Alternative Splicing in Higher Plants

Just Adding to Proteomic Diversity or an Additional Layer of Regulation?

Alternative Splicing

- Alternative splicing is nearly ubiquitous in eukaryotes

- It has been found in plants, flies, worms, mammals, etc.
  - Note: *Saccharomyces cerevisiae* is an intron poor species (95% of genes lack introns)
  - Alternative splicing is rare in this species

- Alternative splicing has recently expanded as an area of research with the advent of high throughput sequencing efforts

- 2 divergent roles
  - Proteomic diversity
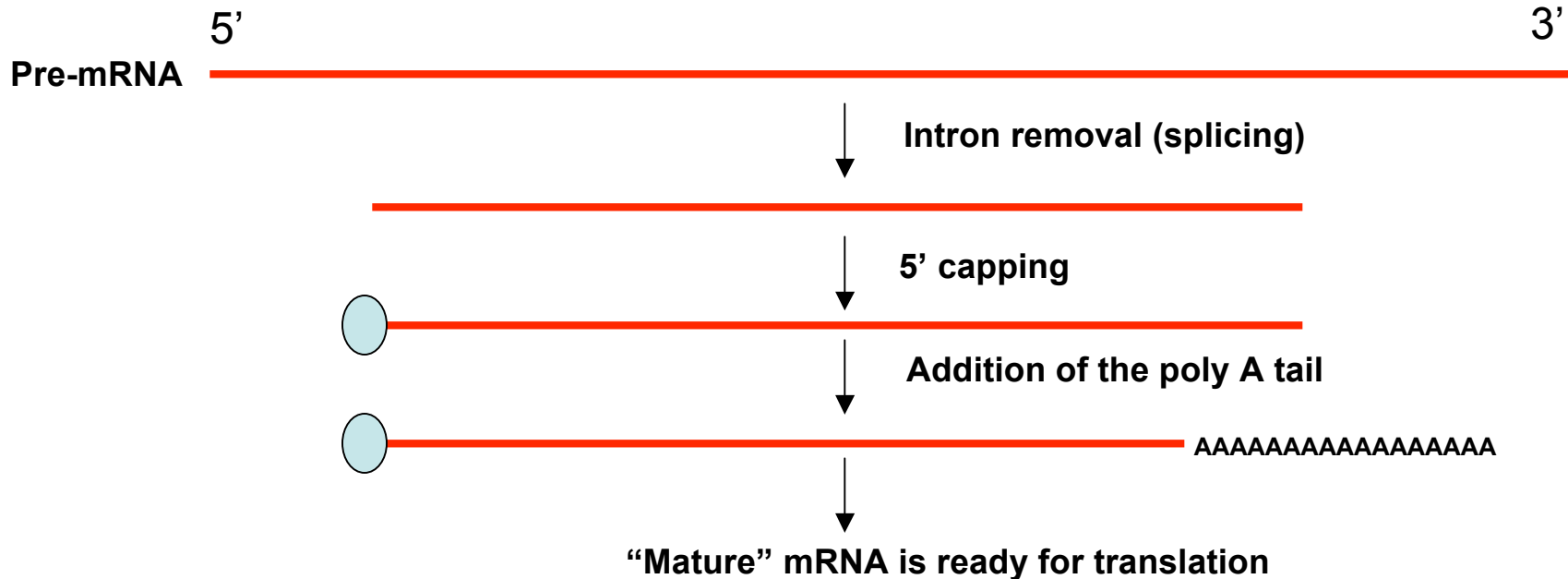 - Post-transcriptional regulation

RNA Polymerase II mediated transcription of a gene

- Alternative splicing is nearly ubiquitous in eukaryotes
- Alternative splicing is defined as different isoforms of mRNA that can originate from a common locus
- mRNA is generated from RNA polymerase II transcription of a gene
- A schema of eukaryotic transcription is given below
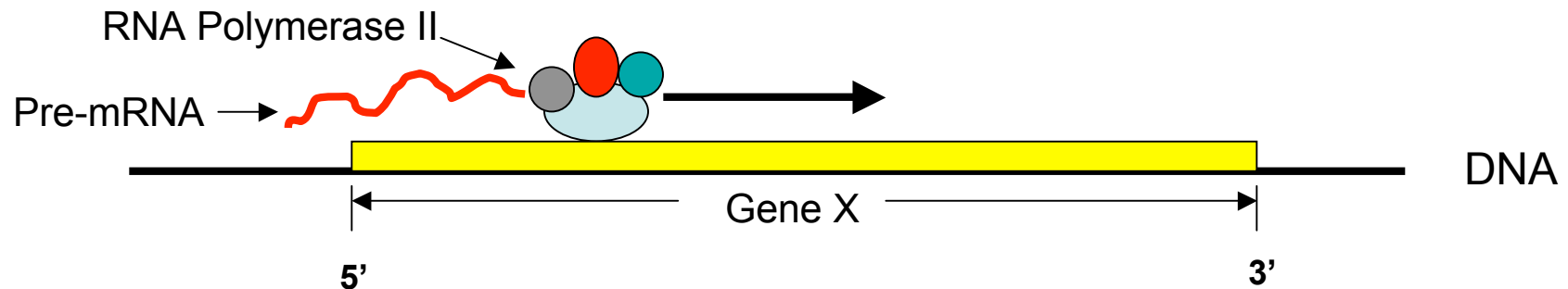
Pre-mRNA processing in a mature mRNA

- Pre-mRNA generated by RNA polymerase II must "processed" to be used for translation

- These steps include:
  - Removal of introns and ligation of the exons
  - Capping of the 5' end of the mRNA
  - Poly A tail

5'   3'

Pre-mRNA

Intron removal (splicing)

5' capping

Addition of the poly A tail

AAAAAAAAAAAAAAAA

"Mature" mRNA is ready for translation

Alternative Splicing

- Altenative splicing is commonly defined as alternatively processed mRNA structure(s) relative to a predominant isoform originating from a common template (locus)

- The predominant isoform is the one most commonly observed

- Note "Predominance" is dependent upon the context (spatial, temporal) of expression

- Alternative splicing is commonly detected by:
  - Aligning ESTs to the genome (or each other) and looking for structural diversity (genome wide)
  - RT-PCR/RACE of a particular locus and looking for alternate isoforms (locus specific)
  - Microarrays that cover the locus can be used to detect alternate isoforms (not common in plants)

Some points to consider

There are two commonly recognized hypotheses about why alternative splicing is employed in eukaryotes
- – Proteomic diversity
- – Post-transcriptional regulation

- Alternative splicing was first postulated in 1978 by Walter Gilbert as a way to create novel isoforms (mRNAs) from a common template (the DNA template for the gene)

- In order to detect alternative splicing, one needs the sequence of both the processed or "mature" mRNA and the original genomic sequence

- Alignment of the mRNA to the genomic sequence is performed *in silico* using alignment programs (not by hand!!!)

- If two or more mRNAs are derived from the same locus but have differing structures, then this gene is said to be alternatively spliced

- Thus alternative splicing is defined by different variations or "isoforms" of the mRNA that come from a shared gene

Alternative Splicing

- Genome wide alternative splicing analyses began in earnest in the late 1990s

- Requirements:
  - Genome sequence (spanning entire gene-space)
  - EST collections
  - Alignment/mapping strategy (e.g. PASA)
  - Detection of alternative isoforms (coordinate based strategy after mapping)

- First genomes that were analyzed using a genome-wide strategy were human and mouse
  - Initial rates of genome wide analysis was 10-20% of genes were alternatively spliced
  - Subsequent rates of alternative splicing have expanded to >60% of alternative splicing in human
  - Humans have > 7million ESTs currently and mouse has >4 million ESTs
  - As EST numbers increase, alternative splicing rates in a genome are also found to increase.
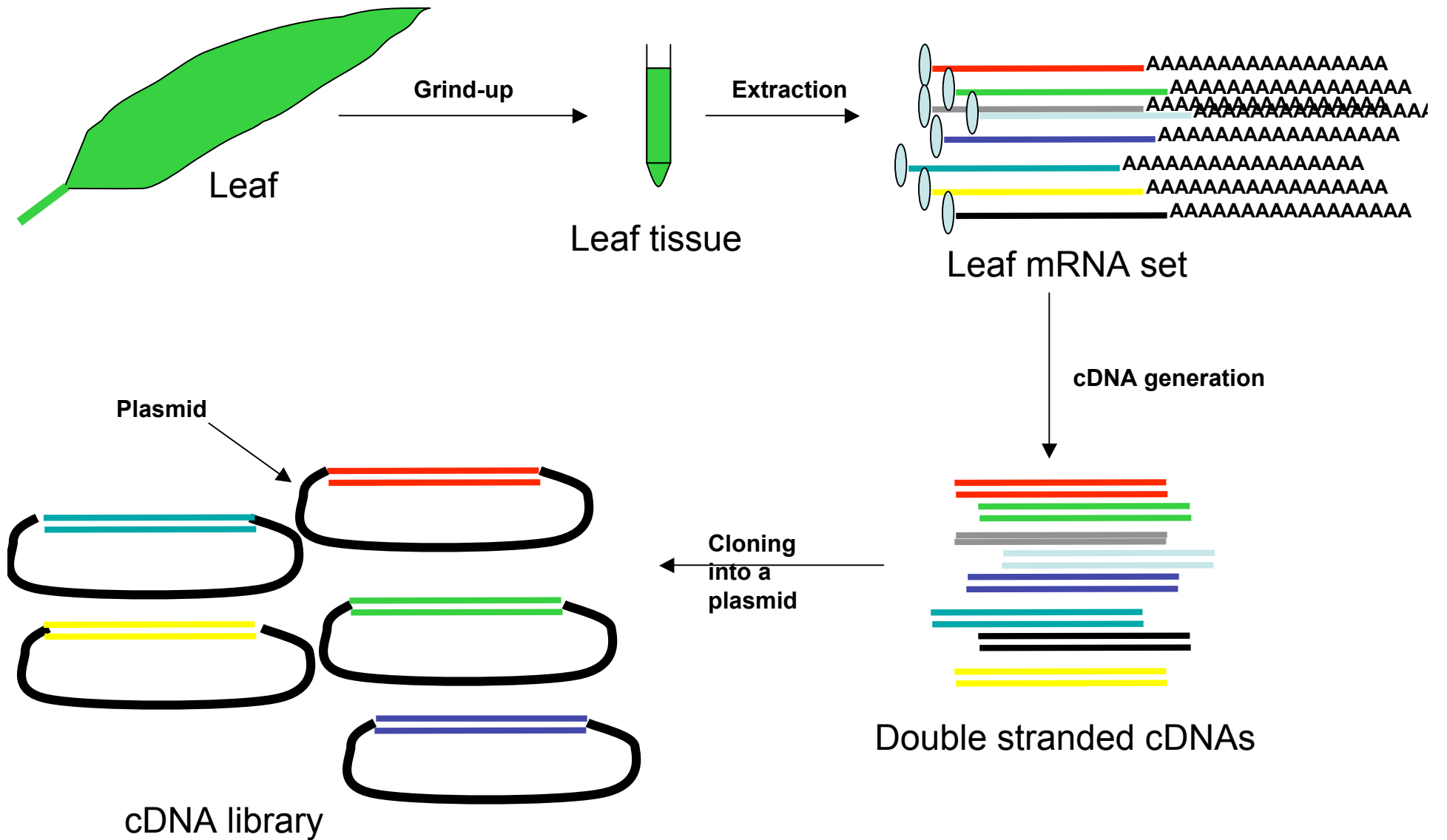
Alternative Splicing

- Interest in alternative splicing analyses in humans became greater with the annotation of the genome and the relatively low gene number

- The complexity of human development is difficult to explain using only 30,000 genes

- Humans have had a recent expansion of the transposable element *Alu* in the genome

    - 4% of human protein coding genes contain at least one *Alu*

    - Comparative analyses among primates have showed that *Alu* elements in introns can be converted into an exon (SLOWLY)

    - Thus far, all trancribed human genes with an *Alu* element are alternatively spliced
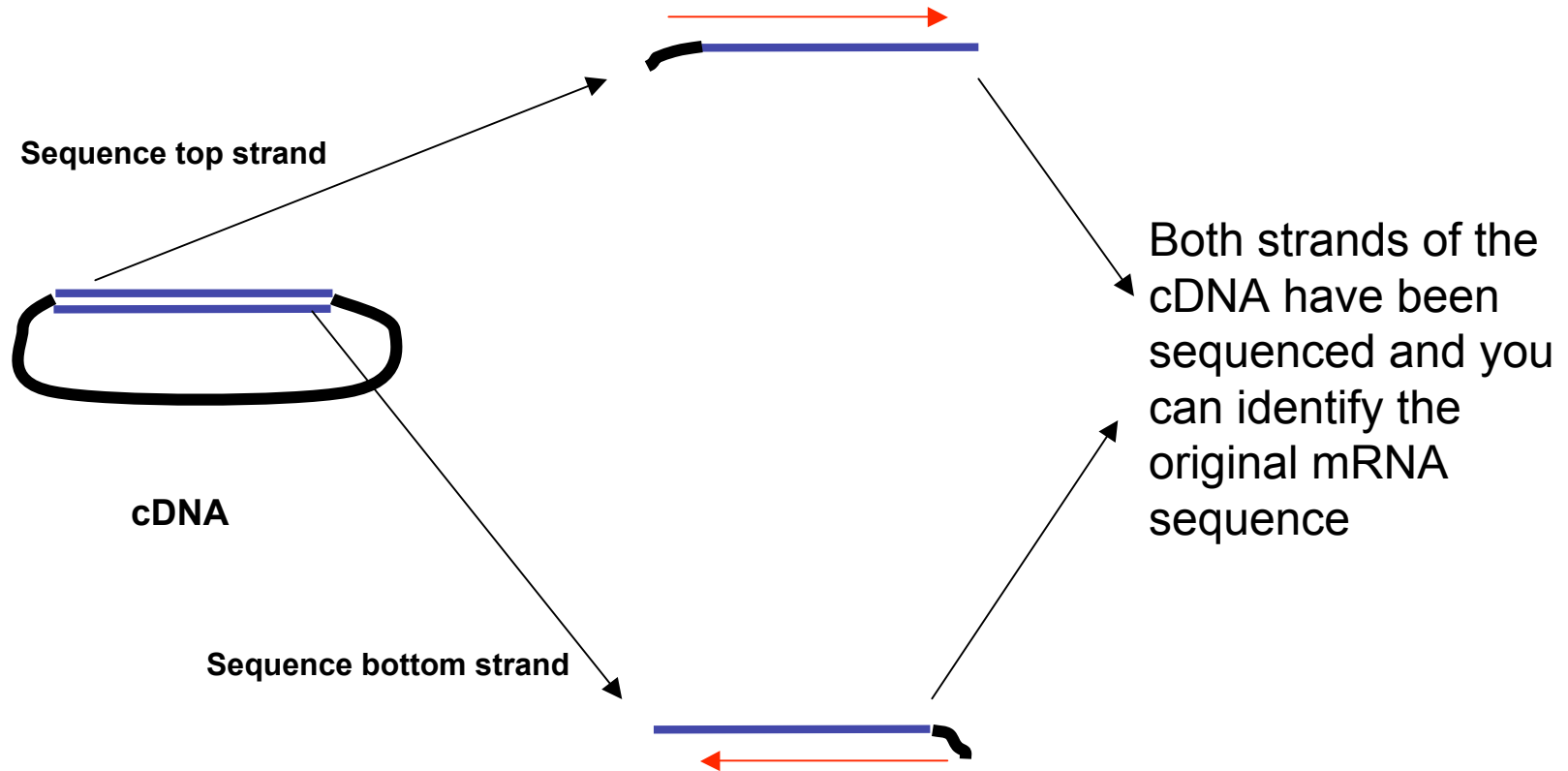
Alternative Splicing

- Phenotypic effects of alternative splicing are most commonly noted in human disease
  - Cystic fibrosis is commonly associated with SNPs that reduce splicing of exons 9 and 12
  - Recent multivariate analysis has supported the intuitive idea that longer genes with higher numbers of introns are more likely to lead to splicing defects
  - Proposed that between 15 and 60% of hereditary diseases are due to splicing mutations
  - Diseases known to be caused by splicing disorders include: cancer, neurodegenerative diseases, ataxia, retardation, hypercholesterolemia,

- Its direct role in phenotypic variation in other species is less clear and has been analyzed in a more circumscribed fasion

Generating a cDNA library

Sequencing a cDNA library

Sequence top strand

cDNA

Sequence bottom strand

Both strands of the cDNA have been sequenced and you can identify the original mRNA sequence

Repeat this procedure thousands of times from a cDNA library to get sample of which genes are expressed and what their structure is

# Expressed Sequence Tag (EST) sequence

Sequence top strand
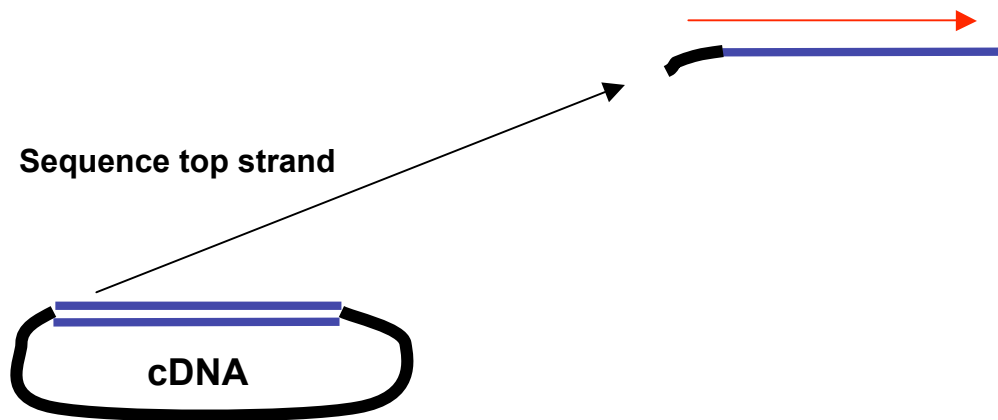
cDNA

This sequence of one end of a cDNA is often called an EST or expressed sequence tag.

Expressed sequence tags (ESTs) are often 500-700bp in length

ESTs reflect a partial sequence of a gene (usually) given that genes are commonly >1000bp in length

ESTs can be used to identify areas of the genomic sequence where there are genes – more intensive strategies are used to identify the full length of the gene

Aligning cDNAs to the genomic sequence

Sequence both strands

Generate a consensus sequence

cDNA

cDNA consensus sequence
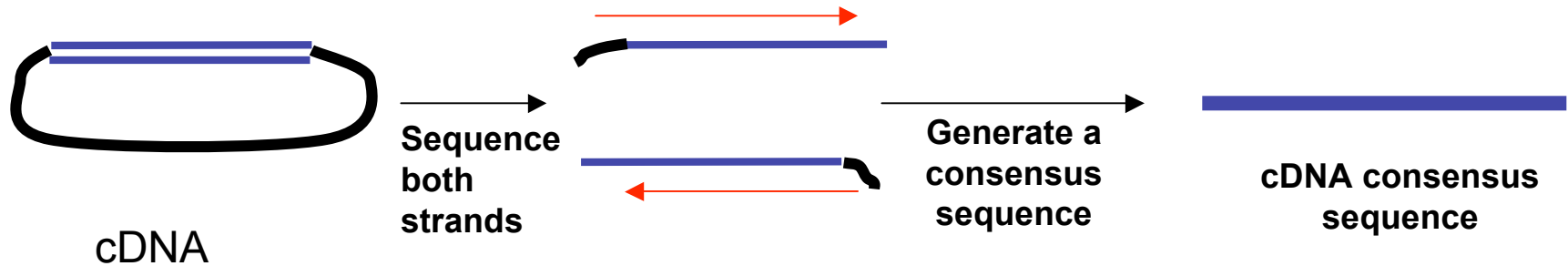
Now use the consensus cDNA sequence to align to the genomic sequence

cDNA

Genomic DNA sequence
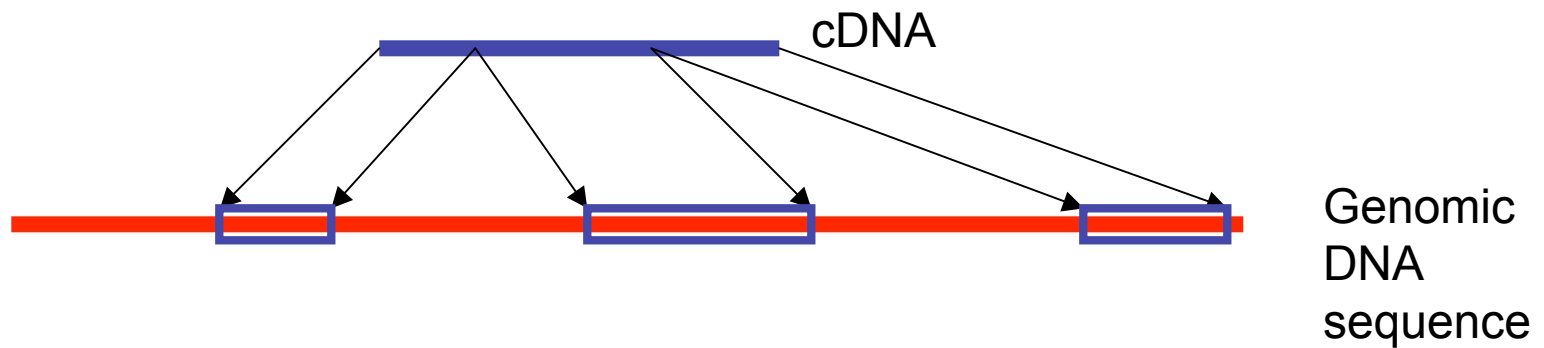
Note that the alignment of the cDNA to the genomic sequence is not contiguous, rather the cDNA sequence appears to be spliced blocks of disparate genomic sequence

Alternative Splicing

- ## Quick refresher of intron splicing
  - Higher eukaryotes use the following dinucleotide pairs in splicing
    - U2 based spliceosome
      - GT – AG (99%)
      - GC – AG (1%)
    - U12 based minor spliceosome
      - AT – AC (<0.1%)
  - The dinucleotides are the two nucleotides in the intron that abut the exon
  - There is a branch point site is upstream of the 3' acceptor in the intron
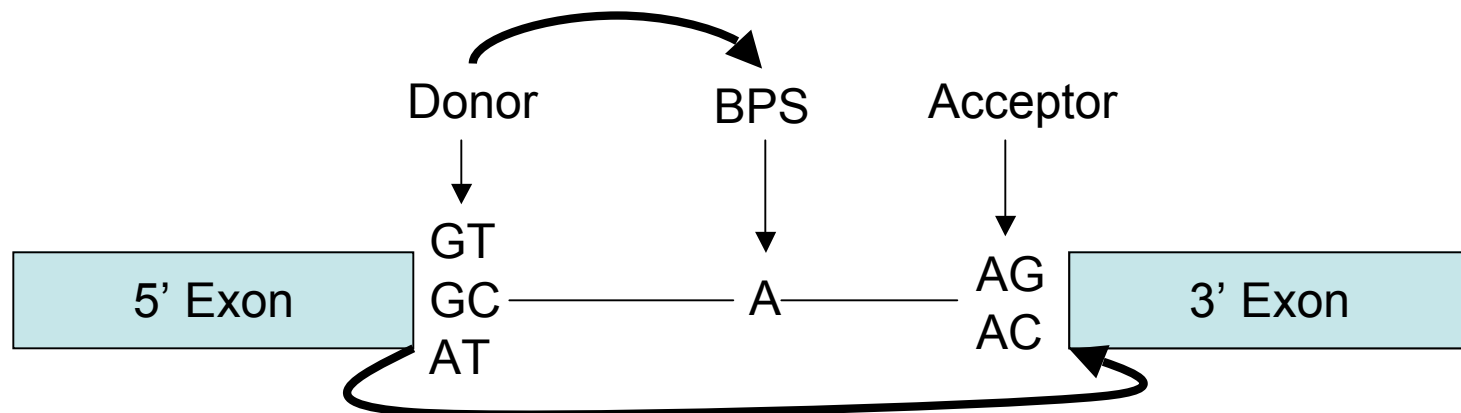
Alternative Splicing

- Consensus sequence for the acceptor site (32 bp around the splice site)

  - Note the polypyrimidine tract upstream
  - Note the AG dinucleotide is invariant
 - Note the sequence downstream of the AG is a randomized mixture (reflective of the coding sequence)

Alternative Splicing

- Consensus sequence for the donor site (32 bp around the splice site)

  – Note the polypyrimidine tract downstream
  – Note the GT dinucleotide is nearly invariant (GC is observed)
  – Note the sequence upstream of the AG is a randomized mixture (reflective of the coding sequence)

Alternative Splicing

- Alternative splicing analyses have a common set of alternative splicing isoforms:

- Alternate acceptor
- Alternate donor
- Alternate Terminal Exon
- Skipped/Retained exon
- Alternate initiation (Initiation within an Intron)
- Alternate polyadenylation (Termination within an Intron)
- Retained Intron

Alternative Splicing
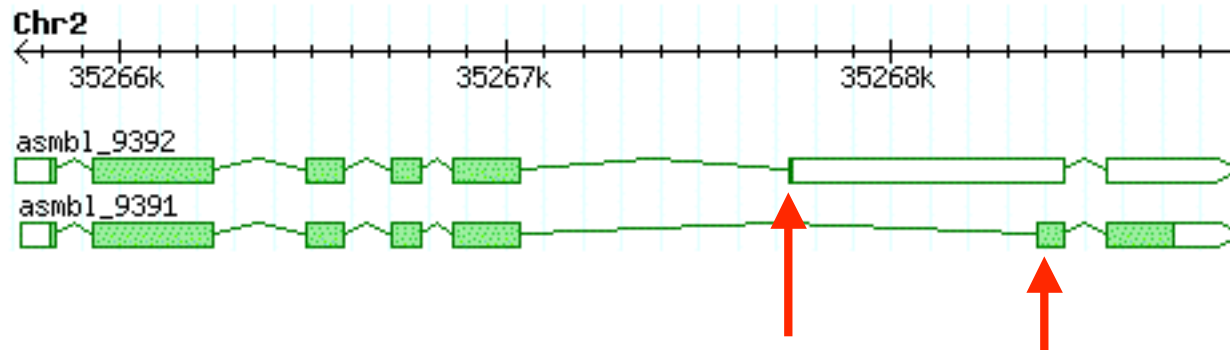
Note: The alternate acceptor has two different acceptor sites that are utilized

The first five exons are identical....but the 3' exons after the fifth exon are quite different

Asmbl_9392 has an upstream acceptor that incorporates a stop codon prematurely

Asmbl_9391 has a downstream acceptor site that extends the coding sequence
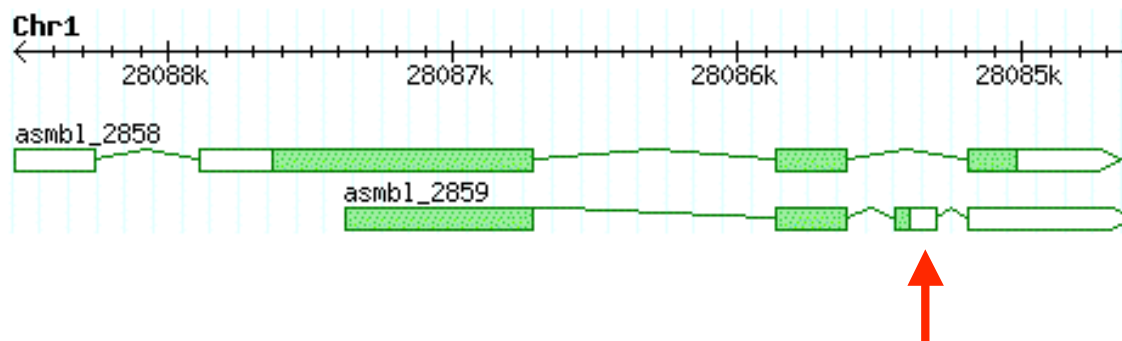
Alternative Splicing
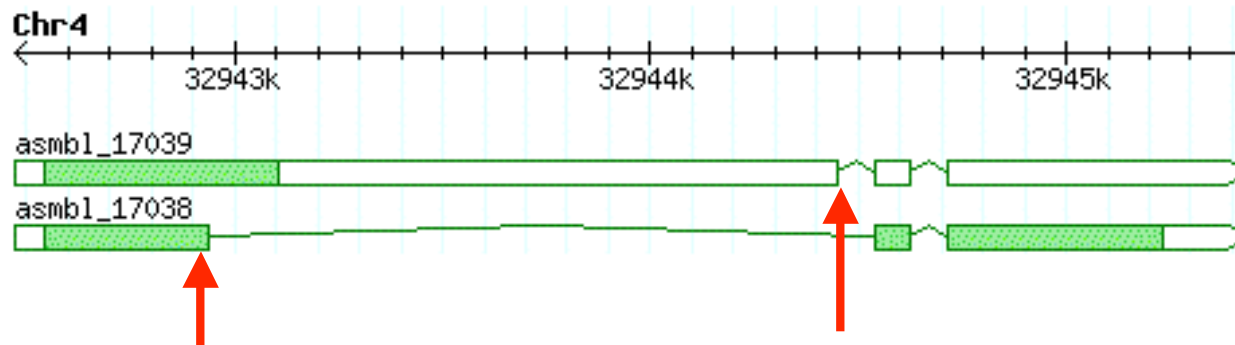
- Note: The alternate donor sites for the first exon are quite different

- Asmbl_17039 utilizes a downstream donor that incorporates a stop codon prematurely

- Asmbl_17038 utilizes an upstream donor site that extends the coding sequence

Alternative Splicing
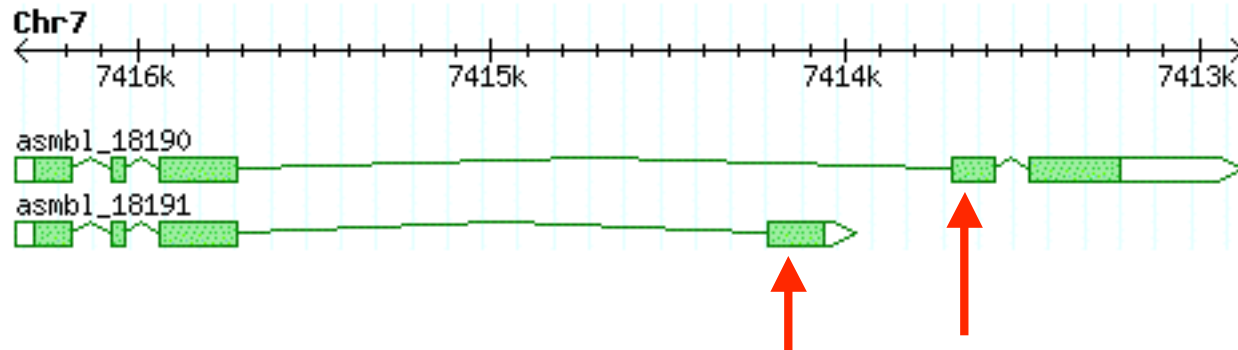
- Note: The alternate terminal exon has 3' terminal exon(s) that are completely distinct (i.e. there is no overlap)

- Asmbl_18191 utilizes a upstream splice site which has a different termination codon and transcriptional termination/polyadenylation site when compared with asmbl_18190

Alternative Splicing

- Note: The skipped/retained exon has differential inclusion of an internal exon

- Here the third exon of asmbl_2859 is not included in its cognate pair of asmbl_2858

Alternative Splicing
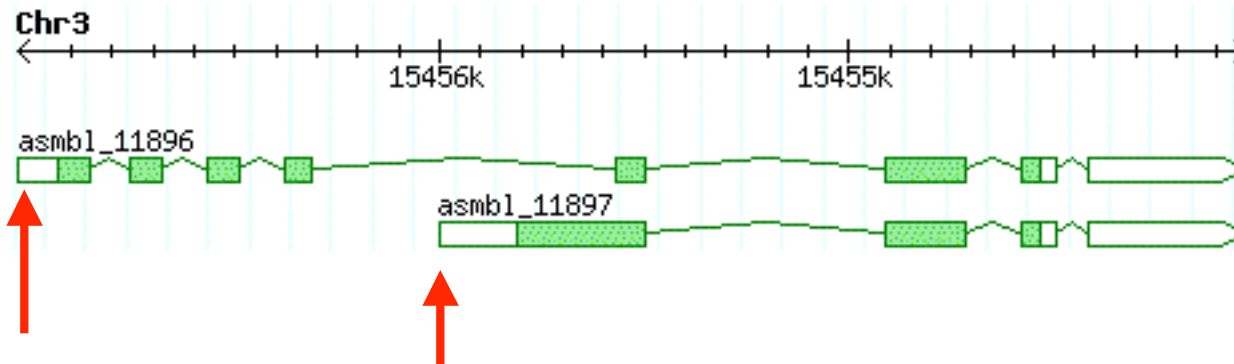
- Note: Alternate initiation (Initiation within an Intron) has two distinct initiation sites for transcription, and more narrowly, the initiation of one isoform is contained wholly within the intron of its cognate pair

- Here transcription of asmbl_11897 is initiated within the fourth intron of asmbl_11896 which leads to very different N-terminal proteins.

Alternative Splicing
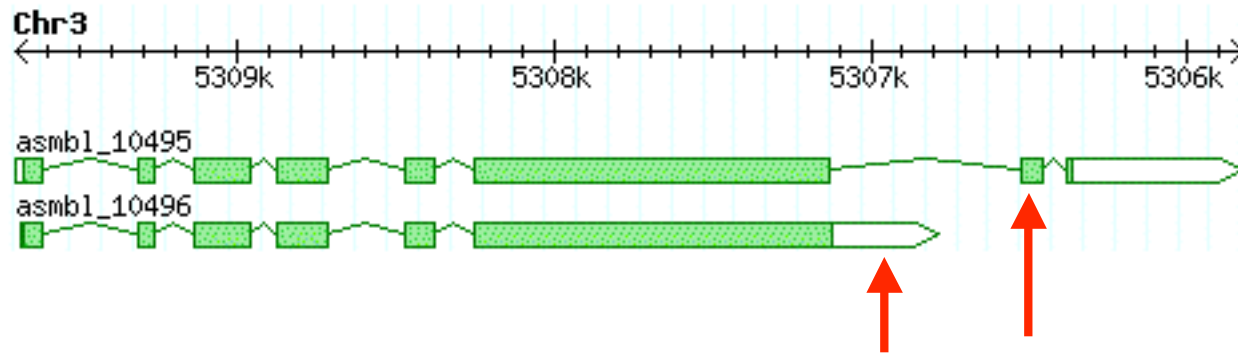
- Note: Alternate polyadenylation (termination within an intron) has two distinct initiation sites for polyadenylation, and more narrowly, the termination of transcription of one isoform is contained wholly within the intron of its cognate pair

- Here transcription of asmbl_10496 is terminated within the sixth intron of asmbl_10495 which leads to slightly different C-terminal proteins

Alternative Splicing
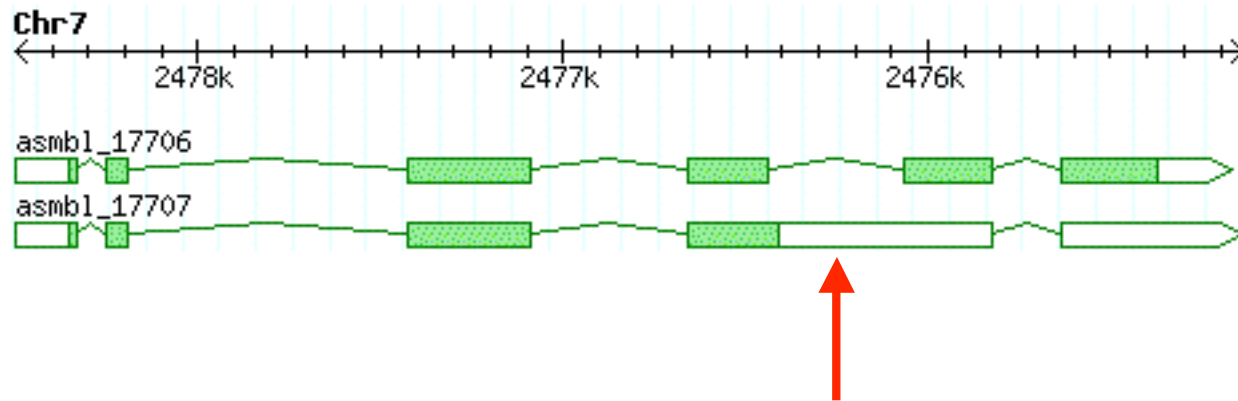
- Note: The retained intron isoform possesses an intron that is not spliced out in one isoform when compared to its cognate pair

- Here, the fourth intron for asmbl_17706 is not spliced out of asmbl_17707 which leads to a premature stop codon being introduced and very distinct C- terminal proteins.
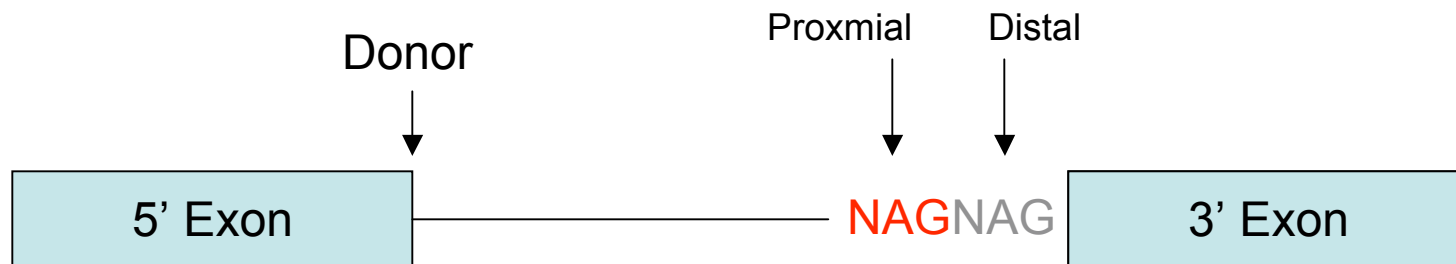
Example: Alternative Splicing in rice

– Rice FL-cDNA Consortium, 2003
  - >28,000 FLcDNAs produces 13.8% alternatively spliced loci

– Campbell et al., 2006
  - 1,156,705 ESTs and FL-cDNAs
  - 15.7% of all annotated rice genes are alternatively spliced
  - The three most frequent classes are alternate acceptor, alternate donor and retained intron

– Wang et al., 2006
  - 283,816 ESTs and FL-cDNAs
  - 21.2% of the transcribed genes are alternatively spliced
  - Retained intron was the most frequent alternative splicing class
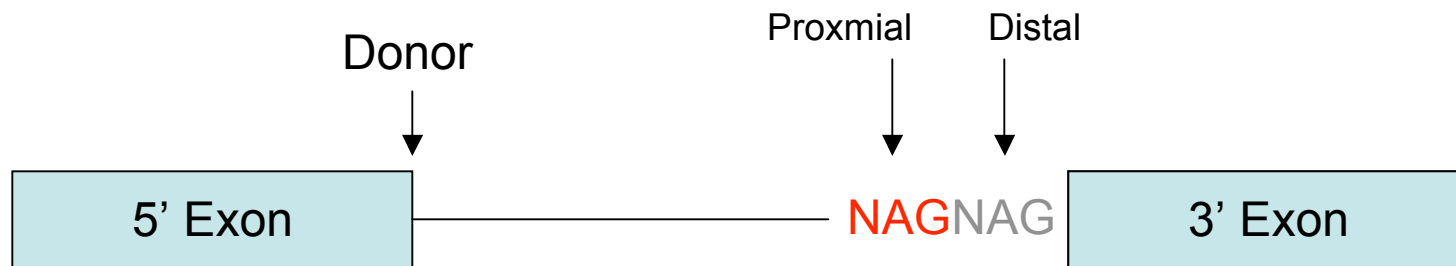
Isoform Variability at a Gene

- How many isoforms can a locus have?

  – Wang and Brendel found that alternative splicing at a locus creates a finite number of isoforms
    - ~60% of the transcribed loci in rice exhibit only a single isoform (no alternative splicing)

  – Campbell found similar results:
    - 68.3% of rice loci have only one isoform (no alternative splicing)
    - For those loci that are alternatively spliced, 57.3% have only two isoforms
    - 2 loci in rice have 17 different isoforms in rice and one Arabidopsis locus has 21 different isoforms

Alternative Splicing

- Analysis in human for the alternate acceptors looked at the rates of retaining frame
  - If the alternate acceptors are separated by a bp count evenly divisible by 3bp (e.g. 3, 6, 9, etc) then the frame of translation will be maintained
  - The authors found that the acceptors separated by 3bp are the most frequently occurring which is also observed in rice and Arabidopsis
  - The separation by 3bp will introduce a single amino acid in the isoform using the proximal acceptor (red) when compared to the isoform using the distal acceptor (grey)
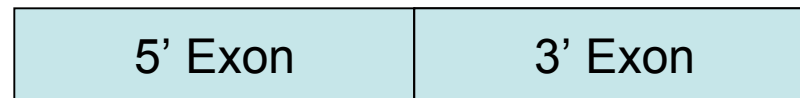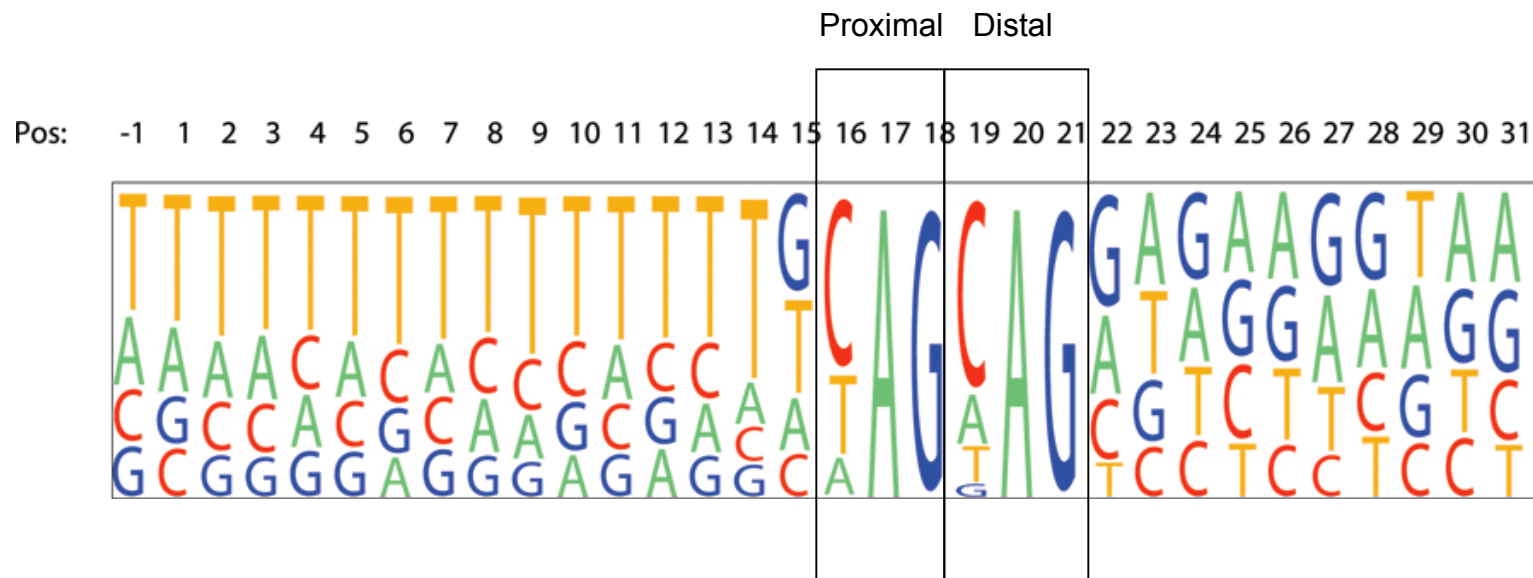 - The motif identifed is called 'NAGNAG' where the N stands for any nucleotide

Alternative Splicing

- Human analysis revealed that the NAGNAG was most commonly the sequence CAGCAG

- Below the CAGCAG motif is also observed in rice when looking at the 3bp difference between alternative acceptors.

- Note the CAGCAG motif is the most common

Alternative Splicing

- For alternate acceptor, alternate donor, retained intron and skipped exon classes where the difference in sequence length is not divisible by 3, what is the effect?

- Frameshift in the translated product!

- Frameshifting of the mRNA can lead to potentially new motifs being translated OR the incorporation of premature termination codons (PTCs)

- The introduction of PTCs into a mRNA has been shown to induce non-sense mediated (NMD) decay pathways

- This proposed NMD degradation offers a post-transcriptional regulatory role

Alternative Splicing

- The rule in mammalian systems has been that NMD pathways are induced when there is a PTC >50bp upstream of an exon/exon junction

- Degradation occurs at the time of translation where the PTC is detected in the ribosome relative to the exon junctions

- Recent work in mammalian systems suggests that over one-third of all alternatively spliced transcripts meet the criteria for NMD

- This phenomenon of high rates of alternative splicing leading to NMD is called Regulated Unproductive Splicing and Translation (RUST)

Alternative Splicing

- Wang and Brendel found:
  - 62% of alternative splicing isoforms in Arabidopsis will alter the frame of translation
  - 56% of alternative splicing isoforms in rice will alter the frame of translation

For these alternative splicing isoforms that have a frameshift:
- In Arabidopsis, 42% of these frameshifted events are found to
        introduce a premature stop codon >50bp upstream of the
        last exon-exon junction and are NMD candidates
-In rice, this percentage is 36%

This level of producing NMD candidates suggests that RUST may be present in higher plants and be linked with alternative splicing

Alternative Splicing

Conclusions:

- Alternative Splicing is a common feature of plant genomes

- As the size of EST collections have expanded, rates of alternative splicing increase as well

- Alternative splicing generally creates a few isoforms at any give locus – restricting its role in expanding proteomic diversity

- Higher plants show evidence of alternative splicing being linked with NMD – similar to mammalian systems