Sequence Analysis of Human Promoter Sequences around TSS and Transcription Factor Binding Site Sequences

A Thesis

submitted to the University of Hyderabad for the award of a Ph.D degree in Dept. of Biochemistry, School of Life Sciences

By

Padmavathi Putta

Department of Biochemistry

School of Life Sciences

University of Hyderabad

Central University (P.O), Gachibowli

Hyderabad – 500046

Andhra Pradesh (India)

DECLARATION

I, Padmavathi Putta, hereby declare that this thesis entitled "Sequence Analysis of Human Promoter Sequences around TSS and Transcription Factor Binding Site Sequences" submitted by me under the guidance and supervision of Prof. Chanchal K. Mitra is an original and independent research work. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

Date:

Name: Padmavathi Putta

Signature of the Student:

Regd.No. 06LBPH04

CERTIFICATE

This is to certify that this thesis entitled “Sequence Analysis of Human Promoter Sequences around TSS and Transcription Factor Binding Site Sequences” is a record of bonafide work done by Ms. Padmavathi Putta, a research scholar for Ph.D. programme in Department of Biochemistry, School of Life Sciences, University of Hyderabad under my guidance and supervision.

The thesis has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Signature of the Supervisor

Head of the Department

Dean of the School

Dedicated To

My Parents & Brother

Acknowledgements

- *I express my deep sense of gratitude to my supervisor Prof. Chanchal K. Mitra, for his encouragement and suggestions throughout my research work. I am privileged to work under his eminent guidance. His wide knowledge and logical way of thinking have been of great value for me. Discussions with him were always enlightening ones in both professional and personal sphere.*
- *I am thankful to Prof. M. Ramanadham, the present Dean and former Deans, School of Life Sciences, for providing me the school facilities.*
- *I thank Prof. K. V. A. Ramaiah the Head, Department of Biochemistry and former Heads, for the facilities provided during my research work.*
- *I owe my deepest gratitude to Dr. S. RajaGopal and his wife Dr. Sridevi who ignited the zeal to opt for a career in research and also for their encouragement and affection towards me.*
- *I am thankful to Prof. Aparna Datta Gupta for her constant encouragement throughout my studies in University of Hyderabad.*
- *I am grateful to Prof. AppaRao Podile, for his support and suggestions from my post graduation onwards. I am also thankful to Late Mrs. Padmaja Podile for her affection.*
- *I am thankful to Prof. N. Shiva Kumar and Prof. P. Reddanna for their suggestions and encouragement.*

- *I thank doctoral committee members, Prof. Anand K. Kondapi and Dr. Mrinal K. Bhattacharyya for their critical analysis and suggestions during my research work.*
- *I am thankful to all the faculty of University of Hyderabad who taught me in my M.Sc and Adv. PG. Diploma in Bioinformatics- who have instilled scientific spirit and guiding me towards research career.*
- *I am grateful to all my teachers from school, college and university who had taught me the value of learning and for encouraging me towards higher studies.*
- *I am thankful to the non-teaching staff from Administration, Department of Biochemistry and School of Life Sciences and Hostel Office for their assistance and help during my stay in the campus.*
- *I am thankful to my seniors for their suggestions and friendly nature. I also thank previous and present project students for their friendly nature and lively environment in the lab. I am thankful to Mr. Dibya Jyothi for his help during work. I thank Mr. Ramesh for his timely help and assistance.*
- *I sincerely acknowledge the financial support from DBT and DBT-CREBB during my research work.*
- *I am thankful to National University of Singapore, for the financial assistance to attend workshop on "Computational Systems Biology approaches to Analysis of Genome Complexity and Regulatory Gene Networks" held in NUS, Singapore from 20-26th Nov, 2010.*
- *I sincerely thank DST, India for the financial assistance to attend workshop and summer school on "IB-PAS 2010: Integrative Biological*

Pathway Analysis and Simulation" held in Bielefeld University, Germany from 21-26th May 2010.

- *I am thankful to DBT-PURSE grant University of Hyderabad, for the financial assistance to attend "ECC10- 9th European Conference on Computational Biology" held in Gent, Belgium from 26-29th Sept, 2010.*
- *It would like to show my gratitude to Dr. Vladimir, Dr. Yury, Dr. Ralf, Lakshmi Narayana, Ramesh, Suguna, Sweta, Siva, Tarak, Ajay and AlaguRaj for their care and hospitality when I visited their places for attending conferences.*
- *I am thankful to all my friends- Sarita, Suma, Uma, Malati, Sirisha, Preeti, Udita, Rajeswari, Sweta, Usha, Shobha, Swati, Vanaja, Sudha, Arundhati, Aruna, ShanthiSri, Neeraja, Sridevi, Bhargavi and well wishers for their love and friendly nature, who left me memorable moments in my hostel life.*
- *I also thank Gnanesh, M. Kishore, RamSuresh for their help and caring towards me. I am thankful to all my classmates & Ph.D. colleagues for their support.*
- *I am indebted to my close friends - Deepthi, Mohan, Sudha, Anu, Kishore, Lakshman, Rajendra, YRK, Praveen, Dileep, Srinivas, Aparna and Bablu for their wonderful friendship, sharing and affection towards me. I also thank all my KC and KYSS friends who shared and cared for me. I am thankful to my cousins Vasu and Hari Prasad for their encouragement.*

- *I am always thankful to my friend AMV, who has been one of the driving forces, made his support available in a number of ways, for his understanding nature and affection.*
- *I am always grateful to my Dad, Mom and my dear Brother Swami - the backbone of my life, for their unending and unconditional love. Thanks shall be a small word before everything they have provided to me.*
- *I am thankful to my Grandparents, aunts and uncles and cousins for their love and affection. I always cherish the memorable moments spent with my cousins, Yaswanth, Mouni, Vindhya and little Mahi and also for their unconditional love.*
- *Finally I thank GOD, for blessing me with enough strength to face the challenges of life and for blessing me with wonderful family.*

Padmavathi Putta

Table of Contents

Abbreviations

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	messenger RNA
tRNA	transfer RNA
rRNA	ribosomal RNA
miRNA	micro RNA
PIC	Pre Initiation Complex
RNA Pol II	RNA Polymerase II
RISC	RNA Induced Silencing Complex
TFs	Transcription Factors
TFBS	Transcription Factor Binding Sites
TSS	Transcriptional Start Site
5'-UTR	5' Untranslated region
3' UTR	3' Untranslated region

Chapter 1

Introduction

1. Introduction

General Introduction

1.1 Significance of Sequence Analysis

Sequence analysis has become an important field in computational biology, as the sequence of DNA or protein itself carries a lot of information about the structural, functional and evolutionary features of biological sequences. In particular, we usually assume that similar sequences hold a similar function; structure and they may have a similar evolutionary history. The tools developed in the field of sequence analysis were aimed to determine the degree of similarity between two sequences.

Since the development of high throughput technologies in sequencing the DNA and protein data from 1990s, the rate of addition of new sequences to the databases is increasing enormously. Rapid deposition of such sequence information helps in understanding the biology of organisms. Sequence analysis provides access to genetic information such as finding common patterns among the sequences, identification of structural features in the genes, to find intron, exon and regulatory elements, finding mutations such as SNP (single nucleotide polymorphisms), to find evolutionary relationships and genetic diversity between different organisms. Another important application of sequence analysis is annotation, which involves computational finding to search for protein-coding genes, RNA genes and other functional elements and assigning their biological functions to the genome. In higher organisms, large parts of DNA do not code for meaningful information, called Junk DNA, but may contain unrecognized functional elements. Sequence analysis aid in unravelling the functional aspects and bridge the gap between genomics and proteomics in the right context.

1.2 Genome

Genome is defined as the entirety of an organism's hereditary information, which is encoded either in the form of DNA or RNA. The genome includes both the coding and non-coding sequences of DNA. The DNA is packaged into thread like structures called chromosomes. Each chromosome consists of linear array of genes, with each gene positioned in a particular location called gene locus. Gene is the hereditary unit of an organism. In modern biology, gene can be defined as a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions or functional sequence regions. A gene may exist in two or more forms termed as alleles that may be expressed as similar or different phenotypes of that particular trait.

1.3 DNA

In majority of living organisms the genome is encoded in long strands of DNA. DNA (Deoxyribonucleic acid) is a double stranded linear polymer of nucleotides that are running anti parallel to each other. Each nucleotide consists of a nucleoside linked with one or more phosphate molecules. Each nucleoside consists of a base linked to a 2'-deoxy ribose sugar. The backbone of the DNA strand is made from alternating phosphate and sugar residues. The sugar in DNA is 2-deoxyribose, which is a pentose sugar.

The four bases found in DNA are: adenine (A), guanine (G) termed as purines and thymine (T) and cytosine (C) referred as pyrimidines. In case of RNA, thymine is replaced by uracil (U). Each type of base on one strand forms a bond with just one type of base on the opposite strand. Purines always form hydrogen bonds with pyrimidines i.e., A bonds only with T and G bonds only with C. This arrangement of nucleotides binding together across the double helix is called as

complementary base pairing. As a result of this complementarity, all the information in double stranded sequence of a DNA helix is duplicated on each strand, which is vital in DNA replication. Indeed, this reversible and specific interaction between complementary base pairs is critical for all the functions of DNA in living organisms. The two strands are stabilized by hydrogen bonds between the bases on opposite strands. The two types of bases form different number of hydrogen bonds. A-T forms two hydrogen bonds and G-C forms three hydrogen bonds.

The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These asymmetric bonds mean a strand of DNA has a direction. In a double helix the direction of the nucleotides in one strand is opposite to their direction in the other strand i.e., the strands are *antiparallel*. The asymmetric ends of DNA strands are called the 5’ and 3’ ends, with the 5’ end having a terminal phosphate group and the 3' end a terminal hydroxyl group. This is important because the new DNA strands are always synthesized in 5’ to 3’ direction. The expression of genes encoded in DNA begins by transcribing the gene into RNA, a second type of nucleic acid that is similar to DNA, which contains ribose sugar and uracil (U) in place of thymine. RNA molecules are less stable than DNA and are typically single stranded molecules. Most biologically active RNAs contain self-complementary sequences that allow parts of the RNA to fold and pair with itself to form double helices.

1.4 Gene Expression

Gene expression is a process by which genetic information residing in a gene is used in synthesis of a functional gene product. The gene product might be a functional protein in case of coding genes or RNA in case of non-coding genes

like rRNA and tRNA. Gene expression (fig1.1) includes several steps like transcription (DNA to RNA), RNA splicing, translation (RNA to Protein), post-translational modifications.

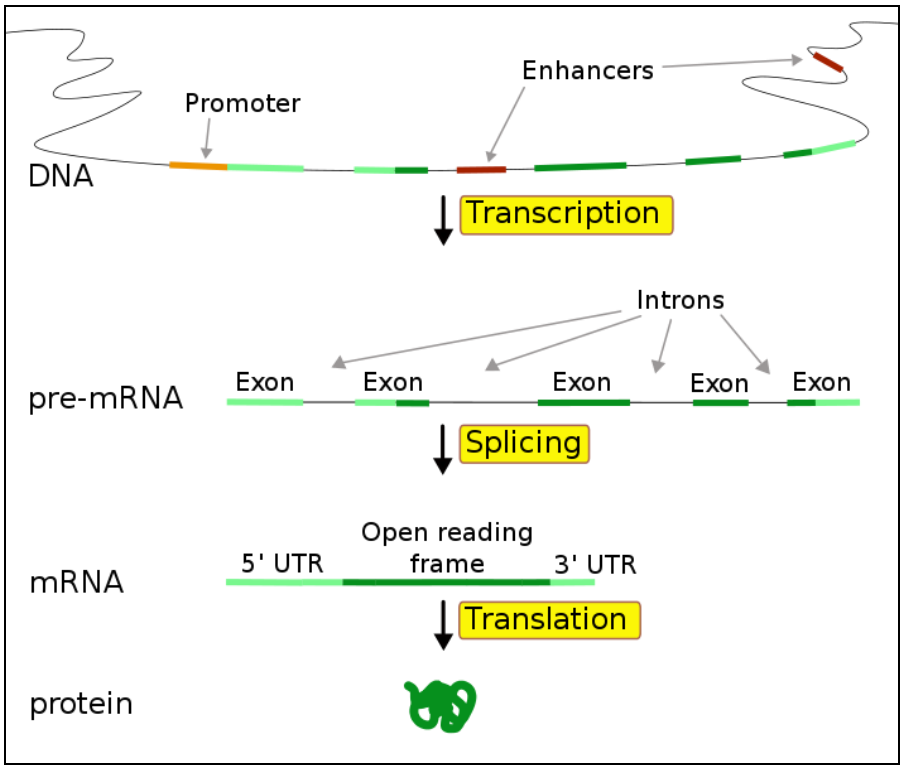


Fig 1.1: Diagram of typical eukaryotic protein-coding gene expression. Promoter and enchancers determine what portions of the DNA will be transcribed into the precursor mRNA (pre-mRNA). The pre-mRNA is then spliced into messenger RNA (mRNA), which is later translated into protein. Picture courtesy: http://en.wikipedia.org/wiki/Gene.

All living organisms and viruses use this process of gene expression to generate the macromolecular machinery of life. Any step in gene expression can be modulated or regulated. Gene regulation controls the over all structure and function of each cell and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism.

1.4.1 Transcription

Transcription is the mechanism by which a template strand of DNA is utilized by specific RNA polymerases to generate one of the three different classes of RNA. The three classes of RNA are:

- Messenger RNAs (mRNAs): This class of RNAs are the genetic coding templates used by the translational machinery to determine the order of amino acids incorporated into an elongating polypeptide in the process of translation.
- Transfer RNAs (tRNAs): This class of small RNAs form covalent attachments to individual amino acids and recognize the encoded sequences of the mRNAs to allow correct insertion of amino acids into the elongating polypeptide chain.
- Ribosomal RNAs (rRNAs): This class of RNAs are assembled together with numerous ribosomal proteins, to form the ribosomes. Ribosomes engage the mRNAs to form a catalytic domain into which the tRNAs enter with their attached amino acids. The proteins of the ribosomes catalyze all the functions of polypeptide synthesis.

RNA polymerase is the catalytic enzyme that transcribes the information residing in the template DNA. To date, four different RNA polymerases have been identified in higher eukaryotes and only one RNA polymerase in prokaryotes and archea. Eukaryotic cells contain three distinct RNA polymerases (I, II and III) that transcribe different classes of genes. RNA polymerase I resides in nucleolus and is primarily involved in transcribing 5.8S, 18S and 28S ribosomal RNAs. RNA polymerase II is located in nucleoplasm and transcribes genes that encode mRNAs and small nuclear RNAs. RNA pol III also located in nucleoplasm and transcribes genes that encode tRNAs, cellular 5S rRNAs and small RNAs involved in splicing and protein transport (snRNAs and scRNAs). New RNA

polymerase, RNA polymerase IV was identified in plants, facilitates the production of siRNA involved in RNA-directed DNA methylation, formation of heterochromatin and transcriptional silencing (Herr *et al.*, 2005, Kanno *et al.*, 2005; Onodera *et al.*, 2005). Another single polypeptide nuclear RNA polymerase IV (spRNAP-IV) was identified in human HeLa cells encoded by alternative transcript of mitochondrial RNA polymerase (Kravchenko *et al.*, 2005). This enzyme is able to transcribe a subset of mRNA-encoding genes that do not contain core promoter elements and regulatory sequences commonly found in RNA polymerase II transcribed genes.

All these polymerases have unique biochemical properties and are complex enzymes with 12-17 subunits. These enzymes do share a common property in transcribing a diverse set of DNA sequences. However, they lack sequence-specific recognition ability to correctly identify the transcription start site unique to each gene. For the site-specific initiation, additional proteins are required to form a stable initiation complex with RNA polymerase. These proteins are sequence-specific DNA-binding proteins called as transcription factors, are required for RNA polymerase stabilization. TFs may vary for each polymerase enzyme both in terms of specificity and order of binding. Here we briefly describe the transcription of protein coding genes by RNA pol II in eukaryotes.

In eukaryotes, the general transcription machinery includes:

- RNA pol II: a 12-subunit enzyme capable of synthesizing RNA and proof reading the transcript.
- General transcription factors: The transcription factors necessary for site-specific initiation by RNA polymerase II are: TFIIA, TFIIB, TFIID, TFIIE, TFIIE and TFIIH, collectively termed as general transcription factors (GTFs).
- Mediator: a 20-subunit complex, which transduces regulatory information

from activators and repressors for RNA pol II.

All these factors perform wide range of functions in stabilizing the pre-initiation complex (PIC), which includes the recognition of the promoter elements to stabilizing the RNA polymerase complex and elongation of transcription.

RNA polymerase II

RNA polymerase II is the key catalytic enzyme in the PIC and responsible for transcription of protein coding genes in eukaryotes. Human RNA pol II contains 12 subunits that are highly conserved among eukaryotes. RPB1 is the largest subunit of RNA pol II which contains a carboxy terminal domain (CTD) composed of up to 52 heptapeptide repeats that are essential for polymerase activity. Other proteins often bind the CTD of RNA polymerase to activate polymerase activity. This domain is involved in the initiation of DNA transcription, the capping of the RNA transcript and attachment to the spliceosome for RNA splicing. The core enzyme can be disassociated into a 10-subunit catalytic core and a heterodimer of RBP4/RBP7. The core enzyme is catalytically active but requires the heterodimer RBP4/7 complex and the general transcription factors for initiation from promoter DNA. During the transcription cycle, the CTD undergoes dynamic phospharylation of serine residues in the heptatetrad repeats. Transcription initiation requires an unphospharylated CTD, whereas elongation requires a phospharylated CTD. For recycling of RNA pol II and reinitiation of transcription, the CTD must again be dephospharylated.

Transcription Factors

Transcription factor is a protein that binds to specific DNA sequences (promoters and enhancers) there by controlling the transcription of genetic information from

DNA to mRNA. Transcription factors influence the rate of transcription of specific genes either positively or negatively by interactions with DNA regulatory elements and by interaction with other proteins. TFs are modular proteins consisting of a number of domains.

a. DNA-binding domain (DBD): that recognizes DNA specific sequences - the promoters or enhancers of the regulatory gene. Based on the three dimensional folding of this motif they have been classified into different families like: helix-turn-helix (HTH), Zinc finger (Zif), basic leucine zipper (bZIP) and basic heli-loop-helix (bHLH).
b. Trans-activation domain (TAD) that contains binding sites for other proteins such as activators or co-regulators of transcription.
c. Dimerization Domain: Majority of transcription factors bind DNA as homo dimers or hetero dimmers. Helix-loop-helix and leucine zipper motifs are the well-known dimerization domains.

In addition, TFs have nuclear localization sequence, and some may have a nuclear export sequence. Some TFs also have ligand-binding domains such as hormone-binding domains, which are essential for controlling their activity.

Assembly of GTFs and PIC formation

Transcriptional initiation is the first step in gene expression and generally constitutes the most important point of control by the sequence-specific binding of transcription factors. The GTF assembly is a stepwise addition of the transcription factors to the promoters and the order of assembly is as follows (fig 1.2).

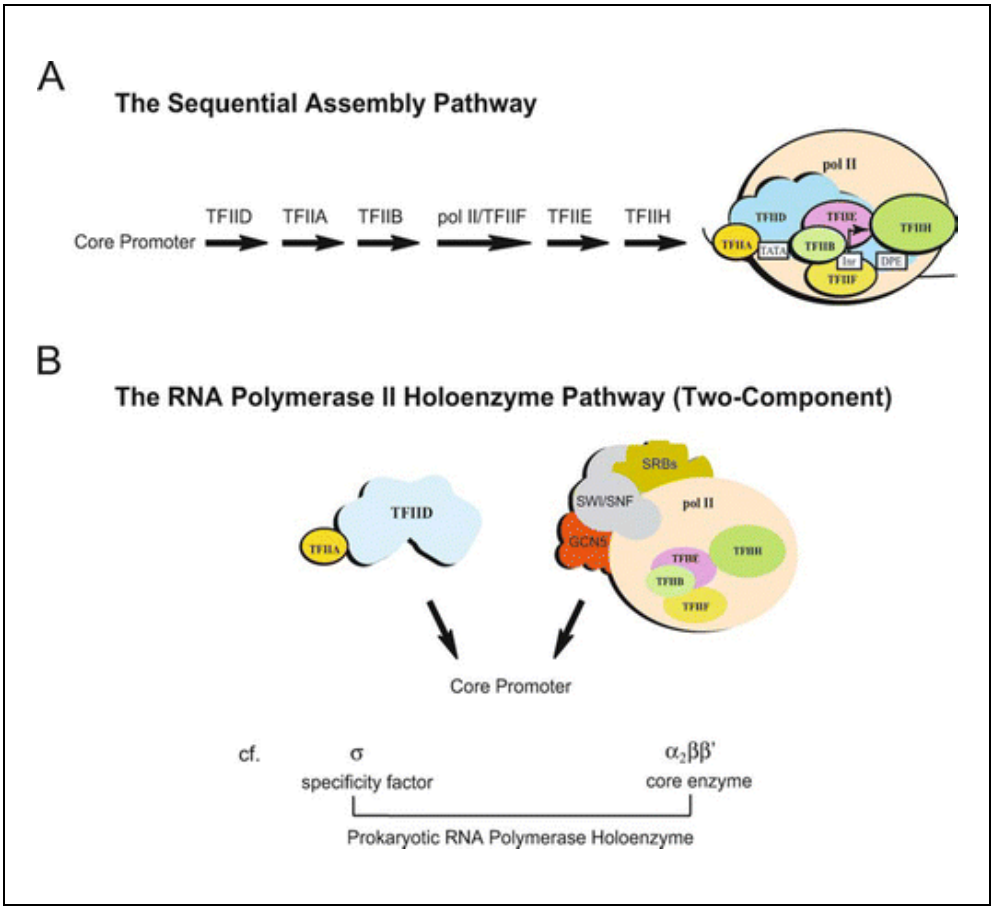


Fig 1.2: Pathways for preinitiation complex assembly. Panel A: represents the sequential assembly pathway. Preinitiation complex *(PIC)* formation may occur by stepwise recruitment of the general transcription machinery. Panel B: represents RNA pol II holoenzyme pathway (two-component) recruitment of preassembled pol II holoenzyme and TFIID complexes. The eukaryotic two-component pathway resembles the prokaryotic RNA polymerase holoenzyme system where a dissociable σ factor directs the entry of the bacterial RNA polymerase core enzyme $\alpha_2\beta\ \beta'$. Picture was adapted from Thomas and Chiang, 2006.

TFIID first binds to the promoter region, followed in a stepwise manner by the entry of TFIIA and TFIIB that help stabilize promoter bound TFIID and the recruitment of pol II and TFIIE. After the formation of stable TFIID-TFIIA-TFIIB-pol II/TFIIF-promoter complex, TFIIE is then recruited with the

subsequent entry of TFIIH. This pathway is known as the sequential assembly pathway (fig 1.2 A).

Another alternative for PIC formation is RNA Pol II holoenzyme pathway (fig 1.2 B). The human RNA pol II holoenzyme complex contains RNA pol II, TFIIB, TFIIF, TFIIH, GCN5 histone acetyl transferase, SWI/SNF chromatin remodelling factor and SRBs but is devoid of TFIID and TFIIA (Wu and Chiang 1998, Wu *et al.,* 1999). TFIID, the core promoter-binding factor facilitates the entry of pol II holoenzyme to the promoter region, which is analogous to the prokaryotic RNA polymerase system. It is likely that both assembly pathways exist in vivo and depending on specific signalling molecules involved and the promoter context, either pathway may be employed in responding to environmental cues. Indeed evidence supporting both models had been reported for different regulatory systems.

Elongation and Termination

Once the PIC complex is assembled at the promoter, TFIIH unwinds the DNA strands using ATP. RNA pol II advances in 3' to 5' direction with incorporation of nucleoside triphosphates (NTPs) to synthesize the RNA transcript in 5' to 3' direction. After the first nucleotide had been added, the RNA pol II has to clear the promoter, which may release truncated RNA transcripts, known as abortive transcription. During RNA elongation, TFIIF remains attached to RNA polymerase while other factors have to be relieved from the PIC complex (promoter clearance). Promoter clearance coincides with the phospharylation of c- terminal domain of RNA pol II by TFIIH. Phospharylation of CTD of RNA polymerase stabilizes the complex and elongation continues with help of elongation factors. Elongation also involves proof reading mechanism that can replace incorrectly incorporated bases. The growing RNA transcript forms

transient base-pairing with the template DNA. However, efficient transcription elongation must overcome several blocks such as transcriptional pause, arrest and termination that are intrinsic to RNA pol II catalytic activity and the chromatinized DNA template. Specific sequences in the template DNA signal the bound RNA polymerase to terminate transcription. At this site, RNA pol II releases the RNA transcript (pre-mRNA) and dissociates from the template DNA.

1.4.2 Post Transcriptional modifications

The pre-mRNA transcripts released after transcription undergoes various modifications and converted to mature mRNAs that occur prior to protein synthesis. The pre-mRNA molecule undergoes three main modifications that include:

i) 5’ capping
ii) 3’ polyadenylation
iii) RNA splicing

i) 5’ capping

Capping of the pre-mRNA involves the addition of 7-methylguanosine (m^7G) to the 5’ end. A phosphatase enzyme removes the terminal 5’ phosphate. The enzyme guanosyl transferase then catalyses the reaction which produces the diphosphate 5’ end. The diphosphate 5' prime end then attacks the α-phosphorus atom of a GTP molecule in order to add the guanine residue in a 5'5' triphosphate link. The enzyme (guanine-N^7-)-methyltransferase transfers a methyl group from S-adenosyl methionine to the guanine ring. The ribose of the adjacent nucleotide may also be methylated at the 2’ OH groups of the ribose sugar. The cap protects the 5’ end of the primary RNA transcript from attack by ribonucleases that have specificity to the 3’ 5’ phosphodiester bonds.

ii) 3’ polyadenylation

The pre-mRNA processing at the 3’ end of the RNA involves cleavage of its 3’ end and addition of about 200 adenine residues to form a poly(A) tail. The cleavage and adenylation reactions occur if a polyadenylation signal sequence (5’-AAUAAA-3’) is located near the 3’ end of the pre-mRNA molecule, which is followed by another sequence, which is usually (5’-CA-3’). The second signal is the site of cleavage. A GU-rich sequence is also usually present further downstream on the pre-mRNA molecule. After the synthesis of the sequence elements, two multi subunit proteins called cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CStF) are transferred from RNA polymerase II to the RNA molecule. The two factors bind to the sequence elements. A protein complex forms that contains additional cleavage factors and the enzyme Polyadenylate Polymerase (PAP). This complex cleaves the RNA between the polyadenylation sequence and the GU-rich sequence at the cleavage site marked by the (5’-CA-3’) sequences. Poly(A) polymerase then adds about 200 adenine units to the new 3' end of the RNA molecule using ATP as a precursor. As the poly(A) tail is synthesised, it binds multiple copies of poly(A) binding protein, which protects the 3’ end from ribonuclease digestion.

iii) RNA splicing

RNA splicing is the process by which introns, regions of RNA that do not code for protein, are removed from the pre-mRNA and the exons are connected to re-form a single continuous mRNA molecule. Although most RNA splicing occurs after the complete synthesis and end capping of the pre-mRNA, transcripts with many exons can be spliced co-transcriptionally. A large protein complex called the spliceosome, assembled from proteins catalyzes the splicing reaction and small

nuclear RNA molecules that recognize splice sites in the pre-mRNA sequence. Many pre-mRNAs, including those encoding antibodies, can be spliced in multiple ways to produce different mature mRNAs that encode different protein sequences. This process is known as alternative splicing, and allows production of a large variety of proteins from a limited amount of DNA.

1.4.3 Translation

For some RNA (non-coding RNA) the mature RNA is the final gene product. In case of coding genes, the messenger RNA (mRNA) produced by transcription is decoded by the ribosome to produce a specific aminoacid chain or a polypeptide that will be folded to form an active protein (fig 1.3). Each mRNA consists of three parts- 5' untranslated region (5' UTR), protein coding region or open reading frame (ORF) and 3' untranslated region (3' UTR). Coding region carries information for protein synthesis encoded by genetic code in the form of triplets. Each triplet of nucleotides of the coding region is called codon. The ribosome facilitates decoding of the codons by inducing the binding of tRNAs with complementary anticodon sequences to that of mRNA. The tRNAs carry specific aminoacids that are chained together into polypeptide as the ribosome passes through the mRNA.

In eukaryotes translation occurs in cytoplasm and across the membrane of endoplasmic reticulum. Translation proceeds in four phases: activation, initiation, elongation and termination. In activation, the aminoacid is covalently bonded to the 3' OH of the correct tRNA. Initiation involves the small subunit of the ribosome binding to the 5'end of mRNA with the help of initiation factors. Termination of the polypeptide occurs when A site of the ribosome faces a stop codon, which induces the binding of a release factor that prompts the disassembly of the entire ribosome/mRNA complex.

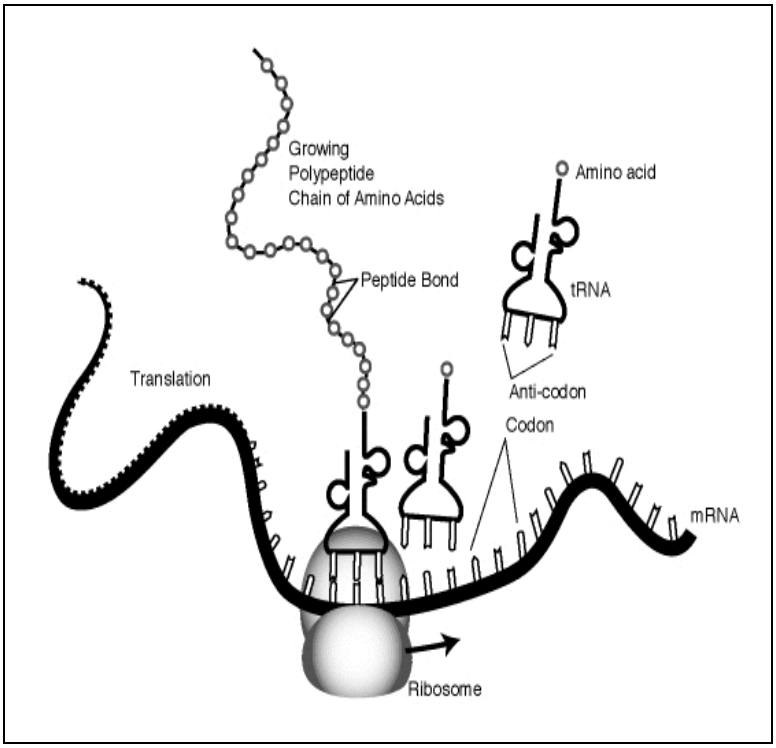


Fig 1.3: Diagram showing the translation of mRNA and synthesis of proteins by a ribosome with aid of tRNAs. The growing polypeptide will be released from the ribosomal complex, when ribosome faces the stop codons in mRNA. Picture courtesy: http://en.wikipedia.org/wiki/File: Translation-genetics.png.

1.4.4 Post-translational modifications

Post-translational modification (PTMs) is the chemical modification of a protein after its translation. Post-translational modification of aminoacids includes extending the functions of the protein by attaching biochemical functional groups (such as acetate, phosphate, various lipids and carbohydrates) or by changing the chemical nature of an aminoacid (for e.g. citrullination) or by making structural changes (for e.g. disulfide brigdes).

1.5 Gene Regulation

Gene regulation at transcriptional level is a complex task in eukaryotes compared to prokaryotes. In prokaryotes transcription and translation are coupled where as in eukaryotes DNA gene expression includes several steps. The activity of a gene might be regulated at numerous stages of its expression. In eukaryotes, gene regulatory mechanisms include:

i) Regulation of transcription in *cis*- mediated by promoters and enhancers.

ii) Regulation in *trans*-, mediated by transcription factors binding to *cis*-sites, RNA interference and miRNAs.

iii) Regulation on the basis of the modification state of the DNA and how it is packaged.

Further modification of gene activity occurs at the level of alternative splicing, RNA stability, translation efficiency, protein stability and protein activity. Of all the above mechanisms, promoter recognition by transcription factors is the initial step in the gene expression cascade and a complex event.

1.6 Work Introduction

The present work focuses on the analysis of promoter sequences that can be recognized by transcription factors during initial stages of the transcription, which is a crucial step in both gene expression and regulatory events. Here we discuss about importance of promoter recognition and the literature evidence from earlier studies on promoter recognition elements. miRNAs, that were known to play a major role in major biological events was also discussed.

1.6.1 Importance of Promoter Recognition

As introduced in earlier sections, it is clear that promoter recognition by transcription factors in one of the important and earliest features in gene expression cascade and also in regulation of gene expression. In computational biology, promoter identification and their functional evaluation is an important task that addresses key findings such as identifying the genes and helps in understanding the complex transcriptional mechanisms.

A promoter may be defined as the region upstream of, and containing the transcription start site (TSS), that is required for basic events of transcriptional initiation (Werner 1999). Promoters contain specific DNA sequences and response elements that provide a secure initial binding site for RNA polymerase and for transcription factors that recruit RNA polymerase. Promoters receive signals from various sources (for e.g., cell receptors) and control the level of transcription initiation that determines a gene expression to a great extent. Thus analysis of promoter region helps in elucidation of transcriptional activation mechanisms in gene expression, annotation of the transcriptional regulatory elements and development of efficient promoter prediction programs.

In prokaryotes, the core promoter elements are located in definite positions at -10 (pribnow box or TATA box) and -35 (-35 element TTGACA) positions upstream of the TSS. Eukaryotic promoters are extremely diverse and difficult to characterize, and it is generally believed that there are no universal core promoter elements as that of prokaryotic promoters. Eukaryotic promoters can be seen as miniature structures of coding regions with few functional elements (exons) interspersed in a larger sequence of unknown function (introns). The promoter exons would be resembled by transcriptional control elements usually-transcription factor binding sites, while the unknown spacers would correspond to

promoter introns. Another crucial obstacle in finding eukaryotic promoters is that they usually do not share extensive sequence similarity even when they are functionally correlated; suggesting that overall sequence similarity in promoters is not a general phenomenon (Werner, 2003). Thus promoter prediction has become an important task in computational biology. To date, in eukaryotes, the known promoter elements that can initiate transcription alone or either in combination with other elements are: TATA box, Inr (initiator), DPE (downstream promoter element), MTE (motif ten element), DCE (downstream core element), BRE^{u} (upstream TFIIB-recognition element) and BRE^{d} (downstream TFIIB recognition element) and CpG elements.

Analysis of known Promoter Elements

Earlier studies have targeted the transcription start sites (TSS) for promoter identification, but the TSS signal alone appears too weak. It has been reported that majority of human RNA pol II binding sites within the promoters have an array of closely located transcription start sites that are spread over 50-100 bp (Frith *et al*., 2008). Broad TSS distributions (dispersed TSS) are correlated with CpG islands and ubiquitously expressed genes, where as promoters with narrow TSS distribution frequently direct tissue-specific genes and often have TATA-boxes (Carninci *et al*., 2006). A number of promoter prediction tools have been developed using different set of characteristics such as TATA boxes (Ohler *et al*., 2002, Knudsen 1999), CpG islands (Davuluri *et al*., 2001, Bajic and Seah, 2003), CAAT boxes (Shahmuradov, 2003), pentamer matrices (Bajic *et al*., 2001), oligonucleotides (Scherf *et al*., 2000) and pattern recognition techniques such as neural networks, markov models, analysis of independent constituents, etc. However, functional analysis studies showed that there is no characteristic feature that describes the whole variety of promoters and each of features revealed during examination of promoter sequences have their own usage and limitations.

With increasing number of computational tools for promoter prediction, one can calculate the abundance of each core promoter element in genome-wide scale. Such analysis of human genes with putative transcription start sites indicated that $TATA^{+}Inr^{+}$, $TATA^{+}Inr^{-}$, $TATA^{-}Inr^{+}$, and $TATA^{-}Inr^{-}$ genes constitute 28%, 4%, 56%, and 12%, respectively (Suzuki *et al.*, 2001). This analysis suggests that 1/3 of human genes contain a functional TATA box, whereas the majority (~68%) of human promoters are in fact TATA-less. It can be seen that Inr alone seems to be able to direct initiation in TATA-less promoters. It can be seen that, in TATA-less promoters, Inr can direct the transcriptional initiation. A more comprehensive promoter analysis of promoters from EPD (eukaryotic promoter database) and DBTSS (database of human transcriptional start sites) found that less than 22% of the human genes contain a TATA box (Gershenzon and Ioshikhes, 2005). This study also showed that more than 78% of the human genes are TATA-less. Another study showed that only a small fraction of RNAP II promoters appears to contain a TATA box. In contrast, a large proportion of RNAP II promoters in metazoan genomes appear to contain an Inr element. Finally, about 25% of human promoters appear to lack known core promoter elements (Gross and Oelgeschläger, 2006). This point may lead to the existence of additional core promoter sequence elements that remain to be identified and functionally characterized.

From functional studies, it seems that the promoters of human housekeeping genes, oncogenes, growth factors, and transcription factors often lack a TATA box (Zhang, 1998; Zhou and Chiang, 2001). Analysis of promoters for known core promoter elements showed that 88% of active promoters were associated with CpG islands (Kim *et al*., 2005) and these promoters were not enriched with TATA box significantly. These studies may infer that TATA box is not conserved in human promoters or this may indicate that TATA box is not a

general promoter motif for human genes. CpG-island associated promoters are most often associated with housekeeping genes or ubiquitous genes, though there are many exceptions including tissue specific genes (Antequera and Bird, 1998).

It has been identified that differential utilization of alternative promoters that use a distinct combination of core promoter elements also plays a critical role in regulating gene expression in a spatial, temporal or lineage-specific manner. The dissection of transcription complex assembly pathways occurring on the core promoter region that serves as a converging point for regulatory events is of vital importance for our understanding of the transcription mechanism unique to each gene. Transcription control also has direct implications for human health, since improper regulation of the transcription of genes involved in cell growth is one of the major causes of all forms of cancers. Analysis of transcriptional regulation on a genomic level will remain the crucial factor for understanding life on molecular basis. In this direction, promoter recognition will be an important contribution towards understanding the transcriptional regulation and provides solid platform to understand the proteomics in the right context.

1.6.2 MicroRNA (miRNA)

Gene expression is largely regulated by the action of *trans*-factors on the *cis*-elements aligning on the regulatory regions of the genes. Among these *trans*-factors and the *cis*-elements, transcription factors (TFs) and their binding sites (TFBS) play important roles in regulation of gene expression. Recently, another group of molecules, namely microRNAs (miRNAs), have been found to regulate gene expression at the post-transcriptional (and translational) levels through base pairing with target messenger RNAs (Cui *et al*., 2007).

microRNAs (miRNAs) are single stranded RNA molecules that are of ~22

nucleotide in length and are known to negatively regulate gene expression at post-transcriptional level by binding to target messenger RNAs (mRNAs). The paradigm for the function and biogenesis of miRNAs has been provided by *lin-4* and *let-7*, which were originally identified during genetic screening of *C. elegans* post-embryonic development. Both *lin-4* and *let-7* are the post-transcriptional regulators of the target mRNA of *lin-14* gene (Lee *et al.*, 1993). Ever since the discovery of miRNA, many researchers have explored the role of miRNA in various biological functions. Here is a brief summary on biogenesis and maturation of miRNA and functional roles of miRNA in regulation of various biological processes.

miRNA Biogenesis and Maturation

Transcription of miRNA genes is mediated by RNA pol II enzyme. A general model for miRNA biogenesis and maturation is depicted in fig 1.4. Transcription of miRNA genes yields long primary transcripts (pri-miRNAs) that are usually several kilobases long that fold back to form hairpin structures. The nascent pri-miRNAs are processed in the nucleus by Drosha, RNase III enzyme to release pre-miRNA, the precursors of miRNA that are of ~70 nucleotides in length. Drosha is a large protein that contains two tandem RNase III domains (RIIIDs) and a double-stranded RNA-binding domain (dsRBD) that are crucial for the catalysis. The central region of the protein is essential for pri-miRNA processing. Drosha interacts with its cofactor, the DiGeorge syndrome critical region 8 (DGCR8) protein for the pri-miRNA processing. Following the nuclear processing by Drosha, pre-miRNAs are exported to the cytoplasm mediated by exportin-5, a Ran-GTP dependent nucleo/cytoplasmic cargo transporter. Following their export to cytoplasm, pre-miRNAs are subsequently processed into ~22 nucleotide miRNA duplexes by the cytoplasmic RNase III Dicer. Dicer contains a putative helicase domain, a DUF283 domain, a PAZ (Piwi-Argonaute-

Zwille) domain, two tandem RNase III domain and a ds RNA-binding domain (dsRBD). PAZ domain recognizes the staggered ends of pre-miRNAs and mediates the cleavage.

As dicer processes the pre-miRNA into the miRNA: miRNA* (mature miRNA strand and its complementary strand) duplex, the stability of the 5' ends of the two arms of the duplex is usually different. Mature miRNA is almost always derived from the strand with less stable 5'end which interacts with the RISC (RNA-Induced Silencing Complex) components for silencing the target mRNA as that of RNAi interference. On the other hand, miRNA* strand (complementary strand of mature miRNA) is probably degraded rapidly. Dicer is associated with several other proteins that play various roles in miRNA stability and effector complex formation and activation. Dicer contains multiple homologues and Dicer2 forms a complex with R2D2 (a dsRNA-binding protein), which enhances sequence-specific mRNA degradation that is mediated by RISC.

The mature miRNA can undergo unwinding and a single strand can enter the RISC complex, where they can act by repressing the translation of their mRNA targets or by inducing their degradation, mediating RNA interference. In mammals, mRNA cleavage is thought to occur by perfect or near perfect interactions of miRNA with their target mRNAs.

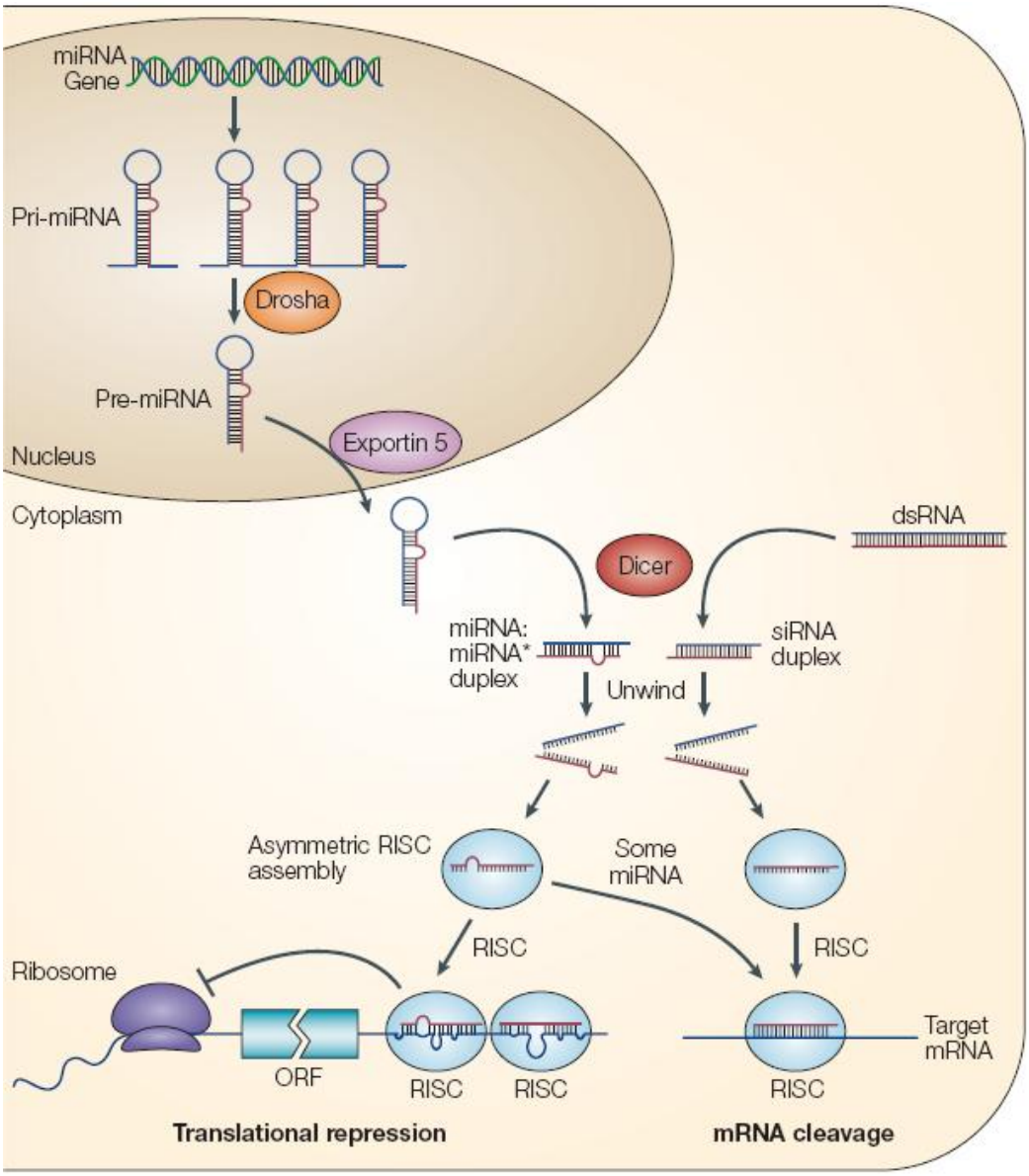


Fig 1.4: Biogenesis of miRNA and post-transcriptional regulation of mRNA by miRNA. The nascent pri-microRNA (pri-miRNA) transcripts are first processed into ~70-nucleotide pre-miRNAs by Drosha inside the nucleus. Pre-miRNAs are transported to the cytoplasm by Exportin 5 and are processed into miRNA:miRNA* duplexes by Dicer. Dicer also processes long dsRNA molecules into small interfering RNA (siRNA) duplexes. Only one strand of the miRNA:miRNA* duplex or the siRNA duplex is preferentially assembled into the RNA-induced silencing complex (RISC), which subsequently acts on its target by

translational repression or mRNA cleavage, depending, at least in part, on the level of complementarity between the small RNA and its target. Open reading frame (ORF). Picture was adapted from He L., and Hannon G. 2004.

miRNA role in Biological Processes

Since the discovery of miRNA in *C. elegans*, role of miRNA in various biological processes has been explored extensively. Computational studies indicate that approximately one third of human genes are potentially regulated by miRNAs and each miRNA on average could target more than 200 genes (Cui *et al*., 2007).

Interaction of miRNA with target mRNAs reported that 5'-UTR motifs interact with the 3'-end of miRNA in all conservation categories. On the other hand, 3'-UTRs show interactions with miRNA only in the case of highly conserved 8-mers (Lee *et al*., 2009). These studies suggest that 3'-end of miRNA family members (interspecies) and those of some miRNAs across species differ, the 3'-end of miRNAs may contribute to the gene- or species specific target site recognition of the 5'-UTR. There are increasing number of evidences suggesting miRNAs play critical roles in many key biological processes, such as cell growth, tissue differentiation, cell proliferation, embryonic development and apoptosis. It has also been reported that miRNA play important roles in cellular signalling networks, cross-species gene expression variation, and coregulation with transcription factors (Ming *et al*., 2008). Studies also showed that miRNAs are aberrantly expressed in different cancers, suggesting their role as a novel class of oncogenes or tumor suppressor genes (Kusenda *et al*., 2006). Many studies have reported that each cancer tissue has a specific miRNA signature and miRNA based classification of cancers can be effective and potential tools (Lu *et al*., 2005). With increasing evidences for diverse roles of miRNAs in many biological functions, we are interested in the role of miRNA in recognition of 6-nt sequences

identified in the present study.

1.7 Objectives

It is generally believed that there are no universal core promoter elements in eukaryotes as that of prokaryotic promoters located at –10 and –35 positions from TSS. Several studies indicate that TATA box is not a conserved promoter element in eukaryotes and promoters with TATA box account for 10-20% of protein-coding genes. From recent studies, miRNAs are known to regulate about 30% of the genes. In this context, we may perhaps suggest that ~50% of the remaining genes can be directly regulated by transcription factors, which recognizes the promoter elements and initiates the gene expression cascade. Transcriptional control has direct implications for human health, since improper regulation of the transcription of genes involved in cell growth is one of the major causes in all forms of cancer.

It was identified that the DNA/promoter elements that can be recognized by protein complexes may vary from 5-11 bp in length (Fickett and Hatzigeorgiou 1997). Longer sequences may have some degeneracy or redundancy or may be just error-prone (Yamamoto *et al*., 2007). We choose to study subsequences of length 6-nt because: i) most common promoter elements in prokaryotes (at –10 and –35 from TSS) are 6-nt in length, ii) most of the restriction enzymes recognize DNA with high accuracy are of 6-nt in length, iii) minimum length of transcription factor binding sites sequences in JASPAR database are above 5-nt nucleotide in length. We implicitly assume that a 6-nt sequence can be recognized by any standard protein motifs without any error.

The present work focuses on analysis of human promoter regions, which might provide some insights in understanding the promoter architecture (in sequence

terms) and their role in gene expression or regulation. I am interested in identifying relatively conserved subsequences in promoter sequences around the transcription start site (TSS) that are 6-7 nt in length. These subsequences were studied for their frequency and positional distributions in promoters, miRNA sequences and transcription factor binding site sequences and to find any correlation among these datasets based on the common subsequences, that can help in understanding the complexity in eukaryotic transcription. In this context, my research work has been presented in the following chapters.

- Chapter 2 describes the materials and methods that were used for the study. Human promoter sequences have been studied for identification of relatively conserved subsequences that are of 6-7 nucleotides in length. The datasets include promoter sequences, miRNA and TFBS sequences of *Homo sapiens*.
- Chapter 3 includes analysis of the results followed by discussion. The results include:
    - Base composition studies in promoter sequences and TFBS sequences.
    - Frequency analysis, distribution and positional distribution of subsequences (6-nt sequences) in three datasets separately.
        - Promoter sequences (Homo sapiens)
        - miRNA sequences (Homo sapiens)
        - Transcription factor binding site sequences (TFBS) (*Homo sapiens*)
- Chapter 4 summarizes the conclusions of my work.

Chapter 2

Materials & Methods

2. Materials & Methods

2.1 Materials

2.1.1 Promoter Sequences (*Homo sapiens*) collected from EPD

Eukaryotic Promoter Database (EPD)

EPD is a non-redundant collection of experimentally characterized eukaryotic RNA POL II promoters, which are experimentally defined by the transcription start sites (TSS). The promoter sequences can be extracted from the pointers to positions in nucleotide sequence entries. EPD provides information about eukaryotic promoters available in the EMBL Data library and facilitates dynamic extraction of biologically meaningful promoter sequences for comparative analysis studies. In the database, the promoter sequences in reference to TSS were collected from the conventional TSS mapping experiments for individual genes and mass genome annotation projects (Cavin Périer *et al*., 1998 and Schmid *et al*., 2006). The database is currently available at http://epd.vital-it.ch/.

For *Homo sapiens*, 1871 promoter sequences were reported from EPD (release 96) in all promoters category. The sequences were extracted with in -100 to +100 (a total of 201 bp in length) relative to transcription start site (TSS) in FASTA format. Another set of promoter sequences was downloaded as only upstream (-100 to +1 wrt TSS) and downstream promoter sequences (-1 to +100 wrt TSS) separately.

1871 human promoter sequences from EPD (release 96) were used for identification of common subsequences (6-7nt). Later these promoter sequences

were looked for studying the frequency and positional distributions of subsequences (6-7nt).

2.1.2 miRNA Sequences (*Homo sapiens*) collected from miRBase

miRBase

The miRBase is a searchable database of published miRNA sequences and annotation (Griffiths-Jones *et al*., 2008). Each entry in the miRBase sequence database represents a hairpin portion of a miRNA and its mature miRNA sequence. Both hairpin and mature sequences are available for download. We downloaded 866 and 695 human mature and stem-loop miRNA sequences from release 12. The database is increasing rapidly and the current release 16 contains 1223 and 1045 mature and stem-loop miRNA sequences respectively. miRBase is available at http://www.mirbase.org/.

Human mature and stem-loop miRNA sequences from miRBase (release 16) were used to study the frequency distributions of 6-nt sequences that were identified in human promoter sequences.

2.1.3 TFBS Sequences (*Homo sapiens*) collected from JASPAR

a. JASPAR

JASPAR is an open-access database of annotated high-quality, matrix-based transcription factor binding site profiles for multicellular eukaryotes (Sandelin *et al*., 2004). The profiles derived in this database were exclusively from sets of nucleotide sequences that were experimentally demonstrated to bind transcription factors either from SELEX experiments, experimentally determined binding

regions of actual regulatory regions and high-throughput technologies like ChIP-seq experiments. JASPAR is a web interface for browsing, searching and subset collection, online sequence analysis and tools for genome wide and comparative genomic analysis of regulatory regions. JASPAR database is available at http://jaspar.genereg.net/.

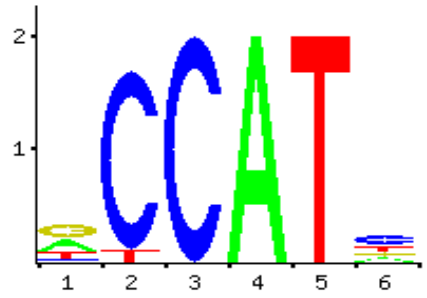
Two sets of transcription factor binding sites of 75 human transcription factors from JASPAR database were used to study the frequency distributions of subsequences.

b. TFBS collection

The TFBS sequences were collected from the frequency matrices given for each transcription factor in JASPAR database. For Homo sapiens, 75 transcription factors were reported in the database. The following figures illustrate the collection of tfbs sequences from JASPAR.

As an example, transcription factor YY1 (ID: MA0095.1) has been described here as reported from the database. The sequence logo and frequency matrix for transcription factor binding site of YY1 were shown in panel A and panel B (figure 2.1). Panel A depicts the sequence logo of transcription factor binding site for YY1 available from JASPAR database. From the frequency matrix (Fig 2.1B), each TFBS sequence was represented based on the highest frequency of nucleotide occurring at each position in the matrix (numbers highlighted in color). From the matrix, TFBS sequence for YY1 factor can be represented as [G/A]CCATC. The frequency matrix was constructed based on 17 experimentally verified tfbs sequences (total of each column in the matrix is 17) that were collected from literature and experimental evidences. All the 17 different TFBS sequences were considered as another set of TFBS sequences for YY1 factor (fig 2.2).

A. Sequence logo for YY1



B. Frequency matrix

A	**[**	**6**	**0**	**0**	**17**	**0**	**3**	**]**
C	**[**	**1**	**16**	**17**	**0**	**0**	**8**	**]**
G	**[**	**7**	**0**	**0**	**0**	**0**	**3**	**]**
T	**[**	**3**	**1**	**0**	**0**	**17**	**3**	**]**

Fig 2.1: Sequence logo and frequency matrix for YY1 (ID. No: MA0095.1) from JASPAR database. Panel A represents the sequence logo for binding site of YY1. Panel B represents the frequency matrix of the binding site sequence for YY1. From the matrix, the binding site sequence for YY1 can be represented as G/ACCATC.

>MA0095 YY1 1
ccgctGCCATCatggttcga
>MA0095 YY1 2
tggcgGCCATCcgccgcctg
>MA0095 YY1 3
tggctGCCATCtgttttacg
>MA0095 YY1 4
ccccgGCCATCttggcggct
>MA0095 YY1 5
gatcgGCCATCttgactcaa
.....
.....
.....
>MA0095 YY1 16
gcgcCCCATAcaaggcg
>MA0095 YY1 17
ggatATCATTttggaag

Fig 2.2. Experimentally verified binding site sequences that were used to construct the frequency matrix of YY1 from JASPAR.

Fig 2.2 depicts list of experimentally verified binding site information available for YY1 in JASPAR database. The 17 sequences were used for generation of frequency matrix and sequence logo of binding site sequence for YY1shown in fig 2.1. Only few of the binding site sequences were represented in figure 2.2. The text highlighted in color was extracted from each sequence and studied as another set of transcription factor binding site sequences for YY1. In JASPAR database, out of 75 TFs, binding site information for 65 TFs is available.

In brief, we have collected the TFBS sequences in two methods.

i) TFBS sequence that represents the frequency matrix for each TF (fig 2.1). A total of 75 TFBS sequences for 75 transcription factors (human) were constructed from JASPAR. These binding site sequences were mentioned as 75 TFBS in the present study.

ii) Another set is: experimentally verified tfbs sequences that were used to construct each of the frequency matrices for each transcription factor in JASPAR (fig 2.2). Out of 75, transcription factors binding site sequence information for 65 transcription factors are available in the database. These 6497 TFBS sequences represent the binding site sequences of 65 transcription factors. This set of binding sites was mentioned as 6497 TFBS sequences in the present study.

2.2 Methods

2.2.1 Identification of Subsequences in Promoter sequences

The 1871 human promoter sequences collected from EPD database were searched pair wise for locating relatively conserved subsequences that are of 6-7 nt in length. We considered all n.(n-1)/2 pairs for this search, where n is number of promoter sequences and n.(n-1)/2 is number of possible pairs. In this study all possible 1871*1890/2 = 1, 749, 385 pairs have been considered. We have modified local alignment algorithm (Smith and Waterman, 1981) to identify relatively conserved subsequences (6-7nts) in the promoter sequences. The algorithm (source code in C compiled in gcc environment) displays the relatively conserved subsequences that were identified in the above pairs. Next, the subsequences were sorted in lexical order in order to arrange all similar

subsequences adjacent to each other. These subsequences were counted for their frequencies. This gives count of all the subsequences (i.e., redundant subsequences) that were identified in promoter sequences (table 3.1). The redundant subsequences were sorted (using sort U function) to get unique subsequences. These distinct subsequences were counted for their frequencies. This set is considered as non-redundant subsequences (table 3.1). Both the redundant and non-redundant subsequences were sorted in their decreasing frequencies and the top 50 most common subsequences were considered for further study (table 3.1). The same procedure is followed for locating the most common subsequences in upstream and downstream promoter sets separately (table 3.2).

Frequency Plots

The top 50 most common subsequences (table 3.2 and 3.3) were searched for their frequency and relative positional distributions in the promoter sequences, miRNA and TFBS sequences using string search function (written in C). The results were analyzed and plotted using SysStat SigmaPlot 9.0.

2.2.2 Identification of Subsequences in TFBS sequences

The 6497 TFBS sequences collected from JASPAR was aligned using local alignment algorithm to identify relatively conserved subsequences that are 4-6 nt in length. We considered all n*(n-1)/2 possible pairs i.e., 6497*6496/2 = 21,102,256 possible pairs for locating the common subsequences. The pair wise search alignment algorithm (source code in C, compiled in gcc environment) displays the subsequences that are common in a pair of TFBS sequences. The subsequences (4-6nt) in the output were sorted in lexical order to arrange all the similar subsequences adjacent to each other and their frequencies were calculated.

These initial subsequences were considered as redundant subsequences. Next the redundant subsequences were sorted to remove the duplicate subsequences in the same pair of TFBS sequences. These distinct subsequences were counted and represented as non-redundant subsequences with non-redundant frequencies. Both the redundant and non-redundant sequences were sorted based on their highest frequencies. Among them, the top 50 most common non-redundant subsequences (4-6nt) were collected for further study (table 3.7).

Frequency Plots

The most common 50 subsequences (table 3.7) were searched in TFBS sequences for their frequency distributions using string search program (source code in C). The results were analyzed and plotted using SysStat SigmaPlot 9.0.

Chapter 3

Results

&

Discussion

3. Results and Discussion

3.1 Frequency Distributions of Subsequences in Promoters

3.1.1 Base Composition in Promoter Sequences

Initially we looked for the base composition of 1871 human promoter sequences within -100 to +100 region relative to TSS (transcription start site) from EPD database. Each promoter sequence contains 201 bp, and a total of 1871*201 = 376,071 nucleotides were present in the selected promoter dataset. Out of them, 605 were represented as 'N', any base. Following are the base compositions observed in the promoter data:

A = 64,984 (17.3%)

T = 70,208 (18.7%)

C = 114,744 (30.56%)

G = 125,530 (33.43%)

Mononucleotide composition suggests that promoter sequences were GC rich in composition. However, distributions of individual nucleotides appear to be non-uniform. We looked for dinucleotide distributions and the results show some interesting features (fig 3.1). There are 1871*(201-1) = 374,200 possible dinucleotides in the present data. We looked for unique dinucleotide combinations (4*4=16). Assuming all are equal in distribution, we expect 374,200/(16*200) = 117 as mean frequency for any dinucleotide at any position within the dataset. Distributions of AT, TA, GC and CG appear more prominent compared to other dincuelotides (fig 3.1). Other 12 dinucleotides were omitted for visual clarity. The very sharp peak observed at position '0' is due to presence of TSS codon.

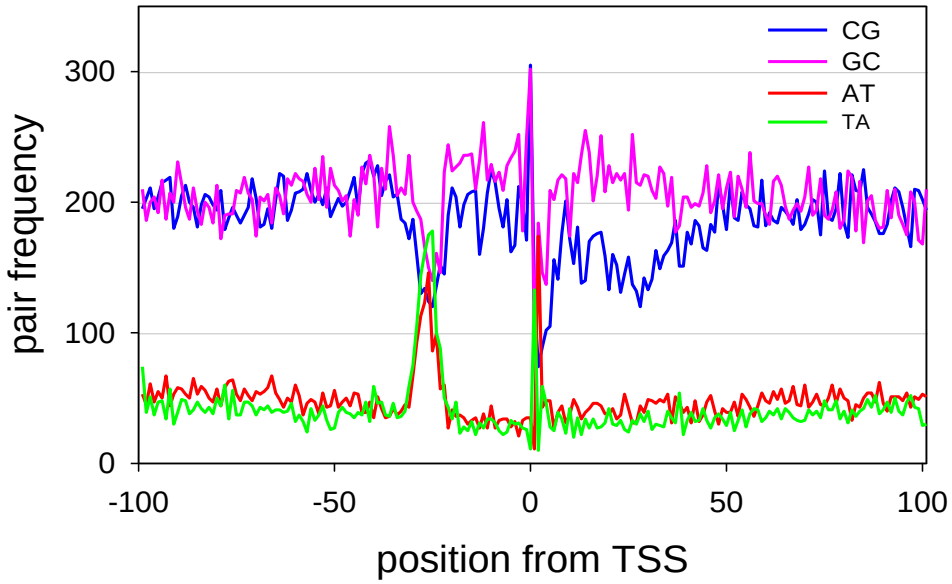


Fig 3.1: Dinucleotide distribution in promoter sequences position wise from -100 to +100 relative to TSS. The lower graphs represent AT/TA distributions and the top graphs represent GC/CG distributions. The sharp peak at position '0' is due to presence of the TSS signal. We suggest that the distributions between -50 and +50 appear to be anomalous in nature.

We can clearly observe that AT and TA frequencies appear below mean frequency values except with a sharp peak at -25 from the TSS. From literature, it was reported that promoters that contain TATA box might account for 10-20% of the protein coding genes in human genome (Frith *et al.*, 2008, Tokusumi *et al.*, 2007 and Bajic *et al.*, 2006). The sharp peak around -25 in TA and AT frequencies may represent the presence of TATA box/AT rich sequences in this region. We noted that CG and GC frequencies appear above the mean values except with a dip around -25 and a broad hump around +25 downstream to the TSS. Sudden dip in CG and GC distributions around -25 may correspond to high frequencies of AT/TA distributions in that region. It should be noted that GC and CG frequencies were not below the mean dinucleotide frequency value, but with

decrease in frequencies around -25 from TSS. The broad hump in downstream of TSS in GC and CG distributions suggests that these signals (promoter elements) are also present in downstream of TSS.

Mononucleotide and dinucleotide distributions within the promoter data (-100 to +100 relative to TSS) suggest that promoter sequences are GC rich in composition in upstream and downstream region from TSS.

3.1.2 Distribution of 6-nt Sequences in Upstream and Downstream of TSS

Using local pair wise alignment, the most common subsequences (6-nt) were identified in upstream and downstream promoter sets separately (Methods: 2.2.1). The distinct subsequences were sorted based on their highest frequencies. Among them, top 50 most common non-redundant 6-nt sequences were selected from both upstream and downstream promoter sets separately. We note that these 6-nt sequences are GC rich in composition. The 6-nt sequences were searched for their frequency distributions in upstream and downstream sequences separately (table 3.2). Frequency of each 6-nt sequence in a promoter set was calculated based on two strategies:

i) The non-redundant frequency represents the single or one time occurrence i.e., multiple occurrence of same 6-nt sequence in a given promoter is considered as one.

ii) Redundant frequency represent sum of total occurrence of each 6-nt sequence in a given promoter sequence (upstream and downstream separately).

The frequencies were shown in descending order in fig 3.2. We note that 6-nt sequences in upstream region are more common but only slightly. The shape of the distribution appears interesting as it reminds one of the typical sigmoidal curves (rotated).

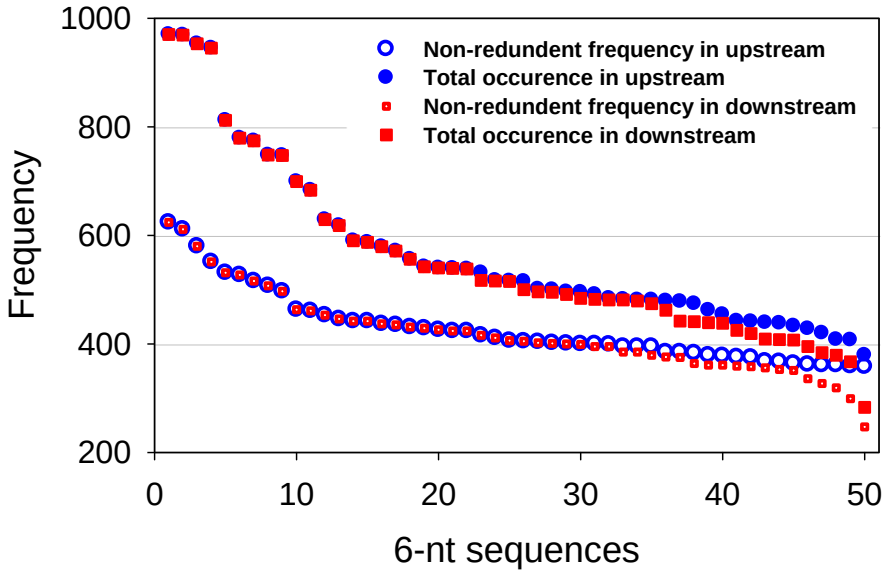


Fig. 3.2: Frequencies of 50 most common 6-nt sequences plotted after arranging their frequencies in descending order. The 6-nt sequences for the upstream and the downstream regions are not necessarily identical. The graph shows typical sigmoidal shape suggesting some internal cooperative nature in 6-nt distribution. Total occurrence corresponds to the frequency of redundant 6-nt sequences.

It perhaps has origin in some internal cooperative process. It is well known that several proteins are needed in many cells and are very common (for example house-keeping genes are expressed regularly in all tissues) and many proteins that are highly specialized are less common (i.e. spacio-temporal expression of tissue specific genes). Therefore we expect some proteins to be synthesized at a fast rate whereas some others may be produced at a relatively slow rate. Fast rate of synthesis can only be achieved using a large number of transcription factors and corresponding binding sites. It should be noted that large copy numbers of genes do not automatically translate to a larger rate of protein synthesis as the transcription factors and TFBS may become rate-limiting factors in the transcription process. The curves shown in figure 3.2 also support this

hypothesis. We suggest that possibly more important genes are having highest frequencies of the corresponding 6-nt sequences.

It was noted that frequencies of 6-nt sequences in both upstream and downstream promoter regions around the TSS appear similar in distribution (fig 3.2). This suggests that transcription factors have to straddle on both sides of TSS, as an essential requirement during initial stages of transcription. X-ray studies also confirmed that transcription factors straddle on both sides of DNA (Muller and Herrman, 1997). However, the transcription factors have to be removed from downstream of TSS for continuation of the transcription process. Mutations in downstream of TSS may preserve the codon table of protein synthesis, but they can severely affect the binding specificity of the transcription factors.

We have calculated total occurrence of each 6-nt sequence in upstream promoter region (-100 to +1 relative to TSS). The frequency of occurrence and their percentile frequencies of each 6-nt sequence in decreasing frequencies (of top 50 6-nt sequences) were shown in fig 3.3. Figure 3.3a and 3.3b represents the non-redundant and redundant frequencies and their percentile frequencies respectively. The nature of distribution resembles the earlier distribution curves in figure 3.2. This clearly suggests the existence of co-operative behavior with in the 6-nt sequences. However, origin of this co-operative nature is unknown at present. We can presume that possibly more important genes will be associated with a higher frequency of the corresponding 6-nt sequences.

Table 3.1: List of top 50 most common redundant and non-redundant 6-nt sequences that were identified in promoter sequences (-100 to +100 relative to TSS) arranged in their decreasing frequencies. NR*: non-redundant 6-nt sequences and R*: Redundant 6-nt sequences.

S.No	6-nt seq (NR*)	#	6-nt seq (R*)	#	S.No	6-nt seq (NR*)	#	6-nt seq (R*)	#
1	GGCGGG	624	GGGCGG	16382	26	GCGGGC	408	GGCCGG	5000
2	GGGCGG	612	GGCGGC	14898	27	GGCGCG	407	GCGGAG	4912
3	GGCGGC	583	GGCGGG	14776	28	GCCCGG	406	GCCCGG	4904
4	GCGGCG	555	CCGCCC	11368	29	GCGCCG	403	CCCGGC	4840
5	GCGGGG	531	GCGGGG	9598	30	GCCCCG	402	CGCCGC	4808
6	CCGCCC	531	GCGGCG	8882	31	CCGGGC	400	CCGGGC	4784
7	CCCGCC	519	CCCGCC	8568	32	CCGCGC	398	GCGCGG	4760
8	GGGGCG	517	GCCGCC	7734	33	GGCTGC	397	GGAGCC	4618
9	CGGCGG	495	GGGGCG	7638	34	CCCGGC	397	CGGCGC	4598
10	GCGGCC	466	CGGCGG	7534	35	GCGCGC	387	GCGGGC	4582
11	CGGGGC	459	GCGGCC	6922	36	CGCGCC	384	CCGCGG	4270
12	GCGCGG	451	CGGGGC	6614	37	GGGCCG	382	CGGAAG	4230
13	CGGCGC	444	GGTGAG	6552	38	GGGGCC	380	CCCGGG	4224
14	GGCCGG	442	GGGAGG	6534	39	GGCTGG	378	GGCGCC	4222
15	GCCGCC	442	GCCCCG	6188	40	GGAGGG	376	TGGCGG	4192
16	CCCCGC	441	CGCCCC	6012	41	GGAGGC	374	GCCGCG	4122
17	GCCGGG	437	GCCGGG	5516	42	CGGCCG	371	CCGCCG	4118
18	CGCCCC	436	GGGGCC	5322	43	GCCGGC	368	CGCGGC	4088
19	CGCCGC	433	GGAGGC	5302	44	CCGGCG	367	CCTCCC	4074
20	CGCGGC	432	GGGGGC	5206	45	AGGCGG	366	GGCCCC	4044
21	GGCCGC	425	GCTGGG	5190	46	GGTGAG	361	CGCGCC	3930
22	GGGAGG	422	GGCCGC	5168	47	CCGCGG	361	GGAAGT	3896
23	GCTGGG	411	GGCTGC	5128	48	CCGCCG	360	GCCCCC	3894
24	GCGGAG	409	CCCCGC	5128	49	GGAGCC	359	GGGCCG	3848
25	GCCGCG	409	GGAGGG	5118	50	CCCGGG	358	GCGCGC	3840

represents frequency of corresponding 6-nt sequence in the promoter sequences (-100 to +100 relative to TSS).

Table 3.2: Frequencies of 6-nt sequences that were identified in upstream (-100 to +1 relative to TSS) and downstream (-1 to +100 relative to TSS) promoter sequences separately after pair wise search in upstream and downstream promoter sequences separately.

S.No	Upstream	R	NR	Downstream	R	NR	S.No	Upstream	R	NR	Downstream	R	NR
1	GGGCGG	970	624	GGCGGC	953	611	26	CGGAAG	515	402	GGCAGC	540	442
2	GGCGGG	953	611	GCGGCG	969	580	27	GCCGGG	531	404	CCCGGC	495	411
3	GGGGCG	969	580	GGTGAG	970	624	28	CGCGGC	484	405	TGCTGC	484	405
4	GCGGGG	945	551	CGGCGG	774	531	29	CCGCGC	478	395	GCGCGG	491	395
5	CCGCCC	779	527	GCCGCC	779	527	30	GGCCGC	539	401	GGCCGG	683	431
6	CCCGCC	774	531	GCGGCC	945	551	31	CCTCCC	482	400	GCCGCG	482	400
7	CGGGGC	748	507	TGGCGG	748	507	32	GGGGCC	501	400	GGAGCC	590	453
8	CGCCCC	747	516	CGCCGC	699	442	33	GCCCGG	462	399	GCAGCC	407	361
9	CCCCGC	812	497	GGCTGC	747	516	34	CGCGCG	491	395	GCGGCT	538	446
10	GCGGCG	579	463	CCGCCG	812	497	35	GCCGCG	571	395	CCGCGG	515	402
11	GGCGGC	629	461	GCTGGG	579	463	36	GGCGCC	481	385	CGGCTG	442	359
12	GGGAGG	590	453	ATGGCG	629	461	37	CCCGGC	474	385	CCGGGC	367	299
13	GCCCCG	538	446	GCTGCC	379	361	38	GGCGGA	441	379	GTGAGT	408	358
14	CGGCGG	540	442	CGCTGC	542	424	39	GCGGCC	438	383	CGGAGC	409	356
15	GGCGCG	699	442	GCTGCT	539	401	40	GAGGGG	481	375	CCCGCC	419	353
16	GCGCGG	618	435	GCCGGG	587	426	41	GCGGAG	439	376	GGGCCG	496	406
17	CGGCGC	516	437	GCTGCG	516	437	42	GGGGAG	454	378	GCCCGC	556	364
18	GCGCGC	587	426	GCGGAG	441	379	43	GGAGGC	428	368	GCAGCT	517	429
19	GGCCGG	683	431	CGCGGC	439	376	44	CCGGGC	420	362	CTCCTG	425	327
20	GGGGGC	517	429	GGCTGG	438	351	45	GGGCCG	433	367	GAGGAG	395	336
21	GGAGGG	500	424	GGCCGC	479	416	46	GCCCCC	379	361	GGCTCC	481	385
22	GCGGGC	542	424	CGGCGC	500	424	47	GAGGCG	442	359	CCGGCC	283	247
23	GCGCCG	479	416	CTGCTG	462	399	48	CGCCGC	556	364	CCGCGC	384	319
24	AGGCGG	495	411	GCCCGG	618	435	49	CCCTCC	407	361	CCCGGG	474	385
25	CGCGCC	496	406	CGGCCG	481	375	50	CGCGGG	408	358	CCCAGC	571	395

NR represents non-redundant frequency i.e., multiple occurrence of 6-nt sequences identified with in single promoter sequence is counted as one. R represents redundant frequency i.e., total count of 6-nt sequence in the promoter sequences. These frequencies were arranged in their descending order and used for frequency plots in fig 3.2.

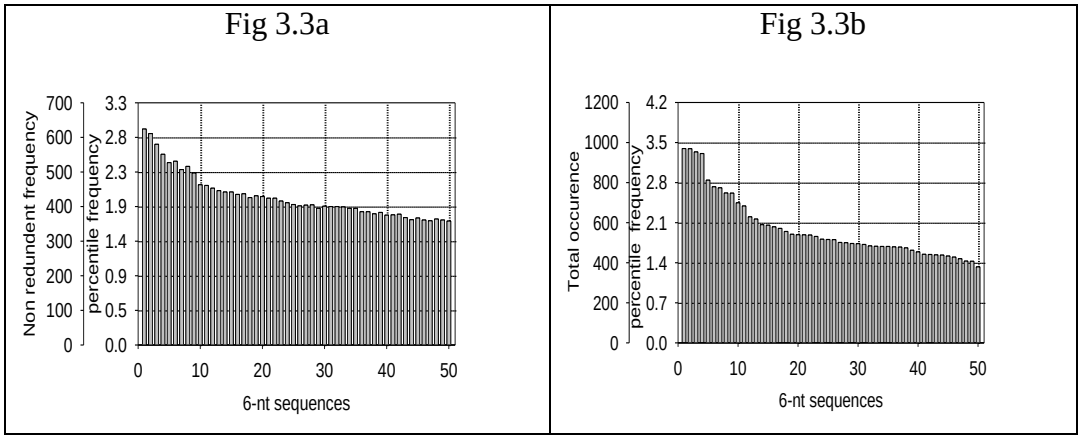


Fig 3.3: Frequency distributions of the most common 50 6-nt sequences from upstream region in the promoter database (-100 to +1 relative to TSS). We have plotted the percentile frequency and total occurrence (on y-axis with decreasing frequencies) of each 6-nt sequence. We can observe the sigmoidal behavior in distribution of the graphs. Fig 3.3a illustrates the non-redundant occurrence of each 6-nt sequence in upstream promoter sequences and their corresponding percentile frequency. Fig 3.3b represents the total occurrence of each of the 6-nt sequence and their corresponding percentile distribution.

6-nt sequence distributions in upstream and downstream promoter sequences suggest the existence of cooperative behavior in 6-nt sequences distributions.

3.1.3 Distribution of all Possible Subsequences (6-7nt sequences) in Promoters

We looked for the distribution of all possible subsequences that are 6-nt and 7-nt length, identified in promoters (Methods: 2.2.1). We have identified 4064 6-nt sequences out of 4^6 = 4096 possible 6-nt sequences in the promoter data (-100 to +100 relative to TSS). We have counted (non-redundant) occurrence of each 6-nt sequence in the promoter sequences. The frequencies were plotted on logarithmic scale for visual clarity (fig 3.4). We can observe that ~25% of 6-nt sequences appear to be random in distribution. We have identified 12,550 7-nt sequences out of 4^7 = 16,384 possible 7-nt sequences in the promoter data. Frequency occurrence of each

7-nt sequence was calculated and the frequencies were plotted (in decreasing order) on logarithmic scale for visual clarity (fig 3.5). Out of these 12,550 sequences, ~23% appear to be random in distribution.

Figure 3.4

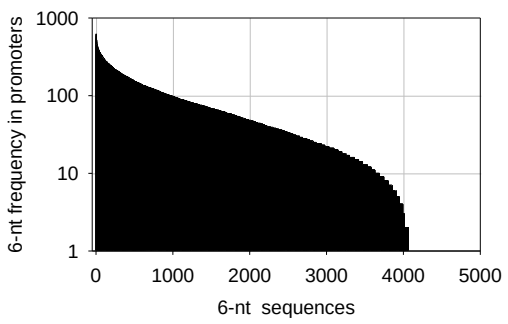


Figure 3.5

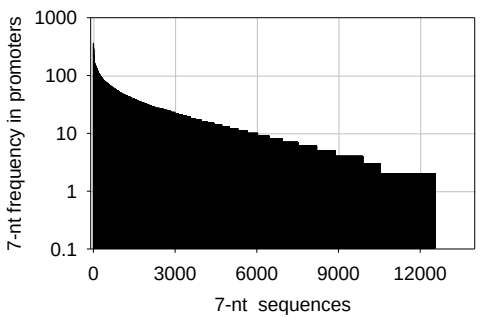


Fig 3.4: Frequency distribution of all possible 6-nt sequences identified in promoter sequences after pair wise search. The frequencies were arranged in decreasing order on logarithmic scale (on y-axis) for visual clarity. ~25% of 6-nt sequences appear to be random in distribution. Fig 3.5: Frequency distribution of all possible 7-nt sequences identified in promoter sequences (-100 to +100 relative to TSS). The frequencies were arranged in decreasing order on logarithmic scale for visual clarity. ~23% of 7-nt sequences appear random in distribution.

From the above distributions, we can note that ~25% of 6-nt sequences appear random in distribution, which can be identified in most important genes with high frequencies.

3.1.4 6-nt Sequence Distributions in Promoters (-100 to +100 relative to TSS)

We looked for frequency distribution of the most common 50 6-nt sequences (Table 3.2) in promoter sequences (-100 to +100 relative to TSS). We counted number of 6-nt sequences that were identified in each promoter sequence (non-redundant frequency). The count was plotted in decreasing order on y-axis (fig 3.6). We noted that one of the promoter sequences has 36 different 6-nt sequences in the promoter

data set. From the graph, it is clear that more than 50% of the promoters contain more than 10 different 6-nt sequences.

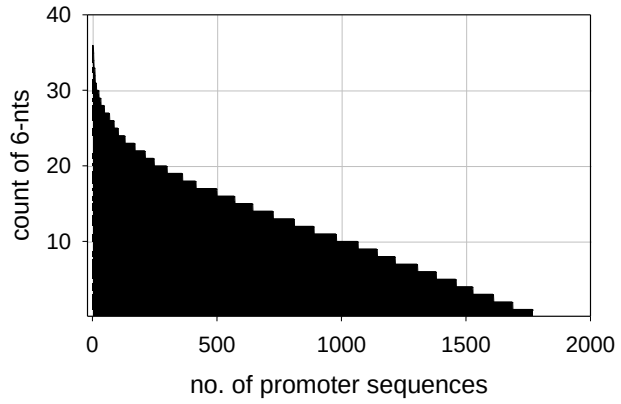


Fig 3.6: Graph represents the number of 6-nt sequences identified in 1871 promoter sequences (-100 to +100 wrt TSS). Number of 6-nt sequences identified in each promoter sequence was calculated (multiple occurrence of same 6-nt sequence is considered as one). We can observe that more than half of the promoter sequences contain more than 10 6-nt sequences.

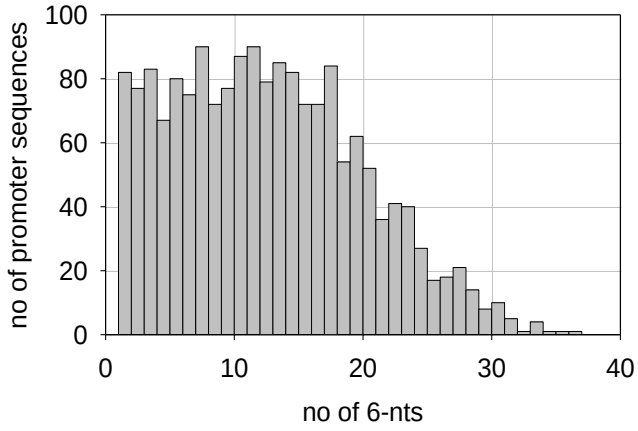


Fig 3.7: Histogram depicts the number of promoter sequences (on Y-axis) that contain number of 6-nt sequences (on X-axis). For example, the first bar represents that 82 of the promoters contain any one of the 50 6-nt sequences. From the graph, it is clear that most of the promoters contain one or more than one occurrence of 6-nt sequences and multiple occurrence of same 6-nt sequence can provide strong signals to TFs.

Table 3.3: List of most common 50 non-redundant 6-nt and 7-nt sequences that were identified in promoter data. Frequency represents the non-redundant frequencies. The roman numbers represent set of complementary, reverse and reverse complementary sequences for a particular subsequence. 6-nt sequences that were highlighted in color represent 13 6-nt sequences for which complementary sets were identified and another 13 sequences with out roman numbers represent sequences that do not have any complementary sets with in the 50 top most common 6-nt sequences.

S.No	6-nt sequence	Frequency	7-nt sequence	Frequency	S.No	6-nt sequence	Frequency	7-nt sequence	Frequency
1	GGCGGG (i)	624	GGCGGGG (i)	364	26	GCGGGC	408	GGCGGGC (ix)	180
2	GGGCGG (i)	612	GGGCGGG (ii)	358	27	GGCGCG (vii)	407	GCGCGGC (x)	180
3	GGCGGC (ii)	583	GGGGCGG (i)	350	28	GCCCGG (xii)	406	GGGGAGG	178
4	GCGGCG (iii)	555	GCGGCGG (iii)	314	29	GCGCCG (viii)	403	GCGCCGC (xi)	178
5	GCGGGG (iv)	531	CCCGCCC (ii)	299	30	GCCCCG (vi)	402	GGGCCGG (vii)	173
6	CCGCCC (i)	531	GGCGGCG (iii)	298	31	CCGGGC (xii)	400	GGGAGGG	173
7	CCCGCC (i)	519	GCGGGGC (iv)	294	32	CCGCGC (vii)	398	GCTGCGG	173
8	GGGGCG (iv)	517	CCGCCCC (i)	285	33	GGCTGC	397	GGGCGGC	172
9	CGGCGG (ii)	495	CGGCGGC (v)	282	34	CCCGGC (x)	397	GCCGGGC	171
10	GCGGCC (v)	466	CCCCGCC (i)	280	35	GCGCGC	387	CCGCGGC (xii)	170
11	CGGGGC (vi)	459	GCGGCGC (xi)	232	36	CGCGCC (vii)	384	CCCGGCC (vii)	170
12	GCGCGG (vii)	451	GCCCCGC (iv)	226	37	GGGCCG (x)	382	CGCCCCC (xiii)	169
13	CGGCGC (viii)	444	GGCGGAG (vi)	203	38	GGGGCC	380	CGGGGCC (xiv)	168
14	GGCCGG (ix)	442	GAGGCGG (vi)	203	39	GGCTGG	378	GGAGGCG	167
15	GCCGCC (ii)	442	GGGGGCG (xiii)	202	40	GGAGGG (xi)	376	CGGCGCG (x)	167
16	CCCCGC (iv)	441	CGCCGCC (iii)	200	41	GGAGGC	374	GCGCGCC	166
17	GCCGGG (x)	437	GCCGCCG (v)	199	42	CGGCCG (xiii)	371	CGCGGCC	166
18	CGCCCC (iv)	436	GGCCGGG (vii)	198	43	GCCGGC (xiii)	368	CCGGAAG	166
19	CGCCGC (iii)	433	GCGGCCG	198	44	CCGGCG (v)	367	GGCCCCG (xiv)	164
20	CGCGGC (viii)	432	GCGCGGG	197	45	AGGCGG	366	GCGGGCG	164
21	GGCCGC (v)	425	CCGCCGC (iii)	196	46	GGTGAG	361	GGGGCCG	163
22	GGGAGG (xi)	422	GGCGCGG (viii)	194	47	CCGCGG	361	GCCGCGG (xii)	161
23	GCTGGG	411	GGCGGCC	192	48	CCGCCG (ii)	360	GGCCCGG	160
24	GCGGAG	409	CGGGGCG (iv)	186	49	GGAGCC	359	GGCTGGG	159
25	GCCGCG (viii)	409	TGGCGGC	183	50	CCCGGG	358	GCCCGCC (ix)	159

The data from fig 3.6 has been plotted as histogram in fig 3.7 to observe the distribution of 6-nt sequences clearly. Histogram depicted in fig 3.7 represents the number of promoter sequences (on Y-axis) that contain any number of given 50 6-nt sequences. Data can be interpreted as for example, 82 promoters (on Y-axis) have one 6-nt sequence i.e. any of the 50 6-nt sequences (on X-axis). From the graph, we can note that 90 promoters contain 8 and another 90 promoters have 12 different 6-nt sequences in each of promoter sequence. Only one promoter contains 36 different 6-nt sequences. In broad sense, a large proportion of promoters (-100 to +100 relative to TSS) contain one or more than one occurrence of 6-nt sequences and multiple occurrence of same 6-nt sequences can provide strong signals for transcription factors.

3.1.5 Presence of Palindromes/Complementary Sequences

In the most common 50 subsequences (6 and 7-nt) that were identified in promoter sequences (-100 to +100 relative to TSS), we observed existence of complementary, reverse and reverse complementary sequences for a given subsequence. Each of 4 such 6-nt sequences were considered as a set and each set is given with corresponding roman numbers (table 3.3). Out of most common 50 subsequences (6-7nt), we can expect ~12 sets (50/4 = 12.5). In 6-nt sequences, we have identified 13 such sets that account for 74% of the 50 6-nt sequences. In case of 7-nt sequences, we have identified 66% of 50 7-nt sequences into 14 sets (including partial sets). Presence of palindrome/complementary and reverse complementary sequences suggests the presence of these subsequences on both strands of DNA. It is clear that transcription factors recognize the promoter recognition elements irrespective of strand orientation, suggesting that both the strands are playing important role in subsequence recognition (i.e. promoter recognition) by the transcription factors during initial stages of transcription. Presence of 6-nt sequences (palindromes) on either strands of DNA, doubles the

concentration of TFs and increases the number of interactions with DNA. Dimeric proteins (TFs) recognize these sequences as they can be viewed on both sides of DNA with high affinity. In support of this conjecture is the evidence from X-ray studies (Muller and Herrmann, 1997) that has confirmed the straddling of dimeric brachury transcription factor on the T-domains of palindromic DNA duplex.

Presence of complementary set of 6-nt sequences suggests that TFs can recognize the promoter elements in either strands of DNA and both the strands plays important role in transcription.

3.1.6 Positional Distribution of 6-nt Sequences in Promoter Sequences

We are interested in positional distribution of the 6-nt sequences along the length of the promoter sequences (-100 to +100 relative to TSS). Positional distribution and corresponding percentile frequencies of most common 50 6-nt sequences in promoter sequences were shown in fig 3.8. We calculated number of 6-nt sequences occurring in each position of promoter sequences from -100 to +100 relative to TSS. The deep cleft at position '0' is due to the TSS signal. We note a peak around -5 to -25 upstream to the TSS and a broad hump after -40 from TSS. This suggests that signals are more prominent with in -100 region from TSS. At the same time we can also observe a broad hump in downstream of TSS. This gives clear idea that these signals are also present in downstream regions of promoter sequences and supports the 6-nt distribution in downstream of promoters as shown in fig 3.2. We observe a dip in 6-nt sequences around -30 from TSS, suggesting that this region might be rich with A/T sequences rather G/C. Though our study did not focus on TATA box, from literature (Yoshiro et al., 2004), it is evident that number of promoters with TATA box is very less in the promoter data (about 97) and their distribution is spanned around -23 to -35

from TSS, that is consistent with the distribution depicted in fig 3.8. It should also be noted that the number of 6-nt sequences was not completely absent in this region. This in turn can be qualitatively correlated and consistent with dinucleotide distribution shown in fig 3.1.

We have looked for the distribution of the 13 particular sequences for which complementary and reverse complementary sequences were identified and another 13 sequences for which the complementary sequences were not identified (table 3.3). We have counted the occurrence of these 26 6-nt sequences and their percentile frequencies were plotted in fig 3.9. It is clear that the positional distribution of these 26 sequences in promoter data is broadly similar to the positional distribution of 50 6-nt sequences as shown in fig 3.8. We suggest that the signals are present within the neighborhood of -100 region relative to the TSS.

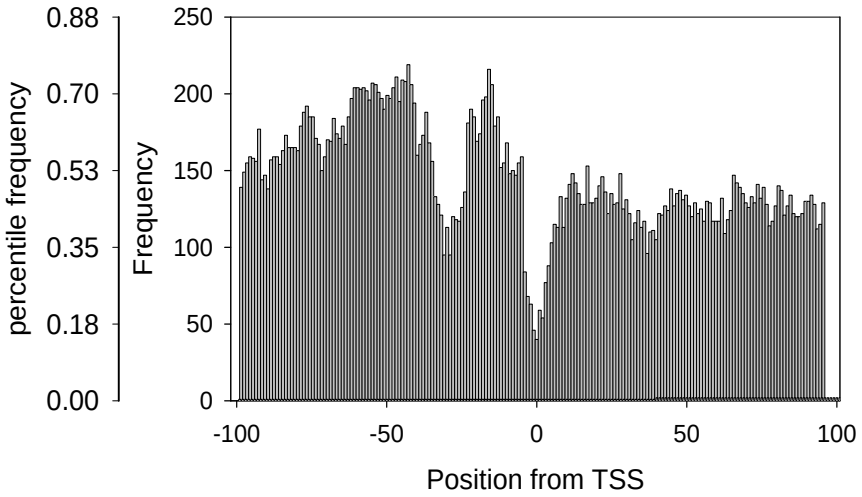


Fig 3.8: Positional distributions of 6-nt sequences occurring in each position along -100 to +100 relative to TSS in promoter sequences. X-axis represents the position and Y-axis corresponds to number of 6-nt sequences occurring at each position and their percentile frequency in promoter sequences. The deep cleft at position '0' is due to TSS signal.

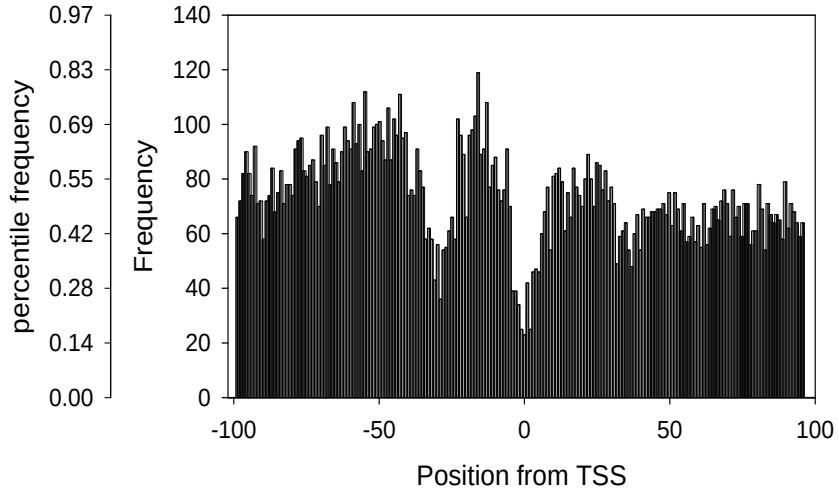


Fig 3.9: Positional distribution of 26 6-nt sequences (table 3.3) in promoter sequences (-100 to +100 relative to TSS). Frequencies of 26 6-nt sequences that are occurring in each position of promoter sequences along -100 to +100 relative to TSS have been calculated and their percentile frequencies were plotted on Y-axis. The sharp dip at position '0' is due to presence of TSS signal. We can clearly see a sharp peak around -5 to -25 and a broad hump after -40 from TSS.

Positional distribution of 6-nt sequences in the promoter sequences (-100 to +100 relative to TSS) reveal that actual physical location of promoter recognition elements can be within the neighborhood of TSS.

3.2 Frequency Distributions of 6-nt Sequences in miRNAs

In addition to post-transcriptional regulation, there are increasing evidences for important role of miRNA in many biological functions such as signal transduction, cell cycle regulation, tissue specific cell differentiation, apoptosis, etc. With these evidences we were interested in checking the role of miRNA in promoter recognition during transcription based on the 6-nt sequences.

Earlier we have worked with the 866 and 695 mature and stem-loop miRNA sequences collected from release 12 (Putta and Mitra, 2010). miRBase is growing rapidly with new miRNA sequence depositions and the current release 16 (as on September 2010) contains 1223 mature miRNA and 1045 stem-loop miRNA sequences. We have updated the frequency distribution of the top 50 most common 6-nt sequences (table 3.3) in 1223 and 1045 mature and stem-loop miRNA datasets here.

Distributions of 6-nt sequences in mature miRNA dataset were shown in fig 3.10. Out of 1223 mature miRNAs, only 123 miRNAs have any one of the 6-nt sequences. We have calculated number of 6-nt sequences that were present in each miRNA sequence and the frequencies were represented in fig 3.10. Next we calculated the total occurrence of each 6-nt sequence appeared in the mature miRNA dataset and the frequencies were shown in fig 3.11. We noted that ~12% of 50 6-nt sequences were absent (not found) in mature miRNAs. This may suggest that perhaps these 12% do not really represent valid promoter recognition elements or it may also mean that miRNAs role is rather restricted to selected promoters as they are playing important roles in other biological functions. The 6-nt sequences were found to appear 233 times in 123 mature miRNA sequences i.e., each 6-nt sequence with a mean of 1.8 times. Frequency distribution of 6-nt sequences in stem-loop sequences was represented in fig 3.12. We have calculated the number of 6-nt sequences present in each stem-loop miRNA and the frequencies were plotted in their decreasing order. We noted that out of 1045 stem-loop miRNA sequences, 312 sequences contain at least one of the 6-nt sequences. 6-nt sequences are found to be 1089 times, i.e. with a mean of ~3.4. We note that ~8% of 6-nt sequences were not identified in stem-loop miRNA sequences.

Fig 3.10

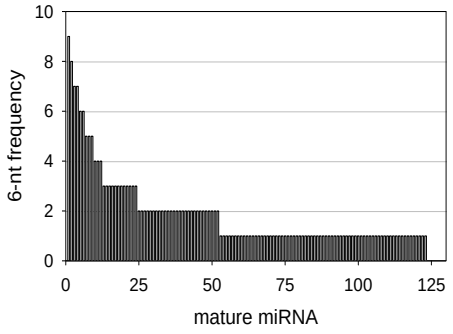


Fig 3.11

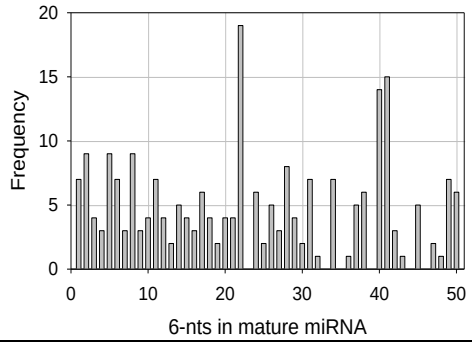


Fig 3.12

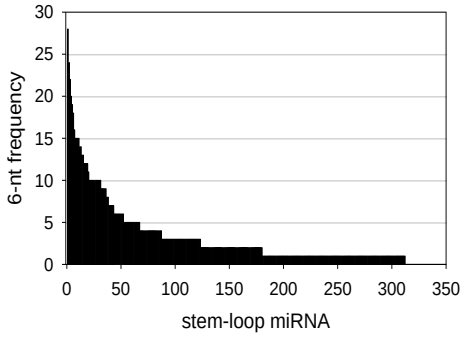


Fig 3.13

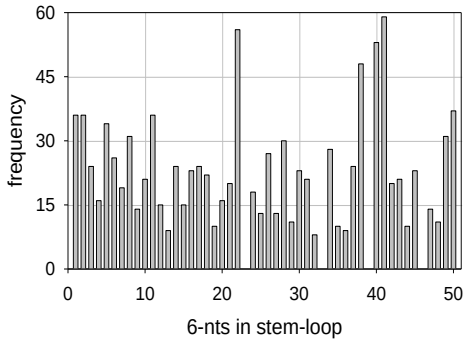


Fig 3.10: Frequency distribution of 6-nt sequences identified in mature miRNA dataset. Number of 6-nt sequences identified in each miRNA sequences were calculated and represented in their decreasing frequencies. We note that only 123 mature miRNAs contain any one of the 6-nt sequences. Fig 3.11: Graph represents the total number of each 6-nt sequence in mature miRNA data. We have calculated occurrence of each 6-nt sequence in mature miRNA sequences. We note that 12% of 6-nt sequences were not identified in mature miRNA sequences. Fig 3.12: Frequency of 6-nt sequences in stem-loop miRNA dataset. We calculated the number of 6-nt sequences identified in stem-loop miRNAs and plotted in their decreasing frequencies (Y-axis). Out of 1045 stem-loop miRNAs, we note that 312 sequences contain any of the 6-nt sequences. Fig 3.13: Graph represents the total count of each 6-nt sequence that was identified in stem-loop miRNA sequences. We note that few 6-nt sequences were not identified in the data.

We consider the average length of miRNA as 20 and 100 for mature and stem-loop miRNA respectively, the ratio of average length of miRNA and the occurrence of 6-nt sequences gives an indicative value for periodicity i.e., how often a 6-nt sequence is present/occurring in miRNA sequences. For mature miRNA, the index value is about ~11 (i.e., 20/1.8 = 11.1) and for stem-loop miRNA the periodicity value is about ~29 (100/3.1 = 29.4). These values indicate that the 6-nt sequences occur/present for every 11 and 29 residues for mature and stem-loop sequences respectively.

In case of mature miRNA sequences, we observe that there is a rapid decrease in the intensity of recognition of 6-nt sequences. This suggests that distributions of 6-nt sequences within miRNA dataset are not of random in distributions. Statistically, assuming uniform distribution of the bases, any given 6-nt sequence is expected to occur every $4^6 = 4096$ bases. For mature miRNA sequences, if we consider the mean length of each miRNA as 20-nts, we can expect at most 4 occurrences of any arbitrary 6-nt sequence. The observed number is far larger and cannot be explained by randomness. We would like to mention that number of miRNA sequences in miRBase has been increasing and there may be an increase in frequencies of 6-nt sequences in both mature and stem-loop miRNA datasets.

To know the density of 6-nt sequences position wise, we calculated the count of 6-nt sequences along the length of mature and stem-loop miRNA datasets. It should be noted that the lengths of both mature and stem-loop sequences are varying from 17-27 nucleotides and 47-150 nucleotides in length for mature and stem-loop miRNA respectively. Figure 3.14 represents the density or positional distribution of 6-nt sequences in mature miRNA sequences. It is clear that density of 6-nt sequences is high in first half compared to second half with in the graph or we can say middle part of the miRNA can recognize the 6-nt sequences. In broad,

we can conclude that 5' ends of the miRNA are most favored region for promoter recognition. Positional distribution of 6-nt sequences in stem-loop miRNAs was shown in fig 3.15. We can see decrease in the frequency of 6-nt sequences along the length of stem-loop miRNA.

Fig 3.14

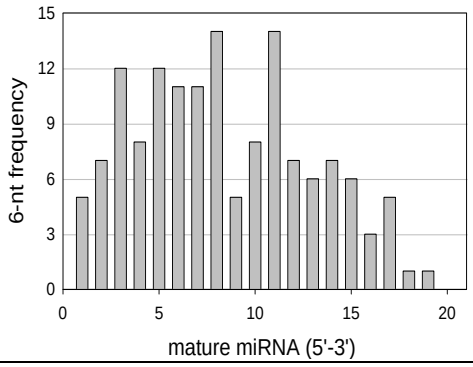


Fig 3.15

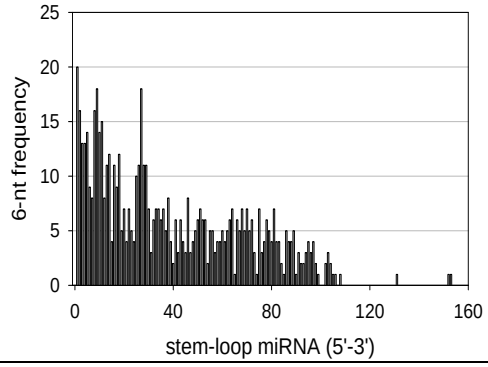


Fig 3.14: Positional frequency of 6-nt sequences in mature miRNA dataset. We have calculated the number of 6-nt sequences present along the length of mature miRNA sequences. We can note that density of 6-nt sequences is high in first half i.e., towards 5'end. Fig 3.15 represents positional frequency of 6-nt sequences in the stem-loop miRNA dataset.

We have attempted to correlate the promoter sequences and miRNA sequences based on the common 6-nt sequences that were identified in both miRNA sets. We have listed the promoter ids, mature miRNA ids and stem-loop miRNA ids with corresponding 6-nt sequences. Only few entries have been tabulated here in table 3.4. From table, it can be noted that most of promoters share a common 6-nt sequence and only few miRNA sequences share common 6-nt sequence. We suggest these mature miRNA that are sharing common 6-nt sequences with promoters may regulate the expression of a group of promoter sequences (either positively or negatively). This data is consistent with the recent reports (Young *et al*., 2009) suggesting that promoters can be potential candidates for regulation by miRNA than the 3'-UTRs of mRNA.

Table 3.4: List of promoter sequences and miRNAs that are sharing common 6-nt sequences. Only few promoters have been included here. Column 1-3 represents the promoter ids and 6-nt sequences, column 4-6 represent mature miRNA ids and 6-nt sequences identified and column 7-9 represent stem-loop miRNA ids and 6-nt sequences identified in the data. # The numbers in this column represent the serial number of 6-nt sequences that was given in table 3.2 and the next column represents the corresponding 6-nt sequence.

Promoter ID	#	6-nt sequence	Mature miRNA ID	#	6-nt sequence	Stem-loop miRNA ID	#	6-nt sequence
EP17036 (+) Hs snRNA U2	1	GGCGGG	hsa-miR-1915 MIMAT0007892	1	GGCGGG	hsa-mir-33b MI0003646	1	GGCGGG
EP49001 (+) Hs histone H1t	1	GGCGGG	hsa-miR-1228* MIMAT0005582	1	GGCGGG	hsa-mir-92b MI0003560	1	GGCGGG
EP15024 (+) Hs histone H33	1	GGCGGG	hsa-miR-638 MIMAT0003308	1	GGCGGG	hsa-mir-135a-1 MI0000452	1	GGCGGG
EP11074 (+) Hs histone H4-A1	1	GGCGGG	hsa-miR-1908 MIMAT0007881	1	GGCGGG	hsa-mir-203 MI0000283	1	GGCGGG
EP31007 (+) Hs HMG-14	1	GGCGGG	hsa-miR-663 MIMAT0003326	1	GGCGGG	hsa-mir-326 MI0000808	1	GGCGGG
EP31009 (+) Hs HMG-17	1	GGCGGG	hsa-miR-1228* MIMAT0005582	2	GGGCGG	hsa-mir-499 MI0003183	1	GGCGGG
EP33038 (+) Hs PRM2	1	GGCGGG	hsa-miR-638 MIMAT0003308	2	GGGCGG	hsa-mir-566 MI0003572	1	GGCGGG
EP37014 (+) Hs[rig] rp S15	1	GGCGGG	hsa-miR-1231 MIMAT0005586	2	GGGCGG	hsa-mir-615 MI0003628	1	GGCGGG
EP14031 (+) Hs b'-tubulin b'2	1	GGCGGG	hsa-miR-638 MIMAT0003308	3	GGCGGC	hsa-mir-636 MI0003651	1	GGCGGG
EP24039 (+) Hs vimentin	1	GGCGGG	hsa-miR-1915 MIMAT0007892	4	GCGGCG	hsa-mir-638 MI0003653	1	GGCGGG
EP33011 (+) Hs DES	1	GGCGGG	hsa-miR-1228* MIMAT0005582	5	GCGGGG	hsa-mir-639 MI0003654	1	GGCGGG
EP16038 (+) Hs fibronectin	1	GGCGGG	hsa-miR-1908 MIMAT0007881	5	GCGGGG	hsa-mir-663 MI0003672	1	GGCGGG

3.3 Frequency Distributions of 6-nt Sequences in TFBS

3.3.1 Distribution of 6-nt Sequences in 75 TFBS Sequences

We have collected two sets of human transcription factor binding site (TFBS) sequences from JASPAR database (Methods: 2.1.3). The initial set contains 75 tfbs sequences. The top 50 common redundant and non-redundant 6-nt sequences (table 3.1) that were identified in human promoter sequences were searched in the 75 TFBS sequences (table 3.5). The distributions of non-redundant 6-nt sequences in 75 TFBS sequences were shown in fig 3.16. We note that a very few 6-nt sequences have been identified in 75 TFBS sequences. Distributions of redundant 6-nt sequences were presented in fig 3.17. Out of 75, 9 TFs recognized 6-nt sequences.

Figure 3.16

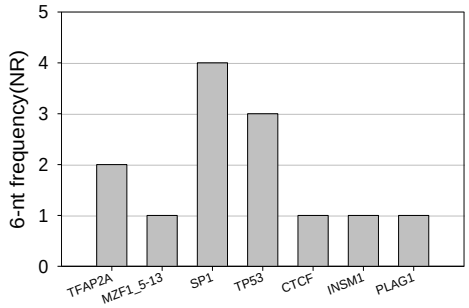


Figure 3.17

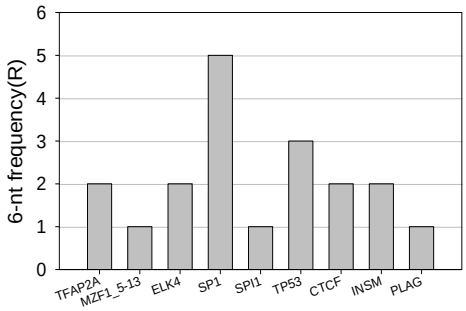


Fig 3.16: Frequency distribution of 6-nt sequences (non-redundant) in 75 transcription factor binding sites. It was observed that out of 75 transcription factors, only 7 transcription factors recognized any of the 50 6-nt sequences. Fig 3.17: Frequency distribution of 6-nt sequences (redundant) in 75 TFBS sequences. The redundant sequences were identified in two more TFs, thus a total of 9 TFs contain any of the 6-nt sequences.

Table 3.5: List of human transcription factor binding site sequences collected from JASPAR database. A total of 75 transcription factor's binding profiles were reported in the database. ID represents the id of TF from JASPAR. Column 3 and 7 represents the total number of experimentally verified TFBS (transcription factor binding site) sequences that were reported for each TF. TFBS sequences listed here were collected based on the highest frequencies of nucleotides reported from frequency matrices of binding profiles for each TF (Methodology 2.1.3).

ID	TF Name	No of TFBS	TFBS sequence	ID	TF Name	No of TFBS	TFBS sequence
MA0003.1	TFAP2A	-	GCCCGGGGG	MA0055.1	Myf	-	CAGCAGCTGCTG
MA0017.1	NRF2F1	13	TGACCTTTGAACCT	MA0056.1	MZF1_1.4	20	TGGGGA
MA0018.2	CREB1	11	TGACGTCA	MA0057.1	MZF1_5-13	16	GGAGGGGGAG
MA0024.1	E2F1	10	TTTGGCGC	MA0058.1	MAX	17	[A/G] ACCACGTGA
MA0025.1	NFIL3	23	TTATGTAAC[G/C]T	MA0059.1	MYC-MAX	21	GA[C/G] CACGTGGT
MA0028.1	ELK1	28	GAGCCGGAAT	MA0060.1	NFYA	116	CCCAACCAATCAGCGC
MA0030.1	FOXF2	20	CAAACGTAAACATT	MA0061.1	NF-kappaB	38	GGGAATTTCC
MA0031.1	FOXD1	20	GTAAACAT	MA0066.1	PPARG	28	GTAGGTCACGGTGACCTACT
MA0032.1	FOXC1	-	GGTAAGTA	MA0069.1	Pax6	43	TTCACGCATGAGTT
MA0033.1	FOXL1	-	TATACATA	MA0070.1	PBX1	18	CCATCAATCAAA
MA0036.1	GATA2	53	GGATA	MA0071.1	RORA_1	25	ATCAAGGTCA
MA0037.1	GATA3	63	AGATAG	MA0072.1	RORA_2	36	TATAAGTAGGTCAA
MA0042.1	FOXI1	-	GGATGTTTGTTT	MA0073.1	RREB1	11	CCCCAAACACCCCCCCCC[C/A]
MA0043.1	HLF	18	GGTTACG[C/T]AATN	MA0074.1	RXRA-VDR	10	GGGTCA [A/T] CG [A/C] GTTCA
MA0048.1	NHLH1	54	GCGCAGCTGCGT	MA0076.1	ELK4	20	ACCGGAAGT
MA0050.1	IRF1	-	GAAAG[C/T]GAAACC	MA0077.1	Sox 9	76	GAACAATGG
MA0051.1	IRF2	12	GGAAAG[C/T]GAAA[C/G]CAAAAC	MA0079.2	SP1	35	CCCCGCCCCC
MA0052.1	MEF2A	58	CTATTTATAG	MA0080.2	SPI1	42	AGGAAGT

Continued…

...continued

ID	TF Name	No of TFBS	TFBS sequence	ID	TF Name	No of TFBS	TFBS sequence
MA0081.1	SPIB	49	AGAGGAA	MA0133.1	BRAC1	43	ACAACAC
MA0083.1	SRF	46	GCCCATATATGG	MA0137.2	STAT1	2085	CATTTCCCGGAAACC
MA0084.1	SRY	28	GTAAACAAT	MA0138.2	REST	874	TTCAGCACCATGGACAGCGCC
MA0090.1	TEAD1	-	CACATTCCTCCG	MA0148.1	FOXA1	897	TGTTTACTTTG
MA0091.1	TAL1-TCF3	44	CGACCATCTGTT	MA0149.1	EWSR1-FLI1	101	GGAAGGAAGGAAGGAAGG
MA0093.1	USF1	30	CACGTGG	MA0150.1	NFE2L2	20	ATGACTCAGCA
MA0095.1	YY1	17	GCCATC	MA0152.1	NFATC2	26	TTTTCCA
MA0098.1	ETS1	40	[T/C]TTCCG	MA0153.1	HNF1B	9	TTAATATTTAAC
MA0099.2	AP1	18	TGACTCA	MA0155.1	INSM1	24	TGTCAGGGGGCG
MA0101.1	REL	17	GGGGATTTCC	MA0156.1	FEV	13	CAGGAAAT
MA0105.1	NFKB1	18	GGGGATTCCCC	MA0157.1	FOXO3	13	TGTAAACA
MA0106.1	TP53	17	CCGGACATGCCCGGGCATGT	MA0158.1	HOXA5	16	CACAAATT
MA0107.1	RELA	18	GGGAATTTCC	MA0159.1	RXR-RAR_DR5	23	AGGTCA[C/T] GGAGAGGTCA
MA0112.2	ESR1	475	GGCCCAGGTCACCCTGACCT	MA0160.1	NR4A2	14	AAGGTCAC
MA0113.1	NR3C1	9	GAGAACATTATGTCCT [A/G][A/T]	MA0161.1	NFIC	-	TTGGCA
MA0115.1	NR1H2 – RXRA	25	AAAGGTCAAAGGTCAAC	MA0163.1	PLAG1	-	GGGGCCCAAGGGGG
MA0119.1	TLX1-NFIC	16	TGGCACCATGCCAA	MA0258.1	ESR2	357	CAAGGTCACGGTGACCTG
MA0124.1	NKX3-1	20	ATACTTA	MA0259.1	HIF1A-ARNT	104	GGACGTGC
MA0130.1	ZNF354C	16	ATCCAC	MA0442.1	SOX10	-	CTTTGT
MA0131.1	MIZF	20	TAACGTCCGC				

3.3.2 Distribution of 6-nt Sequences in 6497 TFBS Sequences

The TFBS list was extended to the entire list of experimentally verified binding site sequences that were used to construct the frequency matrices in JASPAR database (Methods: 2.1.3). A total of 6497 binding site sequences were collected from the database. Binding site sequence information for only 65 transcription factors is available in the database. 6497 binding site sequences correspond to 65 TFs from JASPAR. We searched for the top 50 most common 6-nt sequences identified in promoter sequences (table 3.3) in the 6497 TFBS sequences. Distribution of 6-nt sequences in all 6497 binding site sequences was depicted in fig 3.18. We can observe that 6-nt sequences were not identified in all TFBS sequences. We have calculated the number of 6-nt sequences occurring in each TFBS sequences and their frequencies were plotted (decreasing frequencies on Y-axis) in fig 3.19. X-axis represents the number of TFBS sequences that contain any number of 6-nt sequences. We note that out of 6497 binding sequences, 269 (4.1%) are known to contain one or more than one of any 6-nt sequences. Total frequency of each of 50 6-nt sequences in 269 TFBS was calculated and their frequencies were presented in fig 3.20. We note that two of the 50 6-nt sequences were not identified in TFBS sequences.

Next we have looked for distribution of 6-nt sequences in each transcription factor. For example, SP1 contains a list of 35 experimentally verified binding site sequences from the database. It was noted that a total of 50 (including multiple occurrences) 6-nt sequences were identified in the 35 binding sites of SP1. The number of 6-nt sequences identified in each of the transcription factor was calculated and the frequencies were represented in fig 3.21. From the graph, ESR1 is known to contain highest number of 6-nt sequences (including multiple occurrences). Out of 65 transcription factors, 17 factors (26%) recognize any of the 6-nt sequences. The 6-nt sequences appeared 402 times (including multiple

occurrences) in the binding sequences of the 17 transcription factors. We correlated the promoter sequences, miRNA and tfbs sequences that share common 6-nt sequences. Few entries of these sequence ids were tabulated for demonstration (table 3.6). We suggest that a set/group of promoters can be regulated by the particular set of miRNA and TFs.

Fig 3.18

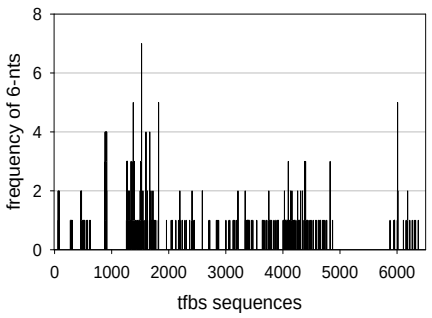
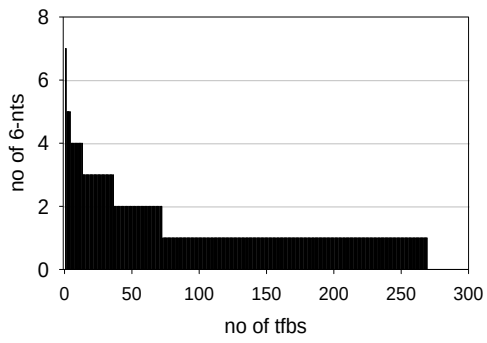


Fig 3.19



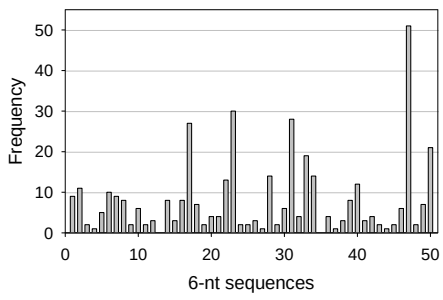


Fig 3.20

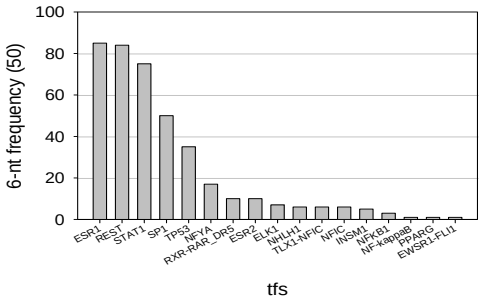


Fig 3.21

Fig 3.18: Distribution of 6-nt sequences in all 6497 binding site sequences. It is noted that most of the TFBS sequences do not contain the top 50 most common 6-nt sequences. Fig 3.19: Frequencies of 6-nt sequences identified in 269 tfbs sequences were plotted as decreasing frequencies on Y-axis. Fig 3.20: Count of each 6-nt sequence (table 3.2) identified in 269 TFBS sequences. Fig 3.21: Frequency distribution of 6-nt sequences identified in 65 transcription factors. The number of TFBS sequences identified with any 6-nt sequences were calculated and represented as corresponding TF (in their decreasing frequencies). We note that binding site sequences of ESR1 has highest number of 6-nt sequences.

Table 3.6: List of promoters, miRNA and TFBS that are sharing common 6-nt sequences. Only few entries have been tabulated here. We suggest that a set or group of promoters can be regulated or expressed by few miRNA and TFs listed here. Numbers in 5th column represent the serial number of corresponding binding site of each TF.

S.no	hexamer	Promoter id	miRNA id	TFBS
1	GGCGGG	EP17036 (+) Hs snRNA U2 EP49001 (+) Hs histone H1t EP15024 (+) Hs histone H33 EP11074 (+) Hs histone H4-A1 EP31007 (+) Hs HMG-14 EP31009 (+) Hs HMG-17	 hsa-miR-1915 hsa-miR-1228* hsa-miR-638 hsa-miR-1908 hsa-miR-663	ELK1 SP1 1,6,15,15 ESR1 303 STAT1 1944 RXR- RAR_DR5 15
2	GGGCGG	EP49001 (+) Hs histone H1t EP30042 (+) Hs histone H1a EP15024 (+) Hs histone H33 EP31007 (+) Hs HMG-14 EP41007 (+) Hs rp S17 EP25034 (+) Hs a'1(I) collagen EP16038 (+) Hs fibronectin	 hsa-miR-1228* hsa-miR-638 hsa-miR-1231	ELK1 NFYA 19 SP1 1,6,14 SP1 15,32 ESR1 41, 227,303 REST 429
3	GGCGGC	EP30042 (+) Hs histone H1a EP15024 (+) Hs histone H33 EP31007 (+) Hs HMG-14 EP31009 (+) Hs HMG-17 EP17045 (+) Hs b'-actin	 hsa-miR-638	 ESR1 41 ESR1 227
4	GCGGCG	EP30042 (+) Hs histone H1a EP15024 (+) Hs histone H33 EP31007 (+) Hs HMG-14 EP31009 (+) Hs HMG-17 EP24040 (+) Hs rp S14	 hsa-miR-1915	 REST 20
5	GCGGGG	EP17030 (-) Hs snRNA U1 (pU1-6) EP17031 (-) Hs snRNA U1(pHU1-1) EP17041 (-) Hs snRNA U4C EP49001 (+) Hs histone H1t EP11073 (+) Hs histone H3b EP15024 (+) Hs histone H33 EP31007 (+) Hs HMG-14 EP33038 (+) Hs PRM2	 hsa-miR-1228* hsa-miR-1908 hsa-miR-663 hsa-miR-885-3p	 SP1 1 SP1 6 SP1 14 SP1 15 SP1 32

The most common 50 6-nt sequences were identified in ~26% of the transcription factors. Based on the common 6-nt sequences among promoter, miRNA and TFBS sequences, we can conclude that a number of promoters or genes can be coregulated by a particular set of miRNA and transcription factors.

3.3.4 Distribution of TFBS in Promoter Sequences

We have looked for the distribution of 75 TFBS sequences (table 3.5) in the human promoter dataset (-100 to +100 relative to TSS). It was observed that 40% of the promoters contain any of the 75 binding site sequences. Out of the 75, binding sites of 31 TFs were recognized in promoter sequences. Number of promoter sequences that contain any TFBS sequences were counted and plotted in decreasing frequencies in fig 3.22. It appears that most of these promoters have binding sites for MZF1 factor.

Fig 3.22

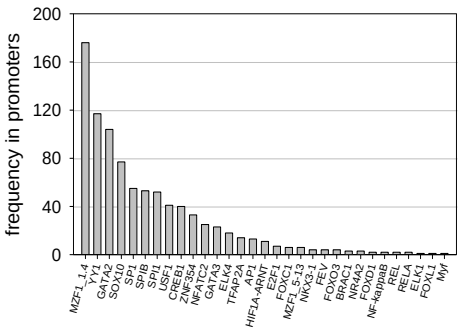


Fig 3.23

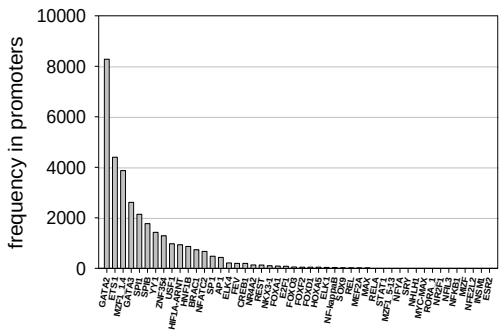


Fig 3.22: Frequency distribution of 75 tfbs in human promoter sequences. Y-axis represents the frequency of TFBS sequences of each transcription factor in the promoter data arranged increasing order. Fig 3.23: Distribution of 6497 TFBS in promoter sequences. GATA2 has highest frequency in the promoter sequences. Out of 65, 48 transcription factors were identified in the promoter data set. The frequencies were arranged in decreasing order.

Next, all the 6497 binding site sequences were searched in the promoter sequences (-100 to +100 relative to TSS). We noted that 99.8% of (1868 out of 1871) promoter sequences have binding sites of 48 transcription factors. Figure 3.23 represents the distribution of all 6497 TFBS in the promoter data. Number of promoter sequences that contain corresponding binding site sequences were

counted and plotted as frequencies. We noted that number of STAT1 binding site sequences identified in promoter data is less despite the fact that STAT1 has highest number of (2085) experimentally verified TFBS sequences among all TFs in the database. GATA2 has highest frequencies in the promoter data among the 48 TFs. We note that TFBS sequence length of GATA2 is 5-nts and it appears that these sequences are abundant in the promoter data.

Figure 3.24 a

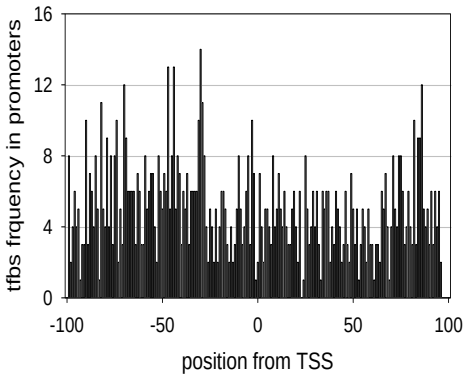


Figure 3.24 b

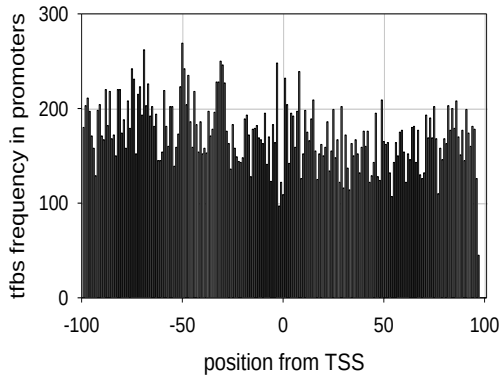


Figure 3.24a represents the positional distribution of 75 TFBS in promoter sequences along –100 to +100 relative to TSS. Fig 3.24b: Positional distribution of all 6497 TFBS within promoter dataset. From the graphs it is evident that binding sites are highly distributed with in –100 region around TSS. We can also observe the presence of TFBS in downstream of TSS, which is in support with the distribution of 6-nt sequences in downstream of TSS shown in figure 3.2.

Positional distributions of the TFBS sequences (both 75 and 6497) in the promoter data were shown in figure 3.24 (a and b). Frequencies of TFBS sequences that are present at each position in promoter sequences (-100 to +100 relative to TSS) were calculated. Y-axis represents the frequency of TFBS sequences in the promoters along the length (on X-axis). Fig 3.24a and fig 3.24b represents the positional distribution of 75 and 6497 TFBS sequences

respectively. It can be noted that both the graphs appear broadly similar in distributions though the frequencies varies and suggest that these sites were present within the neighborhood of TSS.

3.3.4. TFBS Distribution in miRNA Data

Recent computational studies reveal that miRNA potentially regulate one third of the human genes and each miRNA on an average can target more than 200 genes (Griffith-Jones *et al*., 2005). Other studies reveal that genes that have more transcription factor binding sites have a high probability of being targeted by miRNA and have more miRNA binding sites suggesting the existence of highly correlated and coordinated regulation by TFs and miRNAs at transcriptional and post-transcriptional levels (Cui *et al*., 2007).

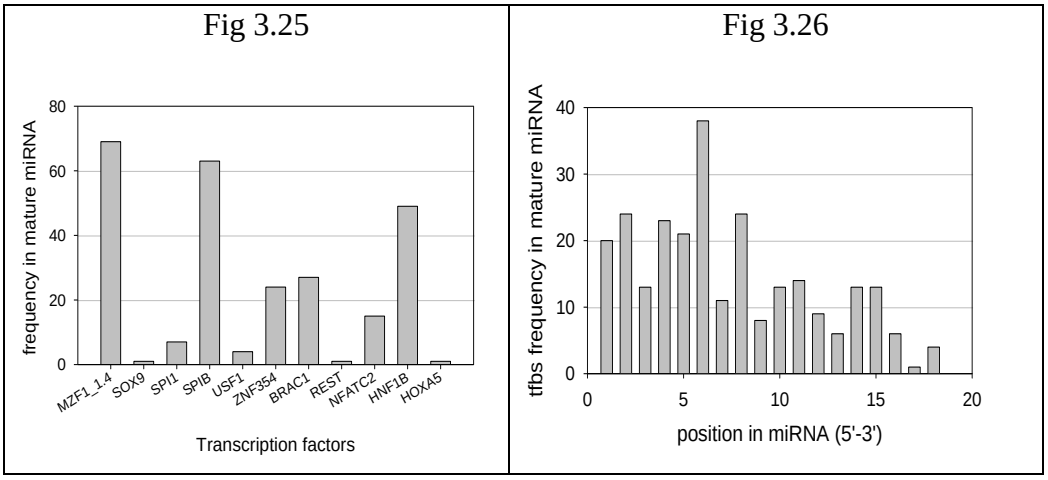


Fig 3.25: Distribution of 6497 TFBS sequences in mature miRNA dataset. Number of TFBS sequences identified in each miRNA sequence were calculated and represented as their corresponding TFs on X-axis. Fig 3.26: Positional distribution of TFBS sequences in mature miRNA sequences. Number of TFBS occurring in each position of miRNA were calculated and represented on Y-axis along the length of miRNA sequence (X-axis).

We have looked for the distribution of 6497 TFBS sequences in mature miRNA dataset. Out of 1223 mature miRNAs, 130 miRNAs have identified 261 TFBS sequences (includes multiple occurrences). Out of 65 TFs, binding sequences of 11 TFs were identified in mature miRNA data. The transcription factor binding site distribution in miRNA was represented with their corresponding TF names in fig 3.25. MZF1 has highest frequency in the miRNA data. Next, we looked for positional distribution of these TFBS in mature miRNA data, represented in fig 3.26. We note that the density of TFBS is more at the 5'end of the miRNA (first half of the graph) and we can conclude that 5'end of miRNA are more favorable binding regions.

3.4 Common Subsequences (4-6nt) in TFBS

From 6-nt sequence distributions (table 3.3) in TFBS it is clear that the top 50 6-nt sequences in transcription factor binding site sequences are not up to random in distributions though their frequencies appear less, which led us to analyze the transcription factor binding site sequences that were reported from JASPAR.

3.4.1 Base Composition in TFBS Sequences

We looked into the base composition of all 6497 transcription factor binding site sequences of 65 human transcription factors from JASPAR. These sequences consists a total of 86,050 characters and there were 118 unknown or any bases (represented as X and N in the database). The base compositions in 6497 tfbs sequences are given below. Interestingly, it was clear that all the four bases showed uniform distribution in the 6497 binding sequences.

Following are the base frequencies observed in 6497 transcription factor binding site sequences from JASPAR.

A = 22707 (26.4 %)

T = 21360 (24.8%)

G = 21203 (24.6%)

C = 20662 (24%)

Next we looked for the dinucleotide frequencies in the tfbs sequences. Expected frequency of any dinucleotide in the given data is 5378. Out of 16 dinucleotides: AA/TT, GG/CC have highest frequencies and AT/TA, CG/GC have lowest frequencies than expected frequency. Following table lists the dinucleotide frequencies. We have also calculated the frequencies of R (A or G) and Y (C or T) combinations, which provided clear idea that the TFBS sequences are rich in RR/YY than RY/YR frequencies. The dinucleotide frequencies were tabulated in table 3.7. We can note the prevalence of RR and YY combinations of dinucleotides.

Base	A	C	G	T	R	Y
A	7096	4668	5661	3765	12757	8433
C	6198	5944	1961	5193	8159	11137
G	5071	2804	6556	4642	11627	7446
T	2795	5150	5538	6393	8333	11543
R	12167	7472	12217	8407	24384	15879
Y	8993	11094	11199	11586	16492	22680

Table 3.7: Dinucleotide frequencies in 6497 TFBS sequences. The expected frequency of any dinucleotide is 5378. R represents purines: A (Adenine) or G (Guanine), Y represents pyrimidines: C (Cytosine) or T (Thymine) according to IUPAC nomenclature for nucleotides. We can note that RR frequencies are having highest frequencies than other combinations.

3.4.2 Identification and Distribution of Subsequences in TFBS Sequences

The 6497 TFBS sequences from JASPAR database were searched pair wise to locate relatively conserved common subsequences that are of 4-6nt in length. We considered all n.(n-1)/2 pairs i.e., 6497.6496/2 = 21,102,256 pairs (Methods: 2.2.2). The redundant subsequences (4-6nt) were sorted in lexical order and the subsequences were counted for their occurrence. The redundant subsequences were sorted to get non-redundant subsequences and their frequencies were calculated. These subsequences were again sorted based on their decreasing frequencies. The top 50 most common subsequences were selected for further study (table 3.8). The 50 subsequences (4-6nt) were looked for their frequency distributions in 6497 TFBS sequences and the results were analyzed and plotted using Sigmaplot 9.0.

3.4.2.1 Frequency Distribution of 6-nt Sequences in TFBS Sequences

We have identified a total of 2588 (out of 4^6 = 4096 possible 6-nt sequences) distinct 6-nt sequences that are common in the 6497 TFBS sequences. Among these sequences, the top 50 6-nt sequences with highest frequencies (table 3.8) were selected and searched for their frequency studies in 6497 TFBS sequences. The number of 6-nt sequences identified in each TFBS sequence were calculated and represented in their decreasing frequencies in fig 3.26. Out of 6497 sequences, 4014 sequences (~62%) contain one or more than one 6-nt sequences. Out of 65 TFs, 46 TFs (of 4014 TFBS) are known to contain any of the 50 6-nt sequences. Total number of each 6-nt sequences occurring in 4014 TFBS sequences was calculated as frequencies in fig 3.27. The distribution of 6-nt sequences appears similar to sigmoidal nature. We note that the frequencies in lower end of the graph are almost half to the initial frequencies (upper end of the graph).

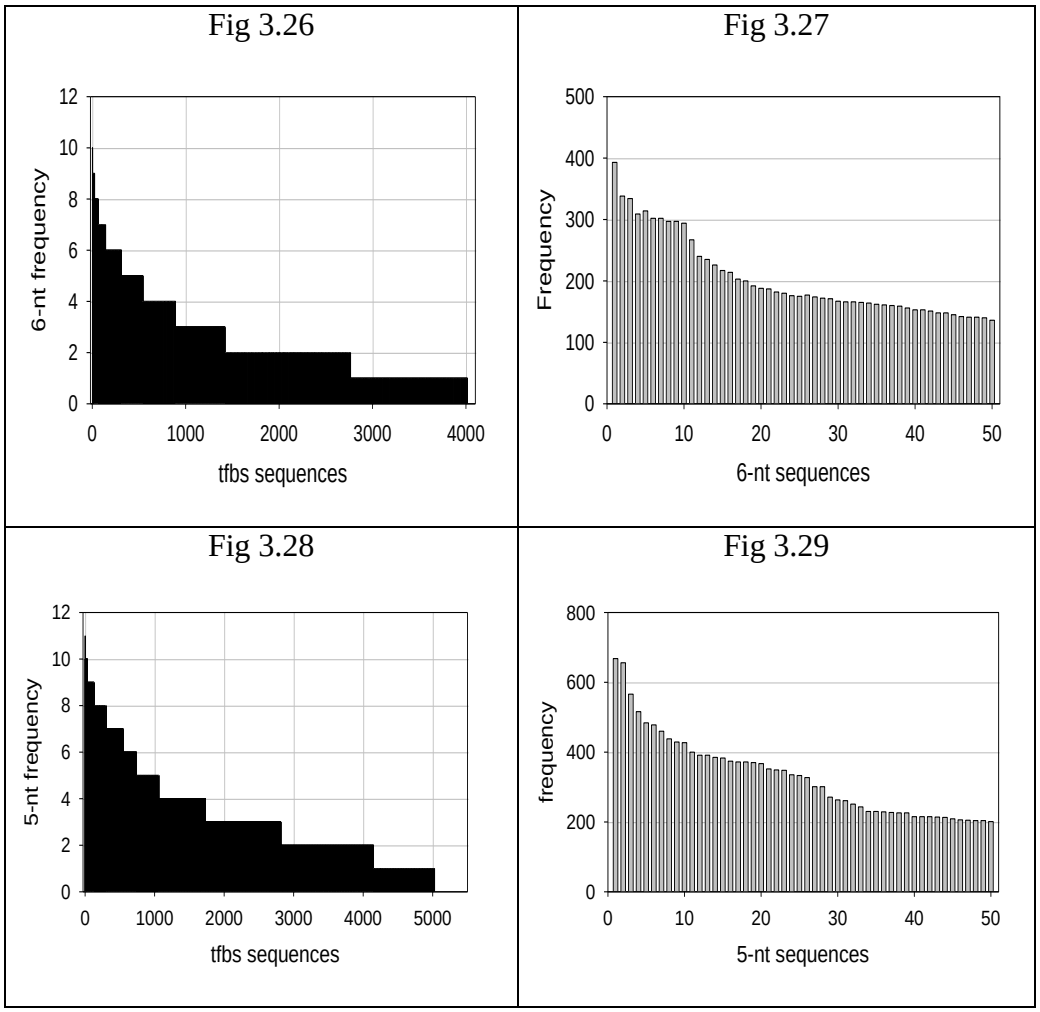


Fig 3.26: Graph represents distribution of 6-nt sequences identified in 6497 TFBS sequences. Out of 6497 tfbs sequences, 4014 (62%) sequences that represent 46 TFs, contain one or more than one 6-nt sequences. Fig 3.27: Graph represents total count of each 6-nt sequence in 4014 tfbs sequences that represent 46 TFs. Fig 3.28: Distribution of top 50 5-nt sequences in all 6497 tfbs sequences. Count of 6-nt sequences were calculated in each tfbs and plotted as the function of decreasing frequencies. 5026 (~78%) sequences that represent 54 TFs, contain one or more than one 5-nt sequence. Fig 3.29: Graph represents the total count of each 5-nt sequence in 5026 tfbs sequences that represent 54 TFs.

3.4.2.2 Frequency Distribution of 5-nt sequences in TFBS Sequences

From the pair wise search, we identified a total of 999 (out of 4^5 = 1024) 5-nt sequences in 6497 TFBS sequences. Among these sequences, top 50 5-nt sequences (table 3.6) were selected to search for their frequency distributions in 6497 TFBS sequences. The distributions were shown in figure 3.28. We counted the number of 5-nt sequences occurring in each TFBS sequence. The TFBS sequences (on X-axis) were arranged according to the decreasing frequencies of 6-nt sequences identified in them. Out of 6497, 5026 (~78%) TFBS sequences that represent binding site sequences of 54 TFs were identified with one or more than one 5-nt sequences. We have calculated total count of each 5-nt sequence in 5026 TFBS sequences and the frequencies were represented in fig 3.29. We note that the final frequencies (last 5-nt sequences of 50) were 1/3 of the initial frequencies.

3.4.2.3 Frequency distribution of 4-nt Sequences in TFBS Sequences

After pair wise alignment, we identified a total of 256 4-nt sequences (4^4 possible combinations) that are distinct in nature. Among these sequences, the top 50 most common 4-nt sequences (table 3.8) were selected and looked for their frequency distributions in 6497 TFBS sequences. We calculated the number of 4-nt sequences that were identified in each of 6497 TFBS sequences. The 4-nt frequencies were arranged in their decreasing order (fig 3.30). Out of 6497, 6039 (~93%) TFBS sequences that belong to 62 TFs contain one or more than one 4-nt sequences. We have also calculated the total frequency of each 4-nt sequence in 6039 TFBS sequences (fig 3.31). We note that first few (1-2) 4-nt sequences have highest frequencies and the last frequencies (in lower end of the graph) are approximately $1/5^{th}$ of the initial higher frequencies.

Fig 3.30

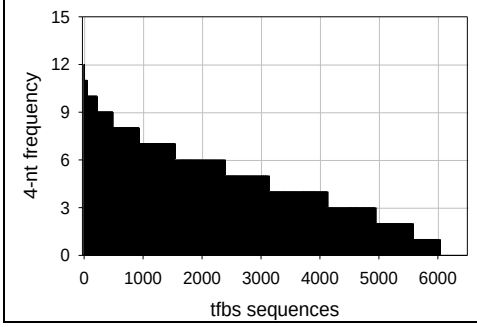


Fig 3.31

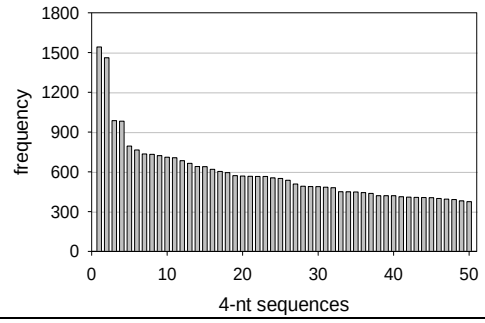


Fig 3.30: Distribution of 4-nt sequences in 6497 TFBS sequences. Out of 6497, 6039 (~93%) TFBS sequences contain one or more than one occurrence of 4-nt sequence. We can note that more than 50% of TFBS sequences contain at least five different 4-nt sequences. Fig 3.31: Graph represents total occurrence of each 4-nt sequence (of 50) in 6039 sequences that represent 62 TFs.

3.4.2.4 Distribution of all Subsequences (4-6nt) in TFBS Sequences

Frequencies of top 50 most common 4, 5 and 6-nt subsequences in 6497 TFBS sequences (table 3.8) were plotted together in fig 3.32 to observe their nature of frequency distributions. We can note that frequencies of 4-nt subsequences were high compared to frequencies of 5 and 6-nt sequences. It can be observed that lower end of 5-nt and 6-nt frequencies (last 21 subsequences) have similar pattern of distribution. We found that few of these sequences share common 5-nt sequences and difference within their frequencies is decreasing in step-wise manner. Differences in frequencies of 5-nt and 6-nt sequences suggest that 5-nt and 6-nt sequences might be conserved in frequency terms and are not up to random distributions.

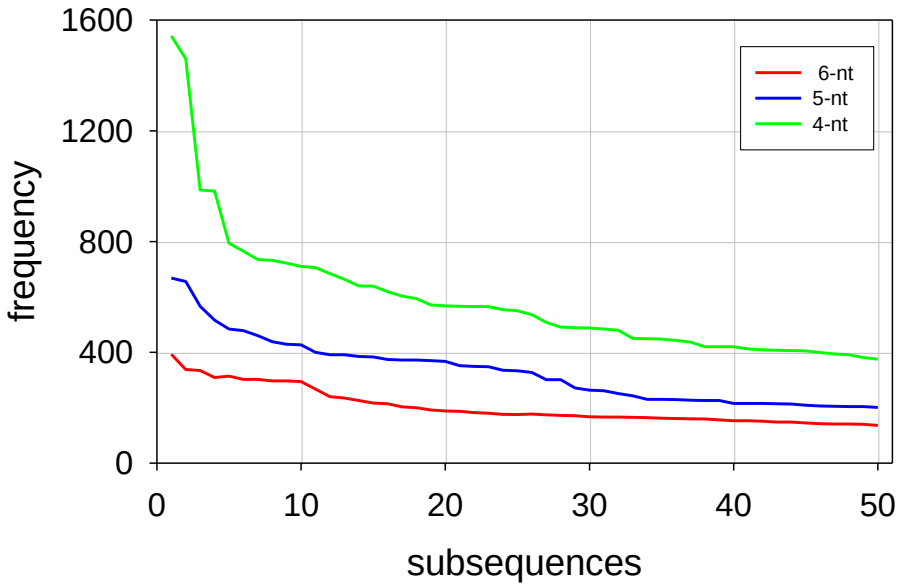


Fig 3.32: Graph represents total occurrence of each subsequence (4-6nt) in 6497 transcription factor binding sequences. Number of subsequences identified in 6497 TFBS sequences were counted and plotted as frequencies against the subsequences.

In broad, we can suggest that frequencies of the top 50 subsequences (4-6 nt) were above their expected frequencies. The frequency distributions of subsequences show that these subsequences are not random in distribution. Studies on functional significance of these subsequences in consideration with promoter elements and other biological molecules/factors can aid in unraveling the combinatorial regulatory events and help in understanding the spatio-temporal and lineage specific gene expression by TFs.

Table 3.8: List of top 50 most common 6-nt, 5-nt and 4-nt sequences that were identified in TFBS sequences that were identified after pair wise alignment. These sequences were searched in the 6497 TFBS sequences and the frequencies were used for plots in fig 3.27, 3.29 and 3.31 for distribution of 6-nt, 5-nt and 4-nt sequences respectively. # represents the non-redundant frequency of subsequences (4-6nt) in the transcription factor binding site sequences.

S.no	6-nt	#	5-nt	#	4-nt	#	S.no	6-nt	#	5-nt	#	4-nt	#
1	CAGGAA	393	TTTCC	668	GGAA	1542	26	TGGGAA	174	CTGTC	327	TGTT	536
2	TTCCTG	338	GGAAA	656	TTCC	1460	27	GGGAAA	174	TGACC	301	AAAG	508
3	GGACAG	334	AGGAA	566	GAAA	986	28	TGTTTA	172	GTAAA	301	AGGT	491
4	CTGTCC	309	GGTCA	516	TTTC	982	29	GGTCAC	171	CTGGG	271	GACA	489
5	TTCCAG	307	TTCCT	484	CAGG	794	30	CTTCCT	167	CATGG	263	TTCT	488
6	TCCAGG	302	AAACA	478	CTGG	765	31	GTTTAC	166	TTTAC	261	TCCC	484
7	AGGTCA	302	GGAAG	460	CCAG	735	32	CATTTC	166	TAAAC	251	TGTC	480
8	CCTGGA	297	CAGGA	438	AAAC	732	33	TGGACA	165	CCATG	243	TGAC	450
9	CCAGGA	297	CCAGG	429	GTCA	722	34	TTCCCA	164	GTCAC	230	AGAA	449
10	CTGGAA	294	CTTCC	427	CCTG	710	35	TCTGGG	162	AGAAA	230	GGAC	448
11	TCCTGG	267	GGGAA	400	AGGA	706	36	GTTTCC	161	TTTCT	229	GTAA	443
12	GGAAAT	240	TTCCA	391	AACA	684	37	TGTCCA	160	GAAAG	227	ACCA	437
13	ATTTCC	235	CCTGG	391	TCCA	664	38	CAAACA	159	GGAAT	226	CTTT	420
14	AGGAAA	226	TGTTT	385	GTTT	640	39	TCCCAG	156	GTTTC	225	CCTT	420
15	TAAACA	217	TCCTG	383	TGGA	639	40	TTCCCG	153	TTCTG	215	ACCT	420
16	CCATGG	214	CTGGA	374	TCCT	619	41	CGGGAA	153	GTCAG	215	GTCC	412
17	GTAAAC	203	TTCCC	372	ACAG	603	42	GGTCAG	151	ATTCC	215	TTTA	409
18	TTTCCT	200	TGGAA	372	GAAG	594	43	GGAAAC	148	GAAAC	214	TCAG	407
19	TTTCCA	192	GGACA	370	ATTT	571	44	CCAGAA	148	CATTT	213	CATG	406
20	TGGAAA	188	TCCAG	367	GGTC	568	45	GGAAAG	145	TTTTC	209	ATGG	405
21	AGGAAG	187	GACAG	352	CTTC	566	46	GAAATG	142	CAGGG	206	TAAA	399
22	CTGGGA	182	GAAAT	349	GGGA	565	47	GGAAGG	141	GTTTA	205	CCGG	394
23	TGACCT	180	TGTCC	348	CTGT	565	48	CCGGAA	141	TCTGG	204	TGGG	391
24	TTTCCC	176	ATTTC	335	AAAT	554	49	CATGGT	140	CCCAG	204	TTAC	381
25	GGGTCA	175	AGGTC	333	AAGG	550	50	GCAAAC	136	TGGGA	201	AATG	375

Chapter 4

Conclusions

4. Conclusions

It is generally believed that eukaryotic promoters are complex and difficult to characterize. And there are no universal or conserved core promoter elements in eukaryotes as that of prokaryotic promoters at definite positions like: –10 (TATAAT) and –35 (TTGACA) elements around the TSS. Computational and functional studies in humans suggest that promoters that contain TATA box account for only 10-20% of protein-coding genes. This may infer that TATA box is not a conserved or general promoter element in eukaryotes. Recent reports suggest that miRNAs, group of small RNA molecules, are known to regulate about 30% of genes. It is clear that there will be other elements that can regulate the gene expression. We perhaps suggest that ~50% of the genes can be directly regulated by transcription factors. Broadly speaking, transcription factors are the rate determining factors, which can regulate the synthesis of a particular protein of interest at transcriptional level.

In this context, we have focused on recognition of promoter elements by transcription factors (in sequence terms) during initial stages of transcription. Earlier studies have targeted the transcription target sites (TSS) to identify promoter elements, but the signal is too weak because of genome complexity. In addition to TSS regions, there are some GC rich regions that are 6-8 nucleotides in length that can be easily recognized by transcription factors. The present study focused on identifying GC rich sequences in human promoter sequences.

Using conventional studies, we have identified a set of relatively conserved GC rich 6-nt sequences around transcription start site (TSS) of human promoter sequences. The top 50 6-nt sequences in the present study account for more than 50% of the promoter sequences in the database with relatively multiple occurrences (more than 10 occurrences); many of these 50 6-nt sequences occur

multiple times in several promoters suggesting that they may provide strong signals for binding of TFs.

From 6-nt sequence distributions in promoters, we conclude that transcription factors recognise 6-nt sequences that are present on downstream of TSS also (fig 3.2). Sigmoidal distribution of 6-nt sequences on both sides of TSS suggests that there exists some internal co-operation in 6-nt sequence distribution around TSS. This suggests that transcription factors can straddle on both sides of TSS during initial stages of promoter recognition in transcription process (fig 4.1). However, transcription factors have to be removed for continuation of transcription (for unwinding of the strands). This clearly suggests that transcription process is much more complex than presently known. In addition, the complexity in transcription process can be enhanced if there are mutations in downstream (coding region) of TSS, which may severely affect the binding specificity of transcription factors, even though mutations may prevent the codon table of protein synthesis.

Identification of the complementary, reverse and reverse complementary sequences for a particular 6-nt sequence suggests that TFs recognize the promoter elements that are present on opposite strand irrespective of strand orientation (fig 4.1). This clearly suggests that both the strands of DNA are playing important role in recognition of promoter elements during initial stages of the transcription. This is well supported with X-ray structures of dimeric T-domain Brachyury transcription factor bound to palindromic duplex DNA (Muller and Herman, 1997). We intend that identification of these common subsequences (binding sites) by particular TF on *ds*DNA is a crucial step in promoter recognition which determines the strand specific direction of transcription that initiates sequential binding of other TFs and RNA pol II, that form pre initiation complex (PIC).

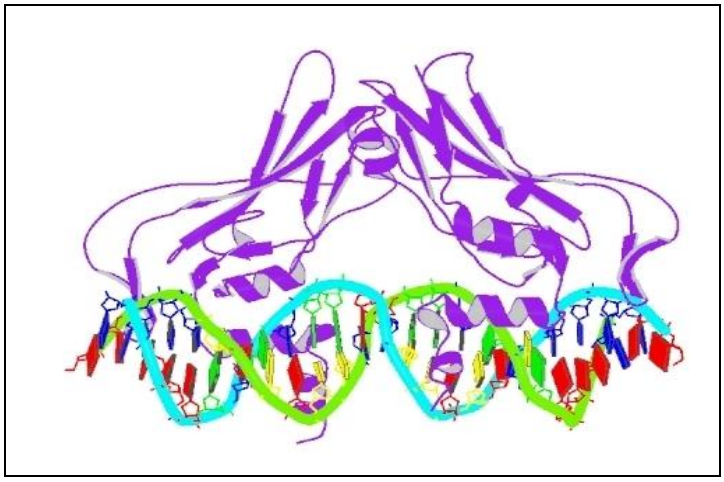


Fig 4.1: Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor. The X-ray structure shows that the T domain is bound as a dimer to the 24-nucleotide palindromic DNA, with the dimer interface lying above the minor groove. The dimer forms a large arc which span over the DNA containing palindromic binding sites for T-domain and allows each monomer to contact its 20-bp DNA-recognition sequence. This picture is in conjecture with our results suggesting that TFs can straddle on both sides of the TSS during promoter recognition, and binding of each T-domain monomer to the palindromic binding sites suggest that both the DNA strands play an important role in promoter recognition during initial stages of transcription. Picture is adopted from-http://www.pdb.org/pdb/explore/explore.do?structureId=1XBR with literature evidence from Müller and Herrmann (1997).

Positional distributions of 6-nt sequences in promoters suggest that actual physical location of these signals is not critical and the promoter recognition elements are within the neighborhood of TSS. We note that the optimal position for these GC rich signals is -30 to -50 from TSS and these signals also present in downstream of TSS. It was evident that TATA elements are not abundant within –100 region from TSS in the promoter data.

From miRNA results, we noted that few of the 6-nt sequences were not identified in the miRNA datasets. We predict that all miRNA are not involved in promoter

recognition, as many miRNA are known to play important role in various biological functions such as tissue differentiation, cell cycle regulation, signal transduction, etc. We can also say that these 6-nt sequences (that are not found) may not really represent valid promoter elements or it is also possible that the role of miRNAs is rather restricted to selected promoters. From our results, we stress that miRNA are playing some significant role in recognition of promoter elements during initial stages of transcription. In addition, the distribution of the 6-nt frequencies in mature miRNA data does not appear to be random. Further, we do not expect the role of all miRNAs in promoter recognition. However, this apparently takes a considerable number of miRNAs, as the database has been growing.

Based on common 6-nt sequences in both miRNA and promoter sequences, we suggest that a group of promoters/genes can be expressed or regulated by miRNA at promoter recognition level during transcription. In conjunction to our results, it was reported that there were potential miRNA targets in the gene promoters and they showed that promoters are strong candidates for miRNA regulation compared to 3'-UTRs. Based on the positional frequency of 6-nt sequences and transcription factor binding sequences in miRNA, we suggest that 5' ends of miRNA are more favorable in recognition of promoter elements. We presume that miRNA can recognize a group of promoters and up regulate or down regulate their expression during initial stages of transcription.

TFBS results revealed that 6-nt sequences were identified in ~26% of TFs in the JASPAR database. Distribution of TFBS sequences in promoter data showed redundancy in their distribution. We may perhaps suggest that the redundancy might be due to presence of alternative promoters/genes and that will be tightly regulated by the TFs. The correlation between promoters, miRNA and TFs based on common 6-nt sequences suggest that a group of genes can be regulated by a

particular set of miRNA and TF. We intend that biological implication of these interactions can aid in development of new therapeutics for diseases such as cancers, which may perhaps target the transcriptional events efficiently.

Base composition within tfbs sequences shows uniform distribution of the four bases and dinucleotide distribution suggest the prevalence of purine rich sequences. GC/CG distributions showed lowest frequencies than expected frequencies. Among the subsequences (4-6 nt) we noted that 4-nts showed highest frequencies. Differences in frequency distribution of 5-nt and 6-nt sequences in TFBS suggest that these sequences do not appear to be random, might be conserved in terms of frequencies. Sequence features from promoters and transcription factor binding sites conclude that ~26% of TFs regulate the GC rich promoters. From TF binding site features, it was also noted that transcription factors exhibit broad recognition towards ATGC distribution.

The present study mainly focused on analysis of promoters and TFBS sequences which suggests that eukaryotic transcription is much more complex event than it is believed presently. The frequency distributions of subsequences in promoter and TFBS sequences show that these subsequences are not random in distribution. Studies on functional significance of these subsequences in consideration with promoter elements and other biological molecules/factors can aid in unraveling the combinatorial regulatory events and help in understanding the spatio-temporal and lineage specific gene expression by TFs.

References

1. Adams MD. (2005). Conserved sequences and the evolution of gene regulatory signals. *Current Opinion in Genetics & Development* 15: 628-633.
2. Akan P and Deloukas P. (2008). DNA sequence and structural properties as predictors of human and mouse promoters. *Gene* 410 (1-2): 165-176.
3. Anish R, Hossain MB, Jacobson RH and Takada S. (2009). Characterization of Transcription from TATA-less Promoters: Identification of a New Core Promoter Element XCPE2 and Analysis of Factor Requirements. *PLoS ONE* 4(4): e5103.
4. Antequera F and Bird A. (1993). Number of CpG islands and genes in human and mouse. *PNAS* 90:11995-11999.
5. Babenko VN, Kosarev PS, Vishnevsky OV, Levitsky VG, Basin VV and Frolov AS. (1999). Investigating extended regulatory regions of genomic DNA sequences. *Bioinformatics* 15 (7/8): 644-653.
6. Bajic VB, Choudhary V, Hock CK. (2004). Content Analysis of the Core Promoter Region of Human Genes. *In Silico Biology* 4: 109-125.
7. Bajic VB and Seah SH. (2003). Dragon Gene Start Finder: An Advanced System for Finding Approximate Locations of the Start of Gene Transcriptional Units. *Genome Res.* 13: 1923-1929.
8. Bajic VB, Tan SL, Christoffels A, Schönbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C, Kawai J, Hume DA, Carninci P

and Hayashizaki Y. (2006). Mice and Men: Their Promoter Properties. *PLoS Genetics* 2(4): 614-626.

9. Baldauf SL. (2003). The Deep Roots of Eukaryotes. *Science* 300(13): 1703-1706.
10. Buchan JR and Parker R (2007). The Two Faces of miRNA. *Science* 318(21): 1877-78.
11. Bucher P and Trifonov EN. (1986). Compilation and analysis of eukaryotic POL II promoters. *Nucleic Acids Res.* 14(24): 10009-10026.
12. Butler JEF and Kadonaga JT. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. *GENES & DEVELOPMENT* 16: 2583-2592.
13. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engstrom PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA and Hayashizaki Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38: 626-635.
14. Carthew RW (2006). Gene regulation by miRNAs. *Current Opinion in Genetics & Development* 16: 203-208.

15. Cui Q, Yu Z, Pan Y, Purisima EO, Wang E. (2007). MicroRNAs preferentially target the genes with high transcriptional regulation complexity. *Biochemical and Biophysical Research Communications* 352: 733-738.
16. Chuang JC and Jones PA. (2007). Epigenetics and MicroRNAs. *Pediatric Research* 61 (5). DOI: 10.1203/pdr.0b013e3180457684.
17. Davuluri RV, Grosse I, Zhang MQ. (2002). Computational identification of promoters and first exons in the human genome. *Nat. Genet.* 29(4): 412-7.
18. Dinger ME, Amaral PP, Mercer TR and Mattick JS. (2009). Pervasive transcription of the genome: functional indices and conceptual implications. *Briefings in Functional Genomics and Proteomics* 8(6): 407-23.
19. Fickett JW and Hatzigeorgiou AG. (1997). Eukaryotic Promoter Recognition. *Genome Res.* 7: 861-878.
20. FitzGerald PC, Shlyakhtenko A, Mir AA and Vinson C. (2004). Clustering of DNA Sequences in Human Promoters. *Genome Res.* 14: 1562-1574.
21. Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, and Sandelin A. (2008). A code for transcription initiation in mammalian genomes. *Genome Res.* 18: 1-12.

22. Fukue Y, Sumida N, Nishikawa Jun-ichi and Ohyama T. (2004). Core promoter elements of eukaryotic genes have a highly distinct mechanical property. *Nucl. Acids Res.* 32(19): 5834-5840.
23. Gershenzon NI and Ioshikhes IP. (2005). Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* 21: 1295-1300.
24. Griffiths-Jones S, Saini HK, van Dongen S, and Enright AJ. (2008). miRBase: tools for miRNA genomics. *Nucleic Acids Res.* 36: D154-D158.
25. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR and Bateman A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. 33: D121-D124.
26. Gross P and Oelgeschläger T. (2006). Core promoter-selective RNA polymerase II transcription. *Biochem Soc Symp.* 73:225-36.
27. Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, Lazarevic D, Hayashizaki Y and Carninci P. (2006). The complexity of the mammalian transcriptome. *J. Physiol.* 575.2: 321-332.
28. He L and Hannon GJ. (2004). MicroRNAs: Small RNAs with a big role in Gene Regulation. *Nature Reviews Genetics* 5: 522-531.
29. Herr AJ, Jensen MB, Dalmay T. and Baulcombe DC. (2005). RNA polymerase IV directs silencing of endogenous DNA. *Science* 308:118-120.

30. Hirose Y and Manley JL. (2000). RNA polymerase II and the integration of nuclear events. *Genes Dev.* 14: 1415-1429.
31. Hobert O. (2008). Gene Regulation by Transcription Factors and MicroRNAs. *Science* 319(28): 1785-1786.
32. Hrabcová I and Kypr J. (2008). The Longest (A+T) and (G+C) Blocks in the Human and other Genomes. *Journal of Biomolecular Structure & Dynamics* 25 (4): 337-346.
33. Iwama H, Murao K, Imachi H and Ishida T. (2011). MicroRNA Networks Alter to Confirm to Transcription Factor Networks Adding Redundancy and Reducing the Repertoire of Target Genes for Coordinated Regulation. *Mol Biol. Evol.* 28(1): 639-646.
34. John B, Enright AJ, Aravin A, Tuschl T, Sander C and Marks DS. (2004). Human MicroRNA Targets. *PLoS Biol.* 2(11): e363.
35. Juven-Gershon T, Hsu JY and Kadonaga JT. (2006). Perspective on the RNA polymerase II core promoter. *Biochemical Society Transactions* 34(6): 1047-1050.
36. Karlin S and Landunga I. (1994). Comparisons of eukaryotic genomic sequences. *PNAS* 91: 12832-12836.
37. Kim DH, Sætrom P, Snøve O Jr, and Rossi JJ. (2008). MicroRNA-directed transcriptional gene silencing in mammalian cells. *PNAS* 105(42): 16230-16235.

38. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD and Ren B. (2005). A high-resolution map of active promoters in the human genome. *Nature* 436(7052): 876-880.
39. Kim VN. (2005). MicroRNA Biogenesis: Coordinated Cropping and Dicing. *Nature Reviews* 6: 376-385.
40. Knudsen S. (1999). Promoter2.0: for the recognition of Pol II Promoters. *Bioinformatics* 15 (5): 365-361.
41. Kravchenko JE, Rogozin IB, Koonin EV and Chumakov PM. (2005). Transcription of mammalian messenger RNAs by a nuclear RNA Polymerase of mitochondrial origin. *Nature* 436: 735-739.
42. Kusenda B, Mraz M, Mayer J, Pospisilova S. (2006). MicroRNA: Biogenesis, Functionality and Cancer Relevance. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub.* 150(2): 205-215.
43. Lai EC. (2005). miRNAs: Whys and Wherefores of miRNA-Mediated Regulation. *Current Biology* 15(12): R458-460.
44. Lee I, Ajay SS, Yook JI, Kim HS, Hong SH, Kim NH, Dhanasekaran SM, Chinnaiyan AM and Athey BD. (2009). New Class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome Research* 19:1175-1183.
45. Lee RC, Feinbaum RL, Ambros V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75(5): 843-54.

46. Lee Y, Jeon K, Lee JT, Kim S and Kim VN. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.* 21(17): 4663-4670.
47. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, and Burge CB. (2003). Prediction of Mammalian MicroRNA Targets. *Cell* 115: 787-798.
48. Littlefield O. and Nelson HC. (1999). A new use for the wing of winged helix-turn-helix motif in HSF- DNA cocrystal. *Nat. Struct. Biol*. 6:464-470.
49. Liu YZ, Yang YC and Wang TM. (2007). Characteristic Distribution of L-tuple for DNA Primary Sequence. *Journal of Biomolecular Structure & Dynamics* 25(1): 85-91.
50. Lizabeth AA. (2007). Fundamental Molecular Biology, Wiley-Blackwell publishers.
51. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR and Golub TR. (2005). MicroRNA expression profiles classify human cancers. *Nature* 435(9): 834-838.
52. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W and Cui Q. (2008). An Analysis of Human MicroRNA and Disease Associations. *PLoS ONE* 3(10): e3420 October.

53. Martianov I, Viville S and Davidson I. (2002). RNA Polymerase II Transcription in Murine Cells Lacking the TATA Binding Protein. *Science* 298: 1036-1039.

54. Mor E, Cabilly Y, Goldshmit Y, Zalts H, Modai S, Edry L, Elroy-Stein O and Shomron N. (2011). Species-specific microRNA roles elucidated following astrocyte activation. *NAR* doi: 10.1093/nar/gkq1325.

55. Mount DW. (2004). Bioinformatics: Sequence and Genome Analysis, 2nd edition, Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.

56. Müller CW and Herrmann BG. (1997). Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor. *Nature* 389(6653): 884-888.

57. Onodera Y, Haag JR., Ream T, Nunes PC, Pontes O, Pikaard CS. (2005). Plant Nuclear RNA Polymerase IV Mediates siRNA and DNA Methylation-Dependent Heterochromatin Formation. *Cell* 120: 613-622.

58. Pedersen AG, Baldi P, Chauvin Y and Brunak S. (1999). The biology of eukaryotic promoter prediction. *Computers & Chemistry* 23: 191-207.

59. Périer RC, Junier T and Bucher P. (1998). The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.* 26: 353-357.

60. Pillai RS. (2005). MicroRNA function: Multiple mechanisms for a tiny RNA. *RNA* 11: 1753-1761.

61. Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y and Sandelin A. (2007). Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biology* 7(8): R78.
62. Prymula K and Roterman I. (2009). Functional Characteristics of Small Proteins (70 Amino Acid Residues) Forming Protein-Nucleic acid Complexes. *Journal of Biomolecular Structure & Dynamics* 26(6): 663-895.
63. Putta P and Mitra CK. (2010). Conserved Short Sequences in Promoter Regions of Human Genome. *Journal of Biomolecular Structure & Dynamics* 27(5): 599-610.
64. Reddy AD, Prasad BVLS and Mitra CK (2006). Functional classification of transcription factor binding sites: Information content as a metric. *Journal of Integrative Bioinformatics* 3(1):20.
65. Reddy AD, and Mitra CK. (2006). Comparative analysis of core promoter region: Information content from mono and dinucleotide substitution matrices. *Genomics, Proteomics and Bioinformatics* 4(3): 189-195.
66. Rekha TS and Mitra CK. (2007). Frequency Analysis of Splice Site Regions in Different Organisms. *Journal of Integrative Bioinformatics* 4(2): 85.
67. Rossi JJ. (2007). Transcriptional activation by small RNA duplexes. *Nature Chemical Biology* 3(3): 136-137.

68. Sandelin A, Alkema W, Engström P, Wasserman WW and Lenhard B. (2004). JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32: D91-D94.
69. Sandelin A, Carninci P, Lenhard Boris, Ponjavic J, Hayashizaki Y and. Hume DA. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics* 8: 424- 436.
70. Scherf M, Klingenhoff A and Werner T. (2000). Highly specific Localization of Promoter Regions in Large Genomic Sequences by PromoterInspector: A Novel Context Analysis Approach. *J. Mol. Biol.* 297: 599-606.
71. Schmid CD, Perier R, Praz V and Bucher P. (2006). EPD in its twentieth year: Towards complete coverage of selected model organisms. *Nucleic Acids Res*. 34:D82-5
72. Shelenkov A and Korotkov E. (2009). Search of regular sequences in promoters from eukaryotic genomes. *Computational Biology and Chemistry* 33: 196-204.
73. Sidow A (2002). Sequence First.Ask Questions Later. *Cell* 111:13-16.
74. Solovyev VV and Shahmuradov IA. (2003). PromH: promoters identification using orthologous genomic sequences. *Nucleic Acids Research* 31 (13): 3540-45.
75. Smale ST (2001). Core Promoters: active contributors to combinatorial gene regulation. *Genes Dev.* 15: 2503-2508.

76. Smith TF and Waterman S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147 (1): 195-197.

77. Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A., Sakaki, Y., Morishita, S., Okubo, K., Sugano, S. (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.* 11: 677-684.

78. Thomas MC and Chiang CM. (2006). The General Transcription Machinery and General Cofactors. *Critical Reviews in Biochemistry and Molecular Biology* 41:105-178.

79. Tokusumi Y., Ma Y., Song X., Jacobson R. H., and Takada S. (2007). The New Core Promoter element XCPE1 directs activator, mediator and TATA-binding protein-dependent but TFIID-independent RNA Polymerase II transcription from TATA-less Promoters. *Molecular and Cellular Biology* 27(5): 1844-1858.

80. Vasudevan S, Tong Y and Steitz JA. (2007). Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation. *SCIENCE* 318: 1931-1934.

81. Werner T. (2003). The state of the art of mammalian promoter recognition. *Brief. Bioinform.* 4(1): 22-30.

82. Werner T. (1999). Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome* 10:169-175.

83. Wu SY and Chiang CM. (2001a). TATA-binding protein-associated factors enhance the recruitment of RNA polymerase II by transcriptional activators. *J. Biol. Chem.* 276: 34235-34243.

84. Wu SY and Chiang CM. (1998). Properties of PC4 and an RNA polymerase II complex in directing activated and basal transcription *in vitro*. *J. Biol. Chem.* 273: 12492-12498.

85. Wu WH and Hampsey M. (1999). An activation-specific role for transcription factor TFIIB *in vivo*. *PNAS* 96: 2764-2769.

86. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. (2005). Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. *Nature* 34: 338-345.

87. Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K and Abe T. (2007). Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* 8(67). doi:10.1186/1471-2164-8-67.

88. Yang C, Bolotin E, Jiang T, Sladek FM and Martinez E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* 389: 52-65.

89. Younger ST, Pertsemlidis A and Corey DR. (2009). Predicting potential miRNA target sites within gene promoters. *Bioorganic & Medicinal Chemistry Letters* 19: 3791-3794.
90. Zhang B, Pan X, Wang Q, Cobba GP and Anderson TA. (2006). Computational identification of microRNAs and their targets. *Computational Biology and Chemistry* 30(6): 395-407.
91. Zhang MQ. (2007). Computational analysis of eukaryotic promoters. *BMC Bioinformatics* 8(6): S3. doi: 10.1186/1471-2105-8-S6-S3.
92. Zhang MQ. (1998). Identification of human gene core promoters in silico. *Genome Res.* 8:319-326.
93. Zheng WX and Zhang CT. (2008). Biological Implications of Isochore Boundaries in the Human Genome. *Journal of Biomolecular Structure & Dynamics* 25(4): 327-336.

Publications, Conferences, Workshops, Oral and Poster Presentations

Publications

1. Padmavathi Putta and Chanchal K. Mitra. (2010). Conserved Short Sequences in Promoter Regions of Human Genome. *Journal of Biomolecular Structure & Dynamics* 27(5): 599-610.
2. Padmavathi Putta and Chanchal K. Mitra (2011). Sequences Analysis of Human Transcription Factor Binding Site Sequences. (Manuscript under preparation).

Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 27, Issue Number 5, (2010)

Conserved Short Sequences in Promoter Regions of Human Genome

http://www.jbsdonline.com

Padmavathi Putta
Chanchal K. Mitra*

Department of Biochemistry,
University of Hyderabad,
Hyderabad - 500 046, India

Abstract

Recognition of promoter elements by the transcription factors is one of the early initial and crucial steps in gene expression and regulation. In prokaryotes, there are clear signals to identify the promoter regions like TATAAT at around –10 and TTGACA at –35 positions from transcription start site (TSS). In eukaryotes the promoter regions are structurally more complex and there are no conserved or consensus sequences similar to the ones found in prokaryotic promoters.

We have located a set of GC rich short sequences (<8 nt) that are relatively common in human promoter sequences around the TSS (±100 relative to TSS). These sequences were sorted based on their frequency of occurrence in the database and the most common 50 sequences were used for further studies. Sigmoidal behavior of the high end of the frequency distribution of these sequences suggests presence of some internal co-operativity. These short sequences are distributed on both sides of TSS, suggesting that probably the transcription factors recognize these sequences on both upstream and downstream of TSS. As eukaryotic promoters lack any conserved sequences, we expect that these short sequences may help in recognition of promoter regions by relevant transcription factors prior to the initiation of transcription process. We postulate that a cluster of genes with common short sequences in the promoter region can be recognized by a particular transcription factor. We also found that most of these short sequences are fairly common within miRNA (both mature and stem-loop sequences). Our studies indicate that eukaryotic transcription is more complex than currently believed.

Introduction

In prokaryotes, there are clear signals within the promoter regions like TATAAT at around -10 and TTGACA at -35 (from TSS). In eukaryotes, the genome consists of introns, exons and promoters and other functional regions (5). The signals in the core promoter region are often fuzzy and difficult to decipher. It is usually believed that there is no universal or conserved core promoter sequence in eukaryotes (9). The eukaryotic promoters are structurally more complex and therefore need to have more complicated way for transcription.

The computational identification of promoter regions and their functional evaluation is an important task in bioinformatics and computational biology. Earlier studies have targeted the TSS as the signal for the recognition of the promoter regions. This works well with prokaryotes but does not work well in eukaryotes, as the signals are very weak or absent. The TSS plays a relatively minor role in the whole transcription process in eukaryotes. In an earlier work, we have reported on the collective behavior of the promoter sequences using information content (4). This approach is useful only to locate gross features but lacks any key details.

*Phone: + 91 40 2313 4668
Fax: + 91 40 23130120
E-mail: c_mitra@yahoo.com

Recent bioinformatics studies have revealed that many mammalian genes do not conform to the simple model in which a TATA box directs transcription from a single defined nucleotide position. Many genes have multiple promoters (within a given region), within which there are multiple start sites (multiple TSS within the same region), and that 72% of human promoters are associated with CpG islands (1, 12, 36). It has been reported that the majority of human RNA polymerase II (RNA pol II) binding sites within the promoter region have an array of closely located transcriptional start sites (TSSs) that are spread over 50–100 bp (13). Broad TSS distributions (''dispersed'' TSSs) are correlated with CpG islands and ubiquitously expressed genes, whereas promoters with a narrow TSS distribution frequently direct tissue-specific genes and often have a TATA box (7). The frequency of TATA box containing promoters among human protein-coding genes is estimated to be 10–20% (12, 34, 36).

It has been well established that mammalian genomes are composed of large sequence segments with fairly homogeneous GC content, namely isochores, which have been linked to many biological functions (39). The AT and GC distribution in the genome confers specific biophysical properties on DNA that are likely to influence genome folding in the nuclei and other functional properties (15). The AT and GC pairs contribute very different properties on the DNA, like thermal stability, conformation, flexibility and binding of various molecules in the major and minor double helix grooves and also in generation of various DNA structures. The GC distribution is extremely significant because it has a phylogenetic meaning and because it even decides the codon usage choice in genes and about the amino acid composition of the encoded proteins (15). In our present work, we have identified the GC rich short sequences in the promoter regions around the TSS.

The DNA elements/ promoter/ nucleotide sequences that can be recognized by the protein complexes range from 5-15 bp long (11). We implicitly assume that a 6-nt sequence can be recognized by standard protein motifs without error. Longer sequences may have some degeneracy or redundancy or may be just error-prone (42). We choose the 6-nt sequences because, (i) the most common promoter recognition elements, TATAAT and TTGACA (-10 and –35 from TSS) in prokaryotes are 6-nucleotide in length (ii) most of the restriction enzymes recognize the DNA sequences with high accuracy are of 6-nt in length (iii) minimum length of the TFBS in JASPAR database are starting with greater than 5-nt sequences. Based on these common observations, we have focused our attention on 6-nt and 7-nt sequences. In this report we have located the common subsequences that may be shared on both sides of the TSS. We presume that these sequences act as strong signal for recognition and binding of the particular TFs initially and enable other factors to bind them sequentially.

Gene expression is largely regulated by the action of *trans*-factors on the *cis*-elements aligning on the regulatory regions of the genes. Among these *trans*-factors and the *cis*-elements, transcription factors (TFs) and their binding sites (TFBS) play important roles in regulation of gene expression. Recently, another group of molecules, namely microRNAs (miRNAs), have been found to regulate gene expression at the post-transcriptional (and translational) levels through base-pairing with target messenger RNAs (26). MicroRNAs are non-coding RNAs of ~22-nucleotides in length that are encoded in the chromosomal DNA and transcribed as longer stem-loop-like precursors. Recent computational studies indicate that approximately one third of human genes are potentially regulated by miRNAs and each miRNA on average could target more than 200 genes (26). There is increasing evidence suggesting that miRNAs play critical roles in many key biological processes, such as cell growth, tissue differentiation, cell proliferation, embryonic development, and apoptosis. It has also been reported that miRNA play important roles in cellular signaling networks, cross-species gene expression variation, and coregulation with transcription factors (21). Based on thermodynamic studies, Lee *et al.,* (14) have

reported that 5'-UTR motifs interact with the 3'-end of miRNA in all conservation categories. On the other hand, 3'-UTRs show interactions with miRNA only in the case of highly conserved 8-mers. Considering that the 3'-end of miRNA family members (interspecies) and those of some miRNAs across species differ, the 3'-end of miRNAs may contribute to the gene- or species specific target site recognition of the 5'-UTR. Prediction of miRNA targets revealed that 29% of miRNAs have unknown functions and 71% have diverse functions, such as DNA binding, transcriptional regulation, signal transducer and kinase activity (19).

We propose that miRNAs are playing some significant role in recognition of the promoter regions during initial stages of transcription, via the transcription factors. We have attempted to correlate the promoter sequences and miRNAs based on these common short sequences. We hope to establish a simple relation about the role of miRNA in recognition of promoter elements during initial stages of transcription.

Materials and Methods

Human promoter sequences were downloaded from SIB-EPD (29) (http://www.epd.isb-sib.ch/). This promoter database was curated with experimentally verified data. The sequences were extracted from +100 to –100 relative to TSS. We have downloaded these sequences in two different sets like: –100 to +1 (referred to as upstream region) and -1 to +100 (referred to as downstream region to TSS). For base composition and distribution studies, we have used the full sequence (–100 to +100: total of 201 nt length). A total of 1871 human promoter sequences were studied (full set for *Homo sapiens*).

We did a pairwise search for all subsequences of 6-nt (7-nt) length between these promoter sequences (separately for the upstream and downstream regions and also for the total 201 nucleotide sequence). We considered all the $n\cdot(n-1)/2$ (n is the number of sequences and $n\cdot(n-1)/2$ is the total number of possible pairs; in this case 1871*1870/2 = 1,749,385 pairs) pairs in this search. We next sort these sequences in lexical order (there are $4^6 = 4096$ possible 6-nt sequences). This gets all similar sequences together (adjacent to each other) that helps in counting in the next step. We counted all distinct sequences and noted their frequencies. This set of sequences was next sorted based on their frequency of occurrence. In the present study, we have selected the most common (top 50 according to the frequency) 50 sequences (6-nt and 7-nt sequences separately). These sequences occur at least 10 times in the promoter region (upstream and downstream regions considered separately). We use these 50 subsequences, searched for their relative positions and occurrence in the promoter sequences (upstream and downstream regions separately and jointly) and the result was used for the final plots.

We have downloaded miRNA sequences (both mature and stem loop miRNA datasets) from miRBase (30) (http://microrna.sanger.ac.uk/). The mature miRNA data contains 866 miRNA sequences (max. length: 27 and min. length: 17) and stem-loop miRNA data contains 695 miRNA sequences (max. length: 150 and min. length: 47). The two-miRNA datasets were searched for the relative positions and occurrence of the most common 50 6-nt sequences.

All the programs were written in C in gcc environment and run in the Linux operating system. The graphs were plotted using SigmaPlot.

Results and Discussion

The promoter sequences (Homo sapiens) from the SIB-EPD have 1871 sequences. We have studied the region –100 to +100 relative to TSS. Therefore each sequence has exactly 201 nucleotides. Out of the 1871•201 = 376,071 total nucleotides, we

had 605 bad characters (N: unknown base). The promoter regions are highly rich in GC composition as seen by the following base composition: A=64984 (17.30%), C = 114,744 (30.56%), G = 125530 (33.43%) and T = 70208 (18.70%). It is important to note that when we select "all promoters", we get a set of 4809 sequences (for all species) and the GC bias in the complete set (results not shown) is far less compared to what we see above. This automatically suggests that the conclusions we derive cannot be extended to other families and we should study them in similar fashion independently (each species should be studied separately).

Distribution of the individual nucleotides in the promoter regions is non- uniform. However, dinucleotide distribution (AT, TA, GC and CG) shows some interesting features (Figure 1). We observed that AT and TA frequencies were uniformly below the mean frequency values (except a peak around –25 from TSS) and GC and CG frequencies were above the mean values (except a negative peak at –25 and a broad hump around +25 from TSS). This suggests that there exists some signal in these regions. Similar features in the nucleotide distributions have been earlier reported by Babenkov *et al.,* (34) who plotted similar graphs for (A + T), (A + G) and (C + G). However, our results show clearly the hump present around +25 in the dinucleotide distribution (GC, CG) appears to be more sensitive.

The AT rich signal may be either functional or structural (or both) as experimental work has shown that many AT-rich sequences of 6-nt or longer can partially replace TATAAT-box motif in the proximity of other control elements (24). They suggest a strong correlation between the GC content and the structural features of DNA, which might be the result of either the dominant effect of the base content on structural parameters or the global DNA structures partially dictating the overall GC content. In our study, we found 6-nt sequences were enriched with GC content compared to AT content. However, we didn't focus on the relation between the GC content and structural features of the promoters in this work. However, other reports suggest that the TATAAT like signals are perhaps only 10-20% of the promoters of the protein coding genes (12, 34, 36).

To study the hexanucleotide distributions, we search the promoter sequences (–100 to +100 relative to TSS) using the most common 50 subsequences (8). We note that the most common set of 50 6-nt sequences are not exactly identical on the upstream and the downstream regions (Table I). Frequency of each subsequence in the promoter database was calculated based on two strategies: The non-redundant frequency represents the one time/single occurrence of each 6-nt sequence (multiple occurrence of the same subsequence in a given promoter sequence is counted as one) in the database and the count corresponds to the sum of total occurrence of each subsequence in all the promoter sequences (upstream and downstream separately). The raw frequencies are shown in the Figure 2. We note that sequences in the upstream regions are usually more common but only slightly. The shape of the distribution is intriguing as it reminds one of the typical sigmoidal curves (rotated). It perhaps has origin in some internal cooperative process. It is well known that several proteins are needed in many cells and are very common and many proteins are highly specialized and are less common. Therefore we expect some proteins to be synthesized at a fast rate whereas some others may be needed to be produced at a relatively slow rate. Fast rate of synthesis can only be achieved using a large number of transcription factors and corresponding binding sites (larger copy number of genes do not automatically translate to a larger rate of protein synthesis as the transcription factors and TFBS may become the rate limiting factor in the transcription process). The curve seen in Figure 2 corresponds and supports this hypothesis. We assume that the most common 50 6-nt sequences correspond to promoters (recognition elements) for more common genes.

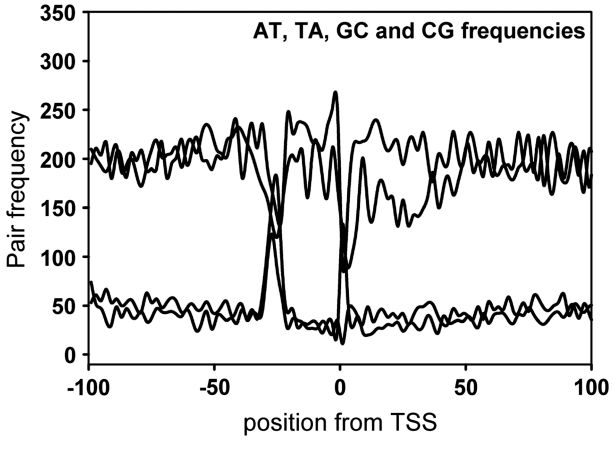


Figure 1: The four dinucleotides AT, TA, GC and CG frequencies have been plotted as a function of their position from the TSS. The 12 other dinucleotides have been omitted from this graph in order to improve the visual clarity. The lower graphs refer to the AT/TA distributions and the upper graphs refer to GC/CG distributions. We claim that the distributions in the region TSS –50 to TSS + 50 are anomalous. The very sharp spike at the TSS is caused by the start codon.

Table I

Frequencies of the 50 most common 6-nt sequences from upstream and downstream sides from TSS in the promoter sequences (+/– 100 with respect to TSS).

S.No	Upstream region	Frequency*	Frequency#	Downstream region	Frequency*	Frequency#
1	GGGCGG	624	970	GGCGGC	611	953
2	GGCGGG	611	953	GCGGCG	580	969
3	GGGGCG	580	969	GGTGAG	624	970
4	GCGGGG	551	945	CGGCGG	531	774
5	CCGCCC	527	779	GCCGCC	527	779
6	CCCGCC	531	774	GCGGCC	551	945
7	CGGGGC	507	748	TGGCGG	507	748
8	CGCCCC	516	747	CGCCGC	442	699
9	CCCCGC	497	812	GGCTGC	516	747
10	GCGGCG	463	579	CCGCCG	497	812
11	GGCGGC	461	629	GCTGGG	463	579
12	GGGAGG	453	590	ATGGCG	461	629
13	GCCCCG	446	538	GCTGCC	361	379
14	CGGCGG	442	540	CGCTGC	424	542
15	GGCGCG	442	699	GCTGCT	401	539
16	GCGCGG	435	618	GCCGGG	426	587
17	CGGCGC	437	516	GCTGCG	437	516
18	GCGCGC	426	587	GCGGAG	379	441
19	GGCCGG	431	683	CGCGGC	376	439
20	GGGGGC	429	517	GGCTGG	351	438
21	GGAGGG	424	500	GGCCGC	416	479
22	GCGGGC	424	542	CGGCGC	424	500
23	GCGCCG	416	479	CTGCTG	399	462
24	AGGCGG	411	495	GCCCGG	435	618
25	CGCGCC	406	496	CGGCCG	375	481
26	CGGAAG	402	515	GGCAGC	442	540
27	GCCGGG	404	531	CCCGGC	411	495
28	CGCGGC	405	484	TGCTGC	405	484
29	CCGCGC	395	478	GCGCGG	395	491
30	GGCCGC	401	539	GGCCGG	431	683
31	CCTCCC	400	482	GCCGCG	400	482
32	GGGGCC	400	501	GGAGCC	453	590
33	GCCCGG	399	462	GCAGCC	361	407
34	CGCGCG	395	491	GCGGCT	446	538
35	GCCGCG	395	571	CCGCGG	402	515
36	GGCGCC	385	481	CGGCTG	359	442
37	CCCGGC	385	474	CCGGGC	299	367
38	GGCGGA	379	441	GTGAGT	358	408
39	GCGGCC	383	438	CGGAGC	356	409
40	GAGGGG	375	481	CCCGCC	353	419
41	GCGGAG	376	439	GGGCCG	406	496
42	GGGGAG	378	454	GCCCGC	364	556
43	GGAGGC	368	428	GCAGCT	429	517
44	CCGGGC	362	420	CTCCTG	327	425
45	GGGCCG	367	433	GAGGAG	336	395
46	GCCCCC	361	379	GGCTCC	385	481
47	GAGGCG	359	442	CCGGCC	247	283
48	CGCCGC	364	556	CCGCGC	319	384
49	CCCTCC	361	407	CCCGGG	385	474
50	CGCGGG	358	408	CCCAGC	395	571

*Counting multiple occurrences of a given 6-nt sequence in a given promoter as one (non-redundant frequency).

#Counting multiple occurrences of a given 6-nt sequence in a given promoter with the actual number (total occurrence).

We have calculated total occurrence of each 6-nt sequence in the promoter sequences (upstream regions). The frequency of occurrence and their percentile distribution were represented in Figure 3. Figure 3a represents the distribution of non-redundant 6-nt frequencies (excluding the multiple occurrence of same 6-nt sequence in the promoter) and their percentile frequency. Figure 3b represents the distribution of total

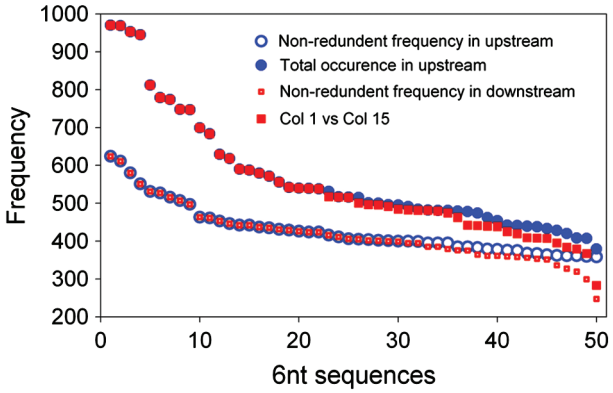


Figure 2: The frequencies of the 50 most common 6-nt sequences plotted after arranging the frequencies in descending order. The 6-nt sequences for the upstream and the downstream regions are not necessarily identical and the x-axis represents the simple ordinality. The graph shows typical sigmoidal shape suggesting some internal cooperative process. Non-redundant frequency means that we count multiple occurrences of the same 6-nt sequence in a given promoter as one.

occurrence of each 6-nt sequence in the promoter database. This clearly suggests the existence of the co-operative behavior with in the 6-nt sequences. However, origin of the co-operative behavior is unknown at present. Possibly more important genes will be associated with a higher frequency of the corresponding 6-nt sequence.

In an earlier work (10), functional classification of transcription factor binding sites based on a measure of information content reveals that diverse factors were placed closely together in the clusters. This suggests functional similarities in TFBS of these factors. This suggests that the binding sites (of these proteins) may be needed together (for functionality) and they may share the same transcription factor. It is apparent that a number of factors share a similar/common binding sites. For example, the TF SOX9 has the highest number (76) of TFBS in human (as reported in JASPAR). Similarly, other factors (*e.g.,* GATA3, MEF2) also have a relatively large number of TFBS candidates (63 and 58 respectively). We see however, that a total of 50 TFs has a total of 1377 TFBS with an average of 27 TFBS for a given TF (data taken from JASPAR) (2). This indirectly suggests that some genes having a larger number of TFBS are perhaps functionally more important (e.g., the relevant gene product may be needed at several points).

Out of the most common 50 subsequences (both 6-nt and 7-nt sequences, Table II), we observed the existence of a complementary, reverse and a reverse complementary

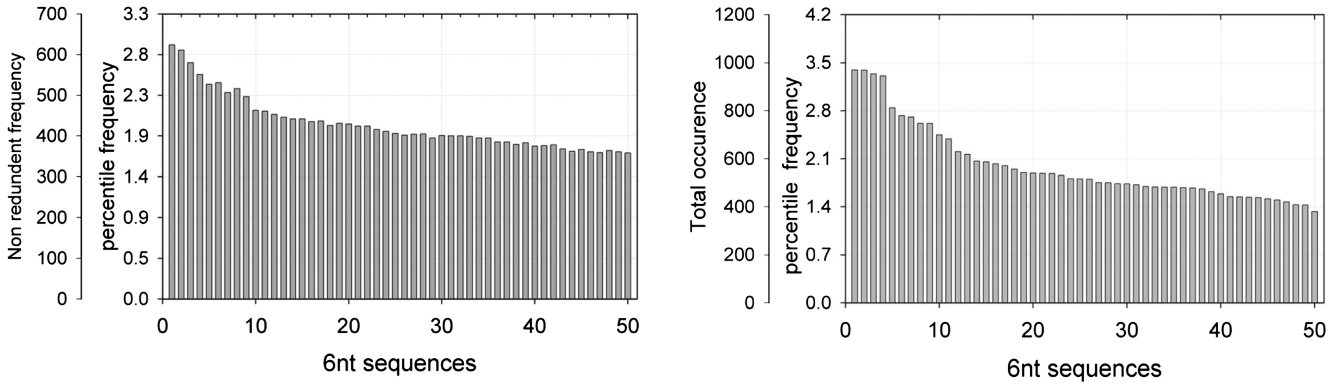


Figure 3: Frequency distributions of the most common 50 6-nt sequences from upstream region in the promoter database (-100 to +1 relative to TSS). We have plotted the percentile frequency and total occurrence (on y-axis with decreasing frequencies) of each subsequence (on x-axis doesn't follow the ordinality) in the upstream promoter sequences. We can observe the sigmoidal behavior in the distribution of the graphs. **3A** illustrates the non-redundant occurrence of the each 6-nt sequence in the promoter database and their corresponding percentile frequency. **3B** represents the total occurrence of each of the 6-nt sequence and their corresponding percentile distribution.

Table II

The frequency distribution of the first 50 most common 6-nt and 7-nt sequences in the promoter sequences (–100 to +100 relative to TSS). The roman numbers represent a set of reverse, complementary and reverse complementary sequences for a particular subsequence.

S.No	6-nt sequence	Frequency	7-nt sequence	Frequency
1	GGCGGG (i)	624	GGCGGGG (i)	364
2	GGGCGG (i)	612	GGGCGGG (ii)	358
3	GGCGGC (ii)	583	GGGGCGG (i)	350
4	GCGGCG (iii)	555	GCGGCGG (iii)	314
5	GCGGGG (iv)	531	CCCGCCC (ii)	299
6	CCGCCC (i)	531	GGCGGCG (iii)	298
7	CCCGCC (i)	519	GCGGGGC (iv)	294
8	GGGGCG (iv)	517	CCGCCCC (i)	285
9	CGGCGG (ii)	495	CGGCGGC (v)	282
10	GCGGCC (v)	466	CCCCGCC (i)	280
11	CGGGGC (vi)	459	GCGGCGC (xi)	232
12	GCGCGG (vii)	451	GCCCCGC (iv)	226
13	CGGCGC (viii)	444	GGCGGAG (vi)	203
14	GGCCGG (ix)	442	GAGGCGG (vi)	203
15	GCCGCC (ii)	442	GGGGGCG (xiii)	202
16	CCCCGC (iv)	441	CGCCGCC (iii)	200
17	GCCGGG (x)	437	GCCGCCG (v)	199
18	CGCCCC (iv)	436	GGCCGGG (vii)	198
19	CGCCGC (iii)	433	GCGGCCG	198
20	CGCGGC (viii)	432	GCGCGGG	197
21	GGCCGC (v)	425	CCGCCGC (iii)	196
22	GGGAGG (xi)	422	GGCGCGG (viii)	194
23	GCTGGG	411	GGCGGCC	192
24	GCGGAG	409	CGGGGCG (iv)	186
25	GCCGCG (viii)	409	TGGCGGC	183
26	GCGGGC	408	GGCGGGC (ix)	180
27	GGCGCG (vii)	407	GCGCGGC (x)	180
28	GCCCGG (xii)	406	GGGGAGG	178
29	GCGCCG (viii)	403	GCGCCGC (xi)	178
30	GCCCCG (vi)	402	GGGCCGG (vii)	173
31	CCGGGC (xii)	400	GGGAGGG	173
32	CCGCGC (vii)	398	GCTGCGG	173
33	GGCTGC	397	GGGCGGC	172
34	CCCGGC (x)	397	GCCGGGC	171
35	GCGCGC	387	CCGCGGC (xii)	170
36	CGCGCC (vii)	384	CCCGGCC (vii)	170
37	GGGCCG (x)	382	CGCCCCC (xiii)	169
38	GGGGCC	380	CGGGGCC (xiv)	168
39	GGCTGG	378	GGAGGCG	167
40	GGAGGG (xi)	376	CGGCGCG (x)	167
41	GGAGGC	374	GCGCGCC	166
42	CGGCCG (xiii)	371	CGCGGCC	166
43	GCCGGC (xiii)	368	CCGGAAG	166
44	CCGGCG (v)	367	GGCCCCG (xiv)	164
45	AGGCGG	366	GCGGGCG	164
46	GGTGAG	361	GGGGCCG	163
47	CCGCGG	361	GCCGCGG (xii)	161
48	CCGCCG (ii)	360	GGCCCGG	160
49	GGAGCC	359	GGCTGGG	159
50	CCCGGG	358	GCCCGCC (ix)	159

sequence for a particular 6-nt sequence. A set of 4 such sequences were numbered next to the corresponding sequences. Based on this criterion, we should expect 50/4 =12 sets from the 50 subsequences. In case of 6-nt sequences we have identified about 13 such sets (including partial set of sequences). In case of 7-nt sequences we have identified about 14 sets.

From these observations, we suggest the presence of these subsequences in both the strands. This clearly says that both the strands are playing an equivalent role in

the initial stages of transcription at the subsequence recognition (*i.e.,* at promoter recognition) level by the transcription factors irrespective of the DNA direction. Existence of the complementary, reverse and reverse complementary sequences for a given 6-nt sequence suggests that the promoter recognition takes place at the double stranded DNA level.

Positional distributions of the common 50 6-nts (from Table II) in the promoter sequences were presented in Figure 4. We have calculated the number of 6-nt sequences occurring in each position of the promoter sequences from +100 to –100

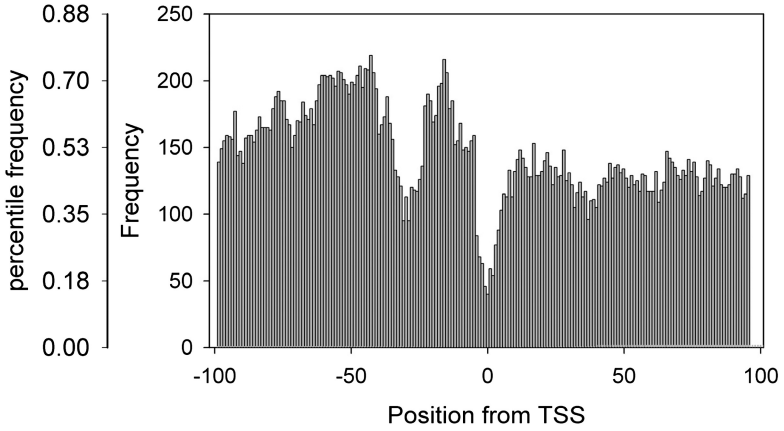


Figure 4: Positional distribution of the top 50 most common 6-nt sequences (Table II) within the promoter database. X-axis represents the position of the 6-nt sequences in the promoter database within ±100 relative to TSS. Y-axis corresponds to the number of 6-nt sequences occurring in particular position in the promoter database and their percentile frequency in the total database. We can clearly note a peak around –5 to –25 from TSS.

wrt to TSS. The deep cleft at position '0' is due to the TSS signal. We note a peak around –5 to –25 upstream to the TSS and a broad hump after –30 from TSS. This suggests that signals are more prominent within the –100 region from the TSS. At the same time we can also observe a broad hump in downstream regions from the TSS. This gives a clear idea that these signals also exist in the downstream regions of the promoter sequences and supports the 6-nt distribution shown in Figure 2. This in turn can be qualitatively correlated (and consistent with) to the dinucleotide distributions seen in Figure 1.

We have also checked the positional distribution of 6-nt sequences (from Table II) in two different sets. One set contains 26 sequences for which we have considered only the main 6-nt sequences (13 sets) and excluded the complementary, reverse and reverse complementary sequences for that particular sequence and also considered other sequences for which we did not find any complementary sequences (13 sequences). Another set consists of the remaining sequences including the sequences that do not contain any complementary sequences. This set contains 38 sequences. Both these sets were checked for positional distribution and their percentile frequencies were broadly similar to Figure 4.

We were interested to know whether these 6-nt sequences can be recognized by the miRNAs. The 50 6-nt sequences (Table II) were searched in the mature miRNA database. The distributions were given in Figure 5(left). We note that ~18% of the 6-nt sequences (23, 33, 35, 36, 39,43, 44, 46 and 48) were absent (not found) in the mature miRNA dataset. This may suggest perhaps these 18% 6-nt sequences do not really represent a valid promoter recognition element or it may also mean that the role of the miRNA is rather restricted to selected promoters. Out of the 866 mature miRNA sequences, 6-nt sequences are found to be 128 times (mean: ~1.7). 75 mature miRNA sequences have at least one of the 50 6-nt sequences. miRNAs

have other known functions and therefore it is not really expected that all the miRNAs will be found to contain the 6-nt sequences. The frequency distribution of 6-nt sequences in stem-loop miRNA dataset is shown in Figure 5(right). Out of the 695 stem-loop miRNA sequences, the 6-nt sequences are found to be present 644 times (with a mean of ~3.1). 203 sequences contain at least one of the 6-nt sequences. We note that ~8% of 6-nt sequences (23, 33, 39 and 46) were not identified in the stem-loop miRNA database. It is also possible that the miRNA database is incomplete and is likely to grow.

If we consider the average length of miRNAs as 20 (mature miRNA) and 100 (stem-loop), the ratio of average length of miRNA and the occurrence of 6-nt sequences (mean) in the miRNA database gives an indicative value for periodicity (how often one 6-nt sequence occurs). For mature miRNA data, the index value/ periodicity is about ~12 (*i.e.,* 20/1.7 = 11.76) and for stem-loop miRNA is about ~32 (100/3.1 = 32.2). These values indicate that a 6-nt sequence may be positioned/ occurring for every 12 and 32 residues in the mature and stem-loop miRNA sequences respectively. We note that number of (concentration of) 6-nt sequences is more in mature sequences compared to stem-loop sequences.

From the miRNA results, it is evident that most of the 6-nt sequences (except a few) can be recognized by miRNAs. We attempted to correlate these miRNAs with the promoter sequences based on the common 6-nt sequences. In Table III, we attempted to correlate the promoter sequences (IDs from the SIB-EPD database) and the 6-nt sequence (cols 1-3). The next three columns (cols 4-6) similarly show the mature miRNA IDs and the 6-nt sequence. Also shown the stem-loop miRNA IDs and the 6-nt sequence (cols 7-9). We have presented only the first 10 entries but it can be clearly seen that the same 6-nt sequence is shared by a number of promoters and miRNA sequences. From the Table, we observe that most of promoter sequences share the same 6-nt sequence (#) but a few miRNA (mature) sequences may recognize the same 6-nt sequence.

Discussion

It is believed that the eukaryotic promoter region is quite complex because of lack of consensus sequences. The TSS is targeted for the identification of the promoter regions, but the signal is too weak or totally absent because of the genome complexity in the eukaryotes. In addition to TSS regions, there may be some GC rich sequences in the promoter regions that are of 6–8 nt in length, which can be easily recognized by the transcription factors. We have identified the distribution of such GC rich 6-nt sequences on both upstream and downstream around the TSS. In humans, from earlier reports it is evident that TATA containing promoters account for 10–20% of the protein coding genes (12, 34, 36). Also reported that miRNA regulate about 30% of the genes (26). We may perhaps suggest that 50% of the genes are directly regulated by the TFs. In biological terms, TFs are the rate determining factors, which can regu-

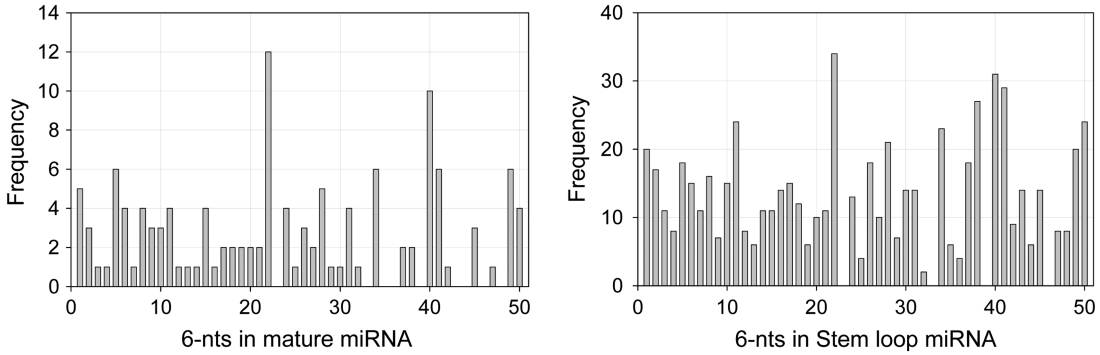


Figure 5: Distributions of 50 6-nt sequences in the mature miRNA and stem-loop miRNA datasets. The 6-nt sequences ordinate correspond to the Table II. We note the adsence of few 6-nt sequences in the miRNA database. See text for details.

late the synthesis of a particular protein of interest at the transcription level. So we have focused on the initial stages of transcription (at sequence level), which involves sequence-specific binding of TFs to the promoter regions.

Table III
The promoter and the miRNA sequences sharing a common 6-nt sequence

Promoter ID	#	6-nt sequence	Mature miRNA ID	#	6-nt sequence	Stem-loop miRNA ID	#	6-nt sequence
EP17036 (+) Hs snRNA U2	1	GGCGGG	hsa-miR-1915 MIMAT0007892	1	GGCGGG	hsa-mir-33b MI0003646	1	GGCGGG
EP49001 (+) Hs histone H1t	1	GGCGGG	hsa-miR-1228* MIMAT0005582	1	GGCGGG	hsa-mir-92b MI0003560	1	GGCGGG
EP15024 (+) Hs histone H33	1	GGCGGG	hsa-miR-638 MIMAT0003308	1	GGCGGG	hsa-mir-135a-1 MI0000452	1	GGCGGG
EP11074 (+) Hs histone H4-A1	1	GGCGGG	hsa-miR-1908 MIMAT0007881	1	GGCGGG	hsa-mir-203 MI0000283	1	GGCGGG
EP31007 (+) Hs HMG-14	1	GGCGGG	hsa-miR-663 MIMAT0003326	1	GGCGGG	hsa-mir-326 MI0000808	1	GGCGGG
EP31009 (+) Hs HMG-17	1	GGCGGG	hsa-miR-1228* MIMAT0005582	2	GGGCGG	hsa-mir-499 MI0003183	1	GGCGGG
EP33038 (+) Hs PRM2	1	GGCGGG	hsa-miR-638 MIMAT0003308	2	GGGCGG	hsa-mir-566 MI0003572	1	GGCGGG
EP37014 (+) Hs[rig] rp S15	1	GGCGGG	hsa-miR-1231 MIMAT0005586	2	GGGCGG	hsa-mir-615 MI0003628	1	GGCGGG
EP14031 (+) Hs b'-tubulin b'2	1	GGCGGG	hsa-miR-638 MIMAT0003308	3	GGCGGC	hsa-mir-636 MI0003651	1	GGCGGG
EP24039 (+) Hs vimentin	1	GGCGGG	hsa-miR-1915 MIMAT0007892	4	GCGGCG	hsa-mir-638 MI0003653	1	GGCGGG
EP33011 (+) Hs DES	1	GGCGGG	hsa-miR-1228* MIMAT0005582	5	GCGGGG	hsa-mir-639 MI0003654	1	GGCGGG
EP16038 (+) Hs fibronectin	1	GGCGGG	hsa-miR-1908 MIMAT0007881	5	GCGGGG	hsa-mir-663 MI0003672	1	GGCGGG

The numbers corresponding to this column correspond to the 6-nt sequences in Table II and the same sequence given in the next columns.

In this study, we have focused on the top most common 50 6-nt and 7-nt sequences. Of course there are more than 50 distinct TFBS and we note from the JASPAR database 1377 TFBS for *Homo sapiens*. This database is also growing and the actual number is likely to be more. Also, the SIB-EPD (experimentally curated) database is likely to be partial and incomplete. We decided to study the 50 most common sequences as (i) they have a relatively higher frequency of occurrence in the selected database (more than 10 occurrences); (ii) the top 50 sequences account for approximately 50% of the all the promoter sequences and (iii) Many of the top 50 sequences occur multiple times in several promoters, suggesting they may provide stronger signal.

From our studies, we conclude that the transcription factors can recognize the common subsequences that are of 6-nt in length that are present in downstream of TSS also (Figure 2). TFs also recognize the complementary sequence, its reverse and reverse complementary sequences (from Table II) that are present on the opposite strand irrespective of the strand orientation. This suggests that both the strands play an equivalent role in recognition of the promoter sequences (binding sites), during the initial stages of transcription. This is well supported in the literature and is consistent with the recent x-ray structures of the several transcription factors available. This suggests that the identification of these common subsequences (binding sites) by TFs (on dsDNA) is the crucial step in the initial stages of transcription, which favors the binding of other transcriptional machinery for the gene expression. Distribution of 6-nt sequences on both sides of TSS suggests that the transcription factors perhaps straddle on the TSS, enabling other factors to be bound sequentially. However, these transcription factors have to be removed from the downstream of TSS prior to the transcriptome complex formation and the strands must separate

subsequently for the transcription process to take place. This again suggests that the assembly of the transcriptosome complex is much more complex than presently known. We intend that initial recognition of promoter elements by the first transcription factor (which recognizes the promoter elements and favors other factors to form the pre-initiation complex) is the crucial step that determines the transcription direction (in the sense of which strand to be transcribed). Positional distribution studies suggest that the actual physical location is not very critical and the promoter recognition elements can be placed in the general neighborhood of the TSS. We note that the optimal position is around –25 and –50, although this is a not a very sharp peak.

From miRNA dataset results, we note that few 6-nt sequences (out of 50) were not recognized by the miRNAs. We predict that all miRNA are not involved in recognition of the promoter elements as miRNAs play critical role in other biological processes such as post-transcriptional regulation, cell proliferation, growth, *etc.,* It is also possible that the few 6-nt sequences we have located may not really be promoter recognition elements. It is also possible that the miRNA database is not complete. From our study, we stress that miRNAs are playing some significant role in the recognition of promoter sequences. We can correlate that a set of 6-nt sequences in promoter sequences can be recognized by miRNAs. In case of mature miRNA data set, we observe that there is a rapid decrease in the intensity of recognition of such peaks. This suggests that the distributions of the 6-nt sequences within the miRNA dataset are not random distributions. Statistically speaking, assuming uniform distribution, any given 6-nt sequence is expected to occur every $4^6 = 4096$ bases. For the mature miRNA, we have 866 sequences with a mean length of 20, and this can explain at most 4 occurrences of an arbitrary 6-nt sequence. The observed number is far larger and cannot be explained by randomness. Also, the distribution of the frequencies in the graphs (Figure 5) does not appear to be random. Further, we do not expect that all the miRNAs in the database to play a role in the recognition of the promoter site. However, this apparently takes a considerable number of miRNAs. In a recent publication (32), it was reported that there were potential miRNA targets in the gene promoters and they demonstrated that promoters are strong candidates for miRNA regulation compared to the 3'-UTRs, based on the minimum free energy and complementarity studies.

From Table III, we can safely conclude that the correspondence between the promoter sequences and the miRNAs are not exactly 1-to-1 but this also suggest other factors may play an important role. We postulate that miRNAs are playing a significant role in recognition of promoter elements and that may result in either expression or regulation of that particular gene of interest. We need to study the relation of the miRNA with the various factors in detail, which can provide a biological significance for these interactions at the initial stages of transcription in eukaryotes (so that a complete picture may be obtained). We believe that 6-nt sequences may help in clustering the factors based not on their structure or function, but on the sequence-specific binding nature of the factors.

Acknowledgements

One of the authors (PP) acknowledges financial support from DBT-CREBB.

References and Footnotes

1. A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizak, and A. David. *Nat Rev Genet 8*, 424-436 (2007).
2. Albin Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. *Nucleic Acids Res 32*, D91-D94 (2004).
3. A. G. Pedersen, P. Baldi, Y. Chauvin, S. Brunak. *Comput Chem 23*, 191-207 (1999).
4. Ashok Reddy D. and Chanchal K. Mitra, *Genomics, Proteomics and Bioinformatics 4*, 189-195 (2006).
5. S. L. Baldauf. *Science 300,* 1703-1706 (2003).

6. B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks. *PLoS Biol 2*, e363 (2004).
7. Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Pa¨r G Engstro¨m, Martin C Frith, Alistair R R Forrest, Wynand B Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M Grimmond, Christine AWells, Valerio Orlando, Claes Wahlestedt, Edison T Liu, Matthias Harbers, Jun Kawai, Vladimir B Bajic, David A Hume, and Yoshihide Hayashizaki. *Nat Genet 38*, 626-635 (2006).
8. C. K. Mitra and L. Milanesi. *Journal of Integrative Bioinformatics 5,* 103 (2008).
9. Chuhu Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez, *Gene 389(1)*, 52-65 (2007).
10. D. A. Reddy, B. V. L. S. Prasad, and C. K. Mitra. *Journal of Integrative Bioinformatics 3(1),* 20 (2006).
11. J. W. Fickett and A. G. Hatzigeorgiou. *Genome Res 7(9)*, 861-878 (1997).
12. P. C. FitzGerald, A. Shlyakhtenko, A. A. Mir, and C. Vinson. *Genome Res 14*, 1562-1574 (2004).
13. M. C. Frith, E. Valen, A. Krogh, Y. Hayashizaki, P. Carninci, and A. Sandelin. *Genome Res 18*, 1-12 (2008).
14. Inhan Lee, Subramanian S. Ajay, Jong In Yook, Hyun Sil Kim, Su Hyung Hong, Nam Hee Kim, Saravana M. Dhanasekaran, Arul M. Chinnaiyan, and Brian D. Athey. *Genome Res 19*, 1175-1183 (2009).
15. I. Hrabcová and J. Kypr. *J Biomol Struct Dyn 25*, 337-346 (2008).
16. J. Ponjavic, B. Lenhard, C. Kai, J. Kawai, P. Carninci, Y Hayashizaki, and A Sandelin. *Genome Biol 7(8)*, (2007).
17. E. Jennifer E., F. Butler, and J. T. Kadonaga. *Genes Dev 16,* 2583-2592 (2002).
18. K. Prymula and I. Roterman. *J Biomol Struct Dyn 26,* 663-895 (2009).
19. L. He and J.Hannon. *Nature Reviews Genetics 5,* 631 (2004).
20. O. Littlefield and H. C. Nelson. *Nat Struct Biol 6*, 464-470 (1999).
21. M. Lu, Q. Zhang, M Deng, J Miao, Y Guo, W. Gao, and Q. Cui. *PLoS ONE 3*, e3420 (2008).
22. D. Matthieu and T. H. Touzet. *BMC Bioinformatics 7*, 396 (2006).
23. M. Q. Zhang. *Genome Res 8*, 319-326 (1998).
24. P. Akan and P. Deloukas. *Gene 410*, 165-176 (2008).
25. P. Bucher and E. N.Trifonov. *Nucleic Acids Res 14*, 10009-10026 (1986).
26. Q. Cui, Z. Yu, Y. Pan, E. O. Purisima, and E. Wang. *Biochemical and Biophysical Research Communications 352,* 733-738 (2007).
27. R. Anish, M. B. Hossain, R. H. Jacobson, and S. Takada. *PLoS ONE 4,* e5103 (2009).
28. R. W Carthew. *Current Opinion in Genetics & Development 16,* 203-208 (2006).
29. R. C. Périer, V. Praz, T. Junier, C. Bonnard, and P. Bucher, *Nucleic Acids Res 26*, 353-357 (1998).
30. S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. *Nucleic Acids Res 36*, D154-D158 (2008).
31. S. Karlin and I. Landunga. *Proc Natl Acad Sci 91*, 12832-12836 (1994).
32. S. T. Younger, A. Pertsemlidis, and D. R. Corey. *Bioorganic & Medicinal Chemistry Letters 19*, 3791-3794 (2009).
33. T. Juven-Gershon, J.-Y. Hsu, and J. T. Kadonaga. *Biochemical Society Transactions 34*, 1047-1050 (2006).
34. Y. Tokusumi, Y. Ma, X. Song, and R. H. Jacobson. *Mol Cell Biol 27*, 1844-1858 (2007).
35. T. S. Rekha and C. K. Mitra. *Journal of Integrative Bioinformatics 4,* 85 (2007).
36. V. B. Bajic, S. L. Tan, A. Christoffels, C. Schönbach, L. Lipovich, L. Yang, O. Hofmann, A. Kruger, W. Hide, C. Kai, J. Kawai, D. A. Hume, P. Carninci, and Y. Hayashizaki. *PLoS Genetics 2*, 614-626 (2006).
37. V. N. Kim. *Nature Reviews Molecular Cell Biology 6,* 376-385 (2005).
38. V. N. Babenko, P. S. Kosarev, O. V. Vishnevsky, V. G. Levitsky, V. V. Basin, and A. S. Frolov. *Bioinformatics 15*, 644-653 (1999).
39. Wen-Xin Zheng and Chun-Ting Zhang. *J Biomol Struct Dyn 25,* 327-336, (2008).
40. X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K.Lindblad-Toh, E. S. Lander, and M. Kellis. *Nature 34*, 338-345 (2005).
41. Y-Z Liu, Y-C. Yang, and T-M Wang. *J Biomol Struct Dyn 25*, 85-91 (2007).
42. Y. Y Yamamoto, H. Ichida, M. Matsui, J. Obokata, T. Sakurai, M. Satou, M. Seki, K. Shinozaki, and T. Abe. *BMC Genomics 8,* (2007).

Date Received: May 24, 2009

Communicated by the Editor Ramaswamy H. Sarma

Conferences, Workshops, Poster and Oral Presentations

1. *Poster presentation* on "Modeling of Metabolic Pathways: Identification of Drug Targets" at BioQuest, University of Hyderabad in March 2008.
2. *Poster presentation* on "Structural Domain Prediction for Transcription factor Binding Sequences" in "13th Human Genome Meet - Genomics and Future Medicine" at HICC, Hyderabad from 27 -30th October, 2008.
3. Attended Workshop on "Computational Systems Biology Approaches to Analysis of Genome Complexity and Regulatory Gene Networks" at NUS, Singapore from 20-25th November 2008.
4. *Poster presentation* entitled "Study on Binding Sites in Promoter Regions of Human Genome" in "NCMB-2009: National Symposium on Cellular and Molecular Biophysics" by Indian Biophysical Society at CCMB, Hyderabad from 22-24th January 2009.
5. *Oral presentation* on "Sequence Analysis of Human Promoter regions" at BioQuest, University of Hyderabad in March 2010.
6. *Oral presentation* on "A study on Mycolic Acid Promoters in Identifying Pathways for New Drug Targets" in workshop "IB PAS-2010: Integrative Biological Pathway Analysis and Simulation" held in Bielefeld University, Germany from 21-26th May, 2010.

7. *Poster presentation* on "Signals in Human Promoter Regions" in "ECCB10- 9th European Conference on Computational Biology" held in Ghent, Belgium from 26-29th September 2010.