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.

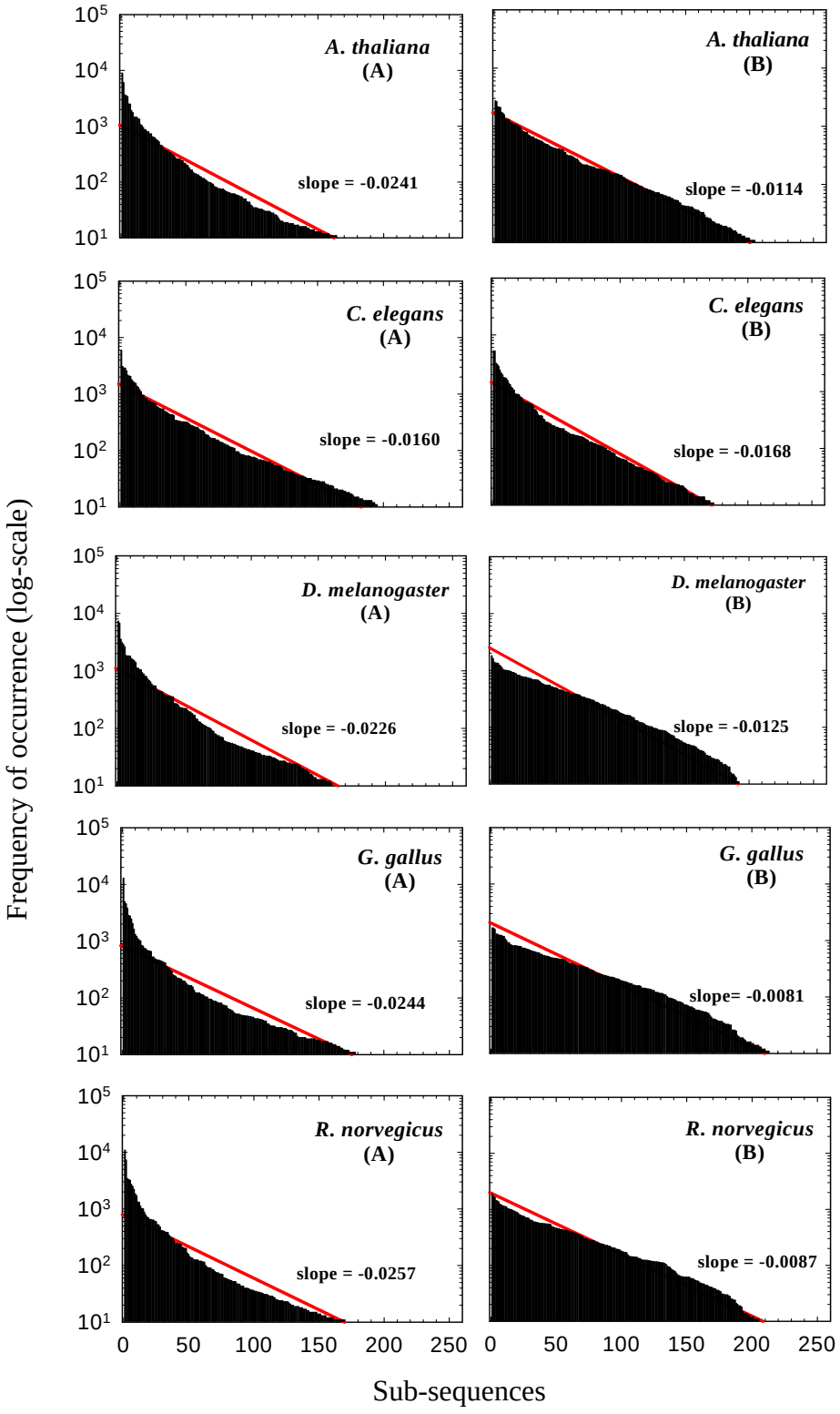

Fig 1. Vertical bar plots of the frequency of occurrence (log-scale) of the unique sub-sequences (arranged in descending order) in each set (first 65,535 sub-sequences considered) of size six of the respective organisms plotted against the corresponding sub-sequences (represented as numbers in linear 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.