoding regions of numerous organisms. Noncoding DNA present in the upstream of the gene is known as promoter which acts as binding elements for regulatory factors as well as transcription machinery to modulate transcription. Structural properties of promoter regions and presence of cis-element act as determining a factor for the interaction of trans-acting elements. The DNA structure of the promoter regions varies according to its length and base composition. These stretches of DNA can have several features for duplex DNA and found to be different from the canonical B-DNA structure. Very often, structural properties of DNA such as DNA stability, bendability and intrinsic curvature have high biological relevance for the protein binding. These structural properties predict the change at local and/or global level gives rise to the binding prevalence for the regulatory factors which can modulate gene expression. In higher eukaryotes such as plant, several other factors also play an important role in gene expression. Inside a gene, the presence of noncoding region can act as a regulatory factor for gene expression. These noncoding elements can give stability to the synthesised RNA as well as modulate the expression level and tissue specificity of a gene.

In this thesis, a thorough study of these properties, which are linked to the gene regulation and expression, was carried out. The majority of the work is focused on the influence of gene architecture and promoter properties on level and breadth of gene expression in plants. The thesis work is divided into four sections, which are

explained in detailed below. The first section is primarily focused on the distribution of coding and noncoding region concerning its length and base composition in dicots of *A. thaliana* and monocots of *O. sativa*, *S. bicolor* and *Z. mays*. An extensive study was also done on the influence of promoter properties as well as gene architecture on the level of expression and on breadth (the number of tissue in which a gene expressed) in the second section. The third part of the work was followed by the strength of promoter properties and gene structure of gene expression in plants. The last section of the work is based on the sequence specificity and structural properties of the binding sites of an OsMADS1 transcription factor in rice.

A detailed overview of functional aspects of coding and noncoding region as well as the structural features of promoter regions on gene expression are discussed in the first chapter (Chapter 1) of the thesis. Brief descriptions of transcription factor binding site and motif enrichment in the binding regions are also explained in this chapter.

Gene architecture varies with GC composition of genes in gramineae (Chapter 2)

The bimodal distribution of GC% of genes in monocots has been reported earlier, whereas dicots have unimodal distribution. In this study, the two populations are divided based on the GC% of genes in three monocots (rice, sorghum and maize) by taking the intersection as a separation point. It was reported earlier that GC% of concatenated intron shows a bimodal distribution, which gives bimodality to the gene. The present study of this thesis highlighted that GC% of exons and concatenated exons have the similar kind of bimodal distribution as like genes. Furthermore, a clear bimodal distribution was also seen for single exon genes. Single exons genes are GC-rich in *O. sativa*, *S. bicolor* and *Z. mays* while the similar scenario is absent in *A.thaliana*. Indeed a high correlation (0.80) was found between GC% of HighGC class genes and the GC% of corresponding concatenated exons in monocots. Genes found within HighGC class and LowGC class were functionally distinguishable by statistical analysis of their gene ontology categories. Thus the new finding of this study is that GC% of exon itself has a bimodal distribution in rice, maize and sorghum which makes the distribution of GC% of gene bimodal. Both the classes of genes are different in their molecular function and cellular component significantly.

Gene structure and promoter architecture influence variation of expression in plants (Chapter 3)

Gene expression is a vital process used by all living cells for their survival, and its modulation at cellular level gives rise to phenotypic variation. Despite the availability of massive gene expression data, the regulation of expression remains unclear. Variation of expression of a gene is evident among individuals, between individuals and tissues. Previous studies on animal and plants have reported the influence of gene structure and promoter architecture on gene expression level and breadth (tissue specificity) however, results were more contradictory in plants. In this study, transcriptome atlas data of different tissue types at various developmental stages of *A. thaliana, O. sativa, S. bicolor* and *Z. mays* were used to understand the relationship between gene components and expression. The results of this chapter revealed that both length and GC% of gene components influences gene expression level and breadth in a similar direction, however, this relationship is plant specific. Expression breadth is positively linked with the gene components like intron content of primary transcript (PT) (%), GC% of $5'$ UTR and $3'$ UTR while negatively related to GC% of PT and difference in GC% of exon and intron. Moreover, Impact of intron content of PT (%) is significantly more for genes with high expression level. Furthermore, rice is found to be an intermediate regarding the variation of gene components on gene expression between dicots and monocots.

Promoter regions are the binding sites of the transcription factor, which have specific structural properties that regulate both gene expression level and breadth. This study has established a relationship between expression parameters and promoter architecture. Structural properties like stability, bendability and curvature of promoter regions were analysed in this chapter. Interestingly promoter regions of tissue-specific and lowly expressed genes are less stable in arabidopsis, rice and sorghum. Bendability of core promoter sequences of constitutively expressed genes has been noticed to be less bendable in plants by using prediction model DNase1 sensitivity and nucleosome positioning preference. Promoter regions of lowly expressed genes are highly curved in sorghum and maize. Although the gene structure is found common in expression study for all plants, maize is different from other monocots in its promoter architecture, which might have evolved separately to regulate gene expression.

Various genomic properties are deterministic factors for gene expression level and breadth (Chapter 4)

In the higher organism, gene expresses differently in different tissues. The number of transcripts and the specificity of expression eventually shape the phenotype. There are many gene components and promoter properties linked to both gene expression level and breadth in similar fashion which has been discussed in chapter 3. The reason could be a common molecular pathway. The relative strength of each genomic component on gene expression parameters would help to clarify the underneath molecular mechanism. In this chapter, a multivariate multiple regression model was built for each plant and influence of every genomic trait has been compared between gene expression level and breadth directly. Results of this analysis revealed the determining factor of gene expression present commonly in the plant. This study uncovered that among gene components, intron content of PT (%) is a powerful determinate of tissue specificity while other gene components are governing expression level and breadth in a diversified way. Similarly, among structural properties of the promoter, free energy has been negatively linked to expression breadth. However, DNase1 sensitivity strongly governed the gene expression breadth in monocots and gene expression level in dicots. This specificity of DNase1 sensitivity in dicot and monocot leads this study to examine the compositional difference and motif enrichment in their promoter regions to underneath the evolutionary relationship. TATA box and Y-patch were preferentially noticed in narrowly expressed data set of all plants. An extensive study on the size of the orthologous group and gene expression was performed to expose the shared molecular pathways. Single copy genes were lowly expressed while multi copy numbers of genes are tissue specific.

Genome-wide binding and recognition signature of OsMADS1 have specific DNA structural features (Chapter 5)

DNA-protein interaction is a fundamental process which often regulates the expression of a gene either by activation or by repression. Proteins bind to the DNA are known as transcription factors (TFs) which are sequence specific and recognise the binding sites are known as consensus motif in dsDNA by its DNA-binding domains. The specificity of DNA and TFs interaction is governed by several factors such as consensus motif, the presence of other cofactors, chromatin environment and structural flexibility or shape of the DNA. Studies on genome-wide binding of TFs have addressed that the transcription factor binding motifs (TFBSs) are not the only recognising feature but how do proteins achieve DNA-binding specificity, are still largely unknown.

In this project, genome-wide binding sites information were gathered from *in vivo* ChIP-seq experiment of the transcription factor OsMADS1 in rice (*Oryza sativa*). MADS domain containing transcription factor is a central regulator that performs the diverse function in development whereas *OsMADS1* controls floral development in rice. The genome-wide binding to the target genes and the underlying mechanisms of gene expression are still not uncovered. Aim of this work was to find out the structural properties of the binding regions for the sequence having consensus motif CArG as well as for the region where the cognate DNA-binding site is absent. The results revealed that total 3112 binding sites associated with the gene were concentrated in the intergenic regions, in the vicinity of transcription start sites (TSSs) and within gene bodies. Among the bound DNA, CArG motif is found in most of the sequences and in several cases only A-tracts. Overall, sequences flanking OsMADS1 binding sites show specific DNA structural properties. These bound DNA sequences are characterised by low stability and flexibility while having a highly intrinsic curvature and enriched with the A-tract motif. In addition to that binding peak shows a narrow minor groove which is previously established from X-ray crystal structure of a dimer of the MADS domain of human serum response factor (SRF) bound to DNA. Moreover, in this chapter, hexamer enrichment analysis also revealed the consensus motif of other TF family like MYC, AP2/ERF, bZIP etc. in OsMADS1 bound DNA. Combining the gene expression data of *OsMADS1* knockdown florets with its DNA binding data, the regulatory network was constructed where transcription factors, such as AP2/ERF, bHLH, HSF and chromatin remodelers are found as direct targets.

General conclusion

Summary of the above results of this thesis is presented in **chapter-6**. This thesis illustrates the importance of the noncoding region in gene regulation of plants. Gene expression and TF binding in plants can be modulated by engineering the promoter as well as by gene architecture.