Computational Analysis of Gene Regulatory Elements using Mutual Information

A dissertation submitted for the Degree of
Doctor of Philosophy (Ph.D.)

by

Dinasarapu Ashok Reddy

Department of Biochemistry
School of Life Sciences
University of Hyderabad
Hyderabad-500046
India

Reg.No: 03LBPH03
(October 2006)

Dedicated to my Late Grand Parents

Sudireddy Nagi Reddy
&
Sudireddy AdiLakshmi

UNIVERSITY OF HYDERABAD

School of Life sciences

Department of Biochemistry

Certificate

This is to certify that the thesis entitled "**Computational analysis of gene regulatory elements using mutual information**" is based on the research work carried out by **Mr. Dinasarapu Ashok Reddy** in fulfillment for the degree of Doctor of Philosophy (Ph.D.) under my guidance. This work has not been submitted for any degree or diploma to any other university.

Supervisor	Head	Dean
	Department of Biochemistry	School of Life Sciences

UNIVERSITY OF HYDERABAD
School of Life sciences
Department of Biochemistry

Declaration

I here by declare that the thesis entitled "**Computational analysis of gene regulatory elements using mutual information**" has been carried out by under the supervision of **Chanchal K. Mitra** and that this work has not been submitted for any degree or diploma to any other university.

Date:

D. Ashok Reddy
(Reg. No: 03LBPH03)

Preface

Acknowledgments This was carried out in the Department of Biochemistry of the School of Life sciences at the University of Hyderabad in India. I thank all past and present colleagues for the good working atmosphere and the scientific-and some times maybe not so scientific-discussions.

Especially, I am grateful to my supervisor Chanchal K. Mitra for suggesting the topic, his scientific support, and the opportunity to write this thesis under his guidance, constant encouragement.

I would also like to thank our collaborators Dr Prasanth K. Panigrahi (Physical Research Laboratory, Ahmedabad) and Dr B.V.L.S Prasad, (Helix Genomics, Hyderabad) for their suggestions.

I would like to thank Prof. A.S. Raghavendra, Dean, School of Life Sciences, and Head, Dept. of Biochemistry Prof. M. Ramanatham for extending the School and Departmental facilities. I would like to thank all the faculty members, School of Life Sciences for their support. I thank the non-teaching staff for their help. I thank all my friends, my family members and relatives for their support.

Publications Parts of this thesis have been published and presented in the International Conference. Parts of results are published in *Computational Biology and Chemistry, Genomics, Proteomics & Bioinformatics* and *Journal of Integrative Bioinformatics*, respectively.

Figures This thesis reproduces few figures and tables in the introduction from other publications (mentioned their references) and from the Internet sources (www-site address has given).

Software All the computations were carried out by software developed in **C++** (LINUX). BLAST and DRAWGRAM of PHYLIP are also used.

Hyderabad, October 2006 **D Ashok Reddy**

Contents

1.Introduction
- 1.1 Genome structure, Organization and Composition
  - 1.1.1 Replication
  - 1.1.2 Transcription
  - 1.1.3 Translation
  - 1.1.4 RNA
  - 1.1.5 Proteins
  - 1.1.6 Nucleoside triphosphotes
  - 1.1.7 Human Genome Project (HGP)
- 1.2 Prokaryotic gene and transcription initiation
- 1.3 Eukaryotic gene and transcription initiation
  - 1.3.1 Structural motifs in transcription factors
- 1.4 Mitochondrial genome and transcription initiation
- 1.5 Information theory
- 1.6 Literature review
- 1.7 Biological Databases
- 1.8 Objectives

2. Materials and Methods
- 2.1 Databases
  - 2.1.1 Databases used in core promoter analysis
  - 2.1.2 Databases used in the TFBS Clustering
  - 2.1.3 Mitochondrial genome sequences
- 2.2 Methods
  - 2.2.1 Sequence Alignment
  - 2.2.2 Construction of nucleotide substitution matrices
    - 2.2.2.1 Neighbor-independent substitution matrices
    - 2.2.2.2 Neighbor-dependent substitution matrices
  - 2.2.3 Average mutual information content
  - 2.2.4 Standard error calculation

2.2.5 Information content calculation with example
2.2.6 BLAST (Basic Local Alignment Search Tool)
2.2.7 Clustering and Classification
2.2.7.1 UPGMA method
2.2.7.2 PHYLIP
2.2.8 C++ / LINUX
3. Results
3.1 Comparative analysis of core promoter region
3.1.1 Blocks of core promoter elements
3.1.2 Information content of core promoter elements
3.1.3 Discussion
3.2 Comparative analysis of TSS
3.2.1 Sequence Data
3.2.2 Information content of TSS
3.2.3 Discussion
3.3 Functional classification of TFBS
3.3.1 Information content of TFBS
3.3.2 Information content of random sequences
3.3.3 TFBS clustering
3.3.4 Discussion
4. Conclusions
5. List of publications
6. References

Abbreviations

TSS	Transcription Start Site
DPE	Downstream Promoter Elements
TF	Transcription Factor(s)
TFBS	Transcription Factor Binding Sites
bp	base pair
Kbp	kilo base pairs
Mbp	mega base pairs
nt	Nucleotide
TBP	TFIIB binding protein
PIC	Pre Initiation Complex
TAFs	TBP-associated factors
BLAST	Basic Local Alignment Search Tool
PAM	Point Accepted Mutation
BLOSUM	Blocks Substitution Matrix
PlantProm DB	Plant Promoter Database
EPD	Eukaryotic Promoter Database
PromEC	*E.coli* Promoter Database
SELEX	Systematic Evaluation of Ligands by Exponential enrichment
TRRD	Transcription Regulatory Regions Database
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
PHYLIP	Phylogeny Inference Package
mtDNA	Mitochondrial DNA
PSSM	Position Specific substitution matrix
E.coli	*Escherichia coli*
UTR	Untranslated Region
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
tRNA	transfer RNA
rRNA	ribosomal RNA
mRNA	messenger RNA

Introduction

1. Introduction

1.1 Genome structure, organization and composition

Deoxyribonucleic acid (DNA) contains the information needed to develop and direct the activities of nearly all-living organisms. DNA is a double stranded molecule and each DNA strand is made up of four chemical units called nucleotide bases. The bases are adenine (A), thymine (T), guanine (G) and cytosine (C) (Figure 1.1). Uracil (U) is a nucleotide base present in ribonucleic acid (RNA), in place of Thymine.

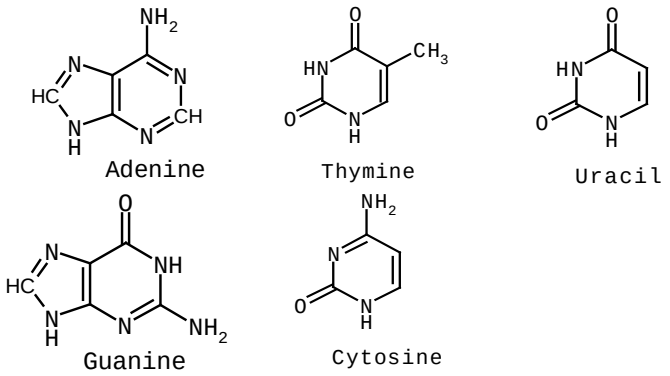


Figure 1.1. Chemical structures of Adenine, Thymine, Uracil, Guanine, and Cytosine (Note: Bond angles and bond lengths are not in scale).

The nucleotide bases can be divided into two chemical classes: purines (A and G, have two joined heterocyclic ring) and pyrimidines (C, U and T, have a single heterocyclic ring). A sugar molecule with an attached base is called a nucleoside. A nucleoside with a phosphate group attached to the sugar constitutes a nucleotide and is the basic repeat unit of a DNA strand. The sugar in DNA is deoxyribose and in RNA the sugar is ribose. The sugar-phosphate bonds of the DNA backbone are phosphodiester bonds linking the 5' carbon of deoxyribose to the 3' carbon of the subsequent deoxyribose. These bonds impose directionality on both

DNA strands. The order or sequence of the nucleotide bases determines the meaning of the information encoded in that part of the DNA molecule. James D. Watson and Francis Crick (1953) succeeded in elucidating the molecular structure of DNA. DNA molecules are paired strands, referred to as a double helix (In the double helix 'A' always pairs with a 'T' and 'C' always pairs with a 'G'). The two DNA strands are kept together by interstrand hydrogen bonding between the bases (Figure 1.2 and Figure 1.3).

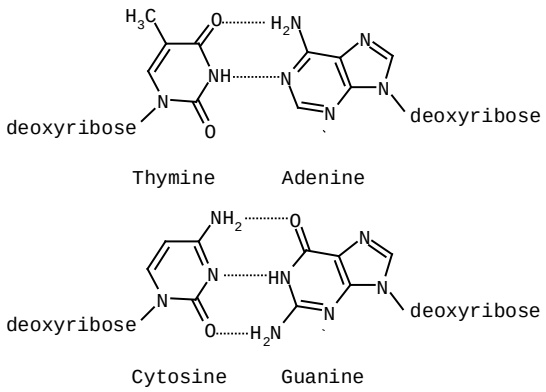


Figure 1.2. Base pairing between A (dATP) and T (dTTP), G (dGTP) and C (dCTP). A and T forms two hydrogen bonds. G and C form three hydrogen bonds. The dotted lines represent hydrogen bond. In case of RNA the sugar is ribose and T is replaced with U (Uracil). (Note: Hydrogen bond lengths are not in scale).

```
.......5'.GTCAGATCGTCGATCGTCG.3'........
          ||||||||||||||||||||
.......3'.CAGTCTAGCAGCTAGCAGC.5'........
```

Figure 1.3. The base pairing between two complementary DNA sequences. The hydrogen bonds are formed as shown in the Figure 1.2. This also illustrates Chargaff's Law.

Although different species have different ratios of pyrimidines or purines, the relative concentration of 'A' is always equal that of 'T'

and ‘G’ equal that of ‘C’ (Chargaff’s Law) (Chargaff, ***et al.***, 1951). Different forms of DNA may be obtained by subjecting DNA fibres to different relative humidities. The DNA helix can assume one of three different geometries, of which the "B" form described by Watson and Crick is believed to predominate in cells (Table 1.1).

Table 1.1. Properties of DNA helical forms, values in parentheses represent the standard deviation from three independently prepared X-ray crystal samples (Dickerson, ***et al.***, 1982).

Geometry attribute	**A-form**	**B-form**	**Z-form**
Helix sense	right-handed	right-handed	left-handed
Repeating Unit	1 bp	1 bp	2 bp
Rotation/bp	33.6^0	$35.9^0(\pm 4.2^0)$	$60^0/2$
Mean bp/turn	10.7	10.0(±1.2)	12
Inclination of bp to axis	$+19^0$	$-1.2^0(\pm 4.1^0)$	-9^0
Rise/bp along axis	0.23 nm	0.332 nm (±0.019 nm)	0.38 nm
Pitch/turn of helix	2.46 nm	3.32 nm (±0.019 nm)	4.56 nm
Mean propeller twist	$+18^0$	$+16^0(\pm 7^0)$	0^0
Glycosyl angle	anti	anti	C: anti, G:syn
Suger pucker	C3’-endo	C2’-endo	C: C2’-endo, G: C2’-exo
Diameter	2.55 nm	2.37 nm	1.84 nm

Each strand in DNA has polarity, such that the 5'-hydroxyl (or 5'-phosphate) group of the first nucleotide begins the strand and the 3'-hydroxyl group of the final nucleotide ends the strand. Accordingly, we say that this strand runs 5' to 3' ("Five prime to three prime"). It is also essential to know that the two strands of DNA run 'antiparallel' such that one strand runs 5'→3' while the other one runs 5'→3' in the opposite direction (hence antiparallel). Along the double-stranded DNA molecule, the nucleotides are complementary (due to base pairing) in nature. The double-stranded, antiparallel, complementary DNA molecule folds to form a helical structure, which resembles a spiral staircase. This is the reason why DNA has been referred to as the "Double Helix" (Figure 1.4). A gene is the part of DNA that carries the information for making a specific protein or set of proteins. In prokaryotes (Bacteria), DNA is present in cytoplasm as a circular molecule. In eukaryotes, individual DNA molecules are found in the chromosomes of the nucleus and also present in mitochondria. In addition to nuclear and mitochondrial DNA, plant cells also have DNA in chloroplasts. Chromosome number varies from organism to organism. Different kinds of organisms have different number of chromosomes. Humans have 23 pairs of chromosomes (44 autosomes and 2 sex chromosomes). The DNA is present in highly compact form in all organisms. However, unlike bacteria, in most eukaryotes, the DNA forms complexes with ***basic*** histone proteins. Histones are the basic nuclear proteins responsible for the nucleosome structure of the chromosomal fiber in eukaryotes (Figure 1.4). The human genome contains approximately 20,000-25,000 genes that must be expressed in specific cells at precise times. Cells manage gene expression by wrapping DNA around clusters (octamers) of globular histone proteins to form nucleosomes (Figure 1.4). These nucleosomes of DNA

are organized into chromatin. Changes in the structure of chromatin structure influence the gene expression. Genes are inactivated (switched off) when the chromatin is condensed (silent), and they are expressed (switched on) when chromatin is uncondenced or extended (active).

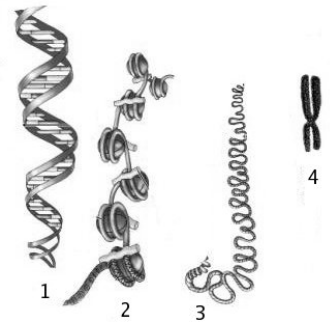


Figure 1.4. DNA double helix with associated histone proteins.
1) DNA helix
2) DNA helix rapped around histone protein complex
3) Super coiled DNA structure
4) Chromosome

The dynamic chromatin states are controlled by the reversible epigenetic patterns of DNA methylation and histone modifications. Epigenetics refers to the study of heritable changes in gene expression that occur without a change in DNA sequence. Enzymes involved in this process (gene expression) include DNA Methyltransferases, Histone deacetylases, Histone acetylases, Histone methyltransferases and the methyl-binding domain protein MECP2. Alterations in these normal epigenetic patterns can deregulate patterns of gene expression, which results in profound and diverse clinical outcomes. All the genetic material (***i.e.,*** DNA) in the chromosomes of a particular organism is called genome. Genome size is generally given as its total number of base pairs. Genes direct the production of proteins with the assistance of enzymes and messenger molecule. Specifically, an enzyme copies the information in a gene's DNA into a molecule called messenger ribonucleic acid (mRNA). The process of mRNA synthesis is called transcription. The mRNA comes out of the nucleus (into the cell's

cytoplasm), where the mRNA is read by tiny molecular machine amino acids in the right order to form a specific protein. The synthesis of protein from mRNA is called translation (Figure 1.5).

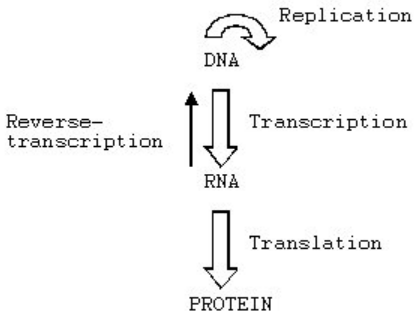


Figure 1.5. The flow of Genetic Information represented by the 'central dogma of molecular biology'. The dogma forms the backbone of molecular biology and is represented by four major stages.

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Figure 1.6. The Universal Genetic Code (Note: the code is an RNA code. The four bases (A, C, G and T or U) present in DNA / RNA form 64 triplet combinations; however, since there are only 20 naturally occurring amino acids, more than one codon may encode an amino acid. This phenomenon is termed the degeneracy of the genetic code. In addition to coding for amino acid, a particular triplet sequences also indicate the beginning (start) and the end (stop) of a particular gene. With the exception of a limited number of differences found in mitochondrial DNA and one or two other species, the genetic code appears to be universal).

One strand of DNA holds the information that codes for various genes; this strand is often called the template strand or antisense strand (containing anticodons). The other, complementary, strand is called the coding strand or sense strand (containing codons). Since mRNA is made from the template strand (because of the 1-to-1 correspondence), it has the same information as the coding strand. The Figure 1.6 refers to triplet nucleotide codons along the sequence of the coding or sense strand of DNA as it runs in 5'→3'; the code for the mRNA would be identical but for the fact that RNA contains U rather than T. An example of two complementary strands of DNA would be:

(5'→3')	ATGGAATTCTCGCTC	Coding, sense strand
(3'←5')	TACCTTAAGAGCGAG	Template, antisense strand
(5'→3')	AUGGAAUUCUCGCUC	mRNA, made from Template strand

Since amino acid residues of proteins are specified as triplet codons, the protein sequence made from the above example would be Met-Glu-Phe-Ser-Leu... (MEFSL...). Codons are "decoded" by transfer RNAs (tRNA), which interact with a ribosome-bound messenger RNA (mRNA) containing the coding sequence. There are 64 different tRNAs, each of which has an anticodon loop (used to recognize codons in the mRNA). 61 of these have a bound amino acyl residue; the appropriate "charged" tRNA binds to the respective next codon in the mRNA and the ribosome catalyzes the transfer of the amino acid from the tRNA to the growing (nascent) protein/polypeptide chain. The remaining 3 codons are used for "punctuation"; ***i.e.***, they signal the termination (the end) of the growing polypeptide chain (Figure 1.6). The Genetic Code is also called "The Universal Genetic Code". It is known as "universal", because all known organisms use it as a code for DNA, mRNA, and tRNA. The universality of the genetic code encompasses animals (including humans), plants,

fungi, archaea, bacteria, and viruses. However, small variations in the genetic code exist in mitochondria and certain microbes. Nonetheless, it should be emphasized that these variances represent only a small fraction of known cases, and that the Genetic Code applies quite broadly, certainly to all known nuclear genes.

1.1.1 Replication

DNA replication or DNA synthesis is the process of copying a double-stranded DNA in a cell, prior to cell division. In eukaryotes, this is during the S phase of the cell cycle, preceding mitosis and meiosis. The two resulting double strands are identical (if the replication went well), and each of them consists of one original and one newly synthesized strand. This is called semiconservative replication.

1.1.2 Transcription

The process of transcription starts in the promoter region of a gene. Promoter elements support the buildup of the RNA Polymerase machinery. A full length RNA copy of the genomic DNA including exons and introns is generated. Transcription termination is linked to 3' polyadenylation of the transcript and involving transitions at the 3' end of genes that may include an exchange of elongation and polyadenylation / termination factors. The transcribed product of a gene, nuclear RNA, is subject to further modifications. A capping component is added to the 5' end and a poly-A tail to the 3' end. The nuclear RNA is additionally shortened by the excision of introns and occasionally further exons. This step is known as 'splicing' and leads to the diversity of transcripts,

mRNAs, by facilitating various exon combinations. This step takes place in nucleus.

1.1.3 Translation

The process of translation takes place after the export of the mature mRNA from the nucleus. Translation means to transfer protein-coding information on mRNA into actual proteins by another synthesis step. Ribosomes, which are large complex of RNA and protein, are the factories of protein biosynthesis that utilize mRNA as a template. There are other parts of the mRNA before and after its start and stop sequences that are not translated. These come from the template DNA strand that the RNA was transcribed from. These regions are known as the 3'UTR and 5'UTR and they do not code for any/part of protein sequences. The stability of the mRNA depends on UTRs. Not all genes that are transcribed are translated. The information in mRNA can be transferred in to complementary DNA (cDNA) by the process of reverse transcription. cDNA may be used for gene cloning or as a gene probe (widely used for protein sequencing).

1.1.4 RNA

RNA is similar to DNA, with the only difference of T in DNA; RNA contains U (Uracil), in the sequence beside the sugar. Most cellular RNA molecules are single stranded. They may form secondary structures such as stem-loop and hairpin. The major role of RNA is to participate in protein synthesis. This requires three classes of RNA are rRNA, tRNA and mRNA.

1.1.5 Protein

Proteins are essentially polymers made up of a specific sequence of amino acids. The details of this sequence are stored in the code of a gene.

Proteins consist of a polypeptide backbone with attached side chains. Each type of protein differs in its sequence and number of amino acids; therefore, it is the sequence of the chemically different side chains that makes each protein distinct. The two ends of a polypeptide chain are chemically different: the end carrying the free amino group (NH_3^+, also written NH_2) is the amino, or N-, terminus, and that carrying the free carboxyl group (COO-, also written COOH) is the carboxyl, or C-, terminus. The amino acid sequence of a protein is always presented in the N→C (N to C) direction, reading from left to right. Different proteins perform a wide variety of biological functions. Some proteins are enzymes, which catalyze chemical reactions and other proteins play structural or mechanical roles, as those that form the joints of the cytoskeleton.

1.1.6 Nucleoside triphosphates

ATP, GTP, CTP, and UTP are used in RNA synthesis, while dATP, dGTP, dCTP, and dTTP are used for DNA replication. dATP has no role outside of DNA, while ATP is essential for many enzymatic processes.

ATP

Adenosine triphosphate plays several important roles in almost every pathway in the cell. ATP can be used as a source of energy, acting alone or combined with other (like niacin in NAD or riboflavin in FAD). ATPases are enzymes that catalize decomposition of ATP in to ADP and Pi. ATPases are found throughout the cell performing a wide variety of

functions (like pumping ions across the membrane and running all of the cytoplasmic motors that shuttle material around the cell and drive cilia, flagella, and muscles). ATP is also used extensively by the cell as a source of phosphates for modifying proteins - several proteins require phosphorylation to be activated or inactivated, and this is used by the cell to control which enzymes are on or off. The enzymes which phosphorylate other proteins are called kinases and all require ATP to function.

GTP

Guanosine triphosphate is similar to ATP, but has fewer roles. There are a few instances in which GTP is used as a phosphate donor or as an energy source. β-tubulin has GTPase activity, which hydrolyze GTP to GDP to form the microtubules of the cytoskeleton. There are several other GTPase enzymes in the cell, however most of these enzymes are not used in enzymatic reactions, but rather are used to transmit signals throughout the cell. G-proteins are a specific class of GTPases which uses GTP to interact with other enzymes and activate various cellular processes. Many hormones and neurotransmitters have receptors that use G-proteins to transmit their signals to the rest of the cell. There are several other GTPases, including the Ras and Rab families of small GTPases, that are all also used to transmit signals and to control other intercellular traffic through their binding to GTP.

UTP

Uridine triphosphate is used for a different purpose. The most common example of this is in glycogen synthesis. Many cells in the body (especially in the liver) store glucose (sugar) in the form of glycogen, a complex starch composed of long, branching chains of glucose

molecules. To enhance this reaction, free glucose molecules coupled with UTP to produce UDP-glucose and free phosphate. This makes the glucose molecules more reactive, since the glucose-phosphate bond in UDP-glucose is a high energy bond. When the UDP-glucose is added to glycogen, the UDP is released, and the energy is used to attach the glucose to the glycogen molecule. In fact, Uridine is used for UDP-glucose, UDP-galactose, UDP-mannose, etc., the building blocks of numerous carbohydrates that are essential for many cellular functions.

CTP

Cytidine triphosphate is used similarly to UTP, however instead of sugars, CTP is needed for synthesis of fats (lipids). CDP-diacylglycerol, CDP-ethanolamine, and CDP-choline are the building blocks of the phospholipids that make up the cell membrane. Since all cells require intact membranes to survive, CTP plays an important cellular function.

Generally speaking, more complex organisms have larger genomes than simpler organisms. But, there is a great deal of variation in the range of genome sizes within a class of organisms (Table 1.2). There is a tremendous diversity in the size and organization of prokaryotic genomes (both Bacteria and Archae). The size of Bacterial chromosomes ranges from 0.6 Mbp to over 10 Mbp, and the size of Archael chromosomes range from 0.5 Mbp to 5.8 Mbp. But prokaryotic genomes are roughly proportional to gene numbers. Gene duplication, small-scale deletions and insertions, transpositions, horizontal transfer, loss of genes in parasitic lines, etc., affect the bacterial genome size. Eukaryotic genome sizes are not proportional to gene numbers or anatomical complexity.

Table 1.2. Genome sizes of some of the model systems.

Model organism	Size (Bases)	No. of Genes
Escherichia coli (Bacterium)	4,639,221	4,377
Phi-X 174 (Bacteriophage)	5,386	10
Oryza setiva (Rice)	3.9×10^8	37,544
Anapheles gambiae (Mosquito)	278,244,063	13,683
Homo sapiens (Human)	3.3×10^9	20,000-25,000
Drosophila melanogaster (Fruit fly)	122,653,977	13,379

The human genome size is ~3.2×10^9 bp. These genes are located on the 23 pairs of chromosomes in the nucleus of a human cell. Each of these estimated 20,000-25,000 genes in the human genome codes for an average of three proteins per gene. The region of DNA that does not carry the information necessary to make a protein is the non-coding DNA. Eukaryotic genome contain various degrees of non-coding repetitive structures viz, satellites, micro/mini satellites, retrotransposons, retrovirus ***etc***. Repetitive sequence size correlates with genome size. Repeats constitute about 45% of the human and the mouse genomes and can be found in both transcribed (intron and UTRs) and non-transcribed intergenic sequences. Two types of repeats are commonly observed. These are Tandemly repeated DNA and the Interspersed repetitive DNA. Tandemly repeated non-coding DNA sequences are composed of a "repeat unit" which is repeated n-times in a head to tail arrangement. Tandom repeats include satellites, micro/mini satellites. LINEs (Long Interspersed Nucleotide Elements), and SINEs (Short Interspersed Nucleotide Elements) are the two major repetitive elements in the human and mouse of type Interspersed repetitive nucleotide elements. ***Alu***

repeats are the SINEs and AT-rich regions are LINEs. The fraction of protein coding DNA in the genome decreases with increasing organismal complexity. In bacteria, about 90% of the genome codes for proteins. This number drops off to 68% in yeast, to 23-24% in nematodes and to 1.5-2% in mammals.

1.1.7 Human Genome Project (HGP)

The Human Genome Project (HGP) produced a human genome sequence and is freely available in public databases. HGP begun in 1986 by US, Department of Energy (DOE) to create an ordered set of DNA segments from known chromosomal locations, develop new computational methods for analyzing genetic map and DNA sequence data, and develop new techniques and instruments for detecting and analyzing DNA. This DOE initiative is now known as the Human Genome Program. The joint national effort, led by DOE and National Institutes of Health (NIH) was taken up initially and is known as the Human Genome Project. Sequence and analysis of the first human genome working draft was published in February 2001 issue of ***Nature*** and ***Science (***Bentley, ***et al.,*** 2001; Venter ***et al***., 2001***)***. The project originally was planned to last 15 years, but rapid technological advances accelerated the completion date to 2003.

Project goals were to

- ***identify*** all the approximately 20,000-25,000 genes in human DNA,
- ***determine*** the sequences of the 3 billion chemical base pairs that make up human DNA,
- ***store*** this information in databases,
- ***improve*** tools for data analysis,
- ***transfer*** related technologies to the private sector, and

- ***address*** the ethical, legal, and social issues (ELSI) that may arise from the project.

To help achieve these goals, researchers also studied the genetic makeup of several non-human organisms. These include the common human gut bacterium ***E.coli***, the fruit fly ***Drosophila melanogaster***, and the laboratory mouse ***Mus musculus***. A unique aspect of the HGP is that it was the first large scientific undertaking to address potential ELSI implications arising from project data.

(http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)

1.2 Prokaryotic gene and transcription initiation

The clustered bacterial genes (operon) that encode proteins are necessary to perform coordinated function, such as biosynthesis of a given amino acid (Figure 1.7). RNA that is transcribed from prokaryotic operon is polycistronic, a term implying that multiple proteins are encoded in a single transcript. Due to their lack of introns, prokaryotes can couple transcription and translation. There are a number of key features to the promoter region that give it the ability to provide a signal in transcription initiation. These signal regions located approximately 10 and 35 bp upstream of the transcription start site (TSS) are two such regions, called the -10 and -35 sequences. Each sequence consists of six base pairs. For an ideal promoter, the -35 region sequence is TTGACA and TATAAT for the -10 region. In addition to the specificity of the bases in these sequences, the spacing between these two is also important. Ideally, this gap is ~17 bp long. Deviations from this spacing have significant effects on the strength of the promoter region. The spacing and sequence similarity of these promoter regions will give the greater strength.

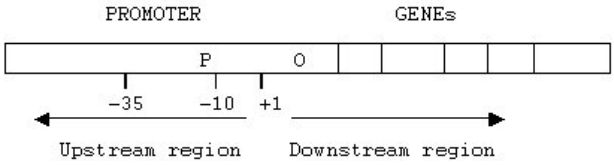


Figure 1.7. The block diagram of prokaryotic operon structure. The start site of transcription is +1. 'o' and 'p' represents operator and promoter respectively.

In prokaryotic cells, all 3 RNA classes (mRNA, tRNA and rRNA) are synthesized by a single polymerase. Prokaryotic RNA polymerase consists of 2α, 1β, 1β' subunits and a σ factor. The activity of RNA polymerase at a given promoter is in turn regulated by interaction with accessory proteins (regulatory proteins), which affect its ability to recognize transcription start sites. These regulatory proteins can act both positively (activators) and negatively (repressors). The accessibility of promoter regions of prokaryotic DNA in many cases is regulated by the interaction of proteins with sequences termed operators. The operator region is adjacent to the promoter elements in most operons and in most cases the sequences of the operator binds with a repressor protein. However, there are several operons in ***E. coli*** that contain overlapping sequence elements (one that binds a repressor and one that binds an activator). French molecular biologists Jacob and Monod were the first to unravel a transcriptionally regulated system, the ***lac*** operon of ***E.coli.*** The ***lac*** operon was proved to be a model for other transcriptionally regulated systems in prokaryotes. The ***lac*** operon functions in the utilization of the sugar lactose by ***E.coli.*** Jacob and Monod demonstrated that the operon was repressed when both glucose and lactose were present, glucose being the preferred energy source. A repressor molecule binding to the operon prevents transcription and provides the repression. When lactose is present, lactose binds to the repressor molecule, changing the molecule's

shape such that the repressor can't bind to operon and causing it to fall off of the operon. Transcription can then occur, and the products necessary for the utilization of lactose are produced. When lactose is absent, the repressor molecule again binds to operator, bringing a halt to transcription. Even though transcription is blocked, RNA polymerase, can still bind to the operon. When the repressor is removed, transcription can thus begin immediately. The positive regulation of ***lac*** operon is dependent on cAMP levels (catabolite repression). Cyclic AMP (cAMP) is made from AMP. When the glucose level in the cell is high, the cAMP level is low, because glucose inhibits synthesis of cAMP. When the glucose level is low, the cAMP level is high. cAMP combines with the CAP (Catabolite Activator Protein) to form a complex that binds to part of the ***lac*** operon promoter. This complex bends the DNA such a way that makes it much easier for RNA polymerase to bind to the promoter. This allows transcription to occur, but only if the ***lac*** repressor is not present. Thus low glucose levels cause high cAMP levels. When cAMP is high, it combines with CAP. The CAP-cAMP complex then binds to the promoter allow transcription to occur.

1.3 Eukaryotic gene and transcription initiation

Complete genome sequences of human and mouse (Venter, ***et al.,*** 2001; Waterston, ***et al.***, 2002) revealed that eukaryotes have very little coding region than expected earlier. Human genome contains approximately 20000-25000 genes, which represent less than 2% of the whole genome. Unlike in most prokaryotic genomes (that contain packed gene units with few intergenic), repeated and non-coding sequences that do not code for proteins make up the remaining part of the human genome. Protein coding genes contain exons, introns and a promoter region (Figure 1.8).

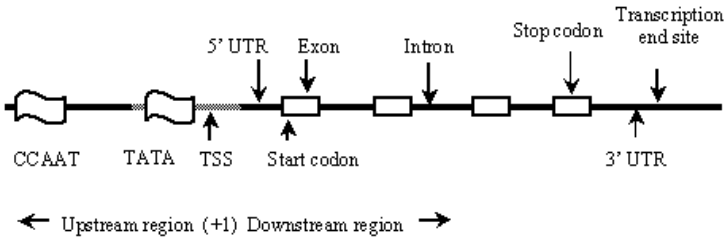


Figure 1.8. Eukaryotic gene structure. The gene typically consists of exons, introns, promoter and other control elements.

Exons

Exons are the regions downstream of the promoter of a gene that are transcribed and exported from the nucleus as part of the mRNA. In many genes, each exon contains part of the open reading frame (ORF) that codes for a specific portion of the complete protein, however, the term exon is often misused to refer only to coding sequences for the final protein. This is not true, since many noncoding exons are known in human genes. Some of the exons will be wholly or part of the 5' untranslated region (5' UTR) or the 3' untranslated region (3' UTR) of each transcript. The untranslated regions are important for efficient translation of the transcript as well as being important for controlling the rate of translation and half life of the transcript. Furthermore, transcripts made from the same gene may not have the same exon structure since parts of the mRNA could be removed by the process of alternative splicing. Some mRNA transcripts have exons with no ORF's and thus are sometimes referred to as non-coding RNA.

Introns

Introns are regions downstream of the promoter of a gene that are also transcribed on to RNA but are excised (spliced) from the maturing

mRNA. Introns are common in eukaryotic RNAs of all types, but are found in prokaryotic tRNA and rRNA genes only. Introns sometimes allow for alternative splicing of a gene, so that several different proteins that share some sections in common can be produced from a single gene (Figure 1.9).

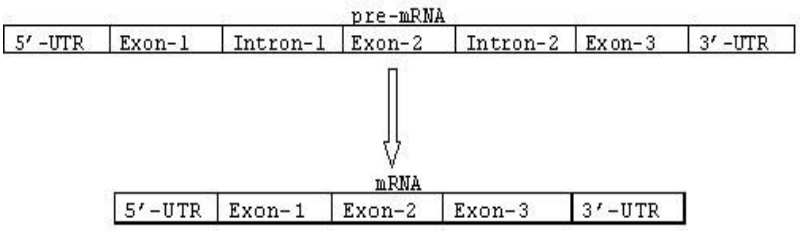


Figure 1.9. Post-transcriptional modifications. Intron and exon structure. Alternative splicing modes are possible in this example.

The part of a gene that contains the information to turn the gene on or off is the promoter region. The process of transcription is initiated at this promoter. Promoter region is a regulatory region and show variations within and also between species to species. The proteins (transcription factors, TFs, which are cell or tissue specific) bind to this promoter region of the gene and subsequently cause efficient binding of RNA polymerase-II to initiate mRNA synthesis. Specific DNA sequence regions (elements) within the promoter region (like, TATA-box, CCAAT-box, Downstream Promoter Element (DPE) and GC-box ***etc***.,) exhibit similarities between different promoters of the same DNA as well as between various species. The core promoter region (which can extend ~35 bp upstream from transcription site and which is a minimal promoter region required to start the pre-initiation complex formation) usually has TATA-box, which is conserved in most of the species (30-50% of promoters) and Transcription Start Site (TSS) region, which usually is

not conserved. The TSS is represented by a single nucleotide. The TSS-region represents the nucleotides around the TSS. Each of these core promoter elements is found in some but not in all core promoters. It appears that there are no universal core promoter elements. Each nucleotide in the consensus sequence motif (TATA-box, CCAAT-box and GC-box) represents the most frequently occurring nucleotide at that position and does not represent an actual sequence. In this case the sequence is represented by the ambiguous nucleotide codes (Table 1.3 and Table 1.4) as a consensus sequence. Reliable identification of the core promoter region by RNA polymerase-II prior to transcription initiation is mandatory for the proper initiation and regulation of mRNA synthesis (Smale and Kadonaga, 2003). The enzyme RNA polymerase-II is a multi-subunit protein that catalyzes the synthesis of mRNA from the DNA template. Accurate and efficient transcription from the core promoter requires the polymerase along with general transcription factors, which include transcription factor IIA (TFIIA), TFIIB, TFIID, TFIIF and TFIIH. With TATA box-dependent core promoters, it has been found that the purified transcription factors can assemble into a transcription pre-initiation complex (PIC) in the following order: TFIID, TFIIB, RNA polymerase II- TFIIF complex, TFIIE and TFIIH. TFIID is a multisubunit protein that consists of a TATA-box binding protein (TBP) and approximately 13 TBP-associated factors (TAFs, Burley and Roeder, 1996). Although TBP can recognize divergent AT-rich sequences because of its DNA binding mechanism, promoters that do not have a TATA-box like sequence at ~30 bp upstream of the TSS would be incompatible with specific binding by TBP (Patikoglou, ***et al***., 1999).

Table 1.3. IUPAC ambiguous nucleotide code (Cornish-Bowden, 1985). Used for the nucleotide consensus representation.

S.No	Ambiguous symbol	Symbols	Name
1	A	A	Adenine
2	C	C	Cytosine
3	G	G	Guanine
4	T	T	Thymine
5	U	U	Uridine
6	R	A or G	puRine
7	Y	C or T	pYrimidine
8	W	A or T	Weak hydrogen bonding
9	S	C or G	Strong hydrogen bonding
10	M	A or C	AMino group at common position
11	K	G or T	Keto group at common position
12	B	C, G or T	not A
13	D	A, G or T	not C
14	H	A, C or T	not G
15	V	A, C or G	not T
16	N	A, C, G or T	Any

In ***D. melanogaster*** and human promoters, many of these non-TATA-containing promoters have some combination of ***Inr*** and DPE elements (Smale and Kadonaga, 2003). In some genes, a pyrimidine-rich consensus sequence, YYAN(t/a)YY, is at or near the TSS. This sequence is called ***Inr.***

Table 1.4. Examples of general and specific control elements (Bucher, 1990). The ambiguous nucleotide codes are used to represent the consensus sequence.

Name	Consensus	Protein factor
TATA-box	TWTWWAW	TFIID (TBP), Histones
Initiator	CWBHY	TFIID (TBP)
CCAAT-box	RRCCAWWSR	CBF, NF-1, C/EBP
GC-box	RGGHK	Sp1
Specific elements		
NF-*k*B	GGGAMTTYCC	p50/p65
CarG-box	$CC(A/T)_6GG$	SRF

Many human and ***D.melanogaster*** promoters have no recognizable TATA element or even AT-rich regions upstream from the TSS (Ohler, ***et al***., 2002), this suggests that if TBP interacts with DNA at these promoters, it must do so by a different mechanism from that seen at classical TATA elements. TFIIB is a single polypeptide that interacts with TBP as well as with the DNA upstream of the TATA-box. In the PIC assembly mechanism TFIID and TFIIB are the first two factors that interact with the core promoter. Accordingly, it appears that these two factors have a critical role in the recognition of core promoter motifs. TFIIF binds RNA polymerase-II as a heteromer and contains two subunits that are conserved among human, insects and yeast. The N-termini of both conserved subunits form a dimerization domain and the C-termini of both subunits are winged helix domains. The general factors TFIIE and TFIIH function primarily in steps after PIC formation. TFIIE

(probably a heterodimer) binds independently to RNA polymerase-II (Bushnell ***et al***., 1996) and is thought to stimulate both the kinase and helicase activities of TFIIH. TFIIH has a duel function in transcription and transcription coupled DNA mismatch repair. The TFIIH is composed of two domains, a core domain (contains two DNA helicase activities) and a kinase domain. These helicase acitivities are supposed to form an open complex formation. Helicases act by destabilizing double stranded nucleic acids through the ATP hydrolysis dependent motion of two separate domains that interact with single- and double- stranded nucleic acids.

The CpG dinucleotide is underrepresented in vertebrate genomes due to methylation at the 5 positions of the cytosine ring and subsequent deamination of the 5-methylcytosine to give a TpG dinucleotide, which is not repaired by the DNA repair machinery. However, there are stretches of DNA, termed CpG islands that are relatively GC-rich and over represented as CpG dinucleotides that are mostly unmethylated (Bird, 1986). CpG islands are generally from 0.5 to 2 kbp in range and contain promoters for a wide variety of genes. CpG islands typically lack TATA or DPE core promoter elements, but contain multiple GC-box motifs that are bound by Sp1 and related transcription factors (Brandeis ***et al.,*** 1994). In addition, transcription from CpG islands initiates from multiple weak start sites that are often distributed over a region of about 100 nt, which is in contrast to transcription from TATA or DPE-dependent core promoters that occurs from a single site or less than 10 nt. It is also possible that CpG island promoters consist of multiple (Sp1+***Inr***) pairs that collectively generate the array of start sites.

Gene expression and its regulation involve the binding of many regulatory TFs to specific DNA elements called Transcription Factor

Binding Sites (TFBS) within the proximal and distal promoter regions (Figure 1.10). A region of 200-300 bp immediately upstream of the core promoter is the proximal promoter that has abundant of TFBS. Further upstream is the distal promoter region that usually contains few TFBS and enhancers. TFBS are represented by relatively short (5-10 bp) nucleotide sequences.

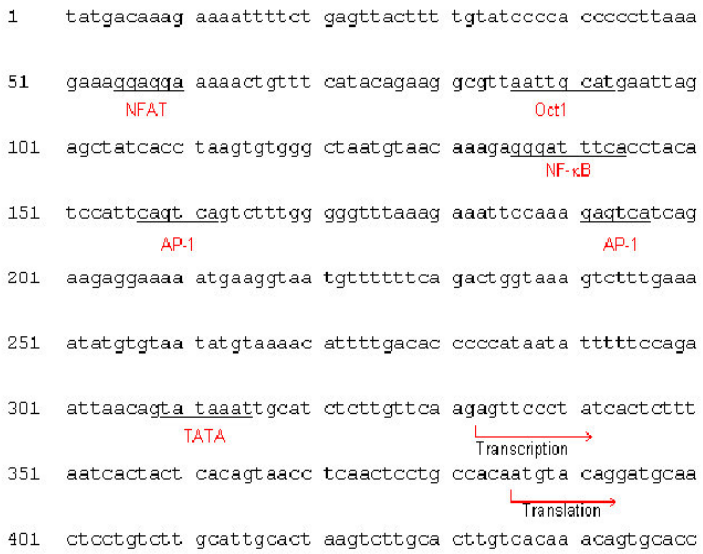


Figure 1.10. Illustration of the binding site arrangement of the human IL2 promoter (Genbank accession number AF031845. Binding sites are underlined. Distinct proteins recognize the different binding sites. Arrows represent Transcription and translation start sites).

Specificity of TF is defined by its interaction with TFBS and it is extremely selective, mediated by non-covalent interactions between appropriately arranged structural motifs of the TF and exposed surfaces of the DNA bases and backbone (Vazquez, ***et al***., 2003). An enhancer is a short region of DNA that can be bound with proteins (namely, the trans-acting factors, much like a set of transcription factors) to enhance transcription levels of genes in a gene-cluster. An enhancer does not need

to be particularly close to the genes it acts on, and need not be located on the same chromosome. An enhancer need not necessorly bind close to the TSS to affect its transcription, as some have been found to bind several hundred thousand base pairs upstream or downstream of the start site. Enhancers can also be found within introns. An enhancer's orientation can also be reversed without affecting its function. Furthermore, an enhancer may be excised and inserted elsewhere in the chromosome, and can still affect the gene transcription. This may be one of the reason for checking of intron polymorphisms though they are not transcribed and translated.

1.3.1 Structural Motifs in Transcription Factors

The proteins (transcription factors) bound to promoter elements through the domains / motifs present in the TF molecule (Table 1.5). Following are some of the motifs illustrations.

Homeodomain

The homeodomain is a highly conserved domain of ~60 amino acids found in a large family of transcription factors. This family was first identified in ***Drosophila*** as a group of genes that, when altered, would cause transformations of one body part for another (e.g., legs for antenna), so called homeotic transformations. This class of genes has been identified in both invertebrate and vertebrate organisms. The homeodomain itself forms a structure, which is highly similar to the bacterial helix-turn-helix proteins such as ***Cro*** and ***lambda*** repressor. The principal function of all homeodomain-containing proteins is in the establishment of patterns with in an organism.

Helix-Loop-Helix (HLH)

The HLH domain is involved in protein dimerization. The HLH motif is composed of two regions of α-helix separated by a region of variable length, which forms a loop between the two α-helices. This motif is quite similar to the Helix-turn-helix motif found in several prokaryotic trascription factors such as the CRP protein involved in the regulation of the ***lac*** operon. The α-helical domains are structurally similar and are necessary for protein interaction with sequence elements that exhibit a two-fold axis of symmetry. This class of transcription factor most often contains a region of basic amino acids located at the N-terminal side of the HLH domain (termed bHLH proteins) that is necessary for the protein to bind DNA at specific sequences. The HLH domain is necessary for homo- and heterodimerization. Examples of bHLH proteins include MyoD (a myogenesis inducing transcription factor) and c-Myc (originally identified as a retroviral oncogene). Several HLH proteins act as repressors, because lack of basic region at N-terminal. These HLH proteins repress the activity of other bHLH proteins by forming heterodimers with them and preventing DNA binding.

Zinc Fingers

The zinc finger domain is a DNA-binding motif consisting of specific spacing of cysteine and histidine residues that allow the protein to bind zinc atoms. The metal atom coordinates the sequences around the cysteine and histidine residues into a finger-like domain. The finger domains can interdigitate into the major groove of the DNA helix. The spacing of the zinc finger domain in this class of transcription factor coincides with a half-turn of the double helix. The classic example is the

RNA polymerase-III transcription factor, TFIIIA. Proteins of the steroid / thyroid hormone family of transcription factors also contain zinc fingers.

Leucine Zipper

The leucine zipper domain is necessary for protein dimerization. It is a motif generated by a repeating distribution of leucine residues spaced at 7 amino acids apart within α-helical regions of the protein. These leucine residues end up with their R-groups protruding from the α-helical domain in which the leucine residues reside. The protruding R-groups are thought to interdigitate with leucine R groups of another leucine zipper domain, thus stabilizing homo- or heterodimerization. The leucine zipper domain is present in many DNA-binding proteins, such as c-Myc, and C/EBP.

Table 1.5. Some of the common motifs found in transcription factors and their functions (see text for further details).

Motif	Function
Helix-turn-helix (HTH)	binds to the major groove of the DNA
Zinc fingers	function as structural platforms for DNA binding
Leucine zippers	Useful in detecting the sequence of binding of several proteins
Basic-helix-loop-helix (bHLH)	Binds to outer region of the DNA non specifically

Winged Helix

The winged helix is a DNA-binding motif composed of an α/β structure. This structure contains 3 N-terminal α-helices and a 3-stranded antiparallel β-sheet. The folding of the β-sheet region about the α-helices

give the appearance of wings on the helices, hence the term winged-helix. This motif was first identified in the transcription factor HNF-3γ. HNF-3γ is a member of a large family of transcription factors that are related to the ***Drosophila*** gene ***forkhead***, hence the gene family is termed the fork head (FKH) family. The nomenclature of the fork head family of transcription factors has been changed so that all members have names that begin with Fox.

Some transcription factors bind to enhancers, that are thousands of base pairs away from the gene they control. This binding enhances the transcription rate of the gene. Enhancers can be located upstream, downstream, or even within the gene they control. The binding of a protein to an enhancer regulates the transcription of a gene thousands of base pairs away from transcription initiation. One possibility is that enhancer-binding proteins have (in addition to their DNA-binding site) have sites that bind to TFs assembled at the promoter of the gene. This would draw the DNA into a loop. Silencers are control regions of DNA that (like enhancers) may be located thousands of base pairs away from the gene they control. However, when transcription factors bind to them, expression of the gene they control is repressed.

1.4 Mitochondrial genome and transcription initiation

The genetic organization of the mitochondrial DNA (mtDNA) is extremely compact and there are no intronic sequences and almost no non-coding nucleotides between genes. Most human cells contain hundreds of mitochondria and each mitochondrion has thousands of mtDNAs. In the mitochondrial genome, the presence of rRNA, protein coding and tRNA genes do not leave any space for promoters comparable to those found in eukaryotic nuclear or bacterial genomes.

The transcription is carried out all around the circle, and a polycistronic RNA (single transcript) is produced (Taanman, 1999). The primary transcript is cleaved afterwards, releasing individual rRNAs, tRNAs and mRNAs. The mtDNA therefore presents the closest analogy to a bacterial operon. The mtDNA molecule encodes 37 genes (2 rRNAs, 22 tRNAs and 13 polypeptides) all of these are components of the oxidative phosphorylation system (Boore, 1999) (Table 1.6). Phylogenetic analysis of the mtDNA suggests that the mitochondrion originates from an α-protobacterium, which early in evolution developed a symbiotic relationship with a primitive eukaryotic cell (Andersson, 2003).

Table 1.6. The polypeptides produced by mitochondrial genome and the proteins involved in the oxidative phosphorylation.

Enzyme complex	**Peptides from mtDNA**	**Total complex proteins**
Complex-I (NADH: ubiquinone oxidoreductase)	7 (ND1-ND6 and ND4L)	46
Complex-III (Ubiquinol: cytochrome-c oxidoreductase)	1 (apocytochrome b)	11
Complex-IV (Cytochrome-c oxidase)	3 (COI-COIII)	13
Complex-V (ATP synthase)	2 (ATP6 and ATP8)	16

Initiation of transcription at mitochondrial promoters in mammalian cells requires the simultaneous presence of a monomeric mitochondrial RNA polymerase (which displays significant sequence similarity to the monomeric RNA polymerases found in bacteriophases), mitochondrial transcription factor A, and either transcription factor B1 or B2. Mitochondrial DNA strands were classified as heavy (H) and light (L)

based on the G+C content of each strand. The heavy strand is rich in G content and light strand is similarly rich in C content. One polypeptide and eight tRNA genes are encoded in the L-strand while the rest are encoded in the H-strand. The non-coding region (control region / D-loop) is of ~1.1 kb, located between the $tRNA^{Phe}$ and $tRNA^{Pro}$ genes, contains the origin of replication for the heavy strand (O_H), the transcription promoters (LSP and HSP, light and heavy strand promoter, respectively) and other regulatory elements for the mtDNA expression (Figure 1.11).

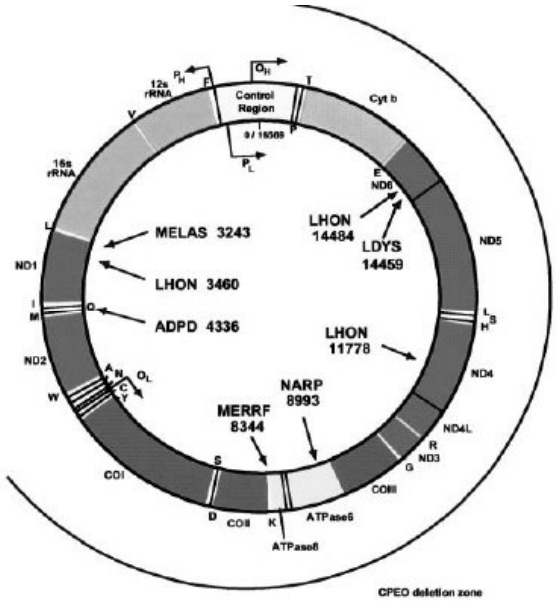


Figure 1.11. MITOMAP: A Human Mitochondrial Genome Database taken from http://www.mitomap.org,.

The length of the mtDNA control region varies a lot among the animal taxa. In vertebrates it ranges from 200 to almost 4000 bp. All studies of human evolution based on mtDNA sequencing have been confined to the control region, which constitutes less than 7% of the mitochondrial

genome (Figure 1.12). Some times there are tandem repeats in the control region / D-loop (most commonly at the beginning and at the end). In some cases there are also other structures (stem loop, etc) that can increase the size. The control region / D-loop is the most variable region in the whole genome. The rate of nucleotide substitution does not correlate with the rate of structural changes in the genome. In mammals, the mtDNA evolves very rapidly in terms of nucleotide substitutions, but the spatial arrangement of genes and the size of the genome are fairly constant among species. The human mitochondrial genome gives three different transcripts from three different points (H1, H2 and L) that transcribe three polycistronic molecules (Montoya, ***et al.,*** 1982). H_1 is located 16 nt upstream of the $tRNA^{Phe}$ gene. The polycistronic primary transcripts synthesized from the three initiation sites are processed, according to the "tRNA punctuation" model, to give rise to the mature rRNAs, mRNAs and tRNAs after the precise endonucleolytic cleavage on both sides of the tRNA molecule (Ojala, ***et al***, 1981).

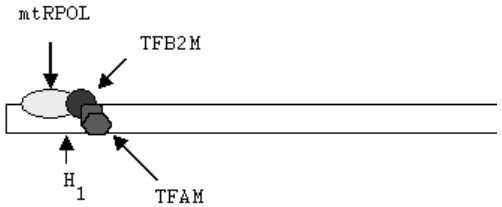


Figure 1.12. The human mitochondrial control region (D-loop). H_1 is the transcription initiation site and mtRPOL, TFAM and TFB2M are the RNA polymerase, and the transcription factors involved in transcription.

According to this model, the tRNA sequences on the nascent RNA chains would acquire the cloverleaf structure and acts as the signals for the processing enzymes. The processing of the polycistronic molecules requires precise endonucleolytic cleavages at the 5'- and 3'- end of the

tRNAs, a polyadenylation activity for mRNAs and rRNAs, and, finally, the post transcriptional modification of a subset of tRNA and rRNA nucleotides. Replication of mammalian mtDNA is linked and dependent on mitochondrial transcription. The human mtDNA has a very high recombination rate, at least 10 times that of nuclear genes. Mutations that arise in the somatic tissues degrade cellular energy production but die with the individual. Mutations that arise in the female germ line are transmitted to the next generation and gives new mtDNA polymorphisms or as a devastating mtDNA disease. As a consequence of maternal transmission (through the oocyte cytoplasm) of mtDNA, the maternal and paternal mtDNAs rarely mix in the same cytoplasm, and no recombination has been detected between different mitochondrial lineages. The only way that the mtDNA sequence can change is by the sequential accumulation of mutations along radiating maternal lineages.

1.5 Information theory

Information theory is the mathematical theory of data communication and storage. The term ***information***, used in a simple sense, refers to the transmitted messages. For example voice transmitted by telephone, images transmitted by television systems and digital data in computer systems and networks can be considered as communicated or stored information. The communications system includes five components: an information source, a transmitter, the medium, a receiver, and a destination (Figure 1.13). Information theory answers two fundamental questions in communication theory.

1.Data compression, which is explained by entropy (H) of the source.

2.Transmission rate of communication, which is explained by the channel capacity (C).

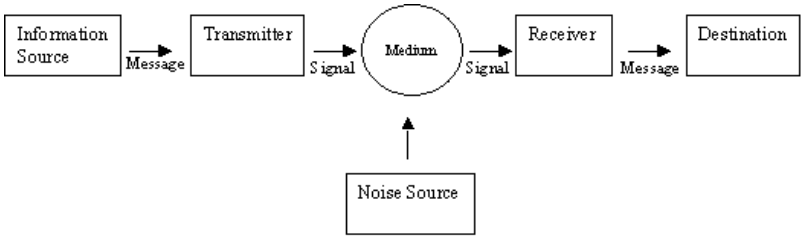


Figure 1.13. Shannon's model of a communications system includes five components: an information source, a transmitter, the medium, a receiver, and a destination. The amount of information that can be transferred from information source to destination is a function of the strength of the signal relative to that of the noise generated by the noise source.

In the early 1940's it was thought that increasing the transmission rate of information over a communication channel has increased the probability of error. Shannon (1948) surprised the communication theory community by providing that this was not true as long as the communication rate was below the channel capacity. The capacity can be computed from the noise characteristics of the channel. The concept of entropy is at the heart of information theory and it is characterized by the quantity of a random process's uncertainty. If the entropy of the source is less than the capacity of the channel, then asymptotically error free communication can be achieved. The entropy of a discrete random variable $\boldsymbol{X}$ with a probability mass function $\boldsymbol{p(x)}$ is defined by $\boldsymbol{H(X) = -\sum_x p(x) \cdot log_2 \cdot p(x).}$ Entropy of two random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ with probability mass functions $\boldsymbol{p(x)}$ and $\boldsymbol{p(y)}$ is defined by (Joint entropy) $\boldsymbol{H(X,Y) = -\sum_{x,y} p(x,y) \cdot log_2 \cdot p(x,y).}$ Conditional entropy $\boldsymbol{H(X|Y)}$ is the entropy of a random variable $\boldsymbol{X}$, given another random variable $\boldsymbol{Y.}$

$\boldsymbol{H(X|Y) = -\sum_{x,y} p(x,y) \cdot log_2 \cdot p(x|y)}$

$\boldsymbol{H(X|Y) = H(X,Y)-H(Y)}$

$\boldsymbol{H(X|Y)}$ and $\boldsymbol{H(X,Y)}$ are Conditional and Joint entropies respectively.

H(X,Y) = H(Y,X)

$H(X|Y) \neq H(Y|X)$ [equality is obtained in and only if $H(X) = H(Y)$]

The relative entropy is a measure of the distance between two distributions. The relative entropy ***D(p||q)*** is a measure of the inefficiency of assuming that the distribution is ***q*** when the true distribution is ***p***. The relative entropy or ***Kullback-Leibler distance*** between two probability mass functions ***p(x)*** and ***q(x)*** is defined as $\boldsymbol{D(p||q) = \sum p(x)\log(p(x)/q(x))}$ (Cover and Thomas, 1991). Relative entropy is always non-negative and is zero if and only if $\boldsymbol{p = q}$. However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. The reduction in uncertainty ***X*** due to the knowledge of ***Y*** random variable is called the mutual information**.** For two random variables ***X*** and ***Y*** this reduction is

$$\boldsymbol{I(X;Y) = H(X) - H(X \mid Y)}$$

$$\boldsymbol{I(X;Y) = H(Y) - H(Y \mid X)}$$

$$\boldsymbol{I(X;Y) = H(X) + H(Y) - H(X,Y)}$$

$$\boldsymbol{I(X;Y) = \sum_{x,y} p(x,y)\log_2(p(x,y)/p(x)p(y))}$$

$\boldsymbol{p(x,y)}$ = Joint probability mass function

$\boldsymbol{p(x)}$, $\boldsymbol{p(y)}$ = Marginal probability mass functions

The mutual information ***I(X;Y)*** is the relative entropy between the Joint distribution and the product distribution (Figure 1.14). ***I(X;Y)*** is a measure of the dependence between the two random variables. It is symmetric in ***X*** and ***Y*** and always non-negative.

$$\boldsymbol{I(X;X) = H(X) - H(X \mid X) = H(X)}$$

Thus the mutual information of a random variable with itself is the entropy of the random variable. This is the reason that entropy is some times referred to as ***self-information***. Mutual information has been used to calculate the channel capacity in communications engineering.

Channel capacity, is the maximum amount of discrete information rate that can be reliably transmitted over a channel

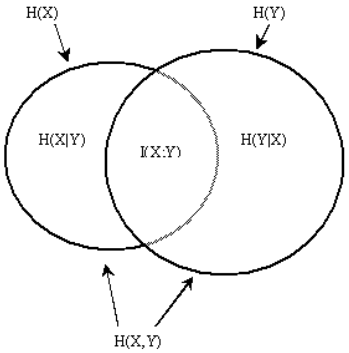


Figure. 1.14. Relationship between entropy and mutual information (Cover and Thomas, 1991)

By the noisy-channel coding theorem, the channel capacity of a given channel is the limiting information transport rate (in units of information per unit time) that can be achieved with vanishingly small error probability. Shannon described the notion of channel capacity and provided a mathematical model by which one can compute the maximal amount of information that can be carried by a channel. The capacity of the channel is given by the maximum of the mutual information between the input and output of the channel, where the maximization is with respect to the input distribution.

1.4 Literature review

Analysis of promoter region (the non-coding) by using the concept of substitution matrix is an important method to identify the similarity between promoter elements of different species. These substitution matrices are used to score sequence similarity, database search (like BLAST and FASTA) and also for finding DNA binding sites in protein

sequences (Ahmad and Sarai, 2005). The elements of these substitution matrices are explicitly calculated from target frequencies of aligned nucleotides and observed frequencies of the nucleotides. The information in these matrices depends on the quantification approach like evolutionary models, structural properties and chemical properties of aligned sequences (Altschul, 1993; Nicholas, ***et al.,*** 2000; Panchenko and Bryant, 2002; Yu, ***et al.,*** 2003; Yu and Altschul, 2005). These are PAM, Point Accepted Mutation (Dayhoff, ***et al.,*** 1978; Schwartz and Dayhoff, 1978), BLOSUM, BLOcks SUbstitution Matrices (Henikoff and Henikoff, 1992), GONNET Matrix (Gonnet, ***et al.,***1992), and DNA Identity matrix. The PAM matrices are based on alignments of closely related sequences and by using these PAM matrices one can estimate target frequencies to any desired evolutionary distance by extrapolation. But in case of BLOSUM the estimation of target frequencies is avoided such extrapolation for different evolutionary distances, it uses the ungapped segments of multiple sequence alignments of protein families.

Neighbor-independent and neighbor-dependent nucleotide substitution matrices have been also used to describe the non-coding sequences (Lunter and Hein, 2004; Stormo, 2000; Arndt and Hwa, 2005) like core promoter region (Reddy, ***et al***., 2006a) and TFBS (Stepanova, 2005; Gershenzon, ***et al.,*** 2005; Staden, 1988; Bulyk, ***et al***, 2002; Reddy, ***et al.,*** 2006b). TFs were also structurally classified based on sequence features of TFBS (Narlikar and Hartemink, 2006) and also showed that a pair of TFs may have a co-localized TFBS (Hannenhalli and Levy, 2002). All protein-coding genes have at least one or more TSS regions, which are active under different conditions. There are several attempts to study the TSS with the help of nucleotide frequencies (Majewski and Ott, 2002;

Bajic, ***et al.,*** 2002, 2003; Aerts, ***et al.,*** 2004) and the DNA weight matrix methods (Bucher, 1990; Down and Hubbard, 2002) around the TSS but it is poorly understood due to the lack of proper signal in the TSS. Mutual information is used to identify the co-evolving functional residues in protein sequences (Martin, ***et al***., 2005), the gene mapping of complex diseases in population based case-control studies (Dawy, ***et al***., 2006) and along with signal processing techniques (Fourier analysis) the mutual information is used to identify homologous DNA sequences (Cristina, ***et al***., 2005). Information content of whole genomes were studied against random sequences and showed that complete genomes have much greater information content than that of random sequences (Chang, ***et al,*** 2005). Information theory based measure is used to score the alignment of ESTs and full-length mRNAs to genomic sequences to study the sequence similarity (Zhang and Gish, 2006). Construction of gene regulatory network from expression data, by using a novel clustering technique has been carried out based on mutual information minimization (Zhou, ***et al***., 2003). The mutual information (provides a general measure of dependencies between variables) from finite data is used in clustering co-expressed genes, which usually requires the definition of 'distance' or 'similarity' between measured datasets (Steuer, ***et al***, 2002).

1.7 Biological Databases

A database is an organized collection of data. A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrive components of the data stored within the system (Table 1.7 and 1.8). The most popular data formats in bioinformatics include `FASTA, PHYLIP, MAML` (Microarray Markup Language), `NEXUS, PAUP, FASTA+GAP`,

Table 1.7. Some of the publicly available Bioinformatics databases accessible ***via*** the Internet.

Database type	Example	Note
Nucleotide Sequence	GenBank	One of the largest Public Sequence Databases
	DDBJ	DNA databank of Japan
	EMBL	European Molecular Biology Laboratory
	MGDB	Mouse Genome Database
	GSX	Mouse Gene Expression Database
	NDB	Nucleic Acid Database
Protein Sequence	SWISS-PROT	Swiss Institute for Bioinformatics and European Bioinformatics Institute
	TrEMBL	Annotated supplement to SWISS-PROT
	PIR	Protein Information Resource
3D Structures	PDB	Protein Databank
	MMDB	Molecular Modeling Database
Sequence motifs (Alignment)	PROSITE	Sequence Motifs
	Pfam	Protein families database of alignments and hidden Markov models
	ProDOM	Protein domains
Molecular Disease	OMIM	Online Mendelian Inheritance in Man
Gene Expression	GEO	Gene Expression Omnibus

Table 1.8. Some of the Bioinformatics tools used to analyze non-coding regions of the genome.

Tools with databases	**www-site**
TRANSFAC (transcription factor database): MATCH and PATCH search tools	http://www.gene-regulation.com/pub/databases.html
MEME and MAST: "biological sequence motif discovery tool" and "Motif Alignment and Search Tool"	http://meme.sdsc.edu/meme/intro.html
Regulatory Sequence Analysis Tools: detect regulatory signals in non-coding sequences	http://rsat.ulb.ac.be/rsat/
EMBOSS: dreg, fuzznuc, tfscan	http://emboss.sourceforge.net/
MatInspector (Genomatix): "fast and versatile tool for detection of consensus matches in nucleotide sequence data"	http://www.genomatix.de/products/MatInspector/index.html
PatScan: search protein or DNA sequence databases for a pattern (ANL)	http://bighost.area.ba.cnr.it/BioWWW/patscanGCG.html
TESS: Transcription Element Search System	http://www.cbil.upenn.edu/cgi-bin/tess/tess

and MmCIF. Some formats are specific to particular data types and applications. For example, MmCIF is used to describe 3D structures,

where as FASTA is used to describe sequences data. The FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (>) symbol.

Although non-coding DNA constitutes majority of most eukaryotic genomes, relatively little is known about its function or the nature of its functional classification. The main focus of our study is to find the statistical behavior of the core promoter elements in different species and the Functional classification of human and mouse TFBS with the help of average mutual information content, which is calculated by using neighbor-independent (4×4 matrices) and neighbor-dependent (16×16 matrices) nucleotide substitutions (Lunter and Hein, 2004; Arndt and Hwa, 2005).

1.8 Objectives

Based on the above knowledge we set few objectives for the present work.

- Comparative analysis of the core promoter region
- Comparative analysis of the TSS
- Functional classification of TFBS

Materials and Methods

2. Materials and methods

2.1 Databases

We have used various (open source) databases that are freely available from the internet. promEC (Hershberg, *et al.,* 2001), PlantPromDB (Shahmuradov, *et al.,* 2003), Eukaryotic Promoter Database-EPD (Périer, *et al.,* 1998), JASPAR (Sandelin, *et al.,* 2004) and TRRD (Kolchanov, *et al*., 2002) contains information on promoter sequences. At NCBI, many of databases are linked through a unique search and retrival system, called *Entrez* (www.ncbi.nlm.nih.gov).

2.1.1 Databases used in core promoter analysis

PlantPromDB

PlantPromDB, a plant promoter database, is an annotated, non-redundant collection of proximal promoter sequences for RNA polymerase-II with experimentally determined TSS from various plant species. PlantPromDB contains entries including promoters from monocot, dicot and other plants. It provides DNA sequence of the promoter regions (-200 : +51) with TSS on the fixed position, taxonomic/ promoter type classification of promoters and Nucleotide Frequency Matrices for promoter elements: TATA-box, CCAAT-box and TSS-motif (*Inr*) (http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom).

EPD

The Eukaryotic Promoter Database (EPD) is an annotated collection of eukaryotic RNA polymerase-II promoters. The TSS have been determined experimentally. EPD is a rigorously selected database in order to include a promoter in EPD, promoter must be recognized by

eukaryotic RNA Polymerase-II, active in a higher eukaryote, experimentally defined, or homologous and sufficiently similar to an experimentally defined promoter, biologically functional, available in the current EMBL release, distinct from other promoters in the database. (http://www.epd.isb-sib.ch/).

PromEC

PromEC is an updated compilation of *E.coli* mRNA promoter sequences. It includes documentation on the location of experimentally identified mRNA transcriptional start sites on the *E. coli* chromosome, as well as the actual sequences in the promoter region. The database currently includes 472 entries (http://margalit.huji.ac.il/).

2.1.2 Databases used in Functional Classification of TFBS

JASPAR

JASPAR is a collection of TFBS that are represented by position-specific weight matrices (PSSMs). JASPAR is small database and contains non-redundant TFBS sequences. All profiles are derived from published collections of experimentally defined TFBS for multicellular eukaryotes. The database represents a curated collection of target sequences. The binding sites were determined either in SELEX experiments, or by the collection of data from the experimentally determined binding regions of actual regulatory regions. It is an open source database (http://mordor.cgb.ki.se/cgibin/jaspar2005/jaspar_db.pl).

TRRD

TRRD is a unique information resource, accumulating information on structural and functional organization of transcription regulatory regions of eukaryotic genes. Only experimentally confirmed information is included into TRRD (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/).

2.1.3 Mitochondrial genome sequences

The NCBI Entrez Genome Project database is intended to be a searchable collection of complete and incomplete (in-progress) large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms. The database is organized into organism-specific overviews that function as portals from which all projects in the database pertaining to that organism can be browsed and retrieved. *Entrez* is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others. The whole genomes of over 1000 organisms can be found in Entrez Genome. All three main domains of life - bacteria, archaea, and eukaryota - are represented, as well as many viruses and organelles (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome).

2.2 Methods

2.2.1 Sequence alignment

Multiple nucleotide or amino acid sequence alignment techniques are usually performed to fit one of following scopes.

1) In order to characterize protein families, identify shared regions of homology in a multiple sequence alignment (this happens

generally when a sequence search revealed homologies to several sequences).

2) Determination of the consensus sequence of several aligned sequences.
3) Help prediction of the secondary and tertiary structures of new sequences and
4) Preliminary step in molecular evolution analysis using phylogenetic methods for constructing phylogenetic trees.

In the present work the promoter sequences obtained from the databases were already aligned sequences. Each promoter sequence has information about the position of TSS. For our computational purpose we have extracted blocks of nucleotides to calculate the nucleotide frequency.

Promoter Sequences from `PlantPromDB`

The promoter sequences obtained from the databases are in `FASTA` format. For the computational purpose we have extracted sequences as a block of different sizes. The single line description of the sequence contains the information of the TSS position. The description line is distinguished from the sequence data by a greater-than ('>') symbol (Figure 2.1).

Aligned Sequences

Aligned sequences are represented by block of nucleotides. Each row in the block represents a different sequence and each column is position of the nucleotide in the sequence (Figure 2.2).

>PLPR0001 ..AC:AB001920 ..OS:Oryza sativa ..GENE:phospholipase D .PROD:phospholipase D .. [-200: +51] ..CDS:+355 ..TSS:201 (+1)

ggcgctggctccgcgacgcgacgtcgcggtcatggagtaaccgcgacggacag
atacttctacccgtttttaacctcgcctcctcctcctcccggctcgagatccg
tggccacgacgcgtggtgggaaaccgggaacgacgtgcacgcacgcacacagg
gcaagtttcagtagaaaaatcgccggcatccagatcgggacagtctctcttct
cccgcaattttataatctcgctcgatccaatctgctccc

>PLPR0002 ..AC:AB004648 ..OS:Oryza sativa ..GENE:RepA ..PROD:cysteine endopeptidase ..[-200: +51] ..CDS:+246 ..TSS:201 (+1)

atgatcgccaaccgccaaaaagtttgtttctaacttcccgccatctcccacct
ctcccgccatctatctctgacctctcctctccacatgctcatcgatcagcttc
cctttcactgctcaaatggcgttttgccgccccacgctacaactcttatcata
tcttcgcctatatatatcgctcaagttccagtttcgtttcagctcattgaact
gaataaagtcggtgtgctatagctaccattcccccggac

>PLPR0003 ..AC:AB013815 ..OS:Arabidopsis thaliana ..GENE:DREB1A ..PROD:DREB1A ..[-200: +51] ..CDS:+140 ..TSS:201 (+1)

gaataagtcaaaaaaagtcttctctggacacatggcagatcttaatgagtgaa
tccttaaactactcattttacaattgcttcgctgtgtatagtttacgtggcat
taccagagacacaaactccgtcttcgccttttcttttgcctctaaaatatctt
ccgccattataaaacagcatgctctcactccaacttttatttatctacaaaca
ttaaatccacctgaactagaacagaaagagagagaaact

>PLPR0004 ..AC:AB013817, AB007789 ..OS:Arabidopsis thaliana ..GENE:DREB1C ..PROD:DREB1C ..[-200: +51] ..CDS:+153

..TSS:201 (+1)

```
cagaaacttcccaagatgggtcaaaggacacatgtcagattctcagtgattga
cagccttgataattacaaaaccgtgggatcgcttagctgtttcttatccacgt
ggcattcacagagacagaaactccgcgttcgaccccacaaatatccaaatatc
ttccggccaatataaacagcaagctctcactccaacatttctataacttcaaa
cacttacctgaattagaaaagaaagatagatagagaaat
```

Figure 2.1. The sequences in database are in FASTA format. This is a standard data format for use with online sequencing databases.

```
Sequence-1   g g g a c a g t c t c
Sequence-2   t t t c a g c t c a t
Sequence-3   t t a t t t a t c t a
Sequence-4   a t t t c t a t a a c
    …        . . . . . . . . . . .
Sequence-n   . . . . . . . . . . .
```

Figure 2.2. The block of 11 nucleotides from the PlantpromDB are aligned based on the position of TSS that represents +1. Each row represents different sequence.

2.2.2 Construction of nucleotide substitution matrices

Sequence comparisons are meaningful only if we have some idea of the similarity between different residues/bases. This information about the similarity of the bases must be derived in a contextual fashion. The coding regions and non-coding regions must be compared using a similarity matrix specifically designed for this purpose. For example, a substitution matrix constructed for the coding regions may perform poorly for the non-coding sequences and vice-versa. For this reason, we have constructed a set of substitution matrices and calculated average mutual information content of core promoter elements (TSS region,

TATA-box, downstream region) and TFBS that are non-coding regions of the protein-coding genes. First we may see the protocol for single base substitution matrices (Figure. 2.3A). These will be 4×4 matrix and lack any preferences (this is the standard assumption made in all sequence alignments that the neighboring bases show no preferences). In other words, adjacent bases are considered independent. Next we see the formulae (they will be very similar except the subscripts will now be pairs) for the base pairs taken together, which correspond to a nearest neighbor preference (Figure. 2.3B). As we are considering a pair, there will be 16×16 matrix. These matrices include adjacent pair preferences explicitly. The blocks of promoter sequences were extracted as 5, 11 and 15 nucleotide blocks for our computational purposes. The following figure shows the principle of counting nucleotide frequencies.

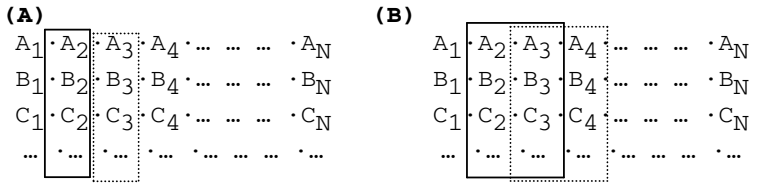


Figure 2.3. The principle of counting the frequencies illustrated diagrammatically. (A) The left side diagram shows the counting principle for neighbor-independent frequency determination. The three lines show the nucleic acid bases corresponding to the already aligned DNA sequences in the database. The solid box is used for determination of the actual frequencies and the counts for A_2-B_2, A_2-C_2 and B_2-C_2 are put in a 4×4 matrix. Then the counting box is shifted by one position (dotted box) and the process is repeated. (B) In the right side illustration, we indicate the counting principle for neighbor-dependent (pair-wise) determination of frequencies. In this illustration, we get the actual counts for A_2A_3-B_2B_3, A_2A_3-C_2C_3, B_2B_3-C_2C_3 and these are placed in a 16×16 matrix. The counting box is next moved right by one base position (shown by the dotted box) and the process continued till the DNA sequence region is completed.

2.2.2.1 Neighbor-independent substitution matrices

For each column of the block, we first count the number of matches and mismatches of each type between the first sequence and every other sequence in the block. This procedure is repeated for all columns of the block with the summed results stored in a 4×4 matrix. For all sequences in the aligned sequences, the same procedure is followed summing these numbers with those that already in the 4×4 matrix. While calculating matches and mismatches the sliding window of one nucleotide along the sequence is used to count the all-possible pairs in a given block. The total number of nucleotide pairs (observed frequency) in a given block is $\frac{1}{2} \cdot w \cdot s \cdot (s-1)$ and the total number of nucleotides (expected frequency) in the block is $w \cdot s$, where s is the number of nucleotides in the given position and w is the block width. The resulting matrix (4×4 matrix) is used to calculate the odds-ratio between those observed frequencies $q(ij)$ and those expected by chance $p(i)$. This odds ratio $q(ij)/(p(i) \cdot p(j))$ is also called a likelihood ratio. Then "log-odds" is calculated (usually logarithm to the base 2) from the odds-ratio and is given by $s(ij)=\log_2(q(ij)/p(i) \cdot p(j))$. Such probabilities (odds ratios) should be multiplied or log-odds can be added to get the probability of their independent occurrence (Karlin and Altschul, 1990; Altschul, 1991).

2.2.2.2 Neighbor-dependent substitution matrices

The incorporation of the pair-preferences into the substitution matrix gives neighbor-dependent substitution matrices. These are very similar to neighbor-independent substitution matrices except that the subscripts will be pairs of nucleotides. While calculating matches and mismatches the sliding window of one nucleotide along the sequence is used to count of

all possible pairs in the given block. The total number of dinucleotide pairs (observed frequency) in a given block is $½ \cdot s \cdot (w-1) \cdot (s-1)$ and the total number of dinucleotides (expected frequency) is given by $s \cdot (w-1)$, where s is the number of sequences and w is the block width. The resulting matrix (16×16 matrix) is used to calculate the odds-ratio between those observed frequencies $q(ij,kl)$ and those expected by chance $p(ij)$. This odds ratio $q(ij,kl)/(p(ij) \cdot p(kl))$ (also called likelihood ratio) is then used to calculate the "log-odds" and is given by $s(ij,kl)=log_2 q(ij,kl)/(p(ij) \cdot p(kl))$

2.2.3 Average mutual information content (*H*)

The comparison of these non-coding regions can be performed either by scores in the substitution matrix themselves or by the information content of these substitution matrices. In information theoretic terms average mutual information content (*H*), is the relative entropy of the target and background pair frequencies and can be thought of as a measure of the average amount of information (in bits) available per nucleotide pair. In neighbor-independent substitution matrices, the log-odds of each nucleotide pair $s(ij)$ (in the units of $\log_2$, called bits) multiplied by the probability of occurrence of that pair $q(ij)$ will give the weighted score and is then summed overall for the nucleotide pairs to produce a score that represents the ability of the average nucleotide pair in the matrix to discriminate the actual alignment from chance alignments. The average mutual information content is given by $H = \sum_{ij} q(ij) \cdot s(ij)$. The higher the value of the relative entropy of target and background distributions, the more easily they are distinguished (Altschul, 1991). The same procedure is applied for calculating the average mutual information content in the

case of neighbor-dependent substitution matrices. The average information content in neighbor-dependent substitution matrices is given by $H = \sum_{ij,kl} q(ij,kl) \cdot s(ij,kl)$. The maximum value of H in DNA is 2 bits in neighbor independent nucleotide substitutions (4 bits in case of nucleotide dependent substitutions) and is calculated by

$$H = -\sum_{i=1}^{4} p_i \log_2 p_i$$

$$H = \sum_{4} \frac{1}{4} \log_2 \frac{1}{4} = 2 \text{ bits}$$

$$H = \sum_{16} \frac{1}{16} \log_2 \frac{1}{16} = 4 \text{ bits}$$

The maximum value of H occurs when all the 4 nucleotides or 16 nucleotide pairs are in equiprobable distribution (completely random).

2.2.4 Standard error calculation

To assess the reliability of our computations, we have performed a simple error analysis of the results. We consider the matrix elements $H(ij)$ of the information content matrix $s(ij) \cdot q(ij)$ as the elements of our data and compute the standard error of the 16 (or 256 in case of the pair preferences) elements using standard techniques. The standard errors are plotted in the graph along with the histograms.

$$\sigma^2 = \frac{1}{n} \sum_{1}^{n} (x - \bar{x})^2$$

$$\text{Standard error} = \frac{\sigma}{\sqrt{n}}$$

x = Sample value; $\bar{x}$ = Sample mean

σ = Standard deviation; σ^2 = Variance; n = Sample size

2.2.5 Information content calculation with example

The information content is calculated either from neighbor dependent or independent nucleotide substitution matrices. The procedure to calculate

the information content of nucleotide neighbor independent substitutions is given below.

Step-1: Sequence alignment

The sequences are aligned with respect to TSS. The following 10 sequences are a block of 6 nucleotides (Figure 2.4).

```
GTACTA
ATCTCC
GTACTC
AACGAT
GGATTC
GTTGAT
CTACTC
TCGCGC
CACATG
GTACCA
```

Figure 2.4. Block of nucleotides (10 sequences of length 6 nt). This is a random example for illustration purpose only.

Step-2: Nucleotide frequency calculation

Consider the first column from aligned block of nucleotides and calculated nucleotide frequencies (Figure 2.5). The nucleotide frequency count is asymmetric; *i.e.,* AT and TA are different. There are 16 possibilities for nucleotide neighbor independent substitutions (256 in case of neighbor dependent substitution). These frequency counts are stored in either 4×4 (in case of neighbor dependent 16×16 matrix). The nucleotide frequency calculation for neighbor-dependent case (16×16 matrix) is not shown in the present example.

Step-3: Observed and expected frequency

The nucleotide independent substitution frequencies are stored in a 4×4 matrix (Table 2.1). These will give observed frequency of each of the possible pairs (16 pairs). The individual nucleotides are also counted, which will give observed frequency (Table 2.2).

Seq.	1st.	Possible nucleotide pairs
Seq1	**G**	GA, GG, GA, GG, GG, GC, GT, GC, GG
Seq2	**A**	AG, AA, AG, AG, AC, AT, AC, AG
Seq3	**G**	GA, GG, GG, GC, GT, GC, GG
Seq4	**A**	AG, AG, AC, AT, AC, AG
Seq5	**G**	GG, GC, GT, GC, GG
Seq6	**G**	GC, GT, GC, GG
Seq7	**C**	CT, CC, CG
Seq8	**T**	TC, TG
Seq9	**C**	CG
Seq10	**G**	

Figure 2.5. Calculation of nucleotide (neighbor independent) frequency. Here we consider first column of the aligned nucleotides. The same procedure is used for all the columns of the block. This construction avoids double counting. (Seq = Sequence; 1st = First column of the block)

	A	**C**	**G**	**T**
A	14	20	15	16
C	20	25	14	16
G	8	17	11	9
T	19	25	14	27

Table 2.1. The calculated nucleotide frequencies are stored in a 4x4 matrix. All the frequencies of a block are added. Total of all the frequencies…4+20+15+16+ 20+….= ½·6·10· 9

A	**C**	**G**	**T**
14	18	11	17

Table 2.2. The total number of individual nucleotides. Total of 14+18+11+17=60.

Total number of nucleotide pairs in the block

$= ½ \cdot w \cdot s \cdot (s-1)$

$= ½ \cdot 6 \cdot 10 \cdot (10-1)$

$= 270$

Total number of nucleotides in the block

$= w \cdot s$

$= 6 \cdot 10$

$= 60$

s = number of nt in the block

w = block width.

Observed frequency ($p(ij)$) of GT $= p(GT)$

$= 9/270$

Expected frequency (p(i).p(j))of GT $= p(G) \cdot p(T)$

$= (11/60).(17/60)$

Odds ratio $= p(GT)/p(G) \cdot p(T)$

Step-4: Odds-ratio calculation

Odds ratios are calculated by observed frequency/expected frequency of nucleotides and stored in the same 4×4 matrix (Table 2.3). log-odds is calculated as a $\log_2$ (Table 2.4).

	A	C	G	T
A	0.9522	1.0582	1.2987	0.8963
C	1.0582	1.0288	0.9427	0.6971
G	0.6926	1.1447	1.1447	0.6417
T	1.0644	0.9982	0.9982	1.2456

Table 2.3. Odds rations are calculated from the aligned sequences.

log odds ratio $s(ij) = \log_2(p(ij)/p(i).p(j))$

	A	C	G	T
A	-0.0703	0.0816	0.3770	-0.1578
C	0.0816	0.0409	-0.0850	-0.5204
G	-0.5298	0.1950	0.2775	-0.6400
T	0.0900	0.1234	-0.0025	0.3169

Table 2.4. log-odds ratios are calculated as $\log_2$

$$H(ij) = p(ij).s(ij)$$

Step-5: Information content

The information content is calculated as $H(ij) = p(ij) \cdot s(ij)$ (Table 2.5). By adding all these $H(ij)$ of a matrix will give H. Standard error is calculated from the $H(ij)$ data.

	A	C	G	T
A	-0.0036	0.0060	0.0209	-0.0093
C	0.0060	0.0038	-0.0044	-0.0308
G	-0.0156	0.0123	0.0113	-0.0213
T	0.0063	0.0114	-0.0001	0.0317

Table 2.5. Information content of $H(ij)$. As we have used $\log_2$, the units are in *bits*.

Average mutual information content and standard error are calculated by using all the $H(ij)$ values in the above table (Table 2.5).

Information content (H) = 0.024464 bits and Standard error = 0.003809

2.2.6 BLAST

Basic Local Alignment Search Tool (BLAST) provides a method for rapid searching of nucleotide and protein databases. The BLAST algorithm detects local as well as global alignments and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify

members of gene families. There are several of different BLAST programs (Table 2.6). The default matrix for all protein-protein comparisons is BLOSUM62.

Table 2.6. The BLAST family of programs allows all combinations of DNA or protein query sequences with searches against DNA or protein databases.

BLAST program	**Description**
blastp	compares an amino acid query sequence against a protein sequence database.
blastn	compares a nucleotide query sequence against a nucleotide sequence database.
blastx	compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
tblastn	compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

2.2.7 Clustering and Classification

Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - some defined distance measure. Data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine

all clusters at once. Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. A key step in a hierarchical clustering is to select a distance measure. A simple measure is manhattan distance, equal to the sum of absolute distances for each variable. A more common measure is Euclidean distance, computed by finding the square of the distance between each variable, summing the squares, and finding the square root of that sum. In the two-variable case, the distance is analogous to finding the length of the hypotenuse in a triangle. The UPGMA is a bottom-up and Hierarchical algorithm. UPGMA -Unweighted Pair Group Method with Arithmetic mean (Sokal and Michener, 1958) is a simple clustering method used for the creation of phylogenetic trees. It is based on a set of pair-wise distances between sequences (a distance matrix d_{ij} is the distance between sequences *i* and *j)*. There are several ways of calculating distances, of which the most common is based on the percentage identity of the sequences.

2.2.7.1 UPGMA algorithm

1. Begin with N sequences.
2. Find two closest sequences *i,j* and define them as a cluster in the first round, Now we now have N-2 sequences and a cluster.
3. Recalculate distances (Distances to a cluster are averages (mean) for sequences in the cluster).
4. Create a new internal node with daughter nodes *i, j* at height $d_{ij}/2$.
5. Iterate (2-4) until only two clusters remain.
6. Place root midway between them the two remaining clusters.

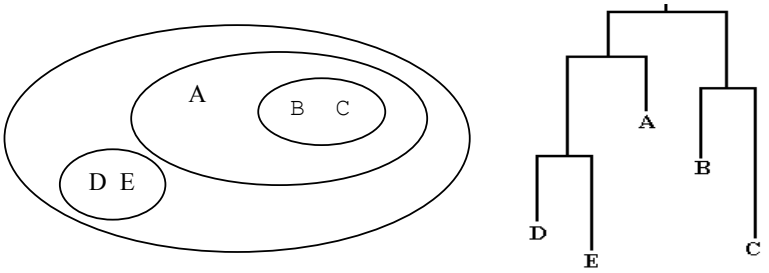


Figure 2.6. The UPGMA method The clustered results are represented by either of these two figures or as a text, like ((DE)(A(BC))).

In summary UPGMA

1. Places all leaves at the same level.
2. When two sequences are joined in a cluster, it is assumed that the common ancestor is equidistant from each sequence.
3. Assumes a "molecular clock" in which rate of evolution is the same on each branch of the tree.

Here the input data is a collection of information content values of TFBS and the output is a rooted tree like structure. Initially, each object is in its own cluster. At each step, the nearest two clusters are combined into a higher-level cluster. The distance *d(ij)* between any two clusters *C(i)* and *C(j)* is taken to be the average of all distances between pairs of objects from each cluster. $d(ij) = (1/(|C(i)|\ |C(j)|))\Sigma_{(p \text{ in } C(i),\ q \text{ in } C(j))}\ d_{pq}$. Where $|C(i)|$ and $|C(j)|$ denote the number of sequences in clusters *i* and *j*, respectively.

2.2.7.2 PHYLIP

PHYLIP is a free package of programs for inferring phylogenies and carrying out certain related tasks (Felsenstein, 2004) At present it contains 31 programs, which carry out different algorithms on different

kinds of data. DRAWGRAM is one of the programs in the package. DRAWGRAM interactively plots a cladogram- or phenogram-like rooted tree diagram, with many options including orientation of tree and branches, style of tree, label sizes and angles, tree depth, margin sizes, stem lengths, and placement of nodes in the tree.

Cladogram

Nodes are connected to other nodes and to tips by straight lines going directly from one to the other. This gives a V-shaped appearance. The default a setting is with no branch lengths yield a V- shaped tree with a 90-degree angle at the base.

Phenogram

Nodes are connected to other nodes and to other tips by a horizontal and then by a vertical line. This gives a particularly precise idea of horizontal levels.

In the present work, We have used only the UPGMA and plotting packages from this suite. Distance matrix is developed for drawing the tree with DRAWGRAM of PHYLIP.

2.2.8 C++ / LINUX

C++ is an object oriented programming language. The object oriented programming involves a concept called a class. A class is like an array; it is a derived type whose elements have other types. But unlike an array, the elements of a class may have different types. An object is a self-contained entity that stores its own data and owns its own functions. All of the above computations (sequence alignments, pair frequency, information content (Figure 2.7 and 2.8) were done in C++.

Part of a C++ Program for Neighbor independent nucleotide substitution calculation

```
char S1* = new char[Z+2];
char S2* = new char[Z+2];
int i = 0;
while(1) // forever loop
{
      if(i == S) break;
      ifile.seekg(i*(Z+1));
      // to reposition the get pointer in the string
      ifile.getline(S1,(Z+2));
      i++;
      int k = i;
      while(1)      // forever loop
      {
            if(k > (S-1)) break; //controls the infile
            infile.seekg(k*(Z+1));
            infile.getline(S2,(Z+2));
            k++;
            for(int q = 0; q <= (Z-1); q++,S1++,S2++)
            {
                  switch(*S1)
                  {
                        case 'A': index1 = 0; break;
                        case 'T': index1 = 1; break;
                        case 'G': index1 = 2; break;
                        case 'C': index1 = 3; break;
                        Default:cout<<"Error"<< endl;
                  }
                  switch(*S2)
                  {
                        case 'A': index2 = 0; break;
                        case 'T': index2 = 1; break;
                        case 'G': index2 = 2; break;
                        case 'C': index2 = 3; break;
                        Default:cout<<"Error"<< endl;
                  }
                  count[index1][index2]++;
                  // frequency count
                  }
            }
```

Figure 2.7. Part of a C++ program for calculating the nucleotide frequencies in nucleotide neighbor independent substitutions.

Part of a C++ Program for Neighbor independent / calculationinformation content / stdard error

```
float A, H = 0, B = 0, xx = 0, size = 0;
float countg[4][4];
float k1, k2, k3, k4;
for(int k = 0; k <= 3; k++)
{
      for(int l = 0; l <= 3; l++)
      {
            k1 = (float) count1[k][l]/seq1;
            k2 = (float) count2[k]/seq2;
            k3 = (float) count2[l]/seq2;
            if(fabs(k1) < 1e-6) k1 = 1e-6;
            if(fabs(k2) < 1e-6) k1 = 1e-6;
            if(fabs(k3) < 1e-6) k1 = 1e-6;
            k4 = log10(k1/(k2*k3))/log10(2.0);
            A = k1*k4;
            H += A;
            B += A;
            xx += A*A;
            size++;
      }
}
float sigma = sqrt(xx/size-(B*B/(size*size)));
cout << "H = " << H << endl;
cout  << "std.err = " << sigma/sqrt(size) << "\n";
```

Figure 2.8. Part of a C++ for calculating the information content and its standard error (neighbor-independent). The zero or less than zero's in the matrix are replace with (1e-6) to avoid the errors in calculation.

Results

3. Results

3.1 Comparative analysis of the core promoter region

To understand the unique features of the core promoter region we have analyzed a set of five organisms by calculating information content of core promoter elements (TATA-box, TSS-region and a Downstream region). We extracted promoter sequences from several sources.

3.1.1 Blocks of nucleotides from core promoter elements

PromEC - ***E.coli*** promoter database (Hershberg, ***et al.,*** 2001), PlantPromDB - a plant promoter database (Shahmuradov, ***et al.,*** 2003) and EPD - Eukaryotic Promoter Database (Périer, ***et al.,*** 1998**)** include sequences that are annotated, non-redundant promoter sequences with experimentally determined TSS (Table 3.1). The EPD sequences included here are "representative sets of not closely related sequences".

Table 3.1. Databases used in the present study (Taxonomic group or organism with number of sequences used).

Database	**Taxonomic group / organism**	**No. of sequences**
PromEC	***E.coli***	472
EPD	Drosophila	1922
PlantPromDB	Plants	305
EPD	Human	1789
EPD	Mouse	118

The promoter sequences obtained from the databases are aligned sequences and can be represented as ungapped blocks with each row a different promoter sequence and each column an aligned base (for each of the five species). From these aligned sequences we extracted set of

blocks (TATA-box region, TSS-region and Downstream region) or columns of different sizes (5, 11 and 15 nucleotide wide) for computational purposes (Table 3.2).

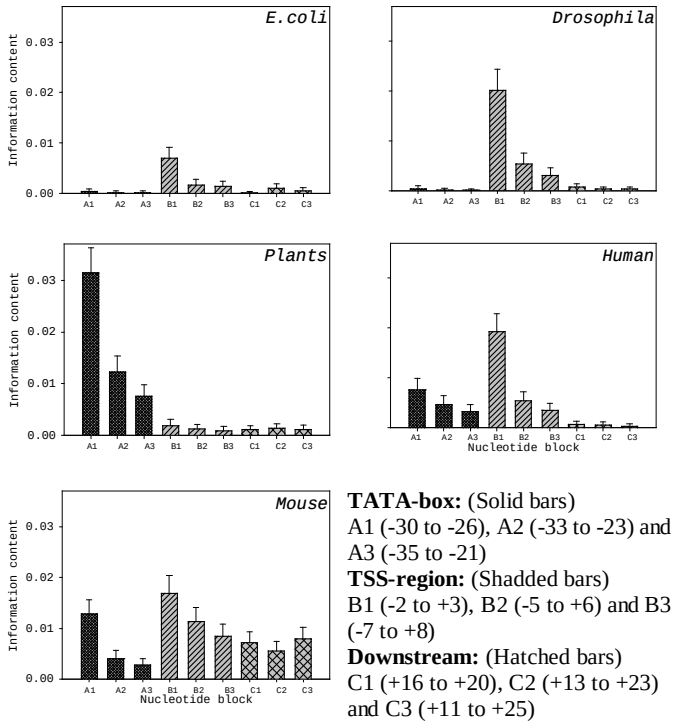
Table 3.2. Blocks of nucleotides (positions are with respect to TSS that represents +1).

Block size	TATA-box	TSS region	Downstream region
5	-30 to -26	-2 to +3	+16 to +20
11	-33 to -23	-5 to +6	+13 to +23
15	-35 to -21	-7 to +8	+11 to +25

In some of the sequences, there are no TATA boxes. For such sequences, the table indicates the expected position.

3.1.2 Information content of core promoter elements

Information content of core promoter elements is calculated from the mono and dinucleotide substitution matrices. This information content of core promoter elements is represented in bar graphs (Figure.3.1 (neighbor-independent) and Figure. 3.2 (neighbor-dependent)). In each graph 'A', 'B' and 'C' group of bars represents block of TATA-box, TSS-region and Downstream region respectively. The subscripts 1, 2 and 3 of A, B and C represent, 5, 11 and 15 nucleotide blocks respectively.



TATA-box: (Solid bars)
A1 (-30 to -26), A2 (-33 to -23) and A3 (-35 to -21)
TSS-region: (Shadded bars)
B1 (-2 to +3), B2 (-5 to +6) and B3 (-7 to +8)
Downstream: (Hatched bars)
C1 (+16 to +20), C2 (+13 to +23) and C3 (+11 to +25)

Figure 3.1. The average mutual information content H (in bits) of core promoter elements (calculated by neighbor-independent nucleotide substitutions) from different datasets. In all the figures 'A', 'B' and 'C' group of bars represent respectively TATA-box, TSS-region and Downstream region. The subscripts 1, 2 and 3 of A, B and C represent, 5, 11 and 15 nucleotide blocks respectively. The bars on top of the histograms represent the standard errors of the 16 $\boldsymbol{H_{ij}}$ values.

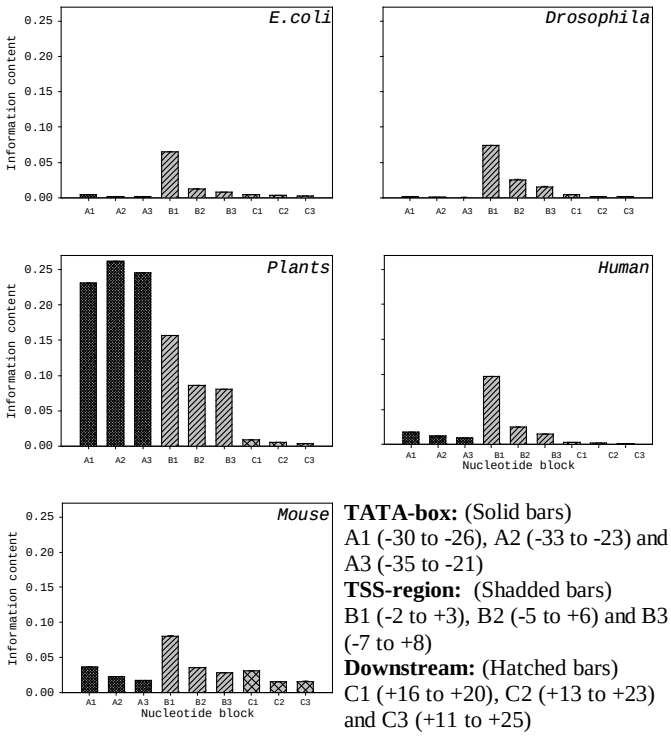


Figure 3.2. The average mutual information content H (in bits) of core promoter elements (calculated by neighbor-dependent nucleotide substitutions) from different datasets. In all the figures 'A', 'B' and 'C' group of bars represent respectively TATA-box, TSS-region and Downstream region. The subscripts 1, 2 and 3 of A, B and C represent, 5, 11 and 15 nucleotide blocks respectively. The bars on top of the histograms represent the standard errors of the 256 H_{ij} values.

3.1.3 Discussion

We notice that the information content decreases with increasing block size. This clearly implies that the TSS-region is likely to be 5-10 bases in size. This pattern is seen in all the species (even in plants where TSS-region evidently play more important role than TATA-box). We also notice that both in the case of mouse and humans, both the TATA-box

and TSS-region are likely to play important roles (probably both are involved in binding with protein factors). We note that TATA-boxes and the TSS are two regions that are physically close together and we do not expect to see a case in which both are relatively less important. The error studies clearly showed that the standard errors are sufficiently small that the overall conclusions are not affected. It is important to note that the different species represent different patterns of binding and it may be futile to look for any consensus sequences that are valid in all the cases. However, it may be still possible to locate some patterns in a very closely related group. In this study, we note that mice and men come close together.

3.2 Comparative analysis of TSS

3.2.1 Sequence data

The promoter sequences of human, mouse, ***E.coli***, and mitochondria (chordate) were obtained from three databases (Table 3.3).

Table 3.3. The number of promoter sequences used with the corresponding databases in the present study.

Organism	**No. Sequences**	**Database**
Human	1789	EPD
Mouse	118	EPD
E.coli	472	PromEC
Mitochondrial	240	Entrez genome

The mitochondrial sequences included here are not more than 1021 bp length that excludes inclusion of tandem repeats, stem loops etc. containing sequences. Sequences from EPD are representative set of not closely related sequences.

3.2.2 Information content of TSS-region

In the present study we have constructed substitution matrices for mono and dinucleotide substitutions in multiple aligned TSS-regions of human, mouse, ***E.coli*** and the mitochondrial control element. Then we calculated the average information content (in bits). Information content of mitochondrial genome (near the control element) as a function of the position is shown in Figure 3.3. In A (left graph), the information content has been computed based on the neighbor independent substitution (4×4) matrix. In B (right graph), neighbor dependent substitutions dinucleotide substitution (16×16) matrix has been used to compute the information content at different positions. The average information content of the

mitochondrial genome near the control element was computed as an average of a 5-nucleotide overlapping block (Figure 3.4). In graph A (left), we have used the neighbor independent substitution matrix (4×4 matrix) whereas in B (right) we have used the neighbor dependent substitution matrix (16×16 matrix) for the computation. The 240 sequences used here are 3' end of control element, which contains the potential promoter region. The average information content of TSS regions of human, mouse and ***E.coli*** as a function of the blocksize is in Figure 3.5. Histograms shown in A_1, B_1 and C_1 (top row) represent information content determined using nucleotide neighbor independent substitution matrix. Similarly, A_2, B_2 and C_2 (bottom row) represent the same as above but using nucleotide neighbor dependent substitution matrix. In each graph the bars represent average information content for blocks of 5 (-2 to +3), 11 (-5 to +6) and 15 (-7 to +8) nucleotides. The positions are with respect to TSS, that represents +1.

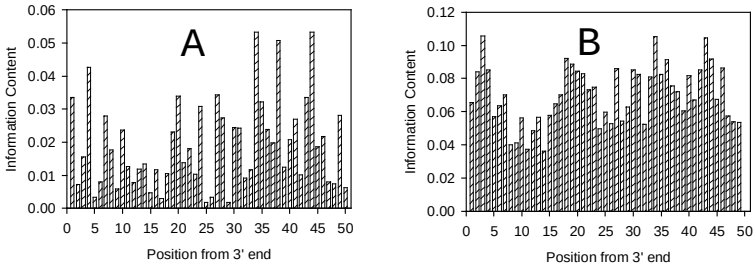


Figure 3.3. Information content of mitochondrial genome (near the control element) as a function of the position is shown. In A (left graph), the information content has been computed based on the single nucleotide using a neighbor independent substitution (4×4) matrix. In B (right graph), neighbor dependent substitutions using a dinucleotide substitution (16×16) matrix has been used to compute the information content at different positions as in A. The 240 sequences used here are 3' end of control element, which contains the potential promoter region.

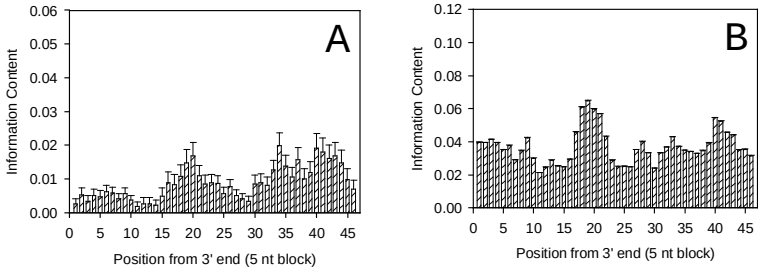


Figure 3.4. The information content of the mitochondrial genome near the control element computed as an average for a 5-nucleotide overlapping block. In graph A (left), we have used the neighbor independent substitution matrix (4×4 matrix) whereas in B (right) we used the neighbor dependent substitution matrix (16×16 matrix) for the computation. The 240 sequences used here are 3' end of control element, which contains the potential promoter region.

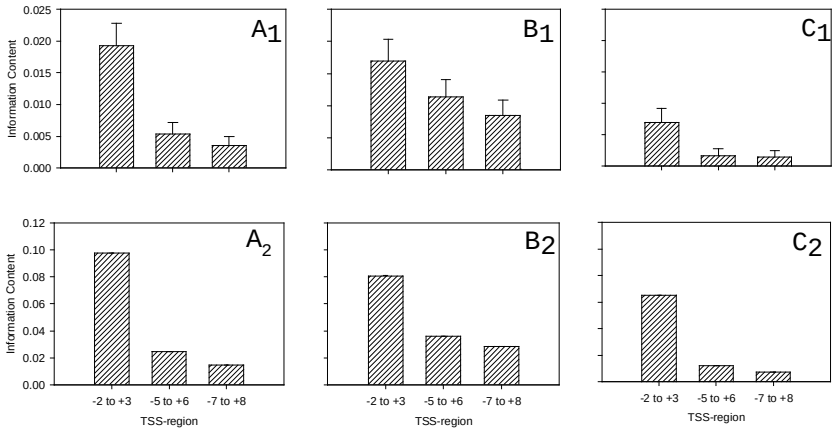


Figure 3.5. The average information content of TSS regions of human, mouse and ***E.coli*** as a function of the blocksize. Histograms shown in A_1, B_1 and C_1 (top row) represent information content determined using nucleotide neighbor independent matrix. Similarly, A_2, B_2 and C_2 (bottom row) represent the same as above but using nucleotide neighbor dependent matrix. In each graph the bars represent average information content for blocks of 5 (-2 to +3), 11 (-5 to +6) and 15 (-7 to +8) nucleotides. The positions are with respect to TSS, that represents +1.

The standard errors are plotted in the in Figure 3.4 and Figure 3.5 (but the error bars cannot be seen in some cases because as they are too small).

3.2.3 Discussion

Analysis of transcriptional regulatory regions in DNA is very important to understand the mechanisms governing gene expression and its regulation. It is well known that the TSS region shows greater variability compared to the other promoter elements and we were interested to search these variable regions by using information content. In this study we observed that the variability is significant in the 5 nt block surrounding the TSS region but with a greater block size (e.g., 15 nt) the variability is not significant. This suggests that the actual region that may be involved in the range of 5-10 nt size. This is apparent from the histograms presented in Figure 3.5. We were also interested to find out whether there are any short-range correlations within the nucleotides in the TSS region. For ***E.coli***, we observed that the information content from dinucleotide substitution matrices clearly show a better discrimination suggesting the presence of some correlation. However, for mouse, this effect is much less and for humans, it is practically absent. We can perhaps safely conclude that the presence of short-range correlations within the TSS region is species dependent and is not universal. We must however, note that we have considered only three species and further studies need to be conducted to establish clear conclusions in this regard. We further note that there are other variable regions in the mitochondrial control element apart from the TSS. In Figure 3.4, we observed a prominent peak around 16-22 nucleotides (for both dependent and independent cases) but similar peaks are also found

near the 38-43, which is known not to contain the TSS. However, this region may also be of interest as other important variable regions. Effective comparisons can only be made on the blocks and the single nucleotide comparison does not give us any detectable signals. The high recombination rate in mitochondrial control region may not alter the overall nature of TSS-region. These results imply a similar regulatory structure in almost all organisms and have been conserved during the evolution due to functional constraints. Finally we came to know that there are well-established tools to locate conserved regions in DNA but looking for variability is also important. We found that information content may be useful to study the variable regions in genome in an efficient manner. We can locate TSS and other variable regions by using this approach.

3.3 Functional classification of TFBS

3.3.1 Information content of TFBS

The sequences in JASPAR database (Sandelin, ***et al.,*** 2004) are annotated and experimentally demonstrated TFBS profiles for multicellular eukaryotes. It is an open-access TF binding profile database that contains more than a hundred TFBS profiles of ***Drosophila melanogaster, Arabidopsis thaliana, Zea mays, Homo sapiens, Mus musculus etc***. We have downloaded and studied only the TFBS for human and mouse (Table 3.4-3.6). The sequences in this database are organized in FASTA format and also contain the frequencies of the four bases for the selected positions.

Table 3.4. Database and the corresponding organism with number of TFBS used in the present study.

Database	Organism	No of TFBS
JASPAR	Human	41
JASPAR	Mouse	13

Each TFBS is a collection of binding sites with already aligned sequences of different lengths. The lengths vary between 5-20 nucleotides and have been used without modifications. This implies that the longer sequences have better recognition properties and are expected to have a lower noise threshold (and vice-versa).

Table 3.5. Human transcription factors with the recognized TFBS and their lengths.

Name of TF	Class of TF	Total of TFBS	Length of TFBS
Elk-1	ETS	28	10
NRF-2	ETS	7	10
SAP-1	ETS	20	9

SPI-1	ETS	57	6
SPI-B	ETS	49	7
FREAC-4	FORKHEAD	20	8
SOX-9	HMG	76	9
SRY	HMG	28	9
Pbx	HOMEO	18	12
MEF2	MADS	58	10
SRF	MADS	46	12
COUP-TF	NUC.RECEPTOR	13	14
PPARgamma-RXRal	NUC.RECEPTOR	41	20
PPARgamma	NUC.RECEPTOR	28	20
RORalfa-1	NUC.RECEPTOR	25	10
RORalfa-2	NUC.RECEPTOR	36	14
RXR-VDR	NUC.RECEPTOR	10	15
p53	P53	17	20
Pax6	PAIRED	43	14
c-REL	REL	17	10
p50	REL	18	11
p65	REL	18	10
AML-1	RUNT	38	9
Irf-2	TRP-CLUSTER	12	18
E2F	Unknown	10	8
MZF_1-4	ZN-FINGER,C2H2	20	6
MZF_5-13	ZN-FINGER,C2H2	16	10
RREB-1	ZN-FINGER,C2H2	11	20
SP-1	ZN-FINGER,C2H2	8	10
Yin-Yang	ZN-FINGER,C2H2	17	6
GATA-2	ZN-FINGER,GATA	53	5

GATA-3	ZN-FINGER,GATA	63	6
Hen-1	bHLH	54	12
Tal 1 beta-E47S	Bhlh	44	12
Thing-E47	bHLH	29	12
Max	bHLH-ZIP	17	10
Myc-Max	bHLH-ZIP	21	11
USF	bHLH-ZIP	30	7
CREB	bZIP	16	12
E4BP4	bZIP	23	11
HLF	bZIP	18	12

Table 3.6. Mouse transcription factors with the recognized TFBS and their lengths.

Name of TF	Class of TF	Total of TFBS	Length of TFBS
SOX17	HMG	31	9
Sox-5	HMG	23	7
EN-1	HOMEO	10	11
Nkx	HOMEO	17	7
S8	HOMEO	59	5
Bsap	PAIRED	12	20
Pax-2	PAIRED	31	8
Brachyur	T-BOX	40	11
Evi-1	ZN-FINGER,C2H2	47	14
ARNT	bHLH	20	20
Ahr-ARNT	bHLH	24	24
n-MYC	bHLH-ZIP	31	31
Spz-1	bHLH-ZIP	12	12

3.3.2 Information content of random sequences

To ascertain the reliability of the results, we have computed the information content based on a sample sequence (of length 20 nucleotides) selected at random. The sample sequence was subjected to a BLAST search against the respective genome (NCBI sample BLAST with default parameters) and BAC clone sequences were excluded from the results. Finally 18 best matches for human genome and 14 best matches for mouse genome were taken. The information content was computed based the neighbor-independent and neighbor -dependent procedures as described above. These values are indicated in the histograms as horizontal dotted lines (Figure 3.6). These values reflect the typical random sequences present in the respective genome as a reference for comparison. The statistical errors (standard errors) are also indicated in the histograms in the conventional way.

3.3.3 TFBS clustering

We have used the distance measures based on the information content to classify the results. We stress that the actual protein sequences were not involved in this computations-only their TFBS. The plots were made using the PHYLIP suite of software (Felsenstein, 2004). The information content calculated for the 41 and 13 TFBS of human and mouse respectively is presented as a histogram in Figure 3.6. Note that the labels in the graphs have been taken from the class-names of the proteins (as given in the database) involved and therefore can occur at multiple places. The dotted line shows typical values of information content for random sequences as a reference of comparison. The error bars (standard errors) calculated on the basis of the elements of the information content

matrix (4×4 matrix for the neighbor-independent and 16×16 matrix for the neighbor- dependent) are shown on the histograms in the usual way (they are sufficiently small to be invisible for the graph on the right Figure. 3.6B). One interesting pattern that was noticed in the two graphs is that they are quite similar but not same (this was also reflected in the next set of figures). In particular when we compare graphs in Figure 3.6 we found that neither the largest peaks nor the smallest peaks correspond with each other. We conclude that the consideration of the neighbor dependence provides additional information but the broad features are similar (strong peaks remain big and weak peaks are also weak in both). We also noted another aspect with respect to the random sequence information content. The random sequence chosen was not expected to correspond any particular TFBS. If we consider the neighbor-independent plot (Figure 3.6: 1A and 2A), the information content of the random sequence is 0.3211 and 0.3863 bits for human and mouse respectively (this corresponds to the dotted horizontal line). We note that these random sequence information content values are close to the mean values of the actual TFBS information content values. When we consider the neighbor-dependent graphs (Figure 3.6: 1B and 2B) the information content of the random sequence is 2.101 and 2.227 bits for human and mouse respectively (this corresponds to the dotted horizontal line). We also note that only one (for humans) or two (for mice) values are above the line. This suggests that there exists a strong and specific correlation between the neighbor nucleotides in the TFBS regions. This correlation is significantly different from the typical genome regions (as represented by the dotted line). The trees (Figure. 3.7 and Figure. 3.8) represent the clustering of TFBS. In a tree, each node with descendants represents the functional group of TFBS that has close information content values. We

believe that one factor may bind to multiple TFBS and cause initiation of transcription of a group of proteins. The results of the clustering suggest that this is likely to be the possible event.

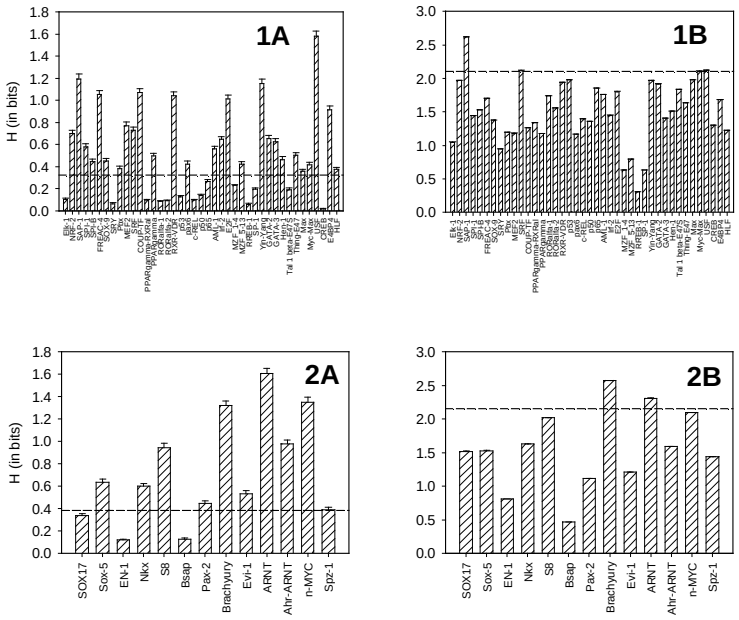
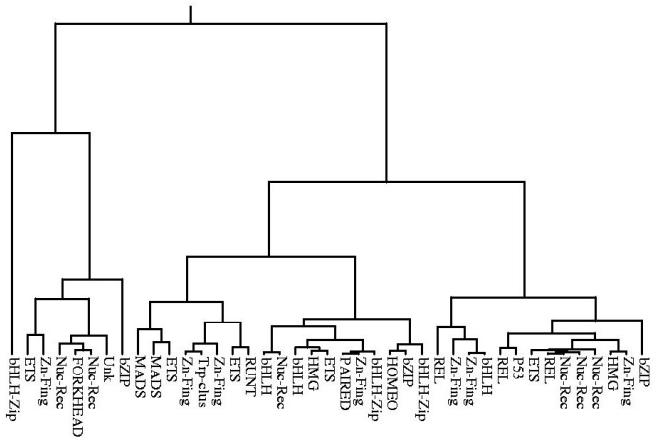


Figure 3.6. The average mutual information content H (in bits) of TFBS (calculated by (A; left) neighbor independent and (B; right) neighbor dependent nucleotide substitutions) of human (1A and 1B) and mouse (2A and 2B). The dotted line represents information content of random sequence of their genome. The bars of the histograms represent the standard errors of the 16 ***H(ij)*** neighbor independent substitution matrices. (256 ***H (ij,kl)*** incase of neighbor dependent). The standard errors have been actually plotted but cannot be seen, as they are too small in case of neighbor-dependent.

A



B

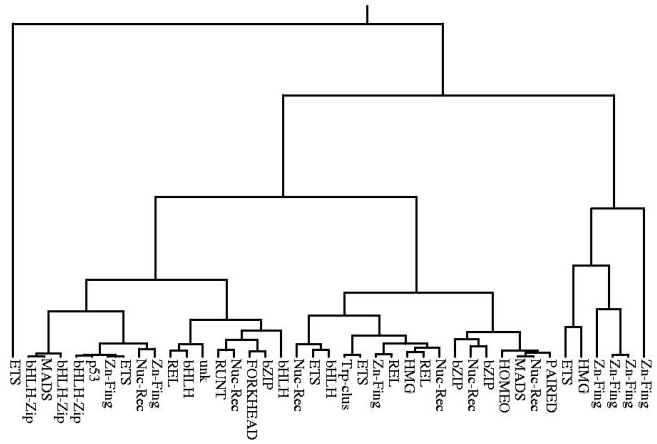
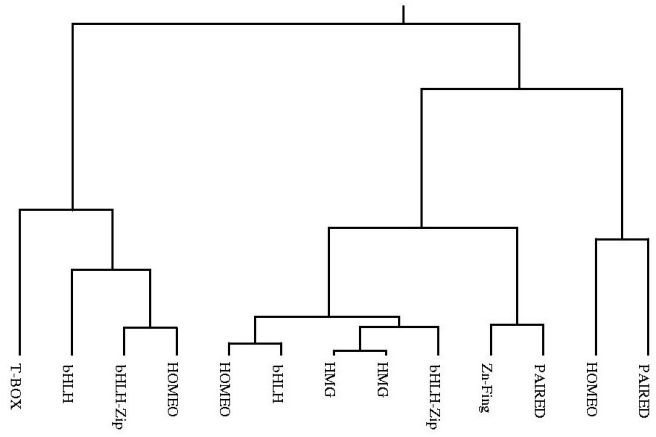


Figure 3.7. Functional classification of TFBS in Human; information content is calculated from nucleotide (A) neighbor-independent and (B) neighbor dependent substitution matrices.

A



B

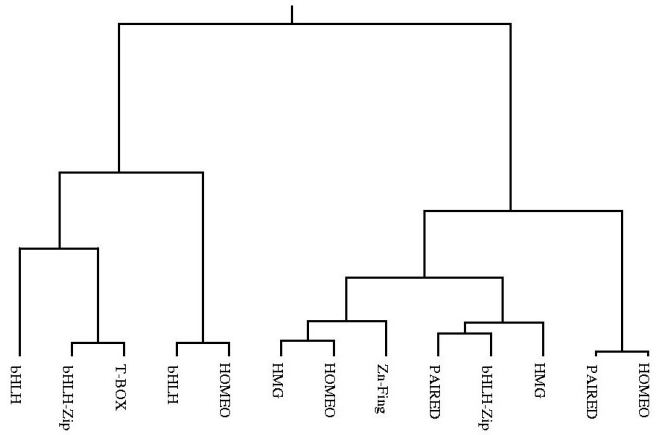


Figure 3.8. Functional classification of TFBS in Mouse; information content is calculated from nucleotide (A) neighbor-independent and (B) neighbor-dependent substitution matrices.

We note that the neighbor-dependent and independent results are different in details but are very similar in broad appearance. In case of

mouse the two results are practically identical particularly at early times (vertical axis). However, we understand that there may be effects due to smaller size of the sample (13 vs 41 in case of human). The information about the TFs that are involved in a specific gene regulation of human (Table 3.7) and mouse (Table 3.8) were collected from the Transcription Regulatory Regions Database (TRRD). The results of hierarchical clustering of TFBS of specific TFs were compared with the TFs that are involved in specific gene regulation in TRRD. These results showed that the computational results are comparable with the experimental results.

Table.3.7. Human gene with number of TFs involved for their regulation and class/family name of TFs (Data is extracted from TRRD database (Kolchanov, ***et al.,*** 2002)).

Gene name	Name of TF	Class/family of TF
AGT	HNF-4	Zn-Finger,C2H2
	COUP-TF	NUCLEAR RECEPTOR
	DBP	bHLH-ZIP
	c/EBPdelta	bZIP
	USF1	bHLH-ZIP
CRYGF	RAR/RXR	NUCLEAR RECEPTOR
	Pax-6	PAIRED
	Prox-1	HOMEO
	Sox-1	HMG
	L-Maf	bZIP
MDR1	HSF1	HSF family
	NF-IL6	bZIP
	NF-R1	bZIP
	NF-kB	bZIP
	NF-Y	CBF family

	YB-1	cold-shock domain factors
	SP1	Zn-Finger,C2H2
	WT1	Zn-Finger,C2H2
COX5B	GABP	ETS
	SP1	Zn-Finger,C2H2
	YY-1	Zn-Finger,C2H2
	USF2	bHLH
CDC25C	SP1	Zn-Finger,C2H2
	NF-Y	CBF family
	p53	P53
	CDF-1	CDF family
	YY-1	Zn-Finger,C2H2

Note: AGT -angiotensinogen; MDR1-multidrug resistance gene; CRYGF – Gamma Crystalline F protein; CDC25C-Cell Division Cycle 25C; OX5B-Cytochrome c oxidase subunit Vb

Table 3.8. Mouse gene with number of TFs involved for their regulation and class/family name of TFs (Data is extracted from TRRD database).

Gene name	Name of TF	Class/family of TF
CRYGF	RAR/RXR	NUCLEAR RECEPTOR
	Pax-6	PAIRED
	Sox1	HMG
	Prox-1	HOMEO
	Six-3	HOMEO
CACNA1S	Sox-5	HMG
	GATA-2	Zn-Finger,GATA
	CREB	bZIP
HOXA7	Antp	HOMEO
	Ftz	HOMEO
	Cad	HOMEO

Note: CRYGF- Gamma F Crystallin gene; HOXA7-Homeobox A7; CACNA1S - Calcium channel, voltage-dependent, L type, alpha 1S

3.3.4 Discussion

The information content (relative entropy) of TFBS is used to cluster the TF classes required to regulate a specific gene expression. When we look at two TFBS, we have information in times of distance between any two TFBS, a measure that we can interpret as the distance between the TFBS**.** If a small distance separates two TFBS then they may have a common TF binding site. In this study, we noted that the diverse proteins are placed closely in the classification. This is not surprising as the TFBS of a particular class of proteins are likely to be very similar. This suggests that these groups of proteins may be needed together and they may share the same transcription factors. Thus, in case of humans out of the 41 TFBS perhaps 5-10 or so transcription factors may be actually needed (instead of 41 different transcription factors). In case of mouse TFBS, out of 13 transcription factors, about 5 TFs needed for regulation. The JASPAR database TFBS are used in this study. The experimental data of TFs of specific gene expression from TRRD database is also coinciding with our computational results. This gives us a new way to look at the protein classification- not based on their structure or function of TFs - but by the nature of their transcription factor binding sites.

Conclusions

4. Conclusions

The present study involves three parts. First, we have studied the core promoter region in five sets (*E.coli*, Plants, Drosophila, Human and Mouse) of promoter sequences by calculating the average mutual information content. Here we have studied substitution matrices (both neighbor independent and neighbor dependent) for the core promoter region and calculated the information content from these substitution matrices to study the TSS-region, TATA-box, and downstream region. Neighbor independent substitutions will give 4×4 matrix and lack any preferences in other words, adjacent bases are considered independent. Next we looked at the formulae for the base pairs taken together, which correspond to a nearest neighbor preference. As we are considering a pair, there will be 16×16 matrix. These matrices include adjacent pair preferences explicitly. The results show that the TSS-region is likely to be 5-10 bases in size. We also noticed that both in the case of mouse and humans, both TATA-box and TSS-region are likely to play important roles in transcription initiation. However, in case of plant, the results showed the importance of TSS-region for transcriptional initiation compared to the TATA-box. Second, we also analyze the mitochondrial genome sequences for the transcription start sites with the information content. In this study, we concluded that the presence of short-range correlations within the TSS region is species dependent and is not universal. The information content of the mitochondrial genome near the control element computed as a 5-nucleotide overlapping block. We further noticed that there are other variable regions in the mitochondrial control element apart from the TSS. We also observed that effective comparisons can only be made on the blocks and single nucleotide comparisons does not give us any detectable signals. These results imply

a similar regulatory structure in almost all organisms and have been conserved during evolution due to functional constraints. As we already know that there are well-established tools to locate conserved regions in DNA but looking for variability is also important. So we have found that information content may be useful to study the variable functional regions in genome in an efficient manner. Third, we present a new way of clustering to classify TFBS. The clustering of TFBS (JASPAR database) with information content suggests that any one of TF can bind to the any one of the corresponding clustered TFBS-class. Thus in JASPAR database, out of the 41 TFBS (in humans), perhaps only 5 -10 TFs may be actually needed and in case of mouse instead of 13 TFs, there are approximately 5 TFs are needed for gene regulation. The experimental data of TFs of specific gene expression from Transcription Regulatory Regions Database (TRRD) also coincides with our computational results. This gives us a new way to look at the protein classification-not based on their structure or function of TFs but by the nature of their TFBS.

Future work directions

By using this preliminary analysis work on gene regulation at the transcription level, we can also analyze the promoters of different combinations having TATA, TATA-less, *Inr* or DPE. This will give better discrimination about different gene expression levels. By using the information in substitution matrices of the promoter we can predict the novel regulatory regions in the whole genome. Finally we can look for the whole gene regulatory network. We can also study the post-transcriptional regulation by using this approach.

5. List of publications

1. **Reddy, D.A.,** Prasad, B.V.L.S., Mitra, C.K., 2006. Comparative analysis of core promoter region: Information content from mono and dinucleotide substitution matrices. *Computational Biology and Chemistry*, **30**, 58-62.
2. **Reddy, D.A.,** Prasad, B.V.L.S., Mitra, C.K., 2006. Functional classification of transcription factor binding sites: Information content as a metric. *Journal of Integrative Bioinformatics*, **3**(1), 0020.
3. **Reddy, D.A.,** Mitra, C.K., 2006. Comparative analysis of transcription start site using mutual information. *Genomics, Proteomics & Bioinformatics* (*In Press*)

Conferences / Poster presentation

1. Wavelet and multifractal analysis 2004 summer school, Institut d'Études Scientifiques de Cargèse, Corsica, France (19-31, July 2004).
2. International workshop on systems biology 2006, Hamilton Institute, National University of Ireland Maynooth, Ireland (17-19, July 2006).

References

6. References

Aerts, S., Thijs, G., Dabrowski, M., Moreau, Y., Moor, B.D., 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. *BMC Genomics* **5,** 34.

Ahmad, S., Sarai, A., 2005. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* **6**, 33.

Altschul, S.F., 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219**, 555-565.

Altschul, S.F., 1993. A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* **36**, 290-300.

Arndt, P., Hwa, T., 2005. Identification and measurement of neighbor-dependent nucleotide substitution process. *Bioinformatics* **21**, 2322-2328.

Bajic, V.B., Seah, S.H., Chong, A., Krishnan, S.P.T., Koh, J.L.T., Brusic, V., 2002. Computer model for recognition of functional transcription start sites in RNA Polymerase II promoters of vertebrates. *J. Mol. Graph.* **21**, 323-332.

Bajic, V.B., Choudhary, V., Hock, C.K., 2003. Content analysis of the core promoter region of human genes. *In Silico Biol.* **4**, 0011.

Bentley, D.R., Deloukas, P., Dunham, A., *et al.,* 2001. The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and x. *Nature* **409**, 942 – 943.

Bird, A.P., 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209-213.

Boore, J.L., 1999. Animal mitochondrial genomes. *Nucleic Acids Res.* **27**, 1767-1780.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., Cedar, H., 1994. Sp1 elements protect a CpG island from de novo methylation. *Nature* **371**, 435-438.

Bucher, P., 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563-578.

Bulyk, M.L., Johnson, P.L.F.,Church, G.M., 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucl. Acids Res.* **30**, 1255-1261.

Burley, S.K., Roeder, R.G., 1996. Biochemistry and Structural biology of transcription factor IID (TFIID). *Annu. Rev. Biochem.* **65**, 769-799.

Bushnell, D.A., Bamdad, C., Kornberg, R.D., 1996. A minimal set of RNA Pol II transcription protein interactions. *J. Biol. Chem.* **271**, 20170-20174.

Chang, C.H., Hsieh, L., Chen, T., Chen, H., Luo, L., Lee, H., 2005. Shannon information in complete genomes. *Journal of Bioinformatics and Computational Biology* **3**, 587-608.

Chargaff, E.R., Lipshitz, R., Green, C., Hodes, M.E., 1951. The composition of the deoxyribonucleic acid of salmon sperm. *J. Biol. Chem.* **192**, 223-230.

Cornish-Bowden., 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations. *Nucleic Acids Res.* **13**, 3021-3030.

Cover, T.A., Thomas, J.A., 1991. Elements of Information Theory. *John Wiley and Sons*, New York.

Cristina, H., Leitao, G., Pessoa, L.S., Stolfi, J., 2005. Mutual information content of homologous DNA sequences. *Genet. Mol. Res.* **4**, 553-562.

Dawy, Z., Goebel, B., Hagenauer, J., Andreoli, C., Meitinger, T., Mueller, J.C., 2006. Gene mapping and marker clustering using

shannon's mutual information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3**, 47-56.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. In Atlas of protein sequence and structure (*Eds*) Dayhoff, M.O., *Natl. Biomed. Res. Found,* Washington, DC, **5**, 345-352.

Dickerson, R.E., Drew, H.R., Conner, B.N., Wing, R.M., Fratini, A.V., Kopka, M.L., 1982. The Anatomy of A-, B-, and Z-DNA. *Science* **216**, 475-485.

Down, T.A., Hubbard, T.J.P., 2002. Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. *Genome Res.* **12**, 458-461.

Felsenstein, J., 2004. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. *Department of Genome Sciences, University of Washington, Seattle.*

Gershenzon, N.I., Stormo, G.D., Ioshikhes, I.P., 2005. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucl. Acids Res.* **33**, 2290-2301.

Gonnet, G.H., Cohen, M.A., Benner, S.A., 1992. Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443-1445.

Hannenhalli, S., Levy, S., 2002. Predicting transcription factor synergism. *Nucleic. Acids. Res.* **30**, 4278-4284.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. *Proc.Natl.Acad.Sci.USA.* **89**, 10915-10919.

Hershberg, R., Bejerano, G., Santos-Zavaleta, A., Margalit, H., 2001. PromEC: An updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.* **29**, 277.

Jacob, F., Monod, J., 1961. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol.Biol.* **3**, 318-356.

Karlin, S., Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA.* **87**, 2264-2268.

Kolchanov, N.A., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Stepanenko, I.L., Merkulova, T.I., Pozdnyakov, M.A., Podkolodny, N.L., Naumochkin, A.N. and Romashchenko, A.G. 2002. Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res.* **30**, 312-317.

Lunter, G., Hein, J., 2004. A nucleotide substitution model with nearest-neighbor interactions. *Bioinformatics* **20**, i216-i223.

Majewski, J., Ott, J., 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12**, 1827-1836.

Martin, L.C., Gloor, G.B., Dunn, S.D., Wahl, L.M., 2005. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **21**, 4116-4124.

Montoya, J., Christianson, T., Levens, D., Rabinowitz, M., Attardi, G., 1982. Identification of initiation sites for heavy strand and light strand transcription in human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA.* **79**, 7195-7199.

Narlikar, L., Hartemink, A., 2006. Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics* **22**, 157-163.

Nicholas Jr, H.B., Deerfield II, D.W., Ropelewski, A.J., 2000. Overviw: Strategies for searching sequence databases. *BioTechniques* **28**, 1174-1191.

Ohler, U., Liao, G.C., Niemann, H., Bubin, G.M., 2002. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol*. **3**, 0087.1-0087.12.

Ojala, D., Montoya, J., Attardi, G., 1981. tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**, 470-474.

Panchenko, A.R., Bryant, S.H., 2002. A comparison of position-specific score matrices based on sequence and structure alignments. *Protein Science* **11**, 361-370.

Patikoglou, G.A., Kim, J.L., Sun, L., Yang, S.-H., Kodadek, T., Burley, S.K., 1999. TATA element recognition by the TATA box binding protein has been conserved throughout the evolution. *Genes Dev.* **13**, 3217-3230.

Périer, C.R., Junier, T., Bucher, P., 1998. The Eukaryotic Promoter Database EPD. *Nucleic Acids Res.* **26**, 353-357.

Reddy, D.A., Prasad, B.V.L.S., Mitra, C.K., 2006a. Comparative analysis of core promoter region: Information content from mono and dinucleotide substitution matrices. *Computational Biology and Chemistry* **30**, 58-62.

Reddy, D.A., Prasad, B.V.L.S., Mitra, C.K., 2006b. Functional classification of transcription factor binding sites: Information content as a metric. *Journal of Integrative Bioinformatics* **3** (1), 0020.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W., Lenhard, B., 2004. JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**(1) Database Issue.

Schwartz, R.M., Dayhoff, M.O., 1978. in Atlas of protein sequence and structure; (*eds*) Dayhoff, M.O., *Natl. Biomed. Res. Found*, Washington, DC, **5**, 353-358.

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M., Solovyev, V.V., 2003. PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.* **31**, 114-117.

Shannon, C. E., 1948. A Mathematical Theory of Communication. *Bell System Technical Journal,* **27**, 379-423, 623-656.

Smale, S.T., Kadonaga, J.T., 2003. The RNA polymerase II core promoter. *Ann. Rev. Biochem.* **72**, 449-479.

Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. *University of Kanas Scientific Bulletin* **28**, 1409-1438.

Staden, R., 1988. Methods to define and locate patterns of motifs in sequences. *Comput. Appl. Biosci.* **4**, 53-60.

Stepanova, M., Tiazhelova, T., Skoblov, M., Baranova, A., 2005. A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas. *Bioinformatics* **21**, 1789 -1796.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., Selbig, J., 2002, The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18**, S231-240.

Stormo, G.D., 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23.

Taanman, J.W., 1999. The mitochondrial genome: Structure, transcription, translation and replication. *Biochim. Biophys. Acta* **1410**, 103-123.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.,* 2001. The sequence of the human genome. *S*cience **291**, 1304-1351.

Vazquez, M.E., Caamano, A.M., Mascarenas, J.L., 2003. From transcription factors to designed sequence-specific DNA-binding peptides. *Chem. Soc. Rev.* **32**, 338-49.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P *et al*., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Watson, J.D., Crick, F.H.C., 1953. Molecular structure of nucleic acids. *Nature* 171, 737-738.

Yu, Y.-K., Altschul, S.F., 2005. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics* 21, 902-911.

Yu, Y.-K., Wootton, J.C., Altschul, S.F., 2003. The compositional adjustment of amino acid substitution matrices. *Proc. Natl. Acad. Sci. USA.* **100**, 15688-15693.

Zhang, M., Gish, W., 2006. Improved spliced alignment from an information theoretic approach. *Bioinformatics* **22**, 13-20.

Zhou, X., Wang, X., Dougherty, E.R., 2003. Construction of genomic networks using mutual-information clustering and reversible-jump Markov-chain-Monte-Carlo predictor design, *Signal Processing* **83**, 745-761.