Chapter 1

Introduction
1 Introduction

1.1 Comparative genomics

Comparative genomics is an exciting new field of biological research, which involves the analysis and comparison of genomes from different species. The purpose of comparative genomics is to gain a better understanding of how species have evolved and to determine the function of genes and noncoding regions of the genome. Researchers have learned a great deal about the function of human genes by examining their counterparts in simpler model organisms such as the mouse. Genome researchers study different features when comparing the genomes such as, sequence similarity, gene location, the length and number of coding regions (called exons) within genes, the amount of noncoding DNA in each genome, and highly conserved regions maintained in organisms as simple as bacteria and as complex as humans.

Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms to infer how selection has acted upon these elements. Those elements that are responsible for similarities between different species should be conserved through time (stabilizing selection), while those elements responsible for differences among species should be divergent (positive selection). Finally, those elements that are unimportant to the evolutionary success of the organism will be unconserved (selection is neutral). Comparative genomics studies of small model organisms are of great importance to advance our understanding of general mechanisms of evolution.

Having come a long way from its initial use of finding functional proteins, comparative genomics is now concentrating on finding regulatory regions and siRNA molecules. Recently, it has been discovered that distantly related species often share long conserved stretches of DNA that do not appear to code for any protein. It is unknown at this time what function such ultra-conserved regions serve.
Computational approaches to genome comparison have recently become a common research topic in computer science. The development of computer-assisted mathematics (using products such as Mathematica or Matlab) has helped engineers, mathematicians and computer scientists to start operating in this domain. A public collection of case studies and demonstrations is growing, ranging from whole genome comparisons to gene expression analysis (Cristianini and Hahn, 2006). This has increased the introduction of different ideas, including concepts from information theory, strings analysis and data mining. It is anticipated that computational approaches will become and remain a standard topic for research and teaching.

1.2 Genome

The hereditary nature of all living organisms is defined by its genome, which contains the complete set of information required to construct any organism. Physically the genome may be divided into a number of different nucleic acid molecules and functionally into genes. Each gene is a sequence within the nucleic acid that represents a single protein. A gene may exist in alternative forms called alleles (different forms of the gene). Chromosomes are organized structures of DNA and proteins that are found in cells. Each chromosome consists of a linear array of genes, with each gene residing in a particular location on the chromosome called the genetic locus. All the alleles of the gene are located in this locus. In diploid organisms, two sets of chromosomes are present, with one copy being inherited from each parent of the organism. A genome consists of an entire set of chromosomes for any particular organism and therefore comprises a series of DNA molecules (one for each chromosome), each of which contains many genes (Cavalier-Smith, 1985).

1.3 DNA

The vast majority of living organisms encode their genes in long strands of DNA. DNA consists of a chain made from four types of nucleotide subunits: adenine, cytosine, guanine, and thymine. Each nucleotide subunit consists of
three components: a phosphate group, a deoxyribose sugar ring, and a nucleobase. The most common form of DNA in a cell is in a double helix structure, in which two individual DNA strands twist around each other in a right-handed spiral. In this structure, the base pairing rules specify that guanine pairs with cytosine and adenine pairs with thymine (each pair contains one purine and one pyrimidine).

The base pairing between guanine and cytosine forms three hydrogen bonds, while the base pairing between adenine and thymine forms two hydrogen bonds. The two strands in a double helix must therefore be complementary, that is, their bases must align such that the adenines of one strand are paired with the thymines of the other strand, and so on.

Due to the chemical composition of the pentose residues of the bases, DNA strands have directionality. One end of a DNA polymer contains an exposed hydroxyl group on the deoxyribose, this is known as the 3’ end of the molecule. The other end contains an exposed phosphate group, this is the 5’ end. All nucleic acid synthesis in a cell occurs in the 5’-3’ direction, because new monomers are added via a dehydration reaction that uses the exposed 3’ hydroxyl as a nucleophile.

The expression of genes encoded in DNA begins by transcribing the gene into RNA, a second type of nucleic acid that is very similar to DNA, but whose monomers contain the sugar ribose rather than deoxyribose. RNA also contains the base uracil in place of thymine. RNA molecules are less stable than DNA and are typically single-stranded. Genes that encode proteins are composed of a series of three-nucleotide sequences called codons, which serve as the “words” in the genetic “language”. The genetic code specifies the correspondence during protein translation between codons and amino acids. The genetic code is nearly the same for all known organisms.

### 1.4 Functional structure of a gene

All genes have regulatory regions in addition to regions that explicitly code for a protein or RNA product. A regulatory region shared by almost all genes is
known as the promoter, which provides a position that is recognized by the transcription machinery when a gene is about to be transcribed and expressed. Although promoter regions have a consensus sequence that is the most common sequence at this position, some genes have “strong” promoters that bind the transcription machinery well, and others have “weak” promoters that bind poorly.

These weak promoters usually permit a lower rate of transcription than the strong promoters, because the transcription machinery binds to them and initiates transcription less frequently. Other possible regulatory regions include enhancers, which can compensate for a weak promoter. Most regulatory regions are “upstream” — that is, before or toward the 5’ end of the transcription initiation site. Eukaryotic promoter regions are much more complex and difficult to identify than prokaryotic promoters (Figure 1.1).

Figure 1.1. Diagram of the “typical” eukaryotic protein-coding gene. Promoters and enhancers determine what portions of the DNA will be transcribed into the precursor mRNA (pre-mRNA). The pre-mRNA is then spliced into messenger RNA (mRNA), which is later translated into protein.

In cells, genes consist of a long strand of DNA that contains coding (exons) and non-coding sequence(introns) (Pearson, 2006). Coding sequence determines what the gene produces, while non-coding sequence can regulate
the conditions of gene expression. When a gene is active, the coding and non-
coding sequence is copied in a process called transcription, producing an RNA copy of the gene’s information. This RNA can then direct the synthesis of proteins via the genetic code (Figure 1.1).

1.4.1 Gene expression

In all organisms, there are two major steps separating a protein-coding gene from its protein: first, the DNA on which the gene resides must be *transcribed* from DNA to messenger RNA (mRNA), and second, it must be *translated* from mRNA to protein. The process of producing a biologically functional molecule of either RNA or protein is called gene expression, and the resulting molecule itself is called a gene product.

1.4.2 Transcription

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

- **Messenger RNAs (mRNAs):** This class of RNAs are the genetic coding templates used by the translational machinery to determine the order of amino acids incorporated into an elongating polypeptide in the process of translation.
- **Transfer RNAs (tRNAs):** This class of small RNAs form covalent attachments to individual amino acids and recognize the encoded sequences of the mRNAs to allow correct insertion of amino acids into the elongating polypeptide chain.
- **Ribosomal RNAs (rRNAs):** This class of RNAs are assembled, together with numerous ribosomal proteins, to form the ribosomes. Ribosomes engage the mRNAs and form a catalytic domain into which the tRNAs enter with their attached amino acids. The proteins of the ribosomes catalyze all of the functions of polypeptide synthesis.
The DNA strand whose sequence matches that of the RNA is known as the coding strand and the strand from which the RNA was synthesized is the template strand. Transcription is performed by an enzyme called an RNA polymerase, which reads the template strand in the 3’ to 5’ direction and synthesizes the RNA from 5’ to 3’. To initiate transcription, the polymerase first recognizes and binds a promoter region of the gene.

In eukaryotic cells there are three distinct classes of RNA polymerase, RNA polymerase (pol) I, II and III. Each polymerase is responsible for the synthesis of a different class of RNA. RNA pol I is responsible for rRNA synthesis (excluding the 5S rRNA). RNA pol II synthesizes the mRNAs and some of the small nuclear RNAs (snRNAs) involved in RNA splicing. RNA pol III synthesizes the tRNAs, the 5S rRNA and some snRNAs.

Signals are present within the DNA template that acts in cis to stimulate the initiation of transcription, that include promoter and enhancer elements, which are important in the control of gene expression.

RNA polymerase is basically composed of five distinct polypeptide chains and the association of several of these generates the RNA polymerase holoenzyme. The sigma subunit is only transiently associated with the holoenzyme and is required for accurate initiation of transcription. In both prokaryotic and eukaryotic transcription the first incorporated ribonucleotide is a purine and it is incorporated as a triphosphate. Elongation involves the addition of the 5’-phosphate of ribonucleotides to the 3’-OH of the elongating RNA with the concomitant release of pyrophosphate. Nucleotide addition continues until specific termination signals are encountered. Transcriptional termination occurs by both factor-dependent and factor-independent means.

In prokaryotes, transcription occurs in the cytoplasm; for very long transcripts, translation may begin at the 5’ end of the RNA while the 3’ end is still being transcribed. In eukaryotes, transcription necessarily occurs in the nucleus, where the cell’s DNA is sequestered; the RNA molecule produced by the polymerase is known as the primary transcript and must undergo post-transcriptional modifications before being exported to the cytoplasm for translation. The genes of eukaryotic organisms contain noncoding regions
(introns) that are removed from the precursor-mRNA in a process called splicing (Pennisi, 2007). The regions encoding the gene products are called exons (Mount, 2004). The splicing of introns present within the transcribed region is a modification unique to eukaryotes and is also a major form of regulation in them (Woodson, 1998). In eukaryotes, a single gene can encode multiple proteins, which are produced through the creation of different arrangements of exons through alternative splicing (Gerstein, 2007).

1.4.3 Post-transcriptional modifications

Post-transcriptional modifications are various RNA modifications by which, the precursor messenger RNA (pre-mRNA) is converted into mature messenger RNA (mRNA) that occur prior to protein synthesis. These processes are vital for the correct translation of the eukaryotic genes, after they have been transcribed. The pre-mRNA molecule undergoes three main modifications, which include,

(i) 5’ capping,
(ii) 3’ cleavage and polyadenylation and
(iii) RNA splicing.

(i) 5’ Capping

Capping of the pre-mRNA involves the addition of 7-methylguanosine (m7G) to the 5’ end of the pre-mRNA. This requires the terminal 5’ phosphate to be removed, which is done with the aid of a phosphohydrolase enzyme. The enzyme guanosyl transferase then catalyses the capping reaction, which produces the diphosphate 5’ end. The diphosphate 5’ end then attacks the α phosphorus atom of a GTP molecule in order to add the guanine residue in a 5’5’ triphosphate link. The enzyme S-adenosyl methionine then methylates the guanine ring at the N-7 position (Figure 1.2).
The ribose of the adjacent nucleotide may also be methylated and this process is continued to form a 5’ cap to the pre-mRNA. In these cases the methyl groups are added to the 2’ OH groups of the ribose sugar and the cap added protects the 5’ end of the primary RNA transcript from the attack of the ribonucleases that have specificity to the 3’-5’ phosphodiester bonds.

(ii) 3’ cleavage and polyadenylation

The pre-mRNA processing at the 3’ end of the RNA molecule involves cleavage of its 3’ end and then the addition of about 200 adenine residues to form a poly(A) tail (Figure 1.3).
The cleavage and adenylation reactions occur if a polyadenylation signal sequence (5’- AAUAAA-3’) is located near the 3’ end of the pre-mRNA molecule, and is followed by another sequence, which is usually (5’-CA-3’) (not shown in the Figure 1.3) the site of cleavage. A GU-rich sequence is also present further downstream on the pre-mRNA molecule. After the synthesis of the sequence elements, two multisubunit proteins called cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CStF) are transferred from RNA Polymerase II to the RNA molecule. The two factors then bind to the sequence elements. A protein complex forms which contains additional cleavage factors and the enzyme Polyadenylate Polymerase (PAP). This complex cleaves the RNA between the
polyadenylation sequence and the GU-rich sequence at the cleavage site marked by the (5’-CA-3’) sequences (not shown in the Figure 1.3). Poly(A) polymerase then adds about 200 adenine units to the new 3’ end of the RNA molecule using ATP as a precursor. As the poly(A) tails is synthesised, it binds multiple copies of poly(A) binding protein, which protects the 3’end from ribonuclease digestion.

(iii) RNA splicing

RNA splicing is the process by which introns (regions of RNA that do not code for protein), are removed from the pre-mRNA and the exons are connected to re-form a single continuous molecule. Although most RNA splicing occurs after the complete synthesis and end-capping of the pre-mRNA, transcripts with many exons can be spliced co-transcriptionally. The splicing reaction is catalyzed by a large protein complex called the “spliceosome” assembled from proteins and small nuclear RNA molecules that recognize splice sites in the pre-mRNA sequence.

1.4.4 Export of mRNA to cytoplasm

After it has been synthesized and processed, mRNA is exported from the nucleus to the cytoplasm in the form of a ribonucleoprotein complex (Le Hir et. al., 2001). Introns may prevent export of mRNA because they are associated with the splicing apparatus. The spliceosome also may provide the initial point of contact for the export apparatus. The complex consists of >9 proteins and is called the EJC (exon junction complex)

The first contact in assembling the EJC is made with one of the splicing factors. Then after splicing, the EJC remains attached to the mRNA just upstream of the exon-exon junction. The EJC is not associated with RNAs transcribed from genes that lack introns, so its involvement in the process is unique for spliced products. The EJC includes a group of proteins called the REF family (the best characterized member is called Aly). The REF proteins in turn interact with a transport protein (TAP and Mex), which has direct responsibility for interaction with the nuclear pore (Figure 1.4).
Figure 1.4. The process of mRNA export that involves key proteins like REF and TAP/ Mex. A REF protein binds to a splicing factor and remains with the spliced RNA product. REF binds to the transport proteins TAP/ Mex that binds to the nuclear pore.

1.4.5 Translation

Translation is the process by which a mature mRNA molecule is used as a template for synthesizing a new protein (Figure 1.5). Translation is carried out by ribosomes, large complexes of RNA and protein responsible for carrying out the chemical reactions to add new amino acids to a growing polypeptide chain by the formation of peptide bonds. The genetic code is read three nucleotides at a time, in units called codons, via interactions with specialized RNA molecules called transfer RNA (tRNA).

Each tRNA has three unpaired bases known as the anticodon that are complementary to the codon it reads; the tRNA is also covalently attached to the amino acid specified by the complementary codon. When the tRNA binds to its complementary codon in an mRNA strand, the ribosome ligates its amino acid sequence to the new polypeptide chain, which is synthesized from amino terminus to carboxyl terminus. During and after its synthesis, the new
protein must fold to its active three-dimensional structure before it can carry out its cellular function.

**Figure 1.5.** A schematic representation of the processes of replication (DNA duplication), transcription (RNA synthesis) and translation (Protein synthesis) all shown together.

### 1.5 RNA Splicing

In molecular biology, splicing is a modification of an RNA after transcription, in which introns are removed and exons are joined. This is needed for the typical eukaryotic messenger RNA before it can be used to produce a correct protein through translation. For many eukaryotic introns, splicing is done in a series of reactions which are catalyzed by the spliceosome, a complex of small nuclear ribonucleoproteins (snRNPs) (Dreyfuss, *et al.*, 1993).

Different classes of organisms contain interrupted genes that represent a minor proportion of the genes in the lowest eukaryotes and a major proportion in higher genomes. Genes might vary according to the length and number of introns in them with exons being relatively shorter in length than introns. The primary transcript obtained by transcription is interrupted by the presence of introns in it and the discrepancy between the interrupted organization of the
gene and the uninterrupted organization of its mRNA requires processing of this transcript. The primary transcript thus obtained has the same organization as the gene, and is also called the pre-mRNA or heterogeneous nuclear RNA (hnRNA).

There are several different classes of introns and the two most common are the group I and group II introns. Many of the group I and II introns are self-splicing and do not require any additional protein factors for splicing but some of them require proteins factors for splicing.

- **Group I introns** are found in nuclear, mitochondrial and chloroplast rRNA genes. These introns require an external guanosine nucleotide as a cofactor. The 3’-OH of the guanosine nucleotide acts as a nucleophile to attack the 5’-phosphate of the 5’ nucleotide of the intron. The resultant 3’-OH at the 3’ end of the 5’ exon then attacks the 5’ nucleotide of the 3’ exon releasing the intron and covalently attaching the two exons together. The 3’ end of the 5’ exon is termed the splice donor site and the 5’ end of the 3’ exon is termed the splice acceptor site.

- **Group II introns** are found in mitochondrial and chloroplast mRNA genes. They are spliced similarly to the group I introns except that instead of an external nucleophile the 2’-OH of an adenine residue within the intron is the nucleophile. This residue attacks the 3’ nucleotide of the 5’ exon forming an internal loop called a lariat structure. The 3’ end of the 5’ exon then attacks the 5’ end of the 3’ exon as in group I splicing, by releasing the intron and covalently attaching the two exons together.

- The **third class of introns** is also the largest class found in nuclear mRNAs. This class of introns undergoes a splicing reaction similar to group II introns in that an internal lariat structure is formed. However, splicing is catalyzed by a specialized RNA-protein complex called the “Spliceosomal complex”, which consists of small nuclear ribonucleoprotein particles (snRNPs).
The fourth class of introns is those found in certain tRNAs. These introns are spliced by a specific splicing endonuclease that utilizes the energy of ATP hydrolysis to catalyze the intron removal and ligation of the two exons together.

1.5.1 An overview of mRNA splicing

Figure 1.6 gives an overview of the various processes like transcription, end modifications, pre-mRNA splicing, mRNA export and translation.

Figure 1.6. Schematic representation of the processes of transcription, end modifications, pre-mRNA splicing, mRNA transport and translation are shown together in this figure.
1.5.2 Splice site (exon-intron) junctions

The splice sites also known as the exon-intron junctions have well conserved, short consensus sequences. The spliceosomal complex recognizes these boundaries during the process of splicing, such that the splicing of the introns and joining of the exons is done at the correct place. The sequences at these junctions which indicate the starting and the ending point of the intron is given as,

\[ \text{GU} \ldots \ldots \text{AG} \]

These junctions thus described are conforming to the GT-AG rule, because the intron defined in this way starts with the dinucleotide GU and ends with the dinucleotide AG. Splice sites are characterized as donor (5’ boundary containing the dinucleotide GT in parent DNA or GU in pre-mRNA) or acceptor (3’ boundary containing the dinucleotide AG) regions. The two sites have different sequences and they define the ends of the intron directionally. They are named proceeding from left to right along the intron from 5’ to 3’ direction i.e., the left (5’) and right (3’) splice sites (Figure 1.7). The subscripts indicate the percent occurrence of the specified base at each consensus position. The only feature, which is 100% conserved at intron/exon junctions is that introns begin with GU and end in AG. There are, however, nucleotides that are found more frequently at particular positions (percentages are shown in figure 1.7).

---

**Figure 1.7.** Diagram of a typical intron, which have intron-exon boundaries that have short consensus sequences in them. The ends of nuclear introns are defined by the GU-AG rule. The branch site is defined by a short consensus present about 18-40 nucleotides upstream of the 3’ splice site.
1.5.3 The branch site

During the process of splicing the consensus sequences at the 5’ and 3’ splice sites and at the branch site (Reed and Maniatis, 1985) are recognized by the spliceosomal complex. The branch site lies 18-40 nucleotides upstream of the 3’ splice site and was found to play an important role in identifying the 3’ splice site as the target for connection to the 5’ splice site. It was found to have a preference for purines/pyrimidines at each position and retains the target A nucleotide, which forms a 2’-5’ bond with the G in the consensus sequences at the 5’ splice site.

1.5.4 Role of snRNAs in splicing

The spliceosomal complex (Grabowski et al., 1985) contains both proteins and RNAs. The RNAs exist as ribonucleoprotein particles (RNPs) both in the nucleus and cytoplasm. The RNAs present in the nucleus are called small nuclear RNAs (snRNA) (Guthrie and Patterson, 1988); and those present in the cytoplasm are called small cytoplasmic RNAs (scRNA). These RNAs (snRNA and scRNA) exist as ribonucleoprotein particles (RNPs), called as small nuclear ribonucleoprotein particles (snRNP) and small cytoplasmic ribonucleoprotein particles (scRNP) depending upon their location.

The spliceosomal complex, which comprises a 50-60S-ribonucleoprotein particle, contains snRNPs and additional proteins. The snRNPs that are involved in splicing are U1, U2, U5, U4, and U6 and are named according to the snRNAs that are present in them. Each snRNP contains a single snRNA and several proteins. A common structural core for each snRNP consists of a group of 8 proteins. The U4 and U6 snRNPs are usually found as a single (U4/U6) particle. Apart from the snRNPs that are involved in splicing, the spliceosome consists of several other proteins known as the splicing factors that are required for the spliceosome assembly, binding of the spliceosome to the RNA substrate and in catalysis. In addition to these proteins, another 30 proteins that are associated with the spliceosome are found to have a role at different stages of gene expression. The major spliceosomal snRNPs (U1, U2,
U5, U4 and U6) are responsible for splicing the vast majority of pre-mRNAs (so-called U2 introns). But a group of less abundant snRNPs, (U11, U12, U4atac, and U6atac), together with the U5 snRNP, are subunits of the so-called minor spliceosome (found in metazoans like, plants, insects and vertebrates), which splices a rare class of pre-mRNA introns called U12-type introns.

The process of splicing requires different types of interactions, between the pre-mRNA and the spliceosomal complex such as, the RNA-RNA, protein-RNA and protein-protein interactions. The RNA-RNA interactions require the direct base pairing of the RNA of pre-mRNA and the RNA of the spliceosomal complex.

1.5.5 Stages of splicing

RNA splicing can be broadly divided into three different stages (Figure 1.8), such as,

(i) Assembly of the spliceosomal machinery onto pre-mRNA,
(ii) Cleavage of the donor site (5’ splice site) and lariat formation,
(iii) Cleavage of the acceptor site (3’ splice site) and exon ligation.

![Figure 1.8. The various steps involved in the processing of the transcribed pre-mRNA (hnRNA), which occurs through the formation of a “Lariat”.

---

18
(i) **Assembly of the Spliceosomal machinery**

Splicing of the precursor mRNA takes place in the spliceosomal complex, which assembles at the splice junctions. Two strategies are given for the assembly of the spliceosomal complex, depending upon the way the complex is formed onto the pre-mRNA. The strategies are given as follows,

(a) The traditional view: A stepwise-assembled complex

(b) The modern view: A preassembled complex

(a) The traditional view: A stepwise assembled complex

According to this model, the spliceosomes are highly dynamic machines, building anew on each pre-mRNA substrate in a highly ordered pathway *in vitro*. Significantly, this order of assembly is conserved from yeast to mammals and even extends to the metazoan non-conventional U12 spliceosome. The various steps involved in this pathway are given in figure 1.9.

(b) The modern view: A preassembled complex

This alternative model, which is found to be present in higher eukaryotes, supports that U2 and U4/U6 and U5 snRNPs functionally associate with the pre-mRNA at an early stage of spliceosome assembly (Will and Luhrmann, 2001). This pathway can be further divided into two, depending on the number of snRNPs (tetra/penta snRNPs) that are present in the pre-spliceosomal complex. The various steps involved in this pathway are given in figure 1.10.

**Penta snRNP complex:**

- According to this model, the U2 snRNP is functionally associated with the pre-mRNA at the time of E complex formation in mammals or commitment complex formation in yeast, together with the U1 snRNP in an ATP independent step (Das *et al.*, 2000). This suggests that the initial steps of the major spliceosome assembly may be similar to those of the minor spliceosome assembly in which U11 and U12 snRNPs,
bind simultaneously to the 5’ splice site and branch site, respectively, under the form of a pre-assembled 18S complex (Frilander and Steitz, 1999) (Figure 1.10).

**Figure 1.9.** The various steps involved in the stepwise assembly of the spliceosomal complex. (1) Binding of U1 snRNP to the 5’ splice site and BBP and Mu2p splicing factors to the branch site of the intron. (2) Formation of pre-spliceosome, which involves the dissociation of BBP and Mu2p splicing factors and binding of U2 snRNP to the branch site. (3) Binding of U4/U6 and U5 snRNPs to the 5’ splice site and dissociation of U1 snRNP. (4) Formation of the spliceosome, which involves the dissociation of U4 snRNP and interaction of U2 and U6 snRNPs. (5) The 1st trans-esterification reaction takes place, which involves the cleavage of the donor (5’) splice site and formation of the lariat. (6) The 2nd trans-esterification reaction takes place, which involves the cleavage of the acceptor (3’) splice site and exon ligation.

**Tetra snRNP complex:**

- In addition, the U4/U6 and U5 snRNP probably also functionally associates with the pre-mRNA at a much earlier stage of spliceosome assembly than previously admitted.
According to recent studies, the 5' splice site is also recognized by the U4/U6 and U5 snRNP, together with U1, at an early stage of spliceosome assembly (i.e. prior to A complex assembly in mammals or pre-spliceosome complex in yeast) in an ATP-dependent manner even in the absence of a stable U2 snRNP/branch site interaction (Figure 1.10).

Figure 1.10. This pathway can be further divided into two depending on the number (tetra/penta) of snRNPs that are present in the pre-assembled spliceosomal complex. The tetra/penta snRNP complex assembles on the pre-mRNA, which is followed by the 1st trans-esterification reaction that involves the splicing of the donor (5') splice site and formation of the lariat. The 2nd trans-esterification reaction takes place, which involves the cleavage of the acceptor (3') splice site and ligation of exons.
(ii) Donor site cleavage and lariat formation

The first step after the assembly of the spliceosomal complex is to make a cut at the 5’ splice site, separating the left exon and the right intron-exon molecule. The left exon takes the form of a linear molecule. The right intron-exon molecule forms a lariat in which the 5’ terminus generated at the end of the intron becomes linked by a 5’—2’ bond to a base ‘A’ in the consensus at the branch site present in the intron, which then leads the formation of a lariat (Figure 1.8).

(iii) Acceptor site cleavage and exon ligation

Cutting at the 3’ splice site releases the free intron in lariat form, while the right exon is ligated (spliced) to the left exon to form the mature mRNA. The cleavage and ligation reactions are shown separately in the figure 1.8 for illustrative purposes, but they actually occur as one coordinated transfer. The lariat is then “debranched” to give a linear excised intron, which is rapidly degraded (Figure 1.8).

1.5.6 Mechanism of splicing

The overall mechanism of splicing involves the formation of many complexes of snRNPs, and each spliceosomal complex has a different composition of snRNPs in them. The initial step of splicing is the binding of U1snRNP to the 5’ splice site, which takes place by the interaction of one of the proteins of U1snRNP and the protein ASF/SF2. The single stranded region at the 5’ terminus of the U1snRNA base pairs with the 5’ splice site by a stretch of 4-6 bases that are complementary with the splice site (Figure 1.11).

(i) Formation of E spliceosomal complex

The first complex formed during the process of splicing is the E (early presplicing) complex, which contains U1 snRNP, the splicing factor U2AF (U2 auxiliary factor), and the members of family-called SR proteins. SR proteins take their name from the presence of an Arg-Ser-rich region and comprise an important group of splicing factors and regulators. They connect
U2AF to U1 snRNP and are essential components of the spliceosome, forming a framework on the RNA substrate. U2AF has a large subunit (U2AF65) that binds to the pyrimidine tract downstream of the branch site, and a small subunit (U2AF35) that binds the dinucleotide AG at the 3’ splice site (Wu et al., 1999). The formation of the E complex is completed by the binding of the U1snRNP and ASF/SF2 to the 5’ splice site and U2AF to the pyrimidine tract (Figure 1.11).

**Figure 1.11.** The splicing reaction proceeds through discrete stages in which spliceosome formation involves the interaction of components that recognize the consenses sequences and the formation of various spliceosomal complexes.
(ii) **Formation of A spliceosomal complex**

The E spliceosomal complex is converted to the A complex when U2 snRNP binds to the branch site. The 5’ end of the U2 snRNA contains sequences that are complementary to the branch site, which base pairs with the branch site. This binding requires ATP hydrolysis and commits a pre-mRNA to the splicing pathway (Figure 1.11).

(iii) **Formation of B spliceosomal complex**

The A complex is converted to the B1 complex when a trimer containing the U5 and U4/U6 snRNPs binds to the A complex (Lamond, 1988). The B1 complex is basically regarded as a spliceosome, since it contains the components that are needed for the splicing reaction to take place. The B1 complex is converted into B2 complex after U1 is released. This dissociation of U1 is necessary to allow other components of the complex to come into juxtaposition with the 5’ splice site (most notably U6 snRNA). Then the U5 snRNA changes its position, so that it shifts to the vicinity of the intron sequences (Sontheimer and Steitz 1993) (Figure 1.11).

(iv) **Formation of C spliceosomal complex**

The rearrangement of the spliceosomal complex leads to the formation of the C1 spliceosomal complex (Figure 1.11). The catalytic reaction is triggered by the release of U4 snRNP, which requires ATP hydrolysis. The U2 snRNP then pairs with the U6 snRNP, which catalyzes the first transesterification reaction. A cleavage is made at the 5’ splice site, which leads to the formation of a lariat and makes the C1 complex to be converted into C2 spliceosomal complex. The U2/U5/U6 snRNPs remain bound to the lariat that catalyzes the second transesterification reaction, which also require ATP hydrolysis. A cleavage occurs at the 3’ splice site and the lariat is released. The lariat thus formed is linearized and degraded and the two linear exons are ligated to form a mature mRNA.
1.5.7 The transesterification reactions

The chemical reactions that lead to the (i) cleavage of the donor (5') splice site and formation of the 5'-2' bond (between the 5' splice site and the ‘A’ of the branch site) and the (ii) cleavage of the acceptor (3’) splice site and the ligation of the 5’-3’ ends of the two exons proceed by two transesterification reactions, in which a bond is transferred from one location to another (Figure 1.12).

**Step 1**: The first step is a nucleophilic attack by the 2’-OH of the invariant A (of the branch site) on the 5’ splice site. This leads to the cleavage to 5’ ss and the formation of the lariat.

**Step 2**: In the second step, the free 3’-OH of the exon1 that was released by the first reaction now attacks the bond at the 3’ splice site of the exon2, which leads to the cleavage of the 3’ ss. The lariat is linearized and degraded and the exon1 and exon2 are ligated. In these two reactions, the number of phosphodiester bonds is

---

**Figure 1.12**. Nuclear splicing occurs by two transesterification reactions in which an OH groups attacks a phosphodiester bond, which leads to the splicing of the 5’ and 3’ ends of the intron and joining of the exons to form of a mature mRNA.
conserved. There were originally two 5′-3′ bonds at the exon-intron splice sites; one has been replaced by the 5′-3′ bond between the exon1-exon2, and the other has been replaced by the 5′-2′ bond that forms the lariat.

1.5.8 Clinical significances of alternative and aberrant splicing

The presence of introns in eukaryotic genes would appear to be an extreme waste of cellular energy when considering the number of nucleotides incorporated into the primary transcript only to be removed later as well as the energy utilized in the synthesis of the splicing machinery. However, the presence of introns can protect the genetic makeup of an organism from genetic damage by outside influences such as chemicals or radiation. An additionally important function of introns is to allow alternative splicing to occur, thereby, increasing the genetic diversity of the genome without increasing the overall number of genes. By altering the pattern of exons, from a single primary transcript, that are spliced together different proteins can arise from the processed mRNA from a single gene. Alternative splicing can occur either at specific developmental stages or in different cell types.

Abnormalities in the splicing process can lead to various disease states. Many defects in the β-globin genes are known to exist leading to β-thalassemias. Some of these defects are caused by mutations in the sequences of the gene required for intron recognition and, therefore, result in abnormal processing of the β-globin primary transcript. Patients suffering from a number of different connective tissue diseases exhibit humoral auto-antibodies that recognize cellular RNA-protein complexes. Patients suffering from systemic lupus erythematosis have auto-antibodies that recognize the U1 RNA of the spliceosome. RNA splicing is not merely a curiosity. Approximately 15% of all genetic diseases are caused by mutations that affect RNA splicing.

So, with this knowledge of the mechanism of “pre-mRNA” splicing and owing to the importance of the splice sites in the mechanism, I carried out my research in order to study the signals (splice sites) that govern the process of
Introduction

It was also important to analyze the various aspects of the genomes of different organisms by comparative genomics, which might provide some insights in understanding their genome architecture. In this regard I have divided my research work into four sections and put forward some of the objectives for each of the work carried out, which are as follows:

I. 1/f correlations in viral genomes- A fast Fourier transformation (FFT) study

- To study the existence of long-range correlations in the complete genomes of viruses.
- To study the correlation between the distribution of the gene length and the “1/f region” of the genomes analyzed.

II. Comparative analysis of splice site regions by information theory

- To characterize signals that govern the process of splicing in different organisms by information theory.
- To study the variability of sequences at the splice sites in the given organisms.

III. Frequency analysis of the splice site regions in different organisms

- To study the frequency distribution of the sub-sequences at the donor/acceptor splice sites.
- To obtain the optimal length of the sub-sequences at the splice sites.
- To obtain sub-sequences that are involved in splicing.
- To discover the pattern of association between the donor and acceptor splice site region.

IV. Frequency studies of the unique sub-sequences at the splice sites
• To carry a comparative study of the sub-sequences at the splice sites in different organisms.

• To study the DNA motifs in the sub-sequences at both the splice sites regions in all the given organisms.

The following chapters give a detailed description (including introduction, methodology, results and discussion and conclusions) of each of the work done.
1.6 References


