Comprehensive Analysis of the Splice Site Regions by Comparative Genomics

Thesis submitted for the Degree of

DOCTOR OF PHILOSOPHY

SHASHI REKHA THUMMALA

Department of Biochemistry
School of Life Sciences
University of Hyderabad
Hyderabad
2008

Comprehensive Analysis of the Splice Site Regions by Comparative Genomics

Thesis submitted for the Degree of

DOCTOR OF PHILOSOPHY

By

SHASHI REKHA THUMMALA

Department of Biochemistry
School of Life Sciences
University of Hyderabad
Hyderabad - 500 046
India

August 2008

Enrollment No: 03LBPH19

Dedicated to my parents

UNIVERSITY OF HYDERABAD

Department of Biochemistry

School of Life Sciences

Certificate

This is to certify that this thesis entitled **"COMPREHENSIVE ANALYSIS OF THE SPLICE SITE REGIONS BY COMPARATIVE GENOMICS"** submitted to the University of Hyderabad, Hyderabad by **Ms. Shashi Rekha Thummala** for the degree of Doctor of Philosophy, is based on the studies carried out by her under my supervision. I declare to the best of my knowledge that this work has not been submitted earlier for the award of degree or diploma from any other University or Institution.

Chanchal K Mitra	**Head**
Supervisor	**Dept. of Biochemistry**

Dean

School of Life Sciences

UNIVERSITY OF HYDERABAD

Department of Biochemistry

School of Life Sciences

Declaration

I hereby declare that the work presented in my thesis entitled **"COMPREHENSIVE ANALYSIS OF THE SPLICE SITE REGIONS BY COMPARATIVE GENOMICS" is entirely original and** has been carried out by me in the Department of Biochemistry, University of Hyderabad, Hyderabad, under the supervision of **Prof. Chanchal K Mitra**. I further declare that this work has not been submitted earlier for the award of degree or diploma from any other University or Institution.

SHASHI REKHA THUMMALA
(Enrollment No. 03LBPH19)

Date:
Place: Hyderabad

Prof. Chanchal K Mitra
(Supervisor)

Acknowledgements

- *I express my sincere gratitude to my teacher and guide Prof. Chanchal K Mitra for his disciplined supervision, motivated guidance, valuable suggestions and constant encouragement throughout my research. His wisdom, knowledge and commitment to the highest standards inspired and motivated me in my work.*
- *I thank the Dean, School of Life Sciences Prof. A. S. Raghavendra and the former Dean Prof. T. Suryanarayana for providing the required facilities for the completion of my research work.*
- *I thank the Head, Department of Biochemistry Prof. M.Ramanadham for extending the facilities required for the completion of my work.*
- *I am thankful to Prof. Aparna Datta Gupta for her encouragement, enthusiasm and zeal, which had been the driving forces for my research.*
- *I am thankful to all the faculty of the University of Hyderabad for instilling in me the scientific spirit and the qualities of being a good scientist in future.*
- *I am grateful to all my teachers who had taught me in school, college and university for their support and encouragement throughout my career.*
- *I thank all the non-teaching staff of the University of Hyderabad for their assistance and help during the course of my work.*
- *I also thank the staff of Centre for Simulation, Modeling and Design (CMSD) for their assistance and support during my work at CMSD.*
- *I thank UNICEF/ UNDP/ World Bank/ WHO for their financial support in attending Workshop on “3rd Regional Training Course*

on Bioinformatics applied to tropical diseases in Southeast Asia", at Mahidol University, Bangkok, Thailand, during 26 July – 6 August 2004.

- *I thank Centre for Development of Advanced Computing (CDAC), Pune for giving me the opportunity to participate in the "Workshop on practical aspects of Parallel Computing on IBM P690 Cluster", held at Centre for Modeling, Simulation and Design (CMSD), University of Hyderabad, Hyderabad, during 21-24 September 2004.*
- *I thank ICGEB and UPE for their financial support in attending the ICGEB workshop on "Chromosome Duplication and Segregation", held at Trieste, Italy, during 2-5 May 2005.*
- *I thank CSIR for granting me fellowship and also supporting me financially for attending many workshops and conferences at Hyderabad and New Delhi.*
- *I thank all my present and former lab mates, friends and well wishers for their support.*
- *I am thankful to my friend Gautham for being very helpful, understanding and supportive in situations of distress.*
- *I am very grateful to my parents and brothers for giving me all their love, support and encouragement throughout my life in having a good carrier.*
- *I am always thankful and grateful to GOD for giving me the courage to overcome the difficulties in my life in order to prove myself.*

Shashi Rekha Thummala

Table of contents

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).

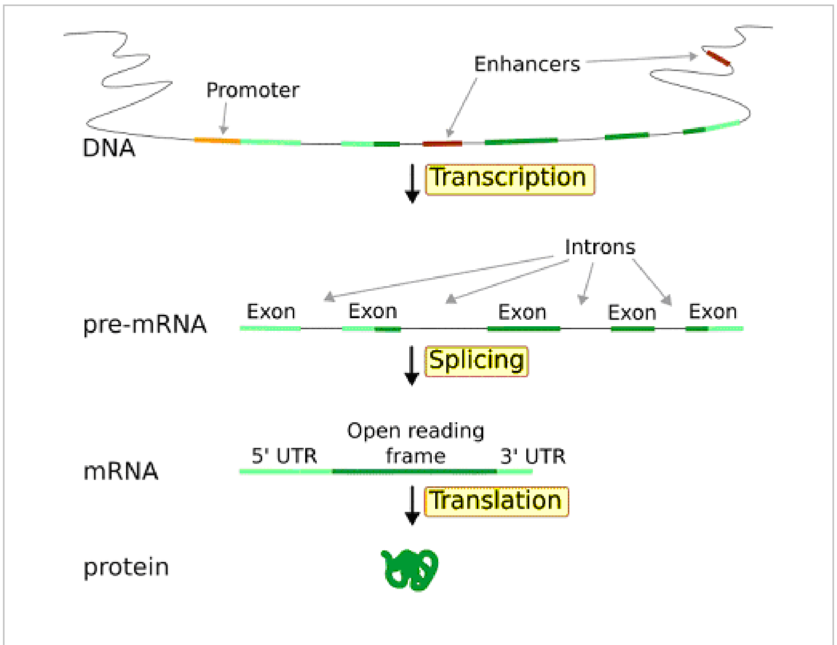


Figure 1.1. Diagram of the "typical" eukaryotic protein-coding gene. Promoters and enhancers determine what portions of the DNA will be transcribed into the precursor mRNA (pre-mRNA). The pre-mRNA is then spliced into messenger RNA (mRNA), which is later translated into protein.

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 5*S* 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 5*S* 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 nonccoding 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 (m^7G) 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).

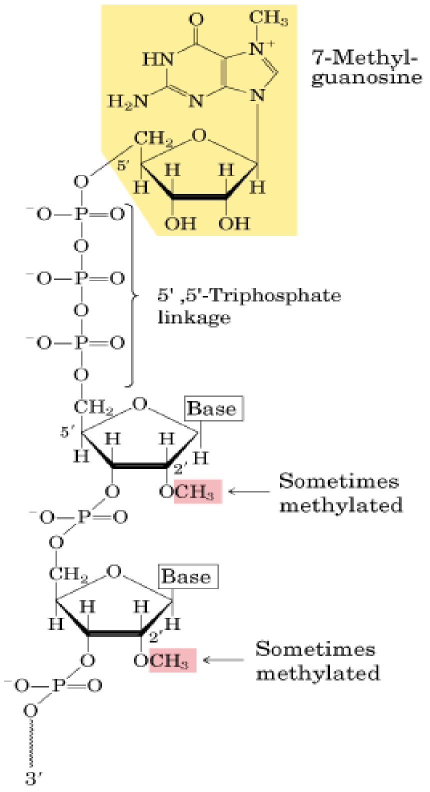


Figure 1.2. The process of 5’ capping of the pre-mRNA, which involves the addition of 7-methylguanosine to the 5’ end. This reaction is catalyzed by the enzyme guanosyl transferase, which adds guanine residue to the 5’ end of the pre-mRNA in a 5’5’ triphosphate link. The enzyme S-adenosyl methionine methylates the guanine at the N-7 position.

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).

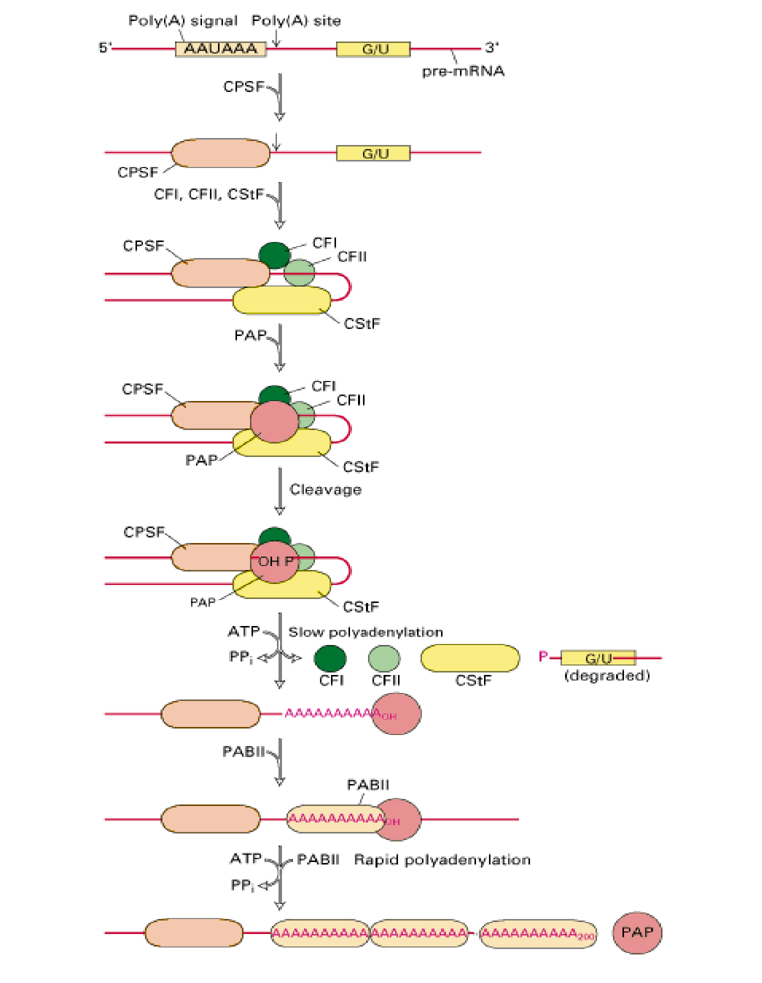


Figure 1.3. The 3' processing of the pre-mRNA that consists of several activities as shown in this figure. The 3' end is cleaved and about 200 adenine residues are added to the poly(A) tail, which involves various cleavage and polyadenylation factors.

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 protiens 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).

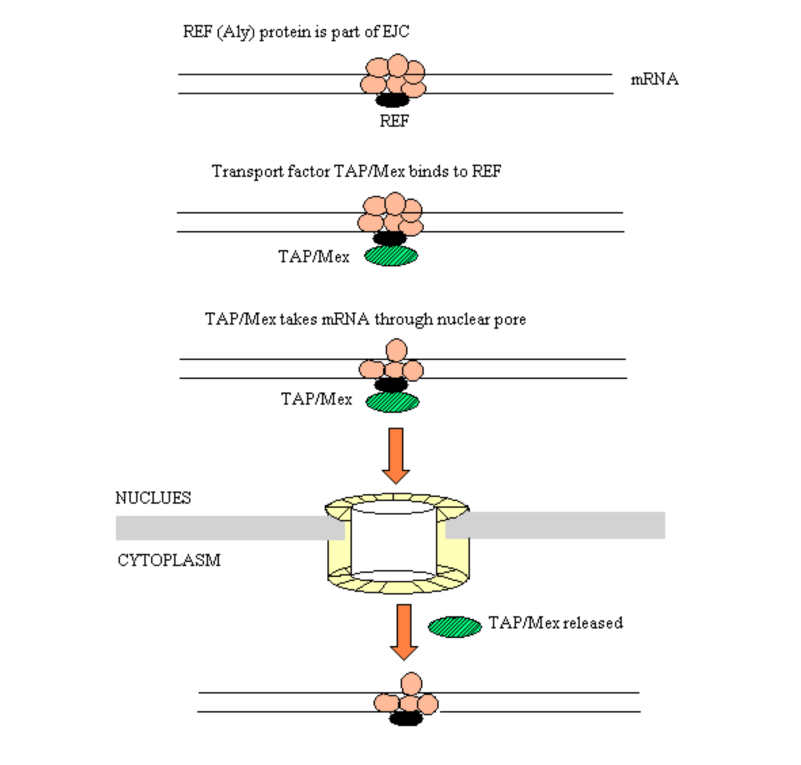

REF (Aly) protein is part of EJC
mRNA
REF
Transport factor TAP/Mex binds to REF
TAP/Mex
TAP/Mex takes mRNA through nuclear pore
TAP/Mex
NUCLUES
CYTOPLASM
TAP/Mex released

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.

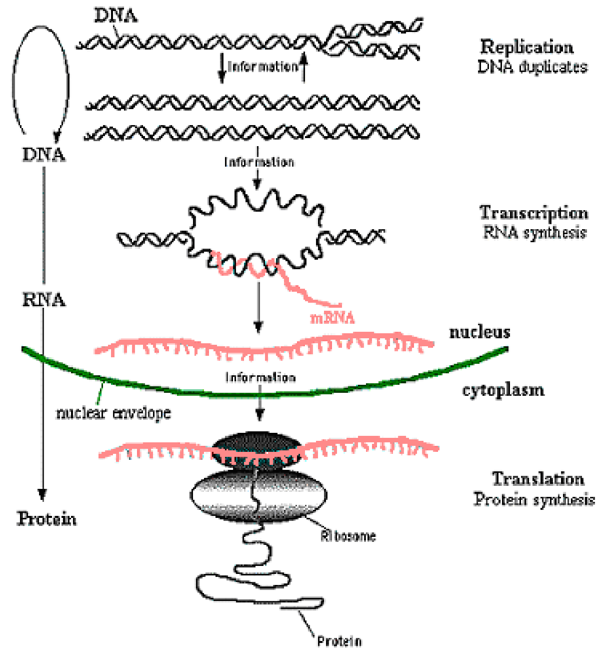


Figure 1.5. A schematic representation of the processes of replication (DNA duplication), transcription (RNA synthesis) and translation (Protein synthesis) all shown together.

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.

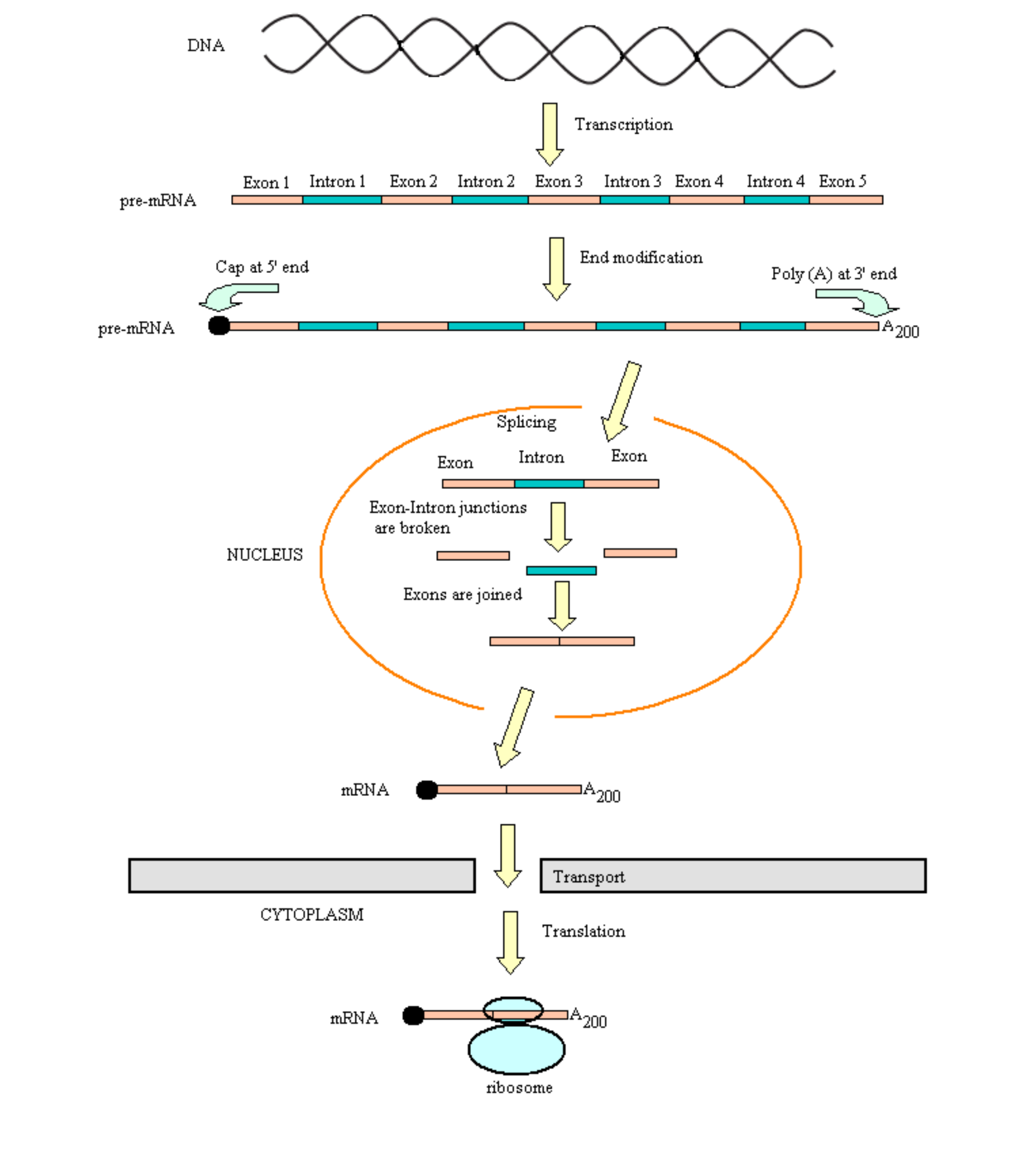


Figure1.6. Schematic representation of the processes of transcription, end modifications, pre-mRNA splicing, mRNA transport and translation are shown together in this figure.

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,

GU…….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).

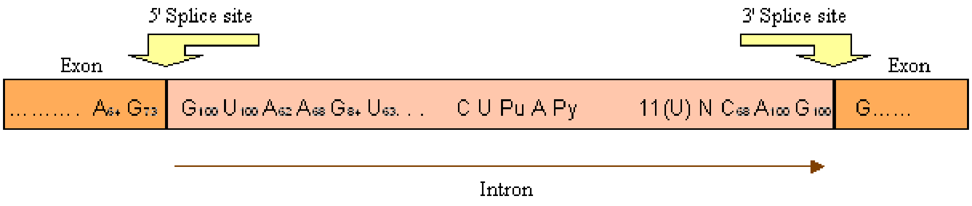

5' Splice site
3' Splice site
Exon
Exon
……… A_{64} G_{73}
G_{100} U_{100} A_{62} A_{68} G_{84} U_{63}. . .
C U Pu A Py
11 (U) N C_{68} A_{100} G_{100}
G……
Intron

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.

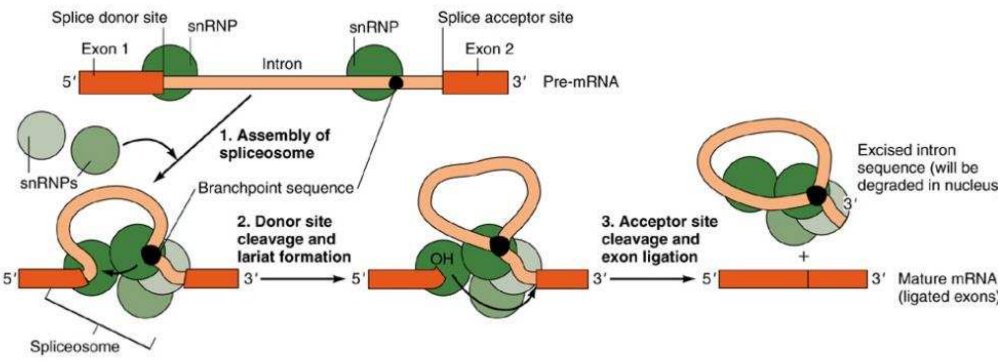

Splice donor site
snRNP
snRNP
Splice acceptor site
Exon 1
Intron
Exon 2
5'
3'
Pre-mRNA
1. Assembly of spliceosome
snRNPs
Branchpoint sequence
2. Donor site cleavage and lariat formation
5'
3'
5'
OH
3. Acceptor site cleavage and exon ligation
3'
5'
Excised intron sequence (will be degraded in nucleus
3'
+
3'
Mature mRNA (ligated exons)
Spliceosome

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

(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).

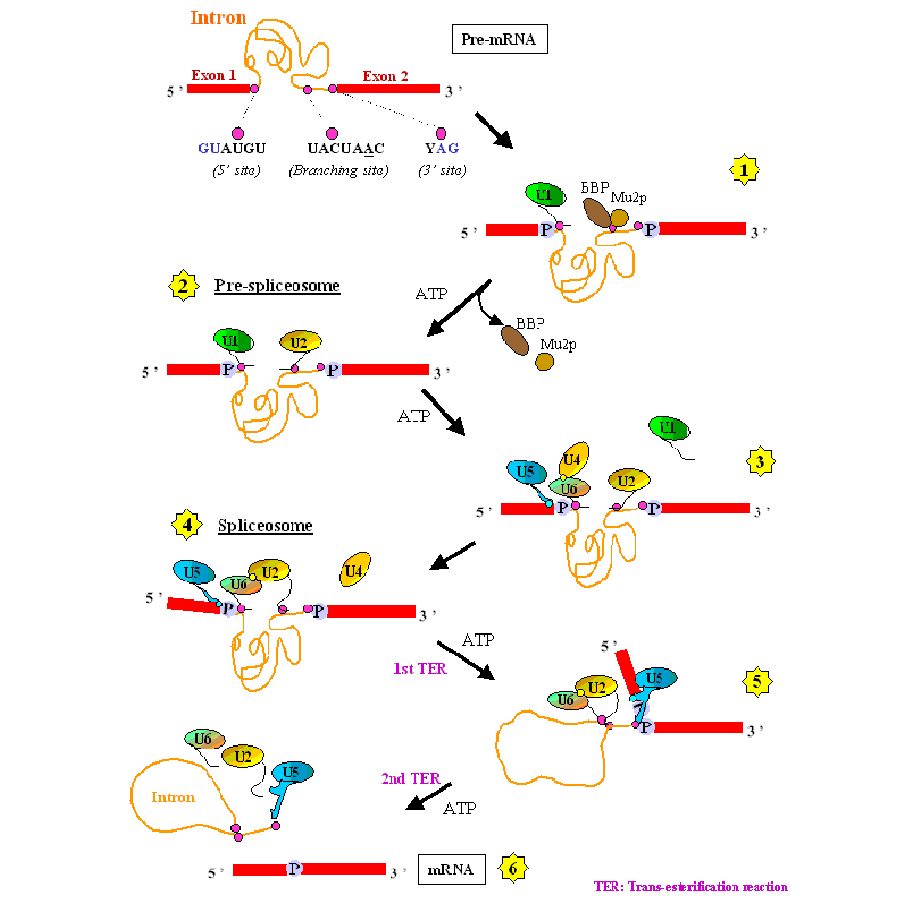


Figure 1.9. The various steps involved in the stepwise assembly of the spliceosomal complex. (1) Binding of U1 snRNP to the 5’ splice site and BBP and Mu2p splicing factors to the branch site of the intron. (2) Formation of pre-spliceosome, which involves the dissociation of BBP and Mu2p splicing factors 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).

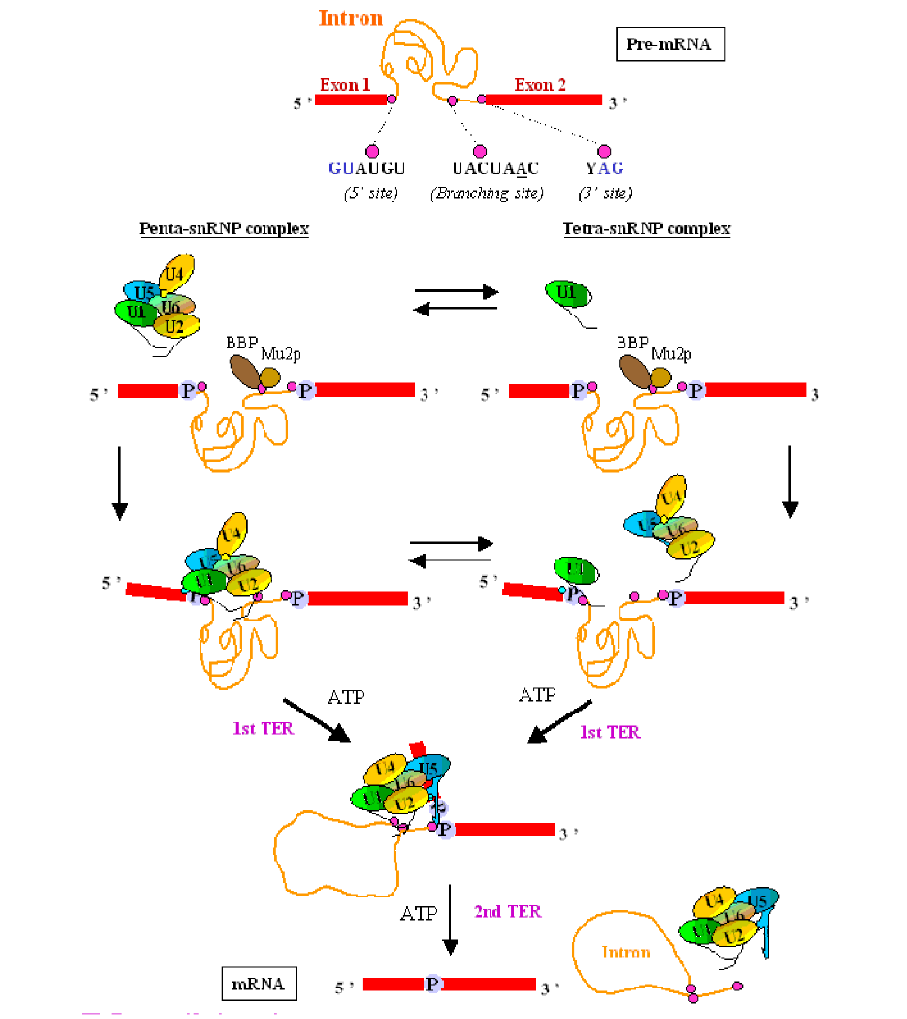


Figure 1.10. This pathway can be further divided into two depending on the number (tetra/penta) of snRNPs that are present in the pre-assembled spliceosomal complex. The tetra/penta snRNP complex assembles on the pre-mRNA, which is followed by the 1st trans-esterification reaction that involves the splicing 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).

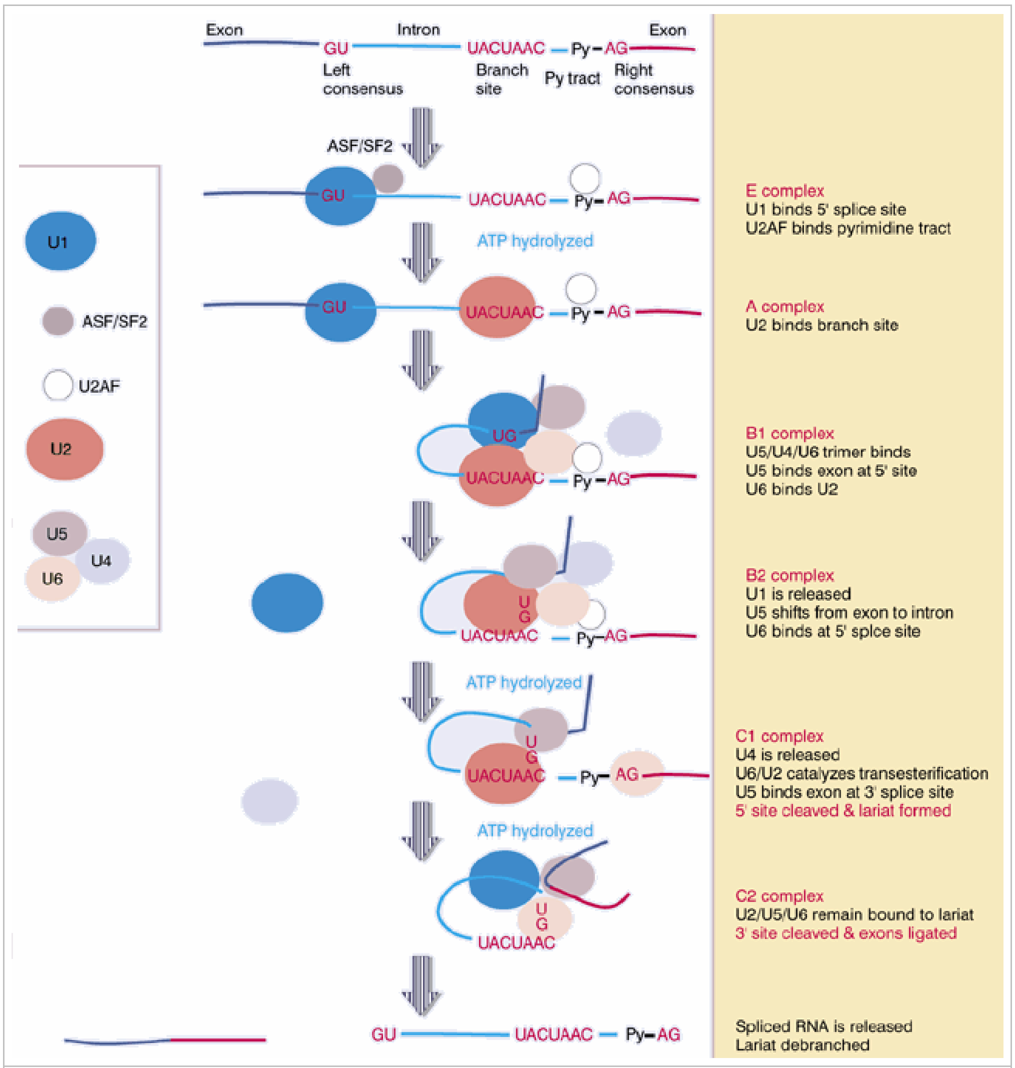


Figure 1.11. The splicing reaction proceeds through discrete stages in which spliceosome formation involves the interaction of components that recognize the consenses sequences and the formation of various spliceosomal complexes.

(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 catalysis 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.

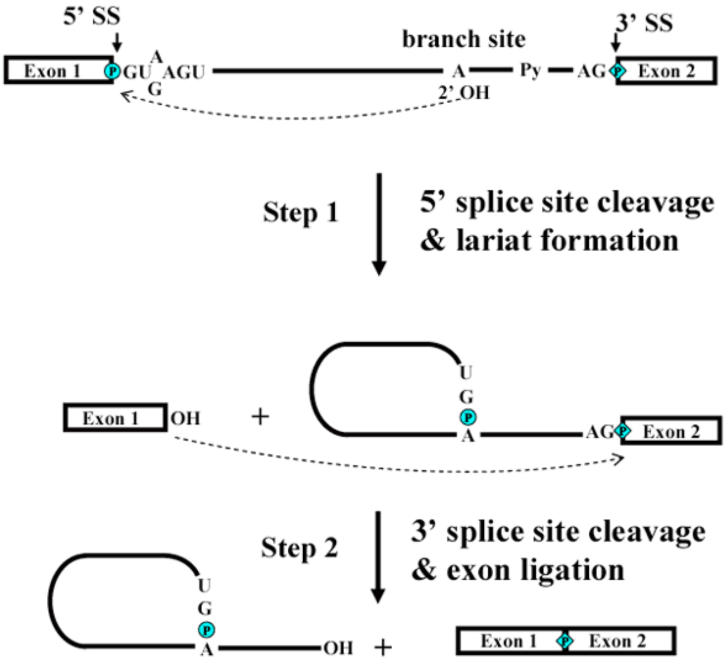


Figure 1.12. Nuclear splicing occurs by two transesterification reactions in which an OH groups attacks a phosphodiester bond, which leads to the splicing of the 5' and 3' ends of the intron and joining of the exons to form of a mature mRNA.

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 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 splicing. 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 governs 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

- Cavalier-Smith, T. (1985). Eukaryotic gene numbers, non-coding DNA, and genome size. In Cavalier-Smith T, ed. *The Evolution of Genome Size* Chichester: John Wiley.
- Cristianini, N. and Hahn, M. (2006). *Introduction to Computational Genomics*, Cambridge University Press.
- Dreyfuss, G., Matunis, M.J., Pinol-Roma, S. and Burd, C.G. (1993). hnRNP proteins and the biogenesis of mRNA. *Ann. Rev. Biochem.* 62, 289-321.
- Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., Snyder, M. (2007). “What is a gene, post-ENCODE? History and updated definition”. *Genome Research* 17 (6): 669-681.
- Grabowski, P. J., Seiler, S. R., and Sharp, P. A. (1985). A multi component complex is involved in the splicing of messenger RNA precursors. *Cell* 42, 345-353.
- Guthrie, C. and Patterson, B. (1988). Spliceosomal snRNAs. *Ann. Rev. Genet.* 22, 387-419.
- Lamond, A. I. (1988). Spliceosome assembly involves the binding and release of U4 small nuclear ribonucleoprotein. *Proc. Nat. Acad. Sci. USA* 85, 411-415.
- Le Hir, H., Gatfield, D., Izaurralde, E., and Moore, M. J. (2001). The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.* 20, 4987-4997.
- Mount, D.W. (2004). *Bioinformatics: Sequence and genome analysis*, 2nd ed., Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.
- Pearson, H. (2006). “Genetics: what is a gene?” *Nature* 441 (7092): 398–401.

- Pennisi, E. (2007). “DNA Study Forces Rethink of What It Means to Be a Gene”. *Science* 316 (5831): 1556–1557.
- Reed, R. and Maniatis, T. (1985). Intron sequences involved in lariat formation during pre-mRNA splicing. *Cell* 41, 95-105.
- Will, C.L. and Luhrmann (2001). Spliceosomal UsnRNP biogenesis, structure and function. *Curr. Opin. Cell Biology*, 13:290-301.
- Woodson, S. A. (1998). “Ironing out the kinks: splicing and translation in bacteria”. *Genes Dev.* 12 (9): 1243–7.
- Wu, S., Romfo, C. M., Nilsen, T. W. and Green, M. R. (1999). Functional recognition of the 3’ splice site AG by the splicing factor U2AF35. *Nature* 402, 832-835.

Chapter 2

1/*f* Correlations in Viral Genomes - A Fast Fourier Transformation (FFT) Study

2.1 Introduction

Complete genome sequences are being generated far more rapidly than our ability to interpret and comprehend the biological meaning of the data. The availability of the complete genome sequences of different organisms paves the way to study the various characteristic features of genome organization and structure. Prokaryotic and eukaryotic genomes have been studied by various methods including auto-correlation function analysis (Herzel *et al.*, 1999), DNA walking (Nee, 1992), Fourier spectral analysis (Fukushima, *et al.*, 2002), (Fukushima, *et al.*, 2001), mutual information function and wavelet translation. Such studies on the complete genomes of different microbes have also elucidated the fractal characteristics of DNA (Vieira, 1999), (Surya Pavan, and Mitra, 2005), (Upadyay, 2003). Fourier analysis has been found to be an important method for studying some of the features of the genome organization of different organisms.

2.1.1 Fourier analysis

Fourier analysis is the decomposition of a function in terms of a sum of sinusoidal functions (called basis functions) of different frequencies that can be recombined to obtain the original function. The recombination process is called Fourier synthesis (in which case, *Fourier analysis* refers specifically to the decomposition process).

2.1.2 Fourier series

In mathematics, the Fourier series is one of the specific forms of Fourier analysis. It is a way of decomposing a function into a superposition of elementary waves. Introduced by Joseph Fourier (1768–1830) for the purpose of solving the heat equation in a metal plate, it led to a revolution in mathematics. Fourier series are named in the honor of Joseph Fourier, who made important contributions to the study of trigonometric series. Although the original motivation was to solve the heat equation, it later became obvious that the same techniques could be applied to a wide array of mathematical and physical problems. Some of the applications of Fourier series are in electrical engineering, vibration analysis, acoustics, optics, signal, image processing and many more.

2.1.3 Fourier transform

Continuous Fourier transform is one of the specific forms of Fourier analysis. As such, it transforms one function into another, which is called the *frequency domain representation* of the original function (where the original function is often a function in the time-domain). In this specific case, both domains are continuous and unbounded. The term Fourier transform can refer to either the frequency domain representation of a function or to the process/formula that "transforms" one function into the other.

(i) Time-domain and frequency-domain representations of sound

Any acoustic signal can be graphically or mathematically depicted in either of two forms, called the time-domain and frequency-domain representations. In the time domain, the amplitude of a signal is represented as a function of time. Such a signal is called a sinusoid because its amplitude is a sine function of time, characterized by some frequency, which is measured in cycles per second, or Hertz (Hz). In the frequency domain, the amplitude of a signal is represented as a function of frequency.

Any sound can be represented as a sum of pure tones (sinusoidal components). Each tone in the series has particular amplitude, relative to the others, and a particular phase relationship. The frequency composition of complex signals is usually not apparent from inspection of the time-domain representation. Fourier (Spectrum) analysis is the process of converting the time-domain of a signal to a frequency-domain representation that shows how different frequency components contributes to the sound. The complete frequency-domain representation of a signal consists of two parts (Figure 2.1).

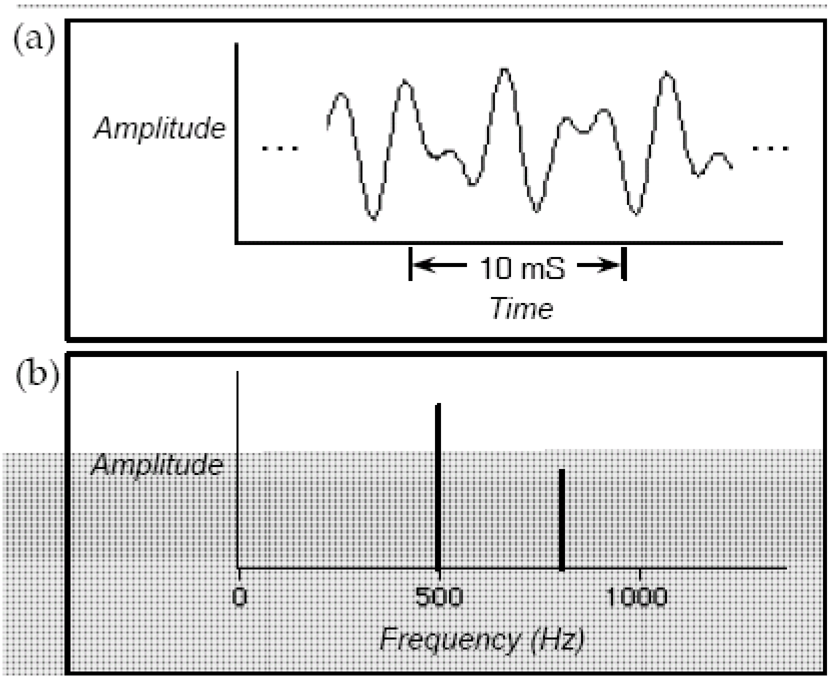


Figure 2.1. Time-domain and frequency-domain representations of an infinitely long sound consisting of two tones, with frequencies of 490 Hz and 800 Hz. (a) Time domain. (b) Frequency domain.

The magnitude spectrum (Figure 2.1b) contains information about the magnitude of each frequency component in the entire signal. The phase spectrum contains information about the phase or timing relationships among the frequency components.

So, Fourier transform is a mathematical function that converts the time-domain form of a signal to a frequency-domain representation or spectrum. When the signal and spectrum are represented as sequence of discrete digital samples, a version of the Fourier transform called the discrete Fourier transform (DFT) is used. The input is a finite sequence of values—the amplitude values of the signal—sampled at regular intervals. The output is a sequence of values specifying the amplitudes of a sequence of discrete frequency components, evenly spaced from zero Hz to half the sampling frequency (Figure 2.2).

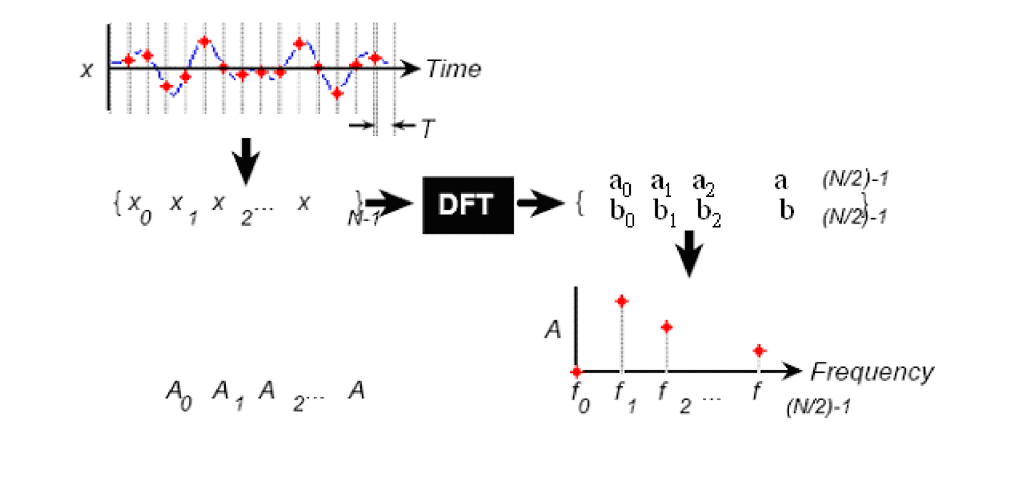


Figure 2.2. Schematic representation of the discrete Fourier transform (DFT) as a black box. The input to the DFT is a sequence of digitized amplitude values (*x0, x1, x2, ... xN-1*) at *N* discrete points in time. The output is a sequence of amplitude values (*a0, a1, a2, ... a(N/2)-1*) and phase values (b*0,b1,b2, ... b(N/2)-1*) at *N/2* discrete frequencies . The highest frequency, *f(N/2)-1*, is equal to half the sampling rate (1/(2T), where T is the sampling period, as shown in the figure). The output can be plotted as a magnitude spectrum.

2.1.4 Fast Fourier transform

A fast Fourier transform (FFT) is an efficient algorithm to compute the discrete Fourier transform (DFT) and its inverse. FFTs are of great importance to a wide variety of applications, from digital signal processing and solving partial differential equations to developing algorithms for quick multiplication of large integers.

2.1.5 Long-range correlations

There are many systems, which are considered to be complex because they have structures at different length scales. The existence of structures at very large scales results in long-range correlations. These long-range correlations can be detected by examining the two-point correlation function to see whether it decays slower than an exponential function, or, whether it reaches the zero value at a very large scale.

If a correlation function decays as a power law, we have a scaling phenomenon. The power spectrum *P(f)*, which is the Fourier transformation of the correlation function, will also be a power law function: $P(f) \sim 1/f^{\alpha}$, where *f* is the frequency and α is the scaling exponent. If the two-point correlation function decays even slower than a power law, such as the case of logarithmic function, the power spectrum is then exactly inversely proportional to the frequency, i.e. $P(f)\sim 1/f^{\alpha}$, or α=1. For the two cases, there is an interesting connection of the $1/f^{\alpha}$ noise—time series with $1/f^{\alpha}$ power spectra—that are quite common in nature. The only difference is that it is the spatial power spectrum instead of the temporal spectral that is calculated.

So, complex systems such as DNA, also have structures at large scales, resulting in statistical patterns like long-range correlations. The long-range correlations, which follow power-law correlation function, are a result of the dependency of a single nucleotide on all other nucleotides over large distances (Figure 2.3).

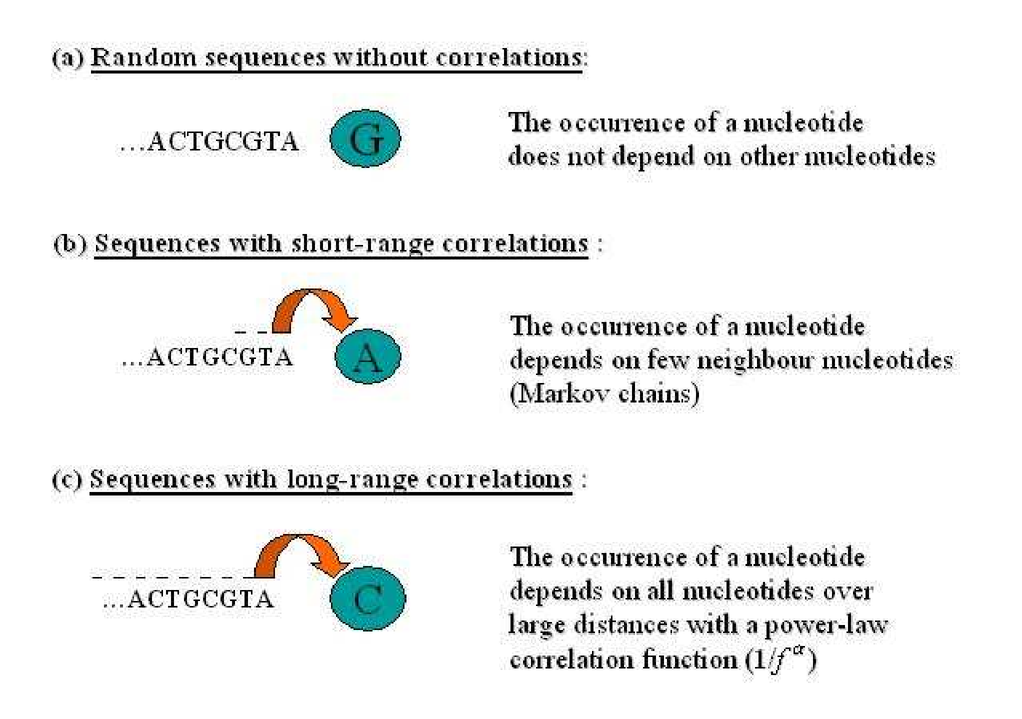

(a) Random sequences without correlations:
…ACTGCGTA
G
The occurrence of a nucleotide does not depend on other nucleotides
(b) Sequences with short-range correlations :
…ACTGCGTA
A
The occurrence of a nucleotide depends on few neighbour nucleotides (Markov chains)
(c) Sequences with long-range correlations :
…ACTGCGTA
C
The occurrence of a nucleotide depends on all nucleotides over large distances with a power-law correlation function ($1/f^{\alpha}$)

Figure 2.3. Illustration showing the dependency on the occurrence of the nucleotides in the (a) random sequence (without correlations), (b) sequence with short-range correlations and (c) sequence with long-range correlations.

In genomic sequences these correlation properties can extend over tens to thousands of base pairs and are found to be scale invariant (Figure 2.4).

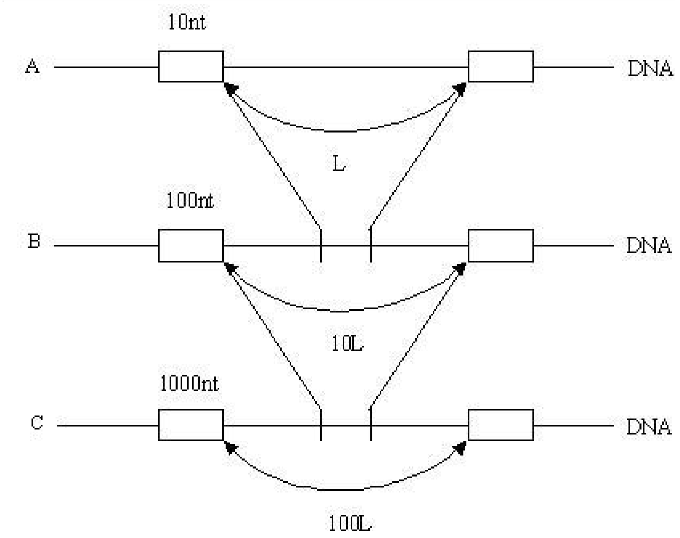


Figure 2.4. Illustration showing that the nucleotide compositions are correlated to each other in the same manner whatever is the scale as shown in (A), (B) and (C).

(i) Long-range correlations in DNA

There has been considerable interest in the finding of long-range correlations in DNA sequences. The possibility of existence of these patterns, hidden in DNA base sequence was also demonstrated by one-over-f (1/*f* noise) spectra. Power-law decay for the correlations as a function of time translates into a power-law decay of the spectrum as a function of frequency, which is called "1/*f* noise". The power spectra *S(f)* as a function of frequency *f* behaves like: $S(f) = 1/f^{\alpha}$, where the exponent α is close to 1. Spectral representation of a DNA sequence has been found to have the following applications i.e.,

- To identify underlying periodic patterns in the sequence, which manifest as peaks at specific frequencies in the power spectrum.
- To ascertain whether a sequence is random lacking any correlation pattern exhibiting flat power spectrum.

(ii) Earlier studies on long-range correlations

Earlier studies suggested that intronic sequences have long-range correlations signifying the power-law correlation function. Two ways have been proposed to generate spatial long-range correlations in DNA, one is by elongation (expansion of the space) i.e., expansion with a small amount of error that leads to spatial scaling and $1/f^{\alpha}$ spectra. Another way is by repetition of the same structure with the repeated structures being separated by some arbitrary long distances.

(iii) Evolution of long-range correlations

So, it was proposed that 1/*f* behavior could be explained by "expansion-modification" model, according to which, the length of the DNA of the present day organisms is longer than the pre-biotic ones. And during course of evolution, the DNA had undergone elongation by repeated process of duplication, followed by mutation, which lead to these long-range correlations (Li and Kaneko, 1992), (Li and Kaneko, 1992). In fact gene duplication is an essential feature of life, in which the elongation has been accomplished by the duplication of the original sequence. The duplication is typically not perfect, but is with a small amount of error. These repetitive structures in DNA sequences are very common, and have been experimentally observed in many different biological systems.

2.1.6 Motivation for the study

So, according to the expansion-modification theory the genomes of the present day organisms are longer than the pre-biotic ones because the elongations and modifications have occurred mostly in the intronic (non-coding) regions (as exonic (coding) regions are more resistant to these changes). This theory clearly emphasizes the fact that these long-range correlations, (which are a result of the rearrangements in DNA) are the characteristic features of the intronic regions only. In view of this hypothesis, it is important to identify the presence of these long-range correlations in the exonic (coding) regions also.

2.2 Methodology

2.2.1 Organisms considered for the present study

We have studied the complete genome sequences of dsDNA viruses to identify the presence of long-range correlations in the intron-less DNA using FFT approach. We have taken the complete DNA sequence of chromosome I of *S. cerevisiae* containing intronic sequences as a control for our study, which is found to exhibit long-range correlations. For this study, we have obtained the complete DNA (RefSeq) sequences of ten different dsDNA viruses and the chromosome I of *S. cerevisiae*, from GenBank database accessed through National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov). The rationale to study these viruses is that all are dsDNA viruses, with their genome length in the range of 0.1-0.3 million base pairs (mbps) and the host they infect range from invertebrates to vertebrates. We have also made an attempt to study the correlation between the distribution of gene length and the domain of 1/*f* region in these organisms. The following gives a detailed review of each of the organisms under study.

(i) ***Amsacta moorei*** **entomopoxvirus**

<u>Scientific classification</u>

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirales*
Family	*Poxviridae*
Sub-family	*Entomopoxvirinae*
Genus	Entomopoxvirus B
Type Species	***Amsacta moorei*** **entomopoxvirus**
Hosts	Invertebrates

Entomopoxviruses (EPVs) are linear dsDNA viruses that infect insects. They were first described by Vago (Vago, 1963) and have been isolated from several insect orders including Coleoptera, Lepidoptera, Orthoptera and Diptera (Arif, 1995). *Amsacta moorei* EPV is probably the most well characterized insect poxvirus, primarily because it replicates in cultured insect cells (Goodwin *et al*., 1990; Marlow *et al*., 1993; Winter *et al*., 1995).

The genome of the genus B entomopoxvirus from *Amsacta moorei* entomopoxvirus virus (AmEPV) was sequenced and was found to contain 232,392 bp with 279 unique open reading frames (ORFs) of greater than 60 amino acids. The central core of the viral chromosome is flanked by 9.4 kb inverted terminal repeats (ITRs), each of which contains 13 ORFs, raising the total number of ORFs within the viral chromosome to 292. The complete genomic sequence of AmEPV paves the way for an understanding and comparison of the molecular properties and pathogenesis between the entomopoxviruses of insects and the more intensively studied vertebrate poxviruses.

(ii) *Cercopithecine* herpesvirus 17

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirales*
Family	*Herpesviridae*
Subfamily	*Gammaherpesvirinae*
Genus	*Rhadinovirus*
Type Species	***Cercopithecine*** **herpesvirus 17**
Hosts	Vertebrates

Rhadinoviruses are a genus of herpesviruses that include the *Cercopithecine* herpesvirus 17. The term rhadino comes from the Latin word - fragile, referring to the tendency of the viral genome to break apart when it is isolated. They are large double-stranded viruses that possess up to 100 genes in a single long chromosome, which is flanked by repetitive DNA sequences called terminal repeats. Rhadinoviruses generally infect B-lymphocytes and once infection occurs, it is generally life-long. Rhadinoviruses have been found in New World monkeys such as the squirrel monkeys (herpesvirus saimiri) and in mice (murine gammaherpesvirus-68). A new form of rhadinovirus called rhesus rhadinovirus have also been discovered in Old World monkeys. The genome of *Cercopithecine* herpesvirus 17 consists of 133,719 bp.

(iii) *Ectocarpus siliculosus* virus

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirales*
Family	*Phycodnaviridae*
Genus	*Phaeovirus*
Type Species	***Ectocarpus siliculosus*** **virus**
Hosts	Algae

Ectocarpus siliculosus Virus (EsV-1) is the founding member of a new genus of marine viruses, the *phaeoviruses*. EsV-1 is endemic in all populations of *Ectocarpus siliculosu*s. EsV-1 has a large circular double-stranded DNA genome of about 320 kb (Delaroque, 2001). *Ectocarpus siliculosus* is a brown alga that takes the form of an uniseriate (arranged in one row), branched filament, which may appear as light brown tangled tufts. It is common throughout the year and is found in pools at high tide levels, on all coasts of temperate climate zones. It is frequently free floating but it may be attached to rocks or other algae. The genome is not segmented and contains a single molecule of linear double-stranded DNA. The complete genome is 335,593 bp).

(iv) Human herpesvirus 4 (Epstein Barr virus – EBV)

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirale*
Family	*Herpesviridae*
Sub-family	*Gammaherpesvirinae*
Genus	Lymphocryptovirus
Type Species	**Human herpesvirus 4**
Hosts	Vertebrates

The Epstein-Barr virus (EBV), also called Human herpesvirus 4 (HHV-4), is a virus of the herpes family (which includes *Herpes simplex virus* and *Cytomegalovirus*), and is one of the most common viruses in humans. EBV is named after Michael Epstein and Yvonne Barr, who together with Bert Achong, discovered the virus in 1964 (Epstein, *et. al.* 1964). Epstein-Barr can cause infectious mononucleosis, also known as “glandular fever”, “Mono” and “Pfeiffer's disease”, Burkitt’s lymphoma, nasopharyngeal carcinoma (Lerner, *et. al.,* 2004), multiple sclerosis and other autoimmune diseases (Lunemann and Munz, 2007).

EBV has a toroid-shaped protein core wrapped with DNA, a nucleocapsid, a protein tegument and an outer envelope. The envelope carries Gp spikes of which the most abundant structural glycoprotein (Gp) is Gp350/220. The EBV genome is a linear, double stranded DNA molecule of 172,281 bp (172 kb) and has reiterated 0.5 kb terminal repeats and a reiterated 3 kb internal direct repeat.

(v) Human herpesvirus 6

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirale*
Family	*Herpesviridae*
Sub-family	*Betaherpesvirinae*
Genus	Roseolovirus
Type Species	**Human herpesvirus 6**
Hosts	Vertebrates

HHV-6 is one of the eight known viruses that are members of the human herpesvirus family. HHV-6 is a member of the betaherpesviridae (subfamily of the herpesvirinae). There are two subtypes of HHV-6 termed HHV-6A and HHV-6B. Various primary infections usually cause fever, with exanthem subitum (Roseola/rash), which is being observed in 10% of cases (Hall, *et.al.,* 1994) and are associated with several more severe complications, such as encephalitis, lymphadenopathy, myocarditis and myelosuppression.

The genome of HHV-6 is 159,321 bp in size, it has a base composition of 43% G+C, and contains 119 open reading frames. The overall structure is 143 kb bounded by 8 kb of direct repeats left (DRL) and direct repeats right (DRR), containing 0.35 kb of terminal and junctional arrays of human telomere-like simple repeats. The genome is considered to contain 102 separate genes likely to encode protein. The genes are arranged colinearly with those in the genome of the previously sequenced betaherpesvirus, human cytomegalovirus (Gompels, *et. al*., 1995).

(vi) Human herpesvirus 6B

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirale*
Family	*Herpesviridae*
Sub-family	*Betaherpesvirinae*
Genus	Roseolovirus
Type Species	**Human herpesvirus 6B**
Hosts	Vertebrates

There are two forms of HHV-6, A and B. A is rare, and acquired in adulthood, B is relatively common, usually acquired in childhood, and associated with roseola. HHV-6B is responsible for up to 93% of primary infections. It causes roseola, febrile illnesses and encephalitis in infants and reactivates in transplant patients, causing complications such as encephalitis, pneumonitis and liver failure. HHV-6B infects close to 100% of children by the age of two, causing mild flu-like symptoms and rash in some, but occasionally progresses to high fever, encephalitis and seizures. In most cases, the virus goes into latency. However, in patients with impaired immune function, the virus may persist in its active state at low levels for years. The genome of HHV-6B is 162,114 bp in length.

(vii) Lymphocystis disease virus 1

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirales*
Family	*Iridoviridae*
Genus	Lymphocystis virus
Type Species	**Lymphocystis disease virus 1**
Hosts	Vertebrates

Lymphocystis disease (LCD) is the viral disease most frequently detected in fish farms located in the Mediterranean area, affecting mainly cultured gilt-head seabream (*Sparus aurata*, L.) (Paperna *et al*., 1982). Lymphocystis disease virus (LCDV) is the etiological agent of this chronic and self-limited disease characterized by the appearance of pearl-like nodules, formed by hypertrophic fibroblastic cells, located on skin and fins (Samalecos, 1986). LCDV has been identified as an iridovirus (Fischer, 1988, Wolf, 1988) and is distributed worldwide. In 1997, the LCDV-1 complete genomic DNA sequence was determined. The genome is 102,653 bp in length and contains 195 open reading frames (ORFs) (Tidona and Darai, 1997).

(viii) Meleagrid herpesvirus 1

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirales*
Family	*Herpesviridae*
Sub-family	*Alphaherpesvirinae*
Genus	*Mardivirus*
Type Species	**Meleagrid herpesvirus 1**
Hosts	Turkey

Herpesvirus of turkey (HVT) or Meleagrid herpesvirus 1 is a naturally occurring, non-pathogenic alphaherpesvirus that was originally isolated from domestic turkeys in the late 1960s (Kawamura *et al.*, 1969, Witter *et al.*, 1970). As a member of the genus *Mardivirus* (Fauqet *et al.*, 2005), HVT is antigenically related to Marek's disease virus (MDV), the causative agent of the highly contagious neoplastic Marek's disease (MD) in chickens. The complete genome sequence of the FC126 strain of HVT has recently been determined (Afonso *et al.*, 2001, Kingham *et al.*, 2001) with the genome length of 159,160 bp. It is estimated to contain at least 99 functional genes, many of which have homologues in MDV and other herpesviruses. The genome is not segmented and contains a single molecule of linear double-stranded DNA. It has a guanine + cytosine content of 56 % and also has terminally redundant sequences, which are reiterated, but not repeated internally.

(ix) Sheep pox virus

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirales*
Family	*Poxviridae*
Sub-family	*Chordopoxvirinae*
Genus	Capripoxvirus
Type Species	**Sheeppox virus**
Hosts	Vertebrates

Sheeppox virus (SPPV), member of the *Capripoxvirus* genus of the *Poxviridae*, is an etiologic agent of important diseases of sheep in northern and central Africa, southwest and central Asia, and the Indian subcontinent. Sheep poxviruses are usually transmitted by the respiratory route, but may also enter the body through abraded skin. Infectious virus is found in all secretions, excretions, and the scabs from skin lesions.

The genome is not segmented and contains a single molecule of linear double-stranded DNA. The complete genome is 144,575 bp long and the DNA is fully sequenced. The genome has a G+C content of 36 % and has termini with cross-linked hairpin ends (i.e. single-stranded loopes thus forming one continuous polynucleotide chain). The genome has terminally redundant sequences, which have reiterated inverted terminal sequences that are tandemly repeated.

(x) Yaba-like disease virus

Scientific classification

Kingdom	Viruses
Phylum	DNA viruses
Class	dsDNA viruses, no RNA stage
Order	*Caudovirales*
Family	*Poxviridae*
Sub-family	*Chordopoxvirinae*
Genus	Yatapoxvirus
Type Species	**Yaba-like disease virus**
Hosts	Vertebrates

The Yatapoxvirus genus of poxviruses is comprised of three virus isolates: Yaba-like disease virus (YLDV), Tanapox virus (TPV), and Yaba monkey tumor virus (YMTV). YLDV belongs to the *Yatapoxvirus* genus of the

Chordopoxvirinae and causes infections in primates (Knight *et al*., 1989) including humans that are characterized by an acute febrile illness accompanied by localized skin lesions.

The genome sequence of Yaba-like disease virus has been determined, which is 230,208 bp in length and contains inverted terminal repeats (ITRs) of 1883 bp. Within 20 nucleotides of the termini, there is a sequence that is conserved in other poxviruses and is required for the resolution of concatemeric replicative DNA intermediates. The nucleotide composition of the genome is 73% A+T, but the ITRs are only 63% A+T. The genome contains 151 tightly packed open reading frames (ORFs) that either are ≥180 nucleotides in length or are conserved in other poxviruses.

(xi) Saccharomyces cerevisiae

Scientific classification

Kingdom	Fungi
Phylum	Ascomycota
Class	Saccharomycetes
Order	Saccharomycetales
Family	Saccharomycetaceae
Genus	*Saccharomyces*
Type Species	*Saccharomyces cerevisiae*

The word "*Saccharomyces*" is derived from Greek, which means "sugar mold" and "*Cerevisiae*" is derived from Latin, which means "of beer". *Saccharomyces cerevisiae* is a species of budding yeast and perhaps the most useful yeast owing to its use since ancient times in baking and brewing. It is one of the most intensively studied eukaryotic model organisms in molecular and cell biology. *Saccharomyces cerevisiae* cells are round to ovoid, 5–10 micrometres (μm) in diameter.

S. cerevisiae was the first eukaryotic genome that was completely sequenced. The yeast genome database (Wickner et al., 2008) is highly annotated and remains a very important tool for developing basic knowledge about the function and organization of eukaryotic cell genetics and physiology. The genome is composed of about 13,000,000 base pairs and 6,275 genes, compactly organized on 16 chromosomes. Only about 5,800 of these are believed to be true functional genes. It is estimated that yeast shares about 23% of its genome with that of humans.

2.2.2 Power spectral analysis

For power spectrum analysis, we have calculated the "adenine plus guanine" (A+G) and "adenine plus thymine" (A+T) proportions of the complete genomes of each of the organisms in a non-overlapping window of size 32, (considering the proportions is a better method than that used in earlier studies). Results were also compared using window sizes of 64 and 128 bases. We have observed that a small window size gives more information at a lower scale, but a larger window provides better signal to noise (because of averaging over a longer region). After comparing the graphs for three base sizes, we have used the window size of 32 in all the following studies. These proportions were Fourier-transformed using the Fast-Fourier-Transform (FFT) algorithm, which calculates the discrete-Fourier transformation (DFT) of a function of *N* points. This algorithm speeds up the calculation of power spectrum by a factor of $N\log_2 N$, but requires the length of sequence being analyzed to be an integral power of two (Press, *et al.*, 1992). To accomplish this we have calculated the respective proportions of the entire length of the genomes and zero padded it to the next integral power of two.

The DFT (H_n) of a function from a finite number of *N* sampled points h_k is given as in Eq. (2.1).

$$H_n = \sum_{k=0}^{N-1} h_k e^{2\pi ikn/N} \qquad \ldots(2.1)$$

DFT maps *N* complex numbers h_k's into *N* complex numbers (H_n's). Eq. (2.1) is periodic in *n*, with a period *N* and the variables *n* and *k* vary from 0 to *N*-1 and $i^2 = -1$. The following gives a detailed description of the algorithm involved in the above mentioned study, which involves the transformation of the data (nucleotide proportions; A+G and A+T) from the time domain to frequency domain.

2.2.3 Fourier transform

A physical process can be described either in the time domain, by the values of some quality *h* as a function of time *t*, e.g., *h(t),* or else in the frequency domain, where the process is specified by giving its amplitude *H* as a function *f*, that is *H(f)*, with $-\infty < f < \infty$. For many purposes it is useful to think of *h(t)* and *H(f)* as being two different representations of the same function. One goes back and forth between theses two representations by means of the Fourier transform equations as given in Eq. (2.2) and (2.3),

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{2\pi i f t} dt \qquad (2.2)$$

$$h(t) = \int_{-\infty}^{\infty} H(f) e^{-2\pi i f t} df \qquad (2.3)$$

If *t* is measured in seconds, then *F* in Eq. (2.2) is in cycles per second, or Hertz.

(i) Fourier transform of discretely sampled data

In the most common situations, function *h(t)* is sampled at evenly spaced intervals in time. Let Δ denote the time interval between consecutive samples, so that the sequence of sampled values is given as in Eq. (2.4),

$$h_n = h(n\Delta) \qquad n = \ldots,-3,-2,0,1,2,3,\ldots. \qquad (2.4)$$

The reciprocal of the time interval Δ is called the sampling rate; if Δ is measured in seconds, for example, then the sampling rate is the number of samples recorded per second.

(ii) Sampling theorem and aliasing

For any sampling interval Δ, there is also a special frequency f_c, called the *Nyquist critical frequency*, given by Eq. (2.5),

$$f_c \equiv \frac{1}{2\Delta} \qquad (2.5)$$

If a sine wave of the Nyquist critical frequency is sampled at its positive peak value, then the next sample will be at its negative trough value, the sample after that at the positive peak again, and so on.

(iii) Discrete Fourier transform

We now estimate the Fourier transform of a function from a finite number of its sampled points. Suppose that we have N consecutive sampled values, which are given as in Eq. (2.6),

$$h_k \equiv h(t_k), \qquad t_k \equiv k\Delta, \qquad k = 0,1,2,..., N\text{-}1 \qquad (2.6)$$

such that the sampling interval is Δ and *N* is even. If the function *h*(*t*) is nonzero only in a finite interval of time, then that whole interval of time is supposed to be contained in the range of the *N* points given. Alternatively, if the function *h*(*t*) goes on forever, then the sampled points are supposed to be at least "typical" of what *h*(*t*) looks like at all other times.

With *N* numbers of input, we will evidently be able to produce no more than *N* independent numbers of output. So, instead of trying to estimate the Fourier transform *H*(*f*) at all values of *f* in the range −*fc* to *fc*, let us seek estimates only at the discrete values as given in Eq. (2.7),

$$f_n \equiv \frac{n}{N\Delta}, \qquad n = -\frac{N}{2},...,\frac{N}{2} \qquad (2.7)$$

The extreme values of *n* in (Eq. (2.7)) correspond exactly to the lower and upper limits of the Nyquist critical frequency range. If we have noticed that there are *N* + 1, but not *N*, values of *n* in (Eq. (2.7)); and it will turn out that the two extreme values of *n* are not independent (in fact they are equal), but all the others are, this reduces the count to *N*.

The remaining step is to approximate the integral in (Eqs. (2.2) and Eq. (2.3)) by a discrete sum as given in Eq. (2.8):

$$H(f_n) = \int_{-\infty}^{\infty} h(t)e^{2\pi i f_n t}dt \approx \sum_{k=0}^{N-1} h_k e^{2\pi i f_n t_k}\Delta = \Delta\sum_{k=0}^{N-1} h_k e^{2\pi i k n/N} \qquad (2.8)$$

Here equations (Eq. (2.6)) and (Eq. (2.7)) have been used in the final equality. The final summation in equation (Eq. (2.8)) is called the *discrete Fourier transform* of the *N* points h_k. Let us denote it by H_n, and it is given as in Eq. (2.9),

$$H_n = \sum_{k=0}^{N-1} h_k e^{2\pi ikn/N} \qquad (2.9)$$

The discrete Fourier transform maps *N* complex numbers (the h_k's) into *N* complex numbers (the H_n's). It does not depend on any dimensional parameter, such as the time scale Δ. The relation (Eq. (2.8)) between the discrete Fourier transform of a set of numbers and their continuous Fourier transform when they are viewed as samples of a continuous function sampled at an interval Δ can be rewritten as in Eq. (2.10),

$$H(f_n) \approx \Delta H_n \qquad (2.10)$$

where f_n is given by (Eq. (2.7)).

The formula for the discrete *inverse* Fourier transform, which recovers the set of h_k's exactly from the H_n's is given as in (Eq. (2.11)):

$$h_k = \frac{1}{N} \sum_{n=0}^{N-1} H_n e^{-2\pi ikn/N} \qquad (2.11)$$

(iv) Fast Fourier Transform (FFT)

Let us define *W* as the complex number, which is written as in Eq. (2.12),

$$W \equiv e^{2\pi i/N} \qquad (2.12)$$

Then the Eq. (2.9) can be written as (Eq. (2.13)),

$$H_n = \sum_{k=0}^{N-1} W^{nk} h_k \qquad (2.13)$$

In other words, the vector of h_k's is multiplied by a matrix whose $(n; k)^{\text{th}}$ element is the constant *W* to the power *n* x *k*. The matrix multiplication produces a vector result whose components are the H_n's. This matrix multiplication evidently requires N^2 complex multiplications, plus a smaller number of operations to generate the required powers of W. So, the discrete Fourier transform appears to be an $O(N^2)$ process. The discrete Fourier transform can, in fact, be computed in $O(N \log 2\ N)$ operations with an algorithm called the *fast Fourier transform*, or *FFT* (Press, *et al*., 1992). With $N = 10^6$, for example, it is the difference between, roughly, 30 seconds of CPU time and 2 weeks of CPU time on a microsecond cycle time computer. The existence of an FFT algorithm became generally known only in the mid-1960s, from the work of J.W. Cooley and J.W. Tukey.

(v) The Cooley-Tukey algorithm

By far the most common FFT is the Cooley-Tukey algorithm. This is a divide and conquer algorithm that recursively breaks down a DFT of any composite size $N = N_1N_2$ into many smaller DFTs of sizes N_1 and N_2, along with O(N) multiplications by complex roots of unity traditionally called twiddle factors. The most well-known use of the Cooley-Tukey algorithm is to divide the transform into two pieces of size $N / 2$ at each step, and is therefore limited to power-of-two sizes, but any factorization can be used in general. These are called the radix-2 and mixed-radix cases, respectively (and other variants such as the split-radix FFT have their own names as well). This algorithm has been used extensively for the quick computation of the FFT of any discrete funtion and has also been applied to identify the occurrence of any statistical pattern in the genome sequences of various organisms.

2.2.4 Calculation of power

The power for each of the proportions was calculated by taking the square root of sum of squares of real and imaginary components of the Fourier amplitude, given as in Eq. (2.14):

$$|H_n(f_j)| = \sqrt{((H_n{}^{real}(f_j))^2 + (H_n{}^{imag}(f_j))^2)} \qquad \ldots(2.14)$$

The power spectrum was visually divided into two linear parts and each section was independently fitted with a linear regression line (using the built-in function of SigmaPlot plotting package). The slope values (-α) are reported in Table 2.3 for comparison and these values were correlated with the gene length distribution graphically.

2.2.5 Calculation of average gene length

The correlation between the distribution of gene length and the domain of 1/*f* region of power spectrum was examined by extracting the average gene length of each species from the same database.

2.3 Results and Discussions

The details of the genomes (NCBI accession number and genome length) analyzed in the present study are given in Table 2.1.

Table 2.1. Genomes analyzed in the present study

	Organisms	NCBI Accession no.	Genome length (bp)
1	*Amsacta moorei* entomopoxvirus	NC_002520	232392
2	*Cercopithecine* herpesvirus 17	NC_003401	133719
3	*Ectocarpus siliculosus* virus	NC_002687	335593
4	Human herpesvirus 4	NC_001345	172281
5	Human herpesvirus 6	NC_001664	159321
6	Human herpesvirus 6B	NC_000898	162114
7	Lymphocystis disease virus 1	NC_001824	102653
8	Meleagrid herpesvirus 1	NC_002641	159160
9	*Saccharomyces cerevisiae*	NC_004002	149955
10	Sheeppox virus	NC_002642	144575
11	Yaba-like disease virus	NC_001133	230208

The log-log plot of Fourier-transforms (Figure 2.5A and B) for the genomes of *Amsacta moorei* entomopoxvirus (a, b and c), human herpesvirus 4 (d, e and f) and chromosome I of *S. cerevisiae* (g, h and i), for A+G and A+T proportions in a window size of 32, 64 and 128 shows the presence of two distinct regions, a power-law and a flat region in their respective power spectra (to conserve space, graphical presentation of only three organisms has been given).

In comparative analysis using window sizes of 32, 64 and 128 bases, the high frequency end at a frequency of 0.5 (corresponding to an angular frequency π) is unaltered, but the low frequency end has moved towards the high frequency range. This is because, as we increase the window size, the number of points contributing to low frequency is being reduced, but the 1/*f* pattern observed in all is the same. The noise reduces by a factor of $N^{-0.5}$ $(N^{-0.5} = N^{-1/2} = 1/N^{1/2} = 1/\sqrt{N})$, so the signal gets doubled in the window size of 128, when compared to the window size of 32.

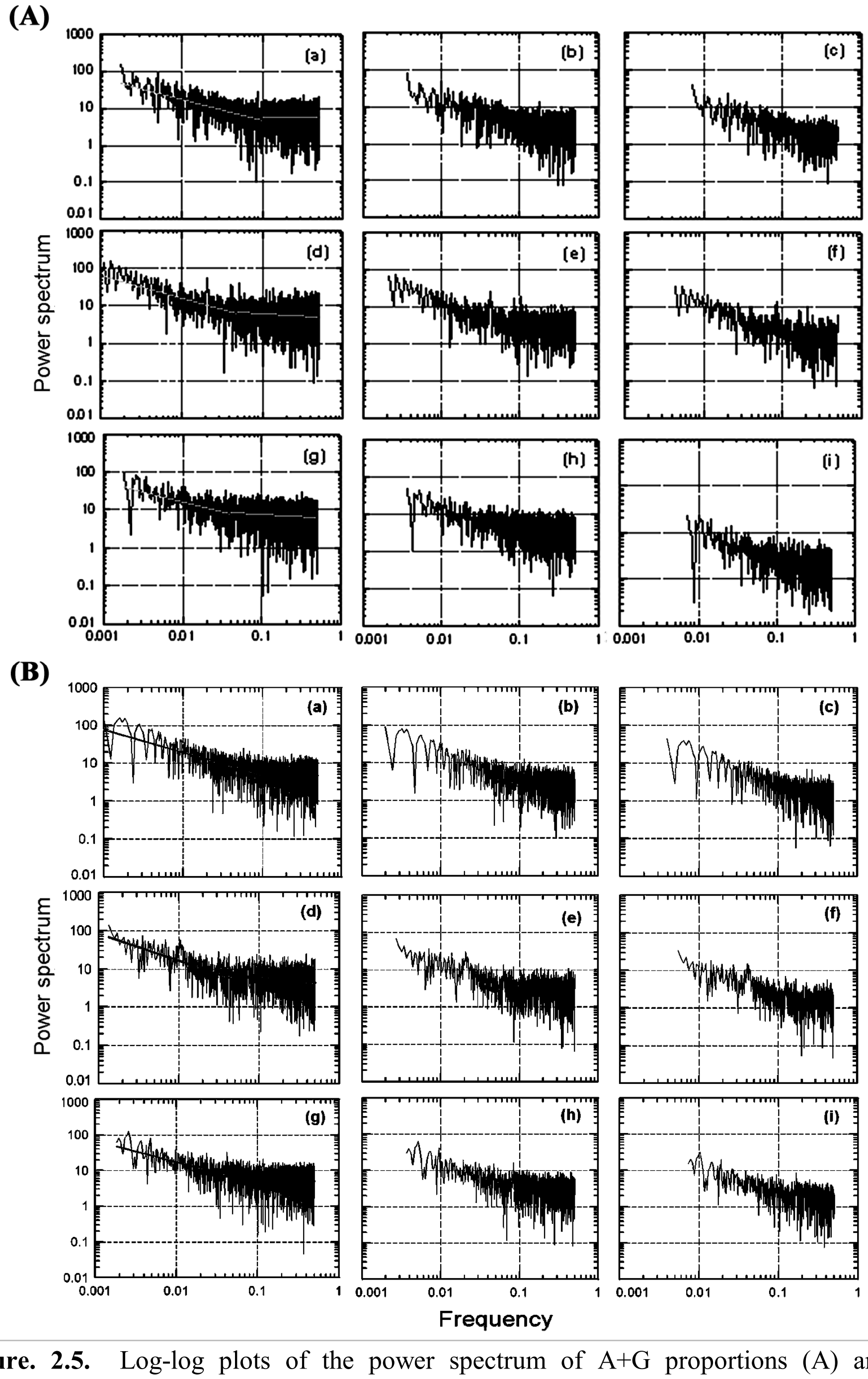


Figure. 2.5. Log-log plots of the power spectrum of A+G proportions (A) and A+T proportions (B) contained in a non-overlapping window of size 32 (a, d and g), 64 (b, e and h) and 128 (c, f, and i) bases for *A. moorei* entomopoxvirus (a, b and c), human herpesvirus 4 (d, e and f) and *S. cerevisia*e chromosome I (g, h and i) with the linear lines of regression for the plots of window size 32 (solid lines in the graph) [For ease of comparison, scales of both axes are set to be same for all the plots].

For the larger window size, intensity of the peaks gets reduced, as the total number of points is reduced, but the pattern observed would remain essentially the same, even though the fine structure is affected. The range of

1/*f* region for the power spectrum of each of the organisms (Table 2.2) is found to be correlated with the gene sizes. The range or the percentage of 1/*f* region for the A+G and A+T proportions of sequence nos 2, 3, 5, 7, 8, 9 and 11 are not the same, which signify that the two proportions are not showing same scaling behavior.

Table 2.2. Range of 1/*f* region of power spectrum of genomes studied

		A+G Proportion		A+T Proportion	
	Organisms	**Range of 1/*f* region (No. of nts)**	**Percentage of 1/*f* region to the total spectrum**	**Range of 1/*f* region (No. of nts)**	**Percentage of 1/*f* region to the total spectrum**
1	*Amsacta moorei* entomopoxvirus	26214	10	26214	10
2	*Cercopithecine* herpesvirus 17	7864	3	13107	5
3	*Ectocarpus siliculosus* virus	10486	2	15729	3
4	Human herpesvirus 4	13107	5	13107	5
5	Human herpesvirus 6	7864	3	13107	5
6	Human herpesvirus 6B	13107	5	13107	5
7	Lymphocystis disease virus 1	7877	6	10499	8
8	Meleagrid herpesvirus 1	7864	3	10486	4
9	*Saccharomyces cerevisiae*	10486	4	7864	3
10	Sheeppox virus	10486	4	10486	4
11	Yaba-like disease virus	10486	4	13107	5

Linear regression analysis was done separately for the two distinct regions ("1/*f* noise" and "white noise") and the significant slope values (> -0.5) were obtained (Table 2.3), suggesting the presence of long-range correlations in their respective genomes. Linear lines of regression for *A. moorei* entomopoxvirus [(Figure 2.5A(a) and 2.5B(a)], human herpesvirus 4 [(Figure 2.5A(d) and 2.5B(d)] and *S. cerevisiae* chromosome I (Figure 2.5A(g) and 2.5B(g)) are also shown.

The slope values of the 1/*f* region for A+G power spectrum ranged from -0.53 to -0.88 and for A+T spectrum from -0.53 to -0.96. Our studies have shown that the behavior of the power spectrum in the low and intermediate regions represented power-law decay, with the slope values close to -1, indicating the presence of long-range correlations in those regions. We can say that greater the exponent, larger is the correlation. Different slopes (differences > 0.1) for (A+G)/(A+T) proportions, are shown in the 1/*f* region for sequence nos 2, 3, 5, 8, 9 and 11 as given in Table 2.3. This observation signifies that the two proportions are not showing same scaling behavior for long-range correlations. Slope is close to -1, for sequence no. 9 for A+T proportion and also for sequence no. 2 for A+G proportions. For Sheeppox virus, power spectrum of A+T proportion shows high correlation, but for *Cercopithecine* herpesvirus 17, A+G proportion is showing high correlation. Power law with a negative value of exponent suggests that a system producing 1/*f* type noise is scale invariant and has long-range time and spatial correlations as a consequence.

Table 2.3. Slopes values (-α) obtained by linear regression analysis of power spectrum of genomes studied

	Organisms	**Slope values of 1/*f* noise**		**Slope values of flat noise**	
		A+G proportions	**A+T proportions**	**A+G proportions**	**A+T proportions**
1	*Amsacta moorei* entomopoxvirus	-0.56	-0.60	-0.01	-0.02
2	*Cercopithecine* herpesvirus 17	-0.88	-0.67	-0.10	-0.12
3	*Ectocarpus siliculosus* virus	-0.70	-0.58	-0.11	-0.13
4	Human herpesvirus 4	-0.62	-0.70	-0.09	-0.18
5	Human herpesvirus 6	-0.67	-0.55	-0.11	-0.11
6	Human herpesvirus 6B	-0.54	-0.53	-0.07	-0.11
7	Lymphocystis disease virus 1	-0.69	-0.63	-0.02	-0.07
8	Meleagrid herpesvirus 1	-0.79	-0.62	-0.05	-0.07
9	*Saccharomyces cerevisiae*	-0.67	-0.96	-0.03	-0.14
10	Sheeppox virus	-0.64	-0.76	-0.02	-0.10
11	Yaba-like disease virus	-0.53	-0.59	-0.13	-0.10

The average gene lengths of each of the organisms considered for the study were obtained from the same database (Table 2.4). We note a correlation between the distribution of the gene length and the domain of “1/*f* region” of the power spectrum. We also note (Table 2.4), that average gene lengths of *A. moorei* entomopoxvirus, human herpesvirus 4 and *S. cerevisiae* are in the increasing order, which is inversely proportional to the end of 1/*f* region (as shown in Figure 2.5A) of power spectra. The 1/*f* region (Figure 2.5A and B) of *A. moorei* entomopoxvirus extends up to 0.1 of angular frequency, signifying the presence of mid to long-range correlations for genes of smaller gene length. Whereas for human herpesvirus 4 and *S. cerevisiae,* the 1/*f* region extends only up to 0.05, which signify the presence of long-range correlations for genes of longer gene length (Figure 2.5A).

Table 2.4. Average gene length of the studied organisms

	Organisms	Average gene length (bp)
1	*Amsacta moorei* entomopoxvirus	800
2	*Cercopithecine* herpesvirus 17	1500
3	*Ectocarpus siliculosus* virus	1000
4	Human herpesvirus 4	1500
5	Human herpesvirus 6	1300
6	Human herpesvirus 6B	2900
7	Lymphocystis disease virus 1	900
8	Meleagrid herpesvirus 1	1700
9	*Saccharomyces cerevisiae*	1100
10	Sheeppox virus	1100
11	Yaba-like disease virus	1600

Histograms were plotted (Figure 2.6) to show the absolute frequencies of the average gene length of all the genomes studied, in a semi-log scale. We note that the genes of the organisms analyzed are not all of the same size (Figure 2.6). The long genes are few in number and smaller ones are more in number, as is apparent from the histograms (Figure 2.6). Also, the more number of small genes in the DNA are expected to increase mid-range correlation length. This shows that genes are having correlations within themselves; however, correlations between the genes are not ruled out as the 1/*f* region covers a scale greater than the longest gene. We also observed that the heights of the vertical bars in the histograms are approximately in exponential decrease; however, the rate of decrease is not the same in all the genomes. To emphasize this fact, we have plotted the graphs in a semi-log scale (Figure 2.6).

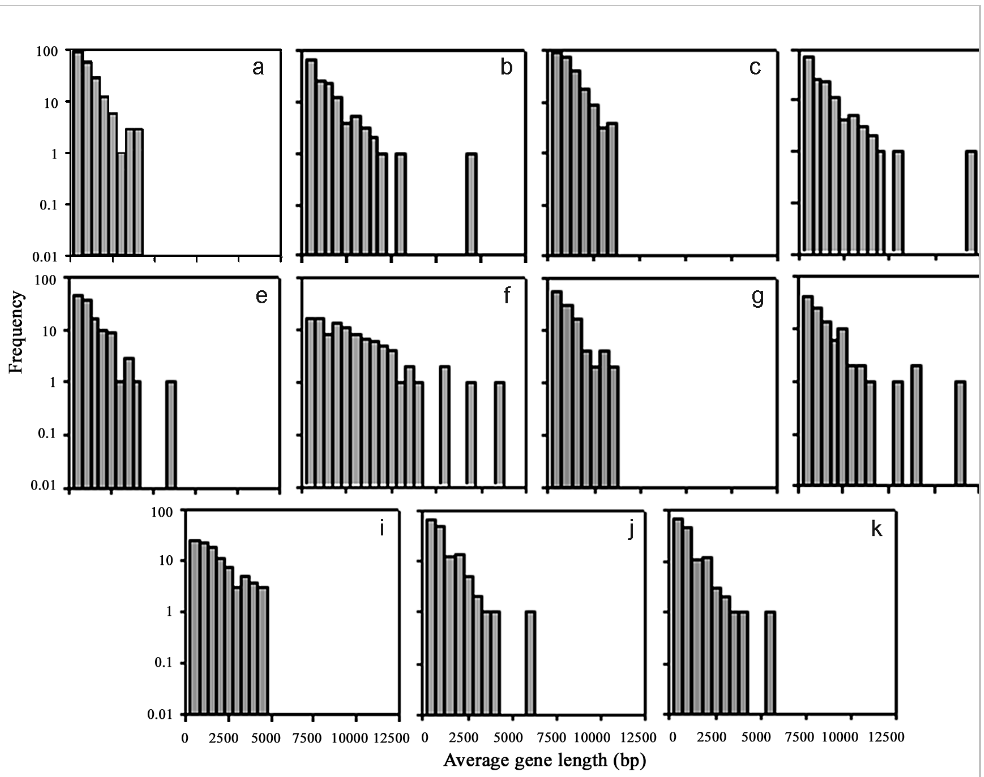


Figure. 2.6. Graphs showing the absolute frequencies of average gene length of the genomes (in alphabetical order from a - k as given in the order for all the tables) in a semi-log scale [Scales are identical to facilitate visual comparison].

2.4 Conclusions

The genomes of the organisms analyzed represent slope values significant in specifying the presence of long-range correlations in their genome organization. It is proposed that the long-range correlations represent an efficient trade-off between efficient information storage and protection against error in genetic code. We conclude that genes are having correlations within themselves and the correlation lengths do not span the complete genome but certainly cross the gene boundaries. On a short length scale, the genes appear noisy, because of lack of short-range correlation. It was also noted that the high frequency end of the 1/*f* region is correlated with the distribution of the gene sizes, and the correlation appears to be related with the scaling exponent α.

As all the DNA sequences studied in this report lack introns (except *S. cerevisiae*) therefore, only very short intergenic regions are expected, which may also contribute to the white noise part of the spectrum. However, *S. cerevisiae* has introns, but this is not apparent in the spectrum either. So, from our studies, we can say that the DNA of the viruses are showing1/*f* noise in their power spectrum, irrespective of the absence of introns in their genomes. Thus, we can assume that 1/*f* noise is not a feature of intron containing regions, but is also exhibited by coding regions of the genome. The occurrence of long-range correlations in exonic sequences might signify the existence of correlation structures or patterns in the gene structure.

2.5 References

- Afonso, C. L., Tulman, E. R., Lu, Z., Zsak, L., Rock, D. L. and Kutish, G. F. (2001). The genome of turkey herpesvirus. *J Virol.* 75: 971–978.
- Arif, B. M. (1995). Recent advances in the molecular biology of entomopoxviruses. *Journal of General Virology.* 76: 1–13.
- Bawden A. L., Glassberg, K. J., Diggans, J., Shaw, R., Farmerie, W. and Moyer R. W. (2000). Complete genomic sequence of the *Amsacta moorei* entomopoxvirus: analysis and comparison with other poxviruses. *Virology*. 274(1): 120–39.
- Delaroque, N., Müller, D.G., Bothe, G., Pohl, T., Knippers, R. and Boland, W. (2001). The Complete DNA Sequence of the *Ectocarpus siliculosus* Virus EsV-1 Genome. *Virology.* 287(1): 112–132.
- Epstein, M. A., Achong, B. G. and Barr, Y. M. (1964). Virus particles in cultured lymphblasts from Burkitt's Lymphoma. *Lancet.* 1: 702–703.
- Fauqet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U. and Ball, L.A. (editors) (2005). *Virus Taxonomy, VIIIth Report of the ICTV*. London: Elsevier.
- Fischer, M., Schnitzler, P., Delius, H. and Darai, G. (1988). Identification and characterization of the repetitive DNA element in the genome of insect iridescent virus type 6. *Virology.* 167:485–496.
- Fukushima, A., Kinouchi, M., Kudo, Y., Kanaya, S., Mori, H. and Ikemura T. (2001). Statistical Analysis of Genomic Information: Various Periodicities in DNA Sequence. *Genome Informatics.* 12: 435–436.
- Fukushima, A., Ikemura, T., Kinouchi, M., Oshima, T., Kudo, Y., Mori, H. and Kanaya, S. (2002). Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. *Gene.* 300: 203–211.

- Gompels, U. A., Nicholas, J., Lawrence, G., Jones, M., Thomson, B.J., Martin, M. E., Efstathiou, S., Craxton, M. and Macaulay, H.A. (1995). The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. *Virology.* 209: 29–51.
- Goodwin, R. H., Adams, J. R. and Shapiro, M. (1990). Replication of the entomopoxvirus from *Amsacta moorei* in serum-free cultures of a Gipsy moth cell line. *Journal of Invertebrate Pathology.* 56: 190–205.
- Hall, C. B., Long, C. E., Schnabel, K. C., Caserta, M. T., McIntyre, K. M., Costanzo, M. A., Knott, A., Dewhurst, S., Insel, R. A. and Epstein, L. G. (1994). Human Herpesvirus-6 Infection in Children - A Prospective Study of Complications and Reactivation,. *N Engl J Med.* 331(7): 432–438.
- Herzel, H., Weiss, O. and Trifonov, E. N. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics.* 15: 187–193.
- http://www.ncbi.nlm.nih.gov.
- Kawamura, H., King, D. J. and Anderson, D. P. (1969). A herpesvirus isolated from kidney cell culture of normal turkeys. *Avian Dis.* 13: 853–863.
- Kingham, B. F., Zelnik, V., Kopacek, J., Majerciak, V., Ney, E. and Schmidt, C. J. (2001). The genome of herpesvirus of turkeys: comparative analysis with Marek's disease viruses. *J Gen Virol.* 82: 1123–1135.
- Knight, J. C., Novembre, F. J., Brown, D. R., Goldsmith, C. S. and Esposito, J. J. (1989). Studies on tanapox virus. *Virology.* 172: 116–124.
- Lerner, A. M., Beqaj, S. H., Deeter, R. G. and Fitzgerald, J.T. (2004). IgM serum antibodies to Epstein-Barr virus are uniquely present in a subset of patients with the chronic fatigue syndrome. *In Vivo*. 18(2): 101–106.

- Li, W. and Kaneko, K. (1992). DNA correlations. *Nature.* 360: 635–636.
- Li, W. and Kaneko, K. (1992). Long-Range Correlation and Partial $1/f^{\alpha}$ Spectrum in a Noncoding DNA Sequence. *Europhys Lett.* 17: 655–660.
- Lünemann, J. D. and Münz, C. (2007). Epstein-Barr virus and multiple sclerosis. *Current neurology and neuroscience reports*. 7(3): 253–258.
- Marlow, S. A., Billam, L. J., Palmer, C. P. and King, L. A. (1993). Replication and morphogenesis of *Amsacta moorei* entomopoxvirus in cultured cells of *Estigmene acrea* (salt marsh caterpillar). *Journal of General Virology*. 74: 1457–1461.
- Nee, S. (1992). Uncorrelated DNA walks. *Nature.* 357: 450.
- Paperna, I., Sabnai I. and Colorni, A. (1982). An outbreak of lymphocystis in *Sparus aurata* L. in the Gulf of Aqaba, Red Sea. *J. Fish Dis.* 5: 433–437.
- Press, W. H. *et al.* (1992). *Numerical Recipes in C,* 12, pp. 496-536 and pp. 537–608, Cambridge University Press.
- Samalecos, C. P. (1986). Analysis of the structure of fish lymphocystis disease virions from skin tumours of *Pleuronectes*. *Arch. Virol.* 91: 1–10.
- Surya Pavan, Y. and Mitra, C. K. (2005). Fractal studies on the protein secondary structure elements. *Indian J Biochem Biophys.* 42: 141–144.
- Tidona, C. A. and Darai, G. (1997). The complete DNA sequence of lymphocystis disease virus. *Virology.* 230: 207–216.
- Upadyay, R. K. (2003). Computation analysis of evolutionary divergence of scorpion toxins. *Indian J Biochem Biophys.* 40: 51–58.
- Vago, C. (1963). A new type of insect virus. *Journal of Invertebrate Pathology.* 5: 275–276.
- Vieira, M. S. (1999). Statistics of DNA sequences: A low-frequency analysis *Phys Rev E.* 60(5): 5932–5937.

- *Wickner, R. B., Tang, J., Gardner, N. A and Johnson, J. E. (2008).* "The Yeast dsRNA Virus L-A Resembles Mammalian dsRNA Virus Cores", Segmented Double-stranded RNA Viruses: Structure and Molecular Biology. *Caister Academic Press.* ISBN 978-1-904455-21-9.
- Winter, J., Hall, R. L. and Moyer, R. W. (1995). The effect of inhibitors on the growth of the entomopoxvirus from *Amsacta moorei* in *Lymantria dispar* (Gypsy moth) cells. *Virology*. 211: 462–473.
- Witter, R. L., Nazerian, K., Purchase, H. G. and Burgoyne, G. H. (1970). Isolation from turkeys of a cell-associated herpesvirus antigenically related to Marek's disease virus. *Am J Vet Res.* 31: 525–538.
- Wolf, K. (1988). Fish virus and fish viral diseases, p. 268–291. Cornell University Press, Ithaca, N.Y.

Chapter 3

Comparative Analysis of the Splice Site Regions by Information Theory

3.1 Introduction

3.1.1 RNA splicing

Eukaryotes undergo the process of “RNA splicing”, which involves the splicing of introns from heterogenous RNA (hnRNA or pre-mRNA) to form mature mRNA. 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. In addition to these dimers, a pyrimidine-rich region precedes AG at the acceptor site, and a short consensus follows GT at the donor site (Lewin, 2000).

The branch site lies 20-50 nts upstream of the 3’ splice site and provides the means by which the 3’ splice site is identified. It is highly conserved and is generally characterized by the occurrence of the consensus sequence (CUPuAPy) in the intronic region. The role of the branch site is to identify the nearest 3’ splice site as the target for connection to the 5’ splice site (Figure 3.1).

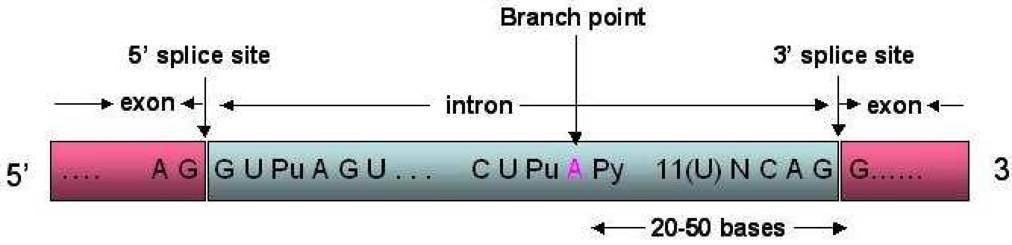


Figure 3.1. Pre-mRNA (hnRNA) showing the two splice site regions — donor (5’- boundary containing the dinucleotide GU) and acceptor (3’- boundary containing the dinucleotide AG). The two boundaries are characterized by two different consensus sequences. The intronic (in blue color) and exonic (pink color) regions are shown in different colors. The branch point is also characterized by a typical consensus sequence (CUPuAPy) located 20-50 nts from the acceptor site. A poly-pyrimidine tract is also present in the intronic region upstream of the acceptor site.

A complex of nucleotide binding proteins and small nuclear RNAs (snRNAs), collectively known as the “spliceosome”, recognizes these splice sites and excises introns by a concerted transesterification reaction. One important consequence of RNA splicing is that one gene can produce several different mRNA variants, or isoforms, simply by joining together different combinations of exons.

3.1.2 Earlier studies on splice site regions

Several earlier studies have been reported for the detection of splice sites using different methods; such as the weight matrix model that uses the position compositional biases in splice sites (Staden, 1984). Artificial neural networks have been applied for the prediction of splice sites in different organisms with confidence levels better than previous methods (Brunak *et al*., 1991). However, the reported results should be interpreted with caution as they were based on small datasets of limited number. A computational tool, GeneSplicer (Pertea *et al*., 2001), was developed based on maximum dependence decomposition and performed better than previous tools. Recently the prediction of splice sites with dependency graphs and their expanded Bayesian networks has gained much importance because of its better performance (Chen *et al*., 2005). Current studies are being carried out to further understand and interpret the information contained in splice sites, as well as to develop a better method for their prediction with better specificity and sensitivity.

3.1.3 Problems involved in the study

Detection of splice sites by using the two dinucleotides (GT/AG) is not meaningful because the frequency of these dinucleotides is very high in genes. Another important aspect to be considered is that the bases flanking them are also involved in the process and are expected to contain information required for splicing. Studying the consensus is also not directly useful, as they are highly variable not only within the species but also between species. Therefore, information theory comes to play a major role for the study of splice sites, which gives a quantitative measure of sequence conservation (or variability).

3.1.4 Shannon's information theory

Information theory is an important tool (Shannon, 1948) that has been often applied for understanding several key concepts in molecular biology (Adami, 2004). Information is defined as the amount of correlation between two random variables (*X* and *Y*), which is measured as the amount of entropy (uncertainty in a random variable) shared by them. This shared entropy is the information that one random variable contains about the other. It is a relative entity and is never absolute. In other words, mutual information is defined as a measure of the amount of information that one random variable contains about the other. It measures exactly the amount by which the entropy of *X* or *Y* is reduced by knowing the other, *Y* or *X* (Durbin *et al*., 1998). This theory has gained much importance in biology by its applications to measure the information content of the nucleotide binding sites (Schneider *et al*., 1986), identification of polymorphisms in DNA (Rogan and Schneider, 1995), prediction of RNA and protein secondary structures (Giraud *et al*., 1998), prediction and analysis of molecular interactions (Adami and Thomson 2005), and drug design (Adami, 2002).

Study of horizontal correlations (between nucleotides along a sequence) is useful to identify features that can distinguish coding and non-coding regions in DNA (Rekha and Mitra 2006). This gives the probability of finding nucleotides in the sequence that are correlated with each other. On the other hand, vertical correlations are important to find the probability of a nucleotide at a particular site by calculating the information content of the aligned set of sequences from its frequency of occurrence. Substitution matrices are thus useful to score these alignments perfectly.

Information theory has also been used for studying the features of spliceosome evolution and function (Stephens and Schneider, 1992). Studies have been carried out to correlate the intron length and the information content of the splice sites (Fields, 1990), suggesting that longer introns contain more information than shorter ones (Mount *et al*., 1992). Recently a comprehensive splice-site analysis using comparative genomics has been performed on different organisms by using the information content of the splice-site motifs, which proves that the identification of broad patterns in naturally-occurring splice sites, through the analysis of genomic datasets, provides mechanistic and evolutionary insights into pre-mRNA splicing (Sheth *et al*., 2006).

3.1.5 Role of substitution matrices

Substitution matrix is a useful tool that scores the similarity between any two nucleic acid bases in terms of their ability to replace each other. By comparing a large number of similar sequences, one can obtain a matrix that describes the probability of a given nucleotide being substituted by another under the conditions of study. As probabilities are multiplicative, the logarithm is used to get an additive formulation. A number of techniques are now available for direct computations of substitution matrices, such as the BLOSUM (blocks substitution matrix) (Henikoff and Henikoff, 1992) and PAM (point accepted mutation) matrices (Dayhoff *et al*., 1978). These matrices have been used extensively for global and local sequence alignments as well as database searches (Altschul, 1991). They were also found to be significant for the study of core promoter regions (Reddy *et al*., 2006).

3.1.6 Motivation for the study

RNA splicing is an important process in eukaryotes, which involves the splicing of introns from pre-mRNA by the spliceosomal proteins that recognize the intron-exon junctions. It is important to understand as to why only the functional splice sites are recognized by the spliceosomal proteins, because similar kind of consensus can be present at non-functional splice sites. This gives us an idea that the functional splice sites do contain the information that is required for the process of splicing to take place efficiently. It is important to characterize signals that govern the process of splicing in different organisms by information theory, which gives a broad idea about the distribution of information around the splice sites in different organisms.

We have studied this aspect by carrying out a comparative analysis of donor and acceptor splice site regions in the gene sequences of five different organisms. We have constructed substitution matrices for the aligned set of sequences in the blocks of 6, 10, and 14 nts around the consensus dinucleotides (gt/ag) and calculated their information content, respectively. The substitution matrix specifically constructed for a given block is expected to work more efficiently than the one constructed for the whole genome sequences. In fact, we expect the difference to be evident among the three block databases. We have performed a broad analysis of the data distribution by calculating the information content at/around the splice sites, and achieved some interesting and informative results.

3.2 Methodology

3.2.1 Exon-Intron Database (EID)

The study of the intron-exon organization of the eukaryotic genes has been facilitated by the presence of enormous amount of the data in the GenBank. The EID database, which contains protein-coding intron-containing gene sequences, has been developed from the eukaryotic subset of GenBank (release 112). This database provides a means to explore the molecular evolution of the "intron-exon architecture" of the eukaryotic genes. Besides providing a well-organized, extensive and experimental data set for studying the features of the introns and exons, it can also be used to improve the accuracy of the gene prediction programs, which aim at the recognition of the intron/exon boundaries in the genome.

(i) Representation of EID

The database contains the gene sequences of different organisms along with their alternative isoforms (Saxonov *et al.*, 2000). The EID (**http://hsc.utoledo.edu/bioinfo/eid/index.html**) released in September 2005 was downloaded for the study. It provides a flat-file distribution of the data for the large-scale analysis of the features of intron-exon organization. For each gene, the database provides eight different files, which contains the DNA sequences, protein sequences, mRNA sequences, exon sequences, intron sequences, header of the sequences, statistics of the sequences and the extensive description of the sequences obtained from the GenBank. Except for the statistics and the description file, all databases were constructed in the FASTA format (Figure 3.2).

We have used the DNA database (containing the DNA sequences), which was also built in the FASTA format, such that each gene has its own entry and the first line of the entry starts with a '>' sign and contains information about the sequence. The information contains a description of the following identifiers: a unique EID index, the GenBank locus, the GenBank protein identifier (protein_id) and a short description extracted from the DEFINITION line of the GenBank entry. In addition the first line also provides the following sequence-specific information: intron phases (positions of introns within codons-could be 0, 1 or 2), intron lengths, exon lengths, the total exon length and the total intron length. If the intron size is unknown for some reasons, the letter 'u' is used instead of the actual intron size. And if the length of any one of the introns is unknown, then the total length is automatically set to –1, denoting that the total length is unknown. The last field of the line contains the splice motif information (for each intron, a four-letter string is provided that is composed of the first two and the last two nucleotides (nts) of the intron).

The DNA database contains splice sites with "gt…ag" exon-intron boundaries (motifs), which accounts to 98% of all the known motifs. But other motifs like "at…ac" and "gc…ag" were also found in the database, which account to a very small percentage compared to "gt…ag". We have considered all gene sequences with only "gt…ag" motifs in all our analysis and excluded sequences containing other motifs and cryptic splice sites. The EID database contained gene sequences being represented along with their alternative isoforms, so we have considered all gene sequences with and without their isoforms. The exon sequences are represented as uppercase letters, and the intron sequences along with the splice site dinucleotides are given as lowercase letters.

```
>1_NT_033778 protein_id:45552443;  Drosophila melanogaster chromosome 2R, complete
sequence. /gene="CG33492"; intron(phase:1,size:224,intr_sum:224);
exon(size:613,392,ex_sum:1005); {splice:gtag}; CDS_start=25, CDS_end=1229, CDS_len=981
GATTGGGGCAAAGTTTATCCAAATATGTCTGGAGATGGTGCTCTTGGTATGCTTATTAATCGTAA
AGCAGATATATGCATTGGAGCTATGTACTCGTGGTACGAAGATTACACATACTTAGACC…..

>4A_NT_033778 protein_id:45550362;  Drosophila melanogaster chromosome 2R, complete
sequence. /gene="CG2944"; intron(phase:21,size:6573,2960,intr_sum:9533);
exon(size:87,890,593,ex_sum:1570); {splice:gtag,gtag}; CDS_start=86, CDS_end=10650,
CDS_len=1032
GTTTTGTGCATCAAAGGTTCCTGAGTGATATTTTGCTTTATGTGTTGTTTTATTTTAATCATAGCAA
ATCGACAAGATAAAAGTAATgtaagtttcacaggaactgataaatgaaattcacacatatctaacgtgggaattgatcaatt…..
.
.

>4B_NT_033778 protein_id:24585809;  Drosophila melanogaster chromosome 2R, complete
sequence. /gene="CG2944"; intron(phase:21,size:330,2960,intr_sum:3290);
exon(size:268,890,593,ex_sum:1751); {splice:gtag,gtag}; CDS_start=249, CDS_end=4588,
CDS_len=1050
GTTTTGTTGGCTCGCTTTTGTAATACCGTTTGCAAGCCTACTTATATTGTGATCGGACAACCCGT
GTGTTATCTTAAATAACAAATTTTTTATTTCGTTGAAAATAGTTGTTTGTTTCGTTTATATTTTAAAT
ACTTTCTACATATTTCTGTGCATAATGAACATAACGATTTCACAACTAAGCAAGAAAAGAAATTAA
CTTCACTCGGTGACGCCATTTTGTCCAATTTTTATATACTGCTTTATAAATGGAATTTGGTAAAAA
AAGgtaaattttgttcaagtttatggccaatttaagaattaaatgctccacatactttttgtgttttgacaaaagtttaaaatatgtaaatatgtacaa
aatgttttataatattaaccagtttatctgtattattgtttaaaggtaatgtaaagtacatggaatgtagaagtgtcttcaacgcgaatgcacatatgc
atatacatgtctatgtatgtatgtatatgtatatttatatgtttttatatgtacgtatgtatgctcttttatgtgtatacataacgtgtttcatttcaaggtaaca
caaaatcgttaattttctttcttttaagGTATAAGGGTGGTCGGACACCTTCTATTAGGTCCAGTGAAAG…..
```

Figure 3.2. Representation of gene sequences in the EID database (built in the FASTA format). The first line of the entry of each gene starts with a '>' sign and contains information about each sequence that includes: a unique EID index along with the protein identifier (protein_id) and a short description extracted from the DEFINITION line of the GenBank entry. It also provides the sequence-specific information, such as: intron phases (positions of introns within codons-could be 0, 1 or 2) along with the intron lengths, exon lengths, the total exon length and the total intron length. Information about the exon size and exon sum is also provided. The last field of the line contains the splice motif information (for each intron, a four-letter string is provided that is composed of the first two and the last two nts of the intron (gtag)). This is followed by the coding sequences (CDS) start, end and length. The actual DNA sequence starts from the next line, which contains intron containing protein coding gene sequences. The DNA database used for the analysis contains splice sites with "gt…ag" exon-intron boundaries (motifs) only. The exon sequences are represented in the uppercase and the intron sequences along with their splice site motifs are represented in the lowercase letters.

3.2.2 Organisms considered in the present study

We have selected five different organisms in our analysis in order to have a broad distribution of the data from plants to mammals, which include *Arabidopsis thaliana* (plant), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (arthropod), *Gallus gallus* (aves) and *Rattus norvegicus* (mammal). Our objective of selecting the given organisms was to have a broad distribution of the data from plants to mammals, otherwise the selection can be considered arbitrary. Therefore the present study can be considered as "typical" or "representative" with a reasonably broad representation. The detailed features of the genome of the given organisms are given below.

(i) A. thaliana:

Scientific classification

Kingdom:	Plantae
Division:	Magnoliophyta
Class:	Magnoliopsida
Order:	Brassicales
Family:	Brassicaceae
Genus:	*Arabidopsis*
Species:	***A. thaliana***

A. thaliana is a species of *Arabidopsis*. It is an annual (rarely biennial) plant growing to 5–30 cm (rarely to 50 cm) tall, belonging to the family Brassicaceae. It is widely used as one of the model organisms for studying plant sciences, including genetics and plant development. It plays the same role for agricultural sciences as that of mice and fruit flies (*Drosophila*) play in human biology. The small size of its genome made it useful for genetic mapping and sequencing. It has about 157 mbps and five chromosomes, making it a small genome for a plant species and it was the first plant genome to be sequenced, in 2000. Much work has been done to assign functions to its 27,000 genes and the 35,000 proteins they have encoded (Lolle *et al*., 2005).

(ii) C. elegans

Scientific classification

Kingdom:	Animalia
Phylum:	Nematoda
Class:	Secernentea
Order:	Rhabditida
Family:	Rhabditidae
Genus:	*Caenorhabditis*
Species:	***C. elegans***

C. elegans is a free-living nematode (roundworm), about 1mm in length, which lives in temperate soil environments. Research into the molecular and developmental biology of *C. elegans* has begun in 1974 by Sydney Brenner and it has since been used extensively as a model organism (*Brenner,1974)*. *C. elegans* has five pairs of autosomes and one pair of sex chromosomes. It is studied as a model organism for a variety of reasons. From a research perspective, it has the advantage of being a multicellular eukaryotic organism which is simple enough to be studied in great detail. It was also the first multicellular organism to have its genome completely sequenced. The genome of *C. elegans* has ~ 100 mbps and contains ~20,000 genes. The vast majority of these genes encode for proteins but there are likely to be as many as 1,000 RNA genes.

(iii) D. melanogaster

Scientific classification

Kingdom:	Animalia
Phylum:	Arthropoda
Class:	Insecta
Order:	Diptera
Family:	Drosophilidae
Genus:	*Drosophila*
Species:	***D. melanogaster***

D. melanogaster is a two-winged insect that belongs to the order, Diptera of the flies. The species is commonly known as the fruit fly, and is one of the most commonly used model organisms in biology (including studies in genetics, physiology and life history evolution). The genome of *D. melanogaster* (sequenced in 2000, and curated at the FlyBase database) contains four pairs of chromosomes: an X/Y pair, and three autosomes labeled 2, 3, and 4. Its sequenced genome of 120 mbps has been annotated and contains ~13,767 protein-coding genes which comprise ~20% of the genome. More than 60% of the genome appears to be functional non-protein-coding DNA involved in gene expression control. About 75% of known human disease genes have a recognizable match in the genetic code of fruit flies, and 50% of fly protein sequences have mammalian analogues. It is being used as a genetic model for several human diseases including the neurodegenerative disorders and also to study mechanisms underlying immunity, diabetes, and cancer, as well as drug abuse (Ashburner *et al.*,2005*)*.

(iv) G. gallus

Scientific classification

Kingdom:	Animalia
Phylum:	Chordata
Class:	Aves
Order:	Galliformes
Family:	Phasianidae
Genus:	*Gallus*
Species:	***G. gallus***

The Red Junglefowl, *G. gallus*, is a tropical member of the Pheasant family and the direct ancestor of the domestic chicken. The chicken (*G. gallus*) is an important model organism for biomedical research, development, and aging, in addition to its importance in agriculture. The chicken genome is the first avian genome to be sequenced, and it has a haploid genome size of 1200 mbps. The chicken genome, similar to other avian genomes, is composed of chromosomes of vastly different sizes, which are identified as either macro- or microchromosomes. The *G. gallus* genome has 38 pairs of autosomes and a pair of sex chromosomes, referred to as Z and W to distinguish them from mammalian sex chromosomes.

(v) R. norvegicus

Scientific classification

Kingdom:	Animalia
Phylum:	Chordata
Class:	Mammalia
Order:	Rodentia
Family:	Muridae
Genus:	*Rattus*
Species:	***R. norvegicus***

The brown rat, common rat, Norway rat, Norwegian rat or wharf rat (*R. norvegicus*) is one of the best-known and common rats, and also one of the largest. Over the years, rats have been used in many experimental studies, which have added to our understanding of genetics, diseases, the effects of drugs, and other topics that have provided a great benefit for the health and wellbeing of humankind. The brown rat is also an important model organism for human physiology and diseases including cardiovascular, diabetes, addiction, arthritis, neurological, and more. It has 21 pairs of chromosomes (Gadaleta *et al.*, 1989).

The DNA database of EID contains the gene sequences of the complete genomes of the given organisms that are considered for the study. Table 3.1 gives the total number of genes and the number of splice sites that are considered for our study. As already mentioned, we have considered only the gt-ag splice sites in our analysis.

Table 3.1 The Number of Genes and Splice Sites of the Five Organisms Studied*

No	Organism	No. of genes	Total no. of genes#	No. of splice sites		Exon/intron boundaries
				Donor	Acceptor	
1	*A. thaliana*	20,716	22,957	130,099	131,229	gt-ag
2	*C. elegans*	18,594	20,470	111,970	112,361	gt-ag
3	*D. melanogaster*	10,612	15,624	72,737	73,167	gt-ag
4	*G. gallus*	16,567	16,568	168,120	169,990	gt-ag
5	*R. norvegicus*	19,146	19,197	181,782	183,476	gt-ag

*The splice sites with only "gt-ag" exon/intron boundaries were considered in our analysis. All other splice sites such as "gc-ag", "at-ac", and all the cryptic ones were excluded in the study. However, we have included all the alternative splice sites in our analysis. #The total number of genes including alternative isoforms.

3.2.3 Construction of the blocks Database (Datasets)

The DNA database containing the gene sequences of the given organisms were considered for the construction of the blocks database. Since the main emphasis of our study is to analyze the sub-sequences at the splice sites, we have to develop our own datasets, such that they can be utilized for further study. We have developed three different databases for the donor (-gt-) and acceptor (-ag-) splice site regions respectively by aligning two, four and six bases flanking on either side of the dinucleotides (-gt- and –ag-) for all the organisms being studied. Consequently, we have constructed three blocks of six (gt±2, ag±2), ten (gt±4, ag±4) and fourteen (gt±6, ag±6) nucleotides for each of the donor and acceptor splice site regions for all the given organisms as illustrated in figure 3.3.

We have used three different block sizes in order to have a comparative analysis of the conservation of bases at the splice sites, which are involved in the process of splicing. This is a better approach when compared to the earlier studies, which gives a good understanding of the distribution of information around the splice sites in the given organisms. Scanning the nucleotides one by one with entropy would have been computationally expensive and the information obtained might have been disproportionately low. The blocks obtained were then used for the computations of the substitution matrix.

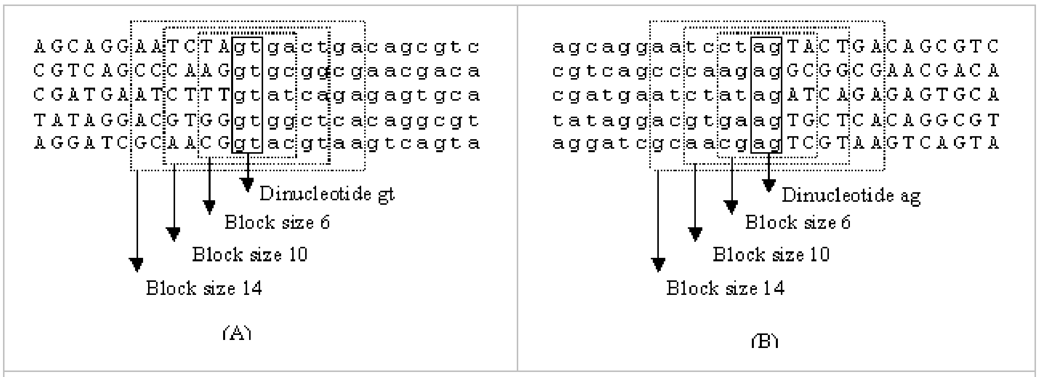

AGCAGGAATCTAgtgactgacagcgtc
CGTCAGCCCAAGgtgcggcgaacgaca
CGATGAATCTTTgtatcagagagtgca
TATAGGACGTGGgtggctcacaggcgt
AGGATCGCAACGgtacgtaagtcagta
Dinucleotide gt
Block size 6
Block size 10
Block size 14
(A)
agcaggaatcctagTACTGACAGCGTC
cgtcagcccaagagGCGGCGAACGACA
cgatgaatctatagATCAGAGAGTGCA
tataggacgtgaagTGCTCACAGGCGT
aggatcgcaacgagTCGTAAGTCAGTA
Dinucleotide ag
Block size 6
Block size 10
Block size 14
(B)

Figure 3.3. Illustrations of the construction of three different block databases for donor (A) and acceptor (B) splice sites. The splice sites are represented as donor (gt) and acceptor (ag) sites and the central dinucleotides (gt/ag) are aligned with 2, 4, or 6 nt taken on both sides. The three blocks are constructed for 6 (gt±2, ag±2), 10 (gt±4, ag±4), and 14 (gt±6, ag±6) nt, respectively. Note that the given sequences are for illustration only and are arbitrary. The exon sequences are represented as uppercase letters, and the intron sequences along with the splice site dinucleotides are given as lowercase letters. The regions enclosed within the boxes are used for the computations of the substitution matrices.

3.2.4 Development of Substitution matrix

We have constructed substitution matrices for the aligned set of sequences of the given block sizes to calculate their mononucleotide substitutions (Henikoff and Henikoff, 1992). For the construction of each substitution matrix, we have counted the number of matches and mismatches of each nucleotide type in each column between the first sequence and every other sequence present in the database. The same procedure was followed for every sequence in the database for all the columns present (Figure 3.4), and the values obtained were stored in a 4x4 frequency table, which gives the number of possible pairs of nucleotides in the database.

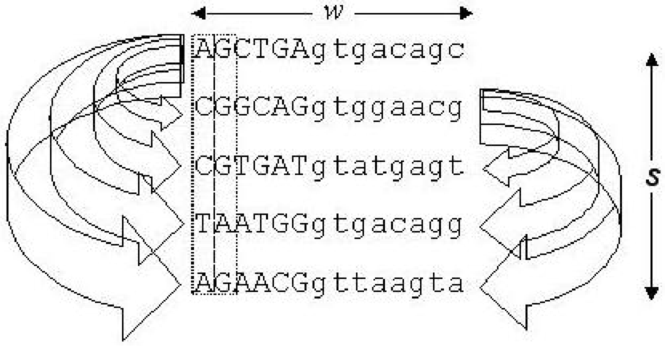
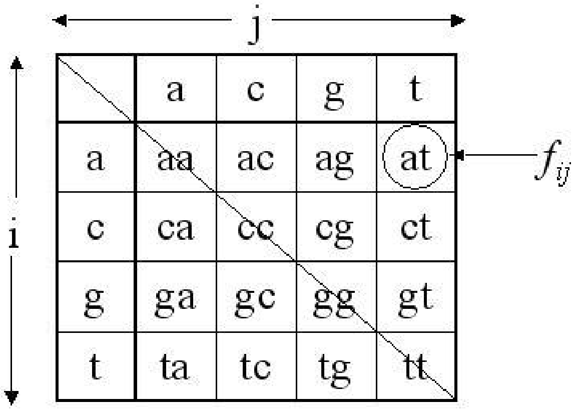


Figure 3.4. Illustration for the calculation of the number of possible pairs of nucleotides in the database, which involves the calculation of the possible pairs in each column of the database. For a database of width *w* nucleotides and a depth of *s* sequences, *w s (s-1)/2* nucleotide pairs can be obtained.

For a database with a width of *w* nucleotides and a depth of *s* sequences, $ws(s-1)/2$ nucleotide pairs can be obtained, giving the frequency of occurrence of each of the 10 (4+3+2+1) different nucleotide pairs in the database. Thus we obtained a 4x4 frequency table, with each of its elements being represented as f_{ij} (Figure 3.5). This table was further utilized for the calculation of log-odds matrix. In our case, *w* is taken to be 6, 10, or 14, while *s* depends on the particular organism (Table 3.1) studied.



	a	c	g	t
a	~~aa~~	ac	ag	at
c	ca	~~cc~~	cg	ct
g	ga	gc	~~gg~~	gt
t	ta	tc	tg	~~tt~~

Figure 3.5. Illustration of the construction of the frequency table (4x4) obtained by calculating the frequency of occurrence of each of the 10 (4+3+2+1) different pairs (of a, c, g and t) in the database.

(i) Log-odds matrix

Log-odds matrix is suitable to score alignments, in which the frequencies of the nucleotides in the aligned sequences are used to construct the substitution matrix. Log-odds values are calculated by taking a logarithm to base 2 ($\log_2$) of the ratio of the observed (target) probability to the expected (background) probability. The observed probability (q_{ij}) for each ij pair is calculated as given in Eq. (3.1):

$$q_{ij} = f_{ij} / \sum_{i=1}^{4} \sum_{j=1}^{i} f_{ij} \quad \longrightarrow \quad (3.1)$$

Then, the probability of occurrence (p_i) of the i^{th} nucleotide in an ij pair is calculated as given in Eq. (3.2):

$$p_i = q_{ij} + \frac{1}{2} \sum_{j \neq i} q_{ij} \quad \longrightarrow \quad (3.2)$$

The expected probability (e_{ij}) for each ij pair is then calculated as $e_{ij} = p_i p_j$ for $i=j$, and $e_{ij} = p_i p_j + p_j p_i = 2p_i p_j$ for $i \neq j$. The likelihood or the odds ratio matrix for each ij pair is calculated as the ratio of the observed probability to the expected probability: q_{ij}/e_{ij}, which gives the likelihood of occurrence of the nucleotides in pairs rather than by chance. The log-odds value of each ij pair is calculated as the logarithm of the odds ratio (s_{ij}), which is given as in Eq. (3.3):

$$s_{ij} = \log_2 (q_{ij} / e_{ij}) \quad \longrightarrow \quad (3.3)$$

(ii) Mutual information content (relative entropy)

The entropy of a random variable is a measure of the uncertainty of the random variable. Thus, it measures the amount of information required on average to describe the random variable. The entropy $H(X)$ of a discrete random variable X with the probability mass (or density) function $p(x)$ is defined as given in Eq. (3.4):

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \quad \longrightarrow \quad (3.4)$$

where the logarithm is taken to the base 2 and the entropy is expressed in bits. The relative entropy is a measure of the distance between two distributions.

The relative entropy or the Kullback-Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as given in Eq. (3.5):

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad \longrightarrow \quad (3.5)$$

Mutual information is defined as a measure of the amount of information that one random variable contains about the other. The mutual information $I(X; Y)$ of two random variables X and Y with a joint probability mass function $p(x; y)$ and marginal probability mass functions $p(x)$ and $p(y)$ is given as the relative entropy between the joint distribution and the product distribution $p(x)p(y)$ (Cover and Thomas, 1991) and is given as Eq. (3.6):

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad \longrightarrow \quad (3.6)$$

We have calculated the mutual information content for each block as the relative entropy H of the observed (target) probability to the expected (background) probability, which is given as Eq. (3.7):

$$H_{ij} = q_{ij} \times s_{ij} \quad \text{or} \quad H_{ij} = q_{ij} \times \log \frac{q_{ij}}{e_{ij}} \quad \longrightarrow \quad (3.7)$$

which is the product of the observed probability (q_{ij}) and the log-odds ratio (s_{ij}). The relative entropy of a log-odds substitution matrix is its ability to distinguish true alignments from other alignments, which appear by chance. We did not take over the sum of all the elements of the H matrix; instead, we plotted them as individual elements (H_{ij}) in the form of box plots.

(iii) Presentation of results

Instead of using a conventional histogram to display the results, we chose a box plot that shows the 25 and 75 percentiles as the box boundaries (Figure 3.6). The median (rather than the mean) value is shown within the box as a solid line. The error bars are shown as the 10 and 90 percentiles. This representation of data is more informative and gives a simple view of the distribution of the given data. All the plots were generated using the commercial software Sigmaplot 9.01 (Systat Software Inc., Richmond, USA).

3.3 Results and Discussions

We calculated the mutual information content (relative entropy *H*) for each of the organisms studied from their log-odds matrices. The log-odds matrices scoring the alignments of the mononucleotide substitutions were obtained from the substitution matrices constructed for the frequency of occurrence of the nucleotide pairs. The information content values for the three blocks of all the five organisms studied are plotted as vertical box plots for both donor and acceptor sites (Figure 3.6). The 16 elements (4x4) of the *H* matrix are plotted to get each box plot. These elements are the mean values of the given block and are directly comparable. Therefore, we are able to identify the contribution of the various elements individually.

The information content derived in this way is obviously a gross feature of the organism and perhaps can be divided into several groups such that the correlations within the groups are much more significant (compared to the whole genome; we expect the correlations between such groups may be quite less). The present plots in figure 3.6 are more informative as they show a better distribution of the given data. We can clearly see the trends by following the median or the other percentiles. In all the plots we note that the 90 percentile bars are far from the median, suggesting that few points have relatively high values. The data points with high values were then examined manually and correlated with the particular elements of the H_{ij} matrix as given in Table 3.2.

The box plots for the donor and acceptor sites of all the organisms studied (Figure 3.6) show interesting aspects that otherwise cannot be observed in the histograms (computed from the sum of H_{ij} matrix elements) of the average mutual information content.

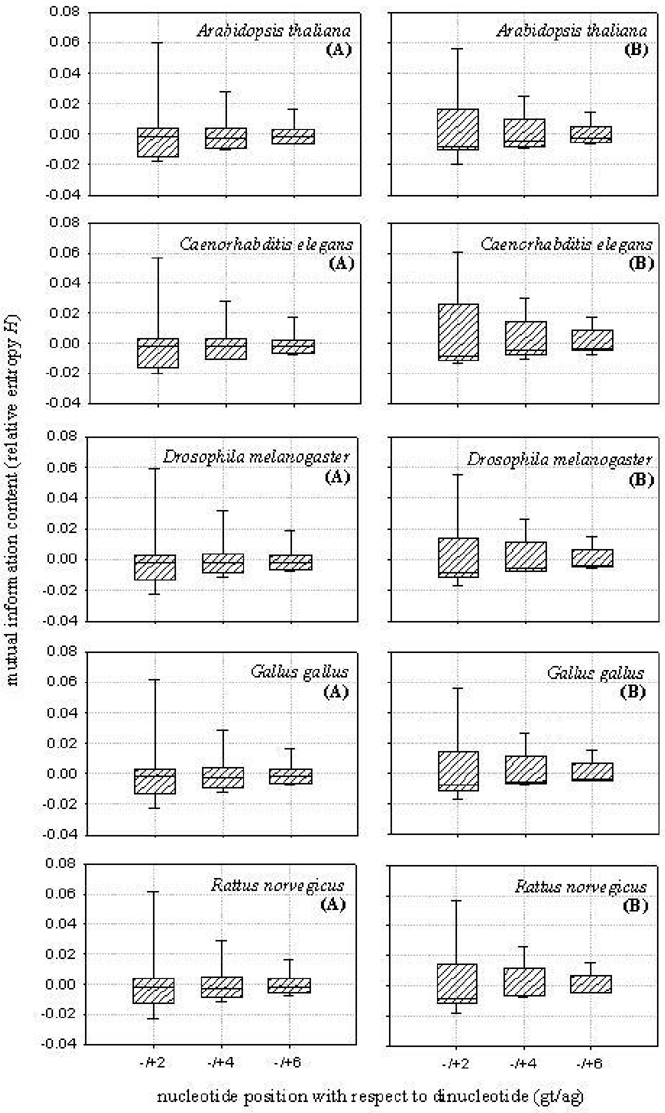


Figure 3.6. The mutual information content (relative entropy) calculated for donor (A; left column) and acceptor (B; right column) splice sites in the block sizes of 6 (gt±2, ag±2), 10 (gt±4, ag±4), and 14 (gt±6, ag±6) nucleotides of the genes of five different organisms studied. The boundaries of the boxes represent the 25 (lower) and 75 (upper) percentile points. The horizontal line within the box represents the median value. The error bars show the 10 (bottom) and 90 (top) percentile points. It is clearly seen that the distribution is highly skewed and all the cases of the 90 percentile points are comparatively high in value. The median values show relatively little variation between the three blocks studied. All the graphs have been plotted on the same scale for ease in visual comparison.

We can see that the information content (the height of the box) decreases with the increasing block size for both donor and acceptor regions in all the organisms studied, suggesting that the distribution of nucleotides around the splice site junctions is more conserved (that is, the splice sites are more variable compared to the neighboring regions).

The 6-nt block has the highest information content, and the information reduces considerably as we move away from the splice site. We speculate that the 6-nt block shows a greater variability (higher information content) and hence a higher selectivity. As we move to a larger window size, the variability decreases accordingly (as expected), suggesting that the selectivity of the spliceosomal binding is mainly dictated by the immediate neighborhood of the splice sites. This result reveals that ~2-3 nts flanking both sides of the splice sites are more important than longer distance nucleotides.

We also find that the median (50 percentile) values are more or less equal for all the plots. There exists a similar pattern of information content for both donor and acceptor sites in all the organisms studied, as they are significant for the binding of different spliceosomal proteins. We note that the values between 10-50 percentiles are very compact (less spread) while the values of 90 percentiles are far away from the median. This suggests that there are 1-2 values that are relatively high, which signify that the corresponding nucleotides are contributing to the high variability.

In order to get a better understanding, we correlated the box plots of each organism with the individual elements of the *H* matrix (H_{ij}, 4x4=16 individual values) to obtain the information about individual base pair preferences as given in Table 3.2.

Table 3.2 Base pair preferences at the donor and acceptor splice site regions obtained from the H matrix calculated

No	Organism	6-nt block	10-nt block	14-nt block
			Donor splice sites	
1	*A. thaliana*	gg>tt>aa>ac>ca>cc	gg>tt>aa>cc>ac>ca	gg>tt>aa>cc
2	*C. elegans*	gg>tt>aa>cc>ca	gg>tt>aa>cc>ac>ca	gg>tt>aa>cc>ac>ca
3	*D. melanogaster*	tt>gg>aa>cc>ac>ca	gg>tt>aa>cc>ac>ca	gg>tt>aa>cc
4	*G. gallus*	tt>gg>aa>ac>ca>cc	gg>tt>aa>cc>ac>ca	gg>tt>aa>cc
5	*R. norvegicus*	tt>gg>aa>ac>ca>cc	gg>tt>aa>cc>ca>ac	gg>tt>aa>cc
			Acceptor splice sites	
1	*A. thaliana*	gg>aa>cc>tt>ct>tc	gg>aa>tt>cc	gg>aa>tt>cc
2	*C. elegans*	gg>aa>cc>tt	tt>gg>aa>cc	tt>gg>aa>cc
3	*D. melanogaster*	gg>aa>cc>tt>ct>tc	gg>aa>tt>cc	gg>aa>tt>cc
4	*G. gallus*	gg>aa>tt>cc>ct>tc	gg>aa>tt>cc	gg>aa>tt>cc
5	*R. norvegicus*	gg>aa>cc>tt>ct>tc	gg>aa>tt>cc>ct>tc	gg>aa>tt>cc

3.3.1 Donor (5' splice site) region

We note from Table 3.2 that in the donor sequences the base pairs “gg” and “tt” have higher information content than “aa” and “cc” for all the cases. This is because the dinucleotide “gt” at the donor splice site is conserved and does not contribute to information content. Thus the high information content is attributed to the variability of the two nucleotides in the flanking regions of “gt”, which suggests a high probability of each of the two nucleotides getting substituted by the other. The probability of adenine getting substituted by cytosine (or *vice versa*) is also significant. We can see from the 6-nt block of donor sites that guanine is more preferred in the flanking regions (1-2 nt) of “gt” in *A. thaliana* and *C. elegans*, while thymine is more preferred in the flanking regions of *D. melanogaster*, *G. gallus*, and *R. norvegicus*. We also see from Table 3.2 that the extent of variability decreases as the block size increases, suggesting that the nucleotides contributing to the variability are present in the neighborhood of the splice sites.

3.3.2 Acceptor (3' splice site) region

We also note that in the acceptor sequences the base pairs "gg" and "aa" have higher information content than "tt" and "cc" for most cases. This is due to the conservation of the dinucleotide "ag" at the acceptor site, which does not contribute to the information content. This observation suggests that the given nucleotides in the decreasing order of their preferences contribute to the variability in the consensus of these sites. In the flanking nucleotides of "ag", the probability of thymine getting substituted by cytosine (or *vice versa*) is also observed. We note that the consensus at the acceptor region is more conserved than that at the donor region as fewer substitutions are observed comparatively, which is also evident from the high information content observed for the 6-nt block (Figure 3.6). It also shows a decreasing order in the preference of nucleotides as the block size increases (Table 3.2). We note from the 10-nt and 14-nt blocks of acceptor sequences that thymine is more preferred in the flanking regions of "ag" in *C. elegans*, which is due to the presence of the short and highly conserved polypyrimidine tract that is adjacent to the acceptor splice site. The consensus sequence TTTTCAG/R at the 3' end has been shown to be critical for its recognition and binding to the U2AF protein during the process of RNA splicing (Hollins *et al.*, 2005). All other organisms show general trends in the distribution of the nucleotides.

3.4 Conclusions

We assume from these observations that even though the nucleotides are showing some degrees of conservation in the flanking regions of the splice sites (gt/ag), there still exists a certain level of variability in the consensus, signifying that some substitutions are found to be tolerable at certain positions. This is presumed to respond to the different spliceosomal factors that lead the splicing process to occur selectively. Our study suggests that the information required for RNA splicing is contained in the consensus of ~6-8 nt at both donor and acceptor regions, which are important for the binding of spliceosomal proteins to the splice sites as expected.

We have developed our own block databases and applied the concepts of information theory for this analysis. Our study gives a broad idea about the distribution of nucleotides at/around the splice sites and also gives a comparative analysis of the consensus sequences at both donor and acceptor regions of the splice sites, which is significant for the process of splicing in terms of their sequence conservation or variability. We assume that our study can provide some insights towards understanding the information-hidden at/around the splice sites that are important for the process of splicing to occur efficiently. We conclude that variability is essential for the selectivity of the splicing process whereas conservation is desirable to restrict the degree of variability.

3.5 References

- Adami, C. (2002). Combinatorial drug design augmented by information theory. *NASA Tech Briefs* 26: 52.
- Adami, C. (2004). Information theory in molecular biology. *Phys. Life Rev.* 1: 3-22.
- Adami, C. and Thomson, S. W. (2005). Predicting protein-protein interactions from sequence data. In *The Chemical Theatre of Biological Systems. Proceedings of the International Beilstein Workshop* (eds. Hicks, M.G. and Kettner, C.). Logos Verlag, Berlin, Germany.
- Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219: 555-565.
- Ashburner, M., Golic, K. G. and Hawley, R. S. (2005). Drosophila: A Laboratory Handbook., 2nd ed., *Cold Spring Harbor Laboratory Press*, pp. 162-4.
- Brenner, S. (1974). The Genetics of *Caenorhabditis elegans. Genetics* 77: 71-94.
- Brunak, S., Engelbrecht, J. and Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220: 49-65.
- Chen, T., Lu, C. and Li, W. (2005). Prediction of splice sites with dependency graphs and their expanded Bayesian networks. *Bioinformatics* 21: 471-482.
- Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc., New York, USA.
- Dayhoff, M. O., Hunt, L. T., Barker, W. C., Schwartz, R. M., Orcutt, B. C. and young, C. L. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (ed. Dayhoff, M.O.), Vol.5, pp.345-352. National Biomedical Research Foundation, Washington DC, USA.

- Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
- Fields, C. (1990). Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Res.* 18: 1509-1512.
- Gadaleta, G., Pepe, G., DeCandia, G., Quagliariello, C., Sbisa, E. and Saccone, C. (1989) "The complete nucleotide sequence of the Rattus norvegicus mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates." *J Mol Evol.* 28(6): 497-516.
- Giraud, B. G., Lapedes, A. and Liu, L. C. (1998). Analysis of correlations between sites in models of protein sequences. *Phys. Rev. E* 58: 6312-6322.
- Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915-10919.
- Hollins, C., Zorio, D. A. R., Macmorris, M. and Blumenthal, T. (2005). U2AF binding selects for the high conservation of the *C. elegans* 3' splice site. *RNA* 11: 248-253.
- Lewin, B. (2000). Nuclear splicing. In *Genes VII.* Oxford University Press, New York, USA.
- Lolle, S. J., Victor, J. L., Young, J. M. and Pruitt, R. E. (2005). Genome-wide non-mendelian inheritance of extra-genomic information in Arabidopsis. *Nature* 434: 505-509.
- Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. and Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* 20: 4255-4262.

- Pertea, M., Lin, X. and Salzberg, S. L. (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* 29: 1185-1190.
- Reddy, D. A. and Mitra, C. K. (2006). Comparative analysis of core promoter region: information content from mono and dinucleotide substitution matrices. *Comput. Biol. Chem.* 30: 58-62.
- Rekha, T. S. and Mitra C. K. (2006). 1/*f* correlations in viral genomes a Fast-Fourier Transformation (FFT) study. *Indian J. Biochem. Biophys.* 43: 137-142.
- Rogan, P. K. and Schneider, T. D. (1995). Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum. Mutat.* 6: 74-76.
- Saxonov, S., Daizadeh, I., Fedorov, A., and Gilbert, W. (2000). EID: the Exon-Intron Database-an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.* 28: 185-190.
- Schneider, T. D., Stormo, G. D. and Gold, L. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188: 415-431.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27: 379-423, 623-656.
- Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R. and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* 34: 3955-3967.
- Staden, R. (1984). Computational methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* 12: 505-519.
- Stephens, R.M. and Schneider, T.D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* 228: 1124-1136.

Chapter 4

Frequency Analysis of the Splice Site Regions in Different Organisms

4.1 Introduction

4.1.1 Consensus sequences at the splice sites

Even though a lot of work has been done to predict splice sites within a gene, studying the sub-sequences at the splice sites has become important for understanding some of the aspects of splicing. The splice site regions are not conserved, as different genes need specific spliceosomes for activation (one spliceosome that activates all the genes is likely to be a very inefficient process). So, we expect a given spliceosomal complex to act on a small number of related genes. The intron boundaries are generally characterized by the presence of the dinucleotides, GU (at the donor) and AG (at the acceptor region). But all the GU…AG present in the genome are not always the integral components of the splice sites. So, it is important to study the sub-sequences at (and around) the splice sites, which contain most of the information required for splicing (attachment of the spliceosomal complex). The recognition of true splice sites was explained to certain extent by the exon-bridging interactions (Robberson *et al*., 1990), where the 5' splice site on the downstream side of an exon can be a crucial determinant in the recognition and splicing of the upstream intron. Earlier work carried out on splice sites also signifies that the distance between the splice sites affect efficient spliceosomal assembly (Fox-Walsh *et al*., 2005). But much remains to be known as to how the two (donor and acceptor) splice sites are paired together, so that they are spliced out efficiently.

4.1.2 Variability of sub-sequences at splice sites

In most higher organisms (metazoans), both the splice sites are generally characterized by the presence of loosely conserved consensus sequences at the junctions of introns and exons (5'-and 3'-splice sites), which are recognized by the snRNA of the spliceosomal complex (Black, 1995). Even though the consensus sequences at the splice sites are variable, they still contain the information required for splicing, which is contained in ~6-8 nucleotides at the donor| acceptor regions (Rekha and Mitra, 2006). It was also observed that the level of variability in them could be compensated by the recognition of different splice sites by different spliceosomal proteins, so that the process of splicing is carried out efficiently (Rekha and Mitra, 2006). One of the earlier models proposed states that the presence of certain nucleotides in certain positions plays a key role in the recognition of the consensus sequences at the splice sites (Milanesi and Rogozin, 1997). It also signifies that the more frequently a consensus is occurring at the splice site the more likely that it is considered to be the functional splice site.

4.1.3 Motivation for the study

Owing to the mechanism of RNA splicing it is interesting to identify the length of the consensus sequences in the given organisms that is optimal for the recognition of the splice sites by the spliceosomal proteins. Comparative analysis of the splice sites can also be helpful in obtaining the splice sites that are highly involved in splicing.

In order to obtain those sequences that are actually involved in splicing, we have obtained all sub-sequences at both donor and acceptor splice site regions (obtained from the protein coding intron containing gene sequences) of five different organisms. We have carried out a comparative study of a few selected sub-sequences that are occurring with a high frequency. We have also analyzed the same sequences to obtain an optimal length of the given sub-sequences that are actually found to be containing the information required for splicing. We have calculated the scores of the alignment of the high frequency donor| acceptor sub-sequences at the splice sites with the different set of sub-sequences (of any particular organism) occurring at the acceptor/donor splice sites and have obtained subsequences that might be paired during the process of splicing.

The basic focus in this work is neither the database nor the sequence analysis. We have looked for conserved regions around the splice sites but if they are too many in number and located at slightly variable locations, it may be difficult to identify all the sequences. We nevertheless could find several small conserved sub-sequences that may act as binding sites for various factors involved in splicing.

4.2 Methodology

4.2.1 Exon-Intron Database

We have downloaded the Exon-Intron Database (EID; release September 2005, http://hsc.utoledo.edu/bioinfo/eid/index.html) for our present analysis. It is a database of protein-coding intron containing gene sequences represented along with their alternative isoforms (Saxonov *et al.*, 2000). It was built in the FASTA format by obtaining the data from the GenBank database. The exon and intron (including the splice site dinucleotides gt| ag) sequences are represented separately as upper and lowercase letters. Gene sequences with three types of splice site (exon| intron) boundaries are given in the database - "gt-ag", "gc-ag" and "at-ac". In the present work, we have considered the gene sequences with "gt-ag" boundaries and have ignored all other splice sites, which were accounting for relatively small proportion. We have selected the gene sequences of five different organisms (with their alternative isoforms); such that we can have a broad distribution of the data from plants to mammals. The choice of organisms can be considered otherwise arbitrary. The selected organisms are *Arabidopsis thaliana* (plant), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (arthropod), *Gallus gallus* (aves) and *Rattus norvegicus* (mammal). The details of the number of gene sequences and splice sites considered in the present study are given in Table 4.1.

Table 4.1. Number of genes and splice sites (ss) of the organisms studied

No	**Organism**	**No. of genes**	**No. of ss**		**Total no of unique ss***	
			Donor	**Acceptor**	**Donor**	**Acceptor**
1	*A. thaliana*	20,716	130,099	131,229	14,082	23,118
2	*C. elegans*	18,594	111,970	112,361	14,231	7,852
3	*D. melanogaster*	10,612	72,737	73,167	7,189	15,058
4	*G. gallus*	16,567	168,120	169,990	17,839	27,813
5	*R. norvegicus*	19,146	181,782	183,476	15,921	28,284

*An unique splice site is defined as the10 nucleotide string xxxx{gt|ag}xxxx, where x can be any one of the nucleotides {A C, G, T}. If we select the 6-nucleotide string, the total number of unique splice sites will be considerably less.

4.2.2 Selection of sub-sequences

All the gene sequences of each of the five different organisms present in the EID database were used for the selection of sub-sequences for the present study. The sub-sequences were obtained by aligning the two centrally conserved dinucleotides (gt| ag) on either side of the donor/acceptor splice site regions of all the gene sequences in each organism separately, by considering two ($n_1n_2\{gt|ag\}n_3n_4$) and four ($n_1n_2n_3n_4\{gt|ag\}n_5n_6n_7n_8$) nucleotides flanking the splice sites. This way four different sets of sub-sequences were obtained for each of the organisms under study with two sets (one each for donor and acceptor) of size six and another two of size ten (Figure 4.1).

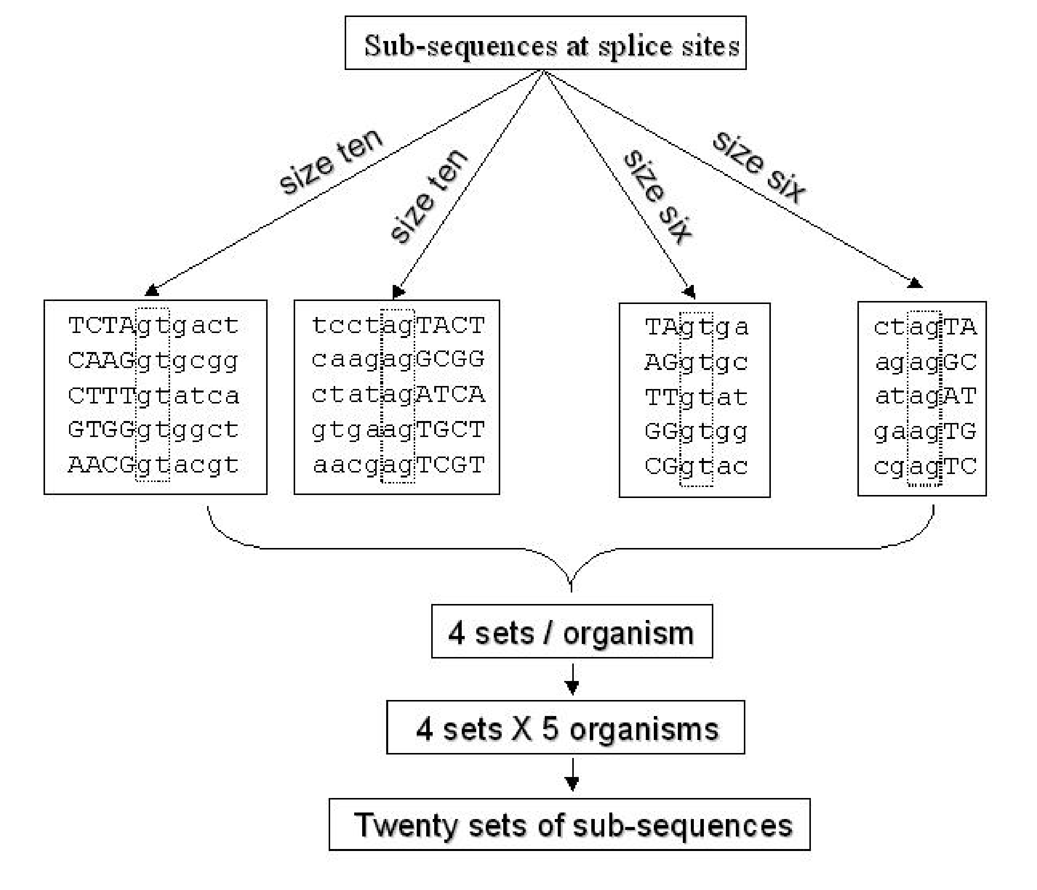


Figure 4.1. Illustration showing the sub-sequences that were obtained by aligning the two centrally conserved dinucleotides (gt| ag) on either side of the donor/acceptor splice site regions of all the gene sequences in each organism separately, by considering two ($n_1n_2\{gt|ag\}n_3n_4$) and four ($n_1n_2n_3n_4\{gt|ag\}n_5n_6n_7n_8$) nucleotides flanking the splice sites. Thus, four different sets of sub-sequences were obtained for each of the organisms under study with two sets (one each for donor and acceptor) of size six and another two of size ten. Totally we have obtained 20 different sets of sub-sequences with four sets for each of the organisms under study.

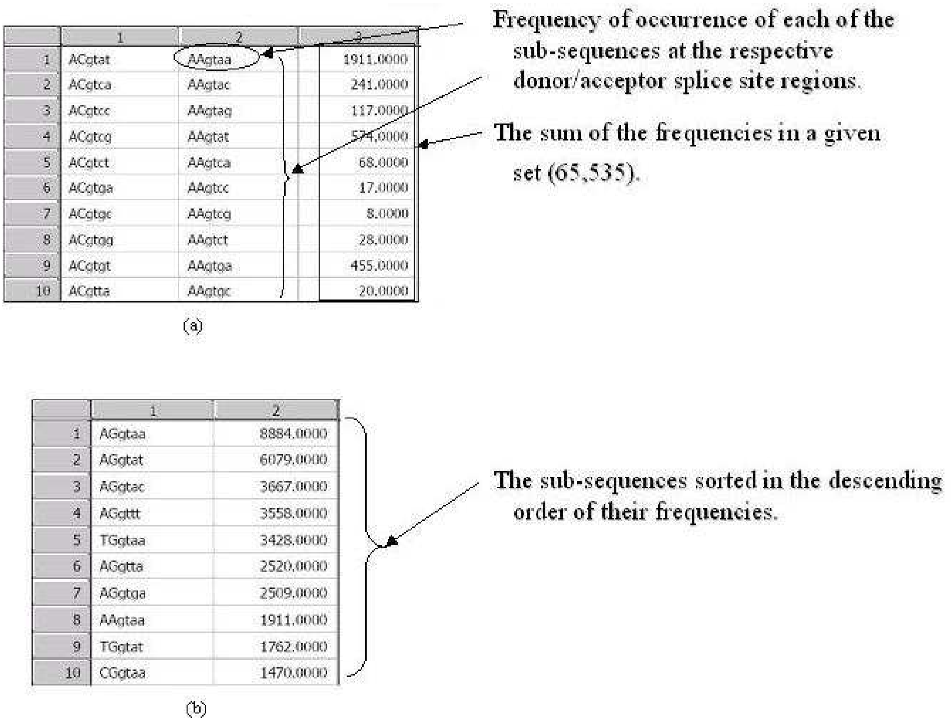
Thus, totally we have obtained 20 different sets of sub-sequences with four sets for each of the organisms under study. We have considered the sizes six| ten only because, from our earlier analysis it was observed that the information required for splicing is contained in ~6-8 nt around (donor| acceptor) the splice sites regions. We have considered only the first 65,535 splice sites of all the organisms in our analysis. This makes all the graphs comparable, as the total frequency is always the same (*vide infra*). The details of the number of unique sub-sequences of length 10 (at the splice sites) of each organism studied are given in Table 4.1.

4.2.3 Frequency distribution of sub-sequences

The 20 [5 (organisms) x 2 (donor| acceptor) x 2 (6| 10 nt length)] different sets of sub-sequences of size six| ten corresponding to the donor| acceptor regions of each of the five organisms considered for the study. Each set was then imported into a worksheet and sorted alphabetically. Each set now has several identical consecutive sub-sequences placed next to each other rather than being arranged in a random manner. The frequency of occurrence of each of the unique sub-sequences was calculated using a script. It is important to note that since, these subsequences were obtained from the splice site regions, so their frequency of occurrence gives their occurrence at the respective splice sites. The sum of the frequencies in a given set now corresponds to the total number of donor| acceptor splice sites for each of the organism under study (65,535 in this case). In the original worksheet, we had several redundancies (multiples) but after this process, all the sequences are now unique (Figure 4.2a).

These sub-sequences were sorted in descending order of their frequencies, so that we now have sub-sequences that are occurring most common at the top followed by the least common at the bottom of the worksheet (Figure 4.2b). We have obtained ~256 unique sub-sequences for the set of size six (for both donor and acceptor sites). In a similar fashion, we obtained ~10,000 unique ones for size 10, at the donor regions of all the organisms (except *D. melanogaster*). And the results were differing at the acceptor region with

~15,000-20,000 different types in all the organisms (except *C. elegans*). Overall, the number of unique splice sites are more in the acceptor region than the donor in all the organisms (except *C. elegans*) for size 10 (the differences are insignificant for size 6).



	1	2	3
1	ACgtat	AAgtaa	1911.0000
2	ACgtca	AAgtac	241.0000
3	ACgtcc	AAgtag	117.0000
4	ACgtcg	AAgtat	574.0000
5	ACgtct	AAgtca	68.0000
6	ACgtga	AAgtcc	17.0000
7	ACgtgc	AAgtcg	8.0000
8	ACgtgg	AAgtct	28.0000
9	ACgtgt	AAgtga	455.0000
10	ACgtta	AAgtgc	20.0000

(a)

Frequency of occurrence of each of the sub-sequences at the respective donor/acceptor splice site regions.

The sum of the frequencies in a given set (65,535).

	1	2
1	AGgtaa	8884.0000
2	AGgtat	6079.0000
3	AGgtac	3667.0000
4	AGgttt	3558.0000
5	TGgtaa	3428.0000
6	AGgtta	2520.0000
7	AGgtga	2509.0000
8	AAgtaa	1911.0000
9	TGgtat	1762.0000
10	CGgtaa	1470.0000

(b)

The sub-sequences sorted in the descending order of their frequencies.

Figure 4.2. Illustration showing the frequency of occurrence of each of the given sub-sequences (that were imported into a worksheet), which was calculated at the respective donor/acceptor splice site regions for each of the given organisms (a). The sum of the frequencies in a given set now corresponds to the total number of donor/acceptor splice sites for each of the organism under study (65,535 in this case). These sub-sequences were sorted in the descending order of their frequencies, so the sub-sequences that are occurring most common are at the top followed by the least common ones that are at the bottom of the worksheet (b).

4.2.4 Splice site utilization factor (*F*)

We have also calculated the splice site utilization factor (*F*), as in Eq. (4.1),

$$F = no.\ of\ splice\ sites\ (donor/acceptor) / No.\ of\ genes \longrightarrow \quad (4.1)$$

in each of the organisms studied, so that we can get an idea about the typical number of splice sites per gene in each organism. The values are tabulated (Table 4.2) for each species studied. We note that more evolved species has a higher value of *F*.

Table 4.2. Splice site utilization factor (*F*) of the organisms studied*

No	Organism	Splice site utilization factor (*F*) (No. of splice sites/No of genes)	
		Donor	**Acceptor**
1	*A. thaliana*	6-7	6-7
2	*C. elegans*	6-7	6-7
3	*D. melanogaster*	6-7	6-7
4	*G. gallus*	10-11	10-11
5	*R. norvegicus*	9-10	9-10

* The *F* values give the no. of splice sites per gene at the donor and acceptor regions of the given organisms.

4.2.5 Frequency plots of sub-sequences

The frequency values of each sub-sequence (arranged in descending order) at the donor| acceptor splice site regions of size six| ten were plotted as vertical bar charts (Figure 4.5 and 4.6) with the number of sub-sequences being plotted on x-axis and their corresponding frequencies on y-axis (using the commercial software Sigmaplot 9.01). We have considered only the first 65,535 number of splice sites of all the organisms in our analysis, such that the total area of all the graphs is the same (in all the plots). The x-axis tick labels are in reality the subsequences (of 6| 10 nts) that have not been shown. In addition, these sequences are not identical in all the species. These plots give us information about the frequency of occurrence of each sub-sequence at the donor| acceptor splice sites regions separately. The frequency axis has been conveniently plotted on a log scale for the ease of study and a regression line (Figure 4.5; in solid line) along with their slope value was also shown to indicate the trends.

(i) Study of the uniqueness of sub-sequences

As we cannot possibly study all unique sub-sequences occurring with different frequencies at the splice sites, we have considered only those sub-sequences of size six| ten, which are occurring with the highest, medium and lowest frequencies as representative to get a comparative analysis of the data. The medium frequency is taken as the 50% frequency of the highest value (median value). We have studied the uniqueness of the sub-sequences by computing the same as *(n/N)*100*, where *n* is the frequency of occurrence of the given subsequence at the splice sites and *N* is the frequency of occurrence of the same sub-sequence in the whole genome (for a given organism). This gives the uniqueness of the given subsequence, with higher the percentage, higher is the uniqueness and lower the percentage lower is the uniqueness. The uniqueness values for the three representative sub-sequences are tabulated in Table 4.3a (size six, donor sites), 3b (size six, acceptor sites), 4.4a (size ten, donor sites) and 4b (size ten, acceptor sites), which gives the details of the frequency of occurrence values of the respective sub-sequences, at the splice sites regions and also in the whole genomes of each of the organism being studied.

We assume from the above data that the sub-sequences, which are occurring more frequently, are the ones that are more commonly involved in the process of splicing. Based on this hypothesis, we have further studied sub-sequences that are occurring with the highest and medium frequencies at the donor| acceptor regions of each of the organisms being studied. As we have observed from our earlier analysis (Rekha and Mitra, 2006) of the splice site regions, that the information required for splicing might be contained in sub-sequences of ~6-8 nucleotides at the donor| acceptor regions. So, we have continued our further study with subsequences of size ten (occurring with highest and medium frequencies).

4.2.6 Identification of optimal length sub-sequences

One of the objectives of our study is to identify the sub-sequences of optimal length at the splice site (donor| acceptor) regions of each organism being studied, which are actually involved in the process of splicing. So, we have considered only sub-sequences of length ten (occurring with highest| medium frequencies at the splice sites) for our further analysis, as it is likely to be greater than the optimal sequence and discarded all sub-sequences of size six. It is not necessarily correct to assume that the information at| around the splice sites would be evenly distributed on both sides, and it is also important to consider the uneven distribution of the nucleotides on either side of the splice sites, so we have trimmed two and four nucleotides from either side of each sub-sequence, in a systematic manner to get sequences of length eight and six. Thus we have only focused on these five different sets of sub-sequences (A1…A10; A1…A8; A1…A6; A3…A10; A5…A10) for our further study (Figure 4.3).

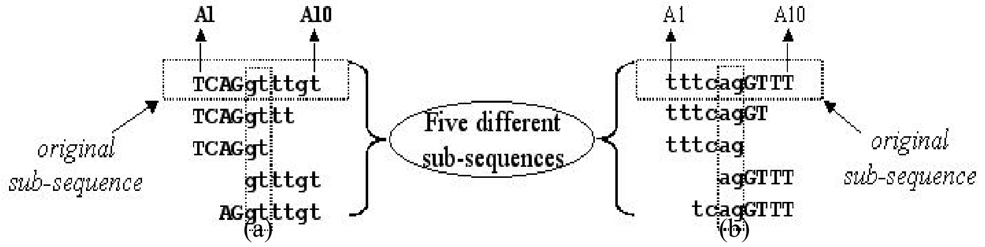


Figure 4.3. Illustration showing the method of trimming the original sub-sequence of size 10 (considering the uneven distribution of information in the sub-sequences at the donor/acceptor splice sites), by trimming two and four nucleotides from either side of the original sub-sequence at the (a) donor and (b) acceptor splice site regions, in order to obtain five different sub-sequences for each organism.

We have searched for these sub-sequences and have calculated their frequency of occurrence at the splice sites and also in the whole genome (in order to obtain only those sequences, which are occurring with the highest frequency at the splice sites with respect to the whole genome) and have recorded the number of matches found. For the ease of comparison, we have reported their percentage of occurrence (uniqueness) at the splice sites being calculated as described earlier in this chapter. Table 4.5a, 4.5b, 4.6a and 4.6b give details of the frequency and the percentage of occurrence (uniqueness) of all sub-sequences at both the splice sites (donor| acceptor) and in the whole genome of each of the organisms studied.

This way, we have identified sub-sequences, which are highly involved in the process of splicing by considering those that are having the highest percentage of occurrence. Table 4.7, gives a list of all sub-sequences whose percentage of occurrence (uniqueness) at the (donor| acceptor) splice sites was found to be the highest in each of the organisms studied.

4.2.7 Scoring the donor/acceptor sub-sequences

We note that the optimal length of the sub-sequences at the donor| acceptor splice site regions of each of the organisms studied is around eight nucleotides (Table 4.7). A unique donor sub-sequence that occurs with a high frequency is likely to be associated with a unique acceptor site occurring with high frequency. However, the frequency distributions for the donor and acceptor subsequences are clearly different and there may be other factors that determine the association between the donors and acceptors. To discover the pattern of association between the donors and acceptors, we use a scoring model. Both donor and acceptor sites are directly recognized by some RNA present in the spliceosomal complex and we hope to look for some correlations between these sequences. We do not imply that the model specifies perfect similarity of the sub-sequences but simply requires that some correlation must be detectable. Therefore the absolute value of the score is less important than the resulting shape of the distribution. With this as objective, we have scored the highest frequency unique sub-sequence (taken from Table 4.7) of the donor regions against the full set of unique sub-sequences at the acceptor sites. This has been done for all the organisms in a systematic manner. We also have carried out the reverse way, i.e., the highest frequency unique acceptor sequence has been scored against the complete set of unique sub-sequences at the donor sites. As the two distributions are clearly dissimilar, the results are expected to be different. As the donor and acceptor must occur in pairs, we are likely to see the correlation between them.

(i) Substitution matrix and Log-odds ratios

For this, we have constructed substitution matrices separately for the aligned set of subsequences of the given size of six/ten for the donor| acceptor regions of each of the organisms, in order to calculate their mononucleotide substitutions (Henikoff and Henikoff, 1992) as described in chapter 3 (Rekha and Mitra, 2006). The log-odds matrix is suitable to score alignments, in which the frequencies of the nucleotides in the aligned sequences were used to construct the substitution matrix and the odds values were calculated by taking the ratio of the observed (q_{ij}) to expected probability (e_{ij}), which is given as q_{ij}/e_{ij}. This ratio gives the likelihood of occurrence of the nucleotides in (*ij*) pairs rather than by chance. The log-odds value of each of the *ij* pair is calculated as the logarithm to base 2 (log2) of the odds ratio (S_{ij}), which is given as: $S_{ij}=log_2\,(q_{ij}/e_{ij})$.

(ii) Calculation of the scores

We have scored four types of alignments, (i) the unique donor sub-sequence (obtained from unique parent sub-sequence of size ten having highest percentage of occurrence) occurring with highest percentage (uniqueness) against each of the unique acceptor set of sub-sequences and (ii) the unique acceptor sub-sequence (obtained from unique parent sub-sequence of size ten having highest percentage of occurrence) occurring with highest percentage (uniqueness) against each of the unique donor set of sub-sequences (Figure 4.4).

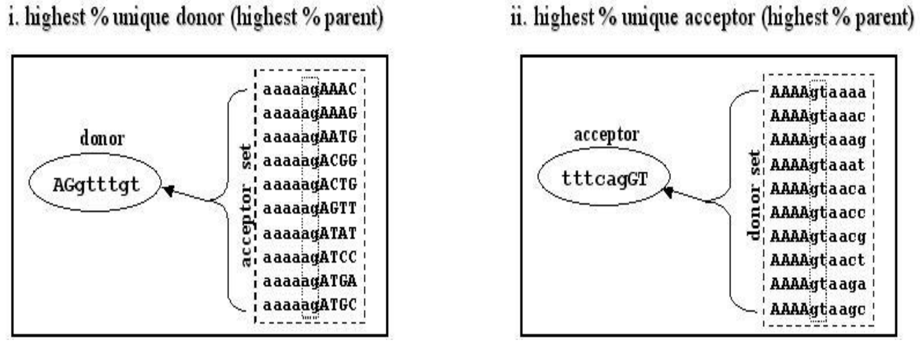


Figure 4.4. Illustration showing the method of scoring the pairs i.e., the unique sub-sequence (of size eight) occurring with the highest percentage of occurrence obtained from the parent (original sub-sequence of size 10) occurring with the highest percentage of occurrence at the donor splice site region paired with the set of sub-sequences (of size 10) at the acceptor region (i), and the unique sub-sequence (of size eight) occurring with the highest percentage of occurrence obtained from the parent (original sub-sequence of size 10) occurring with the highest percentage of occurrence at the acceptor splice site region paired with the set of sub-sequences (of size 10) at the donor region (ii).

Similar type of alignment was also done for the highest percentage of occurring unique donor and acceptor subsequences obtained from the unique parent sub-sequence of size ten occurring with medium frequency (iii) and (iv). All the highest frequency unique sub-sequences (donor/acceptor) aligned were of specific size for each of the organism considered for study (Table 4.7), which were aligned against the same size of the set of unique sub-sequences (acceptor/donor).

These alignments were then scored using the equation as given, $R = \Sigma_{ij} S_{ij}$ where R represents the score of the alignment, and S_{ij} represents the value assigned to the *ith* and the *jth* nucleotide in the log-odds matrix. This way, we have obtained four sets of scores for (i) unique donor-acceptor sub-sequence alignment (highest frequency unique parent) (ii) unique acceptor-donor sub-sequence alignment (highest frequency unique parent) (iii) unique donor-acceptor sub-sequence alignment (medium frequency unique parent) and (iv) unique acceptor-donor sub-sequence alignment (medium frequency unique parent). Thus, we have obtained 20 different sets of score values, which were plotted as histograms (Figures 4.7 and 4.8) using the software Sigmaplot. This way we can identify sub-sequences at the donor and acceptor regions that are actually paired during the process of splicing. The score values help us decide the similarities between the various sub-sequences, e.g., two sub-sequences with near identical scores may be really one sub-sequence. This can be used to reduce further the total number of unique sub-sequences.

4.3 Results and Discussions

4.3.1 Identification of unique sub-sequences

We have obtained unique sub-sequences (occurring with highest frequency) of size six| ten at donor| acceptor splice site regions in all the five organisms studied, which were ~256 in number for the set of size six and ~10,000 in number for the set of size ten. We note that the sub-sequences around the splice sites are highly variable, but far from random. The frequencies of sub-sequences follow an approximate exponential pattern that is common in nature (1/*f* distribution). As the length of sub-sequences increases their frequency of occurrence decreases and the total number of (observed) sub-sequences increases.

4.3.2 Significance of splice site utilization factor (*F*)

The *F* value gives the number of splice sites per gene of the given organisms and these values were found to be same for both the donor and acceptor regions (Table 4.2) (since the difference is small in the number of splice sites of both the donor and acceptor (Table 4.1)). We observe that the difference in the number of unique splice sites of the donor and the acceptor regions is significantly large (Table 4.1), which suggests that more variability is observed in the acceptor region than in the donor. These observations suggest that the given organisms can undergo either the constitutive (since same number of *F* values observed) or alternative splicing (because more variability is observed in the number of unique splice sites at the acceptor region, which suggests that one donor can get paired to more than one acceptor splice site).

4.3.3 Frequency Distribution of sub-sequences

We have calculated the frequency of occurrence of each unique sub-sequence (size six| ten), at the donor| acceptor, splice site regions and also in the whole genome of each organism studied. We note (Figures 4.5 and 4.6) that the frequency distribution is approximately exponential, because the occurrence of certain unique sub-sequences is more common when compared to the other. The distribution of sub-sequences of size six (Figure 4.5) is steeper in donor region, when compared to acceptor in all organisms studied except in *C. elegans*, in which the distribution is more or less equal in both the regions. We have drawn linear lines of regression for all the plots (Figure 4.5) and have obtained their respective slopes. We note that the slopes of the plots of donor region are higher than acceptor in all organisms (except *C. elegans*), which shows equal slopes for both (donor and acceptor) regions. This suggests that the frequency distribution at donor and acceptor regions is equal.

We note (Figure 4.5) that, since the frequency values are high in the donor region, the number of their corresponding (unique) sub-sequences are comparatively low (inverse relation). But the number is more in the acceptor region than the donor (thus their corresponding frequencies are less). This suggests that there are less number of donor and more number of acceptor splice sites in all the four organisms studied, signifying more variability in the acceptor region than the donor. But in *C. elegans*, we note the number to be approximately equal in both regions.

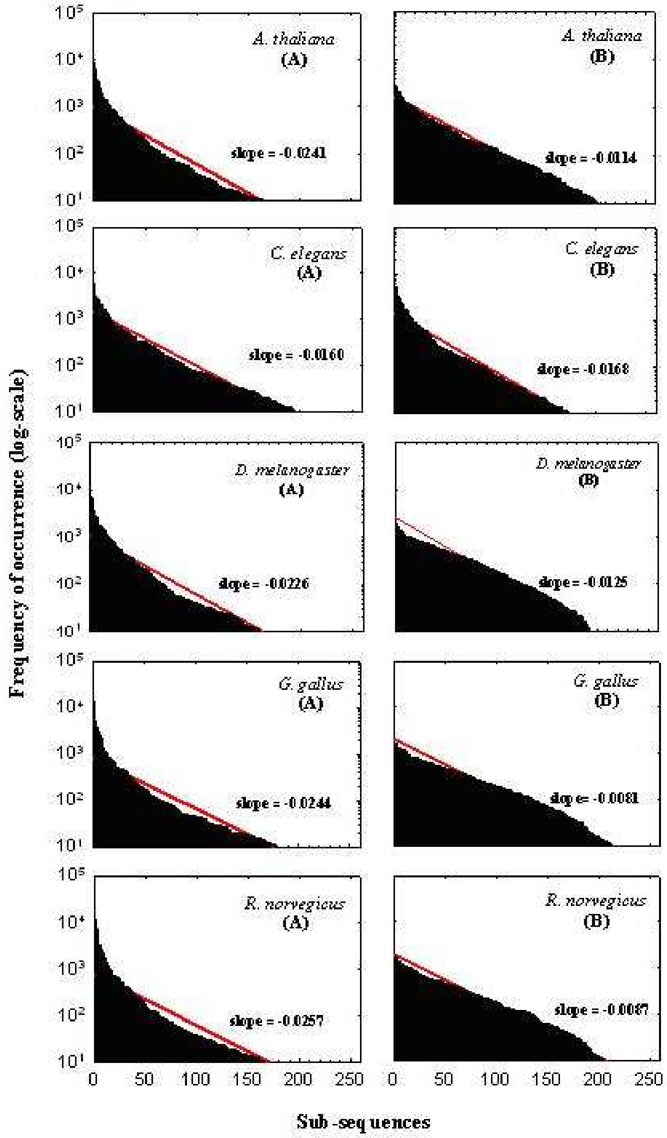


Figure 4.5. Vertical bar plots of the frequency of occurrence (log-scale) of the unique sub-sequences (arranged in descending order) in each set (first 65,535 sub-sequences considered) of size six of the respective organisms plotted against the corresponding sub-sequences (represented as numbers in linear scale) for the (A) donor and (B) acceptor splice site regions. Linear lines of regression are also shown (as solid lines) along with their respective slopes to indicate the trends of each plot. Scales of the axes are shown similar for all the organisms for the ease of comparison. The total area in each of the graphs is the same.

We have also observed a similar trend in the sub-sequences of size ten (Figure 4.6) with their frequencies higher at the donor region than the acceptor,

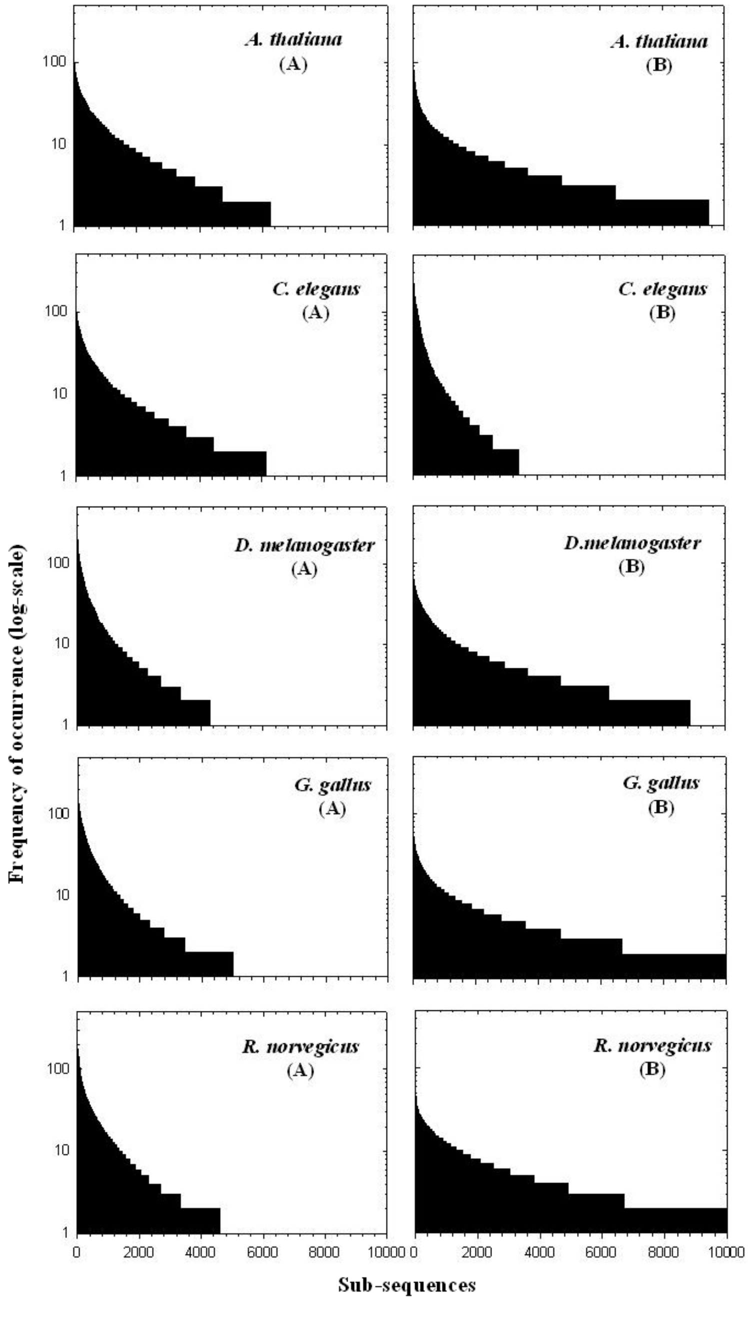


Figure 4.6. Vertical bar plots of the frequency of occurrence (log scale) of the unique sub-sequences (arranged in descending order) in each set (first 65,535 sub-sequences considered), of size ten of the respective organisms plotted against the number of corresponding sub-sequences (represented as numbers in linear scale) for the (A) donor and (B) acceptor splice site regions. Scales of the axes are shown similar for all the organisms for the ease of comparison. The total area in each of the graphs is the same.

(except for *C. elegans*, in which subsequences at the acceptor region are having frequencies higher than the donor). We also observe that the number of unique sub-sequences in the acceptor region are more than the donor, (except for *C. elegans*), which suggests that there is more variability in the acceptor region than the donor. We observe that the sub-sequences at the acceptor region in *C. elegans* are more conserved than the donor. This is because thymine is more preferred in the flanking regions of "ag" in *C. elegans*, which is due to the presence of the short and highly conserved polypyrimidine tract present adjacent to the acceptor splice site. The consensus sequence TTTTCAG/R at the acceptor region of *C. elegans* has been shown to be critical for its recognition and binding by the U2AF protein during the process of RNA splicing (Hollins *et al.*, 2005).

4.3.4 Non-random distribution of sub-sequences

We note that the frequency distribution of the sub-sequences is not uniform. If we consider the set of sub-sequences of size ten (Figure 4.6), the possibility of occurrence of each of the four bases at each of the eight positions (excluding the two central, highly conserved dinucleotides, "ag") would be 4^8 = 65,536, whereas the actual occurrence was found to be ~10,000 for each of the organisms. These observations suggests that there are certain unique sub-sequences, which are occurring more frequently than by random chance, (because certain bases are conserved at certain positions in the sub-sequences studied). But the frequency distribution of the set of sub-sequences of size six (Figure 4.5), was found to be as expected as 256 (i.e., the possibility of occurrence of each of the four bases (A, C, G and T) by random chance, in each of the four positions (excluding the two central, highly conserved dinucleotides, "gt") would be 4^4 = 256). This is in accordance with our earlier work (Rekha and Mitra, 2006), which suggests that there is more variability in the immediate flanking regions of the splice sites and the variability decreases as we move away from these splice sites.

4.3.5 Sub-sequences involved in splicing

From the frequency of occurrence values of sub-sequences of size six| ten at both (donor| acceptor) the splice sites and the whole genome (Table 4.3a, 4.3b, 4.4a and 4.4b) we assume that the sub-sequences with the highest frequency of occurrence at the splice sites are the ones, which are more commonly involved in the process of splicing.

Table 4.3a. Frequency of occurrence of different sub-sequences (size six) at the donor splice site (ss) region and the whole genome of the respective organisms

No	Organism (genome size in nts)	Frequency	Sub-sequences at ss†	Frequency at ss	Frequency in whole genome	Percentage of occurrence (uniqueness) at ss*
1.	*A. thaliana (58,129,057)*	Highest	AG**gt**aa	8,884	25,755	34.495
		Medium	Ag**gt**ac	3,667	13,548	27.067
		Lowest	TC**gt**tc	1	9,243	0.011
2.	*C. elegans (62,321,071)*	Highest	AG**gt**aa	5,916	15,338	38.571
		Medium	TG**gt**aa	2,903	14,907	19.475
		Lowest	TT**gt**cg	1	16,736	0.006
3.	*D. melanogaster (125,309,791)*	Highest	AG**gt**aa	7,362	23,877	30.834
		Medium	TG**gt**aa	3,558	28,006	12.705
		Lowest	TT**gt**tt	1	123,355	0.001
4.	*G. gallus (451,477,660)*	Highest	AG**gt**aa	12,970	137,320	9.446
		Medium	Ag**gt**ga	4,960	163,773	3.029
		Lowest	TC**gt**cg	1	3,918	0.026
5.	*R. norvegicus (867,510,682)*	Highest	AG**gt**aa	11,017	230,685	4.776
		Medium	AG**gt**ga	7,373	272,167	2.709
		Lowest	TA**gt**tc	1	194,280	0.001

*The percentage of occurrence (uniqueness) values are normalized to three decimal points in order to represent even the lowest values. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4.3b. Frequency of occurrence of different sub-sequences (size six) at the acceptor splice site (ss) region and the whole genome of the respective organisms

No	Organism (genome size in nts)	Frequency	Sub-sequences at ss[†]	Frequency at ss	Frequency in whole genome	Percentage of occurrence (uniqueness) at the ss*
1.	*A. thaliana (58,129,057)*	Highest	tc**ag**GT	2,733	23,964	11.405
		Medium	gt**ag**GT	1,321	7,882	16.760
		Lowest	cg**ag**CC	1	3,939	0.025
2.	*C. elegans (62,321,071)*	Highest	tc**ag**AT	5,270	27,645	19.064
		Medium	tc**ag**AC	2,791	14,441	19.327
		Lowest	gg**ag**TC	1	7,786	0.013
3.	*D. melanogaster (125,309,791)*	Highest	gc**ag**AT	1,814	35,990	5.041
		Medium	gc**ag**CA	917	114,441	0.802
		Lowest	tg**ag**TC	1	18,913	0.006
4.	*G. gallus (451,477,660)*	Highest	gc**ag**GT	1,687	139,010	0.214
		Medium	cc**ag**GA	845	140,060	0.604
		Lowest	tg**ag**CA	1	209,553	0.001
5.	*R. norvegicus (867,510,682)*	Highest	cc**ag**GT	1,855	274,033	0.677
		Medium	gc**ag**GT	1,443	222,987	0.648
		Lowest	ag**ag**CA	1	427,990	0.001

*The percentage of occurrence (uniqueness) values are normalized to three decimal points in order to represent even the lowest values. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4.4a. Frequency of occurrence of different sub-sequences (size ten) at the donor splice site (ss) region and the whole genome of the respective organisms

No	Organism (genome size in nts)	Frequency	Sub-sequences at ss[†]	Frequency at ss	Frequency in whole genome	Percentage of occurrence (uniqueness) at ss*
1.	*A. thaliana (58,129,057)*	Highest	TCAG**gt**ttgt	179	435	41.150
		Medium	AAAG**gt**aata	89	194	45.877
		Lowest	TTTT**gt**tttg	1	2,389	0.042
2.	*C. elegans (62,321,071)*	Highest	AAAA**gt**gagt	239	550	43.455
		Medium	AGAT**gt**aagt	120	240	50.000
		Lowest	TTTT**gt**tttt	1	240	0.417
3.	*D. melanogaster (125,309,791)*	Highest	CAAG**gt**gagt	506	615	82.277
		Medium	TGAG**gt**gagt	243	308	78.896
		Lowest	TTTT**gt**tatg	1	486	0.206
4.	*G. gallus (451,477,660)*	Highest	AAAG**gt**aaga	276	1,273	21.682
		Medium	CAAA**gt**aagt	136	892	15.247
		Lowest	TTTT**gt**tttc	1	8,091	0.013
5.	*R. norvegicus (867,510,682)*	Highest	CCAG**gt**gagt	247	1,502	16.445
		Medium	TCAG**gt**gagc	124	1,232	10.065
		Lowest	TTTT**gt**tttt	1	54,854	0.002

*The percentage of occurrence (uniqueness) values are normalized to three decimal points in order to represent even the lowest values. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4.4b. Frequency of occurrence of different sub-sequences (size ten) at the acceptor splice site (ss) region and the whole genome of the respective organisms

No	Organism (genome size in nts)	Frequency	Sub-sequences at ss[†]	Frequency at ss	Frequency in whole genome	Percentage of occurrence (uniqueness) at ss*
1.	*A. thaliana (58,129,057)*	Highest	`tttc`**`ag`**`GTTT`	119	545	21.825
		Medium	`ttgt`**`ag`**`GTGA`	59	176	33.523
		Lowest	`tttt`**`ag`**`TTCC`	1	121	0.827
2.	*C. elegans (62,321,071)*	Highest	`tttc`**`ag`**`AAAA`	651	3,744	17.388
		Medium	`tttc`**`ag`**`ATCA`	328	730	44.932
		Lowest	`tttt`**`ag`**`TGCG`	1	46	2.174
3.	*D. melanogaster (125,309,791)*	Highest	`ttgc`**`ag`**`ATGC`	137	374	36.632
		Medium	`ttgc`**`ag`**`TGCC`	69	248	27.823
		Lowest	`tttt`**`ag`**`TCGG`	1	95	1.053
4.	*G. gallus (451,477,660)*	Highest	`tttc`**`ag`**`GTTT`	99	2,412	4.105
		Medium	`ttgc`**`ag`**`GCAG`	50	1,817	2.752
		Lowest	`tttt`**`ag`**`TTCG`	1	107	0.935
5.	*R. norvegicus (867,510,682)*	Highest	`ctgc`**`ag`**`GTGG`	75	2,223	3.374
		Medium	`tttt`**`ag`**`GTTG`	38	1,494	2.544
		Lowest	`tttt`**`ag`**`TTgt`	1	2,452	0.041

*The percentage of occurrence (uniqueness) values are normalized to three decimal points in order to represent even the lowest values. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

We have obtained similar observations from the percentage of occurrence (uniqueness) values (Table 4.5a, 4.5b, 4.6a and 4.6b) of size ten, for each of the five organisms. We also note that the length of the respective sub-sequences (Table 4.7) occurring with the highest percentage of occurrence (uniqueness) might be optimal for the binding and assembly of the spliceosomal complex during the process of splicing.

Table 4.5a. Percentage of occurrence (uniqueness) of different sub-sequences (with highest frequency) of size ten at the donor splice site (ss) region and the whole genome of the respective organisms*

No	Organism	Sub-sequences at ss[†]	Frequency at ss	Frequency in whole genome	Percentage of occurrence (uniqueness) at ss
1.	*A. thaliana*	TCAG**gt**ttgt	179	435	41.15
		TCAG**gt**tt	1,168	3,122	37.42
		TCAG**gt**	9,302	23,964	38.82
		gtttgt	3,097	41,721	7.43
		AG**gt**ttgt	2,311	3,823	60.45
2.	*C. elegans*	AAAA**gt**gagt	239	550	43.46
		AAAA**gt**ga	716	5,594	12.80
		AAAA**gt**	2,616	71,992	3.64
		gtgagt	14,483	19,302	75.04
		AA**gt**gagt	2,491	2,998	83.08
3.	*D. melanogaster*	CAAG**gt**gagt	506	615	82.28
		CAAG**gt**ga	848	3,183	26.65
		CAAG**gt**	2,822	25,757	10.96
		gtaggt	15,759	36,271	43.45
		AG**gt**gagt	4,817	5,595	86.10
4.	*G. gallus*	AAAG**gt**aaga	276	1,273	21.69
		AAAG**gt**aa	3,802	15,643	24.31
		AAAG**gt**	9,591	160,235	5.99
		gtaaga	9,565	117,941	8.11
		AG**gt**aaga	4,614	11,235	41.07
5.	*R. norvegicus*	CCAG**gt**gagt	247	1,502	16.45
		CCAG**gt**ga	2,413	19,020	12.69
		CCAG**gt**	10,859	274,033	3.97
		gtgagt	24,678	301,321	8.19
		AG**gt**gagt	6,186	18,281	33.84

*The sub-sequences of size ten found with highest frequency at the donor splice site region were trimmed; two/four bases to obtain different sub-sequences of six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at the donor splice site region in all the five organisms studied. The low percentage of occurrence values in *G. gallus* and *R. norvegicus* can be due to the limited data taken for the study. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4.5b. Percentage of occurrence (uniqueness) of different sub-sequences (with highest frequency) of size ten at the acceptor splice site (ss) region and the whole genome of the respective organisms*

No	Organism	Sub-sequences at ss[†]	Frequency at ss	Frequency in whole genome	Percentage of occurrence (uniqueness) at ss
1.	*A. thaliana*	tttc**ag**GTTT	119	545	21.84
		tttc**ag**GT	2,169	4,384	49.48
		tttc**ag**	9,532	36,304	26.26
		agGTTT	2,948	33,668	8.76
		tc**ag**GTTT	534	3,122	17.11
2.	*C. elegans*	tttc**ag**AAAA	651	3,744	17.39
		tttc**ag**AA	7,588	14,916	50.88
		tttc**ag**	53,053	94,019	56.43
		agAAAA	2,369	100,523	2.36
		tc**ag**AAAA	1,250	10,057	12.43
3.	*D. melanogaster*	ttgc**ag**ATGC	137	374	36.64
		ttgc**ag**AT	1,144	3,553	32.20
		ttgc**ag**	9,405	50,070	18.79
		agATGC	676	28,570	23.72
		gc**ag**ATGC	220	3,447	6.39
4.	*G. gallus*	tttc**ag**GTTT	99	2,412	4.11
		tttc**ag**GT	1,896	19,295	9.83
		tttc**ag**	13,530	361,795	3.74
		agGTTT	2,623	185,539	1.42
		tc**ag**GTTT	434	15,976	2.72
5.	*R. norvegicus*	ctgc**ag**GTGG	75	2,223	3.38
		ctgc**ag**GT	1,326	22,362	5.93
		ctgc**ag**	7,858	403,613	1.95
		agGTGG	3,356	288,827	1.17
		gc**ag**GTGG	554	25,010	2.22

*The sub-sequences of size ten found with highest frequency at the acceptor splice site region were trimmed; two/four bases to obtain different sub-sequences of six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at the acceptor splice site region in all the five organisms studied. The low percentage of occurrence values in *G. gallus* and *R. norvegicus* can be due to the limited data taken for the study. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4.6a. Percentage of occurrence (uniqueness) of different sub-sequences (with medium frequency) of size ten at the donor splice site (ss) region and the whole genome of the respective organisms*

No	Organism	Sub-sequences at ss[†]	Frequency at ss	Frequency in whole genome	Percentage of occurrence (uniqueness) at ss
1.	*A. thaliana*	AAAG**gt**aata	89	194	45.88
		AAAG**gt**aa	1,964	2,938	66.85
		AAAG**gt**	8,132	26,204	31.04
		gtaata	2,378	14,535	16.37
		AG**gt**aata	1,315	1,802	72.98
2.	*C. elegans*	AGAT**gt**aagt	120	240	50.00
		AGAT**gt**aa	344	1,185	29.03
		AGAT**gt**	769	18,213	4.23
		gtaagt	16,442	21,110	77.89
		AT**gt**aagt	2,036	2,488	81.84
3.	*D. melanogaster*	TGAG**gt**gagt	243	308	78.90
		TGAG**gt**ga	409	1,466	27.90
		TGAG**gt**	1,330	15,803	8.42
		gtgaGT	15,759	36,271	43.45
		AG**gt**gagt	4,817	5,595	86.10
4.	*G. gallus*	CAAA**gt**aagt	136	892	15.25
		CAAA**gt**aa	584	14,013	4.17
		CAAA**gt**	993	157,546	0.64
		gtaagt	21,815	115,220	18.94
		AA**gt**aagt	2976	11,877	20.06
5.	*R. norvegicus*	TCAG**gt**gagc	124	1,232	10.07
		TCAG**gt**ga	1,808	22,266	8.13
		TCAG**gt**	9,080	269,752	3.37
		gtgagc	7,391	250,002	2.96
		AG**gt**gagc	3,943	17,670	51.41

*The sub-sequences of size ten found with highest frequency at the donor splice site region were trimmed; two/four bases to obtain different sub-sequences of six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at the donor splice site region in all the five organisms studied. The low percentage of occurrence values in *G. gallus* and *R. norvegicus* can be due to the limited data taken for the study. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4.6b. Percentage of occurrence (uniqueness) of different sub-sequences (with medium frequency) of size ten at the acceptor splice site (ss) region and the whole genome of the respective organisms*

No	Organism	Sub-sequences at ss†	Frequency at ss	Frequency in whole genome	Percentage of occurrence (uniqueness) at ss
1.	*A. thaliana*	ttgt**ag**GTGA	59	176	35.53
		ttgt**ag**GT	1,095	1,786	61.32
		ttgt**ag**	5,995	22,839	26.25
		agGTGA	2,711	19,045	14.24
		gt**ag**GTGA	273	688	39.69
2.	*C. elegans*	tttc**ag**ATCA	328	730	44.94
		tttc**ag**AT	7,548	11,078	68.14
		tttc**ag**	53,053	94,019	56.43
		agATCA	1,253	21,616	5.80
		tc**ag**ATCA	682	1,735	39.31
3.	*D. melanogaster*	ttgc**ag**TGCC	69	248	27.83
		ttgc**ag**TG	510	3,389	15.05
		ttgc**ag**	9,405	50,070	18.79
		agGTCC	362	11,165	3.25
		gc**ag**GTCC	58	1,145	5.07
4.	*G. gallus*	ttgc**ag**GCAG	50	1,817	2.76
		ttgc**ag**GC	965	11,154	8.66
		ttgc**ag**	12,489	276,171	4.53
		agGCAG	1,404	214,811	0.66
		gc**ag**GCAG	361	22,684	1.60
5.	*R. norvegicus*	tttt**ag**GTTG	38	1,494	2.55
		tttt**ag**GT	1,064	28,178	3.78
		tttt**ag**	6,137	385,529	1.60
		agGTTG	1,839	211,489	0.87
		tt**ag**GTTG	223	11,539	1.94

*The sub-sequences of size ten found with highest frequency at the acceptor splice site region were trimmed; two/four bases to obtain different sub-sequences of six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at the acceptor splice site region in all the five organisms studied. The low percentage of occurrence values in *G. gallus* and *R. norvegicus* can be due to the limited data taken for the study. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4.7. Sub-sequences at the donor and acceptor splice site regions of different organisms found with the highest percentage of occurrence*

No	Organism	Sub-sequences obtained from original sequence found with respective frequency[†]		Sub-sequences obtained from original sequence found with respective frequency[†]	
		Donor region		Acceptor region	
		Highest	Medium	Highest	Medium
1	*A. thaliana*	AG**gt**ttgt	AG**gt**aata	tttc**ag**GT	ttgt**ag**GT
2	*C. elegans*	AA**gt**gagt	AT**gt**aagt	tttc**ag**	tttc**ag**AT
3	*D. melanogaster*	AG**gt**gagt	AG**gt**gagt	ttgc**ag**ATGC	ttgc**ag**TGCC
4	*G. gallus*	AG**gt**aaga	AA**gt**aagt	tttc**ag**GT	ttgc**ag**GC
5	*R. norvegicus*	AG**gt**gagt	AG**gt**gagc	ctgc**ag**GT	tttt**ag**GT

*The sub-sequences of size ten found with highest and medium frequency at both the splice sites were trimmed; two/four bases to obtain different sub-sequences of ten six, and eight, which were used to calculate their percentage of occurrence at both donor and acceptor splice sites in all the five organisms studied. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

4.3.6 Sub-sequences of optimal length

(i) Sub-sequences at donor region (highest frequency)

From the data (Table 4.5a) obtained, we observe that sub-sequences with highest percentage of occurrence or uniqueness (obtained from parent sub-sequence with highest frequency) containing two bases in the exonic region and six bases in the intronic region (including the two highly conserved dinucleotides "gt") might be highly involved in the process of splicing at the donor region of all the organisms studied (Table 4.7) and contain a length of eight nucleotides, which is optimal for the spliceosomal assembly and binding of the organisms studied.

(ii) Sub-sequences at acceptor region (highest frequency)

The consistency shown in the donor region is not really observed at the acceptor regions of the organisms studied because from the data obtained (Table 4.5b) the optimal length of subsequences (obtained from parent sub-sequence of highest frequency) at the acceptor splice site region is eight with six bases in the intronic region (including the two conserved dinucleotides "ag") and two bases in the exonic region in three of the organisms studied – *A. thaliana*, *G. gallus* and *R. norvegicus*. But in *C. elegans*, the optimal length is found to be six, with all the bases in the intronic region (including the two conserved dinucleotides "ag") only. But in *D. melanogaster*, the optimal length is more than all other species, i.e., ten with six bases in the intronic region (including the two conserved dinucleotides "ag") and four bases in the exonic region. So the optimal length of sub-sequences at the acceptor region required for splicing is highly variable in the organisms studied (Table 4.7). This is perhaps due to the fact that one donor may be able to choose from a number of different acceptors.

(iii) Sub-sequences at donor region (medium frequency)

The data (Table 4.6a) obtained, represents a similar trend (as observed for the donor region discussed earlier) of the sub-sequences (obtained from parent sub-sequence of medium frequency) at the donor regions. These sub-sequences (Table 4.7), with their respective optimal lengths might be moderately involved in splicing in the organisms studied. The difference is in degree and the basic idea remains the same.

(iv) Sub-sequences at acceptor region (medium frequency)

For sub-sequences (obtained from parent sub-sequence of medium frequency) at the acceptor region (Table 4.6b), we observe the optimal length to be eight in the four organisms studied, with six bases in the intronic region (including the two conserved dinucleotides "ag") and two bases in the exonic region (except *D. melanogaster*, where the optimal length was ten, as discussed in [section 4.3.6 (ii)]). We assume that these sub-sequences (Table 4.7) are moderately involved in the process of splicing in the organisms studied.

4.3.7 Scoring the alignments of donor-acceptor sub-sequences

Based on the hypothesis that the certain sub-sequences at the donor region have some similarity with the sub-sequences at the acceptor, we have scored the alignments of the unique donor sub-sequence (occurring with highest percentage of occurrence obtained from parent sub-sequence occurring with highest/medium percentage of occurrence) with each of the subsequences in the set of acceptor region and *vice-versa*. We have observed certain features in the graphs obtained by plotting these score values, which are discussed in detail as follows.

(i) Donor (highest frequency parent sub-sequence) aligned against acceptor set

We observe from the histograms (Figure 4.7A) that the frequency of the score values (represented as percentage of occurrence or uniqueness) obtained by aligning the highest percentage of occurrence donor sub-sequence (obtained from parent sub-sequence occurring with highest percentage of occurrence) with each sub-sequence at the acceptor region is not normal. We have observed that the distribution is multimodal, which signifies that a single graph has a number of normal distributions combined together in it. We have also observed many peaks, which denote that a single donor sub-sequence has different degree of similarity with each of the sub-sequences at the acceptor region. We also assume that the donor subsequences are more crucial in deciding the acceptor region for splicing. The graph shows clustering behavior with each cluster having peaks of different intensity. Different clusters were obtained as the donor sub-sequence is having similarity with different nucleotides in the acceptor sub-sequence. We have obtained negative scores for the sub-sequence similarity, which can be due to mismatches between some nucleotides at the donor and acceptor regions that are making the overall score of the alignment to be negative. But we also observe some positive scores for the alignment, which are found to be very less. This suggests that the similarity between the nucleotides in the donor and acceptor sub-sequences at the splice sites is not very high *per se*. However, it is not expected that the sequence information transmitted from the donor site to the acceptor site via the snRNA will be perfect. In such case, we stress more on the distribution rather than the exact value of the score.

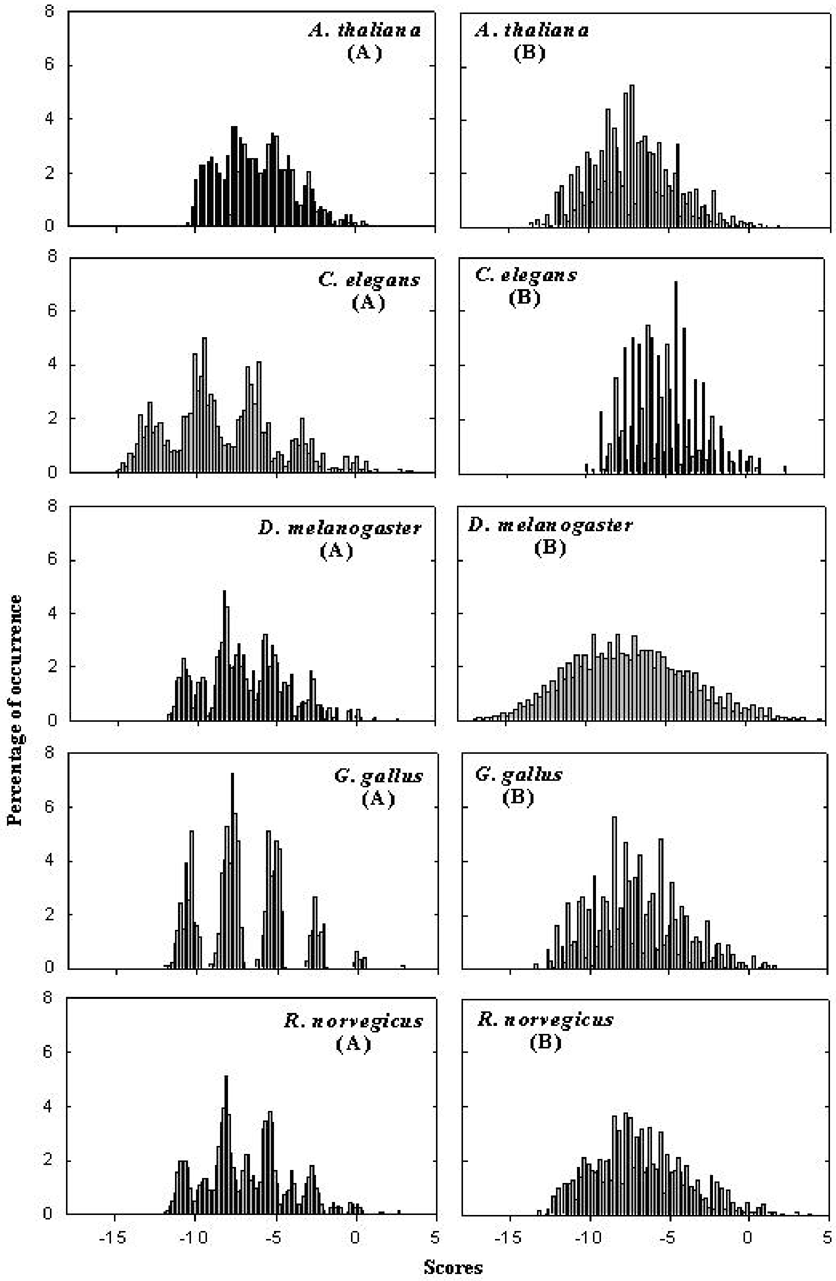


Figure 4.7. Histograms obtained by plotting the scores values (x-axis) against their percentage of occurrence (y-axis). These values were obtained by scoring the alignment of the (A; left column) the highest frequency unique donor sub-sequence against each of the unique acceptor subsequences and similarly by the alignment of the (B; right column) highest frequency unique acceptor sub-sequence against each of the unique donor sub-sequences for each of the organisms under study. The highest frequency donor/acceptor sub-sequences aligned were found to be of specific size for each organism (Table 4.7) and were obtained from the parent sub-sequence of size ten having highest percentage of occurrence.

(ii) Acceptor (highest frequency parent sub-sequence) aligned with donor set

The plots (Figure 4.7B) of the score values obtained by aligning the highest percentage of occurrence (uniqueness) acceptor sub-sequence (obtained from parent sub-sequence occurring with highest percentage occurrence) with each of the sub-sequences at the donor region suggests that the distribution is more or less normal in *A. thaliana*, *D. melanogaster* and *R. norvegicus*. But in *C. elegans* and *G. gallus* it shows the characteristics of a comb distribution with edge peaks. This distribution suggests that the sub-sequence occurring with the highest percentage of occurrence (uniqueness) at the acceptor region do not have proper alignment with sub-sequences at the donor region suggesting that the acceptor regions are not crucial in deciding the splicing process.

(iii) Donor (medium frequency parent sub-sequence) aligned with acceptor set

We observe that the plots (Figure 4.8A) of the score values obtained by aligning the highest percentage of occurrence donor sub-sequence (obtained from parent sub-sequence occurring with medium percentage of occurrence) with the sub-sequences at the acceptor region, show similar trends as discussed earlier [section 4.3.7.(i)] but the patterns seen here are not very clear (well resolved).

(iv) Acceptor (medium frequency parent sub-sequence) aligned with donor set

The plots (Figure 4.8B) of the score values obtained by aligning highest percentage of occurrence of acceptor sub-sequence (obtained from parent sub-sequence occurring with medium percentage occurrence) with the sub-sequences at the donor region, has shown a normal distribution in all the four organisms studied. But in *R. norvegicus*, we observe a comb distribution (with edge peaks), with one set of high values and another set of low values being represented together. This distribution suggests similar conclusions as given earlier [section 4.3.7.(ii)]. Again, we find the behavior broadly similar and it is only different in degree.

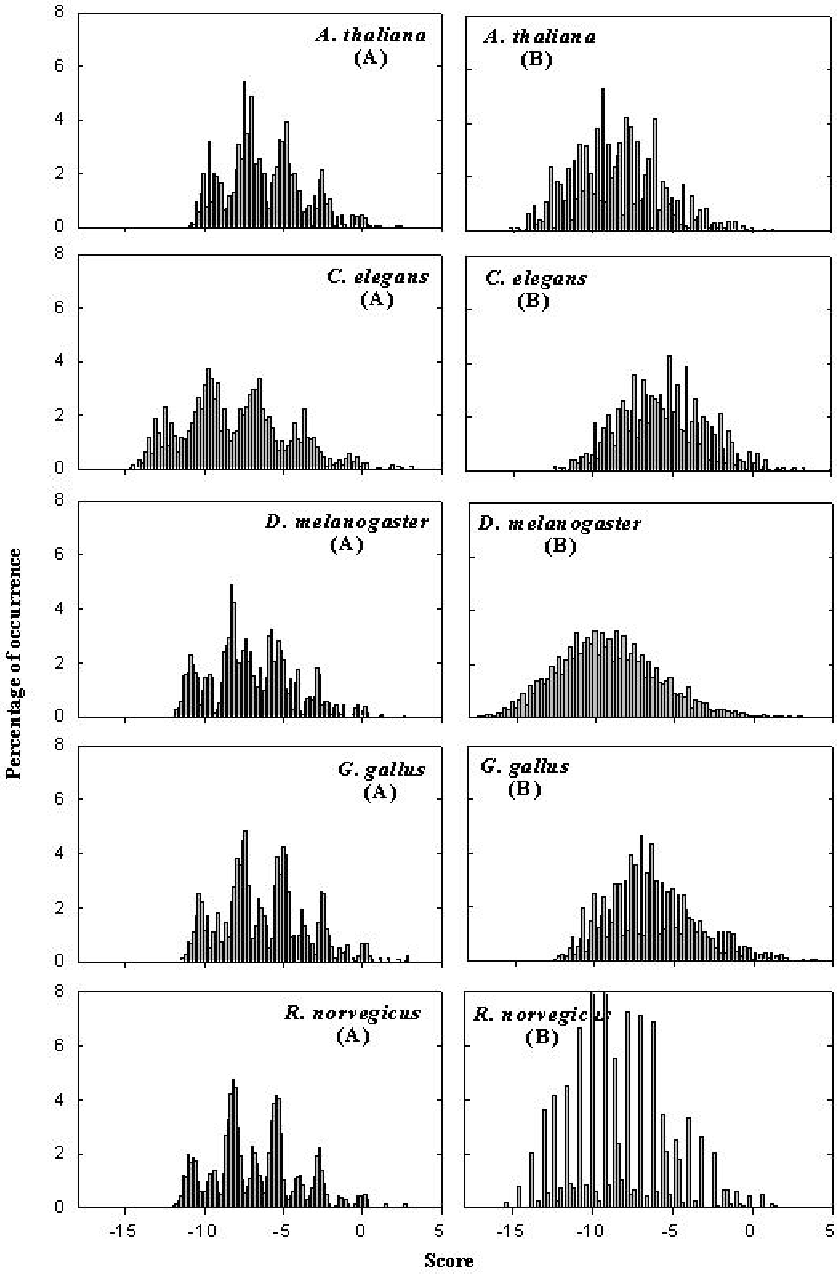


Figure 4.8. Histograms obtained by plotting the scores values (x-axis) against their percentage of occurrence (y-axis). These values were obtained by scoring the alignment of the (A; left column) the highest frequency unique donor sub-sequence against each of the unique acceptor subsequences and similarly by the alignment of the (B; right column) highest frequency unique acceptor sub-sequence against each of the unique donor sub-sequences for each of the organisms under study. The highest frequency donor/acceptor sub-sequences aligned were found to be of specific size for each organism (Table 4.7) and were obtained from the parent sub-sequence of size ten having medium percentage of occurrence.

4.4 Conclusions

We show that the information required for splicing is contained in ~6-8 nt at| around both the donor and acceptor splice sites. This work has given us a better idea about the distribution of information at| around the splice sites suggesting that sub-sequences at the splice sites studied are highly variable. The frequency analysis of these unique sub-sequences also suggests that the distribution is approximately exponential, because of the occurrence of certain high frequency unique sub-sequences more commonly than the other. The percentage of occurrence (uniqueness) values also suggests that sub-sequences with the highest values are the ones, which are highly involved in splicing. We also note that the length of 6-8 nts with six bases in intron (including the two central, conserved dinucleotides) and two bases in exon is optimal for the efficient assembly and binding of the spliceosomal complex during the process of splicing. We assume that the donor sub-sequences are more crucial in pairing with the corresponding acceptor sub-sequences during the process of splicing.

Further this idea can be extended in decoding the information present at the splice sites into distinct groups and classes. The rich variability of the donor and acceptor sites generates greater information and the information may be useful in understanding the language of the DNA at the splice site. Considerable experiments need to be carried out before the problem can be uniquely solved. However, we have clearly identified a number of broad features that can help in this direction. This kind of work can be carried in understanding the information contained in the promoter regions also, which might give some insights into the underlying mechanism.

4.5 References

- Black, D. L. (1995). Finding splice sites within a wilderness of RNA. *RNA*, 1: 763-771.
- Fox-Walsh, K. L., Dou, Y., Lam, B. J., Hung, S., Baldi, P. F. and Hertel, K. J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc. Natl. Acad. Sci. USA*, 102: 16176-16181.
- Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89: 10915-10919.
- Hollins, C., Zorio, D. A. R., Macmorris, M. and Blumenthal, T. (2005). U2AF binding selects for the high conservation of the C. elegans 3' splice site. *RNA*, 11: 248-253.
- Milanesi, L. and Rogozin, I. B. (1997). Analysis of donor splice sites in different eukaryotic organisms. *J. Mol. Evol.*, 45: 50-59.
- Robberson, B. L. *et al.* (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell Biol.*, 10 (1): 84-94.
- Saxonov, S., Daizadeh, I., Fedorov, A. and Gilbert, W. (2000). EID: the Exon-Intron Database - an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, 28 (1): 185-190.
- Rekha, T.S. and Mitra, C. K. (2006). Comparative Analysis of Splice Site Regions by Information Content. *Genomics Proteomics Bioinformatics*, 4: 230-237.

Chapter 5

Frequency Studies of the Unique Sub-sequences at the Splice Sites

5.1 Introduction

The complexity of the eukaryotic genome is not only because of the genome size but is also due to the complexity in the utilization of different splice sites (ss) to generate a diverge range of alternate products. Pre-mRNA splicing is an importance process involving the removal of introns from pre-mRNA by the two concerted transesterification reactions to form mature mRNA, which is later, translated to form the proteins (Lewin, 2000).

Earlier studies signify the variability of sub-sequences at the donor and acceptor splice site regions, which suggest that the information required for splicing is contained in the sub-sequences of ~6-8 nucleotides at the donor and acceptor regions (Rekha and Mitra, 2006). It also suggests that even though the sub-sequences at the splice sites are showing some conservation, a certain degree of variability is observed in them, which might be compensated by the recognition of different splice sites by different spliceosomal proteins. Frequency studies on the recognition of sub-sequences at the splice sites (that are involved in splicing), suggests that the sub-sequences, which are occurring more frequently are the ones that are highly involved in the process of splicing (Rekha and Mitra, 2007). These studies also led to the identification of the optimal length of these sub-sequences that are involved in splicing.

Splicing produces diversity in proteins but preserves the gene codes. In other words, a small number of genes can produce a large number of proteins (Graveley, 2001). To do this efficiently, suitable codes must be present in the gene itself. These are well established in the donor (5') and acceptor (3') sites that mark the boundaries between the introns and exons (Lewin, 2000). As a part of gene regulation, all the genes must not be transcribed all the times and there are factors that are responsible for this regulation (Yeo *et al.*, 2004).

In the case of splicing, all the distinct proteins (required for splicing) are not transcribed simultaneously (else the benefits of splicing would be lost). Therefore we propose that factors are highly responsible for the recognition of the donor and acceptor sites (Smith and Valcarcel, 2000). It is natural to expect that the various factors must be significantly less in number than the total number of splice sites. In other words, we expect one protein factor to work for a number of different (but related) splice sites. We believe that this information is already present at the splice sites and the binding of a given factor is determined by the sequence of nucleotides around the splice sites (Rekha and Mitra, 2006). From our earlier work, we believe this region to be 6-8 nucleotides long (Rekha and Mitra, 2007). As different factors work for different splice sites, we cannot hope to find a consensus. However, we can still look for some specific patterns in the nucleotides at the donor and acceptor splice sites (Ladd and Cooper, 2002). If we assume that the available factors responsible for splicing act on a group of sites, then we expect to find the set of nucleotides (at or near the splice sites) to split into several groups (where each group corresponds to one unique factor). This expectation is borne out in this study, which involves the comparative study of sub-sequences at the splice sites.

Recent methods on the purification of spliceosomes coupled with advances in mass spectrometry have suggested that the spliceosome can be composed of ~300 distinct proteins (Jurica and Moore, 2003; Nilsen 2003). These distinct proteins might be involved in recognizing splice sites that are varying in their consensus, suggesting that the splice sites are not actually conserved, but might have some degree of diversity in them. But there can be some unique patterns in the splice sites that are recognized by the spliceosomal complex.

5.1.1 Motivation of the study

The consensus at the splice sites exhibit a lot of diversity in them, which is evident by the number of consensus sequences obtained at each of the donor/acceptor regions in the organisms studied. In spite of the diversity, there might be some unique patterns present in these sub-sequences, which play an important role in the recognition of the splice sites by the spliceosomal proteins. In order to study the distribution of the sub-sequences at the splice sites, we have carried out a comparative study on the datasets of sub-sequences (of size 12) at the donor/acceptor splice sites in five different organisms. We have considered these five organisms (Table 5.1) for our study such that they represent a broad range of species from plants to mammals. We have also carried out frequency and local pairwise alignment studies on these datasets in order to obtain the occurrence of any specific patterns at both the splice sites regions in the given organisms.

5.2 Methodology

5.2.1 Exon-Intron Database (EID)

The EID database, which contains protein-coding intron-containing gene sequences, has been developed from the eukaryotic subset of GenBank (release 112) (Saxonov, 2000). It is a well-organized, extensive and experimental dataset for studying the features of introns and exons and contains the gene sequences of different organisms along with their alternative isoforms. The EID (http://hsc.utoledo.edu/bioinfo/eid/index.html) released in September 2005 was downloaded for the present study, which provides a flat-file distribution of the data (built in FASTA format). We have used the DNA database, which contains the splice sites with "gt…ag" exon-intron boundaries (motifs), that accounts to 98% of all the known motifs. We have selected the gene sequences of five different organisms; such that we can have a broad distribution of the data from plants to mammals, otherwise the choice can be considered arbitrary. The selected organisms are *Arabidopsis thaliana* (plant), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (arthropod), *Gallus gallus* (aves) and *Rattus norvegicus* (mammal). The details of the number of gene sequences and the splice sites considered in the study are tabulated (Table 5.1).

Table 5.1. Number of genes and the sub-sequences at the splice sites of the organisms studied

No	Organism	No. of genes	Total no. of sub-sequences		Total no. of unique sub-sequences*	
			5' ss	3' ss	5' ss	3' ss
1	*A. thaliana*	20,716	130,099	131,229	52,376	69,254
2	*C. elegans*	18,594	111,970	112,361	45,719	28,256
3	*D. melanogaster*	10,612	72,737	73,167	22,213	32,853
4	*G. gallus*	16,567	168,120	169,990	59,742	93,122
5	*R. norvegicus*	19,146	181,782	183,476	56,714	98,629

*An unique sub-sequence is defined as the 12 nucleotide string xxxxx{gt|ag}xxxxx, (where x can be any one of the nucleotides {A C, G, T}, at the donor or acceptor splice site regions) which is not repeating in the given dataset.

5.2.2 Dataset of sub-sequences

We have used the gene sequences of each of the given five organisms for the selection of sub-sequences of size 12 at both the donor and acceptor splice site regions. Size 12 was considered because it was found to be greater than the optimal length required for the recognition of the splice sites by the spliceosomal proteins (Rekha and Mitra, 2007). A dataset of sub-sequences was constructed separately for each of the organisms by aligning the two centrally conserved dinucleotides (gt|ag) of the donor/acceptor splice site regions of all the gene sequences in a given organism and considering five nucleotides flanking the splice sites ($n_1n_2n_3n_4n_5\{gt|ag\}n_8n_9n_{10}n_{11}n_{12}$). Figure 5.1 describes the construction of the datasets at the splice site regions. Details of the number of sub-sequences at the donor and acceptor splice site regions are also tabulated (Table 5.1) as given earlier.

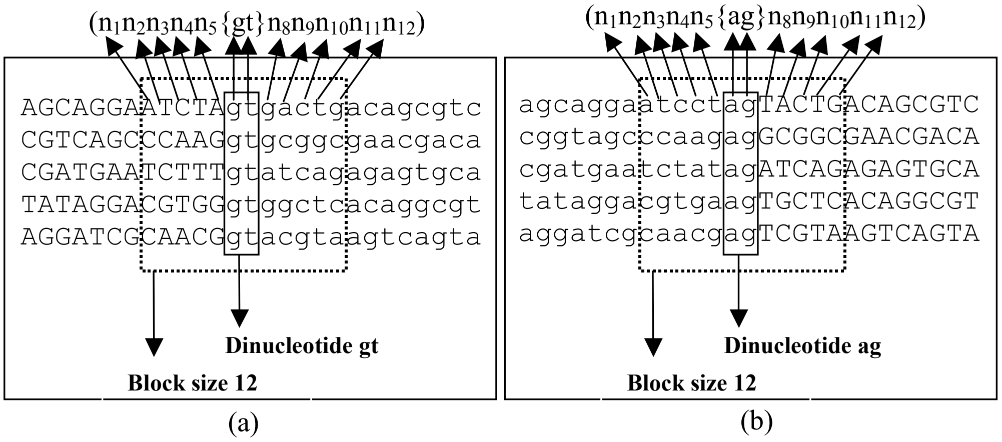

($n_1n_2n_3n_4n_5\{gt\}n_8n_9n_{10}n_{11}n_{12}$)
AGCAGGAATCTAgtgactgacagcgtc
CGTCAGCCCAAGgtgcggcgaacgaca
CGATGAATCTTTgtatcagagagtgca
TATAGGACGTGGgtggctcacaggcgt
AGGATCGCAACGgtacgtaagtcagta
Dinucleotide gt
Block size 12
(a)
($n_1n_2n_3n_4n_5\{ag\}n_8n_9n_{10}n_{11}n_{12}$)
agcaggaatcctagTACTGACAGCGTC
cggtagcccaagagGCGGCGAACGACA
cgatgaatctatagATCAGAGAGTGCA
tataggacgtgaagTGCTCACAGGCGT
aggatcgcaacgagTCGTAAGTCAGTA
Dinucleotide ag
Block size 12
(b)

Figure. 5.1. Illustration of the construction of datasets for the (a) donor and (b) acceptor splice site regions of the organisms studied. The splice sites are represented as donor (gt) and acceptor (ag) regions, in which the two central dinucleotides (gt|ag) are aligned with five nucleotides flanking on both sides. Each dataset was constructed for 12 (gt±5, ag±5) nucleotides. Note that the given sequences are for illustration purpose only and are arbitrary. The exon sequences are represented as uppercase letters, and the intron sequences along with the splice site dinucleotides (gt|ag) are given as lowercase letters. The regions enclosed within the dotted boxes were used for further study.

5.2.3 Frequency analysis of the sub-sequences

Frequency analysis is a simple study, which can be used to obtain information about the data from their frequency distribution. In order to obtain the frequency distribution of the sub-sequences (of size 12) at the splice sites, we have further sorted them separately and calculated the frequency of occurrence of each of the sub-sequence. This gives us the information about the frequency of each of the sub-sequence thus removing the redundancy in the data for further analysis. Each sub-sequence thus obtained is unique being represented only once in the dataset. Thus we have now obtained a dataset, which is much smaller than the earlier one. This way we have reduced the size of each dataset of sub-sequences by 60-70% in the donor sub-sequences and 45-75% in the acceptor sub-sequences of the total size (Table 5.1). We have plotted the unique sub-sequences (on x-axis) against their corresponding frequencies (on y-axis) to obtain the frequency plots for each organism (Figure 5.2), which gives the distribution of the sub-sequences at the splice sites. This degeneracy is not uniform as it is evident from the graphs obtained (Figure 5.2). We have plotted only the first 65,536 motifs of all th*e* organisms in our analysis such that all the graphs are comparable for further discussions (but their actual numbers are given in Table 5.1).

5.2.4 Local pairwise alignment

Local pairwise alignment is an important method to obtain the local similarity between the aligned pair of sequences. From the frequency analysis being carried out, we have obtained a varied number of unique sub-sequences occurring at the splice sites (Table 5.1), thus suggesting the non-random distribution of the unique sub-sequences at the splice sites. These unique sub-sequences must contain some motifs that are conserved in them, which are recognized by the given set of proteins in the spliceosomal complex. It is also important to obtain the frequency distribution of these motifs in order to study the nature of their nucleotide distribution. To study these aspects, we have carried out a local pairwise alignment of all the unique sub-sequences at the splice sites of the given organisms. For this purpose, we have developed a simple algorithm (based on the Smith-Waterman algorithm) for calculating the local pairwise alignment of the sub-sequences at the splice sites. We used a simple scoring model of assigning 1 for every match and 0 for every mismatch for constructing a scoring matrix. We have not introduced any gaps for the alignment and no gap penalties were defined. All motifs whose scores are >=6 were taken into consideration for obtaining the frequency distribution.

5.2.5 Frequency of occurrence of unique motifs

Although the Table 5.1 and Figure 5.2 clearly suggest the presence of degeneracy at the splice site regions, the numbers may be unreliable as we have considered a fixed length of nucleotides (12 in number) centered on the donor and acceptor sites. In earlier studies we note that such an assumption is usually not valid. So, in order to obtain only the unique motifs (locally aligned pairs of size >=6), we have calculated the frequency of occurrence of each of the motifs obtained (Table 5.2).

Table 5.2. Number of locally aligned pairs (motifs) of the organisms studied*

No	Organism	Total no. of motifs		No. of unique motifs	
		5' ss	3' ss	5' ss	3' ss
1	*A. thaliana*	95,426,257	101,989,564	76,529	109,236
2	*C. elegans*	65,068,944	34,706,251	71,381	42,286
3	*D. melanogaster*	25,446,885	25,075,268	35,649	61,600
4	*G. gallus*	111,758,051	152,654,064	88,603	142,957
5	*R. norvegicus*	113,412,019	171,669,729	81,168	149,132

*A local pairwise alignment of all the unique sub-sequences of size 12 was carried out and all locally aligned pairs (motifs) of size >=6 were obtained as given in this table.

This (Table 5.2) gives us the number of unique motifs at the donor and acceptor splice site regions of each of the given organisms. We have plotted the same as vertical bar plots, with unique motifs plotted on x-axis and their corresponding frequencies on y-axis (Figure 5.3). We have plotted only the first 65,536 motifs of all th*e* organisms in our analysis such that all the graphs are comparable for further discussions (but their actual numbers are given in Table 5.2).

5.3 Results and Discussions

5.3.1 Frequency plots of the unique sub-sequences

We note from Table 5.1 that the total number of sub-sequences at the donor/acceptor regions ranges from 72,737 to 183,476 and the number of unique sub-sequences range from 22,213 to 98,629. This shows that the distribution of the sub-sequences at the splice sites is not random, because if it was random then the possibility of occurrence of each of the four nucleotides (A, C, G and T) in each of the given 10 positions (excluding the two central conserved dinucleotides (gt|ag) in a sub-sequence of size 12) would be 4^{10}=1,048,576. But the observed results showed a data, which is much less than expected. We have also observed that the data of unique sub-sequences got reduced by a factor of 60-70% in the donor region and 45-75% in the acceptor region, when compared to the total number of sub-sequences at the splice sites (Table 5.1). This suggests that a lot of redundancy is observed in the sub-sequences (of the plotted data), which show that some unique sub-sequences are occurring more frequently than by random chance. The frequency plots (Figure 5.2) suggests that ~50% of the unique sub-sequences (of the plotted data) are occurring only once and the other 50% (of the plotted data) are having frequencies ranging from one to hundreds. We have restricted the upper limit of the x-axis to 65,536 in the graphs in order to have a good comparison of all the plots. However the actual number has been tabulated (Table 5.1) as given earlier.

From the frequency plots (Figure 5.2) we observe that all the plots show an exponential decay in their frequency distribution (1/*f* distribution), which suggests that, some unique sub-sequences are occurring more common when compared to the other. We have observed that the number of unique sub-sequences at the acceptor region is more than that at the donor, in all the organisms studied except in *C. elegans*. This observation suggests that a single donor sub-sequence might be paired with different acceptor sub-sequences, during alternative splicing. But *C. elegans* show less number of donor sub-sequences than the acceptor, which signify that different sub-sequences at the donor get paired with some common acceptor sub-sequences during the process of splicing.

5.3.2 Occurrence of unique motifs

From our earlier analysis we have identified that the optimal length of the sub-sequences (that are required for the binding of the spliceosomal proteins), was found to be ~6-8 nucleotides at both the splice site regions (Rekha and Mitra, 2007). We have also observed that the information required for splicing is unevenly distributed around the splice sites in the given organisms and the sub-sequences thus identified were found to be of varying length at the splice sites.

By local pairwise alignment, we have obtained all unique motifs of size >=6 (Table 5.2). The frequency of occurrence of each of these motifs were calculated and the number of unique motifs thus obtained were tabulated (Table 5.2). These observations suggest that there is a lot of redundancy in the occurrence of these motifs and their frequency studies have reduced the redundancy by a very large percentage. The number of unique motifs are much less when compared to the total number of motifs, which suggests that there are certain patterns that are conserved in these motifs. The frequency distribution of these unique motifs was also found to be non-random, which signify that some motifs are conserved in them, which are recognized by a given set of proteins. From the local pairwise alignment studies we observe similar trends as observed earlier (Figure 5.2), suggesting the presence of some correlations in them. We also observe (Table 5.2) that the local pairwise aligned sub-sequences are more in the acceptor region than in the donor in all the organisms, except *C. elegans*. We have restricted the upper limit of the x-axis to 65,536 in the graphs in order to have a good comparison of all the plots. However the actual number has been tabulated (Table 5.2) as given earlier.

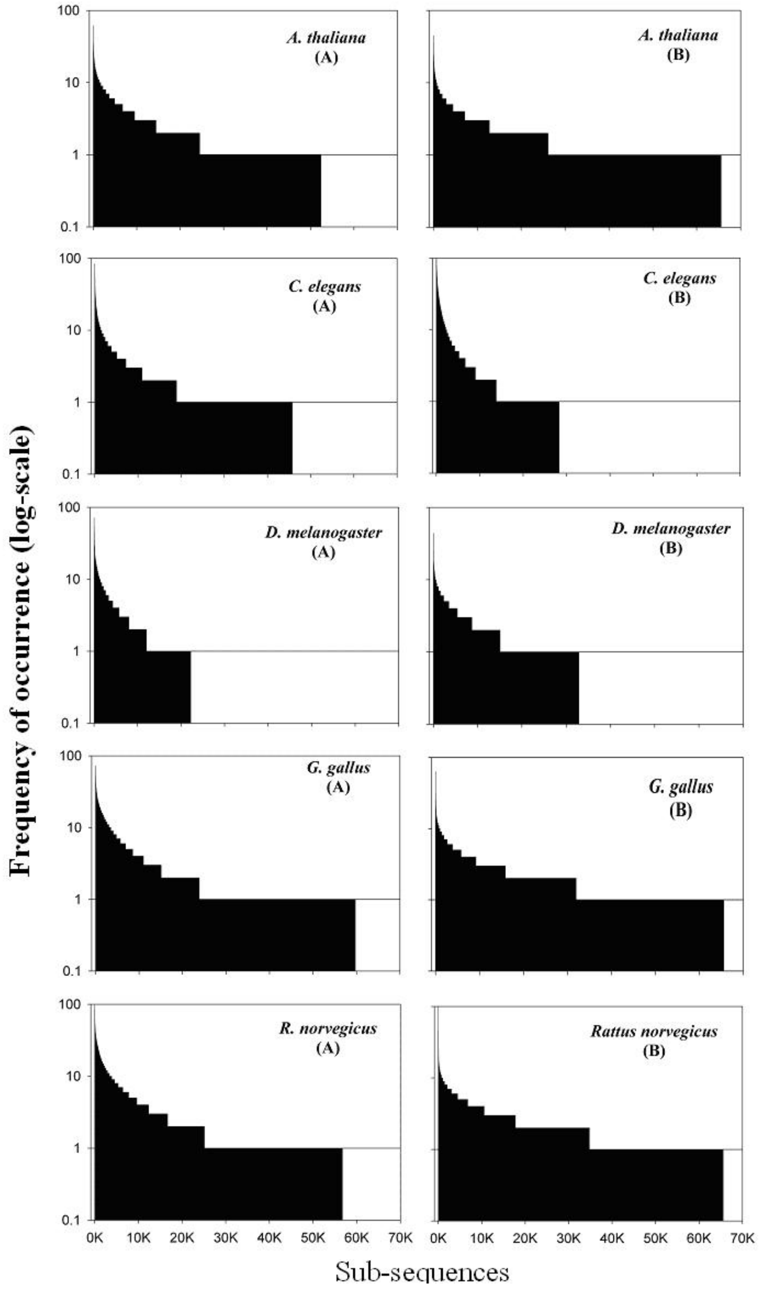


Figure. 5.2. Vertical bar plots of the frequency of occurrence (represented on log-scale, on y-axis) of the unique sub-sequences (arranged in descending order) of size 12 of the respective organisms plotted against their corresponding sub-sequences (represented as numbers on linear scale, on x-axis) for the (A) donor and (B) acceptor splice site regions. Scales of the axes are shown similar for all the organisms for the ease of comparison. The total area in each of the graphs is the same in each organism (first 65,536 sub-sequences are plotted).

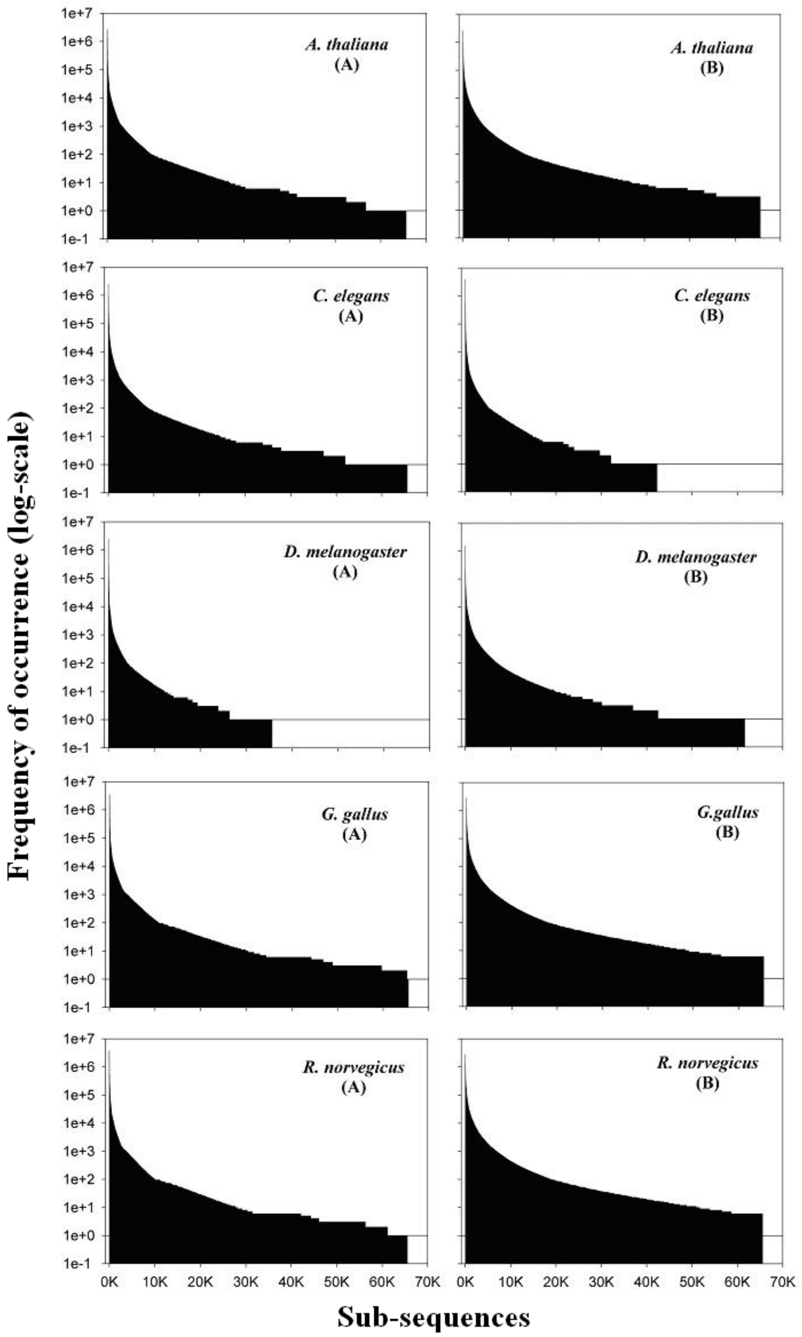


Figure. 5.3. Vertical bar plots of the frequency of occurrence (represented on log-scale, on y-axis) of the unique motifs (obtained by the local pairwise alignment and arranged in descending order), whose size is >=6, which are being plotted against their corresponding motifs (represented as numbers on linear scale, on x-axis) for the (A) donor and (B) acceptor splice site regions of the respective organisms. Scales of the axes are shown similar for all the organisms for the ease of comparison. The total area in each of the graphs is the same in each organism (first 65,536 motifs are plotted).

These plots (Figure 5.3) suggest that even though there are different number of unique sub-sequences at the splice sites, they do contain nucleotides conserved in them, which is evident from the graphs. These motifs are recognized by snRNAs and a complex of spliceosomal proteins collectively called as spliceosomes. In order to accomplish the splicing process efficiently, the complex must assemble and bind the sub-sequences efficiently. But from our analysis, we have observed that the number of these motifs are thousands in number and each of them have to be recognized by the spliceosomal complex. Since the number is more, we expect that either the snRNAs or the spliceosomal proteins might have to show some significant diversity in them. Since a single set of spliceosomal proteins cannot always recognize a diverge set of splice sites; we suggest the occurrence of different sub-sets of the spliceosomal proteins in the given organisms.

Our results are also in agreement of some earlier work on the annotation of the spliceosomal proteins, which suggests that the evolution of novel members of splicing regulatory protein families permitted the diversification of their canonical binding sites in pre-mRNAs, giving the cell the potential to produce new transcripts by altering splice choices (Barbosa-Morais *et al.*, 2006). So, our study has shown that the motifs at the splice sites show some diversity in their distribution.

5.4 Conclusions

This study gives an idea about the distribution of sub-sequences at the splice sites and suggests the presence of certain correlations in them. Local pairwise alignment studies signify the occurrence of some unique motifs in the sub-sequences at the splice sites of different organisms. Since we have observed some diversity in these motifs, we assume the evolution of different spliceosomal protein families might have led to the diversification of the binding sites in the given organisms. We also suggest that the existence of different spliceosomal protein families could also be regulating the level of gene expression by involving spliceosomal proteins specific only to a class of introns.

Summary

Summary

I. 1/*f* Correlations in Viral Genomes - A Fast Fourier Transformation (FFT) Study

- The genomes of the viruses analyzed represented slope values significant in specifying the presence of long-range correlations in their genome organization.
- It was suggested that long-range correlations represent an efficient trade-off between efficient information storage and protection against error in genetic code.
- It was also suggested that the genes are having correlations within themselves and the correlation lengths do not span the complete genome but certainly cross the gene boundaries.
- It was noted that the high frequency end of the 1/*f* region is correlated with the distribution of the gene sizes, and the correlation appears to be related with the scaling exponent α.
- We can say from our study, that the DNA of the viruses are showing 1/*f* noise in their power spectrum, irrespective of the absence of introns in their genomes.
- Thus, we assume that 1/*f* noise is not a feature of intron containing regions, but is also exhibited by coding regions of the genome.
- We suggest that the occurrence of long-range correlations in exonic sequences might signify the existence of correlation structures (patterns) in the gene structure.

II. Comparative Analysis of the Splice Site Regions by Information Theory

- The study suggests that even though the nucleotides are showing some degrees of conservation in the flanking regions of the splice sites (gt/ag), there still exists a certain level of variability in the consensus of the different organisms studied.

- This signify that some substitutions are found to be tolerable at certain positions of the splice site junctions, which is presumed to respond to the different spliceosomal factors that lead the splicing process to occur selectively.
- Our study suggests that the information required for RNA splicing is mainly contained in the consensus of ~6-8 nt at both the donor and acceptor regions, which are important for the binding of spliceosomal proteins to the splice sites as expected.
- Our study gives a broad idea about the distribution of nucleotides at/around the splice sites and also gives a comparative analysis of the consensus sequences at both donor and acceptor regions of the splice sites in different organisms.
- All the organisms studied except C. elegans show similar kind of patterns in their genome architecture at/around the splice sites.

III. <u>Frequency Analysis of the Splice Site Regions in Different Organisms</u>

- The study suggests that the information required for splicing is contained in ~6-8 nucleotides (nts) at/around both the donor and acceptor splice sites.
- This work has given a better idea about the distribution of information at/around the splice sites suggesting that the sub-sequences (at the splice sites) are highly variable.
- The frequency analysis of the unique sub-sequences suggests that the distribution is approximately exponential and the percentage of occurrence (uniqueness) values suggests that sub-sequences with the highest values are the ones, which are highly involved in splicing.

- We note that the length of 6-8 nts with six bases in intron (including the two central, conserved dinucleotides (gt|ag)) and two bases in exon is optimal for the efficient assembly and binding of the spliceosomal complex during the process of splicing.
- We also suggest that the donor sub-sequences are more crucial in pairing with the corresponding acceptor sub-sequences during the process of splicing.
- This study also revealed that the all the organisms studied except *C. elegans* showed similar kind of patterns in their genome organization.

IV. <u>Frequency Studies of the Unique Sub-sequences at the Splice Sites</u>

- The study suggested the presence of certain correlations in them.
- Local pairwise alignment studies were also carried out, which signify the occurrence of some unique motifs in these sub-sequences at the splice sites.
- Since we have observed some diversity in these motifs, we assume the evolution of different spliceosomal protein families might have led to the diversification of the binding sites in the given organisms.
- We suggests that the existence of different spliceosomal protein families in these organisms could be regulating the level of gene expression by involving spliceosomal proteins specific only to a class of introns.

The rich variability of the donor and acceptor sites generates greater information, which can be useful in understanding the language of DNA at the splice sites. Further this idea can be extended in decoding the information at the splice sites into distinct groups and classes. However, we have clearly identified a number of broad features that can help in this direction, but considerable experiments need to be carried out before the problem can be uniquely solved. This kind of work can also be carried in understanding the information contained in the promoter regions, which might give some insights into the underlying mechanism.

Publications

Indian Journal of Biochemistry & Biophysics
Vol. 43, June 2006, pp. 137-142

1/*f* Correlations in viral genomes – A Fast-Fourier Transformation (FFT) study

T Shashi Rekha and Chanchal K Mitra*

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

Received 29 September 2005; revised 05 April 2006

We have studied the presence of long-range correlations in the complete genomes of ten different dsDNA viruses and *Saccharomyces cerevisiae* (bakers' yeast) chromosome I. We have also studied the correlation between the distribution of the gene length and the domain of "1/*f* region" of their genomes. Linear regression analysis was done for the power-law region of these organisms and the slope values obtained were ~ -1, which signify the existence of "1/*f* noise" in the low and medium (intermediate) frequency regions. This suggests the presence of long-range correlations in their genomes. The presence of 1/*f* noise in a given frequency interval indicates the existence of a fractal (self-similar) structure in the corresponding range of wavelengths. The results of our study suggest that genes have correlations within themselves, and the correlations appear to be related with the scaling exponent α.

Keywords: 1/*f* noise, DNA sequence, Power spectrum, Long-range correlations.

Complete genome sequences are being generated far more rapidly than our ability to interpret and comprehend the biological meaning of the data. The availability of the complete genome sequences of different organisms paves the way to study the various characteristic features of genome organization and structure. Prokaryotic and eukaryotic genomes have been studied by various methods including autocorrelation function analysis[1], DNA walking[2], Fourier spectral analysis[3,4], mutual information function and wavelet translation. Studies on the complete genomes of different microbes have elucidated the fractal characteristics of DNA[5-7].

Statistical analysis is an important tool for studying the structural and functional characteristics of DNA. Complex systems such as DNA, which shows non-linear behavior, have structures at large scales, resulting in statistical patterns like long-range correlations. The long-range correlations, which follow power-law correlation function, are a result of the dependency of a single nucleotide on all other nucleotides over large distances and are found to be scale invariant. The possibility of existence of these patterns, hidden in DNA base sequence is demonstrated by one-over-f (1/*f* noise) spectra. Power-law decay for the correlations as a function of time translates into a power-law decay of the spectrum as a function of frequency, which is called "1/*f* noise". The power spectra *S(f)* as a function of frequency *f* behaves like: $S(f) = 1/f^{\alpha}$, where the exponent α is close to 1. Spectral representation of a DNA sequence has the following applications,

- To identify underlying periodic patterns in the sequence, which manifest as peaks at specific frequencies in the power spectrum.
- To ascertain whether a sequence is random lacking any correlation pattern exhibiting flat power spectrum.

Earlier studies suggested that intronic sequences have long-range correlations signifying the power-law correlation function. It was proposed that 1/*f* behavior could be explained by "expansion-modification" model, according to which, the length of the DNA of the present day organisms is longer than the pre-biotic ones. And during course of evolution, the DNA had undergone elongation by repeated process of duplication, followed by mutation, which lead to these long-range correlations[8,9].

We have studied the complete genome sequences of dsDNA viruses to identify the presence of long-range correlations in intron-less DNA using FFT approach. We have taken the complete DNA sequence of chromosome I of *S. cerevisiae* with intronic sequences as a control for our study, which is found to exhibit long-range correlations in them. For this study, we have obtained the complete DNA (RefSeq) sequences of ten different dsDNA viruses and the chromosome I of *S. cerevisiae*, from GenBank

*Author for correspondence
Tel: (091) (40) 23134668
Fax: (091) (40) 23010120
E-mail: c_mitra@yahoo.com

database accessed through National Center for Biotechnology Information (NCBI)[10]. The rationale to study these viruses is that all are dsDNA viruses, with their genome length in the range of 0.1-0.3 mbps and the host they infect range from invertebrates to vertebrates. We have also made an attempt to study the correlation between the distribution of gene length and the domain of 1/*f* region in these organisms.

Methods

Power spectral analysis

For power spectrum analysis, we calculated the "adenine plus guanine" (A + G) and "adenine plus thymine" (A + T) proportions of the complete genomes of each of the organisms in a non-overlapping window of size 32, considering the proportions is a better method than that used in earlier studies. Results were also compared using window sizes of 64 and 128 bases. A small window size gives us more information at a lower scale, but a larger window provides better signal to noise (because of averaging over a longer region). After comparing the graphs for three base sizes, we used the window size of 32 in all the following studies. These proportions were Fourier-transformed using the Fast-Fourier-Transform (FFT) algorithm, which calculates the discrete-Fourier transformation (DFT) of a function of N points. This algorithm speeds up the calculation of power spectrum by a factor of $N\log_2 N$, but requires the length of sequence being analyzed to be an integral power of two[11]. We calculated the respective proportions of the entire length of the genomes and zero padded it to the next integral power of two.

The DFT (H_n) of a function from a finite number of N sampled points h_k is given as in Eq. (1).

$$H_n = \sum_{k=0}^{N-1} \mathrm{h}e^{2\pi ikn/N} \qquad \ldots(1)$$

DFT maps N complex numbers h_k's into N complex numbers (H_n's). Eq. (1) is periodic in n, with a period N. Variables n and k vary from 0 to N-1 and $i^2 = -1$. The power for each of the proportions was calculated by taking the square root of sum of squares of real and imaginary components of the Fourier amplitude, given as in Eq. (2):

$$\left|H_n(f_j)\right| = \sqrt{((H_n^{real}(f_j))^2 + (H_n^{imag}(f_j))^2)} \qquad \ldots(2)$$

The power spectrum was visually divided into two linear parts and each section was independently fitted with a linear regression line (using the built-in function of SigmaPlot plotting package). The slope values (-α) are reported in Table 3 for comparison. These values were correlated with the gene length distribution graphically.

Calculation of average gene length

The correlation between the distribution of gene length and the domain of 1/*f* region of power spectrum was examined by extracting the average gene length of each species from the same database.

Results and Discussion

The details of the genomes analyzed in the present study are given in Table 1. The log-log plot of Fourier-transforms (Fig. 1A and B) for the genomes of *Amsacta moorei* entomopoxvirus (a, b and c), human herpesvirus 4 (d, e and f) and chromosome I of *S. cerevisiae* (g, h and i), for A+G and A+T proportions in a window size of 32, 64 and 128 shows the presence of two distinct regions, a power-law and a flat region in their respective power spectra (to conserve space, graphical presentation of only three organisms has been given).

In comparative analysis using window sizes of 32, 64 and 128 bases, the high frequency end at a frequency of 0.5 (corresponding to an angular frequency π) is unaltered, but the low frequency end has moved towards the high frequency range. This is because, as we increase the window size, the number of points contributing to low frequency is being reduced, but the 1/*f* pattern observed in all is the same. The noise reduces by a factor of $\mathrm{N}^{-0.5}$ $(\mathrm{N}^{-0.5} = \mathrm{N}^{-1/2} = 1/\mathrm{N}^{1/2} = 1/\sqrt{\mathrm{N}})$, so the signal gets doubled in the window size of 128, when compared to the window size of 32. For the larger window size, intensity of the peaks gets reduced, as

Table 1—Genomes analyzed in the present study

	Organisms	NCBI Accession no.	Genome length (bp)
1	*Amsacta moorei* entomopoxvirus	NC_002520	232392
2	*Cercopithecine* herpesvirus 17 (B-virus)	NC_003401	133719
3	*Ectocarpus siliculosus* virus	NC_002687	335593
4	Human herpesvirus 4 (Epstein Barr virus – EBV)	NC_001345	172281
5	Human herpesvirus 6	NC_001664	159321
6	Human herpesvirus 6B	NC_000898	162114
7	Lymphocystis disease virus 1	NC_001824	102653
8	Meleagrid herpesvirus 1	NC_002641	159160
9	*Saccharomyces cerevisiae*	NC_004002	149955
10	Sheeppox virus	NC_002642	144575
11	Yaba-like disease virus	NC_001133	230208

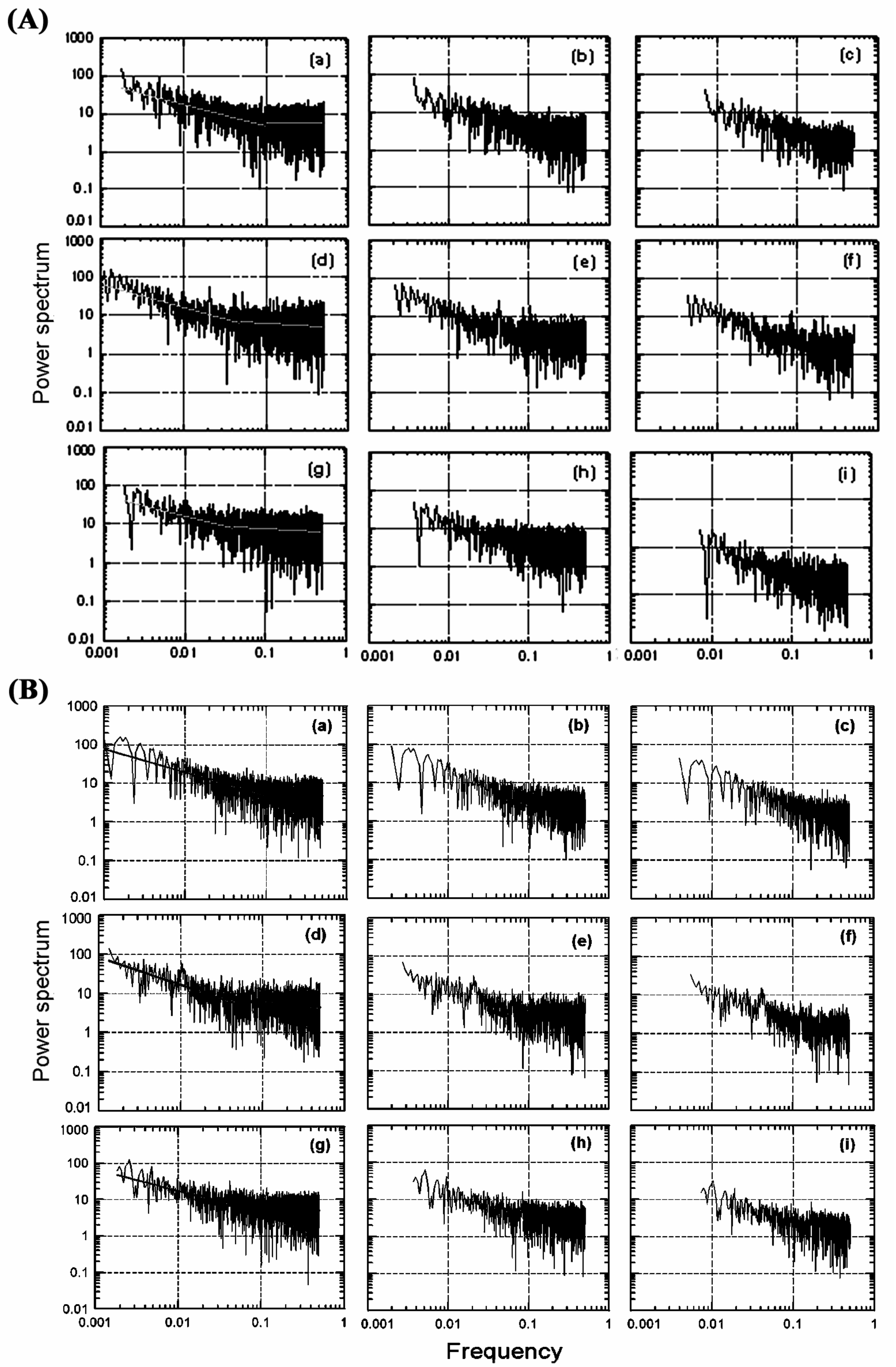


Fig. 1—Log-log plots of the power spectrum of A+G proportions (Fig. 1A) and A+T proportions (Fig. 1B) contained in a non-overlapping window of size 32 (a, d and g), 64 (b, e and h) and 128 (c, f, and i) bases for *A. moorei* entomopoxvirus (a, b and c), human herpesvirus 4 (d, e and f) and *S. cerevisiae* chromosome I (g, h and i) with the linear lines of regression for the plots of window size 32 (solid lines in the graph) [For ease of comparison, scales of both axes are set to be same for all the plots]

Table 2—Range of 1/*f* region of power spectrum of genomes studied

	Organisms	A+G Proportion		A+T Proportion	
		Range of 1/*f* region (No. of nucleotides)	% of 1/*f* region to the total spectrum	Range of 1/*f* region (No. of nucleotides)	% of 1/*f* region to the total spectrum
1	*Amsacta moorei* entomopoxvirus	26214	10	26214	10
2	*Cercopithecine* herpesvirus 17	7864	3	13107	5
3	*Ectocarpus siliculosus* virus	10486	2	15729	3
4	Human herpesvirus 4	13107	5	13107	5
5	Human herpesvirus 6	7864	3	13107	5
6	Human herpesvirus 6B	13107	5	13107	5
7	Lymphocystis disease virus 1	7877	6	10499	8
8	Meleagrid herpesvirus 1	7864	3	10486	4
9	*Saccharomyces cerevisiae*	10486	4	7864	3
10	Sheeppox virus	10486	4	10486	4
11	Yaba-like disease virus	10486	4	13107	5

Table 3—Slopes values (-α) obtained by linear regression analysis of power spectrum of genomes studied

	Organisms	Slope values of 1/*f* noise		Slope values of flat noise	
		A+G proportions	A+T proportions	A+G proportions	A+T proportions
1	*Amsacta moorei* entomopoxvirus	-0.56	-0.60	-0.01	-0.02
2	*Cercopithecine* herpesvirus 17	-0.88	-0.67	-0.10	-0.12
3	*Ectocarpus siliculosus* virus	-0.70	-0.58	-0.11	-0.13
4	Human herpesvirus 4	-0.62	-0.70	-0.09	-0.18
5	Human herpesvirus 6	-0.67	-0.55	-0.11	-0.11
6	Human herpesvirus 6B	-0.54	-0.53	-0.07	-0.11
7	Lymphocystis disease virus 1	-0.69	-0.63	-0.02	-0.07
8	Meleagrid herpesvirus 1	-0.79	-0.62	-0.05	-0.07
9	*Saccharomyces cerevisiae*	-0.67	-0.96	-0.03	-0.14
10	Sheeppox virus	-0.64	-0.76	-0.02	-0.10
11	Yaba-like disease virus	-0.53	-0.59	-0.13	-0.10

the total number of points is reduced, but the pattern observed would remain essentially the same, even though the fine structure is affected. The range of 1/*f* region for the power spectrum of each of the organisms (Table 2) is found to be correlated with the gene sizes. The range or the percentage of 1/*f* region for the A+G and A+T proportions of sequence nos 2, 3, 5, 7, 8, 10 and 11 are not the same, which signify that the two proportions are not showing same scaling behaviour.

Linear regression analysis was done separately for the two distinct regions ("1/*f* noise" and "white noise") and the significant slope values (> -0.5) were obtained (Table 3), suggesting the presence of long-range correlations in their respective genomes. Linear lines of regression for *A. moorei* entomopoxvirus [(Fig. 1A(a) and 1B(a)], human herpesvirus 4 [(Fig. 1A(d) and 1B(d)] and *S. cerevisiae* chromosome I (Fig. 1A(g) and 1B(g)) are also shown.

The slope values of the 1/*f* region for A+G power spectrum ranged from -0.53 to -0.88 and for A+T spectrum from -0.53 to -0.96. Our studies have shown that the behaviour of the power spectrum in the low and intermediate regions represented power-law decay, with the slope values close to -1, indicating the presence of long-range correlations in those regions. We can say that greater the exponent, larger is the correlation. Different slopes (differences > 0.1) for (A+G)/(A+T) proportions, are shown in the 1/*f* region for sequence nos 2, 3, 5, 8, 10 and 11 as given in Table 3. This observation signifies that the two proportions are not showing same scaling behavior for long-range correlations. Slope is close to -1, for sequence no. 9 for A+T proportion and also for sequence no. 2 for A+G proportions. For Sheeppox virus, power spectrum of A+T proportion shows high correlation, but for *Cercopithecine* herpesvirus 17, A+G proportion is showing high correlation. Power law with a negative value of exponent suggests that a system producing 1/*f* type noise is scale invariant and has long-range time and spatial correlations as a consequence.

The average gene lengths of each of the organisms considered for the study were obtained from the same database (Table 4). We note a correlation between the distribution of the gene length and the domain of "1/*f* region" of the power spectrum. We also note (Table 4), that average gene lengths of *A. moorei* entomopoxvirus, human herpesvirus 4 and *S. cerevisiae* are in the increasing order, which is inversely proportional to the end of 1/*f* region (as shown in Fig. 1A) of power spectra. The 1/*f* region (Fig. 1A and 1B) of *A. moorei* entomopoxvirus extends up to 0.1 of angular frequency, signifying the presence of mid to long-range correlations for genes of smaller gene length. Whereas for human herpesvirus 4 and *S. cerevisiae*, the 1/*f* region extends only up to 0.05, which signify the presence of long-range correlations for genes of longer gene length.

Histograms were plotted (Fig. 2) to show the absolute frequencies of the average gene length of all the genomes studied, in a semi-log scale. We note that the genes of the organisms analyzed are not all of the same size (Fig. 2). The long genes are few in number and smaller ones are more in number, as is apparent from the histograms (Fig. 2). Also, the more number of small genes in the DNA are expected to increase mid-range correlation length. This shows that genes are having correlations within themselves; however,

Table 4—Average gene length of the studied organisms

	Organisms	Average gene length (bp)
1	*Amsacta moorei* entomopoxvirus	800
2	*Cercopithecine* herpesvirus 17	1500
3	*Ectocarpus siliculosus* virus	1000
4	Human herpesvirus 4	1500
5	Human herpesvirus 6	1300
6	Human herpesvirus 6B	2900
7	Lymphocystis disease virus 1	900
8	Meleagrid herpesvirus 1	1700
9	*Saccharomyces cerevisiae*	1100
10	Sheeppox virus	1100
11	Yaba-like disease virus	1600

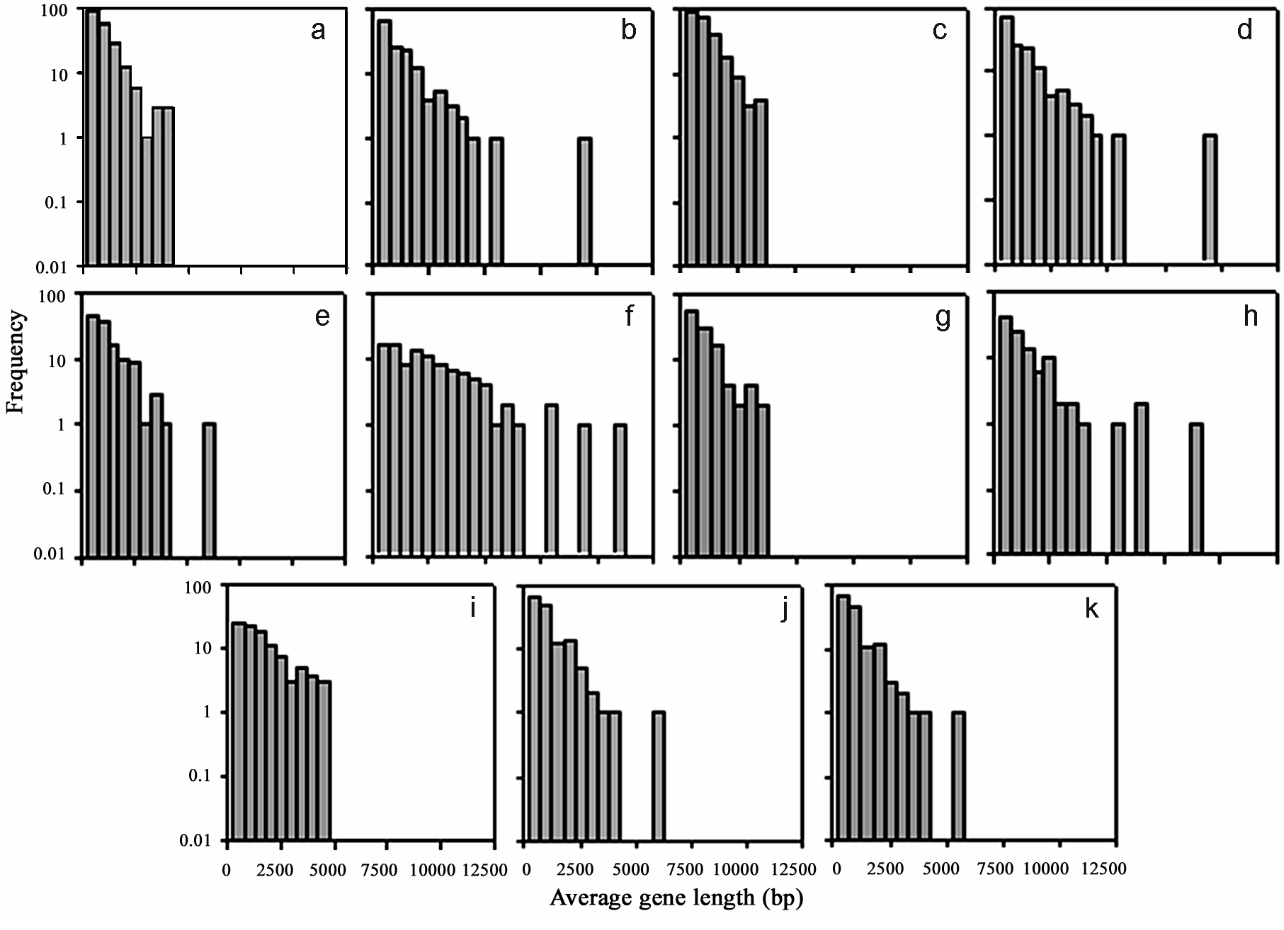


Fig. 2—Graphs showing the absolute frequencies of average gene length of the genomes (in alphabetical order from a - k as given in the order for all the tables) in a semi-log scale [Scales are identical to facilitate visual comparison]

correlations between the genes are not ruled out as the 1/*f* region covers a scale greater than the longest gene. We also observed that the heights of the vertical bars in the histograms are approximately in exponential decrease; however, the rate of decrease is not the same in all the genomes. To emphasize this fact, we have plotted the graphs in a semi-log scale (Fig. 2).

Conclusions

The genomes of the organisms analyzed represent slope values significant in specifying the presence of long-range correlations in their genome organization. It is proposed that the long-range correlations represent an efficient trade-off between efficient information storage and protection against error in genetic code. We conclude that genes are having correlations within themselves and the correlation lengths do not span the complete genome but certainly cross the gene boundaries. On a short length scale, the genes appear noisy, because of lack of short-range correlation. It was also noted that the high frequency end of the 1/*f* region is correlated with the distribution of the gene sizes, and the correlation appears to be related with the scaling exponent α.

As all the DNA sequences studied in this report lack introns (except *S. cerevisiae*) therefore, only very short intergenic regions are expected, which may also contribute to the white noise part of the spectrum. However, *S. cerevisiae* has introns, but this is not apparent in the spectrum either. So, from our studies, we can say that the DNA of the viruses are showing1/*f* noise in their power spectrum, irrespective of the absence of introns in their genomes. Thus, we can assume that 1/*f* noise is not a feature of intron containing regions, but is also exhibited by coding regions of the genome. The occurrence of long-range correlations in exonic sequences might signify the existence of correlation structures or patterns in the gene structure.

Acknowledgement

One of the authors (SRT) thanks Council of Scientific and Industrial Research (CSIR), Govt. of India, N. Delhi for award of Junior Research Fellowship. The authors also thank the anonymous referee who has painstakingly gone through the manuscript and has made very constructive suggestions.

References

1 Herzel H, Weiss O & Trifonov E N (1999) *Bioinformatics* 15, 187-193

2 Nee S (1992) *Nature* 357, 450

3 Fukushima A, Ikemura T, Kinouchi M, Oshima T, Kudo Y, Mori H & Kanaya S (2002) *Gene* 300, 203-211

4 Fukushima A, Ikemura T, Kinouchi M, Kudo Y, Kanaya S, Mori H & Ikemura T (2001) *Genome Informatics* 12, 435-436

5 Vieira M S (1999) *Phys Rev E* 60, 5, 5932-5937

6 Surya Pavan Y & Mitra C K (2005) *Indian J Biochem Biophys* 42, 141-144

7 Upadyay R K (2003) *Indian J Biochem Biophys* 40, 51-58

8 Li W & Kaneko K (1992) *Nature* 360, 635-636

9 Li W & Kaneko K (1992) *Europhys Lett* 17, 655-660

10 http://www.ncbi.nlm.nih.gov.

11 Press W H, Teukolsky S A, Vetterling W T & Flannery B P (1992) *Numerical Recipes in C,* 12, pp. 496-536 and pp. 537-608, Cambridge University Press

Comparative Analysis of Splice Site Regions by Information Content

T. Shashi Rekha and Chanchal K. Mitra*

Department of Biochemistry, University of Hyderabad, Hyderabad 500046, India.

We have applied concepts from information theory for a comparative analysis of donor (gt) and acceptor (ag) splice site regions in the genes of five different organisms by calculating their mutual information content (relative entropy) over a selected block of nucleotides. A similar pattern that the information content decreases as the block size increases was observed for both regions in all the organisms studied. This result suggests that the information required for splicing might be contained in the consensus of ∼6–8 nt at both regions. We assume from our study that even though the nucleotides are showing some degrees of conservation in the flanking regions of the splice sites, certain level of variability is still tolerated, which leads the splicing process to occur normally even if the extent of base pairing is not fully satisfied. We also suggest that this variability can be compensated by recognizing different splice sites with different spliceosomal factors.

Key words: splice site, substitution matrix, mutual information content, relative entropy

Introduction

Eukaryotes undergo the process of “RNA splicing”, which involves the splicing of introns from heterogenous RNA (hnRNA or pre-mRNA) to form mature mRNA. 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. In addition to these dimers, a pyrimidine-rich region precedes AG at the acceptor site, and a short consensus follows GT at the donor site, while a very weak consensus appears at the branch point ∼30 nucleotides (nt) upstream of the acceptor site. A complex of nucleotide binding proteins and small nuclear RNAs (snRNAs), collectively known as the “spliceosome”, recognizes these splice sites and excises introns by a concerted transesterification reaction (*1*). One important consequence of RNA splicing is that one gene can produce several different mRNA variants, or isoforms, simply by joining together different combinations of exons.

Several earlier studies have been reported for the detection of splice sites using different methods, such as the weight matrix model that uses the position compositional biases in splice sites (*2*). Artificial neural networks have been applied for the prediction of splice sites in different organisms with confidence levels better than previous methods (*3*). However, the reported results should be interpreted with caution as they were based on small datasets of limited number. A computational tool, GeneSplicer (*4*), was developed based on maximum dependence decomposition and performed better than previous tools. Recently the prediction of splice sites with dependency graphs and their expanded Bayesian networks has gained much importance because of its better performance (*5*). Current studies are being carried out to further understand and interpret the information contained in splice sites, as well as to develop a better method for their prediction with better specificity and sensitivity.

Detection of splice sites by using the two dinucleotides (GT/AG) is not meaningful because the frequency of these dinucleotides is very high in genes. Another important aspect to be considered is that the bases flanking them are also involved in the process and are expected to contain information required for splicing. Studying the consensus is also not directly useful, as they are highly variable not only within the species but also between species. Therefore, information theory comes to play a major role for the study of splice sites, which gives a quantitative measure of sequence conservation (or variability).

Information theory is an important tool (*6*) that has been often applied for understanding several key concepts in molecular biology (*7*). Information is defined as the amount of correlation between two random variables (X and Y), which is measured as the

***Corresponding author.**
E-mail: c_mitra@yahoo.com

amount of entropy (uncertainty in a random variable) shared by them. This shared entropy is the information that one random variable contains about the other. It is a relative entity and is never absolute. In other words, mutual information is defined as a measure of the amount of information that one random variable contains about the other. It measures exactly the amount by which the entropy of X or Y is reduced by knowing the other, Y or X (*8*). This theory has gained much importance in biology by its applications to measure the information content of the nucleotide binding sites (*9*), identification of polymorphisms in DNA (*10*), prediction of RNA and protein secondary structures (*11*), prediction and analysis of molecular interactions (*12*), and drug design (*13*).

Study of horizontal correlations (between nucleotides along a sequence) is useful to identify features that can distinguish coding and non-coding regions in DNA (*14*). This gives the probability of finding nucleotides in the sequence that are correlated with each other. On the other hand, vertical correlations are important to find the probability of a nucleotide at a particular site by calculating the information content of the aligned set of sequences from its frequency of occurrence. Substitution matrices are thus useful to score these alignments perfectly.

Substitution matrix is a useful tool that scores the similarity between any two nucleic acid bases in terms of their ability to replace each other. By comparing a large number of similar sequences, one can obtain a matrix that describes the probability of a given nucleotide being substituted by another under the conditions of study. As probabilities are multiplicative, the logarithm is used to get an additive formulation. A number of techniques are now available for direct computations of substitution matrices, such as the BLOSUM (blocks substitution matrix) (*15*) and PAM (point accepted mutation) matrices (*16*). These matrices have been used extensively for global and local sequence alignments as well as database searches (*17*). They were also found to be significant for the study of core promoter regions (*18*).

Information theory has also been used for studying the features of spliceosome evolution and function (*19*). Studies have been carried out to correlate the intron length and the information content of the splice sites (*20*), suggesting that longer introns contain more information than shorter ones (*21*). Recently a comprehensive splice-site analysis using comparative genomics has been performed on different organisms by using the information content of the splice-site motifs, which proves that the identification of broad patterns in naturally-occurring splice sites, through the analysis of genomic datasets, provides mechanistic and evolutionary insights into pre-mRNA splicing (*22*).

It has become an important topic of research to characterize signals that govern the process of splicing in different organisms by information theory, which gives a broad idea about the distribution of information around the splice sites in different organisms. We have studied this aspect by carrying out a comparative analysis of donor and acceptor splice site regions in the genes of five different organisms (Table 1). We have constructed substitution matrices for the aligned set of sequences in the blocks of 6, 10, and 14 nt around the consensus dinucleotides (gt/ag) and calculated their information content, respectively (Figure 1). The substitution matrix specifically constructed for a given block is expected to work more efficiently than the one constructed for the whole genome sequences. In fact, we expect the difference to be evident among the three block databases. We have performed a broad analysis of the data distribution by calculating the information content at/around the splice sites, and achieved some interesting and informative results.

Table 1 The Number of Genes and Splice Sites of the Five Organisms Studied*

No.	Organism	No. of genes	Total No. of genes#	No. of splice sites		Exon/intron boundaries
				Donor	Acceptor	
1	*Arabidopsis thaliana*	20,716	22,957	130,099	131,229	gt-ag
2	*Caenorhabditis elegans*	18,594	20,470	111,970	112,361	gt-ag
3	*Drosophila melanogaster*	10,612	15,624	72,737	73,167	gt-ag
4	*Gallus gallus*	16,567	16,568	168,120	169,990	gt-ag
5	*Rattus norvegicus*	19,146	19,197	181,782	183,476	gt-ag

*The splice sites with only “gt-ag” exon/intron boundaries were considered in our analysis. All other splice sites such as “gc-ag”, “at-ac”, and all the cryptic ones were excluded in the present study. However, we have included all the alternative splice sites in our analysis. #The total number of genes including alternative isoforms.

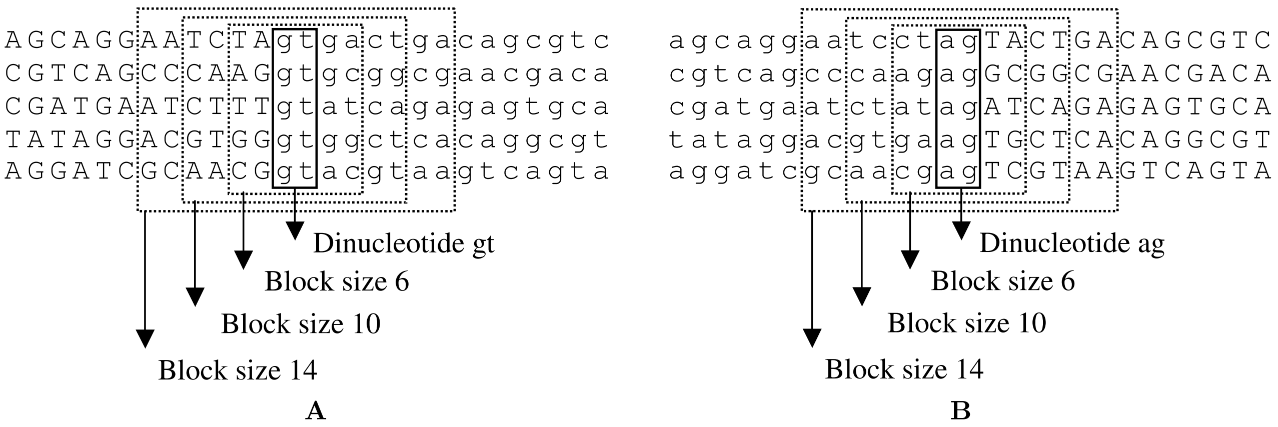

AGCAGGAATCTAgtgactgacagcgtc
CGTCAGCCCAAGgtgcggcgaacgaca
CGATGAATCTTTgtatcagagagtgca
TATAGGACGTGGgtggctcacaggcgt
AGGATCGCAACGgtacgtaagtcagta
Dinucleotide gt
Block size 6
Block size 10
Block size 14
A
agcaggaatcctagTACTGACAGCGTC
cgtcagcccaagagGCGGCGAACGACA
cgatgaatctatagATCAGAGAGTGCA
tataggacgtgaagTGCTCACAGGCGT
aggatcgcaacgagTCGTAAGTCAGTA
Dinucleotide ag
Block size 6
Block size 10
Block size 14
B

Fig. 1 Illustrations of the construction of three different block databases for donor (**A**) and acceptor (**B**) splice sites. The splice sites are represented as donor (gt) and acceptor (ag) sites and the central dinucleotides (gt/ag) are aligned with 2, 4, or 6 nt taken on both sides. The three blocks are constructed for 6 (gt±2, ag±2), 10 (gt±4, ag±4), and 14 (gt±6, ag±6) nt, respectively. Note that the given sequences are for illustration only and are arbitrary. The exon sequences are represented as uppercase letters, and the intron sequences along with the splice site dinucleotides are given as lowercase letters. The regions enclosed within the boxes are used for the computations of the substitution matrices.

Results and Discussion

We calculated the mutual information content (relative entropy H) for each of the organisms studied from their log-odds matrices. The log-odds matrices scoring the alignments of the mononucleotide substitutions were obtained from the substitution matrices constructed for the frequency of occurrence of the nucleotide pairs. The information content values for the three blocks of all the five organisms studied are plotted as vertical box plots for both donor and acceptor sites (Figure 2). The 16 elements (4×4) of the H matrix are plotted to get each box plot. These elements are the mean values of the given block and are directly comparable. Therefore, we are able to identify the contribution of the various elements individually.

The information content derived in this way is obviously a gross feature of the organism and perhaps can be divided into several groups such that the correlations within the groups are much more significant (compared to the whole genome; we expect the correlations between such groups may be quite less). The present plots in Figure 2 are more informative as they show a better distribution of the given data. We can clearly see the trends by following the median or the other percentiles. In all the plots we note that the 90 percentile bars are far from the median, suggesting that few points have relatively high values. The data points with high values were then examined manually and correlated with the particular elements of the H_{ij} matrix as given in Table 2.

The box plots for the donor and acceptor sites of all the organisms studied (Figure 2) show interesting aspects that otherwise cannot be observed in the histograms (computed from the sum of H_{ij} matrix elements) of the average mutual information content. We can see that the information content (the height of the box) decreases with the increasing block size for both donor and acceptor regions in all the organisms studied, suggesting that the distribution of nucleotides around the splice site junctions is more conserved (that is, the splice sites are more variable compared to the neighboring regions). The 6-nt block has the highest information content, and the information reduces considerably as we move away from the splice site. We speculate that the 6-nt block shows a greater variability (higher information content) and hence a higher selectivity. As we move to a larger window size, the variability decreases accordingly (as expected), suggesting that the selectivity of the spliceosomal binding is mainly dictated by the immediate neighborhood of the splice sites. This result reveals that the nucleotides of ~2–3 nt flanking both sides of the splice sites are more important than longer distance nucleotides.

We also find that the median (50 percentile) values are more or less equal for all the plots. There exists a similar pattern of information content for both donor

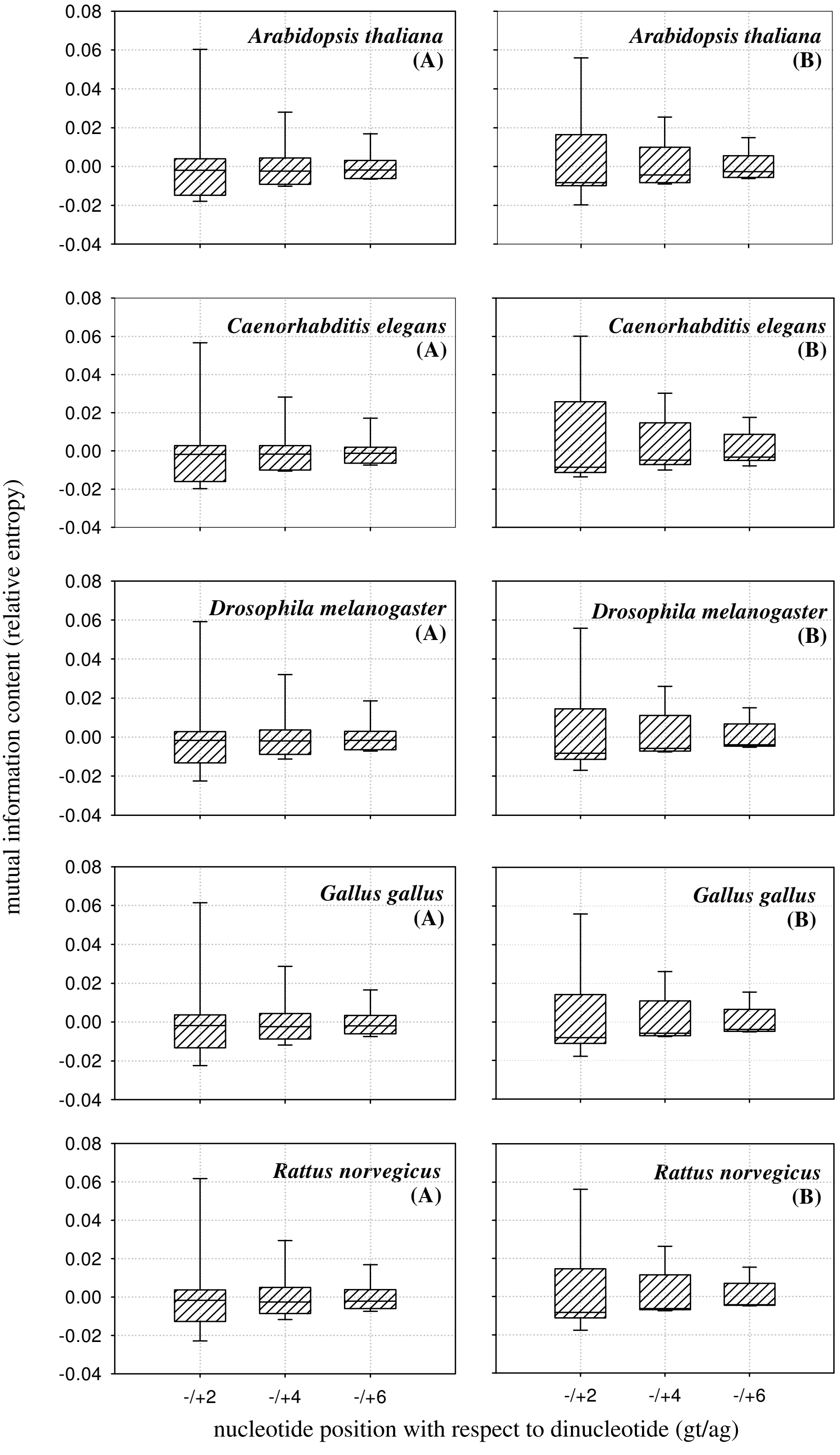


Fig. 2 The mutual information content (relative entropy) calculated for donor (A; left column) and acceptor (B; right column) splice sites in the block sizes of 6 (gt±2, ag±2), 10 (gt±4, ag±4), and 14 (gt±6, ag±6) nt of the genes of five different organisms studied. The boundaries of the boxes represent the 25 (lower) and 75 (upper) percentile points. The horizontal line within the box represents the median value. The error bars show the 10 (bottom) and 90 (top) percentile points. It is clearly seen that the distribution is highly skewed and all the cases of the 90 percentile points are comparatively high in value. The median values show relatively little variation between the three blocks studied. All the graphs have been plotted on the same scale for ease in visual comparison.

Table 2 Base Pair Preferences at Donor and Acceptor Splice Site Regions

Organism	Donor splice site region		
	6-nt block	10-nt block	14-nt block
A. thaliana	gg>tt>aa>ac>ca>cc	gg>tt>aa>cc>ac>ca	gg>tt>aa>cc
C. elegans	gg>tt>aa>cc>ca	gg>tt>aa>cc>ac>ca	gg>tt>aa>cc>ac>ca
D. melanogaster	tt>gg>aa>cc>ac>ca	gg>tt>aa>cc>ac>ca	gg>tt>aa>cc
G. gallus	tt>gg>aa>ac>ca>cc	gg>tt>aa>cc>ac>ca	gg>tt>aa>cc
R. norvegicus	tt>gg>aa>ac>ca>cc	gg>tt>aa>cc>ca>ac	gg>tt>aa>cc
	Acceptor splice site region		
	6-nt block	10-nt block	14-nt block
A. thaliana	gg>aa>cc>tt>ct>tc	gg>aa>tt>cc	gg>aa>tt>cc
C. elegans	gg>aa>cc>tt	tt>gg>aa>cc	tt>gg>aa>cc
D. melanogaster	gg>aa>cc>tt>ct>tc	gg>aa>tt>cc	gg>aa>tt>cc
G. gallus	gg>aa>tt>cc>ct>tc	gg>aa>tt>cc	gg>aa>tt>cc
R. norvegicus	gg>aa>cc>tt>ct>tc	gg>aa>tt>cc>ct>tc	gg>aa>tt>cc

and acceptor sites in all the organisms studied, as they are equally significant for the binding of different spliceosomal proteins. We note that the values between 10–50 percentiles are very compact (less spread) while the values of 90 percentiles are far away from the median. This suggests that there are 1–2 values that are relatively high, which signify that the corresponding nucleotides are contributing to the high variability. In order to get a better understanding, we correlated the box plots of each organism with the individual elements of the H matrix (H_{ij}, 4×4=16 individual values) to obtain the information about individual base pair preferences as given in Table 2.

Donor (5′ splice site) region

We note from Table 2 that in the donor sequences the base pairs "gg" and "tt" have higher information content than "aa" and "cc" for all the cases. This is because the dinucleotide "gt" at the donor splice site is conserved and does not contribute to information content. Thus the high information content is attributed to the variability of the two nucleotides in the flanking regions of "gt", which suggests a high probability of each of the two nucleotides getting substituted by the other. The probability of adenine getting substituted by cytosine (or *vice versa*) is also significant. We can see from the 6-nt block of donor sites that guanine is more preferred in the flanking regions (1–2 nt) of "gt" in *A. thaliana* and *C. elegans*, while thymine is more preferred in the flanking regions of *D. melanogaster*, *G. gallus*, and *R. norvegicus*. We also see from Table 2 that the extent of variability decreases as the block size increases, suggesting that the nucleotides contributing to the variability are present in the neighborhood of the splice sites.

Acceptor (3′ splice site) region

We also note that in the acceptor sequences the base pairs "gg" and "aa" have higher information content than "tt" and "cc" for most cases. This is due to the conservation of the dinucleotide "ag" at the acceptor site, which does not contribute to the information content. This observation suggests that the given nucleotides in the decreasing order of their preferences contribute to the variability in the consensus of these sites. In the flanking nucleotides of "ag", the probability of thymine getting substituted by cytosine (or *vice versa*) is also observed. We note that the consensus at the acceptor region is more conserved than that at the donor region as fewer substitutions are observed comparatively, which is also evident from the high information content observed for the 6-nt block (Figure 2). It also shows a decreasing order in the preference of nucleotides as the block size increases (Table 2). We note from the 10-nt and 14-nt blocks of acceptor sequences that thymine is more preferred in the flanking regions of "ag" in *C. elegans*, which is due to the presence of the short and highly conserved polypyrimidine tract that is adjacent to the acceptor splice site. The consensus sequence TTTTCAG/R at the 3′ end has been shown to be critical for its recognition and binding to the U2AF protein during the process of RNA splicing (*23*). All other organisms show general trends in the distribution of the nucleotides.

Conclusion

We assume from these observations that even though the nucleotides are showing some degrees of conservation in the flanking regions of the splice sites (gt/ag), there still exists a certain level of variability in the consensus, signifying that some substitutions are found to be tolerable at certain positions. This is presumed to respond to the different spliceosomal factors that lead the splicing process to occur selectively. Our study suggests that the information required for RNA splicing is contained in the consensus of ∼6–8 nt at both donor and acceptor regions, which are important for the binding of spliceosomal proteins to the splice sites as expected.

We have developed our own block databases and applied the concepts of information theory for this analysis. Our study gives a broad idea about the distribution of nucleotides at/around the splice sites and also gives a comparative analysis of the consensus sequences at both donor and acceptor regions of the splice sites, which is significant for the process of splicing in terms of their sequence conservation or variability. We assume that our study can provide some insights towards understanding the information hidden at/around the splice sites that are important for the process of splicing to occur efficiently. We conclude that variability is essential for the selectivity of the splicing process whereas conservation is desirable to restrict the degree of variability.

Materials and Methods

Database

The Exon-Intron Database (EID) released in September 2005 (http://hsc.utoledo.edu/bioinfo/eid/index.html) was downloaded for the present study. This database was built in FASTA format by utilizing the data obtained from GenBank. It is a database of protein-coding intron-containing genes, which contains gene sequences of different organisms along with their alternative isoforms (*24*). The splice sites with only "gt-ag" exon/intron boundaries were considered in our analysis. All other splice sites such as "gc-ag", "at-ac", and all the cryptic ones were excluded in the present study. However, we have included all the alternative splice sites in our analysis. The exon sequences are represented as uppercase letters, and the intron sequences along with the splice site dinucleotides are given as lowercase letters. We selected the gene sequences of five different organisms in order to have a broad data distribution, including *Arabidopsis thaliana* (plant), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (arthropod), *Gallus gallus* (aves), and *Rattus norvegicus* (mammal). Table 1 gives the details of the number of gene sequences and splice sites analyzed in the present study. Our objective was to select a broad range of species but otherwise the selection may be considered arbitrary. Therefore the present study can be considered as "typical" or "representative" with a reasonably broad representation.

Construction of block databases

The databases of splice sites containing the gene sequences of the given organisms were used for the construction of block databases. We developed three different databases for the donor (gt) and the acceptor (ag) splice sites respectively by aligning 2, 4, and 6 bases flanking on either side of the dinucleotides (-gt- and -ag-) for all the organisms being studied. Consequently, we constructed three blocks of 6 (gt±2, ag±2), 10 (gt±4, ag±4), and 14 (gt±6, ag±6) nt for each of the donor and the acceptor regions as illustrated in Figure 1. We have used the three different block sizes in order to have a comparative analysis of the conservation of bases at the splice sites, which are involved in the process of splicing. This is a better approach when compared to earlier studies, which gives a good understanding of the distribution of information around the splice sites. Scanning the nucleotides one by one with entropy would have been computationally expensive and the information obtained might have been disproportionately low. The blocks obtained were then used for the computations of the substitution matrix.

Substitution matrix

We constructed substitution matrices for the aligned set of sequences of the given block sizes to calculate their mononucleotide substitutions (*15*). For the construction of each substitution matrix, we counted the number of matches and mismatches of each nucleotide type in each column between the first sequence and every other sequence present in the database. The same procedure was followed for every sequence in the database for all the columns present, and the values obtained were stored in a 4×4 frequency table, which gives the number of possible pairs of nucleotides in the database. For a database with a width of w nucleotides and a depth of s sequences, $ws(s-1)/2$ nucleotide pairs can be obtained, giving the frequency of occurrence of each of the 10 (4+3+2+1) different nucleotide pairs in the database. Thus we obtained a 4×4 frequency table, with each of its elements being represented as f_{ij}. This table was further utilized for the calculation of log-odds matrix. In our case, w is taken to be 6, 10, or 14, while s depends on the particular organism (Table 1) studied.

Log-odds matrix

Log-odds matrix is suitable to score alignments, in which the frequencies of the nucleotides in the aligned sequences are used to construct the substitution matrix. Log-odds values are calculated by taking a logarithm to base 2 ($\log_2$) of the ratio of the observed (target) probability to the expected (background) probability. The observed probability (q_{ij}) for each ij pair is calculated as:

$$q_{ij} = f_{ij} \,/\, \sum_{i=1}^{4} \sum_{j=1}^{i} f_{ij}$$

Then, the probability of occurrence (p_i) of the i^{th} nucleotide in an ij pair is calculated as:

$$p_i = q_{ij} + \frac{1}{2} \sum_{j \neq i} q_{ij}$$

The expected probability (e_{ij}) for each ij pair is then calculated as $e_{ij} = p_i p_j$ for $i = j$, and $e_{ij} = p_i p_j + p_j p_i = 2p_i p_j$ for $i \neq j$. The likelihood or the odds ratio matrix for each ij pair is calculated as the ratio of the observed probability to the expected probability: q_{ij}/e_{ij}, which gives the likelihood of occurrence of the nucleotides in pairs rather than by chance. The log-odds value of each ij pair is calculated as the logarithm of the odds ratio (s_{ij}), which is given as: $s_{ij} = \log_2(q_{ij}/e_{ij})$.

Mutual information content (relative entropy)

The entropy of a random variable is a measure of the uncertainty of the random variable. Thus, it measures the amount of information required on average to describe the random variable. The entropy $H(X)$ of a discrete random variable X with the probability mass (or density) function $p(x)$ is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

where the logarithm is taken to the base 2 and the entropy is expressed in bits. The relative entropy is a measure of the distance between two distributions. In statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy or the Kullback-Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as:

$$D(p \parallel q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}$$

Mutual information is defined as a measure of the amount of information that one random variable contains about the other. The mutual information $I(X;Y)$ of two random variables X and Y with a joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$ is given as the relative entropy between the joint distribution and the product distribution $p(x)p(y)$ (*25*):

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

We calculated the mutual information content for each block as the relative entropy H of the observed (target) probability to the expected (background) probability:

$$H_{ij} = q_{ij} \times s_{ij} \quad \text{or} \quad H_{ij} = q_{ij} \times \log_2 \frac{q_{ij}}{e_{ij}}$$

which is the product of the observed probability (q_{ij}) and the log-odds ratio (s_{ij}). The relative entropy of a log-odds substitution matrix is its ability to distinguish true alignments from other alignments, which appear by chance. We did not take over the sum of all the elements of the H matrix; instead, we plotted them as individual elements (H_{ij}) in the form of box plots.

Presentation of results

Instead of using a conventional histogram to display the results, we chose a box plot that shows the 25 and 75 percentiles as the box boundaries (Figure 2). The median (rather than the mean) value is shown within the box as a solid line. The error bars are shown as the 10 and 90 percentiles. This representation of data is more informative and gives a simple view of the distribution of the given data. All the plots were generated using the commercial software Sigmaplot 9.01 (Systat Software Inc., Richmond, USA).

Acknowledgements

This work was supported by the Council of Scientific and Industrial Research, Government of India (Senior Research Fellowship to TSR).

Authors' contributions

TSR carried out the computations. Both authors participated in the discussion of the results and the interpretation. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Lewin, B. 2000. Nuclear splicing. In *Genes VII.* Oxford University Press, New York, USA.
2. Staden, R. 1984. Computational methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* 12: 505-519.
3. Brunak, S., *et al.* 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* 220: 49-65.
4. Pertea, M., *et al.* 2000. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* 29: 1185-1190.
5. Chen, T., *et al.* 2005. Prediction of splice sites with dependency graphs and their expanded Bayesian networks. *Bioinformatics* 21: 471-482.
6. Shannon, C.E. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27: 379-423, 623-656.
7. Adami, C. 2004. Information theory in molecular biology. *Phys. Life Rev.* 1: 3-22.
8. Durbin, R., *et al.* 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, UK.
9. Schneider, T.D., *et al.* 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188: 415-431.
10. Rogan, P.K and Schneider, T.D. 1995. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. *Hum. Mutat.* 6: 74-76.
11. Giraud, B.G., *et al.* 1998. Analysis of correlations between sites in models of protein sequences. *Phys. Rev. E* 58: 6312-6322.
12. Adami, C. and Thomson, S.W. 2005. Predicting protein-protein interactions from sequence data. In *The Chemical Theatre of Biological Systems. Proceedings of the International Beilstein Workshop* (eds. Hicks, M.G. and Kettner, C.). Logos Verlag, Berlin, Germany.
13. Adami, C. 2002. Combinatorial drug design augmented by information theory. *NASA Tech Briefs* 26: 52.
14. Rekha, T.S. and Mitra C.K. 2006. $1/f$ correlations in viral genomes—a Fast-Fourier Transformation (FFT) study. *Indian J. Biochem. Biophys.* 43: 137-142.
15. Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915-10919.
16. Dayhoff, M.O., *et al.* 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (ed. Dayhoff, M.O.), Vol.5, pp.345-352. National Biomedical Research Foundation, Washington DC, USA.
17. Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219: 555-565.
18. Reddy, D.A., *et al.* 2006. Comparative analysis of core promoter region: information content from mono and dinucleotide substitution matrices. *Comput. Biol. Chem.* 30: 58-62.
19. Stephens, R.M. and Schneider, T.D. 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* 228: 1124-1136.
20. Fields, C. 1990. Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Res.* 18: 1509-1512.
21. Mount, S.M., *et al.* 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* 20: 4255-4262.
22. Sheth, N., *et al.* 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* 34: 3955-3967.
23. Hollins, C., *et al.* 2005. U2AF binding selects for the high conservation of the *C. elegans* $3'$ splice site. *RNA* 11: 248-253.
24. Saxonov, S., *et al.* 2000. EID: the Exon-Intron Database—an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.* 28: 185-190.
25. Cover, T.M. and Thomas, J.A. 1991. *Elements of Information Theory.* John Wiley & Sons, Inc., New York, USA.

Frequency Analysis of the Splice Site Regions in Different Organisms

T. Shashi Rekha and Chanchal K Mitra*

University of Hyderabad, Hyderabad 500 046, India.

e-mail: c_mitra@yahoo.com

We have carried out a comparative analysis of the sub-sequences of size six| ten at the (donor| acceptor) splice site regions of five different organisms. The frequency analysis of the unique sub-sequences at the donor and acceptor regions suggests that the distribution of their occurrence is approximately exponential. We have observed that the number of unique sub-sequences (occurring with different frequencies) at the donor region are less than at the acceptor, suggesting that the sub-sequences at the acceptor region are more variable. The sub-sequences with high percentage of occurrence (uniqueness) are considered to be highly involved in splicing. Our analysis suggests that sub-sequences of length ~6-8 nucleotides (nt) at the splice sites - with six bases in intron (including the two central, conserved dinucleotides) and two bases in exon are optimal for the efficient assembly and binding of the spliceosomal complex during the process of splicing. The score pattern obtained by the alignment of the nucleotides at the donor region with the acceptor and vice-versa also suggests that a single sub-sequence at the donor region have different degree of similarity with sub-sequences at the acceptor thus determining that the donor sub-sequences are more crucial in pairing with the corresponding acceptor sub-sequences during the process of splicing.

1 Introduction

The mechanism of splicing is directed by the recognition of the donor (5’-splice site), acceptor (3’-splice site) and the branch point consensus sequences by the catalytic particles of the splicing apparatus called the “spliceosomal complex”. The splicing apparatus contains both proteins and RNAs, which takes the form of small ribonucleoprotein particles in nucleus and cytoplasm. Those restricted to the nucleus are called small nuclear RNAs (snRNAs) and exist as ribonucleoprotein (snRNP) particles. The snRNPs involved in splicing (U1, U2, U5, U4 and U6) together with some additional proteins form a large particulate complex at the splice sites, called the “spliceosome” (Wassarman, 1992). The mechanism of splicing takes place in two concerted transesterification reactions as described in the given stages:

Stage I: In the first stage, a cut is made at the 5’ end of the 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’ phosphodiester bond to a base (‘A’) present in the branch point consensus of the intron.

Stage II: In the second stage, cutting at the 3’ splice site releases the free intron in lariat form, while the right exon is ligated (spliced) to the left exon (Lewin, 2000).

1.1 Consensus sequences at the splice sites

Even though a lot of work has been done to predict splice sites within a gene, studying the sub-sequences at the splice sites is an important topic of research for understanding some of the aspects of splicing. The splice site regions are not conserved, as different genes need specific spliceosomes for activation (one spliceosome that activates all the genes is likely to be a very inefficient process). So, we expect a given spliceosomal complex to act on a small number of related genes. The intron boundaries are generally characterized by the presence of the dinucleotides, GU (at the donor) and AG (at the acceptor region). But all the GU…AG present in the genome are not always the integral components of the splice sites. So, it is important to study the sub-sequences at (and around) the splice sites, which contain most of the information required for splicing (attachment of the spliceosomal complex). The recognition of true splice sites was explained to certain extent by the exon-bridging interactions (Robberson et al., 1990), where the 5' splice site on the downstream side of an exon can be a crucial determinant in the recognition and splicing of the upstream intron. Earlier work carried out on splice sites also signifies that the distance between the splice sites affect efficient spliceosomal assembly (Hertel, 2005). But much remains to be known as to how the two (donor and acceptor) splice sites are paired together, so that they are spliced out efficiently.

1.2 Variability of sub-sequences at splice sites

In most higher organisms (metazoans), both the splice sites are generally characterized by the presence of loosely conserved consensus sequences at the junctions of introns and exons (5'- and 3'-splice sites), which are recognized by the snRNA of the spliceosomal complex (Black, 1995). Even though the consensus sequences at the splice sites are variable, they still contain the information required for splicing, which is contained in ~6-8 nucleotides at the donor| acceptor regions (Rekha and Mitra, 2006). It was also observed that the level of variability in them could be compensated by the recognition of different splice sites by different spliceosomal proteins, so that the process of splicing is carried out efficiently (Rekha and Mitra, 2006). One of the earlier models proposed states that the presence of certain nucleotides in certain positions plays a key role in the recognition of the consensus sequences at the splice sites (Milanesi, 1997). It also signifies that the more frequently a consensus is occurring at the splice site the more likely that it is considered to be the functional splice site.

In order to obtain those sequences that are actually involved in splicing, we have obtained all sub-sequences at both donor and acceptor splice site regions (obtained from the protein-coding intron containing gene sequences) of five different organisms (Table 1). We have carried out a comparative study of a few selected sub-sequences that are occurring with a high frequency. We have also analyzed the same sequences to obtain an optimal length of the given sub-sequences that are actually found to be containing the information required for splicing. We have calculated the scores of the alignment of the high frequency donor| acceptor sub-sequences at the splice sites with the different set sub-sequences (of any particular organism) occurring at the acceptor/donor splice sites and have obtained sub-sequences that might be paired during the process of splicing. Thus, analysis of the splice sites has become an important aspect of study in the field of computational biology because of their role in the prediction of exon-intron architecture of the protein coding genes.

It is common to use substitution matrices to compare similarity, and they are widely available for different kind of situations. For example, PAM and BLOSUM are very common but the basic assumptions in deriving these matrices are considerably different. We want to confine ourselves to the region around the splice sites but the usual substitution matrices are computed for the complete genome. Features specific to the splice sites are likely to get lost if we consider the substitution matrix computed for the complete genome. We have therefore attempted to construct a specific substitution matrix from the regions around the splice sites of the database. Any specific preferences will then show up in our matrix.

The basic focus in this work is neither the database nor the sequence analysis. We have looked for conserved regions around the splice sites but if they are too many in number and located at slightly variable locations, it may be difficult to identify all the sequences. We nevertheless could find several small conserved sub-sequences that may act as binding sites for various factors involved in splicing.

2 Materials and methods

2.1 Exon-Intron Database

We have downloaded the Exon-Intron Database (EID; release September 2005, http://hsc.utoledo.edu/bioinfo/eid/index.html) for our present analysis. It is a database of protein-coding intron containing gene sequences represented along with their alternative isoforms (Saxonov, 2005). It was built in the FASTA format by obtaining the data from the GenBank database. The exon and intron (including the splice site dinucleotides gt| ag) sequences are represented separately as upper and lowercase letters. Gene sequences with three types of splice site (exon| intron) boundaries are given in the database - “gt-ag”, “gc-ag” and “at-ac”. In the present work, we have considered the gene sequences with “gt-ag” boundaries and have ignored all other splice sites, which were accounting for relatively small proportion. We have selected the gene sequences of five different organisms (along with their alternative isoforms); such that we can have a broad distribution of the data from plants to mammals. The choice of organisms can be considered otherwise arbitrary. The selected organisms are *Arabidopsis thaliana* (plant), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (arthropod), *Gallus gallus* (aves) and *Rattus norvegicus* (mammal). The details of the number of gene sequences and splice sites considered in the present study are given in Table 1.

Table 1. Number of genes and splice sites of the organisms studied

No	Organism	No. of genes	No. of splice sites		Total no of unique splice sites*	
			Donor	Acceptor	Donor	Acceptor
1	*Arabidopsis thaliana*	20,716	130,099	131,229	14,082	23,118
2	*Caenorhabditis elegans*	18,594	111,970	112,361	14,231	7,852
3	*Drosophila melanogaster*	10,612	72,737	73,167	7,189	15,058
4	*Gallus gallus*	16,567	168,120	169,990	17,839	27,813
5	*Rattus norvegicus*	19,146	181,782	183,476	15,921	28,284

*An unique splice site is defined as the10 nucleotide string xxxx{gt|ag}xxxx, where x can be any one of the nucleotides {A C, G, T}. If we select the 6 nucleotide string, the total number of unique splice sites will be considerably less.

2.2 Selection of sub-sequences

All the gene sequences of each of the five different organisms present in the EID database were used for the selection of sub-sequences for the present study. The sub-sequences were obtained by aligning the two centrally conserved dinucleotides (gt| ag) on either side of the donor/acceptor splice site regions of all the gene sequences in each organism separately, by considering two ($n_1n_2\{gt|ag\}n_3n_4$) and four ($n_1n_2n_3n_4\{gt|ag\}n_5n_6n_7n_8$) nucleotides flanking the splice sites. This way four different sets of sub-sequences were obtained for each of the organisms under study with two sets (one each for donor and acceptor) of size six and another two of size ten. Thus, totally we have obtained 20 different sets of sub-sequences with four sets for each of the organisms under study. We have considered the sizes six| ten only because, from our earlier analysis it was observed that the information required for splicing is contained in ~6-8 nt around (donor| acceptor) the splice sites regions. We have considered only the first 65,535 splice sites of all the organisms in our analysis. This makes all the graphs comparable as the total frequency is always the same (*vide infra*). The details of the number of unique sub-sequences of length 10 (at the splice sites) of each organism studied are given in Table 1.

2.3 Frequency distribution of sub-sequences

Thus we have obtained 20 [5 (organisms) x 2 (donor| acceptor) x 2 (6| 10 nt length)] different sets of sub-sequences of size six| ten corresponding to the donor| acceptor regions of each of the five organisms. Each set was then imported into a worksheet and sorted alphabetically. Each set now has several identical consecutive sub-sequences placed next to each other rather than being arranged in a random manner. The frequency of occurrence of each of the unique sub-sequences was calculated using a script. It is important to note that since, these sub-sequences were obtained from the splice site regions, so their frequency of occurrence gives their occurrence at the respective splice sites. The sum of the frequencies in a given set now corresponds to the total number of donor| acceptor splice sites for each of the organism under study (65,535 in this case). In the original worksheet, we had several redundancies (multiples) but after this process, all the sequences are now unique.

These sub-sequences were sorted in descending order of their frequencies, so that we now have sub-sequences that are occurring most common at the top followed by the least common at the bottom of the worksheet. We have obtained ~256 unique sub-sequences for the set of size six (for both donor and acceptor sites). In a similar fashion, we obtained ~10,000 unique ones for size 10, at the donor regions of all the organisms (except *D. melanogaster*). And the results were differing at the acceptor region with ~15,000-20,000 different types in all the organisms (except *C. elegans*). Overall, the number of unique splice sites are more than in the acceptor region than the donor in all the organisms (except *C. elegans*) for size 10 (the differences are insignificant for size 6).

2.4 Splice site utilization factor (*F*)

We have also calculated the splice site utilization factor (*F*), as *F = (no. of splice sites (donor/acceptor) /No. of genes)* in each of the organisms studied, so that we can get an idea about the typical number of splice sites per gene in each organism. The values are tabulated (Table 2) for each species studied. We note that more evolved species has a higher value of *F*.

Table 2. Splice site utilization factor of the organisms studied

No	Organism	Splice site utilization factor (*F*) No of splice sites/No of genes	
		Donor	Acceptor
1	*A. thaliana*	6-7	6-7
2	*C. elegans*	6-7	6-7
3	*D. melanogaster*	6-7	6-7
4	*G. gallus*	10-11	10-11
5	*R. norvegicus*	9-10	9-10

2.5 Frequency plots of sub-sequences

The frequency values of each sub-sequence (arranged in descending order) at the donor| acceptor splice site regions of size six| ten were plotted as vertical bar charts (Figure 1 and 2) with the number of sub-sequences being plotted on x-axis and their corresponding frequencies on y-axis (using the commercial software Sigmaplot 9.01). We have considered only the first 65,535 number of splice sites of all the organisms in our analysis, such that the total area of all the graphs is the same (in all the plots). The x-axis tick labels are in reality the sub-sequences (of 6| 10 nts) that have not been shown. In addition, these sequences are not identical in all the species. These plots give us information about the frequency of occurrence of each sub-sequence at the donor| acceptor splice sites regions separately. The frequency axis has been conveniently plotted on a log scale for the ease of study and a regression line (Figure 1; in red) along with their slope value was also shown to indicate the trends.

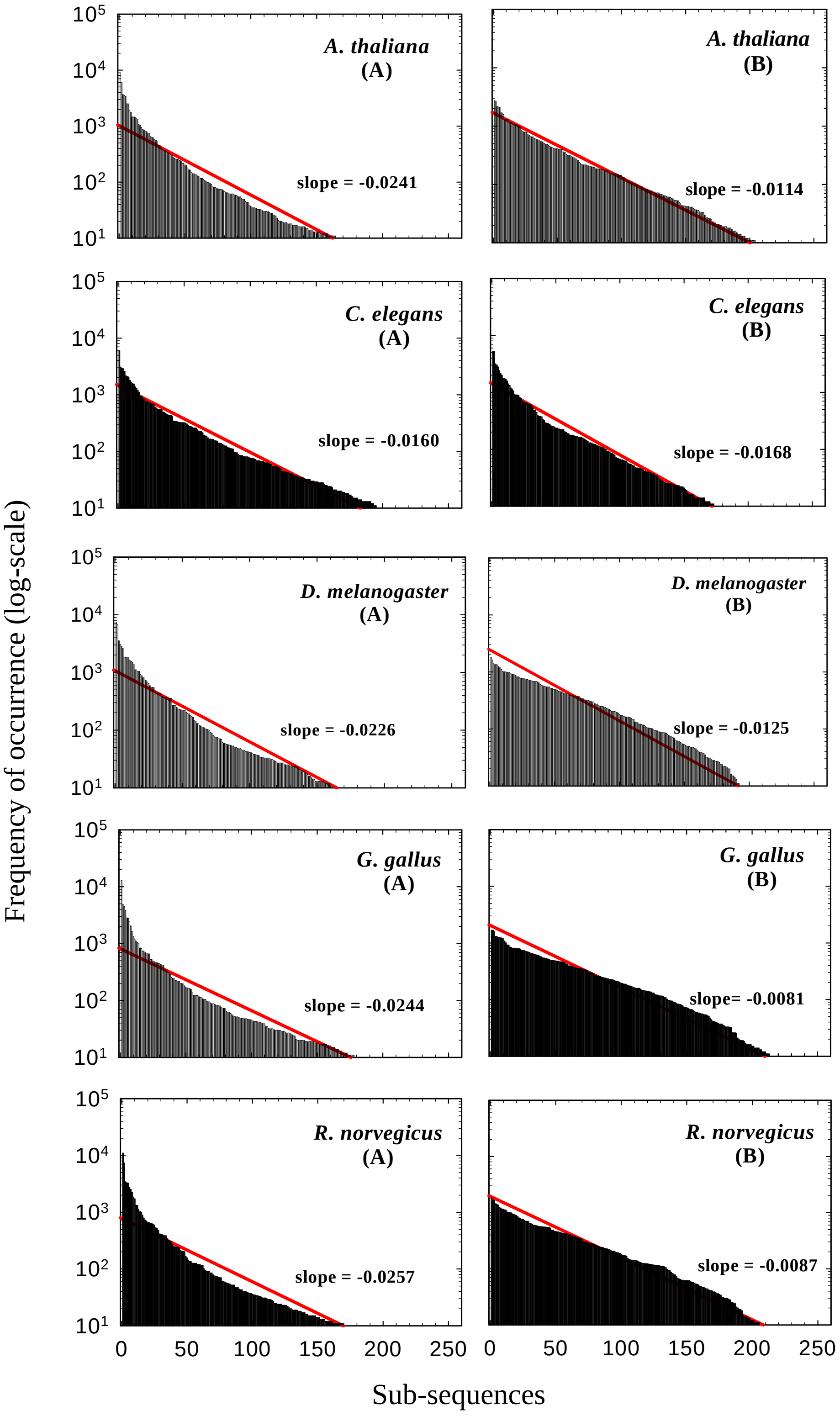


Fig 1. Vertical bar plots of the frequency of occurrence (log-scale) of the unique sub-sequences (arranged in descending order) in each set (first 65,535 sub-sequences considered) of size six of the respective organisms plotted against the corresponding sub-sequences (represented as numbers in linear scale) for the (A) donor and (B) acceptor splice site regions. Linear lines of regression are also shown (in red color) along with their respective slopes to indicate the trends of each plot. Scales of the axes are shown similar for all the organisms for the ease of comparison. The total area in each of the graphs is the same.

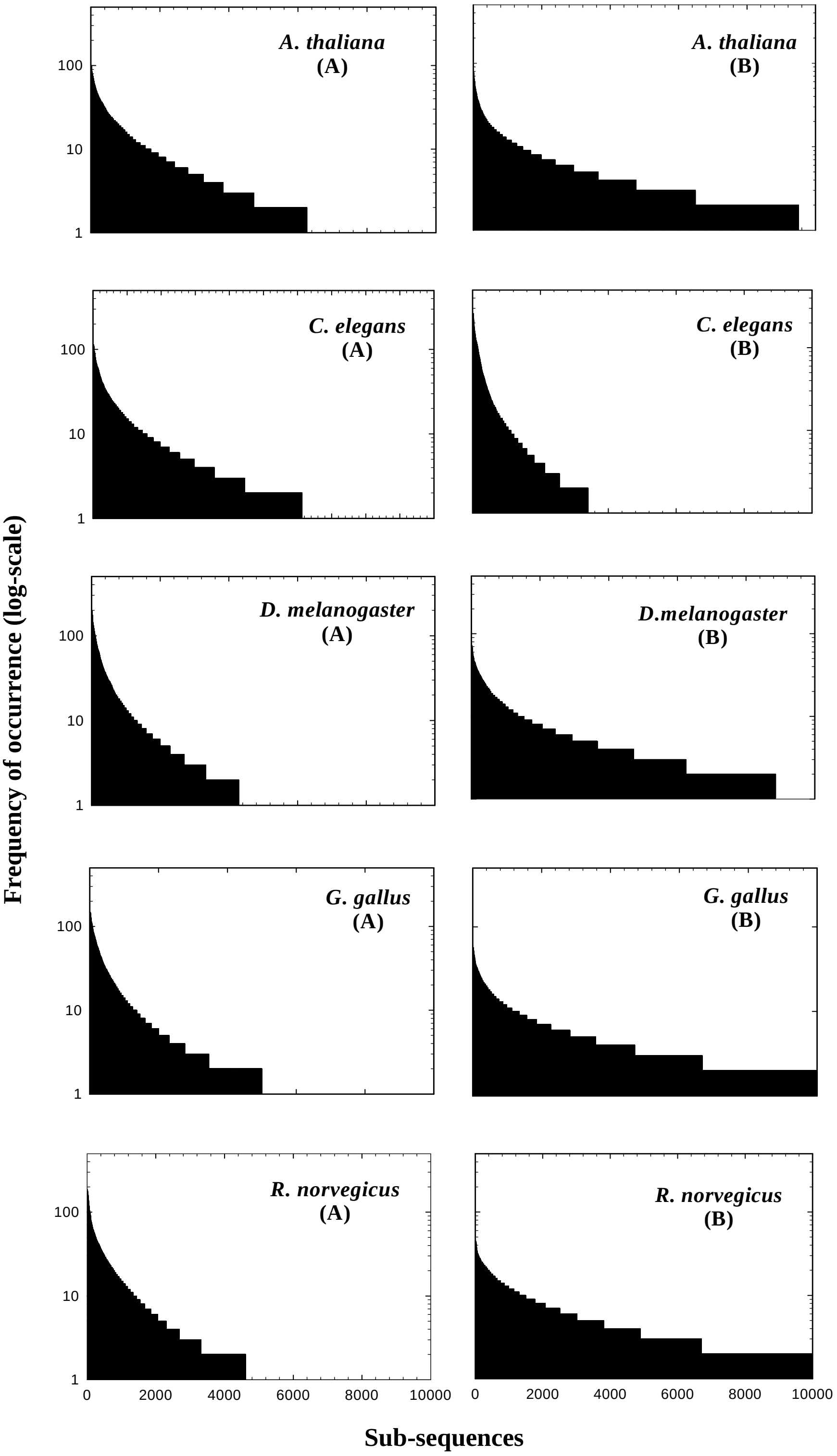


Fig 2. Vertical bar plots of the frequency of occurrence (log scale) of the unique sub-sequences (arranged in descending order) in each set (first 65,535 sub-sequences considered), of size ten of the respective organisms plotted against the number of corresponding sub-sequences (represented as numbers in linear scale) for the (A) donor and (B) acceptor splice site regions. Scales of the axes are shown similar for all the organisms for the ease of comparison. The total area in each of the graphs is the same.

2.5.1 Study of the uniqueness of sub-sequences

As we cannot possibly study all unique sub-sequences occurring with different frequencies at the splice sites, we have considered only those sub-sequences of size six| ten, which are occurring with the highest, medium and lowest frequencies as representative to get a comparative analysis of the data. The medium frequency is taken as the 50% frequency of the highest value (median value). We have studied the uniqueness of the sub-sequences by computing the same as *(n/N)*100*, where *n* is the frequency of occurrence of the given sub-sequence at the splice sites and *N* is the frequency of occurrence of the same sub-sequence in the whole genome (for a given organism). This gives the uniqueness of the given sub-sequence, with higher the percentage, higher is the uniqueness and lower the percentage lower is the uniqueness. The uniqueness values for the three representative sub-sequences are tabulated in Table 3a (size six, donor sites), 3b (size six, acceptor sites), 4a (size ten, donor sites) and 4b (size ten, acceptor sites), which gives the details of the frequency of occurrence values of the respective sub-sequences, at the splice sites regions and also the whole genomes of each of the organism being studied.

Table 3a. Frequency of occurrence of different sub-sequences (size six) at the donor splice site region and the whole genome of the respective organisms

No	Organism (genome size in nts)	Frequency	Sub-sequences at splice site†	Frequency at splice sites	Frequency in whole genome	Percentage of occurrence (uniqueness) at splice sites*
1	*A. thaliana*	Highest	AG**gt**aa	8,884	25,755	34.495
·	(58,129,057)	Medium	AG**gt**ac	3,667	13,548	27.067
		Lowest	TC**gt**tc	1	9,243	0.011
2	*C. elegans*	Highest	AG**gt**aa	5,916	15,338	38.571
·	(62,321,071)	Medium	TG**gt**aa	2,903	14,907	19.475
		Lowest	TT**gt**cg	1	16,736	0.006
3	*D. melanogaster*	Highest	AG**gt**aa	7,362	23,877	30.834
·	(125,309,791)	Medium	TG**gt**aa	3,558	28,006	12.705
		Lowest	TT**gt**tt	1	123,355	0.001
4	*G. gallus*	Highest	AG**gt**aa	12,970	137,320	9.446
·	(451,477,660)	Medium	AG**gt**ga	4,960	163,773	3.029
		Lowest	TC**gt**cg	1	3,918	0.026
5	*R. norvegicus*	Highest	AG**gt**aa	11,017	230,685	4.776
·	(867,510,682)	Medium	AG**gt**ga	7,373	272,167	2.709
		Lowest	TA**gt**tc	1	194,280	0.001

*The percentage of occurrence (uniqueness) values are normalized to three decimal points in order to represent even the lowest values. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 3b. Frequency of occurrence of different sub-sequences (size six) at the acceptor splice site region and the whole genome of the respective organisms

No	Organism (genome size in nts)	Frequency	Sub-sequences at splice site†	Frequency at splice sites	Frequency in whole genome	Percentage of occurrence (uniqueness) at splice sites*
1.	*A. thaliana*	Highest	`tc`**`ag`**`GT`	2,733	23,964	11.405
	(58,129,057)	Medium	`gt`**`ag`**`GT`	1,321	7,882	16.760
		Lowest	`cg`**`ag`**`CC`	1	3,939	0.025
2.	*C. elegans*	Highest	`tc`**`ag`**`AT`	5,270	27,645	19.064
	(62,321,071)	Medium	`tc`**`ag`**`AC`	2,791	14,441	19.327
		Lowest	`gg`**`ag`**`TC`	1	7,786	0.013
3.	*D. melanogaster*	Highest	`gc`**`ag`**`AT`	1,814	35,990	5.041
	(125,309,791)	Medium	`gc`**`ag`**`CA`	917	114,441	0.802
		Lowest	`tg`**`ag`**`TC`	1	18,913	0.006
4.	*G. gallus*	Highest	`gc`**`ag`**`GT`	1,687	139,010	0.214
	(451,477,660)	Medium	`cc`**`ag`**`GA`	845	140,060	0.604
		Lowest	`tg`**`ag`**`CA`	1	209,553	0.001
5.	*R. norvegicus*	Highest	`cc`**`ag`**`GT`	1,855	274,033	0.677
	(867,510,682)	Medium	`gc`**`ag`**`GT`	1,443	222,987	0.648
		Lowest	`ag`**`ag`**`CA`	1	427,990	0.001

*The percentage of occurrence (uniqueness) values is normalized to three decimal points in order to represent even the lowest values. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4a. Frequency of occurrence of different sub-sequences (size ten) at the donor splice site region and the whole genome of the respective organisms

No	Organism (genome size in nts)	Frequency	Sub-sequences at splice site†	Frequency at splice sites	Frequency in whole genome	Percentage of occurrence (uniqueness) at splice sites*
1.	*A. thaliana*	Highest	`TCAG`**`gt`**`ttgt`	179	435	41.150
	(58,129,057)	Medium	`AAAG`**`gt`**`aata`	89	194	45.877
		Lowest	`TTTT`**`gt`**`tttg`	1	2,389	0.042
2.	*C. elegans*	Highest	`AAAA`**`gt`**`gagt`	239	550	43.455
	(62,321,071)	Medium	`AGAT`**`gt`**`aagt`	120	240	50.000
		Lowest	`TTTT`**`gt`**`tttt`	1	240	0.417
3.	*D. melanogaster*	Highest	`CAAG`**`gt`**`gagt`	506	615	82.277
	(125,309,791)	Medium	`TGAG`**`gt`**`gagt`	243	308	78.896
		Lowest	`TTTT`**`gt`**`tatg`	1	486	0.206
4.	*G. gallus*	Highest	`AAAG`**`gt`**`aaga`	276	1,273	21.682
	(451,477,660)	Medium	`CAAA`**`gt`**`aagt`	136	892	15.247
		Lowest	`TTTT`**`gt`**`tttc`	1	8,091	0.013
5.	*R. norvegicus*	Highest	`CCAG`**`gt`**`gagt`	247	1,502	16.445
	(867,510,682)	Medium	`TCAG`**`gt`**`gagc`	124	1,232	10.065
		Lowest	`TTTT`**`gt`**`tttt`	1	54,854	0.002

*The percentage of occurrence (uniqueness) values is normalized to three decimal points in order to represent even the lowest values. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 4b. Frequency of occurrence of different sub-sequences (size ten) at the acceptor splice site region and the whole genome of the respective organisms

No	Organism (genome size in nts)	Frequency	Sub-sequences at splice site†	Frequency at splice site	Frequency in whole genome	Percentage of occurrence (uniqueness) at splice sites*
1.	*A. thaliana*	Highest	tttc**ag**GTTT	119	545	21.825
	(58,129,057)	Medium	ttgt**ag**GTGA	59	176	33.523
		Lowest	tttt**ag**TTCC	1	121	0.827
2.	*C. elegans*	Highest	tttc**ag**AAAA	651	3,744	17.388
	(62,321,071)	Medium	tttc**ag**ATCA	328	730	44.932
		Lowest	tttt**ag**TGCG	1	46	2.174
3.	*D. melanogaster*	Highest	ttgc**ag**ATGC	137	374	36.632
	(125,309,791)	Medium	ttgc**ag**TGCC	69	248	27.823
		Lowest	tttt**ag**TCGG	1	95	1.053
4.	*G. gallus*	Highest	tttc**ag**GTTT	99	2,412	4.105
	(451,477,660)	Medium	ttgc**ag**GCAG	50	1,817	2.752
		Lowest	tttt**ag**TTCG	1	107	0.935
5.	*R. norvegicus*	Highest	ctgc**ag**GTGG	75	2,223	3.374
	(867,510,682)	Medium	tttt**ag**GTTG	38	1,494	2.544
		Lowest	tttt**ag**TTGT	1	2,452	0.041

*The percentage of occurrence (uniqueness) values is normalized to three decimal points in order to represent even the lowest values. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

We assume from the above data that the sub-sequences, which are occurring more frequently, are the ones that are more commonly involved in the process of splicing. Based on this hypothesis, we have further studied sub-sequences that are occurring with the highest and medium frequencies at the donor| acceptor regions of each of the organisms being studied. As we have observed from our earlier analysis (Rekha and Mitra, 2006) of the splice site regions, that the information required for splicing might be contained in sub-sequences of ~6-8 nucleotides at the donor| acceptor regions. So, we have continued our further study with sub-sequences of size ten (occurring with highest and medium frequencies).

2.6 Identification of optimal length sub-sequences

One of the objectives of our study is to identify the sub-sequences of optimal length at the splice site (donor| acceptor) regions of each organism being studied, which are actually involved in the process of splicing. So, we have considered only sub-sequences of length ten (occurring with highest| medium frequencies at the splice sites) for our further analysis, as it is likely to be greater than the optimal sequence and discarded all sub-sequences of size six. It is not necessarily correct to assume that the information at| around the splice sites would be evenly distributed on both sides, and it is also important to consider the uneven distribution of the nucleotides on either side of the splice sites, we have trimmed two and four nucleotides from either side of each sub-sequence, in a systematic manner to get sequences of length eight and six. Thus we have only focused on these five different sets of sub-sequences (A1…A10; A1…A8; A1…A6; A3…A10; A5…A10) for our further study.

We have searched for these sub-sequences and have calculated their frequency of occurrence at the splice sites and also in the whole genome (in order to obtain only those sequences, which are occurring with the highest frequency at the splice sites with respect to the whole genome) and have recorded the number of matches found. For the ease of comparison, we have reported their percentage of occurrence (uniqueness) at the splice sites being calculated as described earlier in this paper. Table 5a, 5b, 6a and 6b give details of the frequency and the percentage of occurrence (uniqueness) of all sub-sequences at both the splice sites (donor| acceptor) and in the whole genome of each of the organisms studied.

Table 5a. Percentage of occurrence (uniqueness) of different sub-sequences (with highest frequency) of size ten at the donor splice site region and the whole genome of the respective organisms*

No	Organism	Sub-sequences at splice site[†]	Frequency at splice site	Frequency in whole genome	Percentage of occurrence (uniqueness) at splice sites
1.	*A. thaliana*	TCAG**gt**ttgt	179	435	41.15
		TCAG**gt**tt	1,168	3,122	37.42
		TCAG**gt**	9,302	23,964	38.82
		gtttgt	3,097	41,721	7.43
		AG**gt**ttgt	2,311	3,823	60.45
2.	*C. elegans*	AAAA**gt**gagt	239	550	43.46
		AAAA**gt**ga	716	5,594	12.80
		AAAA**gt**	2,616	71,992	3.64
		gtgagt	14,483	19,302	75.04
		AA**gt**gagt	2,491	2,998	83.08
3.	*D. melanogaster*	CAAG**gt**gagt	506	615	82.28
		CAAG**gt**ga	848	3,183	26.65
		CAAG**gt**	2,822	25,757	10.96
		gtaggt	15,759	36,271	43.45
		AG**gt**gagt	4,817	5,595	86.10
4.	*G. gallus*	AAAG**gt**aaga	276	1,273	21.69
		AAAG**gt**aa	3,802	15,643	24.31
		AAAG**gt**	9,591	160,235	5.99
		gtaaga	9,565	117,941	8.11
		AG**gt**aaga	4,614	11,235	41.07
5.	*R. norvegicus*	CCAG**gt**gagt	247	1,502	16.45
		CCAG**gt**ga	2,413	19,020	12.69
		CCAG**gt**	10,859	274,033	3.97
		gtgagt	24,678	301,321	8.19
		AG**gt**gagt	6,186	18,281	33.84

*Sub-sequences of size ten found with highest frequency at the donor splice site region were trimmed; two| four bases to obtain different sub-sequences of six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at the donor splice site region in all the five organisms studied. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 5b. Percentage of occurrence (uniqueness) of different sub-sequences (with highest frequency) of size ten at the acceptor splice site region and the whole genome of the respective organisms*

No	Organism	Sub-sequences at splice site†	Frequency at splice site	Frequency in whole genome	Percentage of occurrence (uniqueness) at splice sites
1.	*A. thaliana*	tttc**ag**GTTT	119	545	21.84
		tttc**ag**GT	2,169	4,384	49.48
		tttc**ag**	9,532	36,304	26.26
		agGTTT	2,948	33,668	8.76
		tc**ag**GTTT	534	3,122	17.11
2.	*C. elegans*	tttc**ag**AAAA	651	3,744	17.39
		tttc**ag**AA	7,588	14,916	50.88
		tttc**ag**	53,053	94,019	56.43
		agAAAA	2,369	100,523	2.36
		tc**ag**AAAA	1,250	10,057	12.43
3.	*D. melanogaster*	ttgc**ag**ATGC	137	374	36.64
		ttgc**ag**AT	1,144	3,553	32.20
		ttgc**ag**	9,405	50,070	18.79
		agATGC	676	28,570	23.72
		gc**ag**ATGC	220	3,447	6.39
4.	*G. gallus*	tttc**ag**GTTT	99	2,412	4.11
		tttc**ag**GT	1,896	19,295	9.83
		tttc**ag**	13,530	361,795	3.74
		agGTTT	2,623	185,539	1.42
		tc**ag**GTTT	434	15,976	2.72
5.	*R. norvegicus*	ctgc**ag**GTGG	75	2,223	3.38
		ctgc**ag**GT	1,326	22,362	5.93
		ctgc**ag**	7,858	403,613	1.95
		agGTGG	3,356	288,827	1.17
		gc**ag**GTGG	554	25,010	2.22

*Sub-sequences of size ten found with highest frequency at the acceptor splice site region were trimmed; two| four bases to obtain different sub-sequences of six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at the acceptor splice site region in all the five organisms studied. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 6a. Percentage of occurrence (uniqueness) of different sub-sequences (with medium frequency) of size ten at the donor splice site region and the whole genome of the respective organisms*

No	Organism	Sub-sequences at splice site†	Frequency at splice site	Frequency in whole genome	Percentage of occurrence (uniqueness) at splice sites
1.	*A. thaliana*	AAAG**gt**aata	89	194	45.88
		AAAG**gt**aa	1,964	2,938	66.85
		AAAG**gt**	8,132	26,204	31.04
		gtaata	2,378	14,535	16.37
		AG**gt**aata	1,315	1,802	72.98
2.	*C. elegans*	AGAT**gt**aagt	120	240	50.00
		AGAT**gt**aa	344	1,185	29.03
		AGAT**gt**	769	18,213	4.23
		gtaagt	16,442	21,110	77.89
		AT**gt**aagt	2,036	2,488	81.84
3.	*D. melanogaster*	TGAG**gt**gagt	243	308	78.90
		TGAG**gt**ga	409	1,466	27.90
		TGAG**gt**	1,330	15,803	8.42
		gtgaGT	15,759	36,271	43.45
		AG**gt**gagt	4,817	5,595	86.10
4.	*G. gallus*	CAAA**gt**aagt	136	892	15.25
		CAAA**gt**aa	584	14,013	4.17
		CAAA**gt**	993	157,546	0.64
		gtaagt	21,815	115,220	18.94
		AA**gt**aagt	2976	11,877	20.06
5.	*R. norvegicus*	TCAG**gt**gagc	124	1,232	10.07
		TCAG**gt**ga	1,808	22,266	8.13
		TCAG**gt**	9,080	269,752	3.37
		gtgagc	7,391	250,002	2.96
		AG**gt**gagc	3,943	17,670	51.41

*Sub-sequences of size ten found with highest frequency at the donor splice site region were trimmed; two| four bases to obtain different sub-sequences of six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at the donor splice site region in all the five organisms studied. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

Table 6b. Percentage of occurrence (uniqueness) of different sub-sequences (with medium frequency) of size ten at the acceptor splice site region and the whole genome of the respective organisms*

No	Organism	Sub-sequences at splice site†	Frequency at splice site	Frequency in whole genome	Percentage of occurrence (uniqueness) at splice sites
1.	*A. thaliana*	ttgt**ag**GTGA	59	176	35.53
		ttgt**ag**GT	1,095	1,786	61.32
		ttgt**ag**	5,995	22,839	26.25
		agGTGA	2,711	19,045	14.24
		gt**ag**GTGA	273	688	39.69
2.	*C. elegans*	tttc**ag**ATCA	328	730	44.94
		tttc**ag**AT	7,548	11,078	68.14
		tttc**ag**	53,053	94,019	56.43
		agATCA	1,253	21,616	5.80
		tc**ag**ATCA	682	1,735	39.31
3.	*D. melanogaster*	ttgc**ag**TGCC	69	248	27.83
		ttgc**ag**TG	510	3,389	15.05
		ttgc**ag**	9,405	50,070	18.79
		agGTCC	362	11,165	3.25
		gc**ag**GTCC	58	1,145	5.07
4.	*G. gallus*	ttgc**ag**GCAG	50	1,817	2.76
		ttgc**ag**GC	965	11,154	8.66
		ttgc**ag**	12,489	276,171	4.53
		agGCAG	1,404	214,811	0.66
		gc**ag**GCAG	361	22,684	1.60
5.	*R. norvegicus*	tttt**ag**GTTG	38	1,494	2.55
		tttt**ag**GT	1,064	28,178	3.78
		tttt**ag**	6,137	385,529	1.60
		agGTTG	1,839	211,489	0.87
		tt**ag**GTTG	223	11,539	1.94

*Sub-sequences of size ten found with highest frequency at the acceptor splice site region were trimmed; two| four bases to obtain different sub-sequences of six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at the acceptor splice site region in all the five organisms studied. †The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

This way, we have identified sub-sequences, which are highly involved in the process of splicing by considering those that are having the highest percentage of occurrence. Table 7, gives a list of all sub-sequences whose percentage of occurrence (uniqueness) at the (donor| acceptor) splice sites was found to be the highest in each of the organisms studied.

Table 7. Sub-sequences at the donor and acceptor splice site regions of different organisms found with the highest percentage of occurrence (uniqueness)*

No	Organism	Sub-sequences obtained from original sequence found with respective frequency[†]		Sub-sequences obtained from original sequence found with respective frequency[†]	
		Donor region		Acceptor region	
		Highest	Medium	Highest	Medium
1	*A. thaliana*	AG**gt**ttgt	AG**gt**aata	tttc**ag**GT	ttgt**ag**GT
2	*C. elegans*	AA**gt**gagt	AT**gt**aagt	tttc**ag**	tttc**ag**AT
3	*D. melanogaster*	AG**gt**gagt	AG**gt**gagt	ttgc**ag**ATGC	ttgc**ag**TGCC
4	*G. gallus*	AG**gt**aaga	AA**gt**aagt	tttc**ag**GT	ttgc**ag**GC
5	*R. norvegicus*	AG**gt**gagt	AG**gt**gagc	ctgc**ag**GT	tttt**ag**GT

*Sub-sequences of size ten found with highest and medium frequency at both the splice sites were trimmed; two| four bases to obtain different sub-sequences of ten six, and eight, which were used to calculate their percentage of occurrence (uniqueness) at both donor and acceptor splice sites in all the five organisms studied. [†]The two central, highly conserved dinucleotides in the sub-sequences are shown in bold.

2.7 Scoring the donor/acceptor sub-sequences

We note that the optimal length of the sub-sequences at the donor| acceptor splice site regions of each of the organisms studied is around 8 nucleotides. A unique donor sub-sequence that occurs with a high frequency is likely to be associated with a unique acceptor site occurring with high frequency. However, the frequency distributions for the donor and acceptor sub-sequences are clearly different and there may be other factors that determine the association between the donors and acceptors. To discover the pattern of association between the donors and acceptors, we use a scoring model. Both donor and acceptor sites are directly recognised by some RNA present in the spliceosomal complex and we hope to look for some correlations between these sequences. We do not imply that the model specifies perfect similarity of the sub-sequences but simply requires that some correlation must be detectable. Therefore the absolute value of the score is less important than the resulting shape of the distribution. With this as objective, we have scored the highest frequency unique sub-sequence (taken from Table 7) of the donor regions against the full set of unique sub-sequences at the acceptor sites. This has been done for all the organisms in a systematic manner. We also have carried out the reverse way, i.e., the highest frequency unique acceptor sequence has been scored against the complete set of unique sub-sequences at the donor sites. As the two distributions are clearly dissimilar, the results are expected to be different. As the donor and acceptor must occur in pairs, we are likely to see the correlation between them.

2.7.1 Substitution matrix and Log-odds ratios

For this, we have constructed substitution matrices separately for the aligned set of sub-sequences of the given size of six/ten for the donor| acceptor regions of each of the organisms, in order to calculate their mononucleotide substitutions (Henikoff and Henikoff, 1992) as described in our earlier paper (Rekha and Mitra). The log-odds matrix is suitable to score alignments, in which the frequencies of the nucleotides in the aligned sequences were used to construct the substitution matrix and the odds values were calculated by taking the ratio of the observed (q_{ij}) to expected probability (e_{ij}), which is given as q_{ij}/e_{ij}. This ratio gives the likelihood of occurrence of the nucleotides in (ij) pairs rather than by chance. The log-odds value of each of the ij pair is calculated as the logarithm to base 2 (log2) of the odds ratio (S_{ij}), which is given as: $S_{ij}=log_2\,(q_{ij}/e_{ij})$.

2.7.2 Calculation of the scores

We have scored four types of alignments, (i) the highest percentage of occurring (uniqueness) unique donor sub-sequence (obtained from unique parent sub-sequence of size ten having highest percentage of occurrence) against each of the unique acceptor set of sub-sequences and (ii) the highest percentage of occurring (uniqueness) unique acceptor sub-sequence (obtained from unique parent sub-sequence of size ten having highest percentage of occurrence) against each of the unique donor set of sub-sequences. Similar type of alignment was also done for the highest percentage of occurring unique donor and acceptor sub-sequences obtained from the unique parent sub-sequence of size ten occurring with medium frequency (iii) and (iv). All the highest frequency unique sub-sequences (donor/acceptor) aligned were of specific size for each of the organism considered for study (Table 7), which were aligned against the same size of the set of unique sub-sequences (acceptor/donor).

These alignments were then scored using the equation as given, $R = \Sigma_{ij} S_{ij}$ where R represents the score of the alignment, and S_{ij} represents the value assigned to the *ith* and the *jth* nucleotide in the log-odds matrix. This way, we have obtained four sets of scores for (i) unique donor-acceptor sub-sequence alignment (highest frequency unique parent) (ii) unique acceptor-donor sub-sequence alignment (highest frequency unique parent) (iii) unique donor-acceptor sub-sequence alignment (medium frequency unique parent) and (iv) unique acceptor-donor sub-sequence alignment (medium frequency unique parent). Thus, we have obtained 20 different sets of score values, which were plotted as histograms (Figures 3 and 4) using the software Sigmaplot. This way we can identify sub-sequences at the donor and acceptor regions that are actually paired during the process of splicing. The score values help us decide the similarities between the various sub-sequences, e.g., two sub-sequences with near-identical scores may be really one sub-sequence. This can be used to reduce further the total number of unique sub-sequences.

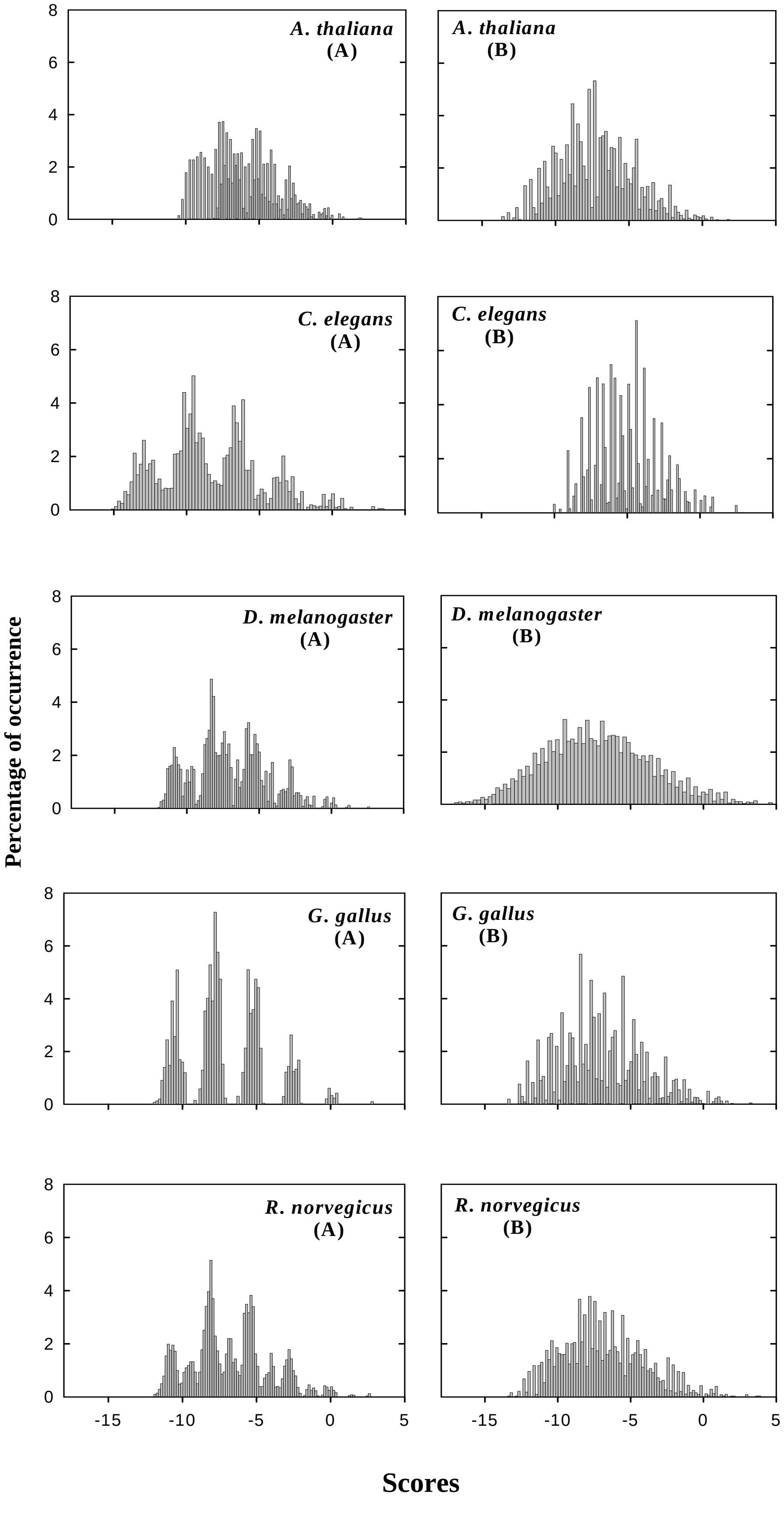


Fig 3. Histograms obtained by plotting the scores values (x-axis) against their percentage of occurrence (y-axis). These values were obtained by scoring the alignment of the (A; left column)

the highest frequency unique donor sub-sequence against each of the unique acceptor sub-sequences and similarly by the alignment of the (B; right column) highest frequency unique acceptor sub-sequence against each of the unique donor sub-sequences for each of the organisms under study. The highest frequency donor/acceptor sub-sequences aligned were found to be of specific size for each organism (Table 7) and were obtained from the parent sub-sequence of size ten having highest percentage of occurrence.

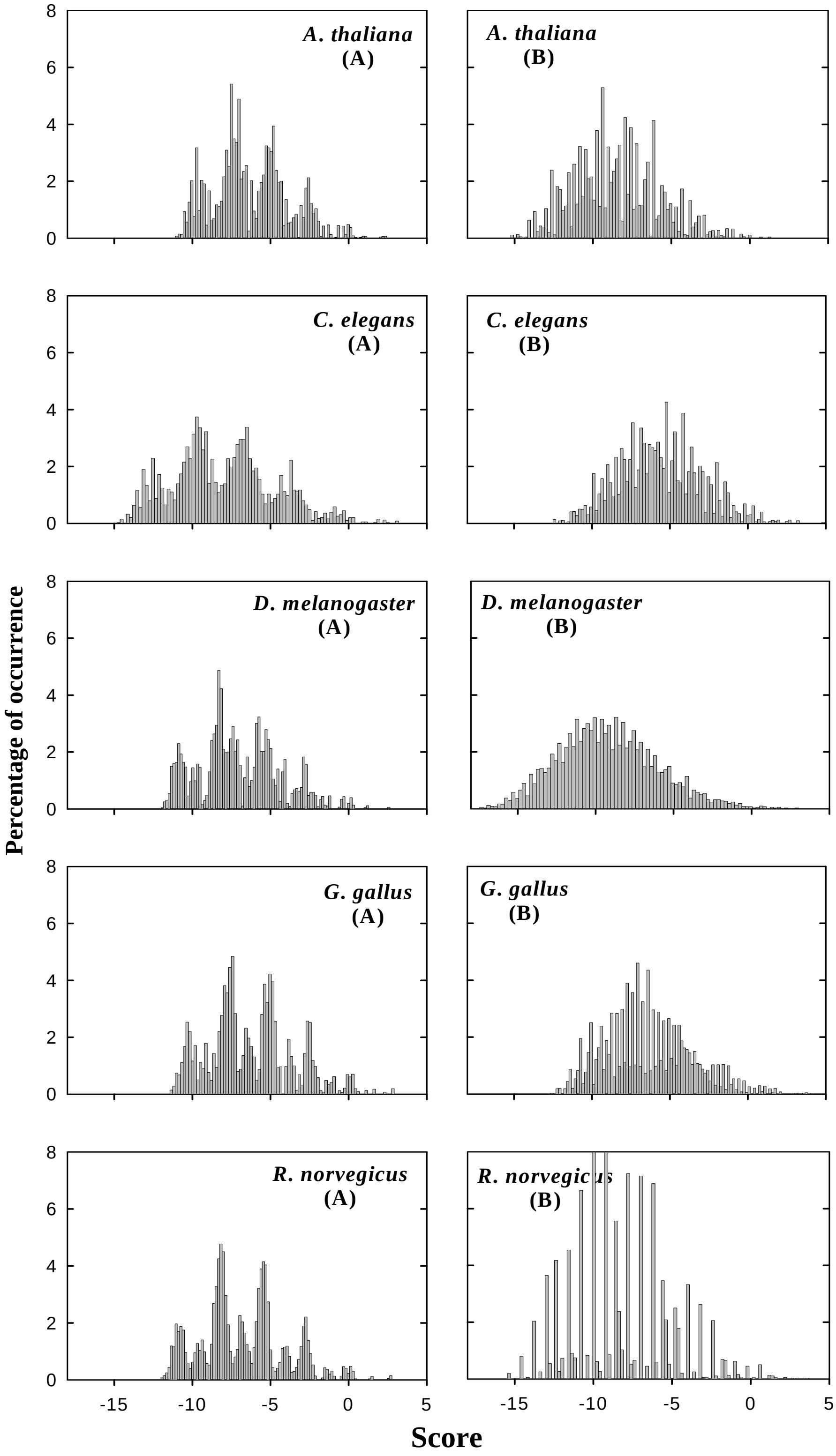


Fig 4. Histograms obtained by plotting the scores values (x-axis) against their percentage of occurrence (y-axis). These values were obtained by scoring the alignment of the (A; left column) the highest frequency unique donor sub-sequence against each of the unique acceptor sub-sequences and similarly by the alignment of the (B; right column) highest frequency unique acceptor sub-sequence against each of the unique donor sub-sequences for each of the organisms under study. The highest frequency donor/acceptor sub-sequences aligned were found to be of specific size for each organism (Table 7) and were obtained from the parent sub-sequence of size ten having medium percentage of occurrence.

3 Results and Discussions

3.1 Identification of unique sub-sequences

We have obtained unique sub-sequences (occurring with highest frequency) of size six| ten at donor| acceptor splice site regions in all the five organisms studied, which were ~256 in number for the set of size six and ~10,000 in number for the set of size ten. We note that the sub-sequences around the splice sites are highly variable, but far from random. The frequencies of sub-sequences follow an approximate exponential pattern that is common in nature (1/*f* distribution). As the length of sub-sequences increases their frequency of occurrence decreases and the total number of (observed) sub-sequences increases.

3.2 Frequency Distribution of sub-sequences

We have calculated the frequency of occurrence of each unique sub-sequence (size six| ten), at the donor| acceptor, splice site regions and also in the whole genome of each organism studied. We note (Figures 1 and 2) that the frequency distribution is approximately exponential, because the occurrence of certain unique sub-sequences is more common when compared to the other. The distribution of sub-sequences of size six (Figure 1) is steeper in donor region, when compared to acceptor in all organisms studied except in *C. elegans*, in which the distribution is more or less equal in both the regions. We have drawn linear lines of regression for all the plots (Figure 1) and have obtained their respective slopes. We note that the slopes of the plots of donor region are higher than acceptor in all organisms (except *C. elegans*), which shows equal slopes for both (donor and acceptor) regions. This suggests that the frequency distribution at donor and acceptor regions is equal.

We note (Figure 1) that, since the frequency values are high in the donor region, the number of their corresponding (unique) sub-sequences are comparatively low (inverse relation). But the number is more in the acceptor region than the donor (thus their corresponding frequencies are less). This suggests that there are less number of donor and more number of acceptor splice sites in all the four organisms studied, signifying more variability in the acceptor region than the donor. But in *C. elegans*, we note the number to be approximately equal in both regions.

We have also observed a similar trend in the sub-sequences of size ten (Figure 2) with their frequencies higher at the donor region than the acceptor, (except for *C. elegans*, in which sub-sequences at the acceptor region are having frequencies higher than the donor). We also observe that the number of unique sub-sequences in the acceptor region are more than the donor, (except for *C. elegans*), which suggests that there is more variability in the acceptor region than the donor. We observe that the sub-sequences at the acceptor region in *C. elegans* are more conserved than the donor. This is because thymine is more preferred in the flanking regions of “ag" in *C. elegans*, which is due to the presence of the short and highly conserved polypyrimidine tract present adjacent to the acceptor splice site. The consensus sequence TTTTCAR at the acceptor region of *C. elegans* has been shown to be critical for its recognition and binding by the U2AF protein during the process of RNA splicing (Blumenthal, 2005).

3.3 Non-random distribution of sub-sequences

We note that the frequency distribution of the sub-sequences is not uniform. If we consider the set of sub-sequences of size ten (Figure 2), the possibility of occurrence of each of the four bases at each of the eight positions (excluding the two central, highly conserved dinucleotides, “ag”) would be 4^8 = 65,536, whereas the actual occurrence was found to be ~10,000 for each of the organisms. These observations suggests that there are certain unique sub-sequences, which are occurring more frequently than by random chance, (because certain bases are conserved at certain positions in the sub-sequences studied). But the frequency distribution of the set of sub-sequences of size six (Figure 1), was found to be as expected as 256 (i.e., the possibility of occurrence of each of the four bases (A, C, G and T) by random chance, in each of the four positions (excluding the two central, highly conserved dinucleotides, “gt”) would be 4^4 = 256). This is in accordance with our earlier work (Rekha and Mitra), which suggests that there is more variability in the immediate flanking regions of the splice sites and the variability decreases as we move away from these splice sites.

3.4 Sub-sequences involved in splicing

From the frequency of occurrence values of sub-sequences of size six| ten at both (donor| acceptor) the splice sites and the whole genome (Table 3a, 3b, 4a and 4b) we assume that the sub-sequences with the highest frequency of occurrence at the splice sites are the ones, which are more commonly involved in the process of splicing.

We have obtained similar observations from the percentage of occurrence (uniqueness) values (Table 5a, 5b, 6a and 6b) of size ten, for each of the five organisms. We also note that the length of the respective sub-sequences (Table 7) occurring with the highest percentage of occurrence (uniqueness) might be optimal for the binding and assembly of the spliceosomal complex during the process of splicing.

3.5 Sub-sequences of optimal length

3.5.1 Sub-sequences at donor region (highest frequency)

From the data (Table 5a) obtained, we observe that sub-sequences with highest percentage of occurrence or uniqueness (obtained from parent sub-sequence with highest frequency) containing two bases in the exonic region and six bases in the intronic region (including the two highly conserved dinucleotides “gt”) might be highly involved in the process of splicing at the donor region of all the organisms studied (Table 7) and contain a length of eight nucleotides, which is optimal for the spliceosomal assembly and binding of the organisms studied.

3.5.2 Sub-sequences at acceptor region (highest frequency)

The consistency shown in the donor region is not really observed at the acceptor regions of the organisms studied because from the data obtained (Table 5b) the optimal length of sub-sequences (obtained from parent sub-sequence of highest frequency) at the acceptor splice site region is eight with six bases in the intronic region (including the two conserved dinucleotides “ag”) and two bases in the exonic region in three of the organisms studied – *A. thaliana*, *G. gallus* and *R. norvegicus*. But in *C. elegans*, the optimal length is found to be six, with all the bases in the intronic region (including the two conserved dinucleotides “ag”) only. But in *D. melanogaster*, the optimal length is more than all other species, i.e., ten with six bases in the intronic region (including the two conserved dinucleotides “ag”) and four bases in the exonic region. So the optimal length of sub-sequences at the acceptor region required for splicing is highly variable in the organisms studied (Table 7). This is perhaps due to the fact that one donor may be able to choose from a number of different acceptors.

3.5.3 Sub-sequences at donor region (medium frequency)

The data (Table 6a) obtained, represents a similar trend (as observed for the donor region discussed earlier) of the sub-sequences (obtained from parent sub-sequence of medium frequency) at the donor regions. These sub-sequences (Table 7), with their respective optimal lengths might be moderately involved in splicing in the organisms studied. The difference is in degree and the basic idea remains the same.

3.5.4 Sub-sequences at acceptor region (medium frequency)

For sub-sequences (obtained from parent sub-sequence of medium frequency) at the acceptor region (Table 6b), we observe the optimal length to be eight in the four organisms studied, with six bases in the intronic region (including the two conserved dinucleotides “ag”) and two bases in the exonic region (except *D. melanogaster*, where the optimal length was ten, as discussed earlier). We assume that these sub-sequences (Table 7) are moderately involved in the process of splicing in the organisms studied.

3.6 Scoring the alignments of donor-acceptor sub-sequences

Based on the hypothesis that the certain sub-sequences at the donor region have some similarity with the sub-sequences at the acceptor, we have scored the alignments of the unique donor sub-sequence (occurring with highest percentage of occurrence obtained from parent sub-sequence occurring with highest/medium percentage of occurrence) with each of the sub-sequences in the set of acceptor region and vice-versa. We have observed certain features in the graphs obtained by plotting these score values, which are discussed in detail as follows.

3.6.1 Donor (highest frequency parent sub-sequence) aligned against acceptor set

We observe from the histograms (Figure 3A) that the frequency of the score values (represented as percentage of occurrence or uniqueness) obtained by aligning the highest percentage of occurrence donor sub-sequence (obtained from parent sub-sequence occurring with highest percentage of occurrence) with each sub-sequence at the acceptor region is not normal. We have observed that the distribution is multimodal, which signifies that a single graph has a number of normal distributions combined together in it. We have also observed many peaks, which denote that a single donor sub-sequence has different degree of similarity with each of the sub-sequences at the acceptor region. We also assume that the donor sub-sequences are more crucial in deciding the acceptor region for splicing. The graph shows clustering behaviour with each cluster having peaks of different intensity. Different clusters were obtained as the donor sub-sequence is having similarity with different nucleotides in the acceptor sub-sequence. We have obtained negative scores for the sub-sequence similarity, which can be due to mismatches between some nucleotides at the donor and acceptor regions that are making the overall score of the alignment to be negative. But we also observe some positive scores for the alignment, which are found to be very less. This suggests that the similarity between the nucleotides in the donor and acceptor sub-sequences at the splice sites is not very high *per se*. However, it is not expected that the sequence information transmitted from the donor site to the acceptor site via the snRNA will be perfect. In such case, we stress more on the distribution rather than the exact value of the score.

3.6.2 Acceptor (highest frequency parent sub-sequence) aligned with donor set

The plots (Figure 3B) of the score values obtained by aligning the highest percentage of occurrence (uniqueness) acceptor sub-sequence (obtained from parent sub-sequence occurring with highest percentage occurrence) with each of the sub-sequences at the donor region suggests that the distribution is more or less normal in *A. thaliana*, *D. melanogaster* and *R. norvegicus*. But in *C. elegans* and *G. gallus* it shows the characteristics of a comb distribution with edge peaks. This distribution suggests that the sub-sequence occurring with the highest percentage of occurrence (uniqueness) at the acceptor region do not have proper alignment with sub-sequences at the donor region suggesting that the acceptor regions are not crucial in deciding the splicing process.

3.6.3 Donor (medium frequency parent sub-sequence) aligned with acceptor set

We observe that the plots (Figure 4A) of the score values obtained by aligning the highest percentage of occurrence donor sub-sequence (obtained from parent sub-sequence occurring with medium percentage of occurrence) with the sub-sequences at the acceptor region, show similar trends as discussed earlier (3.6.1) but the patterns seen here are not very clear (well resolved).

3.6.4 Acceptor (medium frequency parent sub-sequence) aligned with donor set

The plots (Figure 4B) of the score values obtained by aligning highest percentage of occurrence of acceptor sub-sequence (obtained from parent sub-sequence occurring with medium percentage occurrence) with the sub-sequences at the donor region, has shown a normal distribution in all the four organisms studied. But in *R. norvegicus*, we observe a comb distribution (with edge peaks), with one set of high values and another set of low values being represented together. This distribution suggests similar conclusions as given earlier (3.6.2). Again, we find the behaviour broadly similar and it is only different in degree.

4 Conclusions

We show that the information required for splicing is contained in ~6-8 nt at| around both the donor and acceptor splice sites. This work has given us a better idea about the distribution of information at| around the splice sites suggesting that sub-sequences at the splice sites studied are highly variable. The frequency analysis of these unique sub-sequences also suggests that the distribution is approximately exponential, because of the occurrence of certain high frequency unique sub-sequences more commonly than the other. The percentage of occurrence (uniqueness) values also suggests that sub-sequences with the highest values are the ones, which are highly involved in splicing. We also note that the length of 6-8 nt with six bases in intron (including the two central, conserved dinucleotides) and two bases in exon is optimal for the efficient assembly and binding of the spliceosomal complex during the process of splicing. We assume that the donor sub-sequences are more crucial in pairing with the corresponding acceptor sub-sequences during the process of splicing.

Further this idea can be extended in decoding the information present at the splice sites into distinct groups and classes. The rich variability of the donor and acceptor sites generates greater information and the information may be useful in understanding the language of the DNA at the splice site. Considerable experiments need to be carried out before the problem can be uniquely solved. However, we have clearly identified a number of broad features that can help in this direction. This kind of work can be carried in understanding the information contained in the promoter regions also, which might give some insights into the underlying mechanism.

5 Acknowledgements

TSR thanks the Council of Scientific and Industrial Research, Government of India, for a Senior Research Fellowship.

6 References

[1] D. A. Wassarman and J. A. Steitz. Interactions of small nuclear RNAs with precursor messenger RNA during in vitro splicing. Science, 257: 1918-1925, 1992.

[2] B. Lewin. Nuclear splicing. In Genes VII. Oxford University Press, New York, USA, 2000.

[3] B. L. Robberson, G. J. Cote, and S. M. Berget. Exon definition may facilitate splice site selection in RNAs with multiple exons. Molecular Cell Biology, 10 (1): 84-94, 1990.

[4] K. L. Fox-Walsh, Yimeng Dou, B. J. Lam, She-pin Hung, P. F. Baldi and K. J. Hertel. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proceedings of the National Academy of Sciences of the United States of America, 102: 16176-16181, 2005.

[5] D. L. Black. Finding splice sites within a wilderness of RNA. RNA, 1: 763-771, 1995.

[6] T. Shashi Rekha and C. K. Mitra. Comparative Analysis of Splice Site Regions by Information Content. Genomics, Proteomics & Bioinformatics, 4: 230-237, 2006.

[7] L. Milanesi and I. B. Rogozin. Analysis of donor splice sites in different eukaryotic organisms. Journal of Molecular Evolution, 45: 50-59, 1997.

[8] S. Saxonov, I. Daizadeh, A. Fedorov and W. Gilbert. EID: the Exon-Intron Database - an exhaustive database of protein-coding intron-containing genes. Nucleic Acids Research, 28 (1): 185-190, 2000.

[9] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America, 89: 10915-10919, 1992.

[11] C. Hollins, Diego A. R. Zorio, M. Macmorris and T. Blumenthal. U2AF binding selects for the high conservation of the C. elegans 3’ splice site. RNA, 11: 248-253, 2005.