Computational Analysis of Alternative Splicing and Transcriptional Initiation Patterns

A Dissertation Submitted to
the Graduate School of Life and Environmental Sciences,
the University of Tsukuba
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Science

Hideki NAGASAKI

Table of Contents

1 ABBREVIATIONS

AS	Alternative Splicing
ATI	Alternative Transcriptional Initiation
ASTI	Alternative Splicing and Transcriptional Initiation
DNA	Deoxyribonucleic Acid
cDNA	Complementary DNA
EST	Expression Sequence Tag
RNA	Ribonucleic Acid
mRNA	Messenger RNA
miRNA	Micro RNA
CDS	Coding Sequence
CAGE	Cap-analysis Gene Expression
UTR	Untranslated Region
NMD	Nonsense-mediated mRNA Decay
GO	Gene Ontology
EDI	FN Extra Domain I
ESE	Exonic Splicing Enhancer
ESS	Exonic Splicing Silencer
SR protein	Serine/Arginine-rich RNA-binding Protein
snRNP	Small Nuclear Ribonucleoprotein
hnRNP	Heterogeneous Nuclear Ribonucleoprotein

2 ABSTRACT

The genome sequencing of various organisms was completed recently, and the total number of genes in each organism has been disclosed. In human, the number of protein-coding genes is now estimated to be approx. 25,000, which is much smaller than earlier estimates. Therefore, scientists are shifting their interest to the structural and functional diversity of transcripts (isoforms) produced from a single gene by the mechanisms of alternative splicing (AS) and alternative transcriptional initiation (ATI). This thesis is devoted to the genome-wide detection and analysis of alternative splicing and alternative transcriptional initiation (ASTI), which have become feasible by using newly developed algorithms and computational tools as described here in detail.

We have developed a computational system that automatically performs genome-wide detection and classification of ASTI events, and applied it to six eukaryotes (human, mouse, fruit fly, nematode, cress and rice) whose genome sequencing has been completed or nearly completed. Transcriptional isoforms were collected by mapping a batch of full-length cDNA sequences onto respective cognate genomic sequences. Isoforms mapped on the same gene locus were compared pair-wise, ASTI patterns were segmented into minimal spans, and the minimal patterns (ASTI units) were classified into unique types such as the cassette type or the alternative donor site. All these procedures were performed automatically under the same conditions so that the results obtained from different species could be compared directly. The fraction of loci that underwent ASTI of the total mapped loci was the largest in mammals and fruit fly, and the smallest in plants. Exactly the same trend was observed for the number of unique ASTI types found in each species. The observed fractional representations of the ASTI types were similar between evolutionarily close species such as human and mouse or cress and rice. On the other hand, the relative orders of abundance of individual ASTI types were considerably different between evolutionarily distant species such as mammals and plants. In human and mouse, AS other than the retained introns tended to occur within the protein coding sequence (CDS) regions rather than within the untranslated regions (UTRs), whereas this tendency was obscure in the other four species. In all the species examined, the difference in alternative exon lengths was most likely in multiples of three, and this tendency was most prominent when the alternative exons were embedded within the CDSs. These observations are consistent with the idea

that more complex organisms such as mammals utilize the ASTI mechanisms more extensively and in a more complicated manner than less complex organisms such as nematodes and plants, and that ASTI actively participates in the enhancement of the functional and structural diversity of products generated from a limited number of genes on a genome.

These results of ASTI pattern analysis are open to the public through ASTRA (Alternative Splicing and TRanscription Archives), a database equipped with a JAVA applet viewer that enables users to display the most complex ASTI patterns in the order of their demand (http://alterna.cbrc.jp).

3 INTRODUCTION

3.1 Summary of ASTI

A eukaryotic gene is composed of two nucleotide sequence parts called “exons” and “introns.” An mRNA precursor, the primary transcript from a gene, contains the copy of all parts of the gene. The parts corresponding to exons are retained in mature mRNA, whereas the parts corresponding to introns are removed at the stage of mRNA processing called “splicing.” The 5’ and 3’ ends of an intron are called donor and acceptor splice sites, respectively. For some genes, the exon-intron organization is not unique due to alteration in the selection of splice sites. Therefore, more than one form of mature mRNA can be generated from a single gene. This mechanism is called alternative splicing (AS).

Another mechanism called alternative transcriptional initiation (ATI) also participates in the diversification of transcripts from a single gene. ATI (also called alternative promoter or alternative 5’ end) is the phenomenon by which RNA polymerase starts transcription from different sites driven by different promoters [1]. The promoter activities depend on tissue type, developmental stage and other cellular conditions. The variant 5’ untranslated regions (UTRs) might differ in their secondary structure and/or in the presence or absence of upstream ORFs, which can affect the structure of translation products or translation rates [1] (Figure 3.1).

Although the generation mechanisms of AS and ATI (collectively called ASTI hereafter) are different, both events, independently or in concert, play an important role in increasing the diversity of transcripts of eukaryotic genes.

3.2 Typical examples of AS transcripts

The AS phenomenon is widely observed in eukaryotic cells. Various AS transcripts can be generated depending on tissue type, developmental stage and other cellular conditions [2, 3]. Black [4] has recently reviewed a number of AS events and their regulatory mechanisms. Well-known examples of AS products that carry different functional roles are the SXL proteins related to the sex determination of *Drosophila* [5, 6] and Down’s syndrome cell adhesion molecule (DSCAM) of human and *Drosophila* belonging to the immunoglobulin superfamily [7, 8]. The BMAL2 circadian clock gene of human is also influenced by AS [9]. In addition, many cancer-associated genes such

as Bcl-2, Fas and caspase 2 (Ich-1) are alternatively spliced [10].

AS variation within the coding sequence (CDS) of a gene brings about altered amino acid sequences that eventually lead to structural changes in the protein. In the case of caspase-2 (Ich-1), a member of the caspase family of proteinase, the inclusion or exclusion of an alternatively expressed exon of 61 bp forms two AS variants, Ich-1L and Ich-1S. Inclusion of the 61 bp exon results in translation termination due to a stop codon located upstream of Ich-1S. Therefore, the longer Ich-1L transcript encodes a shorter protein product. The roles of the two isoforms show stark contrast; Ich-1L causes apoptosis whereas Ich-1S prevents it [11].

Functional diversification can be achieved not only at the protein level but also at the mRNA level. For example, an exon insertion into the 5' UTR region inhibits the translation of mRNA of rat DNase I, which is related to the induction of DNA fragmentation in apoptosis [12]. Moreover, an exon insertion into the 3' UTR region of mouse Bcl-xγ mRNA, which carries an important molecular link with the CD28-dependent co-stimulatory pathway, is predicted to change the stability of the mRNA as a result of altered stem and loop structures [13]. Thus, AS in the 5' or 3' UTR region may cause significant changes in the transcriptional or translational efficiency of the gene.

3.3 Genome-wide detection of AS events

The genome sequencing of various organisms was completed recently, and the total number of genes in each organism has become known. When the draft genome sequence of human was opened to the public, the number of genes was estimated to be approx. 35,000, which was much smaller than the earlier consensus of around 100,000 genes [14]. When more accurate sequences were published later, the number of estimated genes was decreased to 20,000-25,000 [15]. This small number of protein coding genes suggests that AS variation might contribute, at least in part, to the functional diversity of human genes [14]. Therefore, scientists have gradually shifted their interest to the structure of transcriptional products and the diverse manners of gene expression that are influenced by AS.

Before human genome sequencing was completed, the methods for detecting AS variants in large scale were based on either clustering of Expression Sequence Tags

(ESTs)/cDNAs or MEDLINE search using “alternative splicing” as the keyword [16, 17]. The completion of genome sequencing has facilitated accurate determination of gene structures through the mapping of ESTs/cDNAs sequences onto the genomic sequences, and has drastically accelerated AS investigations. For example, in earlier studies based on human ESTs, the fraction of genes that generate AS variants was estimated to be approx. 35% of the total genes [18, 19, 20]. More recently, however, Johnson *et al*. [2] reported that at least 74% of human multi-exon genes are alternatively spliced on the basis of the results of microarray analysis with splice junction probes. Furthermore, a variety of genome-wide AS analyses have been developed, including the discovery of regulatory motifs responsible for AS, the prediction of alternatively spliced exons, the comparison of orthologous AS genes, and insights into evolutionary relationships between AS variants and gene duplications [21, 22, 23, 24, 25, 26].

Genome-wide AS analysis has also provided novel knowledge of cancer-associated genes. Xu and Lee [27] conducted genomics-based analyses of the role of AS in cancer according to the human curated EST dataset. A large number (190) of cancer-associated genes show previously uncharacterized splice forms in tumor cells that are distinct from the predominant forms in normal tissues.

3.4 AS patterns and typical AS types

Pair-wise comparison of AS variants has indicated that several distinct patterns are frequently observed in their exon-intron organizations. These typical AS patterns are traditionally classified into seven types, i.e., cassette (one exon insertion), retained intron, alternative donor (shift in 5’ splice site) and acceptor (shift in 3’ splice site), two kinds of alternative polyadenylation site (termination) and mutually exclusive exons [28]. Figure 3.2 shows these typical types, although slightly different classifications are also used in the literature. It has been consistently observed that the cassette type is the most abundant AS type in human transcripts [17, 18]. In most cases, AS patterns other than these typical types are collectively classified as "miscellaneous" or "others." Sharov *et al*. [29] newly detected eight "others" types from mouse mapping results. Thanaraj *et al*. [30] suggested that the AS types classified as "others" may be decomposed into several simpler elementary units.

3.5 Novel knowledge from genome-wide AS analysis

Genome-wide AS analysis has dramatically increased the detection of instances of AS from various species and genes. These observations have significantly contributed to the accumulation of knowledge about the AS phenomenon that extends from the molecular mechanisms of AS selection to medical applications. Some representative findings are introduced below.

3.5.1 NAGNAG motif

Alternative acceptor sites shifted by only three bases were quite frequently detected in mouse transcripts [31]. Those acceptors form the "NAGNAG" motif, where the right side "NAG" represents the constitutive acceptor motif, and the left side "NAG" represents the alternative acceptor motif (N stands for A, C, G or T) [32] (Fig. 3.3). Such motifs are also observed in ruminant, chicken, and tomato, among other organisms [33, 34, 35]. Hiller *et al*. [32] scanned 20,213 human mRNAs from the RefSeq division of GenBank, and found that 5% (8,105 of 152,288) of the splice acceptors contained a NAGNAG motif. Their observations are summarized as follows. (1) 7,326 of 8,105 (90.4%) NAGNAG acceptors were situated upstream of an exon annotated as part of the CDS. (2) The tandem acceptors are biased toward intron phase 1 (40% phase 0, 43% phase 1 and 17% phase 2), which is significantly different from that of all human introns (46% phase 0, 33% phase 1 and 21% phase 2). (3) The intron phase determines the outcome of a NAG insertion/deletion (indel) that includes a single-amino acid insertion/deletion (indel), exchange between a single amino acid and an unrelated dipeptide, and creation or destruction of a stop codon. (4) 73% of orthologous NAGNAG acceptor pairs were conserved between human and mouse (Fig. 3.3).

Gene Ontology (GO) [36] analyses suggest that a significant portion of genes containing the NAGNAG motif encode DNA-binding proteins. This observation implies that the insertion or deletion of a single amino acid may play a role in fine tuning the binding properties of the different DNA-binding protein isoforms [37].

3.5.2 AS diversity according to Alu insertion and exon duplication

Alu repetitive elements are short interspersed elements (SINEs) typically 300 nucleotides long, and account for >10% of the human genome [14, 38]. While some alternative exons include Alu elements, the vast majority (84%) of the Alu-containing exons that appear within the coding region of mRNAs cause a frame shift or a premature termination codon. Alu-containing exons are included in transcripts at lower frequencies than alternatively spliced exons that do not contain the Alu sequence. These results indicate that internal exons that contain the Alu sequence are predominantly, if not always, alternatively spliced [39]. Alu elements are rarely found in constitutive exons, indicating their evolutionary disadvantage. Alu elements are composed of several subfamilies of which the J and S subfamilies are most prevalent. Mutations within these Alu elements sometimes convert the alternative exon in which they reside into a constitutive one. In some unfortunate cases, such mutations lead to disease [40]. Thus, "exonization" of Alu or other transposable elements may accelerate evolution through the creation of new AS exons, but it may also have deleterious effects on the host organism.

Exon duplication is considered to be the principal path for the functional diversification of proteins and the emergence of new functions. Kondrashov and Koonin [41] reported some of the mutually exclusive exons involved in substitution-type AS, and speculated that the alternative exons have evolved by exon duplication.

Considering that the prediction of eukaryotic genes is not perfect, Letunic *et al.* [42] searched for "unannotated" duplex exons from adjacent intron sequences in human, fly and nematode genomes. They identified a total of 2,438 cases, and suggested that exon duplication tends to lead to mutually exclusive AS events. They also presented the idea that AS is a general mechanism for the modulation of protein functions, as suggested by the 3D structural analysis of human glycine receptor alpha-2 and *Drosophila* trypsin-like serine protease [42].

3.5.3 Nonsense-mediated mRNA decay (NMD)

Nonsense-mediated mRNA decay (NMD) is a eukaryotic mRNA surveillance pathway that reduces errors in gene expression by eliminating aberrant mRNAs that encode incomplete polypeptides. Recent experiments suggest a working model whereby

premature and normal translation termination events are distinct as a consequence of the spatial relationship between the termination codon and mRNA-binding proteins, a relationship partially established by nuclear pre-mRNA processing [43].

During pre-mRNA processing, the spliceosome removes intron sequences. As this occurs, a set of proteins called the exon-junction complex are deposited 20-24 nucleotides upstream of the sites of intron removal. For normal mRNAs whose termination codons are on or near the final exon, the ribosome will have displaced all exon-junction complexes. By contrast, if any exon-junction complexes remain when the ribosome reaches the stop codon, a series of interactions ensue, leading to decapping and degradation of the RNA. This model explains the basis of the "50 nucleotide rule" for mammalian NMD: if a termination codon is more than approximately 50 nucleotides upstream of the final exon, it is a premature termination codon and the mRNA that harbors it will be degraded (Fig. 3.4) [44]. Lewis *et al.* [45] noted that approximately 1/3 of the alternative transcripts they examined contained premature termination codons, and are candidate targets of NMD. The widespread coupling of AS and NMD indicates that the cell possesses a large number of irrelevant mRNA isoforms that must be eliminated [45].

3.6 Relationship between ATI and AS events

Although the principal mechanisms of ATI differ from those of AS, there are several lines of evidence suggesting strong coupling between these phenomena. Supportive experiments have been conducted, which involve transient transfection of mammalian cells with minigenes carrying the FN Extra Domain I (EDI) exon that encodes a facultative repeat of fibronectin (FN) [46, 47]. Strong transcriptional activators such as VP16 of class IIB induce skipping of an exon that is included without such an activator [48]. Promoters and enhancers are cis-acting elements that control gene transcription via complex networks of protein-DNA and protein-protein interactions. Both promoters and enhancers can control transcriptional initiation and elongation, of which the elongation rate is tightly correlated with the inclusion or exclusion of the AS exons [46, 49]. Transcriptional coregulators recruited by some steroid hormone receptors are also involved in the control of AS via the same molecular mechanisms [50, 51]. Yet another

line of evidence for the transcriptional control of AS comes from the observation that mutants of polII RNA polymerase with a low elongation rate cause exon insertion, as reviewed in Kornblihtt *et al*. [47].

3.7 Detection of alternative promoters and ATI analysis

A tremendous number of cDNAs have been cloned, sequenced and submitted to public databases energetically from across the world. Although these sequences comprise the primary resources for ASTI analysis, a large fraction of them are copies from truncated mRNAs. Therefore, ATI analysis that needs precise information of transcriptional initiation sites has fallen behind AS analysis, and promoter analysis has long been carried out as an unfixed subject for biologists and bioinformatics researchers [52, 53, 54, 55]. Not only do promoters serve as key to the transcription of genes, they also link to AS, as described above. Hence, detection of the transcriptional initiation site is essential to understand cooperation among these mechanisms. Recent technologies have made it feasible to construct full-length cDNA libraries [56, 57] and Cap-analysis gene expression (CAGE) tag resources [58]. Thus, the most difficult obstacles for ATI analysis have been eliminated. Very recently, Kimura *et al*. [59] analyzed 1,780,295 5'-end sequences of human full-length cDNAs and found that at least 50% of human genes are subject to regulation by alternative promoters, that there are on average 3.1 putative alternative promoters per gene, and that ATI-dependent transcripts are most abundant in testis and brain compared with other tissues. Furthermore, their analysis based on GO indicated that alternative promoters are enriched in the genes encoding signal-transduction-related proteins, but are rare in genes encoding extracellular proteins. These observations suggest varied functional requirements for these classes of genes, *i.e.,* enhanced diversity for signal transduction genes and preserved sorting signals in extracellular protein coding genes, respectively.

3.8 ASTI pattern analysis

3.8.1 Complex mechanisms that control ASTI phenomena

According to recent studies, splicing mechanisms have been elucidated to some degree. The splicing reaction is carried out by the spliceosome that consists of five small

nuclear ribonucleoprotein (snRNP) complexes U1, U2, U4, U5 and U6 and a large number of non-snRNPs. The spliceosome acts through a multitude of RNA-RNA, RNA-protein and protein-protein interactions to precisely excise each intron and to join the exons in the correct order [60]. The SR protein family of non-snRNPs, except for some members such as U2AF35 and 65 that participate in constitutive splicing events, plays important roles in AS [61]. Some consensus motif sequences on pre-mature mRNAs bind such SR proteins and act as exonic splicing enhancers (ESEs). On the other hand, other motifs known as exonic splicing silencers (ESSs) bind another class of RNA-protein complexes called heterogeneous nuclear ribonucleoprotein (hnRNP) complexes. A considerable number of these elements have been identified in human or mouse genes [61, 62].

If some computational methods succeed in the *ab initio* detection of consensus motifs responsible for ASTI [17, 63], alternative exon-intron structures would be predicted [21, 22]. However, the ASTI mechanisms have remained unclear for several reasons. First, the ASTI mechanisms are extraordinarily complicated and thus preclude *ab initio* approaches. Second, all the components of spliceosome have not yet been isolated. Finally, the detection and classification of ASTI patterns from multiple splicing variants have not been well established. Thus, the primary task in investigating the ASTI phenomena is simply to know which parts of the gene (or premature mRNA precursor) are subject to AS variations, and to classify ASTI patterns according to the possible mechanisms of their generation.

3.8.2 Classification of ASTI patterns and AS databases

The splicing events are normally classified as: (i) a cassette or skipped exon, where an entire exon is inserted/deleted; (ii) exon/intron isoforms, where a different donor/acceptor splice site leads to alternative initiation, truncation, or extension of an exon; and (iii) intron retention [64]. In human transcripts, it is commonly recognized that the cassette type is the most popular and each cassette exon affects protein structure without changing the reading frame in a majority of cases. However, the contributions of other AS types have not been extensively studied.

Several research groups have reported the large-scale detection of AS. For

instance, Modrek *et al*. [65] mapped human mRNA and EST sequences on draft genomic sequences, and identified alternative exons as well as constitutive ones. Heber *et al*. [66] applied graph representation to assembled ESTs mapped on genome sequences. By tracing paths of the "splicing graphs," they inferred the exon-intron structure of each AS variant. Zavolan *et al*. [67] aligned full-length cDNA sequences of mouse against the genome sequence. Genomic exons thus identified were compared to one another to detect AS variations. With the increase in the number of mapped cDNA sequences, methods for detecting AS forms have become considerably elaborate, and several databases that present detected AS variants are now publicly accessible [30, 31, 65, 68].

To obtain deeper insights into the mechanisms and biological significance of ASTI events, however, we need more advanced analyses of ASTI patterns. The first step towards this direction is the classification of AS events according to type. So far, there has been no systematic way to do this, which has hampered direct comparison of the results obtained by different groups with different data resources. In the following sections, we propose our algorithm for the detection and classification of ASTI types and its application to six representative eukaryotes. Through the analyses, we show that ASTI actively participates in the diversification of transcriptomes and proteomes derived from a limited number of genes on a genome, and that structurally or behaviorally more complex organisms utilize these mechanisms more extensively than less complex organisms.

4 Materials and Methods

4.1 Algorithms for detection and classification of ASTI patterns

4.1.1 Definitions and algorithms

ASTI variants (isoforms) are mature mRNA forms that retain partially different portions of the same template gene after transcription and processing. For simplicity, we exclude partial, immature, or degraded products from our consideration. In addition, we concentrate most of our attention to pair-wise comparison, although complicated combinatory patterns may appear when many isoforms from a single gene are compared simultaneously.

Consider some isoforms that are aligned onto the genomic sequence. From the alignment, exon-intron structures of respective isoforms are immediately derived. Our algorithm starts with labeling each nucleotide in each variant either 1 or 0 depending on whether that nucleotide lies in an exon or a non-exon (intron or extragenic region), respectively. This procedure produces two-dimensional bit arrays in which each column corresponds to a position in the genomic sequence, each row corresponds to a distinct mRNA, and the label indicates the exonic status. The bit pattern is then compressed so that adjacent columns with the same values are combined (Figure 4.1-I, I' and I").

Isoforms usually have a majority of exons and introns in common; however, we are interested in only the difference between them. Moreover, apparently complicated variations in exon-intron organizations may be decomposed into simpler units. Hence, we define an "ASTI unit" as a minimal span of distinct exon-intron structures flanked by either common exon(s) or extragenic region(s). With the compressed bit arrays mentioned above, the region is represented by a pair of non-identical shortest bit series flanked by either (1, 1) or "extragenic" (0, 0) at an end of a transcript. (To save space, bit arrays are shown side by side rather than up and down.) In the example shown in Figure 4.1, we find four ASTI units indicated by solid bidirectional arrows (the left solid arrow corresponds to three pair-wise ASTI units), whereas the regions indicated by broken arrows are not ASTI units because the bit series are identical (the right broken arrow) or as discussed in the next paragraph (the left broken arrow). ASTI units may be classified into many types, some of which correspond to typical AS patterns such as alternative donor and acceptor splice sites, cassette, mutually exclusive exons, terminal exons with alternative polyadenylation sites, and retained intron [28]. With our system,

these typical patterns are represented by relatively short bit arrays, as shown in Figure 4.2a. Note that bit arrays encoding an AS type are flanked by (1, 1) at both ends, whereas those encoding variants with different transcriptional initiation and termination sites have (0, 0) at the left and right ends, respectively. In fact, the terminal (0, 0) column is dispensable for the unique identification of distinct types because its immediate neighbor cannot be (1, 1) and hence never denotes an AS type. A pair of bit arrays flanked by (0, 0) at both ends without any (1, 1) in between correspond to nested genes and are not counted as an ASTI type.

For better human recognition, we convert a bit series into the corresponding decimal number (Figure 4.1-IV), in which the smaller number is the first component and the larger one is the second component of the two-dimensional integer vector. In this conversion, we omit the terminal (0, 0) columns without loss of information, as noted above. Each decimal representation is specific to each ASTI type. For instance, the aforementioned typical patterns, i.e., alternative donor and acceptor splice sites, cassette, mutually exclusive exons, two patterns of terminal exons with alternative polyadenylation sites, and retained intron, are represented by (9, 13), (9, 11), (17, 21), (69, 81), (17, 20), (9, 12) and (5, 7), respectively (Figure 4.2a). In a computer program, we can use a hash function or an associative array to compactly deal with such an extendable set of multi-component variables. This description system can uniquely and compactly encode not only typical patterns but also any rare patterns that are usually collectively assigned to “others” (Figure 4.2b).

ATI sites are usually regulated by different promoters of a gene, and the molecular mechanisms of transcriptional initiation are quite different from those of splicing [1]. On the other hand, transcriptional termination is tightly coupled with the upstream splicing patterns. Hence, we consider variations in transcriptional initiation separately from the other variations in the analyses described below, although a common classification system is used in both cases (Figure 4.2c).

4.1.2 Conversion of mapping data into bit arrays

The primary information obtained by mapping cDNA sequences onto genomic sequence is the coordinates (positions) of exon boundaries. Base-wise conversion of this

information into bit arrays, as suggested above, is obviously inefficient. Thus, we developed a simple algorithm similar to that used in merge sort to combine two or more already sorted arrays into a single array. We treat a set of isoforms pair-wise or collectively. In either case, the boundary coordinates are processed from left to right with a set of switches that indicate exonic status. In the collective procedure, a priority queue is used to indicate the exon boundary to be processed next (Figure 3.1). This procedure converts the boundary coordinates into the compressed form of bit arrays in $O(KN\log(N))$, where N and K denote the number of isoforms and the average number of boundaries per isoform, respectively. Regions delineated by exonic regions common to all isoforms are treated separately. Rows with the same bit series within such a region are merged to eliminate redundant computations involved in the all-by-all comparisons for the detection of ASTI units. Thus, although the overall computational complexity is $O(KN^2)$, practical computational time may be considerably shortened when ASTI units are sparsely distributed. Our implementation of this algorithm contributed to a 6.6-fold reduction in execution time compared to the original implementation by round robin comparisons of splice variants.

4.2 *In silico* mapping of mRNA sequences onto genomic sequences

The ASTI variants of six organisms, human (*Homo sapiens*), mouse (*Mus musculus*), fruit fly (*Drosophila melanogaster*), nematode (*Caenorhabditis elegans*), cress (*Arabidopsis thaliana*), and rice (*Oryza sativa*), were obtained in two steps: collection of transcriptional isoforms and their classification according to ASTI patterns. The classified categories are referred to as the ASTI types. For the five organisms other than rice, cDNA sequences were obtained from UniGene database. For the UniGene cDNAs, we chose those sequences that presumably code for mature protein CDSs according to the annotation. For rice, a full-length 32k cDNA clone set and information of CDSs were obtained from the Laboratory of Gene Expression, Department of Molecular Genetics, National Institute of Agrobiological Sciences [69] (ftp://cdna01.dna.affrc.go.jp/pub/data/CURRENT). For comprehensible detection of ASTI events, it is obviously disadvantageous not to use a full spectrum of transcripts including ESTs. However, we decided to use only full-length cDNA sequences because

they provide accurate and reliable information about the structure and function of genes as discussed in Seki *et al*. [70]. The genomic sequences of *H. sapiens*, *M. musculus*, *D. melanogaster*, and *A. thaliana* were obtained from NCBI (ftp://ftp.ncbi.nih.gov/genomes/). The genomic sequences of *C. elegans* and the draft contigs of *O. sativa* were obtained from Sanger Center (ftp://ftp.sanger.ac.uk/pub/) and TIGR Institute (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotaion_dbs/pseudomolecules/version_3.0/), respectively.

The cDNA sequences were first mapped onto completed or draft genomic sequences by "MEGABLAST" (ver. 2.2.1) [71]. Table 4.1 shows the data sets used in the present analyses. Hit sequences (>96% identity) were re-computed for consistent alignment with our "ALN" program, which uses a dynamic programming algorithm for spliced alignment [72, 73]. ALN aligned genomic DNA sequences against cDNA sequences allowing for long gaps, taking both nucleotide matches and splicing signals into account. After the initial alignment, the cDNAs were discarded if the aligned regions were less than 85% of their total length or if the average nucleotide identities were less than 97% of the aligned regions. To proceed further from the alignment, we adopted a relatively conservative view as follows. (i) A potential intron must be longer than 30 bp. (ii) Its terminal dinucleotides must be GT..AG, GC..AG or AT..AC. (iii) Each matched region flanking a putative intron must be at least 25 bp long, and at most one mismatch or single-base indel is allowed in a total of 50 bp boundary region. A transcript containing a potential intron(s) that violated one of the above conditions was excluded from the present analyses.

5 RESULTS

5.1 Genes that undergo AS and ATI

The mapping results are summarized in Table 4.1. The percentages of genes that were found to undergo AS (100 x the number of loci that generate AS variants/the total number of successfully mapped loci) varied from 32.1% in human and 23.0% in mouse to 18.6% in fruit fly, 14.1% in cress, 8.1% in rice, and 4.9% in worm. Thus, human and mouse exhibited apparently larger numbers of AS genes and their transcriptional variants than the other organisms. However, it has been argued that the estimated frequency of genes with AS patterns should increase with the number of cDNA samplings [31]. Thus, we conducted a simulation study in which fixed numbers (say 10,000, 20,000…) of cDNA clones were randomly chosen and only the loci that associated with the chosen clones were counted. The results indicated that the fractions of AS genes of mapped loci in human, mouse and fruit fly were 2~2.5 times as large as those in worm, cress and rice (Fig. 5.1). This tendency was observed over the available range of 10,000 to 50,000 chosen clones, suggesting the association of the number of AS events with biological complexity [74]. Interestingly, the number of AS variants per locus (the number of distinct AS variants divided by the number of loci that generate AS variants) was surprisingly uniform in all the six species, ranging between 2.5 in human and 2.1 in rice.

As for ATI, the percentages of relevant genes (100 x the number of loci that generate ATI variants/the total number of successfully mapped loci) varied from 16.1% in human to 1.9% in rice. As most worm cDNA sequences lack 5' UTR (see next subsection), its ATI patterns were omitted from the present analyses. The percentage of ATI loci in human is in good agreement with a manually validated report [1]. In the case of mouse, the percentage was 10.1%, which is in good agreement with the value of 9% reported by Zavolan *et al*. [67]. In mammals, the fractions of ATI genes were approximately half those of AS genes, whereas in plants, they were approximately one-quarter of the AS gene fractions. Insights into this difference require more detailed sequence analysis because alternatively promoted genes in plants are not well documented. The numbers of ATI variants per locus were similar to those of AS variants in all the six species, ranging between 2.3 in fruit fly and 2.0 in rice.

5.2 Distributions of AS types in six species

The AS patterns are conventionally classified into seven representative types: alternative donor splice site, alternative acceptor splice site, cassette (exon skipping or cryptic exon), mutually exclusive exon, two types of terminal exon with alternative polyadenylation sites, and retained intron [28]. Although our scheme classified minor types impartially and equally as these representative types, atypical types were analyzed separately from these representative types for ease of understanding.

Figure 5.2 summarizes the results of classification of the AS patterns for the six eukaryotes studied. At first glance, two points are notable. First, the general features observed in human genes are not necessarily conserved in the other organisms. Second, the overall representations of the AS types of evolutionarily close species, i.e., human and mouse or cress and rice, are similar to each other. Recent genome-wide studies on AS transcripts in mammals have consistently shown that the cassette type has the highest frequency [3, 21, 31, 64, 65, 75]. Our results for human and mouse coincided with those already reported. The cassette type represents 28.8 and 25.3% of the human and mouse AS patterns observed, respectively. On the other hand, the cassette type showed a decrease in the other species, ranging from 20.3% in *C. elegans*, 13.6% in *D. melanogaster*, 8.8% in *A. thaliana*, and 3.4% in *O. sativa* (purple circles in Fig. 5.2). In contrast, fractions of the retained intron were increased in all the species except *C. elegans* (from 30.8% to 55.0%; orange circles in Fig. 5.2).

One feature commonly observed in all the species examined is the difference in relative abundance of the two types of alternative polyadenylation sites: one is the type generated by a permeable donor site leading to the last exon (alternative polyadenylation site 1 in Fig. 5.2), and the other is the alternative acceptor site constituting the last exon (alternative polyadenylation site 2 in Fig. 5.2). The former was consistently much more abundant than the latter, presumably because a simpler process generates more products. Consistent with this idea, mutually exclusive exon, the most complex pattern, was the rarest among all the typical AS types in all the species examined.

Quite notable is that *C. elegans* has an AS type distribution that differs from those of the other species. However, the results of *C. elegans* should be interpreted with

caution. Probably because a considerable fraction of the mRNAs of *C. elegans* are generated by the trans-splicing mechanism [76], most of its "full-length" cDNA sequences from UniGene start from "ATG" and lack the 5' UTR sequences. When we verified the locations of the AS units classified in the five typical AS types except alternative polyadenylation, we became aware that 81.7 to 99.0% of the AS units classified in the respective types were located in the CDSs completely (Table 5.1). This fact may cause some bias in the quantification of its AS types. Indeed, the number of alternative polyadenylation sites seems to be overestimated in *C. elegans*.

5.3 Influence of AS on coding capacity in mammalian genes

For a CDS, it is important to know whether AS affects its downstream reading frame. Hence, we examined the difference in exon length in each AS unit in terms of its effects on the coding capacity. Variable transcriptional initiation or termination was not considered because their impact on coding capacity was indeterminable.

Table 5.1 shows the classification of the five common AS types according to their location within the transcripts. The overall features for human and mouse (upper two panels) are quite similar to each other. Approximately 30-40% of the alternative donor sites were found in the 5' UTRs, whereas the fractions of the other types except the retained introns within the 5' UTR were around 10%. These numbers are much less than those reported by Mironov *et al*. [18]. If we disregard this high proportion of alternative donor sites in the 5' UTRs, only the retained introns would have features distinct from the other types; namely, only 12-14% of the AS units of the retained-intron type are located within the CDSs, whereas more than 65% of the AS units of the other types are within the CDSs. If the AS units located at the boundaries of the CDSs were also included, more than 80% of the AS units other than the retained introns would contribute to the diversification of the protein sequences. The observed fractions of AS units within the CDSs are significantly greater than those estimated from the random distribution of AS units within transcripts, as the average proportions of CDSs in human and mouse mRNAs in our data set are 59.6% and 59.3%, respectively. Our finding that most AS units affect CDSs is consistent with the recent observations in human [65] and mouse [31].

When the difference in length of the alternative exons in each AS unit (Δe) was examined, the cassette-type exons were found to be significantly more enriched with the 0/3-type than the general internal exons ($p < 10^{-36}$ with χ^2-val = 162, df = 1), which are known to be significantly biased for the 0/3-type [77] (Fig. 5.3). [We refer to an AS unit as the r/3-type (r = 0, 1 or 2) when the remainder of the division of Δe by three is r.] When only AS units embedded within the CDSs were considered, the fraction of the 0/3-type was further increased from 52.8 to 60.2% in human (from 57.1 to 63.7% in mouse). The same trend was observed for all AS types, most prominently for mutually exclusive exons.

The length distribution of the retained introns was much broader than that of the general internal exons (data not shown) and the mammalian retained introns were not enriched with the 0/3-type. In fact, the fraction of the 0/3-type of the retained introns was significantly *less* than that of the general internal exons of human genes ($p < 10^{-4}$ with χ^2-val = 18.3, df = 1). This observation, together with the distinct distribution of AS sites within transcripts (Fig. 5.3), strongly indicates that the mechanical and functional properties of the retained introns in mammals are quite different from those of the other AS types.

5.4 Influence of AS on coding capacity in plants and lower animals

Most of the above-mentioned features, i.e., (1) high proportion of alternative donor sites in the 5' UTRs, (2) preferred location of the AS units in the CDSs, (3) tendency of difference in length of the AS exons divided by three, (4) enhancement of this tendency for the AS units within the CDSs, and (5) unique characteristics of the retained introns, are also preserved in the other four species. However, the second feature is not apparent except for the cassette exons in *Arabidopsis*. For *Drosophila*, by far the most frequent location of the retained introns is the 5' UTR (Table 5.1). The 5' UTR is also the most frequent location of the alternative donor sites. The number of variants whose first and second AS units are located on the 5' UTR is 877, which is more than 35% of the total AS units of 3226. For all the species examined except *C. elegans*, which lacks virtually the entire 5' UTR, the 5' UTR is preferred to the 3' UTR as the location of all the AS types except the retained intron. For all the species except *Drosophila*, on the other hand,

the retained introns are found more frequently in the 3' UTRs than in the 5' UTRs. Thus, the retained intron again shows a unique characteristic distinct from the other AS types, suggesting its unique functional roles and generating mechanisms.

The retained intron is the most abundant AS type in fruit fly, cress and rice (Fig. 5.2). Formerly, some AS units classified as retained introns were suspected to be pseudo splicing units generated by such artificial events as alignment of pre-mature cDNAs or contamination of cDNA libraries with genomic DNAs. For plants, however, the retained introns were shown to be a prominent feature of the AS. Ner-Gaon *et al*. [78] recently confirmed by RT-PCR that at least 75% of the sampled intron-retention candidates in *Arabidopsis* (18 of 24 samples) were actually found in transcripts, and that the retained introns were prominent in photosynthesis, stress response and stimulus response. Their results coincide with our observation (Fig. 5.2) and together indicate that the retained introns are abundant in plants and lower animals. It is also noteworthy that the lengths of the retained introns of these species tend to be in multiples of three significantly more frequently than those observed in mammals.

5.5 AS patterns classified as "others"

The AS transcripts classified as "others" in Figure 5.2 consist of complex patterns such as multi-exon insertion or a combination of seven representative AS types. The number of AS patterns classified as "others" is the highest in human (117) and the lowest in rice (17), presumably reflecting the complexity of biological systems and evolutionary relationships.

Figure 5.4 shows the five most abundant "others" types in the six eukaryotes. As in Fig. 5.2, the overall representation of the AS types in evolutionarily close species is similar to each other. The high ranking types tend to be more complicated for relatively complex organisms, i.e., human and mouse. For example, the two-exon insertion is the most abundant type in both human and mouse (33.5 and 33.5%, respectively), and the three-exon insertion is the second most abundant in human and the third most abundant in mouse (9.7 and 6.4%, respectively). On the other hand, such multiple-exon insertions, similar to the cassette type, decrease their share in the rest of the species. In fact, the three-exon insertion is quite rare in fruit fly and plants: the numbers of three-exon

insertions are 10 (1.6% of "others," 0.3% of total), 3 (1.9% of "others," 0.1% of total), and 2 (0.5% of "others," 0.1% of total), respectively. Details on the statistics of the AS types including "others" are available through our Web site (ASTRA database) at http://alterna.cbrc.jp .

5.6 Distribution of ATI patterns in six species

Figure 5.5 shows the five most abundant ATI types for the five eukaryotes. The results of *C. elegans* are omitted for the reasons described above. The general tendency is in parallel to that for the AS types; evolutionarily close species use similar ATI types of similar fractions, whereas the overall features of the relative frequencies are variable among remote species although more conservative than those observed in the AS types. The most abundant type in human, mouse and fruit fly is the pattern that the first exon is located inside the other variant's intron (purple circles in Fig. 5.5). The second most abundant type in those species is the upstream extension of the first exon from the consensus exon part (orange circles in Fig. 5.5). In plants, the order of these two types is reversed, and they account for more than 95% of all the ATI units. It is noteworthy that more complex patterns including exon skipping event(s) are more abundant in the three animals. The fifth most abundant type in rice represents a transcription start point immediately downstream of the acceptor site of the other transcript. Although a few ATI units showing this pattern were detected in four species (1 in *H. sapiens*, 2 in *A. thaliana* and 3 each in *D. melanogaster* and *O. sativa*), the significance of these observations remains to be confirmed.

5.7 ASTRA, a visual database of ASTI patterns

The data mentioned above about the ASTI units for the six organisms (*H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *A. thaliana*, and *O. sativa*) were stored in a database named ASTRA (Alternative Splicing and TRanscription Archives). The ASTI units were pre-computed from full-length cDNA sequences in the UniGene Database and from genomic sequences at NCBI, Sanger Center, and TIGR Institute. All data can be searched by their ASTI types, gene names, GO terms, GenBank accession numbers, UniGene IDs, OMIM IDs, and some properties of AS such as association with NMD

[45] and NAGNAG [31, 32, 37] (Fig. 5.6). On the front page, ASTRA also reports some statistical features characteristic to each species, such as the fractional representations of ASTI types.

ASTI units represent the most elementary localized information about ASTI. However, when a single gene undergoes widely different ASTI events, their underlying mechanisms might be better understood with a graphical display of the ASTI patterns whose alignment order can be customized depending on user's personal purposes. To satisfy this requirement, we designed the graphical interface of ASTRA. The system consists of two components: (1) an SQL-based database engine to provide visual classification of ASTI patterns and (2) an interactive Java-based browser to rearrange ASTI patterns on the client side at the user's command.

The browser is launched when a user chooses a specific UniGene locus in the database. The browser supports zoom in/out of the overview of the chosen locus. By double-clicking the exon and intron boxes, the user can retrieve DNA sequences and the amino acid translations of the chosen splicing variant. Since the order of the splicing diagram can be rearranged, the user can focus only on the splicing patterns of interest.

6 DISCUSSION

6.1 Standardization of classification system

Several research groups have reported the large-scale detection of AS [3, 21, 64, 65, 66, 67, 75]. The classification of observed events according to their patterns should precede detailed studies of the functional implication and mechanisms of AS [79] and ATI. However, different research groups have adopted different categories for the classification [17, 21, 28, 65, 80, 81]. Variants of transcriptional initiation were included in the categories of Kan *et al*. [21] and Zavolan *et al*. [31]; only terminal variations were included in the original category of Breitbart *et al*. [28]; or both were excluded from the categories of Huang *et al*. [80] and Modrek *et al*. [65]. This anarchic situation impedes the direct comparison of independent observations and brings about unnecessary confusion. We consider two factors responsible for this confusion. First, there is no established notion of the "unit" of ASTI. Second, there are no good procedures to consider multiple transcripts at a time. Our proposed notation realizes a systematic and objective description of ASTI patterns based only on sequences, and renders their automatic classification feasible.

6.2 Diversity in ASTI patterns in human transcripts

We found as many as 124 distinct types of elementary AS patterns in human mRNAs. This striking diversity was revealed for the first time by the automatic classification system we have developed. A vast majority of these divergent types are considered genuine for three reasons. First, the transcripts we used are full-length or nearly full-length mRNAs as annotated in the UniGene database. This subset of UniGene entries of high quality may effectively prevent various troubles associated with ESTs and other imperfect sequences. Second, we adopted stringent criteria for the identification of cognate transcripts to obtain a relatively conservative view. Finally, of the 124 AS types, 68 (58%) were represented by more than one independent AS unit. Moreover, 17,883 of 20,330 (88.0%) intron-flanking boundary pairs that comprise the 9,498 AS units were supported by multiple mRNA or EST sequences when we searched the 50-bp-long boundary sequences against the entire UniGene entries by BLASTN [82]. These observations indicate that experimental error or mistakes in data processing are not likely to account for the observed diversity in AS patterns.

6.3 AS types classified as “others”

Atypical AS types are observed in all the six species examined, and are most popular in human with respect to both number and kind, as described in the previous section. Recently, Sharov *et al*. [29] also confirmed 23 distinct ASTI types in mouse transcripts detected by a method based on our preliminary proposal [83]. Thus, atypical ASTI events are not exceptional but rather common phenomena observed widely in eukaryotes. Our observation that approximately 16% of AS units found in human mRNAs belong to “others” suggests their significant contribution to the multi-functionality of many genes. Because these atypical AS patterns produce structural variations that are generally more complicated than common patterns, they may have a greater impact on the diversification of products generated from a single gene. Then, what is the biological significance of atypical AS types? Here we introduce a few interesting examples.

Monarch-1 (NBD-LRR/NACHT/PYPAF) gene on human chromosome 19 is genetically linked to immunological disorders [84]. A cDNA clone, Hs#S4623591 (accession no.: AY116294), derived from Monarch-1 gene has three leucine-rich repeat domains encoded in the 7th to 9th exons inside the CDS region (Fig. 6.1a). By comparison with the transcript Hs#S4623588 (accession no.: AY116207), our algorithm detected a rare type of AS denoted as (100000001, 101010101). The lengths of all the exons are 171 bp, causing no frame shift. The Etandem program included in EMBOSS package 2.9.0 diagnosed that those exons are tandem repeats, although the mutual similarities are weak, ranging from 61 to 67% identities at the nucleotide sequence level, and from 62 to 74% identities at the translated amino acid sequence level (Fig. 6.1b). AS variants that lack one or two of the tandem repeat exons were also found (Fig. 5.1a). Hence, in addition to the above-mentioned AS type with three exon insertions, there exist one (100000101, 101010001), three (1000001, 1010101), one (1010001, 1000101), and four (10001, 10101)-type distinct AS units within this genomic region. Each tandem repeat exon encodes a single domain called ribonuclease-inhibitor (RI)-like leucine-rich repeat as suggested by CD Search [85] on NCBI online BLASTP search site (Fig. 6.1c). Although Williams *et al*. reported that Monarch-1 transcripts encode leucine-rich

repeats in the AS position [84], they did not refer to the tandem repeats of alternative exons. Monarch-1 gene is known as a global inducer of MHC-I (major histocompatibility complex, class I) genes. As the leucine-rich repeat works as a protein recognition motif, the variation in the number of repeat exons may modulate the induction levels of MHC-I genes.

An atypical AS unit of (100011, 101001) type was found in human interleukin 28 receptor alpha (IL-28RA) gene. There is another isoform that lacks both internal exonic regions of these variants, and hence (10001, 10101)-type and (1001, 1011)-type AS units are also present in this region. The middle exon of the former encodes a transmembrane domain, and the extra exonic region in the latter encodes an intracellular domain [86].

A similar situation was observed in human Wilim's tumor 1 (WT1) gene, where an atypical AS unit of (100001, 110101) type is associated with one (1101, 1001)-type and one (10001, 10101)-type AS unit. The internal AS exon encodes 17 amino acids that supposedly modify the transcriptional regulatory property of WT1 [87].

In all the cases of Monarch-1, IL-28RA and WT1 genes, each tally comprising an atypical AS unit (each splice variant of the gene) is also involved in typical AS units, such as cassette and alternative donor/acceptor sites, if a third isoform is used as the counterpart (Fig. 6.2). Thanaraj *et al.* reported complex AS events caused by their flanking exons' extension or truncation [30]. Of the 912 atypical internal AS units we found, 400 are isolated ones whereas 512 share the same genomic region with some other typical/atypical AS units. If we also take atypical alternative transcriptional terminations into account, 446 units are isolated ones whereas 586 units are associated with some other AS units. Thus, more than 55% of the genomic regions corresponding to atypical AS units are associated with other typical/atypical AS units as well. In this sense, a majority of atypical AS units may be viewed as composite products of simpler AS events, often observed at "hot spots" of AS events (Fig. 6.2).

6.4 Gene duplication vs. ASTI

In chapter 5.1, we have shown that the fractions of genes that undergo ASTI vary considerably among species, and also indicated that plant and lower animal (nematode)

genes are less likely involved in ASTI phenomena than higher animal (human, mouse, and fruit fly) genes. Then, how did these species acquire functional diversity to respond to environmental variations? The answer to this question seems to exist in the recently reported genome sequences of plants such as cress and rice. In this subsection, we will discuss alternative strategies of plants and higher animals for the enhancement of functional diversity from a limited number of primordial genes.

6.4.1 Gene diversification due to gene duplication

Gene duplication often generates gene families in an organism to enhance functional diversity. A well-known example is the human globin family that is composed of several members (including pseudo genes) with different expression properties depending on the developmental stage [88].

Analysis of the *Arabidopsis* genome revealed 1,528 tandem arrays containing 4,140 individual genes, the largest cluster of which consists of 23 adjacent members. It is estimated that 17% of all *Arabidopsis* genes are arranged in tandem arrays, while 35% of the predicted *Arabidopsis* proteins are unique in their genome [89]. The proportion of proteins belonging to families of more than five members is 37.4% in *Arabidopsis*. Similar to *Arabidopsis*, rice genome also contains 14% of tandemly duplicated genes [90]. A family of genes encoding glycine-rich protein with 27 copies and one that encodes TRAF/BTB domain protein with 48 copies were found by manual curation from rice [91].

Extensive gene duplications are also observed in *C. elegans*. Although the fraction of duplicated genes in *C. elegans* is not as large as that in *Arabidopsis* genome, only 55.2% of all the genes are unique, and the proportion of proteins belonging to families of more than five members is 24.0%. In addition, 402 duplicated clusters were detected in *C. elegans* genome [92]. On the other hand, in *Drosophila* that generates ASTI genes at a significantly higher rate than that in plants or nematode, gene duplication seems to occur at a much lower rate; it has been reported that 72.5% genes are unique and the proportion of proteins belonging to families of more than five members is only 12.1% [89].

In mammalian genomes, similarly to *C. elegans* and plant genomes, some large

clusters such as variable regions of immunoglobulin and T cell receptor and V1R vomeronasal olfactory receptor genes [93] are present. Are there any differences among organisms in some properties of gene duplication, similar to those observed for ASTI? Transcription factor families are known to be members of multigene families. Shiu *et al*. [94] reported that those families have much higher expansion rates in higher plants such as rice and cress than in animals such as human, fruit fly and nematode. Because new proteins are largely produced by gene duplication, Vogel and Chothia [95] examined the correlation between the expansion of the domain superfamily and the complexity of 38 eukaryotes, where they adopted the number of different cell types as a measure of the complexity of an organism. As a result, they found a strong correlation between the sizes of 194 superfamilies and the complexity of the organism, although half of the 1219 superfamilies they investigated showed no significant correlation. They also recognized three large clusters of correlated expansions, two of which represent specificities to either vertebrates or plants. The above reports suggest that expanding gene superfamilies are considerably different among species across the evolution, especially between vertebrates and plants.

6.4.2 Rice class III chitinase family: An example with paralogues scattered over plant genome

Besides tandemly duplicated genes, some families such as class III chitinase are composed of paralogues that are scattered over the rice genome. Chitinase (EC 3.2.1.14), which hydrolyzes poly-beta-1,4-N-acetyl glucosamine (chitin), is commonly found in prokaryotes, yeasts and higher plants. In plants, chitinase is thought to act as one of the defenses against invading fungi that have cell walls made of chitin, and is induced by plant hormone and abiotic factors such as heavy metals [96]. Plant chitinases are classified into seven classes [97]. Class III chitinase has no significant sequence similarity to other types of plant chitinase, but has a region with weak similarity to prokaryote chitinases. It has been identified in cucumber, *Arabidopsis* [98], tobacco [99] and chick pea [100].

We isolated seven kinds of cDNA clones putatively identified as encoding class III chitinases from cDNA libraries constructed from dichlorophenoxyacetic acid (2,4-

D)- and benzyl adenine (BA)-treated rice callus. Those clones are named C10501, C10728, C00481, C10122, C10150, C10701, and C10923 (accession nos.: D55711, D55713, D55708, D55709, D55710, D55712 and D55714, respectively) (Fig. 6.3).

Figure 6.4 shows the alignment of the deduced amino acid sequences of rice class III chitinases alongside those of other plant class III chitinases. Plant class III chitinases have six conserved cysteine residues that are considered to play an important role in establishing and maintaining the three-dimensional structures of these proteins. The six cysteine residues are also conserved in rice class III chitinases encoded in C10501 and C10728 at exactly the same positions. On the other hand, proteins encoded in C10122, C10701, and C10923 contain only four cysteine residues. The remaining two clones, C00481 and C10150, also lack two of the conserved cysteine residues but have one cysteine residue at a different position. The glutamic acid at the 170th position in the alignment (Fig. 6.4) is considered to be essential for chitinase activity. Putative proteins encoded in C10501 and C10728 have glutamic acid at this position and are thought to be true homologues of class III chitinase. The proteins encoded in the other five cDNAs, which lack the two cysteine residues typical of class III chitinases as described above, include an aspartic acid residue at the usual glutamic acid position within their plausible active sites. This indicates a possible mutation from glutamic acid to aspartic acid at the active site. This kind of mutation has been reported in *Bacillus circulans* WL-12 chitinase A1 and induced a 17,000-fold decrease in Km [101].

Thus, it is unknown whether the rice class III chitinase homologues in which aspartic acid substitutes for active glutamic acid really have chitinase activities. The differences in position and type of key amino acid residues raise the question of whether these five rice cDNAs are the exact homologues of class III chitinases. The overall sequence homology of these five rice class III chitinase homologues to the most similar class III chitinase is around 40%. The proteins encoded in the other two cDNAs, which are thought to be true homologues to class III chitinases, show higher similarity: 68% for the putative protein of C10501 to rubber chitinase and 61.0% for the putative protein of C10728 to chick pea chitinase. The differences between true homologues and relatives of rice class III chitinases might be explained by the concept of superfamily proteins. One well-known example of superfamily proteins is the antitrypsin family

[102]. The amino acid sequences of proteins within a superfamily are similar despite different functions. In the case of the rice cDNAs described here, two of the seven clones encode true homologues of class III chitinases. The other five cDNA clones however, may correspond to members of a superfamily of class III chitinase proteins.

By mapping the 3' UTR sequences of these clones onto the rice linkage map, we found that class III chitinase homologues are scattered over the rice genome (Fig. 6.5). Other examples of paralogous genes that are scattered over the rice genome are the MADS box family, which is involved in flower development [103], and the Zinc finger protein family including CONSTANS, the orthologues of which are known to control flowering in *Arabidopsis* [104].

6.4.3 Relationship between ASTI and gene duplication: alternative methods for production of gene diversity

AS and gene duplication are distinct evolutionary mechanisms that provide the host organism with raw materials for new biological function. Kopelman *et al*. [25] have reported an inverse relationship between the size of a gene's family and its use of alternatively spliced isoforms in human, mouse, puffer fish, fruit fly and yeast. They described that the inverse relationship between gene duplication and AS was independent of family-size binning, number of exons per gene and expressed-sequence tag coverage.

We used full-length cDNA data sets from six eukaryotes to identify the structural conformation of respective splice variants. Our estimations of ASTI genes described in Subsection 5.1 were based on the assumption that these cDNA clones were isolated from a random pool. In reality, however, some cDNA clone libraries were constructed after an artificial normalization process that facilitated faster coverage of entire genes with decreased redundancy [104]. On the other hand, the normalization makes it difficult to detect all the hidden internal AS forms and to correctly estimate the number of ASTI genes. The bias may differ among organisms and may influence the estimation of ASTI genes to various degrees. Nevertheless, the differences in the fractions of ASTI genes observed among the six species would be significant.

Considering our observations that the fractions of ASTI genes in *C. elegans* and

plants are less than half of those in human, mouse and fruit fly, as well as the reported numbers of gene duplications in plant genomes [88, 89, 93, 94], we propose the hypothesis that the smaller numbers of ASTI genes in plants compared with those in more complex organisms might be compensated for by more frequent gene duplications. The hypothesis is consistent with the observations of Kopelman *et al*. [25] as mentioned above.

6.5 Utility of ASTRA for analysis of complex ASTI patterns

The primary task of ASTRA database is to store ASTI units detected in various organisms and to provide a user-friendly graphic interface for the retrieval of these data together with related information such as the relevant genomic sequence. In addition, one important function of ASTRA is to provide the user with an interactive tool for the analysis of complex ASTI patterns.

As discussed above, some genomic regions can be regarded as ASTI hot spots within which various splicing variations are observed (*e.g*., Figure 6.2). Because our algorithm for detecting ASTI units is based on the pair-wise comparison of isoforms, it is not convenient to grasp an overall ASTI pattern represented by many isoforms. The visual interface of ASTRA can compensate for the limitation of pair-wise analyses. For example, the pair-wise analysis detects ten AS units in the genomic region shown in Figure 4a, while their mutual relationships are easily recognized by the graphical representation. The flexible user interface of ASTRA (*e.g*., easy access to nucleotide/amino acid sequences, and rearrangement of the order of the aligned splice variants) would help researchers gain new hints from their investigations.

Another function of ASTRA is to present species-specific statistical properties, such as fractional representations of various ASTI types. Because ASTI is a mechanism that enhances transcriptome complexity of an organism with a limited number of genes, the complexity in ASTI is expected to be correlated with the functional and structural complexity of the organism, as shown in Section 4. We are in the process of adding more information to ASTRA to deepen our understanding of species specificity of ASTI mechanisms.

6.6 Applications and future directions

In this thesis, I have presented a new computational algorithm for the automatic detection and classification of ASTI patterns, and its application to the analysis of species-specific characteristics of ASTI phenomena in six eukaryotes. Through the analyses, it has become clear that ASTI events play a vital role in the structural and functional diversification of gene products to variable extents depending on the complexity of the host organism. However, we are only starting to understand the extremely complex mechanisms and the profound biological influence of ASTI. There are many hidden facts to be discovered in the near future. Some of the problems that need to be tackled immediately are discussed below.

(1) Since an AS hot spot can generate various transcripts, it is expected to contribute significantly to the functional diversity of a gene. It is highly likely that regulatory *cis* elements are concentrated around such an AS hot spot, and hence it is a good target for the discovery of specific consensus motifs or some other common features responsible for AS variation.

(2) In the present investigation, we have concentrated our attention to the classification of ASTI types without examining the expression levels of individual variants. Obviously, however, ASTI phenomena are tightly coupled with tissue-specific or stage-specific expression levels of the variants. Thus, it is important to investigate ASTI classification in relation to their expression patterns. We are planning to incorporate EST information for the analysis of ASTI transcript expression.

(3) In relation to the above problem, one important next direction is to analyze the potential cooperation between AS and ATI from the viewpoint of ASTI classification. As our algorithm detects not only AS units but also ATI units in the same framework, we are ready to perform such studies with CAGE and other resources now available for promoter analysis.

(4) Another emerging problem in relation to alternative transcription is cooperation between micro RNAs (miRNAs) and alternative polyadenylation sites. MiRNAs are 21-25-nucleotide-long RNAs expressed in a wide variety of organisms ranging from plants to worms and mammals. Some miRNAs such as *C. elegans* lin-4 inhibit protein synthesis by binding incomplete complementary sequences at multiple

locations in the 3' UTRs of specific mRNAs [105]. As the target sites for miRNAs on 3' UTRs are altered by 3' terminal AS events, the lifetime of each AS variant is supposed to be drastically changed. The third and fourth items would follow the expression analysis of ASTI variants described in the second item.

(5) Genome sequencing and cDNA sequencing projects have been carried out in many species. Thus, we are currently planning to extend our analysis of ASTI patterns to a wider range of eukaryotes including chicken, *Xenopus*, and zebrafish as representatives of bird, amphibian, and fish, respectively, and also various fungi and protists. By increasing the number of species, we can enlarge the samples used for comparative analysis of ASTI patterns among organisms. As a result, the difference in ASTI mechanisms among species will be understood more completely. In particular, it is urgent to study the relationships among the extent of gene duplication, the quality and quantity of ASTI, and the complexity of the organisms. In order to cope with the increasing amount of data obtained from various sources, we are also planning to improve our computer system for (semi)-automatic update of the ASTRA database.

(6) One important area of ASTI studies is the application to medical sciences. It is now well known that ASTI, especially AS, is deeply involved in diseases such as cancer and neurological diseases. ASTI is investigated not only as a diagnostic resource but also as a target of therapeutic treatment, *e.g.,* interference of pathogenic AS variants by antisense oligonucleotides. Such methods have been tried to control the expression levels of AS variants related to genetic disease such as Duchenne muscular dystrophy (DMD) [106] and Bcl-x [107] genes. We want to incorporate such disease-related information into a future version of ASTRA database. We hope that accumulating knowledge and improvement in analytical methods of alternative transcripts will eventually contribute to the advancement of human welfare.

7 REFERENCES

1. Landry, J. R., Mager, D. L., Wilhelm, B. T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet.*, **19**, 640-648.

2. Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., Shoemaker, D. D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141-2144.

3. Xu, Q., Modrek, B., Lee, C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754-3766.

4. Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291-336.

5. Sosnowski, B. A., Belote, J. M., McKeown, M. (1989) Sex-specific alternative splicing of RNA from the transformer gene results from sequence-dependent splice site blockage. *Cell*, **58**, 449-459.

6. Inoue, K., Hoshijima, K., Sakamoto, H., Shimura, Y. (1990) Binding of the Drosophila sex-lethal gene product to the alternative splice site of transformer primary transcript. *Nature*, **344**, 461-463.

7. Yamakawa, K., Huot, Y. K., Haendelt, M. A., Hubert, R., Chen, X. N., Lyons, G. E., Korenberg, J. R. (1998) DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Hum. Mol. Genet.*, **7**, 227-237.

8. Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., Zipursky, S. L. (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, **101**, 671-684.

9. Schoenhard, J. A., Eren, M., Johnson, C. H., Vaughan, D. E. (2002) Alternative splicing yields novel BMAL2 variants: tissue distribution and functional characterization. *Am. J. Physiol. Cell Physiol.*, **283**, C103-C114.

10. Mercatante, D., Kole, R. (2000) Modification of alternative splicing pathways as a potential approach to chemotherapy. *Pharmacol. Ther.*, **85**, 237-243.

11. Jiang, Z. H., Zhang, W. J., Rao, Y., Wu, J. Y. (1998) Regulation of Ich-1 pre-mRNA alternative splicing and apoptosis by mammalian splicing factors. *Proc. Natl. Acad. Sci. U S A*, **95**, 9155-9160.

12. Basnakian, A. G., Singh, A. B., Shah, S. V. (2002) Identification and expression of deoxyribonuclease (DNase) I alternative transcripts in the rat. *Gene*, **289**, 87-96.

13. Yang, X. F., Ye, Q., Press, B., Han, R. Z., Bassing, C. H., Sleckman, B. P., Alt, F. W., Cantor, H. (2002) Analysis of the complex genomic structure of Bcl-x and its relationship to Bcl-x (gamma) expression after CD28-dependent costimulation. *Mol. Immunol.*, **39**, 45-55.

14. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

15. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.

16. Ji, H., Zhou, Q., Wen, F., Xia, H., Lu, X., Li, Y. (2001) AsMamDB: an alternative

splice database of mammals. *Nucleic Acids Res.*, **29**, 260-263.

17. Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., Zhang, M. Q. (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol.*, **19**, 739-756.

18. Mironov, A. A., Fickett, J. W., Gelfand, M. S. (1999) Frequent alternative splicing of human genes. *Genome Res.,* **12**, 1288-1293.

19. Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., Luft, F., Reich, J., Bork, P. (1999) Alternative splicing of human genes: more the rule than the exception? *Trends Genet*. **15**, 389-390.

20. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., Bork, P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83-86.

21. Kan, Z., Rouchka, E. C., Gish, W. R., States, D. J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889-900.

22. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G., Shamir, R. (2004) A non-EST-based method for exon-skipping prediction. *Genome Res.*, **14**, 1617-1623.

23. Modrek, B., Lee, C. J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**, 177-180.

24. Thanaraj, T. A., Clark, F., Muilu, J. (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Res.*, **31**, 2544-2552.

25. Kopelman, N. M., Lancet, D., Yanai, I. (2005) Alternative splicing and gene

duplication are inversely correlated evolutionary mechanisms. *Nat. Genet.*, **37**, 588-589.

26. Su, Z., Wang, J., Yu, J., Huang, X., Gu, X. (2006) Evolution of alternative splicing after gene duplication. *Genome Res.*, **16**, 182-189.

27. Xu, Q., Lee, C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, **31**, 5635-5643.

28. Breitbart, R. E., Andreadis, A., Nadal-Ginard, B. (1987) Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu. Rev. Biochem.*, **56**, 467-495.

29. Sharov, A. A., Dudekula, D. B., Ko, M. S. (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. *Genome Res.*, **15**, 748-754.

30. Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J. J., Le Texier, V., Muilu, J. ASD: the Alternative Splicing Database. (2004) *Nucleic Acids Res.*, **32**, D64-D69.

31. Zavolan, M., Kondo, S., Schoenbach, C., Adachi, J., Hume, D. A., Hayashizaki, Y., Gaasterland, T. (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.*, **13**, 1290-1300.

32. Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., Platzer, M. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.*, **36**, 1255-1257.

33. Ferranti, P., Lilla, S., Chianese, L., Addeo, F. (1999) Alternative nonallelic deletion is constitutive of ruminant alpha(s1)-casein. *J. Protein Chem.*, **18**, 595-602.

34. Rogina, B., Upholt, W. B. (1995) The chicken homeobox gene Hoxd-11 encodes two alternatively spliced RNA species. *Biochem. Mol. Biol. Int.*, **35**, 825-831.

35. Li, L., Howe, G. A. (2001) Alternative splicing of prosystemin pre-mRNA produces two isoforms that are active as signals in the wound response pathway. *Plant Mol. Biol.*, **46**, 409-419.

36. Lewis, S., Ashburner, M., Reese, M. G. (2000) Annotating eukaryote genomes. *Curr. Opin. Struct. Biol.*, **10**, 349-354.

37. Akerman, M., Mandel-Gutfreund, Y. (2006) Alternative splicing regulation at tandem 3' splice sites. *Nucleic Acids Res.*, **34**, 23-31.

38. Li, W. H., Gu, Z., Wang, H., Nekrutenko, A. (2001) Evolutionary analyses of the human genome. *Nature*, **409**, 847-849.

39. Sorek, R., Ast, G., Graur, D. (2002) Alu-containing exons are alternatively spliced. *Genome Res.*, **12**, 1060-1067.

40. Kreahling, J., Graveley, B. R. (2004) The origins and implications of alternative splicing. *Trends Genet.*, **20**, 1-4.

41. Kondrashov, F. A., Koonin, E. V. (2001) Origin of alternative splicing by tandem exon duplication. *Hum. Mol. Genet.*, **10**, 2661-2669.

42. Letunic, I., Copley, R. R., Bork, P. (2002) Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.*, **11**, 1561-1567.

43. Baker, K. E., Parker, R. (2004) Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr. Opin. Cell Biol.*, **16**, 293-299.

44. Hillman, R. T., Green, R. E., Brenner, S. E. (2004) An unappreciated role for RNA surveillance. *Genome Biol.*, **5**, R8.1-8.16.

45. Lewis, B. P., Green, R. E., Brenner, S. E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U S A*, **100**, 189-192.

46. Kadener, S., Fededa, J. P., Rosbash, M., Kornblihtt, A. R. (2002) Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. *Proc. Natl. Acad. Sci. U S A*, **99**, 8185-8190.

47. Kornblihtt, A. R., de la Mata, M., Fededa, J. P., Munoz, M. J., Nogues, G. (2004) Multiple links between transcription and splicing. *RNA*, **10**, 1489-1498.

48. Nogues, G., Kadener, S., Cramer, P., Bentley, D., Kornblihtt, A. R. (2002) Transcriptional activators differ in their abilities to control alternative splicing. *J. Biol. Chem.* **277**, 43110-43114.

49. Yankulov, K., Blau, J., Purton, T., Roberts, S., Bentley, D. L. (1994) Transcriptional elongation by RNA polymerase II is stimulated by transactivators. *Cell*, **77**, 749-759.

50. Auboeuf, D., Honig, A., Berget, S. M., O'Malley, B. W. (2002) Coordinate regulation of transcription and splicing by steroid receptor coregulators. *Science*, **298**, 416-419.

51. Auboeuf, D., Dowhan, D. H., Kang, Y. K., Larkin, K., Lee, J. W., Berget, S. M., O'Malley, B. W. (2004) Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes. *Proc. Natl. Acad. Sci. U S A*, **101**, 2270-2274.

52. Scherf, M., Klingenhoff, A., Werner, T. (2000) Highly specific localization of

promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.*, **297**, 599-606.

53. Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Nakai, K., Sugano, S. (2004) Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. *Genome Res.*, **14**, 1711-1718.

54. FitzGerald, P. C., Shlyakhtenko, A., Mir, A. A., Vinson, C. (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, **14**, 1562-1574.

55. Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338-345.

56. Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., Sugano, S. (1997) Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene*, **200**, 149-156.

57. Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocher, M., Blocker, H., Bauersachs, S., Blum, H., Lauber, J., Dusterhoft, A., Beyer, A., Kohrer, K., Strack, N., Mewes, H. W., Ottenwalder, B., Obermaier, B., Tampe, J., Heubner, D., Wambutt, R., Korn, B., Klein, M., Poustka, A. (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.*, **11**, 422-435.

58. Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., Hayashizaki, Y. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U S A*, **100**, 15776-15781.

59. Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H., Ishii, S., Sugiyama, T., Saito, K., Isono, Y., Irie, R., Kushida, N., Yoneyama, T., Otsuka, R., Kanda, K., Yokoi, T., Kondo, H., Wagatsuma, M., Murakawa, K., Ishida, S., Ishibashi, T., Takahashi-Fujii, A., Tanase, T., Nagai, K., Kikuchi, H., Nakai, K., Isogai, T., Sugano, S. (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.*, **16**, 55-65.

60. Hastings, M. L., Krainer, A. R. (2001) Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.*, **13**, 302-309.

61. Graveley, B. R. (2000) Sorting out the complexity of SR protein functions. *RNA*, **6**, 1197-1211.

62. Zhu, J., Mayeda, A., Krainer, A. R. (2001) Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol. Cell*, **8**, 1351-1361.

63. Nurtdinov, R. N., Artamonova, I. I., Mironov, A. A., Gelfand, M. S. (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.*, **12**, 1313-1320.

64. Clark, F., Thanaraj, T. A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451-464.

65. Modrek, B., Resch, A., Grasso, C., Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850-2859.

66. Heber, S., Alekseyev, M., Sze, S. H., Tang, H., Pevzner, P. A. (2002) Splicing graphs and EST assembly problem. *Bioinformatics*, **18,** Suppl 1, S181-188.

67. Zavolan, M., van Nimwegen, E., Gaasterland, T. (2002) Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Res.*, **12**, 1377-1385.

68. Lee, B. T., Tan, T. W., Ranganathan, S. (2004) DEDB: a database of Drosophila melanogaster exons in splicing graph form. *BMC Bioinformatics*, **5**, 189.

69. Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., Hotta, I., Kojima, K., Namiki, T., Ohneda, E., Yahagi, W., Suzuki, K., Li, C. J., Ohtsuki, K., Shishiki, T., Otomo, Y., Murakami, K., Iida, Y., Sugano, S., Fujimura, T., Suzuki, Y., Tsunoda, Y., Kurosaki, T., Kodama, T., Masuda, H., Kobayashi, M., Xie, Q., Lu, M., Narikawa, R., Sugiyama, A., Mizuno, K., Yokomizo, S., Niikura, J., Ikeda, R., Ishibiki, J., Kawamata, M., Yoshimura, A., Miura, J., Kusumegi, T., Oka, M., Ryu, R., Ueda, M., Matsubara, K., Kawai, J., Carninci, P., Adachi, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Hayatsu, N., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Kondo, S., Konno, H., Miyazaki, A., Osato, N., Ota, Y., Saito, R., Sasaki, D., Sato, K., Shibata, K., Shinagawa, A., Shiraki, T., Yoshino, M., Hayashizaki, Y., Yasunishi, A. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, **301**, 376-379.

70. Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., Muramatsu, M., Hayashizaki, Y., Kawai, J., Carninci, P., Itoh, M., Ishii, Y., Arakawa, T., Shibata, K., Shinagawa, A., Shinozaki, K. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141-145.

71. Zhang, Z., Schwartz, S., Wagner, L., Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203-214.

72. Gotoh, O. (1990) Optimal sequence alignment allowing for long gaps. *Bull. Math. Biol.*, **52**, 359-373.

73. Gotoh, O. (2000) Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps. *Bioinformatics*, **16**, 190-202.

74. Kim, H., Klein, R., Majewski, J., Ott, J. (2004) Estimating rates of alternative splicing in mammals and invertebrates. *Nat. Genet.*, **36**, 915-916.

75. Black, D. L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367-370.

76. Blumenthal, T. (1995) Trans-splicing and polycistronic transcription in Caenorhabditis elegans. *Trends Genet.*, **11**,132-136.

77. Sakharkar, M., Passetti, F., de Souza, J. E., Long, M., de Souza, S. J. (2002) ExInt: an Exon Intron Database. *Nucleic Acids Res.*, **30**, 191-194.

78. Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R., Fluhr, R. (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J.*, **39**, 877-885.

79. Smith, C. W, Valcarcel, J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381-388.

80. Huang, Y. H., Chen, Y. T., Lai, J. J., Yang, S. T., Yang, U. C. (2002) PALS db: Putative Alternative Splicing database. *Nucleic Acids Res.*, **30**, 186-190.

81. Hide, W. A., Babenko, V. N., van Heusden, P. A., Seoighe, C., Kelso, J. F. (2001)

The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res.*, **11**, 1848-1853.

82. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

83. Nagasaki, H., Suwa, M. Gotoh, O. (2003) An algorithm for classification of alternative splicing and transcriptional initiation and its genome-wide application. *Genome Inform.*, **14**, 424-425.

84. Williams, K. L., Taxman, D. J., Linhoff, M. W., Reed, W., Ting, J. P. (2003) Cutting edge: Monarch-1: a pyrin/nucleotide-binding domain/leucine-rich repeat protein that controls classical and nonclassical MHC class I genes. *J. Immunol.*, **170**, 5354-5358.

85. Marchler-Bauer, A., Bryant, S. H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.*, **32**, W327-W331.

86. Sheppard, P., Kindsvogel, W., Xu, W., Henderson, K., Schlutsmeyer, S., Whitmore, T. E., Kuestner, R., Garrigues, U., Birks, C., Roraback, J., Ostrander, C., Dong, D., Shin, J., Presnell, S., Fox, B., Haldeman, B., Cooper, E., Taft, D., Gilbert, T., Grant, F. J., Tackett, M., Krivan, W., McKnight, G., Clegg, C., Foster, D., Klucher, K. M. (2003) IL-28, IL-29 and their class II cytokine receptor IL-28R. *Nat. Immunol.*, **4**, 63-68.

87. Wang, Z. Y., Qiu, Q. Q., Huang, J., Gurrieri, M., Deuel, T. F. (1995) Products of alternatively spliced transcripts of the Wilms' tumor suppressor gene, wt1, have altered DNA binding specificity and regulate transcription in different ways. *Oncogene*, **10**, 415-422.

88. Johnson, R. M., Gumucio, D., Goodman, M. (2002) Globin gene switching in primates. *Comp. Biochem. Physiol. A. Mol. Integr. Physiol.*, **133**, 877-883.

89. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.

90. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793-800.

91. Song, R., Llaca, V., Messing, J. (2002) Mosaic organization of orthologous sequences in grass genomes. *Genome Res.*, **12**, 1549-1555.

92. *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012-2018.

93. Del Punta, K., Leinders-Zufall, T., Rodriguez, I., Jukam, D., Wysocki, C. J., Ogawa, S., Zufall, F., Mombaerts, P. (2002) Deficient pheromone responses in mice lacking a cluster of vomeronasal receptor genes. *Nature*, **419**, 70-74.

94. Shiu, S. H., Shih, M. C., Li, W. H. (2005) Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol.*, **139**, 18-26.

95. Vogel, C., Chothia, C. (2006) Protein family expansions and biological complexity. *PLoS Comput. Biol.*, **2**, 0370-0382.

96. Collinge, D. B., Kragh, K. M., Mikkelsen, J. D., Nielsen, K. K., Rasmussen, U., Vad, K. (1993) Plant chitinases. *Plant J.*, **3**, 31-40.

97. Kasprzewska A. (2003) Plant chitinases - regulation and function. *Cell Mol. Biol. Lett.*, **8**, 809-824.

98. Samac, D. A., Hironaka, C. M., Yallaly, P. E., Shah, D. M. (1990) Isolation and characterization of the genes encoding basic and acidic chitinase in *Arabidopsis*

thaliana. *Plant Physiol*., **93**, 907-914.

99. Lawton, K., Ward, E., Payne, G., Moyer, M., Ryals, J. (1992) Acidic and basic class III chitinase mRNA accumulation in response to TMV infection of tobacco. *Plant Mol. Biol*., **19**, 735-743.

100. Vogelsang, R., Barz, W. (1993) Cloning of a class III acidic chitinase from chickpea. *Plant Physiol*., **103**, 297-298.

101. Watanabe, T., Kobori, K., Miyashita, K., Fujii, T., Sakai, H., Uchida, M., Tanaka, H. (1993) Identification of glutamic acid 204 and aspartic acid 200 in chitinase A1 of Bacillus circulans WL-12 as essential residues for chitinase activity. *J. Biol. Chem*., **268**, 18567-18572.

102. Narumi, H., Hishida, T., Sasaki, T., Feng, D. F., Doolittle, R. F. (1993) Molecular cloning of silkworm (Bombyx mori) antichymotrypsin. A new member of the serpin superfamily of proteins from insects. *Eur. J. Biochem*., **214**, 181-187.

103. Shinozuka, Y., Kojima, S., Shomura, A., Ichimura, H., Yano, M., Yamamoto, K., Sasaki, T. (1999) Isolation and characterization of rice MADS box gene homologues and their RFLP mapping. *DNA Res*., **6**, 123-129.

104. Song, J., Yamamoto, K., Shomura, A., Itadani, H., Zhong, H. S., Yano, M., Sasaki, T. (1998) Isolation and mapping of a family of putative zinc-finger protein cDNAs from rice. *DNA Res*., **30**, 95-101.

105. Pillai, R. S. (2005) MicroRNA function: multiple mechanisms for a tiny RNA? *RNA*, **11**, 1753-1761.

106. Goyenvalle, A., Vulin, A., Fougerousse, F., Leturcq, F., Kaplan, J. C., Garcia, L., Danos, O. (2004) Rescue of dystrophic muscle through U7 snRNA-mediated exon

skipping. *Science*, **306**, 1796-1799.

107. Wilusz, J. E., Devanney, S. C., Caputi, M. (2005) Chimeric peptide nucleic acid compounds modulate splicing of the bcl-x gene in vitro and in vivo. *Nucleic Acids Res.*, **33**, 6547-6554.

108. Metraux, J. P., Burkhart, W., Moyer, M., Dincher, S., Middlesteadt, W., Williams, S., Payne, G.., Carnes, M., Ryals, J. (1989) Isolation of a complementary DNA encoding a chitinase with structural homology to a bifunctional lysozyme/chitinase. *Proc. Natl. Acad. Sci.U S A,* **86**, 896-900.

109. Jekel, P. A., Hartmann, B. H., Beintema, J. J. (1991) The primary structure of hevamine, an enzyme with lysozyme/chitinase activity from Hevea brasiliensis latex. *Eur. J. Biochem.*, **200**, 123-130.

110. Kurata, N., Nagamura, Y., Yamamoto, K., Harushima, Y., Sue, N., Wu, J., Antonio, B. A., Shomura, A., Shimizu, T., Lin, S.-Y., Inoue, T., Fukuda, A., Shimano, T., Kuboki, Y., Toyama, T., Miyamoto, Y., Kirihara, T., Hayasaka, K., Miyao, A., Monna, L., Zhong, H. S., Tamura, Y., Wang, Z.-X., Momma, T., Umehara, Y., Yano, M., Sasaki, T., Minobe, Y. (1994) A 300 kilobase interval genetic map of rice including 883 expressed sequences. *Nat. Genet.*, **8**, 365-372.

8 FIGURES AND TABLES

Figure 3.1. Outline of generation of ASTI transcripts. (1) An example of exon-intron structure in eukaryote. (2) Two mRNA precursors having different transcriptional initial sites. (3) Splicing of different exons. (4) Generation of one pair of splice variants.

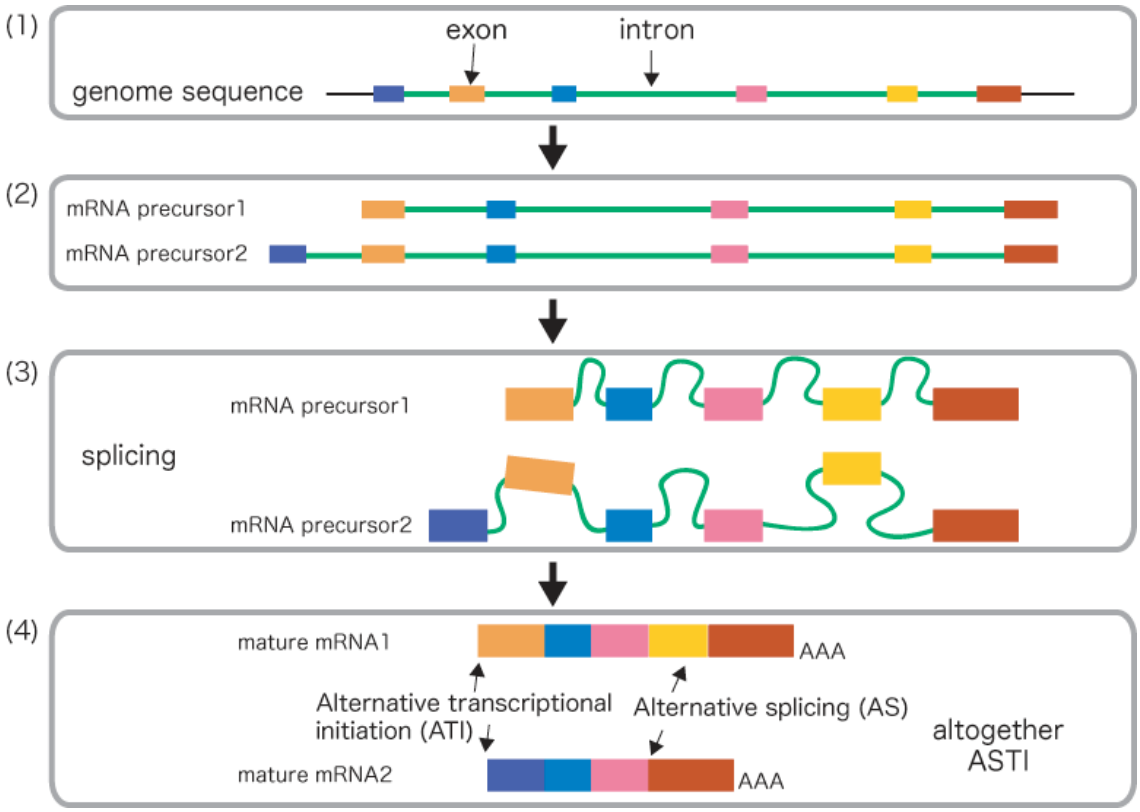
(1)
exon
intron
genome sequence
(2)
mRNA precursor1
mRNA precursor2
(3)
mRNA precursor1
splicing
mRNA precursor2
(4)
mature mRNA1
AAA
Alternative transcriptional initiation (ATI)
Alternative splicing (AS)
altogether ASTI
mature mRNA2
AAA

Figure 3.1

Figure 3.2. Representation of typical AS types proposed by Breitbart *et al*. [28]. Those were composed in the minimal scale of pair-wise exon-intron structures. A black or gray box indicates an alternative exon or exon part, whereas a white box indicates a constitutive exon or exon part. The upright and inverted V-shaped lines indicate the linkage of exons. (a) Cassette type (exon skipping); (b) retained intron, which is indicated by a thick line; (c) alternative donor; (d) alternative acceptor; two kinds of alternative polyadenylation (A) sites (e) and (f); and (g) mutually exclusive.

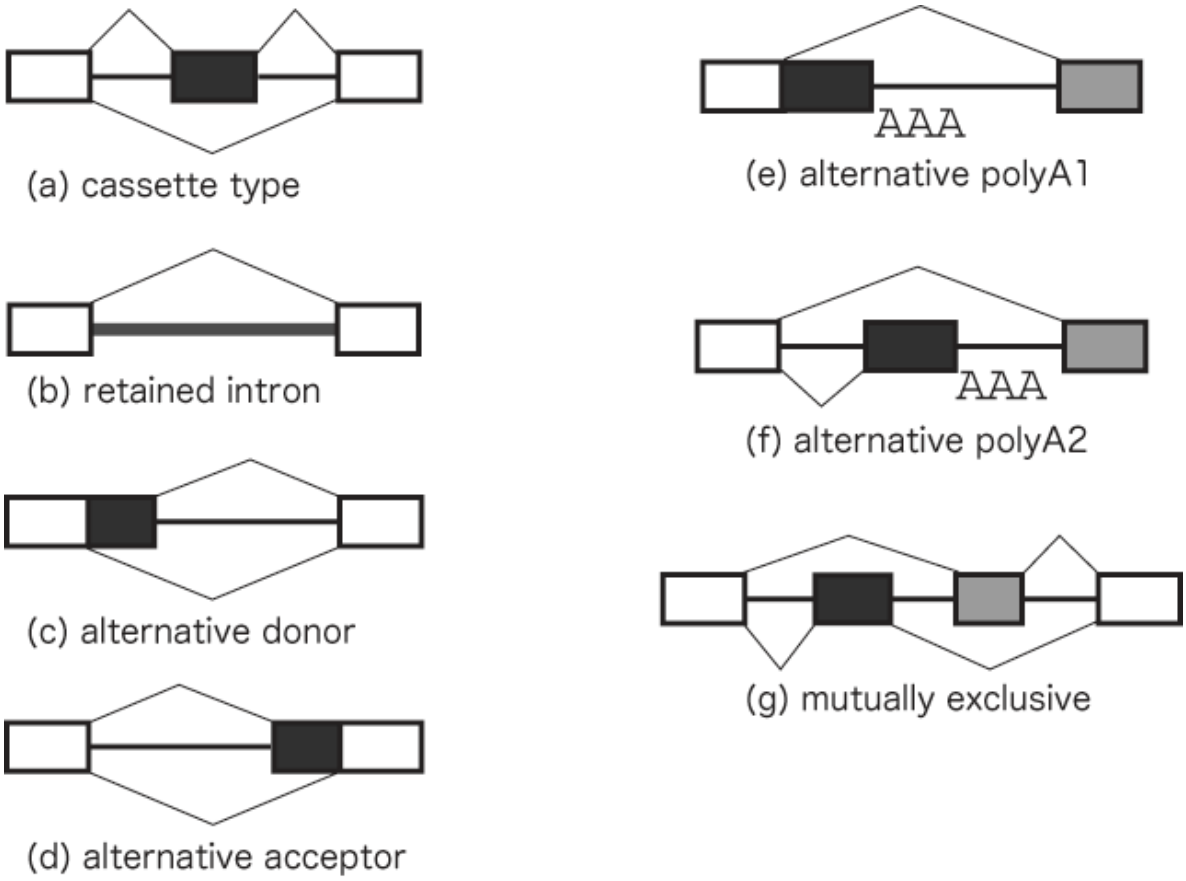
(a) cassette type
(b) retained intron
(c) alternative donor
(d) alternative acceptor
AAA
(e) alternative polyA1
AAA
(f) alternative polyA2
(g) mutually exclusive
Figure 3.2

Figure 3.3. Alternative splicing at NAGNAG acceptors (summarized from Hiller *et al*. [32]). (a) Proposed nomenclature for NAGNAG acceptors and transcripts. E, 3' half of the NAGNAG motif becomes part of the exon; I, the NAGNAG motif is completely retained in the intron. (b) Protein variability caused by alternative splicing at tandem acceptors according to intron phases 0, 1 and 2. I, single–amino acid indels; II, exchange of a single amino acid and an unrelated dipeptide; III, indel of a termination codon. Exonic nucleotides are shown in uppercase letters and intronic nucleotides, in lowercase letters. The RefSeq ID is given for annotated transcripts; NA, not annotated in RefSeq.

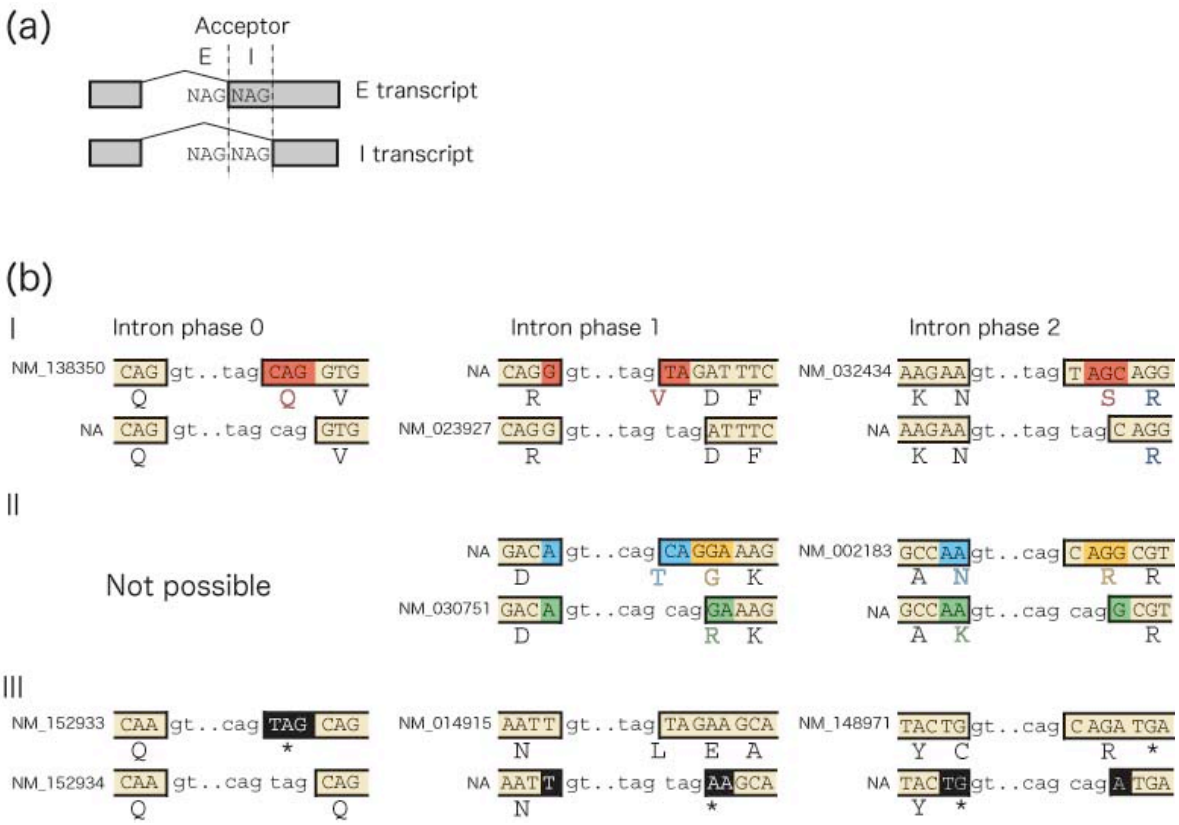
(a)
Acceptor
E
I
NAG
NAG
E transcript
NAG
NAG
I transcript
(b)
I
Intron phase 0
Intron phase 1
Intron phase 2
NM_138350 CAG gt..tag CAG GTG
Q Q V
NA CAG gt..tag cag GTG
Q V
NA CAGG gt..tag TA GAT TTC
R V D F
NM_023927 CAGG gt..tag tag ATTTC
R D F
NM_032434 AAGAA gt..tag T AGC AGG
K N S R
NA AAGAA gt..tag tag C AGG
K N R
II
Not possible
NA GACA gt..cag CA GGA AAG
D T G K
NM_030751 GACA gt..cag cag GA AAG
D R K
NM_002183 GCCAA gt..cag C AGG CGT
A N R R
NA GCCAA gt..cag cag G CGT
A K R
III
NM_152933 CAA gt..cag TAG CAG
Q * Q
NM_152934 CAA gt..cag tag CAG
Q Q
NM_014915 AATT gt..tag TA GAA GCA
N L E A
NA AATT gt..tag tag AA GCA
N *
NM_148971 TACTG gt..cag C AGA TGA
Y C R *
NA TACTG gt..cag cag A TGA
Y *
Figure 3.3

Figure 3.4. Summary of the generation of nonsense-mediated mRNA decay (NMD) [43, 44]. Some AS events (exon part insertion or removal, shown by shaded boxes of exon-intron structures in this figure) generate premature stop codons that cause NMD. (a) The case that the stop codon is on the last exon. During pre-mRNA processing, introns are removed and a set of proteins called the exon-junction complex is deposited. This complex serves to facilitate transport from the nucleus and to remember the gene structure. During the first pioneering round of translation, while the ribosome move from upstream onto mRNA, and will displace all exon-junction complexes in its path until it researches a stop codon. If the termination codon is on or near the final exon, as is the case for most genes, the ribosome will have displaced all exon-junction complexes. The mRNA will then undergo multiple rounds of translation. (b) If the termination codon is sufficiently far upstream of the final intron position, the exon-junction complexes will remain. Interactions ensue to result in degradation of the mRNA by NMD.

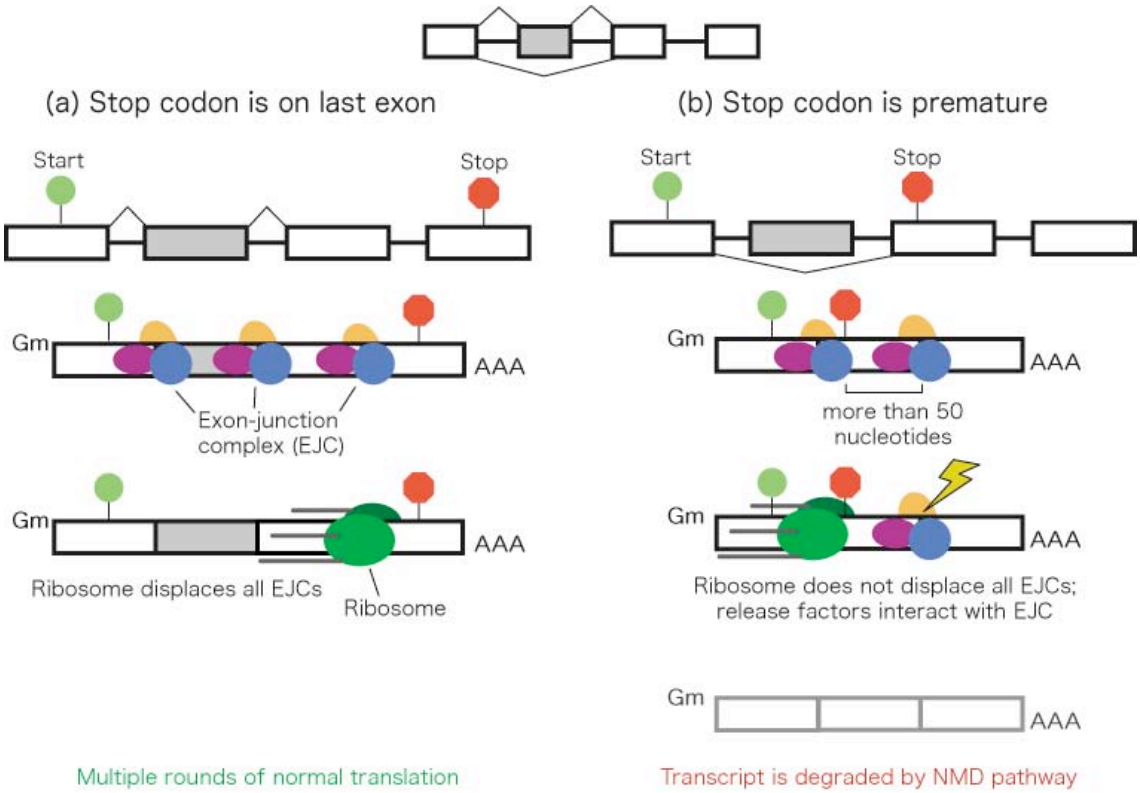
(a) Stop codon is on last exon
(b) Stop codon is premature
Start
Stop
Gm
AAA
Exon-junction complex (EJC)
Ribosome displaces all EJCs
Ribosome
more than 50 nucleotides
Ribosome does not displace all EJCs; release factors interact with EJC
Multiple rounds of normal translation
Transcript is degraded by NMD pathway

Figure 3.4

Figure 4.1. Outline of the algorithm for the classification of ASTI patterns. The algorithm first converts each exon–intron structure into a binary (0 or 1) array (I, I' and I"). Redundant rows with the same bit series within a delineated region are thinned out (II). Each pair of non-redundant bit series is compressed again, *e.g.*, two consecutive (0, 0) columns in II are combined into the underlined column (III). Finally, binary series are converted into decimal numbers to identify an ASTI type by a two-dimensional integer vector (IV). Solid arrows indicate the regions defined as AS units, whereas broken arrows indicate those not defined as AS units. In a more efficient procedure actually used, exon boundaries are processed from left to right in the order determined by a priority queue. Numbers in circles indicate the order of visits in this example. A flip-flop switch indicates the exonic status of each isoform.

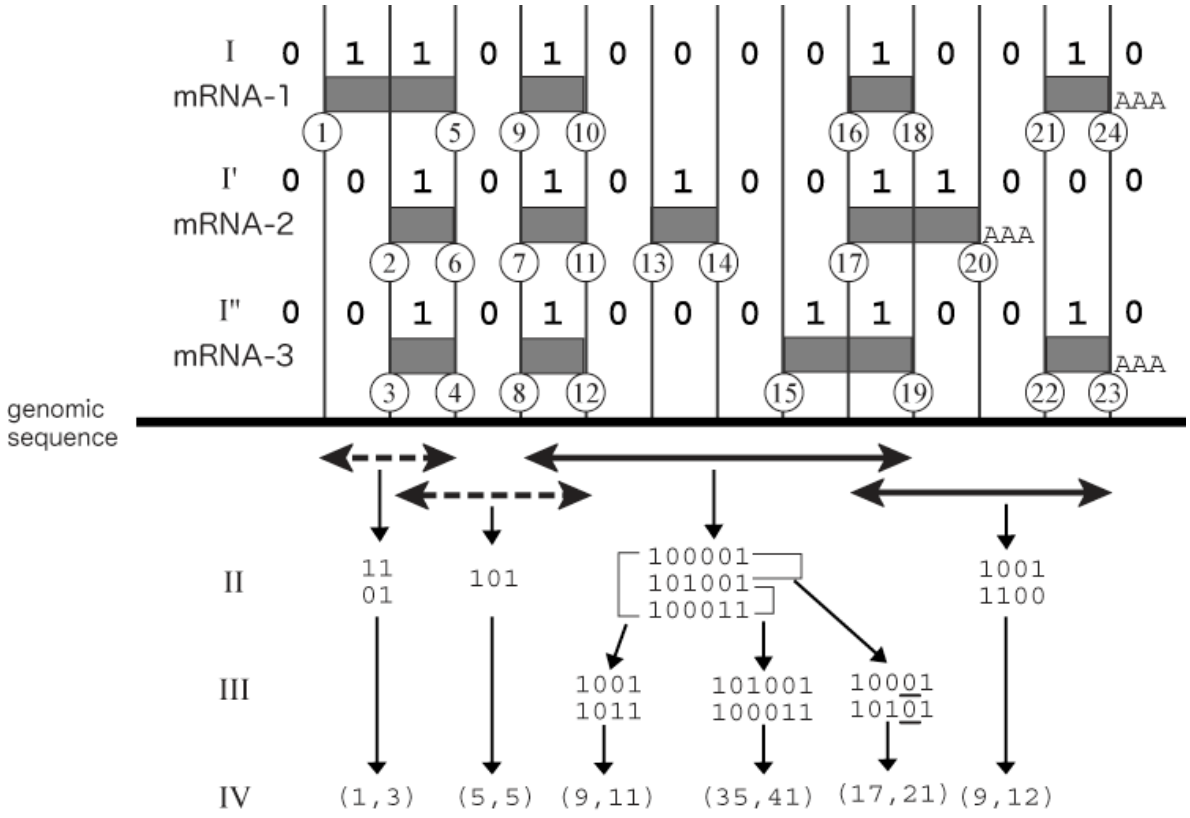
I 0 1 1 0 1 0 0 0 0 1 0 0 1 0
mRNA-1
AAA
1 5 9 10 16 18 21 24
I' 0 0 1 0 1 0 1 0 0 1 1 0 0 0
mRNA-2
AAA
2 6 7 11 13 14 17 20
I" 0 0 1 0 1 0 0 0 1 1 0 0 1 0
mRNA-3
AAA
3 4 8 12 15 19 22 23
genomic sequence
II
11 01
101
100001 101001 100011
1001 1100
III
1001 1011
101001 100011
10001 10101
IV
(1,3) (5,5) (9,11) (35,41) (17,21) (9,12)

Figure 4.1

Figure 4.2. Examples of ASTI types detected in human transcripts. (a) The seven representative AS types proposed by Breitbart *et al.* [28] displayed in the descending order of abundance in human transcripts. From left to right: the AS type, the binary representation, the decimal representation, the number of AS units detected and the relative abundance of the AS type. (b) The 10 most abundant AS types classified as "others." (c) The five most abundant ATI types. A black or gray box indicates an alternative exon or exon part, whereas a white box indicates a constitutive exon or exon part. A retained intron is indicated by a thick line.
Representation rules followed those in Figure 3.2.

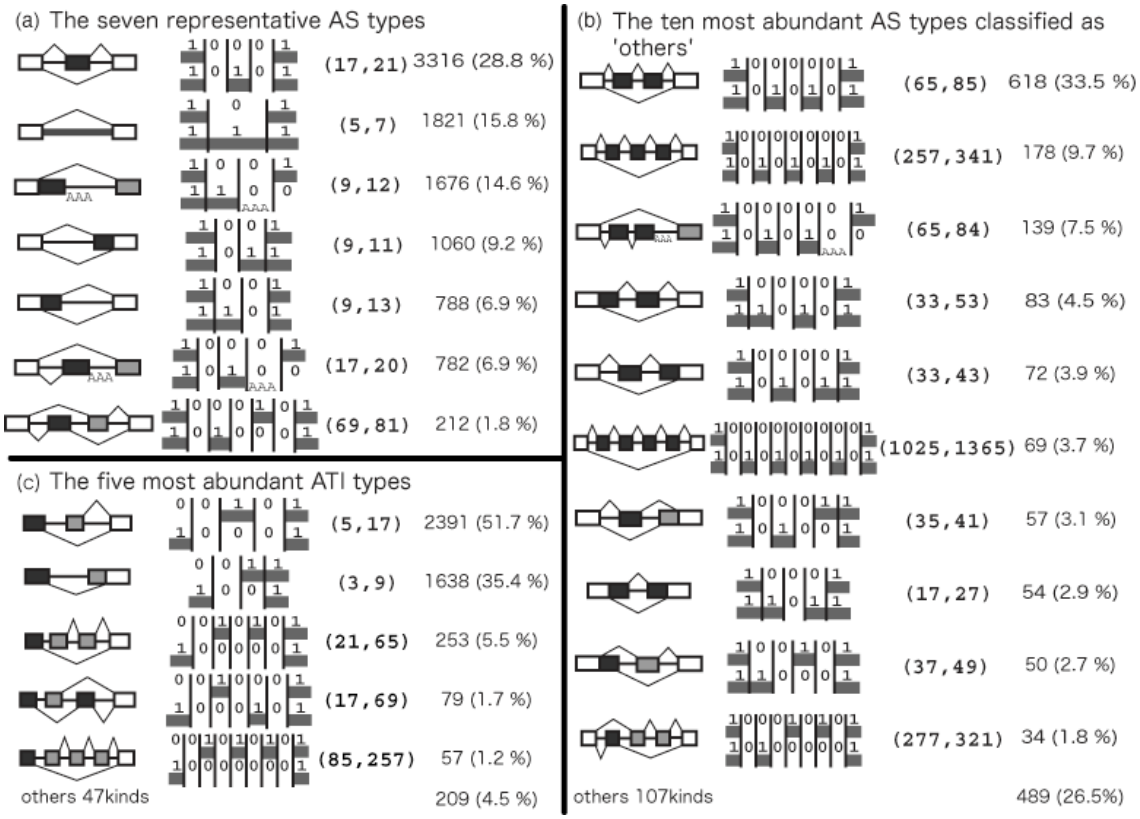
(a) The seven representative AS types
(17,21) 3316 (28.8 %)
(5,7) 1821 (15.8 %)
(9,12) 1676 (14.6 %)
(9,11) 1060 (9.2 %)
(9,13) 788 (6.9 %)
(17,20) 782 (6.9 %)
(69,81) 212 (1.8 %)
(b) The ten most abundant AS types classified as 'others'
(65,85) 618 (33.5 %)
(257,341) 178 (9.7 %)
(65,84) 139 (7.5 %)
(33,53) 83 (4.5 %)
(33,43) 72 (3.9 %)
(1025,1365) 69 (3.7 %)
(35,41) 57 (3.1 %)
(17,27) 54 (2.9 %)
(37,49) 50 (2.7 %)
(277,321) 34 (1.8 %)
others 107kinds 489 (26.5%)
(c) The five most abundant ATI types
(5,17) 2391 (51.7 %)
(3,9) 1638 (35.4 %)
(21,65) 253 (5.5 %)
(17,69) 79 (1.7 %)
(85,257) 57 (1.2 %)
others 47kinds 209 (4.5 %)

Figure 4.2

Figure 5.1. Variation in the fraction of ASTI genes as a function of the restricted number of randomly chosen cDNA clones. (a) Fraction of AS genes and (b) fraction of ATI genes. The indicated numbers of cDNA clones were randomly chosen, and the ASTI pairs and the ASTI loci were detected with the chosen clones alone. The fraction of ASTI loci shown is the average of ten trials. The standard deviations were smaller than the size of the symbols. Blue squares, orange circles, purple stars, yellow crosses, orange diamonds and green triangles indicate the results of human, mouse, fruit fly, nematode, cress and rice, respectively.

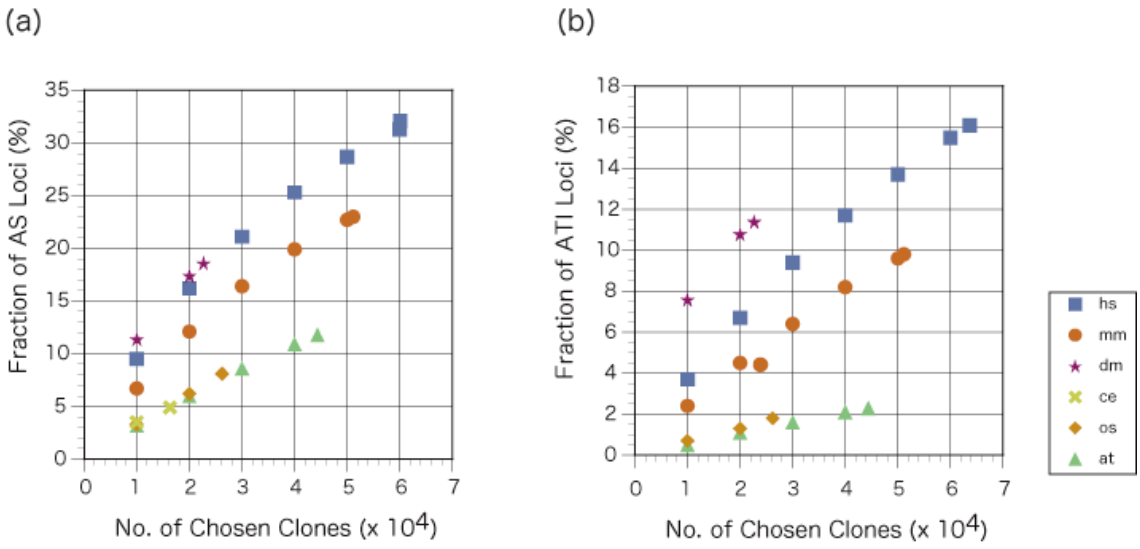
(a)
(b)
Fraction of AS Loci (%)
No. of Chosen Clones (x 10^4)
Fraction of ATI Loci (%)
No. of Chosen Clones (x 10^4)
hs
mm
dm
ce
os
at

Figure 5.1

Figure 5.2. Representations of the seven representative alternative splicing (AS) types in six eukaryotes. The classified AS units are categorized according to the general scheme by Breitbart *et al.* [28], and are arranged in the order of their abundance. Representative AS types are denoted as a) cassette exon, b) retained intron, c) alternative polyadenylation site 1 d) alternative acceptor, e) alternative polyadenylation site 2, f) alternative donor, and g) mutually exclusive exon. The number of categorized units in each type and its percentile share are shown under each illustrated pattern.

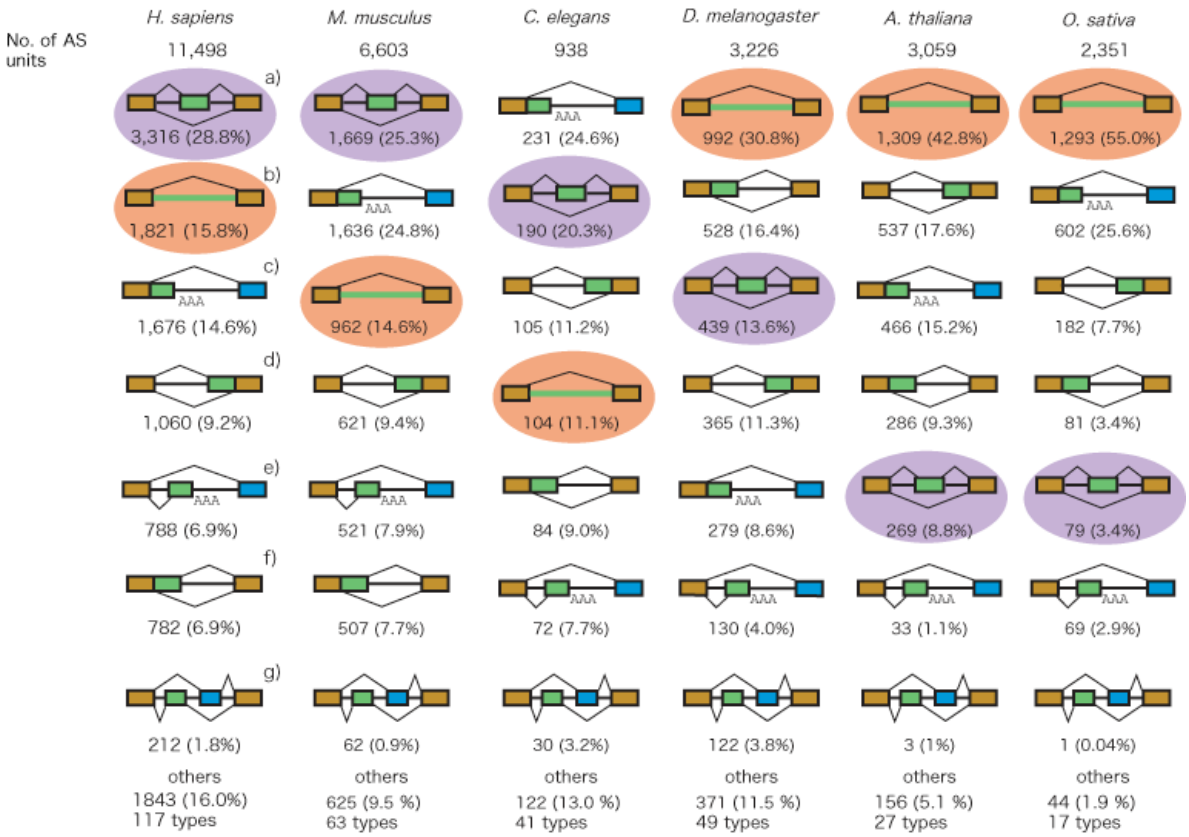
No. of AS units
H. sapiens
11,498
M. musculus
6,603
C. elegans
938
D. melanogaster
3,226
A. thaliana
3,059
O. sativa
2,351
a)
3,316 (28.8%)
1,669 (25.3%)
231 (24.6%)
992 (30.8%)
1,309 (42.8%)
1,293 (55.0%)
b)
1,821 (15.8%)
1,636 (24.8%)
190 (20.3%)
528 (16.4%)
537 (17.6%)
602 (25.6%)
c)
1,676 (14.6%)
962 (14.6%)
105 (11.2%)
439 (13.6%)
466 (15.2%)
182 (7.7%)
d)
1,060 (9.2%)
621 (9.4%)
104 (11.1%)
365 (11.3%)
286 (9.3%)
81 (3.4%)
e)
788 (6.9%)
521 (7.9%)
84 (9.0%)
279 (8.6%)
269 (8.8%)
79 (3.4%)
f)
782 (6.9%)
507 (7.7%)
72 (7.7%)
130 (4.0%)
33 (1.1%)
69 (2.9%)
g)
212 (1.8%)
62 (0.9%)
30 (3.2%)
122 (3.8%)
3 (1%)
1 (0.04%)
others
1843 (16.0%)
117 types
others
625 (9.5 %)
63 types
others
122 (13.0 %)
41 types
others
371 (11.5 %)
49 types
others
156 (5.1 %)
27 types
others
44 (1.9 %)
17 types

Figure 5.2

Figure 5.3. Classification of AS units by difference in length of alternative exons. Alternative splicing (AS) units are classified according to the difference in length (Δe) of alternative exons: 0/3, Δe mod 3 = 0; 1/3, Δe mod 3 = 1; and 2/3, Δe mod 3 = 2. The last column indicates the percentage of the 0/3-subtype AS units within each AS type. (a) The total number of observed AS units. (b) The number of AS units embedded within the CDS. (c) General internal exons were obtained from a well-curated, non-redundant human gene collection prepared by M. Mizuno *et al.* (personal communication). The classification depends on the exon length itself rather than the length difference, similar to the cases of cassette and retained intron AS types.

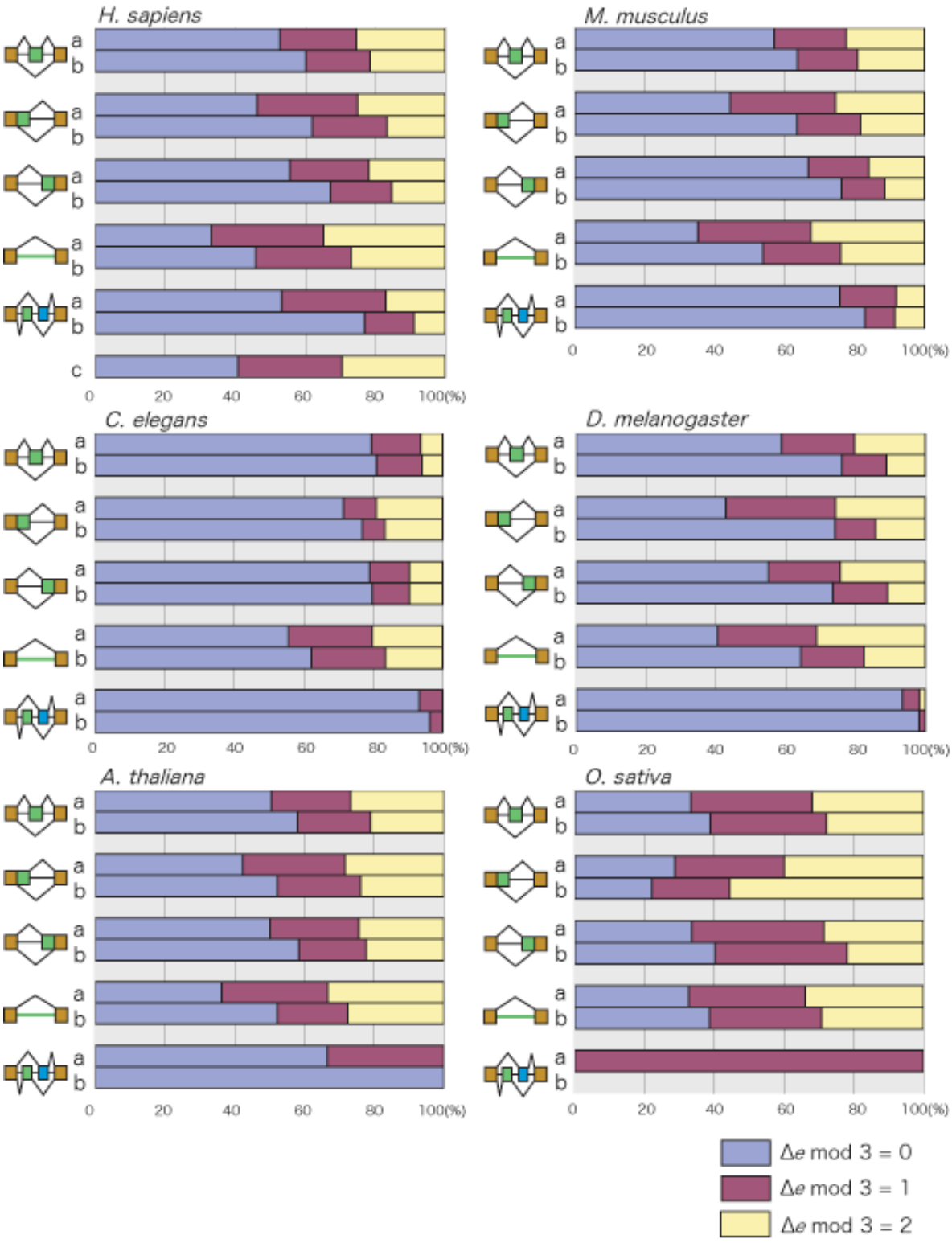
H. sapiens
M. musculus
C. elegans
D. melanogaster
A. thaliana
O. sativa
a
b
c
0
20
40
60
80
100(%)
Δe mod 3 = 0
Δe mod 3 = 1
Δe mod 3 = 2

Figure 5.3

Figure 5.4. The five most abundant alternative splicing (AS) types classified as "others." The types are arranged in the order of their abundance. The number of categorized units in each type and its percentile share are shown under each illustrated pattern.

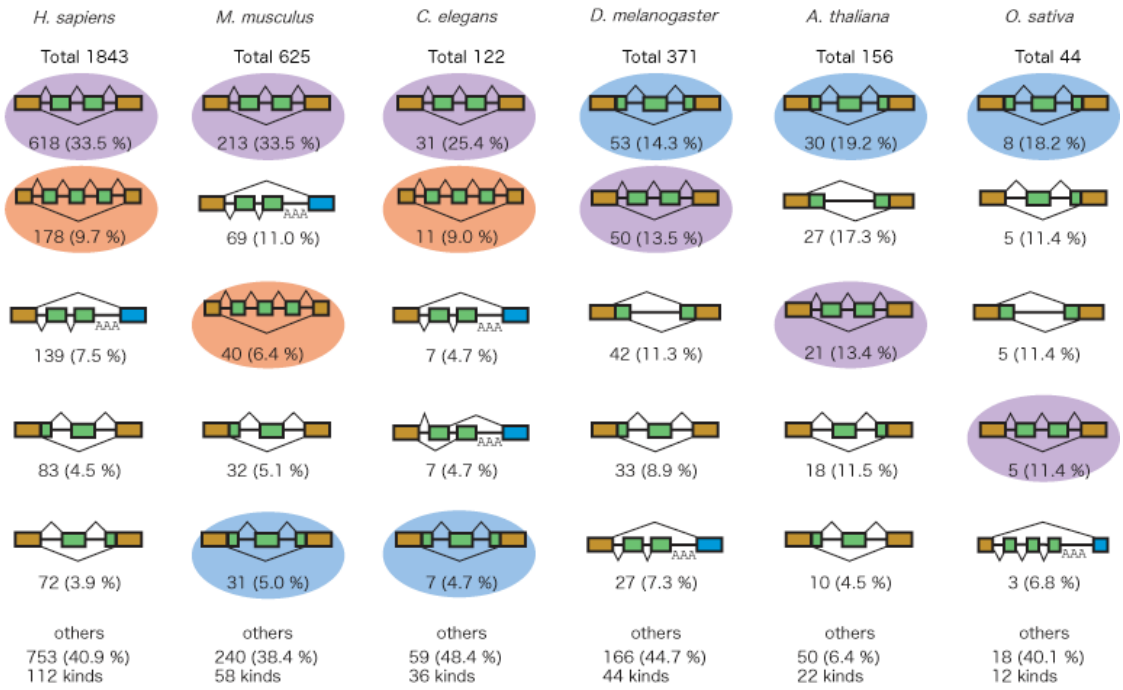
H. sapiens
Total 1843
618 (33.5 %)
178 (9.7 %)
139 (7.5 %)
83 (4.5 %)
72 (3.9 %)
others
753 (40.9 %)
112 kinds
M. musculus
Total 625
213 (33.5 %)
69 (11.0 %)
40 (6.4 %)
32 (5.1 %)
31 (5.0 %)
others
240 (38.4 %)
58 kinds
C. elegans
Total 122
31 (25.4 %)
11 (9.0 %)
7 (4.7 %)
7 (4.7 %)
7 (4.7 %)
others
59 (48.4 %)
36 kinds
D. melanogaster
Total 371
53 (14.3 %)
50 (13.5 %)
42 (11.3 %)
33 (8.9 %)
27 (7.3 %)
others
166 (44.7 %)
44 kinds
A. thaliana
Total 156
30 (19.2 %)
27 (17.3 %)
21 (13.4 %)
18 (11.5 %)
10 (4.5 %)
others
50 (6.4 %)
22 kinds
O. sativa
Total 44
8 (18.2 %)
5 (11.4 %)
5 (11.4 %)
5 (11.4 %)
3 (6.8 %)
others
18 (40.1 %)
12 kinds

Figure 5.4

Figure 5.5. The five most abundant alternative transcriptional initiation (ATI) types. The number of categorized units in each type and its percentile share are shown under each illustrated pattern.

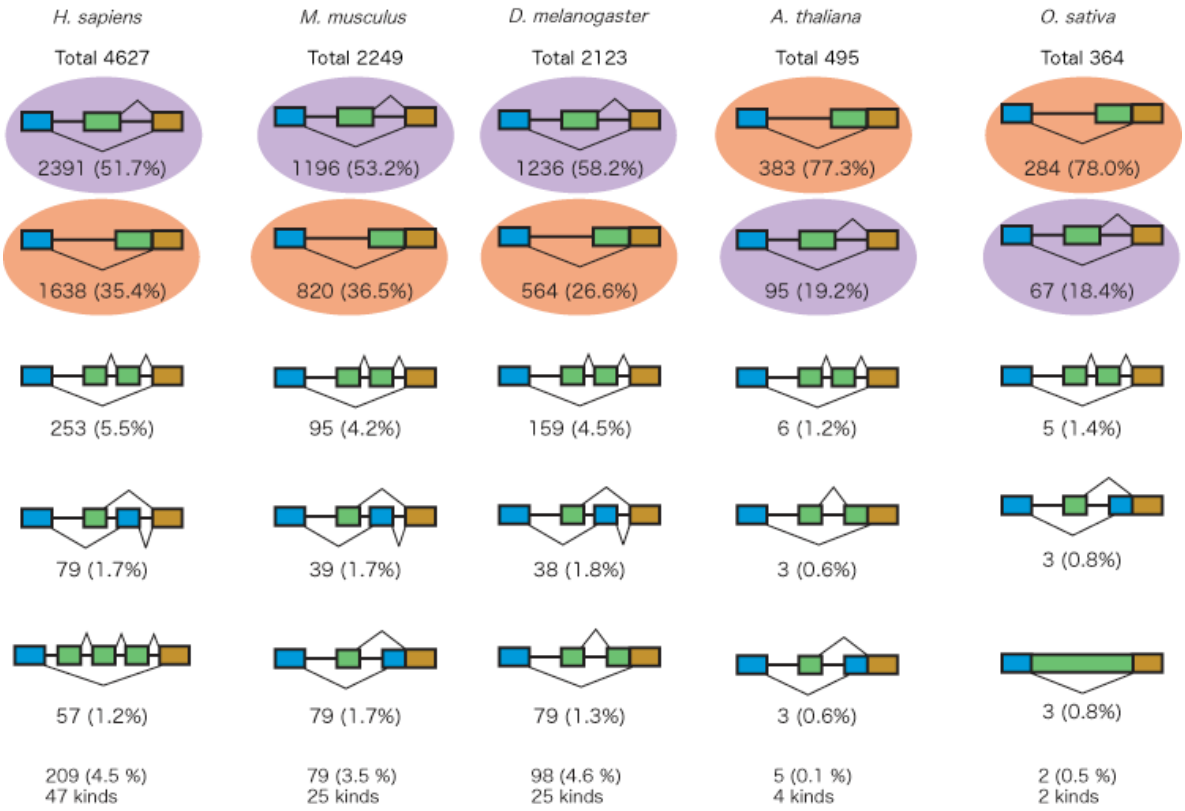
H. sapiens
Total 4627
2391 (51.7%)
1638 (35.4%)
253 (5.5%)
79 (1.7%)
57 (1.2%)
209 (4.5 %)
47 kinds
M. musculus
Total 2249
1196 (53.2%)
820 (36.5%)
95 (4.2%)
39 (1.7%)
79 (1.7%)
79 (3.5 %)
25 kinds
D. melanogaster
Total 2123
1236 (58.2%)
564 (26.6%)
159 (4.5%)
38 (1.8%)
79 (1.3%)
98 (4.6 %)
25 kinds
A. thaliana
Total 495
383 (77.3%)
95 (19.2%)
6 (1.2%)
3 (0.6%)
3 (0.6%)
5 (0.1 %)
4 kinds
O. sativa
Total 364
284 (78.0%)
67 (18.4%)
5 (1.4%)
3 (0.8%)
3 (0.8%)
2 (0.5 %)
2 kinds
Figure 5.5

Figure 5.6. Snapshots of the graphical interface of ASTRA. (1) "Gene Viewer" representing exon-intron structures defined by genome-cDNA mapping. (2) "Navigation Window" showing all variants of which those within the yellow-colored area are presented in the upper frame of the Gene Viewer. (3) "Control Panel" used to scroll and zoom in/out of Gene Viewer. (4) "Annotation Window" indicating annotation and sequence of the relevant cDNA. Links to GenBank and Ensembl databases are also indicated.

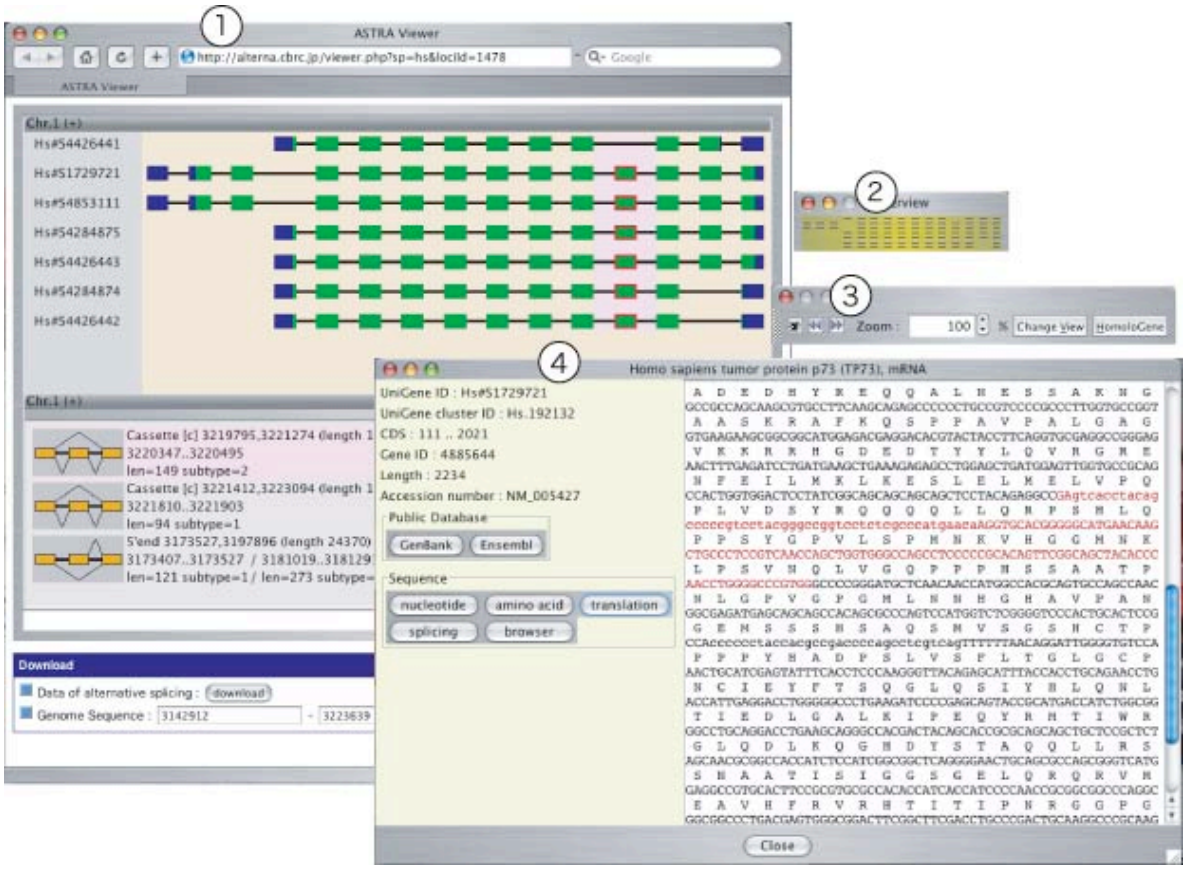
1
ASTRA Viewer
http://alterna.cbrc.jp/viewer.php?sp=hs&locId=1478
Chr.1 (+)
Hs#54426441
Hs#51729721
Hs#54853111
Hs#54284875
Hs#54426443
Hs#54284874
Hs#54426442
Cassette [c] 3219795,3221274 (length 1
3220347..3220495
len=149 subtype=2
Cassette [c] 3221412,3223094 (length 1
3221810..3221903
len=94 subtype=1
3'end 3173527,3197896 (length 24370
3173407..3173527 / 3181019..318129
len=121 subtype=1 / len=273 subtype=
Download
Data of alternative splicing :
download
Genome Sequence :
3142912
3223639
2
3
Zoom :
100
%
Change View
HomoloGene
4
Homo sapiens tumor protein p73 (TP73), mRNA
UniGene ID : Hs#51729721
UniGene cluster ID : Hs.192132
CDS : 111 .. 2021
Gene ID : 4885644
Length : 2234
Accession number : NM_005427
Public Database
GenBank
Ensembl
Sequence
nucleotide
amino acid
translation
splicing
browser
Close

Figure 5.6

Figure 6.1. Analysis of tandem repeat exons of human Monarch-1 gene. Each exon encodes a leucine-rich repeat with a conserved sequence motif. (a) Exon-intron structures of human Monarch-1 gene. Black bars, blue boxes, green boxes and green boxes with red frame indicate introns, 5' and 3' UTRs, CDSs and the tandem repeat exons, respectively. The three tandem repeats are labeled repeat1, 2 and 3, and are indicated by blue, green and orange arrows, respectively. (b) Dot matrix plots between tandem repeat exons of human Monarch-1 gene. (c) Multiple alignment of translated tandem repeat sequences. The conserved motif (LxxLxLxxN/CxL) is boxed.

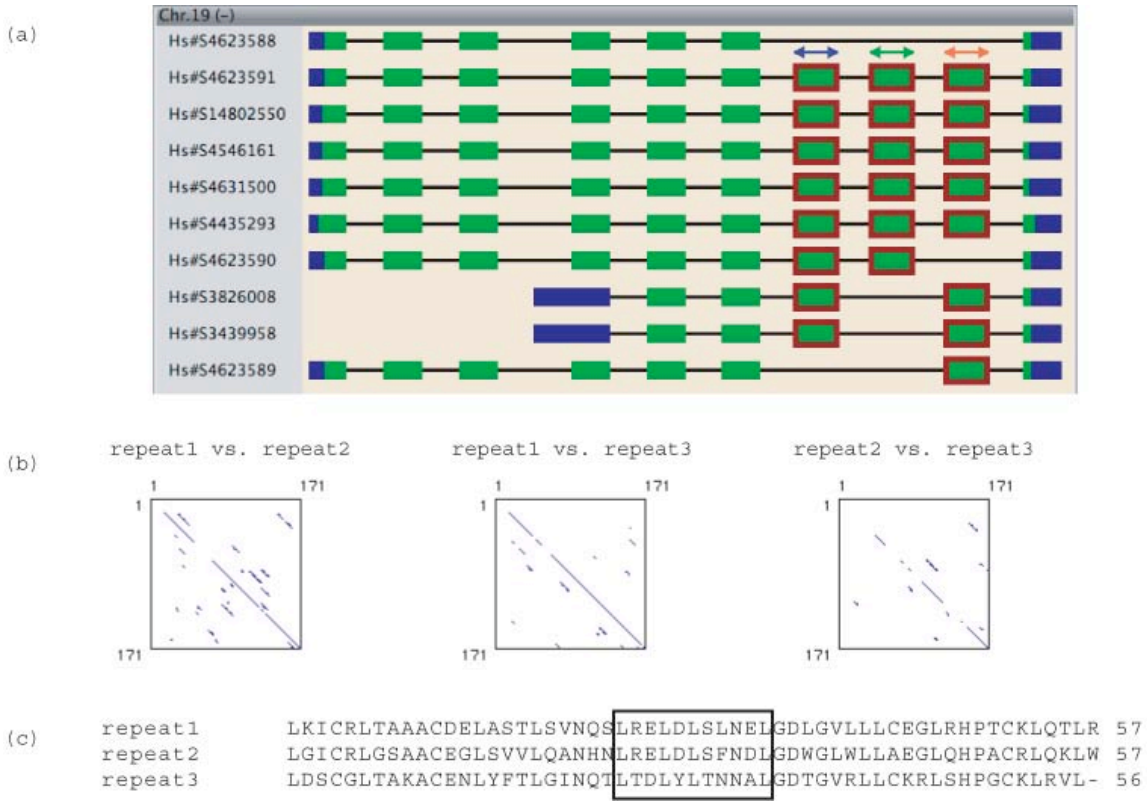
(a)
Chr.19 (-)
Hs#S4623588
Hs#S4623591
Hs#S14802550
Hs#S4546161
Hs#S4631500
Hs#S4435293
Hs#S4623590
Hs#S3826008
Hs#S3439958
Hs#S4623589
(b)
repeat1 vs. repeat2
repeat1 vs. repeat3
repeat2 vs. repeat3
1
171
(c)
repeat1 LKICRLTAAACDELASTLSVNQSLRELDLSLNELGDLGVLLLCEGLRHPTCKLQTLR 57
repeat2 LGICRLGSAACEGLSVVLQANHNLRELDLSFNDLGDWGLWLLAEGLQHPACRLQKLW 57
repeat3 LDSCGLTAKACENLYFTLGINQTLTDLYLTNNALGDTGVRLLCKRLSHPGCKLRVL- 56

Figure 6.1

Figure 6.2. An atypical AS unit of type (100011, 101001) in human interleukin 28 receptor alpha (IL-28RA) gene. There is another isoform that lacks both internal exonic regions of these variants, and hence typical AS units, which are classified as cassette and alternative acceptor types represented by (10001, 10101) and (1001, 1011), are also present in this region.

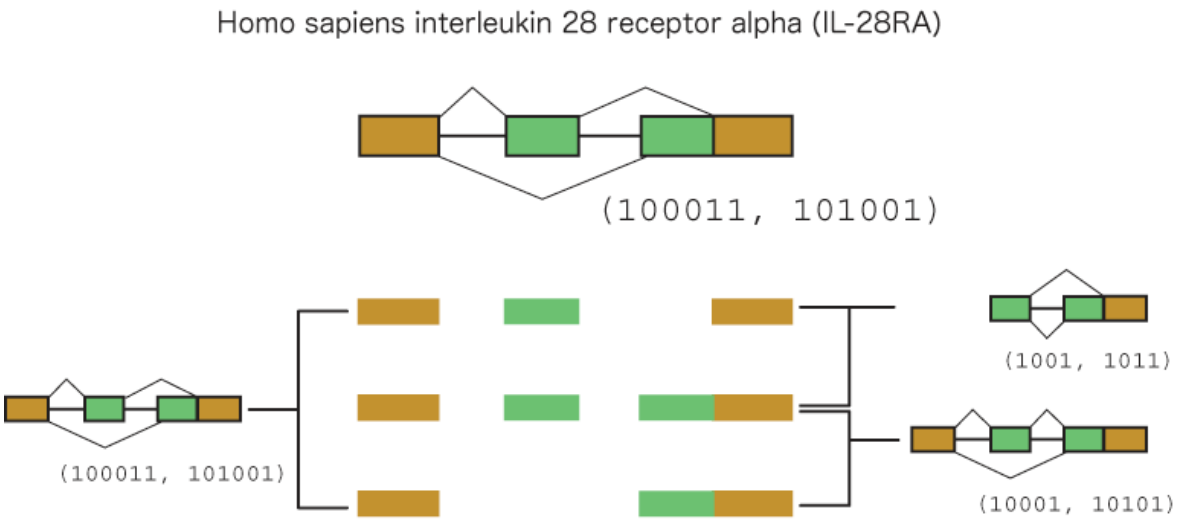
Homo sapiens interleukin 28 receptor alpha (IL-28RA)
(100011, 101001)
(100011, 101001)
(1001, 1011)
(10001, 10101)
Figure 6.2

Figure 6.3. The structure of rice class III chitinase cDNAs. Open part of the bars indicates the open reading frames for respective cDNAs, and shaded part indicates the putative signal sequence estimated by sequence comparison with those of class III chitinase of cucumber [108] and hevamine of para rubber tree [109]. All the cDNAs contain 5' and 3' UTRs as indicated. Accession numbers granted by the DNA Data Bank of Japan (DDBJ) for rice class III chitinase cDNAs are as follows: D55708 for C00481; D55709 for C10122; D55710 for C10150; D55711 for C10501; D55712 for C10701; D55713 for C10728; and D55714 for C10923.

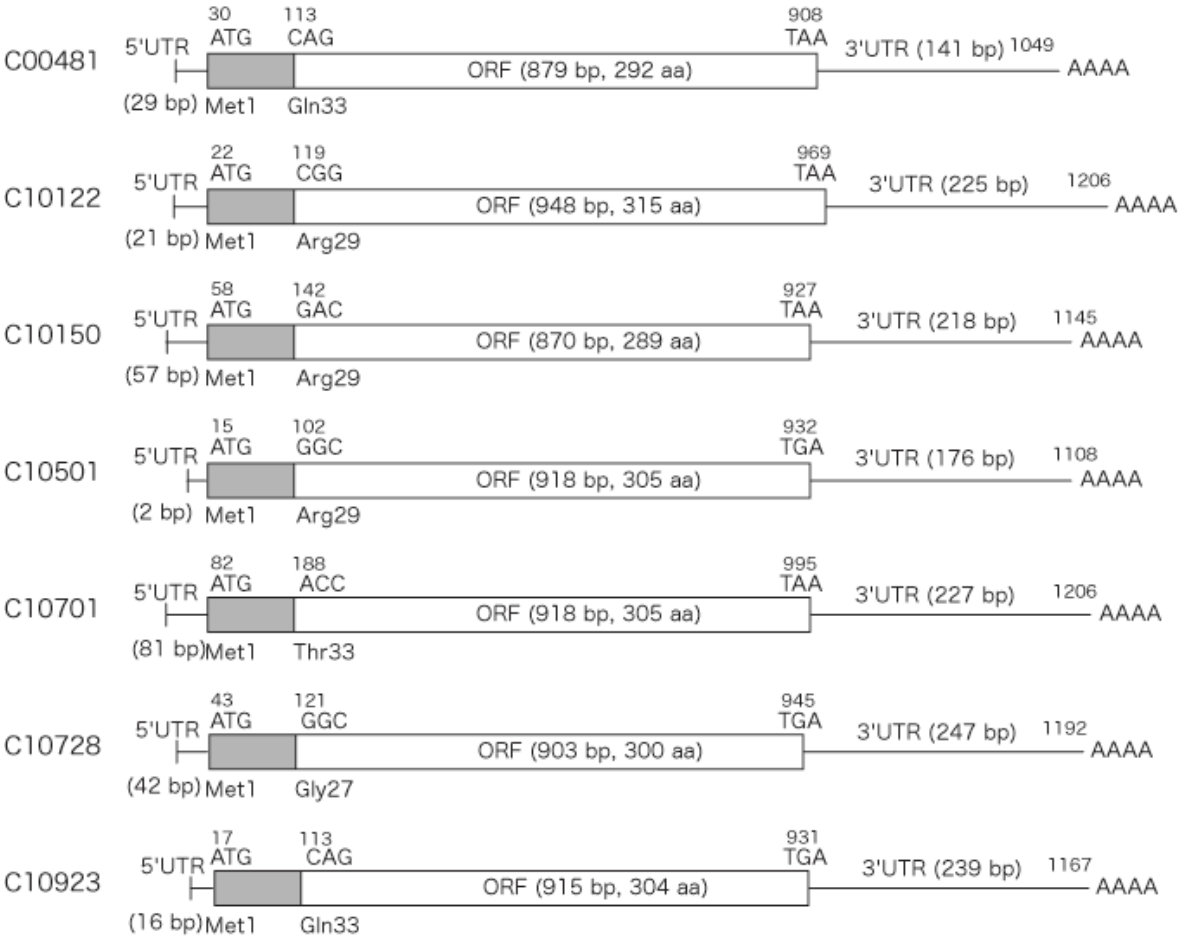
C00481
30
ATG
113
CAG
908
TAA
5'UTR
3'UTR (141 bp)
1049
AAAA
ORF (879 bp, 292 aa)
(29 bp)
Met1
Gln33
C10122
22
ATG
119
CGG
969
TAA
5'UTR
3'UTR (225 bp)
1206
AAAA
ORF (948 bp, 315 aa)
(21 bp)
Met1
Arg29
C10150
58
ATG
142
GAC
927
TAA
5'UTR
3'UTR (218 bp)
1145
AAAA
ORF (870 bp, 289 aa)
(57 bp)
Met1
Arg29
C10501
15
ATG
102
GGC
932
TGA
5'UTR
3'UTR (176 bp)
1108
AAAA
ORF (918 bp, 305 aa)
(2 bp)
Met1
Arg29
C10701
82
ATG
188
ACC
995
TAA
5'UTR
3'UTR (227 bp)
1206
AAAA
ORF (918 bp, 305 aa)
(81 bp)
Met1
Thr33
C10728
43
ATG
121
GGC
945
TGA
5'UTR
3'UTR (247 bp)
1192
AAAA
ORF (903 bp, 300 aa)
(42 bp)
Met1
Gly27
C10923
17
ATG
113
CAG
931
TGA
5'UTR
3'UTR (239 bp)
1167
AAAA
ORF (915 bp, 304 aa)
(16 bp)
Met1
Gln33
Figure 6.3

Figure 6.4. Comparison of amino acid sequences deduced from rice class III chitinase cDNAs and those of other plants. The deduced amino acid sequences of obtained cDNAs are compared with those of class III chitinases of Arabidopsis [98], chick pea [100], cucumber [108], tobacco [99], and hevamine of para rubber tree [109]. The starting positions of mature proteins of rice class III chitinase were estimated by comparison with those of cucumber [108] and para rubber tree [109]. Asterisks indicate amino acid residues identical to those of C10501. The N-terminus of mature proteins starts from the 40th amino acid position. E (Glu) or D (Asp) marked by a black dot at the 170th position is the supposed active residue of class III chitinase [101].

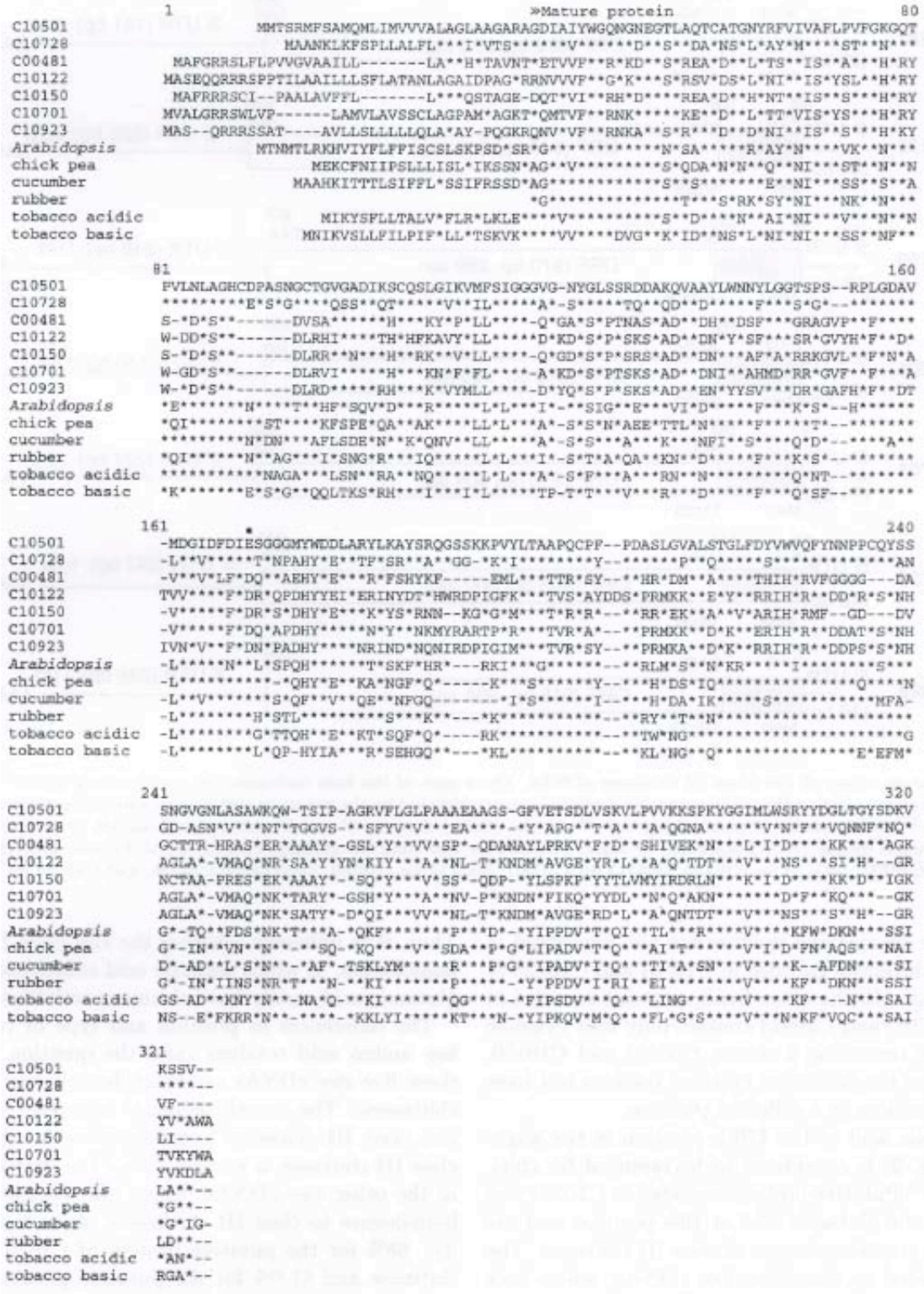
1
»Mature protein
80
C10501
C10728
C00481
C10122
C10150
C10701
C10923
Arabidopsis
chick pea
cucumber
rubber
tobacco acidic
tobacco basic
81
160
161
240
241
320
321

Figure 6.4

Figure 6.5. The loci of class III chitinases on our linkage map of rice. Linkage analysis was performed as described in Kurata *et al*. [110]. The loci of obtained rice class III chitinase cDNAs are indicated by white letters with black background. Designations of markers are described in Kurata *et al*. [110].

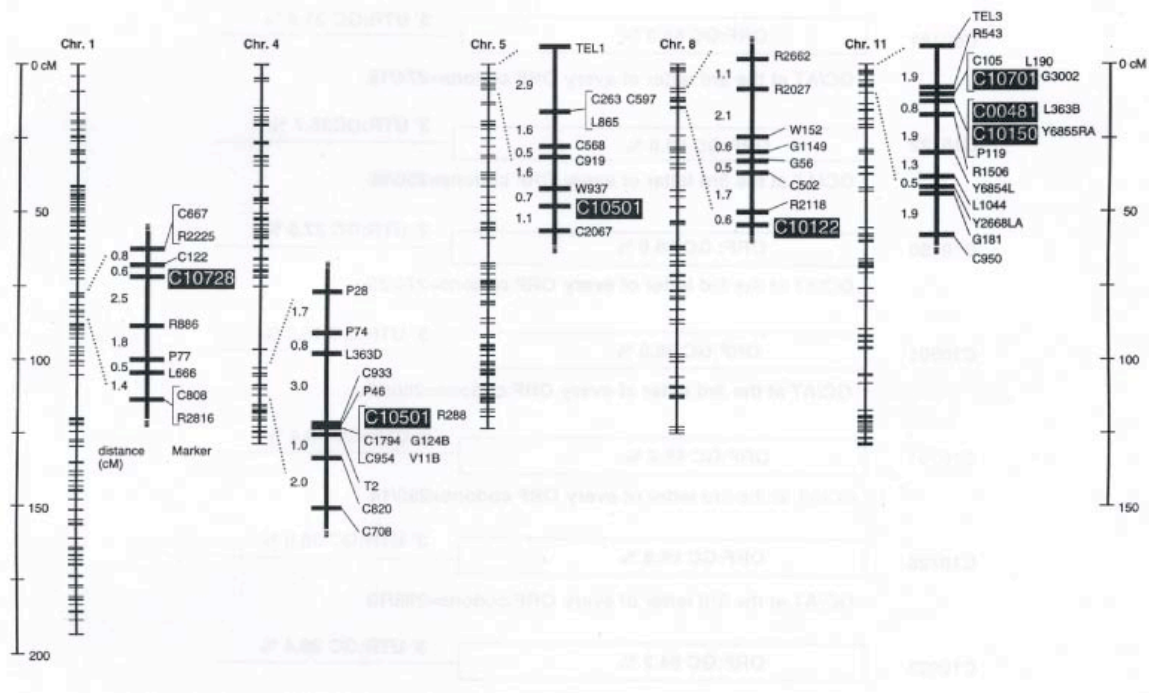
Chr. 1
Chr. 4
Chr. 5
Chr. 8
Chr. 11
0 cM
50
100
150
200
C667
R2225
C122
C10728
R886
P77
L666
C808
R2816
distance (cM)
Marker
P28
P74
L363D
C933
P46
C10501
R288
C1794
G124B
C954
V11B
T2
C820
C708
TEL1
C263
C597
L865
C568
C919
W937
C10501
C2067
R2662
R2027
W152
G1149
G56
C502
R2118
C10122
TEL3
R543
C105
L190
C10701
G3002
C00481
L363B
C10150
Y6855RA
P119
R1506
Y6854L
L1044
Y2668LA
G181
C950

Figure 6.5

Table 4.1. Materials used to detect alternative splicing (AS) and alternative transcriptional initiation (ATI) patterns and the mapping results. Shown are the statistics of cDNA sequences, AS and ATI patterns for each species. See Materials and Methods for data sources and detection methods.

Species	Genome sequence	cDNA data set	Number of cDNA sequences	Number of mapped cDNA sequences	Number of mapped loci	Number of loci that generate AS variants	Number of distinct AS variants	Number of detected AS units
H. sapiens	build 33	UniGene build#172	109,452	65,041	15,371	4,931	12,470	11,498
M. musculus	build 32	UniGene build#139	89,002	56,310	15,878	3,656	8,456	6,603
D. melanogaster	build 3	UniGene build#34	31,583	22,890	8,695	1,619	3,920	3,226
C. elegans	WS133	UniGene build#18	18,100	16,780	13,358	649	1,407	938
A. thaliana	Updated in 2000	UniGene build#46	58,376	44,618	15,516	2,195	4,642	3,059
O. sativa	pseudomolecule version_3.0	32K cloneset	32,127	27,175	18,796	1,514	3,158	2,351

Species	Number of loci that generate ATI variants	Number of distinct ATI variants	Number of detected ATI units
H. sapiens	2,474	5,521	4,627
M. musculus	1,611	3,393	2,249
D. melanogaster	992	2,276	2,123
A. thaliana	435	881	495
O. sativa	331	670	364

Table 4.1

Table 5.1. Location of AS units with respect to coding regions for genes of the six species studied. Alternative splicing (AS) units are classified according to the type and location within or outside the CDSs. The upper value in each cell indicates the observed number of units, and the lower value shows its percentage in the total number of AS units of the same type. An AS unit is labeled "Within CDS" when both exon variants in the mutually exclusive type or additional exon(s) in the other types are completely included in an ORF(s), irrespective of preservation of the downstream reading frames. Similarly, "5' UTR" or "3' UTR" indicates that both exon variants and additional exon(s) are completely included in a respective UTR. If the transcription of an exon starts or ends within the CDS, the units are labeled "5'-CDS" or "CDS-3'," respectively. In this table, the percentages do not total 100%, because mutually exclusive AS units whose exons interleave the CDSs are not included.

H. sapiens

Type / Location	Within CDS	5' UTR	5'-CDS	3' UTR	CDS-3'	Total
Cassette	2356 71.0%	387 11.7%	189 5.7%	101 3.0%	283 8.5%	3316 100.0%
Alt donor site	384 48.7%	249 31.6%	59 7.5%	32 4.1%	64 8.1%	788 100.0%
Alt acceptor site	699 65.9%	137 12.9%	24 2.3%	62 5.8%	138 13%	1060 100.0%
Retained intron	253 13.9%	325 17.8%	168 9.2%	527 28.9%	548 30.1%	1821 100.0%
Mutually exclusive	113 53.3%	16 7.5%	1 0.5%	1 0.5%	7 3.3%	138 65.1%

M. musculus

Type / Location	Within CDS	5' UTR	5'-CDS	3' UTR	CDS-3'	Total
Cassette	1259 75.4%	224 13.4%	73 4.4%	20 1.2%	93 5.6%	1669 100.0%
Alt donor site	214 42.2%	212 41.8%	31 6.1%	10 2.0%	40 7.9%	507 100.0%
Alt acceptor site	431 69.4%	84 13.5%	18 2.9%	23 3.7%	65 10.5%	621 100.0%
Retained intron	117 12.2%	190 19.8%	55 5.7%	220 22.9%	380 39.5%	962 100.0%
Mutually exclusive	47 75.8%	3 4.8%	1 1.6%	0 0.0%	1 1.6%	52 83.9%

C. elegans

Type / Location	Within CDS	5' UTR	5'-CDS	3' UTR	CDS-3'	Total
Cassette	184 96.8%	0 0.0%	0 0.0%	1 0.5%	5 2.6%	190 100.0%
Alt donor site	78 92.9%	1 1.2%	1 1.2%	2 2.4%	2 2.4%	84 100.0%
Alt acceptor site	104 99.0%	1 1.0%	0 0.0%	0 0.0%	0 0.0%	105 100.0%
Retained intron	85 81.7%	1 1.0%	2 1.9%	3 2.9%	13 12.5%	104 100.0%
Mutually exclusive	27 90.0%	0 0.0%	0 0.0%	0 0.0%	2 6.7%	29 96.7%

D. melanogaster

Type / Location	Within CDS	5' UTR	5'-CDS	3' UTR	CDS-3'	Total
Cassette	251 57.2%	94 21.4%	59 13.4%	7 1.6%	28 6.4%	439 100.0%
Alt donor site	112 21.2%	367 69.5%	39 7.4%	2 0.4%	8 1.5%	528 100.0%
Alt acceptor site	185 50.7%	138 37.8%	16 4.4%	7 1.9%	19 5.2%	365 100.0%
Retained intron	188 19.0%	510 51.4%	68 6.9%	78 7.9%	148 14.9%	992 100.0%
Mutually exclusive	112 91.8%	0 0.0%	4 3.3%	0 0.0%	0 0.0%	116 95.1%

A. thaliana

Type / Location	Within CDS	5' UTR	5'-CDS	3' UTR	CDS-3'	Total
Cassette	205 76.2%	25 9.3%	10 3.7%	8 3.0%	21 7.8%	269 100.0%
Alt donor site	159 55.6%	81 28.3%	11 3.8%	13 4.5%	22 7.7%	286 100.0%
Alt acceptor site	354 65.9%	120 22.3%	17 3.2%	23 4.3%	23 4.3%	537 100.0%
Retained intron	276 21.1%	375 28.6%	80 6.1%	178 13.6%	400 30.6%	1309 100.0%
Mutually exclusive	2 66.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 66.7%

O. sativa

Type / Location	Within CDS	5' UTR	5'-CDS	3' UTR	CDS-3'	Total
Cassette	36 52.2%	11 15.9%	6 8.7%	9 13.0%	7 10.1%	69 100.0%
Alt donor site	27 33.8%	31 38.8%	4 5.0%	11 13.7%	7 8.7%	80 100.0%
Alt acceptor site	82 45.1%	55 30.2%	6 3.3%	26 14.3%	13 7.1%	182 100.0%
Retained intron	106 8.2%	392 30.3%	142 11.0%	341 26.4%	312 24.1%	1293 100.0%
Mutually exclusive	0 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	1 100.0%

Table 5.1

9 ACKNOWLEDGEMENTS

First, I would like to express my special thanks to Dr. M. Arita of Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, and Mr. T. Nishizawa of Information and Mathematical Science Laboratory Inc., who have patiently contributed to my investigation. I also thank Drs. K. Fukui, T. Aita, T. Kin, K. Asai, and M. Suwa of Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Dr. M. Nakao of Kazusa DNA Institute, and Dr. T. Yada of Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University for helpful discussions, Mr. M. Mizuno for providing AQUA database, Mr. T. Kumagai for his help in setting up ASTRA hardware, Drs. J. Yazaki, K. Satoh and S. Kikuchi of the Laboratory of Gene Expression, Department of Molecular Genetics, National Institute of Agrobiological Sciences, for their suggestion to apply the data derived from *O. sativa* to our investigation, and my family for unfailing support.

I am deeply indebted to Dr. K. Kawashima of National Cancer Center Research Institute, who inspired me to become a researcher, and Drs. T. Sasaki of the National Institute of Agrobiological Sciences and O. Gotoh of Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, who left great accomplishments as the researcher, and exist like great walls hard for me to get over or break through.

This work was partially supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) “Genome Information Science” from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and by the Institute for Bioinformatics Research and Development (BIRD) of Japan Science and Technology Agency (JST).