Distribution of Base Composition around the Splice Sites in Different Species

Masahiko Mizuno
mizuno@kuicr.kyoto-u.ac.jp

Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Uji, Kyoto 611, Japan

Abstract

We have analyzed the distribution of base composition around the 5' and 3' splice sites in genomic DNA sequences of different species. A set of sequences belonging to one species is aligned at the 5' and 3' splice sites, respectively, and the average of base composition is calculated for 10 base windows over the range of 100 bases each for upstream and downstream regions. In consistent with the previous observations that coding regions are more guanine-cytosine (GC) rich than noncoding regions, we observe a jump in the GC content at the splice sites, except for vertebrate sequences. In addition, introns are Uracil (U) rich rather than Adenine-Uracil (AU) rich, especially in plants and invertebrates. It is also found that the pyrimidine rich regions preceding the 3' splice site in mammals extend upstream over the consensus sequences, while the polypyrimidine tracts in plants and invertebrates are much shorter than in mammals. Furthermore, the size of increase in pyrimidine content is more striking at the 3' splice site in mammalian, but is smaller in plants and invertebrates. Thus, we consider that the broad and intensive polypyrimidine tract is required for the recognition of the 3' splice site in the higher eucaryotes, where introns are GC rich, and that more AU rich intron is important in the lower eucaryotes.

1 Introduction

Protein coding regions tend to be more GC rich than noncoding regions in genomic DNA sequences. We have previously examined the distribution of GC content around the translation initiation site in different species [1]. Our results suggest that the GC content profiles are characteristic of the species. The overall shape of the GC content profile is similar within each organism group even though the average GC contents can be very different. Furthermore, by

水野　政彦、金久　實：京都大学化学研究所，〒611 京都府宇治市五ヶ庄

examining different profiles for different species, we have found a negative correlation between the average GC content and the jump size at the translation initiation site. Apparently, more AT rich genomes, which tend to lack macroscopic mosaic structures, exhibit more marked differences in the GC content at the microscopic level. Here, we extended our analysis to the 5' and 3' splice sites.

In *Saccharomyces cerevisiae* the 3' region of the intron is recognized by a highly conserved branch point UACUAAC [2]. Vertebrate introns do not have highly conserved branch point sequences, but contain instead a pyrimidine rich tract usually positioned upstream of the 3' splice site [3]. The only sequence elements necessary for intron processing in protoplasts from a dicot plant, *Nicotiana plumbaginifolia* are the splice site and a high A+U nucleotide content in the intron. Neither conserved branch point sequences nor a polypyrimidine tract are found to be essential [4, 5]. Mammalian introns, which are usually not AU rich, are generally not spliced when their processing is tested in transgenic plants or protoplasts [6]. Monocot introns (maize and wheat) are not as efficiently spliced as a dicot intron in transgenic tobacco plants. Maize, but not *N. plumbaginifolia*, can splice natural and synthetic introns which are GC rich, or introns which contain stem-loop structures [7]. These observations suggest the possibility of differences in pre-mRNA splicing between the species.

Although these sequence elements have been extensively studied, they do not seem to be sufficient to identify the splice sites accurately. The evaluation of fitness for the consensus sequences do not always distinguish authentic splice sites from cryptic ones. Also, it has been shown that the context where the consensus is located affects the selection of a splice site [8]. These observations suggest that additional sequence elements may be required for splicing of pre-mRNAs. A purine rich sequence located within the last exon, M2, of the mouse immunoglobulin μ (IgM) gene plays an essential role in the splicing of this gene [9, 10]. This sequence, designated ERS (exon recognition sequence), stimulated splicing of a distant upstream intron. Several exon sequences whose mutation or deletion affects splice site selection contain purine rich sequences similar to the ERS of IgM exon M2 in different species [11]: for example, the avian sarcoma-leukosis virus gene [12], the chicken cardiac troponin T gene [13], the human hypoxanthine-guanine phosphoribosyltransferase gene [14], the bovine growth hormone gene [15], and the rat beta-tropomyosin gene [16]. Some of the polypurine tracts is located between about +20 to +130 base positions, where the first nucleotide of the exon at the 3' splice site is +1. Thus, additional elements may be involved in splice site selection.

In this study, we investigated the distributions of base composition, specifically in the region surrounding the 5' and 3' splice sites, respectively, over the range up to 200 bases long, and considered the differences of splicing in different species. Furthermore, we examined if additional sequence elements and/or the context for splicing could be found.

2 Materials and Methods

2.1 Selection of data

We select from the GenBank database (releases 72 and 73) those genomic sequence entries that contain complete protein coding sequences. Also included are the coding sequences of 500 bases or longer without a stop codon. We exclude mRNA entries, pseudogenes, and genes coded by mitochondria, chloroplasts, kinetoplasts, and plasmids.

Data sets are selected for the following organism groups and species: mammals (human, mouse, rat, bovine, and rabbit), other vertebrates (chicken and *Xenopus*), invertebrates (*Drosophila*, *Caenorhabditis*, and *Plasmodium*), plants (*Arabidopsis, Zea*, rice, and tomato), and fungi (*Saccharomyces* and *Dictyostelium*). In each organism group, most of the species were selected according to the abundance of GenBank entries.

In order to examine possible statistical bias of these data sets, homologous sequences are removed given a threshold level of amino acid sequence homology. For each pair of DNA sequences, the pairwise alignment is made for the translated amino acid sequences using the Dayhoff PAM250 matrix. The DNA sequences coding less than 20 amino acids were not considered here. The degree of homology is defined by the number of exact matches divided by the length of the shorter sequence. If any pair of sequences exceeds the threshold homology level, only the longer sequence is retained in the data set. We choose 25% and 75% as threshold values. All the splice sites are investigated in the nonhomologous data sets, but the splice sites where both sides are noncoding in the 5' or 3' untranslated regions are excluded. Table 1 shows the numbers of sequences and the base composition in the data sets utilized in the present analysis.

2.2 Averaging method

A set of DNA sequences are aligned at the 5' and 3' splice sites and the distribution of base composition is calculated for both upstream and downstream of the sites. A window length of 10 bases is shifted by 5 bases at a time along the nucleotide sequence from the -100 base position to the +100 base position, where the nucleotide immediately downstream of the 5' and 3' splice sites are designated by +1, and the preceding base position by -1. The average of base composition in each of the 10 base windows is calculated accordingly. When a window contains sequences of less than 10 bases, they are not included in the averaging. If a neighbor splice site exists between the base positions -100 and +100, the sequence is truncated at the splice site. The average of base composition in a 10 base window is plotted at the base position in the center of the window.

3 Results

Figure 1 shows the distribution profiles of base composition around the 5' and 3' splice sites, respectively, in the four organism groups. Since the patterns of base composition are similar in the same organism group, we show the results on one or two species in each organism group. The following results are almost the same as in the data sets with 25% as a value of homology threshold. Therefore, we demonstrate the results in the case of 75%. There are increases in purine (AG) content over the 5' splice site for all the species in Figure 1. This result is consistent with the observation that the consensus sequences at the 5' splice site, AG|GU[A/G]AGU (where | designates the splice site) contain an abundance of A or G [17]. Figure 1(a)-(d) contain the profiles for human and chicken, respectively. The GC content profiles are more or less flat. In contrast, there are marked decreases in the GC content around the 5' splice site, while there are marked increases around the 3' splice site, in lower organisms as shown in Figure 1(e)-(l). In other words, plant and invertebrate introns are more AU rich than their coding regions. Moreover, these introns are U rich rather than AU rich as shown in

Figure 1 (e)-(l). The changes in U content are much more than those in A content around the 5' and 3' splice sites. This observation is also confirmed by the comparison between differences in GC content and in U content in Table 1. The differences in GC and U content are characteristic of each organism group.

The last column of Table 1 shows the length of the pyrimidine rich tract preceding the 3' splice site in each of the 16 data sets. For vertebrates in Figure 1(b) and (d), the pyrimidine rich regions extend upstream beyond the consensus region up to around -50 base position. However, for plants and invertebrates in Figure 1(f), (h), (j), and (l), the pyrimidine rich regions are much shorter than those in vertebrates. This result is consistent with the observation that a polypyrimidine tract is not required for splicing in dicot plants [4, 5]. The second last column of Table 1 shows the size of increase in pyrimidine content: the difference between the average and the maximum of the pyrimidine content from the -100 to +100 base position around the 3' splice site. The size of the factor corresponds to the length of the pyrimidine rich region. The higher organisms exhibit the more increase in the pyrimidine content.

4 Discussion

In consistent with our previous observations about the region surrounding the translation initiation site, this study confirms that coding regions are generally GC rich even around the 5' and 3' splice sites. Moreover, the GC profile change between noncoding and coding regions is apparently different for different organism groups. There are more marked changes in GC content in lower organisms than in mammals and other vertebrates. Also, the difference in GC or U content between exon and intron is similar within each organism group, except for bovine, even though the average GC content could be very different. These findings are consistent with our previous study that the distribution profile of GC content around the translation initiation site is characteristic of the organism group which the species belongs to [1].

As shown in Table1 and Figure 1, Plant introns are generally U rich around the splice sites. An U rich tract enhances the recognition of the 3' splice site downstream of the tract in yeast [18]. The tract should be U rather than A rich. These results suggest that the U rich sequences function in the 3' splice site selection as enhancers.

Our results do not show the existence of splicing enhancer elements, ERS [9, 10]. The purine rich sequences are identified in the exon upstream of the 3' splice site. If they distribute uniformly in the upstream exon, our method may not find the presence. However, for the 5' splice sites of human in Figure 1(a), there are somewhat guanine rich regions from around +10 to +60 base position. Mammals except for bovine show the same tendency (data not shown), although other vertebrates, invertebrates, and plants exhibit no G rich tracts in the region as shown in Figure 1(c), (e), (g), (i), and (k). According to the base composition of the regions in Figure 1(a), the regions are presumed to have more G rich and less A rich oligonucleotides. Furthermore, we divided human genes into two groups, relatively AU rich (146 genes, 671 sites) and GC rich ones (116 genes, 429 sites) according to the GC average value (56.1%), and calculated the base compositions for both groups. The G rich region appears more noticeably in the GC rich group and disappears in the AU rich ones. Actually, for the human GC rich group, the most frequent tetranucleotides and pentanucleotides in this region is UGGG and CUGGG, respectively, which is found in 354 and 195, respectively, out of 671 sequences. Those

nucleotides are also G rich. Recognition sequences for splicing may exist in such compositionally biased regions. The regions should be investigated further to examine novel splicing elements.

In conclusion, we suggest the difference of base composition for splicing between vertebrates and lower organisms, i.e., the widely and intensively pyrimidine rich region is necessary for the selection of the 3' splice site in vertebrates, while the AU, especially U, rich intron is important in lower organisms. In addition, the relative profile of base composition around the 5' and 3' splice sites is roughly invariant in the same organism group, although the average GC contents can be very different in different species.

Acknowledgement

This work was supported by the Grant-in-Aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture of Japan. The computation time was provided by the Supercomputer Laboratory, the Institute for Chemical Research, Kyoto University.

References

[1] Mizuno, M. and Kanehisa, M. (1994) *FEBS letters*, in press.

[2] Green, M. R. (1986) *Annu. Rev. Genet.*, 20:671-708.

[3] Sharp, P. A. (1987) *Science*, 235:766-771.

[4] Goodall, G. J. and Filipowicz, W. (1989) *Cell*, 58:473-483.

[5] Goodall, G. J., Kiss, T. and Filipowicz, W. (1991) *Oxford Surveys Plant Mol. Cell Biol.*, 7:255-296.

[6] Pautot, V., Brzezinski, R. and Tepfer, M. (1989) *Gene*, 77:133-140.

[7] Goodall, G. J. and Filipowicz, W. (1991) *EMBO*, 10:2635-2644.

[8] Nelson, K. K. and Green, M. R. (1988) *Genes Dev.*, 2:319-329.

[9] Watakabe, A., Sakamoto, H. and Shimura, Y. (1991) *Gene Expression*, 1:175-184.

[10] Watakabe, A., Tanaka, K. and Shimura, Y. (1993) *Genes Dev.*, 7:407-418.

[11] Tanaka, K., Watakabe, A. and Shimura, Y. (1994) *Mol. Cell. Biol.*, 14:1347-1354.

[12] Fu, X. D., Kats, R. A., Skalka, A. M. and Maniatis, T. (1991) *Genes Dev.*, 5:211-220.

[13] Xu, R., Teng, J. and Cooper, T. A. (1993) *Mol. Cell. Biol.*, 13:3660-3674.

[14] Steingrimsdottir, H., Rowley, G., Dorado, G., Cole, J. and Lehmann, A. R. (1992) *Nucleic Acids Res.*, 20:1201-1208.

[15] Hampson, R. K., Follette, L. La and Rottman, F. M. (1989) *Mol. Cell. Biol.*, 9:1604-1610.

[16] Helfman, D. M., Rich, W. M. and Finn, L. A. (1988) *Genes Dev.*, 2:1627-1638.

[17] Senapathy, P., Shapiro, M. B. and Harris, N. L. (1990) *Methods Enzymol.*, 183:252-278.

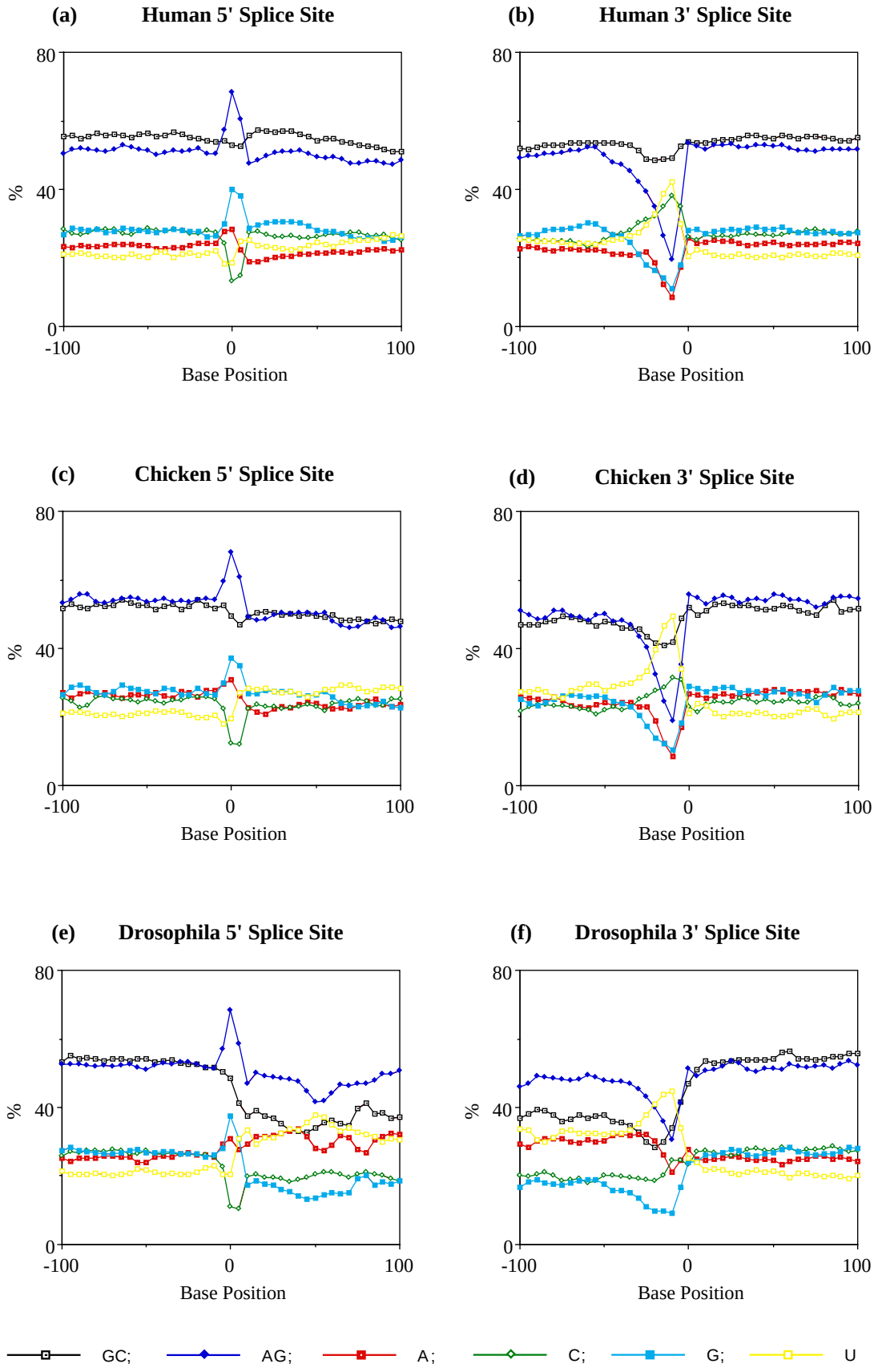
[18] Patterson, B. and Guthrie, C. (1991) *Cell*, 64:181-187.

Table 1. The data sets used for the analysis of the 5' and 3' splice sites.

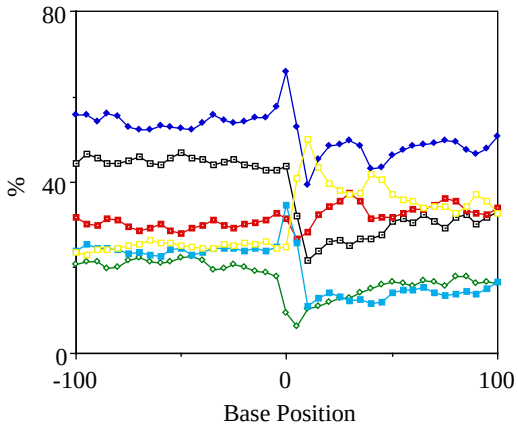
Organism Groups	Genus (Species)	Number of Genes	Number of Sites	Average GC	S.D.	Difference[a] in GC	Difference[a] in U	Py[b] Increase	Py[c] Length
Mammals	*Homo sapiens* 5'	262	1100	54.9	11.1	-1.4	4.0		
	Homo sapiens 3'			53.3	10.8	3.6	-6.8	29.0	55
	Mus 5'	168	654	53.8	7.3	-1.7	4.1		
	Mus 3'			52.4	7.4	4.0	-6.8	29.5	50
	Rattus 5'	106	513	53.5	6.0	-1.3	4.0		
	Rattus 3'			52.5	6.5	3.4	-6.6	29.0	50
	Bos 5'	27	90	49.0	14.2	-11.4	9.1		
	Bos 3'			48.2	15.1	12.1	-10.2	32.6	45
	Oryctolagus cuniculus 5'	14	46	58.5	10.6	0.4	0.3		
	Oryctolagus cuniculus 3'			55.5	12.1	4.2	-5.5	29.4	50
Other Vertebrates	*Gallus gallus* 5'	43	294	50.9	10.9	-3.3	7.0		
	Gallus gallus 3'			49.2	10.5	5.9	-10.3	30.3	40
	Xenopus 5'	17	95	39.4	5.2	-13.0	10.9		
	Xenopus 3'			38.9	4.3	14.9	-14.0	26.1	55
Invertebrates	*Drosophila* 5'	131	363	45.4	6.6	-14.8	9.1		
	Drosophila 3'			45.7	6.5	18.0	-13.6	18.5	40
	Caenorhabditis 5'	59	332	37.9	5.2	-12.6	10.6		
	Caenorhabditis 3'			38.5	4.7	16.5	-11.6	11.6	20
	Plasmodium 5'	22	52	23.7	6.7	-16.7	6.9		
	Plasmodium 3'			25.1	6.5	18.0	-19.5	30.9	45
Plants	*Zea* 5'	22	100	49.4	7.8	-12.8	12.7		
	Zea 3'			49.1	7.6	15.8	-13.5	12.8	45
	Oryza 5'	25	94	44.9	7.3	-19.2	15.9		
	Oryza 3'			44.0	5.8	18.5	-15.2	13.0	25
	Arabidopsis thaliana 5'	47	186	39.1	4.1	-14.5	14.2		
	Arabidopsis thaliana 3'			38.8	4.1	15.5	-16.1	12.2	30
	Lycopersicon 5'	26	89	34.3	6.5	-15.9	15.0		
	Lycopersicon 3'			34.2	6.0	15.1	-15.2	13.1	30
Fungi	*Saccharomyces* 5'	41	43	35.5	4.5	-2.0	6.4		
	Saccharomyces 3'			36.2	4.1	11.8	-10.9	14.3	70
	Dictyostelium 5'	20	32	17.4	4.7	-20.6	17.0		
	Dictyostelium 3'			18.6	3.7	20.5	-18.1	14.7	95

a) The differences in GC and U contents are calculated by the two 100-base regions immediately upstream and downstream of the splice sites.
b) The difference between the average and the maximum of the pyrimidine content around the 3' splice site.
c) The rough estimate of the length of the polypyrimidine tract preceding the 3' splice site.

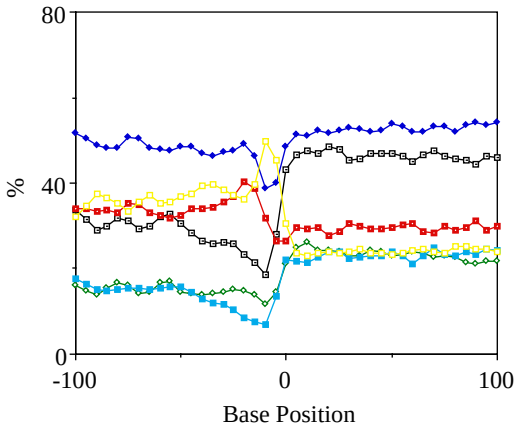
Figure 1 (Next two pages). The distribution profiles of base composition around the 5' and 3' splice sites, respectively, in different species. Each genus or species is classified into one of the four organism groups: (a) and (b) mammals, (c) and (d) other vertebrates, (e)-(h) invertebrates, and (i)-(l) plants. The average of base composition in a ten-base window is plotted against the base position, where the center corresponds to the 5' or 3' splice sites.


(a) Human 5' Splice Site
(b) Human 3' Splice Site
(c) Chicken 5' Splice Site
(d) Chicken 3' Splice Site
(e) Drosophila 5' Splice Site
(f) Drosophila 3' Splice Site
%
80
40
0
-100
0
100
Base Position
GC;
AG;
A;
C;
G;
U

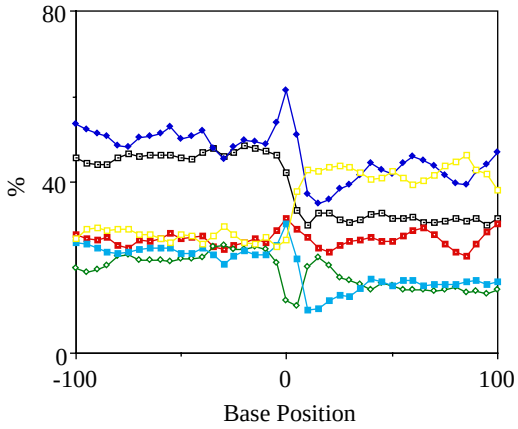
(g) Caenorhabditis 5' Splice Site



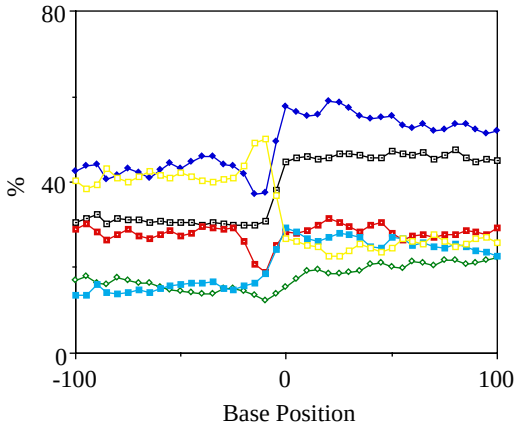
(h) Caenorhabditis 3' Splice Site



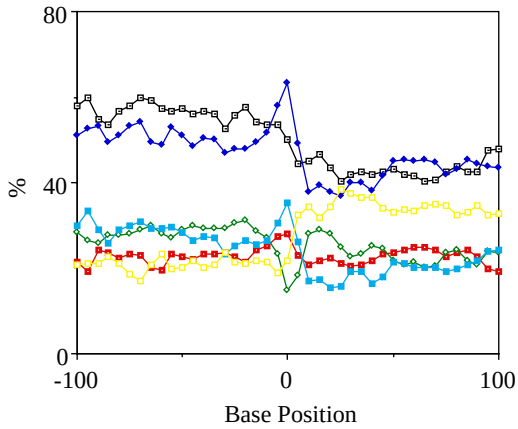
(i) Arabidopsis thaliana 5' Splice Site



(j) Arabidopsis thaliana 3' Splice Site



(k) Zea 5' Splice Site



(l) Zea 3' Splice Site