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.

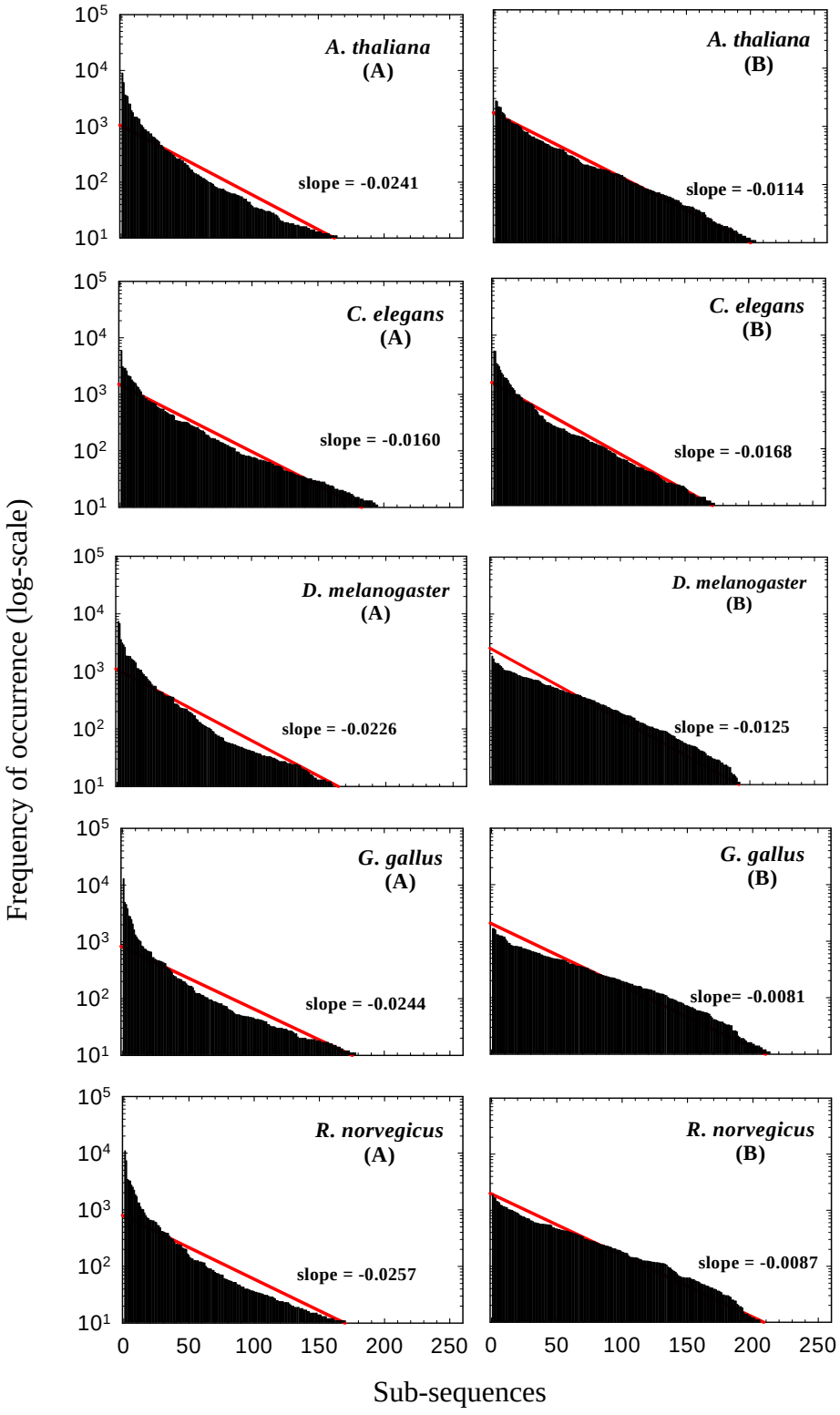

Fig 1. Vertical bar plots of the frequency of occurrence (log-scale) of the unique sub-sequences (arranged in descending order) in each set (first 65,535 sub-sequences considered) of size six of the respective organisms plotted against the corresponding sub-sequences (represented as numbers in linear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

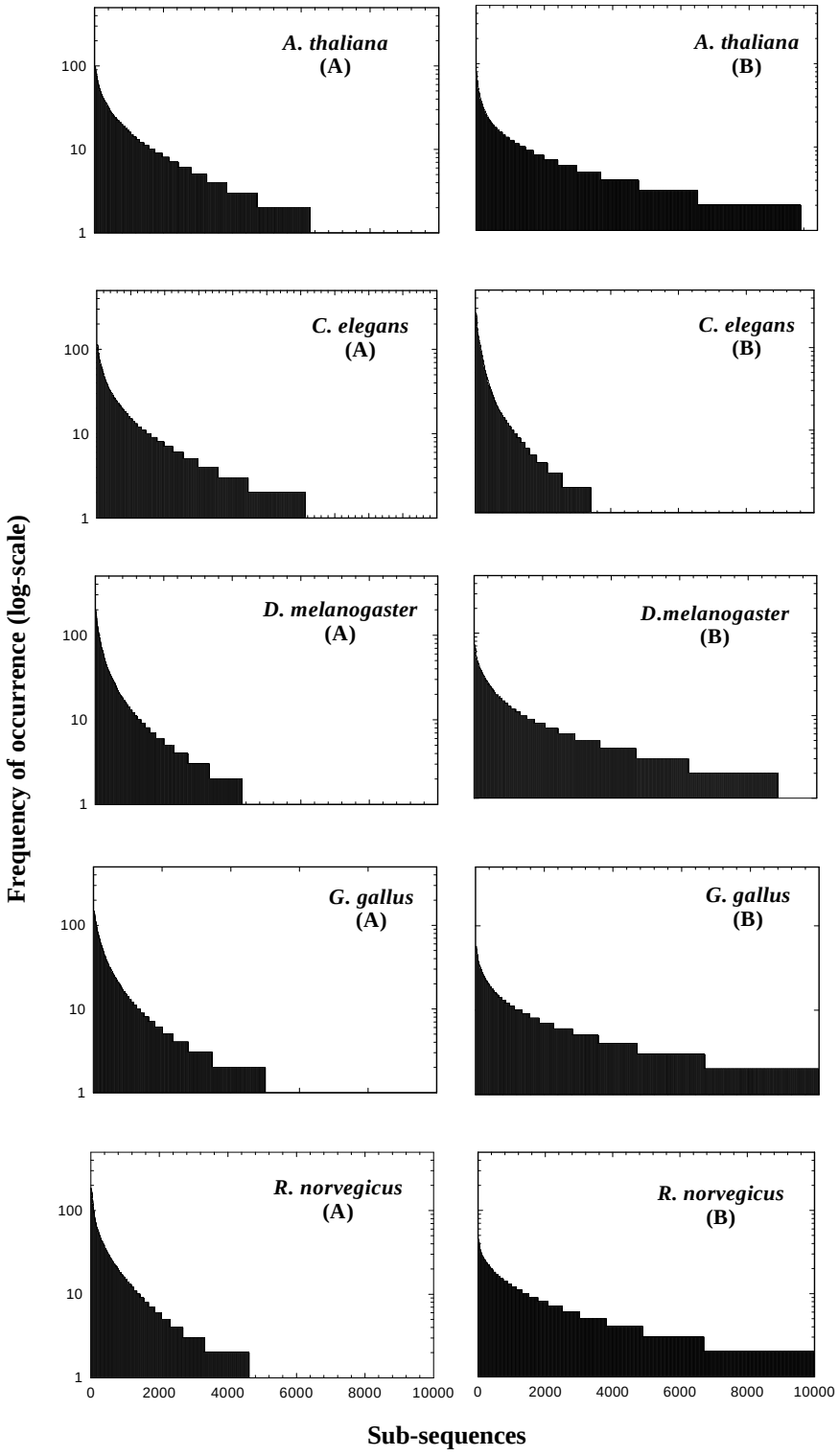

Fig 2. Vertical bar plots of the frequency of occurrence (log scale) of the unique sub-sequences (arranged in descending order) in each set (first 65,535 sub-sequences considered), of size ten of the respective organisms plotted against the number of corresponding sub-sequences (represented as numbers 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**tttgt	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* (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 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	ttt**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.

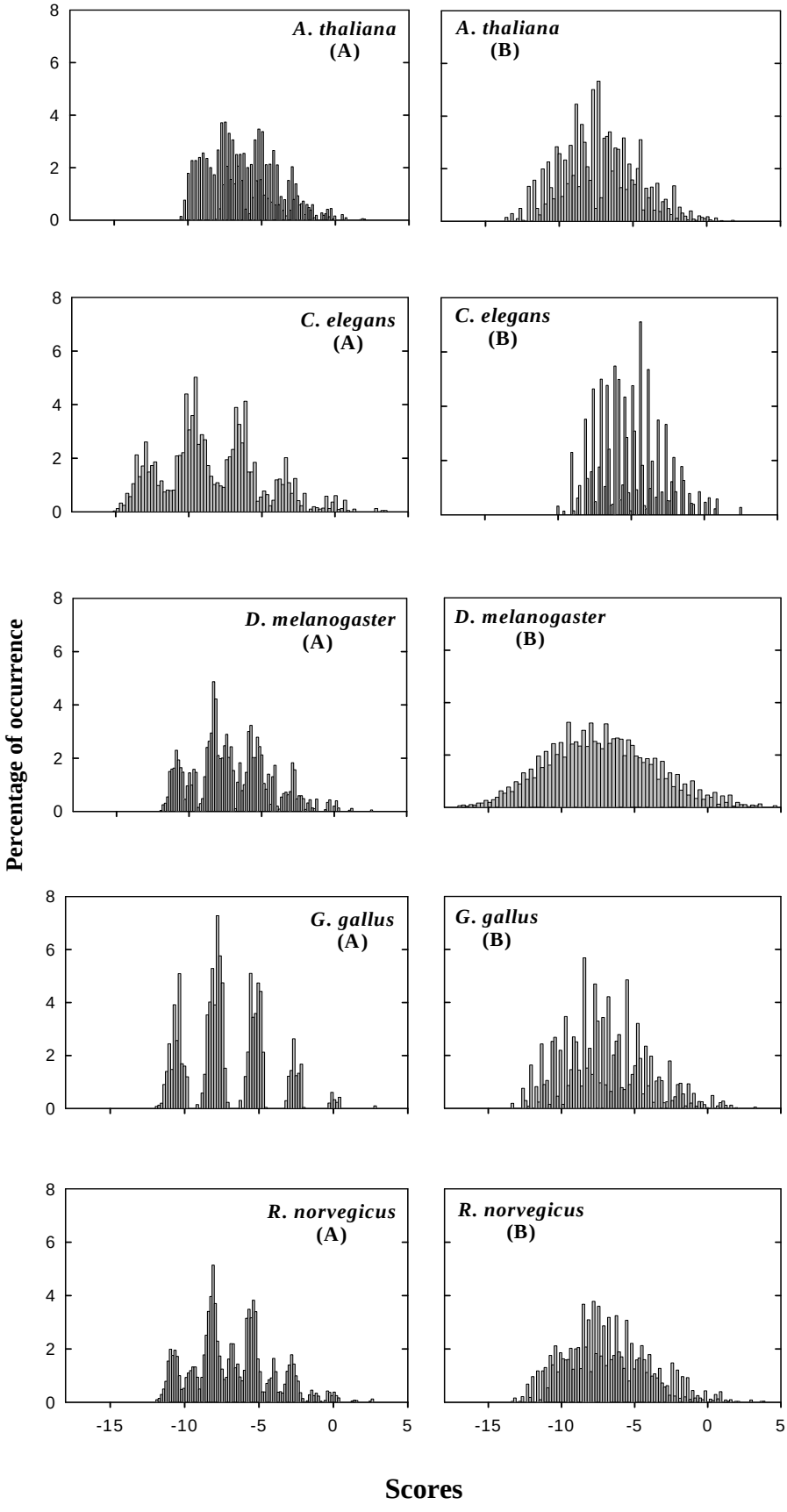

Fig 3. Histograms obtained by plotting the scores values (x-axis) against their percentage of occurrence (y-axis). These values were obtained by scoring the alignment of the (A; left column)

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.

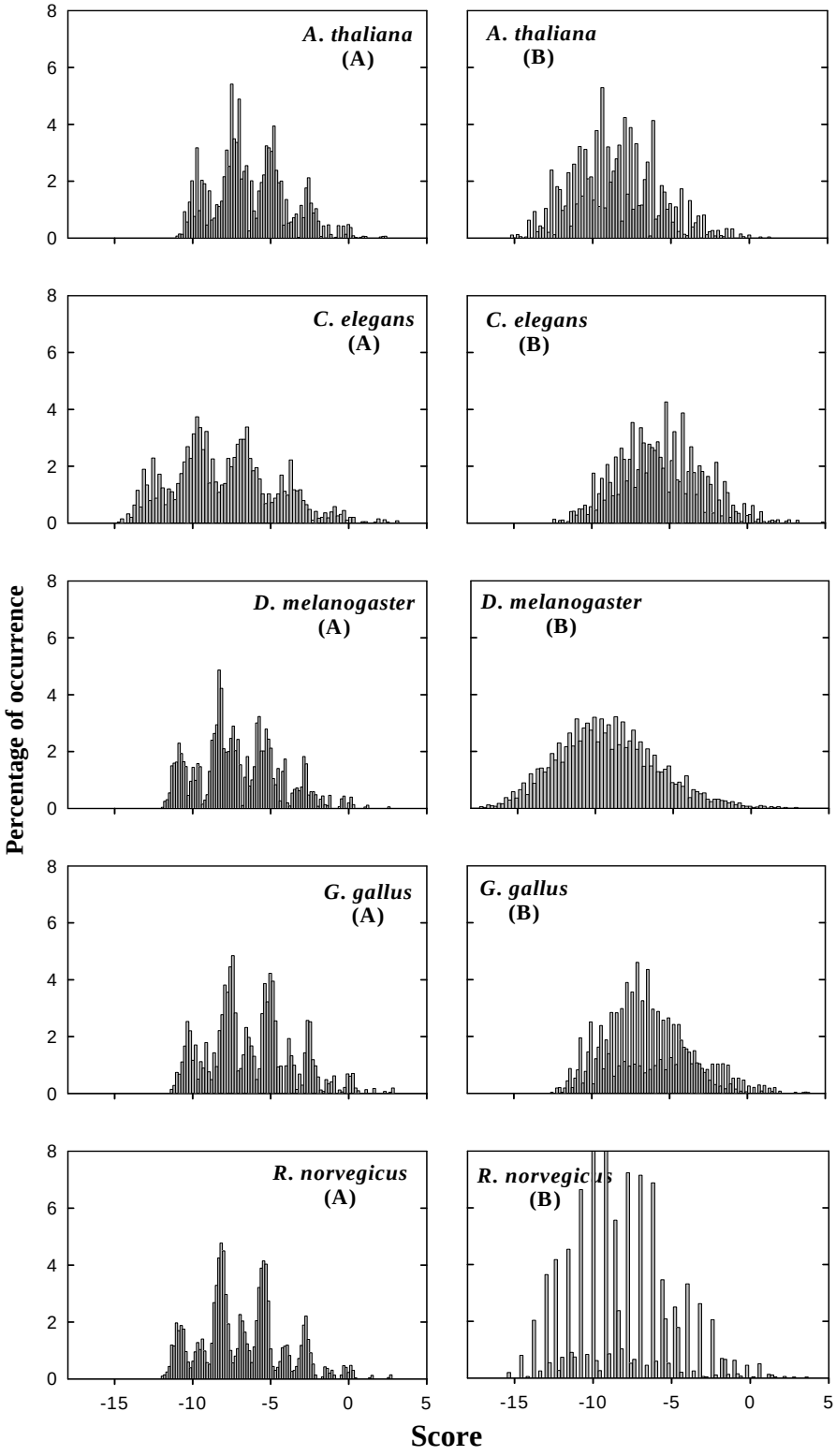

Fig 4. Histograms obtained by plotting the scores values (x-axis) against their percentage of occurrence (y-axis). These values were obtained by scoring the alignment of the (A; left column) the highest frequency unique donor sub-sequence against each of the unique acceptor sub-sequences and similarly 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.