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>
<thead>
<tr>
<th>No</th>
<th>Organism</th>
<th>No. of genes</th>
<th>Total no. of sub-sequences</th>
<th>Total no. of unique sub-sequences*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>5’ ss</td>
<td>3’ ss</td>
</tr>
<tr>
<td>1</td>
<td><em>A. thaliana</em></td>
<td>20,716</td>
<td>130,099</td>
<td>131,229</td>
</tr>
<tr>
<td>2</td>
<td><em>C. elegans</em></td>
<td>18,594</td>
<td>111,970</td>
<td>112,361</td>
</tr>
<tr>
<td>3</td>
<td><em>D. melanogaster</em></td>
<td>10,612</td>
<td>72,737</td>
<td>73,167</td>
</tr>
<tr>
<td>4</td>
<td><em>G. gallus</em></td>
<td>16,567</td>
<td>168,120</td>
<td>169,990</td>
</tr>
<tr>
<td>5</td>
<td><em>R. norvegicus</em></td>
<td>19,146</td>
<td>181,782</td>
<td>183,476</td>
</tr>
</tbody>
</table>

*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.

![Figure 5.1](image.png)

**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 \((gt\pm5, ag\pm5)\) 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 the 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 $\geq 6$ 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 $\geq 6$), we have calculated the frequency of occurrence of each of the motifs obtained (Table 5.2).

<table>
<thead>
<tr>
<th>No</th>
<th>Organism</th>
<th>Total no. of motifs</th>
<th>No. of unique motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5’ ss</td>
<td>3’ ss</td>
</tr>
<tr>
<td>1</td>
<td><em>A. thaliana</em></td>
<td>95,426,257</td>
<td>101,989,564</td>
</tr>
<tr>
<td>2</td>
<td><em>C. elegans</em></td>
<td>65,068,944</td>
<td>34,706,251</td>
</tr>
<tr>
<td>3</td>
<td><em>D. melanogaster</em></td>
<td>25,446,885</td>
<td>25,075,268</td>
</tr>
<tr>
<td>4</td>
<td><em>G. gallus</em></td>
<td>111,758,051</td>
<td>152,654,064</td>
</tr>
<tr>
<td>5</td>
<td><em>R. norvegicus</em></td>
<td>113,412,019</td>
<td>171,669,729</td>
</tr>
</tbody>
</table>

*A local pairwise alignment of all the unique sub-sequences of size 12 was carried out and all locally aligned pairs (motifs) of size $\geq 6$ 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 the 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 \( \geq 6 \) (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).
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.