THESIS

# A COMPARATIVE ANALYSIS OF ALTERNATIVE SPLICING IN *A. THALIANA* AND *C. REINHARDTII*

Submitted by

Adam Labadorf

Department of Computer Science

In partial fulfillment of the requirements

for the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2010

COLORADO STATE UNIVERSITY

March 24th, 2010

WE HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER OUR SUPERVISION BY ADAM LABADORF ENTITLED A COMPARATIVE ANALYSIS OF ALTERNATIVE SPLICING IN *A. THALIANA* AND *C. REINHARDTII* BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

Committee on Graduate Work

Sanjay Rajopadhye

Anireddy Reddy

Advisor: Asa Ben-Hur

Department Head: Darrell Whitley

ABSTRACT OF THESIS

# A COMPARATIVE ANALYSIS OF ALTERNATIVE SPLICING IN *A. THALIANA* AND *C. REINHARDTII*

The extent of and mechanisms causing alternative splicing in plants are not currently well understood. A recent study in the model organism *Arabidopsis thaliana* estimates that approximately 42% of intron-containing genes are alternatively spliced and it is speculated that this number may be much higher [1]. Results from our pevious studies showed that the single celled alga *Chlamydomonas reinhardtii* also exhibits alternative splicing characeristic of plants [2]. In this work we present the results of a comprehensive alternative splicing analysis using the largest Expressed Sequence Tag (EST) datasets available for both of these organisms, describe an analysis pipeline tailored to these large datasets, and conduct a cross-organism comparative analysis of aspects related to alternative splicing.

Adam Labadorf
Department of Computer Science
Colorado State University
Fort Collins, CO 80523
Spring 2010

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Asa Ben-Hur, for his constant support and guidance rendered during this thesis work. I also extend special thanks to my colleague Mark Rogers for his assistance and support with strategies for detecting alternative splicing. I am grateful to Dr. Anireddy Reddy for providing guidance with the biological aspects of this work and to Dr. Sanjay Rajopadhye for serving on my committee. Lastly I would like to thank my colleagues, friends, and family for their support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# Chapter 1

# Introduction

Alternative splicing (AS) is a biological mechanism that has important implications in gene regulation and protein diversity within a cell [3]. Alternative splicing analyses are typically facilitated by comparing DNA sequences obtained from ESTs or cDNAs against a known genomic sequence. The task of determining where a EST originates in the genome is followed by successive steps to filter, refine, and collate the data in such a way as to allow detection of alternative splicing events. The ultimate goal of alternative splicing analyses is to identify and understand the underlying biological mechanisms that cause alternative splicing. Alternative splicing has been heavily studied in animals but its mechanisms are not as well understood in plants. This study conducts an AS analysis on the largest available EST datasets for the two model organisms *Arabidopsis thaliana* and *Chlamydomonas reinhardtii*.

The remainder of this document is organized as follows. In the next chapter is a detailed explanation of alternative splicing and how it can be detected, including a brief overview of the biology, algorithms, and strategies used. The following chapter discusses the specific datasets and analysis pipeline developed for this study and how design choices were made to deal with issues related to using these particular data. The results of the analysis are then presented followed by a discussion of splice site motifs for both organisms. The document concludes with ideas on further work.

## 1.1 Glossary of Terms

Because there are many biological terms in this work, a glossary of terms has been added at the end of this document to help better follow the text.

# Chapter 2

# Background

## 2.1 Genes, Gene Splicing, and Protein Synthesis

An organism's genome is a double-stranded molecule made of deoxyribonucleic acid (DNA) whose sequence stores information about how a cell functions. [4]. Specifically, the genome contains a number of distinct *genes* whose primary role is to code for proteins that control nearly all aspects of a cell. DNA, and therefore a gene's sequence, is made up of a combination of four molecules called *nucleotides* or *bases* - adenine (A), cytosine (C), guanine (G), and thymine (T). In the first step of coding for a protein, genes undergo the process of *transcription* whereby the gene's nucleotide sequence is copied by cellular machinery into a *precursor messenger RNA* or *pre-mRNA* molecule. The pre-mRNA that is copied from a gene exhibits *base complementarity* with respect to the gene's sequence, meaning that a cytosine in the original gene pairs with a guanine (C → G), a guanine pairs with a cytosine (G → C), a thymine pairs with an adenine (T → A), and an adenine pairs with a uracil (A → U), uracil being the RNA equivalent of thymine.

Once a pre-mRNA molecule has finished being copied from a gene, it then undergoes a separate process to prepare it for *translation* into a protein. In higher eukaryotes, many genes' coding sequences are interrupted by non-coding subsequences called *introns* that are excised from pre-mRNA before translation into a protein. This excision of introns and joining of coding sequences from pre-mRNA molecules is called *splicing* and transforms the pre-mRNA into mature *messenger RNA* or *mRNA*. The mRNA subsequences that remain after splicing, called *exons*, are concatenated together and it is this sequence that is translated into a protein. Proteins are composed of sequences of molecules called *amino acids*, where a single amino acid is specified by three consecutive bases of RNA called *codons* (*e.g.* the RNA codon CAU codes for the

amino acid Histadine). The pattern of introns and exons of a gene, called a *splice form*, dictates the final sequence of amino acids and therefore the protein's function. Figure 2.1 is a cartoon illustration of the process whereby a gene's sequence is ultimately translated into a protein.

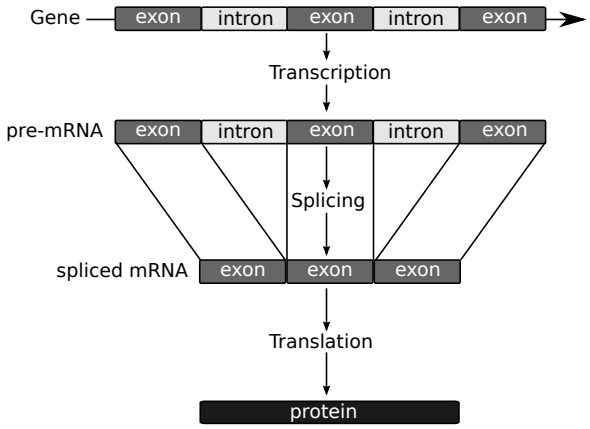


Figure 2.1: Simplified illustration of how a gene's sequence is translated into a protein.

The splicing machinery, or *spliceosome*, of a cell splices introns by recognizing splicing *signals* in the sequence of bases in pre-mRNA. The most common splicing signals are the dinucleotide pairs GT at the beginning of an intronic sequence and AG at the end of an intronic sequence, concisely written as GT/AG. The second most frequent splice site dinucleotide combination is GC/AG. These two splice site markers, GT/AG and GC/AG, are found in the vast majority of known introns and thus are referred to here as *canonical* splice site signatures. These dinucleotide pairs very often define the precise locations in a gene where pre-mRNA splicing will occur after transcription. The locations in a gene that define the boundaries between exons and introns are commonly called *splice sites* or *splice junctions*. Figure 2.2 is a cartoon illustration of an intron with canonical splice site junctions.

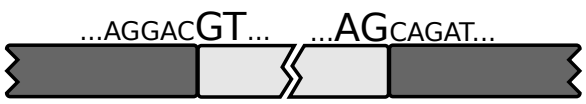


Figure 2.2: Illustration of an intron with canonical GT/AG splice site junctions. Exons are dark grey and the intron is light gray.

The start and end of a DNA molecule are labeled the *5'* and *3'* end, respectively. In a canonical intron like the one in Figure 2.2, the 5' end of the intron has the GT dinucleotide and

the 3' end has the AG dinucleotide. The same terminology is used to express the beginning and end of genes.

## 2.2 Expressed Sequence Tags

*Expressed Sequence Tags* (ESTs) are relatively short (usually no longer than 800 bases) DNA sequences that encode either the 5' or 3' end of a transcribed mRNA [5]. In a process called *reverse transcription,* mRNA molecules are first transcribed back into a single strand of *complementary DNA,* or cDNA (in the sense that the DNA contains the complementary bases of the mRNA, and thus the same sequence as the originating gene) and then that single stranded cDNA is converted into a double stranded DNA sequence. The DNA molecule can then be sequenced using a number of techniques [6] into a digital form and used for a large number of genetic analyses. When used in conjunction with a published genome, ESTs can be used for analyses that study how genes are spliced.

## 2.3 Alternative Splicing

When a single gene has more than one splice form it is said to exhibit *alternative splicing* (AS). AS has been implicated in many regulatory functions within a cell [3]. The important role and widespread occurrence of AS is now well documented and understood in animals but the mechanisms and extent of AS in plants are still comparatively unknown. Different splice forms of a single gene thus result in different final proteins whose function may be different, inhibited, or completely deactivated.

There are five primary forms of alternative splicing: donor site (Alt5'), acceptor site (Alt3'), simultaneous Alt3'/Alt5' (AltB), Exon Skipping (ES), and Intron Retention (IR). Many more complicated forms of AS exist [3], but the focus of this work is on these five forms because they are simple and the most prevalent. An AS *event* is defined as the occurrence of differential splicing involving introns and exons that correspond to the same region of a gene. More detail on how alternative splicing events are counted is found in section 3.4. Figure 2.3 is a cartoon illustration of types of alternative splicing.

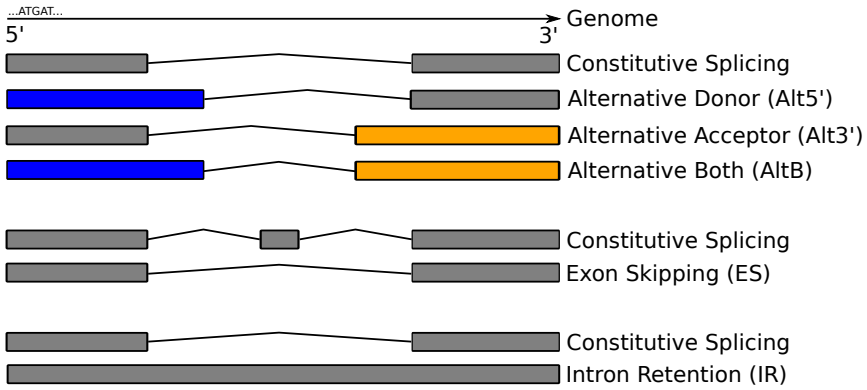


Figure 2.3: Illustrated examples of the primary forms of AS. Boxes and edges represent exons and introns, respectively. The top form is considered the constitutive, or most common, splice form. Other forms span the same genetic sequence but have different splice patterns as indicated.

## 2.4 Sequence Alignment

Sequence alignment is the most critical component of any AS analysis. In its simplest form, *aligning* two sequences of characters amounts to finding the longest common subsequence between the two sequences. For DNA sequences, alignment entails pairing two sequences of nucleotides such that the most nucleotides match and also such that the alignment is biologically relevant (*e.g.* it respects canonical splice site boundaries when appropriate). In the context of AS, the alignment of nucleotide sequences (*e.g.* ESTs) to the genome facilitates the detection of regions where a gene is alternatively spliced. Figure 2.4 is a simple illustration of two aligned DNA sequences.

Sequence 1: AGACATTAGATAGA
Sequence 2: --ACAGTAG---GA

Figure 2.4: Example of two aligned sequences. The complete Sequence 2 is ACAGTAGGA with gaps inserted to align the most characters between sequences.

Sequence alignment algorithms typically employ a dynamic programming strategy (*e.g.* Needleman-Wunsch [7], Smith-Waterman [8], PALMA [9]), a graphical model based approach (*e.g.* GENESEQER [10], HMMER [11]), a sequence indexing strategy as used in high-throughput alignment tools (*e.g.* BLAST [12], blat [13], PASS [14]), or some combination of these techniques (*e.g.* GMAP [15]). Each of these methods produces an alignment *score*

that is some measure of how well two sequences align. A common way to score an alignment is to factor the number of characters that match, the number that mismatch, and the number of gaps (indicated with a - in Figure 2.4) into a number that allows alignments to be distinguished from each other based on their biological significance. The proportion of characters that match in an alignment is often called the *sequence identity* or *percent identity* of an alignment. In the simplest algorithms, every gap inserted into one sequence or another incurs a penalty to the alignment score. However, since it is known that the sequences used in AS analysis have substrings removed from them (*i.e.* introns), penalizing for every gap individually is an inappropriate strategy. Therefore, the more specific problem of *gapped* or *spliced* alignment is solved [16], whereby long consecutive stretches of gaps in one sequence are allowed with little penalty to the final score.

One heuristic an alignment algorithm can use to guide alignment is favoring alignments with gaps marked by canonical splice site signatures. More generally, organism-specific splice site models can be constructed from known introns that often include a large number of bases up- and downstream of splice sites to improve alignments further. PALMA [9] and HMMer [11] are two examples of alignment programs that utilize sophisticated, organism-specific splice site models to guide sequence alignment.

## 2.5 Computational AS Pipelines

Datasets used for AS analysis are typically made of DNA sequences that have been sequenced before they have been translated into a protein but after introns have been spliced out of them. As the introns have already been spliced, the first step is to perform a spliced alignment against a reference genome to determine the sequence's origin. Once alignments have been obtained, they are often filtered and edited to ensure the highest quality results in the final AS analysis. Typical filtering and editing steps are rejecting alignments that do not have sufficient sequence identity, and altering alignments to prevent spurious AS detection on account of alignment artifacts. After filtering, alignments are grouped together if their alignment coordinates overlap or correspond to a single gene, a process referred to here as *clustering*. The filtered, edited clusters of alignments are ready for AS detection and analysis. Specific details of the pipeline in this work is described in chapter 3.

## 2.6 Splice Graphs

Detecting alternative splicing requires determining where a set of alignments' splice sites disagree. One way to perform this analysis is to explicitly compare pairs of introns and exons between alignments for differences, but complicated splicing patterns make this approach difficult and computation-intensive. An alternative representation of a set of alignments is a graph called a *splice graph* [17] where exons are nodes and introns are edges. Figure 2.5 is an example of a splice graph constructed from a set of alignments.

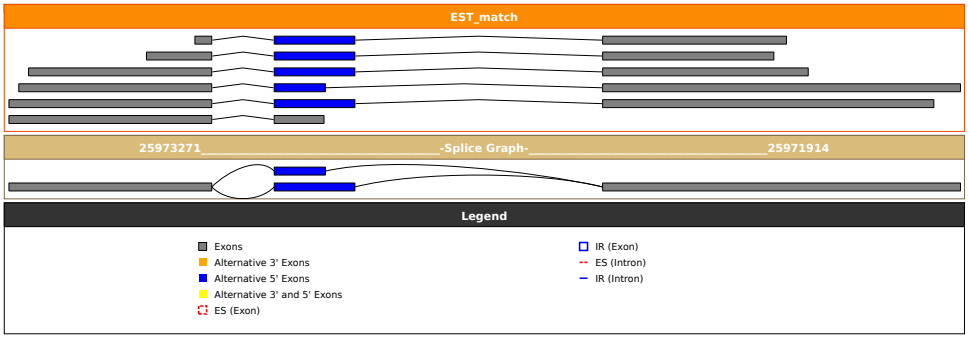


Figure 2.5: A splice graph constructed from a number of distinct alignments. Input alignments are in the top frame, the splice graph is in the bottom frame as labeled. Boxes and edges represent exons and introns, respectively.

A splice graph is constructed by consolidating alignments with splice forms that agree between alignments such that only non-redundant splicing information remains. In Figure 2.5, six input alignments are collapsed into a graph with just four nodes, where the middle two exons represent different splice patterns. Since only non-redundant information remains in a splice graph, AS event detection can be thought of simply as an analysis of the graph's nodes and edges. The splice graph representation of alignments also allows other analyses such as the effect of AS on the potential protein products of a gene as well as quantifying the complexity of a gene's splicing patterns. Splice graphs are used for AS event detection and analysis in this work.

## 2.7 Previous Studies

There have been several AS studies conducted in *Arabidopsis*. Iida *et. al* [18] aligned approximately 280,000 full length cDNA sequence against the *Arabidopsis* genome using BLAST [12] and discovered IR as the most prevalent form of AS in a modest set of detected events. In 2006,

Wang *et. al* [19] determined that approximately one fifth of all expressed genes in *Arabidopsis* demonstrate AS by using GENESEQER [10] to align the largest EST dataset available at the time of publication to the *Arabidopsis* genome. More recently, Filichkin *et. al* [1] used high-volume, next generation sequence data in *Arabidopsis* to suggest that approximately 42% of all intron-containing genes show evidence of AS. These and other studies indicate that the more data used in the analysis, the more AS events can still be discovered. This study furthers the effort of AS detection in *Arabidopsis* by using a large dataset of ESTs combined with our chosen methods for sequence alignment, filtering, and AS detection.

To date, little AS analysis has been conducted on *Chlamydomonas reinhardtii*. This model organism is particularly interesting because it is a single celled eukaryote that exhibits properties of both plants and animals [20]. A single celled alga, it has a chloroplast and undergoes photosynthesis in the presence of light but can survive in total darkness when alternative nutrients are available. *Chlamydomonas* also has a light-sensitive spot and two flagella that allow movement, attributes typically only found non-photosynthetic organisms. In our previous work [2], we found that *Chlamydomonas* exhibits AS patterns similar to that of plants using the largest available dataset of ESTs. The pipeline developed for this work is applied to the same dataset with the goal of extending this preliminary analysis and results.

# Chapter 3

# Methods

Our pipeline is depicted in the schematic in Figure 3.1 and generally follows the pattern of other AS pipelines.

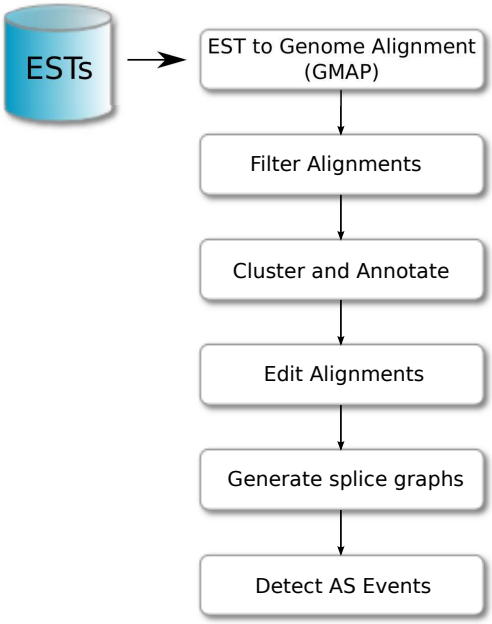


Figure 3.1: Steps in our AS pipeline.

The pipeline consists of first aligning each organism's ESTs against its genome. The resulting alignments are filtered based on quality and to isolate a single best alignment for each well-aligned input EST. Next, the unique EST alignments are clustered based on the known genes they overlap or, if no gene is found for an alignment, whether alignments overlap each other. After clustering, the alignments are edited to remove alignment artifacts and improve overall quality. The edited, clustered alignments are then analyzed for AS events using a modified version of the

Sircah software package [21]. Each of these steps is now discussed in detail.

## 3.1 Sequence Alignment and Alignment Filtering

Sequence alignment is the most important aspect of any AS analysis and is typically the first step of AS detection pipelines. The alignment tool GMAP [15] was chosen to perform the alignments because it is particularly well suited for performing spliced alignments on large datasets. GMAP is designed to find correct alignments for sequences that other popular high-throughput alignment tools like blat [13] have trouble finding. Specifically, GMAP has a sophisticated intron and consensus splice site model that favors aligning sequences to canonical splice site junctions at the possible expense of introducing mismatches or gaps into the alignments. Because the fully dynamic programming algorithm has high computational complexity, GMAP first performs a coarse alignment of sequences using an efficient hashing scheme similar to that which blat uses. As a result of combining these strategies, GMAP very quickly and accurately aligns very large datasets. The datasets in this work were also aligned with blat so the effects of the different alignment tools could be explicitly compared.

The dataset sequences were aligned against the TAIR9 *Arabidopsis* genome [22] and Chlre4 *Chlamydomonas* genome [23], respectively. Alignments were required to have 80% identity using GMAP, and alignments were further filtered to require 90% identity for each exon. Due to gene duplication events, it is often the case that one RNA sequence aligns well in more than one location across the genome. To avoid ambiguous alignments, the single best alignment by percent identity was identified for every EST.

## 3.2 Clustering and Annotation Mapping

The filtered alignments were next *clustered* using the organisms' respective genomes. Clustering in this pipeline refers to the process of determining which alignments overlap in genomic coordinates and mapping the resulting sets of alignments to known genes where possible. A set of overlapping alignments, called a *cluster*, maps to a known gene if any portion of the cluster overlaps a gene's genomic coordinates as specified in the genome annotation. Figure 3.2 is a cartoon illustration of the ways alignments are clustered.

In Figure 3.2 short horizontal lines represent alignments and the alignments in each box

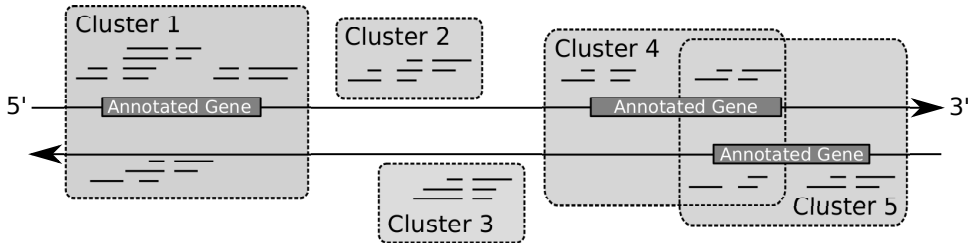


Figure 3.2: Examples of alignment clustering. Short horizontal black lines are alignments and all alignments in a box are in the same cluster.

belong to the indicated clusters. Cluster 1 illustrates how the alignments that overlap a single annotated gene are grouped into a single cluster irrespective of their strand. Clusters 2 and 3 overlap by genomic coordinate, but as they do not overlap an annotated gene the alignments on opposite strands remain on in separate clusters. There are instances in both the *Arabidopsis* and *Chlamydomonas* genomes that two genes overlap each other on opposite strands as illustrated on the right side of Figure 3.2. Some of the alignments in this illustration overlap both genes, some on the top strand and some on the bottom. In these cases, the alignments that overlap both genes are included in clusters corresponding to both genes. A later editing step described in section 3.3.5 will resolve which alignments should be associated with which gene in these cases. The most current gene annotations for TAIR9 and Chlre4 genomes were used for *Arabidopsis* and *Chlamydomonas*, respectively [22, 23].

## 3.3 Alignment Editing

Even though the alignments produced by GMAP are generally good, there are some alignment edits we can perform to further reduce the likelihood of detecting spurious AS events. Each of these edits is now discussed in detail.

### 3.3.1 Filtering Short Introns

The genomes of different ecotypes of the same species often contain small *indels* (*i.e.* insertions or deletions of nucleotides) unique to those ecotypes. Additionally, sequencing errors can introduce bases into EST sequences as an artifact of the sequencing process. A short deletion from one query sequence when compared to a reference genome of a different ecotype can result in an apparent short intron where there is none. To address this issue, introns shorter than 10 base pairs

were assumed to be the result of such deletions and therefore filtered from alignments. Figure 3.3 is an illustration of this issue.

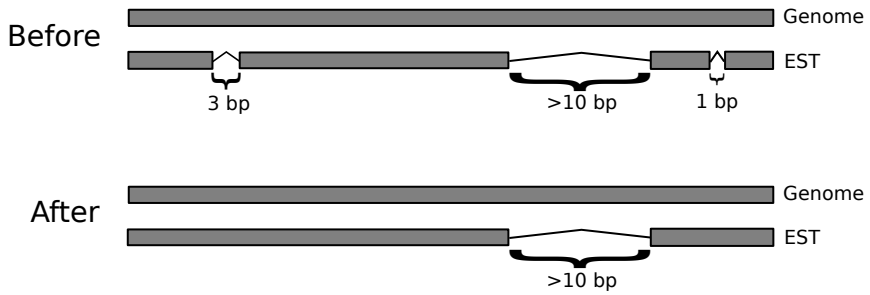


Figure 3.3: Filtering short introns to eliminate spurious AS events.

### 3.3.2 Removing Short Initial and Terminal Exons

Short sequences are statistically more likely to align to a reference genome by chance than longer sequences. A short portion of an EST that is aligned at the beginning or end of an alignment may therefore be aligned incorrectly, potentially generating a spurious alternative splicing event. Since ESTs often span multiple exons, it is beneficial to retain the portion of an alignment that has longer and more confidently aligned subsequences. This edit removes exon alignments of $< 10$ base pairs at the beginning and end of alignments leaving the rest of the alignment intact. Figure 3.4 is an illustration of this edit.

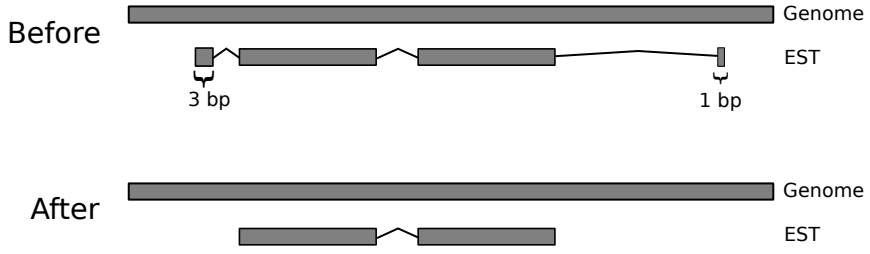


Figure 3.4: Removing short initial and terminal exons from an alignment.

### 3.3.3 Shifting Intron Coordinates

Examining clusters revealed situations where an intron within an alignment is shifted a small number of base pairs to the left or right of a true splice site. Figure 3.5 is an illustration of this situation.

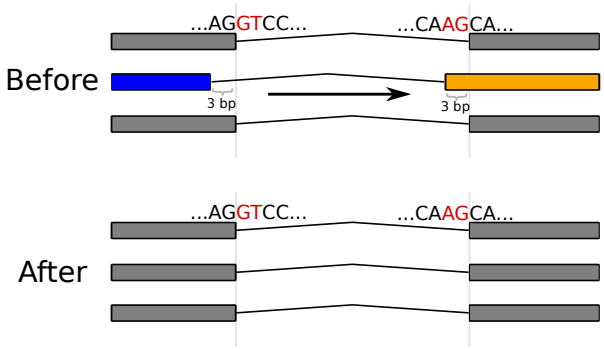


Figure 3.5: Illustration of an intron that is shifted by a small number of bases

The figure represents an alignment artifact in the sense that if the intron is shifted a small number of bases to the right the intron boundaries fall on a canonical (*i.e.* GT/AG or GC/AG) splice site and eliminate a spurious AS event. In this edit, all introns are checked whether they fall on a canonical splice site boundary. If not, a 20 base pair margin around the existing alignment splice sites is examined for canonical splice site signatures that are the same distance apart as the length of the intron. If a set of canonical splice sites is found the intron alignment is adjusted to reflect these new splice sites. This edit may introduce mismatches into the alignment, but a canonical splice site junction is more likely to result in a correct alignment than sequence identity and so is preferred over alignment identity.

### 3.3.4 Filtering Alignments with Short Exons

Similarly to that described in section 3.3.2, this filtering step is also related to the fact that short sequences have a high probability of alignment by chance. In some situations, an alignment favors a small exon in the middle of an existing intron rather than on the end of a flanking exon if the percent identity is higher. In most cases this is a spurious alignment, and should therefore be filtered. Figure 3.6 is an illustration of the problem.

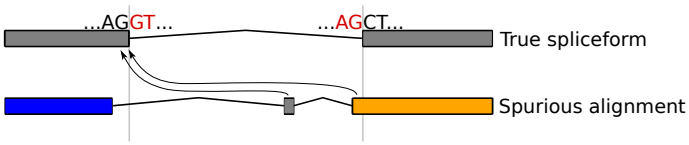


Figure 3.6: Illustration of filtering of short exon alignments. Potentially due to mismatches between two sequences, a small portion of one end of an exon may be split up and aligned spuriously within an intron, or even attached to the opposite exon.

While it would be possible to adjust such a spurious alignment to its true position attached to a flanking exon, there are many possible cases to consider in such a situation that complicate the task. Therefore, alignments with complete exons shorter than 10 base pairs are removed from the final dataset.

### 3.3.5 Adjusting Alignment Strands

The clustering protocol described in section 3.2 maps alignments to annotated genes if the alignments overlap the gene's coordinates on either strand. Due to the way ESTs are prepared and sequenced, it is often the case that the bases of a sequenced EST are reverse complemented with respect to the gene from which it originates. Since the alignments have been mapped to a gene whose strand is known, we can easily distinguish between alignments that map to the annotated strand of the gene and those that map to the opposite strand. An alignment that is reverse-complemented with respect to its originating gene may have had introns spliced out of it that correspond to true splicing events of the gene. It is therefore beneficial to determine whether an alignment should be considered on the same strand as a gene when it is mapped to the opposite strand. Additionally, there are situations in the *Arabidopsis* genome where two genes overlap on opposite strands, so it is necessary to explicitly distinguish which alignments should be assigned to which gene. This section describes two ways of accomplishing this assignment of strand for alignments that have been mapped to annotated genes.

In the first edit, an alignment mapping to the opposite strand of a gene is added to the set of alignments used for AS detection if it has any intron whose coordinates match an intron on the annotated strand. The probability of true splice sites having the exact same coordinate on opposite strands is extremely small, so this strategy is not likely to introduce spurious AS events. Figure 3.7 is an illustration of this edit.

In the second edit, alignment on the opposite that do not share any intron boundaries with those on the annotated strand have their splice site sequences examined for canonical junctions. If any intron's splice site nucleotides match a canonical signature on the annotated strand of the gene that alignment's strand is changed to match the mapped gene. Given the extremely high prevalence of canonical splice site junctions, it is unlikely that a true splice site on the opposite strand will have a canonical signature on the annotated strand. It is therefore also unlikely that

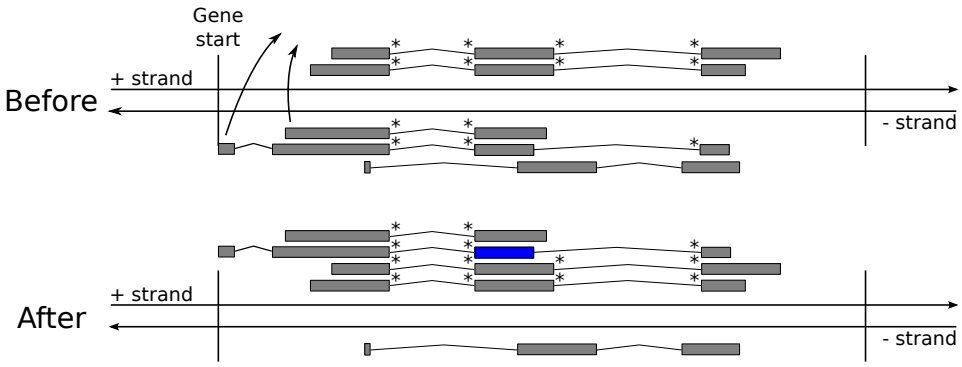


Figure 3.7: Adjusting alignment strands based on splice site coordinates. Starred splice sites have the same genomic coordinates. By including alignments that exactly share introns from the opposite strand, legitimate AS events may be detected. In this example an Alt5' AS event is evident only when strands are adjusted.

this edit will introduce spurious AS events. Figure 3.8 is an illustration of the canonical splice site strand adjustment.

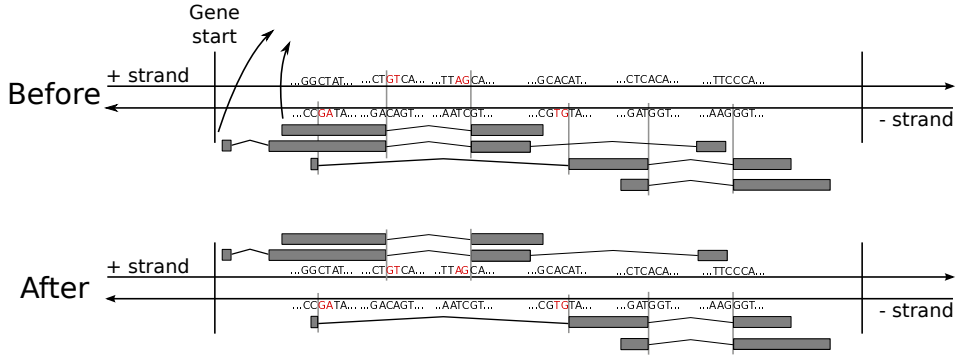


Figure 3.8: Splice site signature alignment strand adjustment. Canonical splice sites are highlighted in red for each strand. The mapped gene is on the + strand. The first two - strand alignments each have one intron that maps exactly to a canonical GT/AG splice site boundary and thus have their strands adjusted. The second two alignments have no introns that map to a + strand canonical splice site and so are unedited.

### 3.3.6 Filtering Insertions in EST Sequences

Insertions in EST sequences are the result of either sequencing errors or genetic differences between the reference genome and the genome where the EST originated, neither of which should result in the detection of AS events. Because AS events can only be detected against a single reference genome and not between genomes that differ, any insertions in EST sequences must be removed from consideration when conducting AS analysis. Therefore, only the portions

of EST sequences that align to the reference genome are used for AS detection, and alignment regions associated with insertions in EST sequences are removed from the resulting alignments.

## 3.4 Alternative Splicing Detection

The filtered, clustered, and edited alignments were analyzed for alternative splicing using a modified version of the software package Sircah [21]. The original software was modified for our pipeline because it counted AS events differently than we do in this analysis and more statistics needed to be extracted from splice graphs than were in the available implementation. The software constructs a splice graph from the alignment information in a cluster and detects discrepancies in the splice patterns. Figure 3.9 is an example splice graph of the *Arabidopsis* gene AT1G74650.

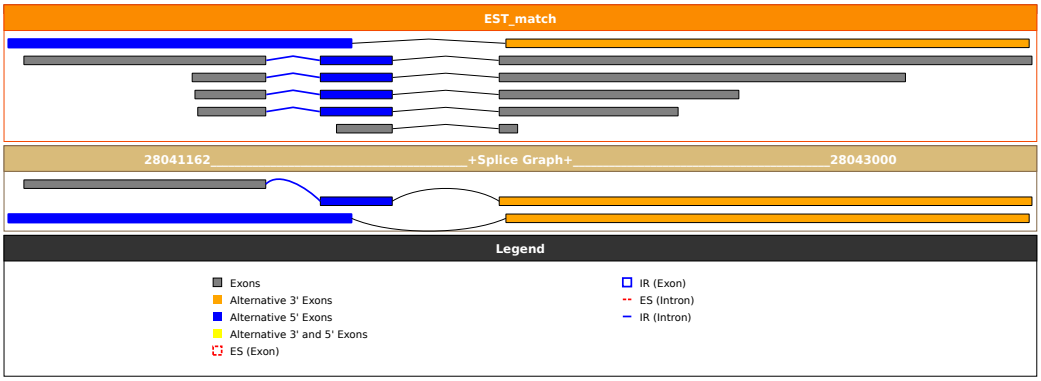


Figure 3.9: Example of a Sircah splice graph for a real cluster

In the example splice graph there are three distinct AS events: one Alt5', one IR, and one Alt3'. Counting AS events in this simple example is elementary, but how to count events in a more complicated splice graph is a little less clear. Since there are some ambiguous event counting cases, the rationale for counting AS events is now discussed.

Counting AS events in this analysis is linked to counting the number of distinct biological events that result in different splicing patterns. AS events are also counted from the perspective of the introns involved in the alternative splicing, rather than the exons. Figure 3.10 contains an example illustrating the approach.

In Figure 3.10, the intron is retained with respect to two exons and therefore may result in three distinct biological events - one with respect to the shorter exon, one with respect to

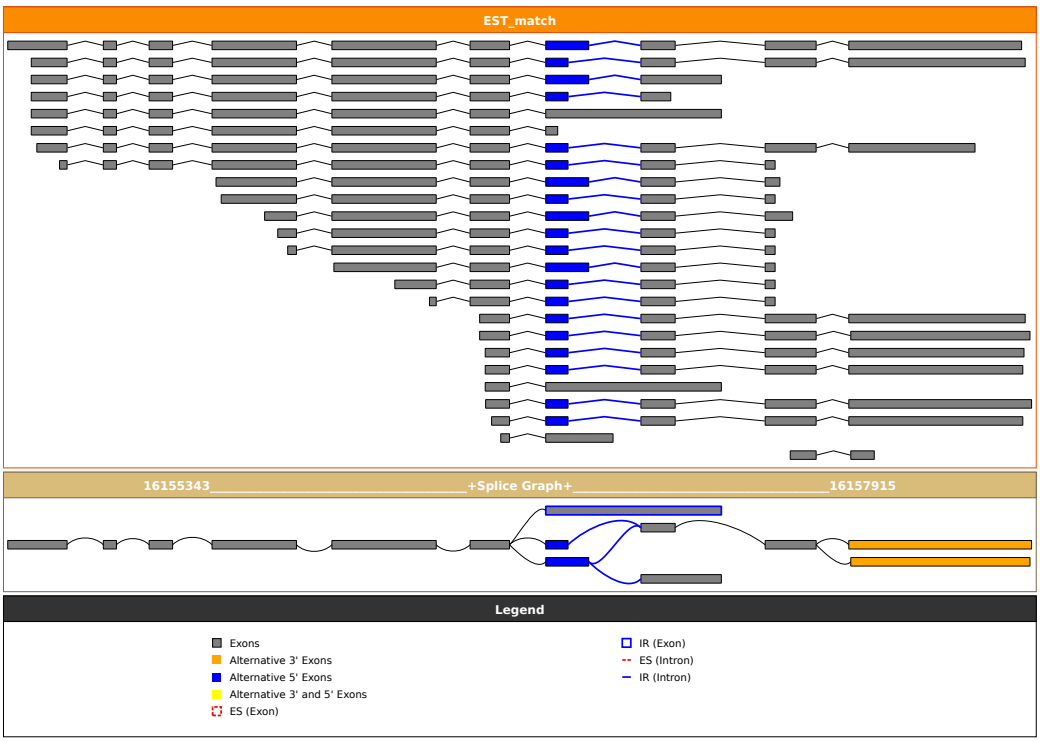


Figure 3.10: Counting AS events in terms of the number of resulting protein products. This gene is considered to have two IR events, one Alt5' event, and one Alt3' event.

the longer exon, and one with the intron totally retained - and therefore two splice forms are considered alternative. When considering the terminal Alt3' AS event, only two spliceforms result as a consequence of this AS and thus only one event is observed.

## 3.5 Implementation Details

The above pipeline was developed using the `python` programming language. GMAP was compiled from source and run using the command line options `--findcanonical --trimexonpct=0.8 --batch=2 -S --summary -f 1 --maxintron=<length>`. The `--maxintron` parameter was given as 6,000 and 2,000 for *Arabidopsis* and *Chlamydomonas*, respectively, as determined from each organism's gene annotations. Our modified Sircah software produced the splice graph images and provided all AS statistics.

# Chapter 4

# Results and Discussion

## 4.1 Datasets

All available *Arabidopsis* EST sequences were downloaded from NCBI's dbEST database [24], which at the time contained 1,527,299 sequences. Additionally, the *Arabidopsis* EST dataset used by Wang *et al* [19] included 71,806 sequences not found in dbEST and thus were also included, raising the total number of *Arabidopsis* sequences to 1,599,105. The *Chlamydomonas* EST database available [25] contained 252,484 ESTs, all of which were used in the analysis.

## 4.2 Pipeline Comparison

A fundamental problem in spliced alignment is that, in general, we cannot be sure that any given alignment is biologically correct. Every biological sequence alignment tool makes some assumptions about the biology underlying the data, and it is not always clear whether those assumptions are appropriate or what biases they introduce. Since it is generally accepted that most known legitimate splice site junctions are canonical, we may be reasonably confident in the AS events we detect if we require that all alignments contributing to detection demonstrate canonical junctions. However, recent studies suggest that there may be more legitimate non-canonical splite site junctions than previously thought [1], so enforcing this strict criterion of all-canonical splice site junctions may miss legitimate events. Thus, four versions of the pipeline were run on each organism's dataset. The first version aligned the entire dataset using GMAP without filtering based on canonical splice sites. The second version also did not filter on canonical splice sites but used blat as the alignment too. The third version also aligned the dataset with GMAP but filtered alignments based on the criterion that all introns in an alignment must have canonical

splice site junctions. The fourth version aligned the dataset using blat also with only alignments with all canonical introns. By comparing the AS statistics of the three pipeline versions to each other and to previous studies, we can make a judgment on which version of the pipeline is likely to give us the most reliable results. Table 4.1 contains AS event and gene counts for the three pipelines run on both organisms.

| *Arabidopsis* | | | | |
|---|---|---|---|---|
| **Event** | **GMAP w/o CSS** | **blat w/o CSS** | **GMAP w/ CSS** | **blat w/ CSS** |
| IR | 3,580 (36.0) | 5,767 (18.0) | 2,506 (44.0) | 2,373 (44.6) |
| ES | 1,159 (11.7) | 2,068 (6.4) | 635 (11.7) | 491 (9.2) |
| Alt3' | 2,974 (29.9) | 10,141 (31.7) | 1,683 (29.6) | 1,592 (29.9) |
| Alt5' | 1,772 (17.9) | 9,277 (28.9) | 815 (14.6) | 765 (14.3) |
| AltB | 422 (4.4) | 4,750 (14.8) | 84 (1.7) | 99 (1.8) |
| Total | 9,928 | 32,003 | 5,723 | 5,221 |
| *Chlamydomonas* | | | | |
| **Event** | **GMAP w/o CSS** | **blat w/o CSS** | **GMAP w/ CSS** | **blat w/ CSS** |
| IR | 429 (34.0) | 843 (16.2) | 262 (45.5) | 246 (43.2) |
| ES | 158 (12.5) | 327 (6.2) | 83 (14.4) | 63 (11.0) |
| Alt3' | 299 (23.7) | 1,572 (30.2) | 140 (24.3) | 154 (27.0) |
| Alt5' | 271 (21.5) | 1,520 (29.2) | 83 (14.4) | 90 (15.8) |
| AltB | 102 (8.1) | 939 (18.0) | 8 (1.4) | 16 (2.8) |
| Total | 1,259 | 5,201 | 576 | 569 |

Table 4.1: AS event count comparison for different alignment tools and filtering. CSS stands for Canonical Splice Sites. Canonical SS experiments were conducted with only alignments where every intron exhibits canonical splice site boundaries.

From Table 4.1, we first notice that the numbers of events for the GMAP and blat alignment pipelines without strictly canonical splice sites are much higher than those with only canonical junctions for either alignment tool. This is not surprising, considering that requiring *all* introns in an alignment to be canonical likely eliminates many canonical introns in alignments that contain only one non-canonical intron. It is certainly the case, however, that some portion of the events detected in the datasets with non-canonical splice sites are spurious. Also, we note that the proportion of IR events with the all-canonical pipelines is significantly higher than the pipeline that does not require canonical splice sites, while the proportion and number of Alt5' drops. It may be the case that some legitimate Alt5' events have non-canonical splice site junctions, but the fact that the number of detected Alt5' events decreases by more than half between datasets makes it seem unlikely that the differences are all true AS events. Interestingly, the propor-

tion of Alt3' events stays relatively constant across all experiments, and for *Arabidopsis* those proportions are significantly higher than reported in previous studies [19], which found ∼20%. This suggests that the proportion of Alt3' events is legitimately higher than previously thought. The version using blat and not filtering on canonical splice sites has vastly different proportions of events than the other pipeline results. This is consistent with previous observations of blat's performance, particularly concerning how blat often poorly aligns splice site junctions.

The GMAP alignments with all canonical splice site junctions were used for further analysis because we felt the most confident in the AS events detected with that combination. The event proportions from this version of the pipeline most closely match previously published work, which is discussed in more detail in the sections 4.3 and 4.4. We also have greater confidence in the alignments GMAP produces over blat. Since blat does not use a very sophisticated splice site model during alignment, there are often many spurious alignments around splice junctions when there are indels in query sequences. GMAP, on the other hand, was designed to favor canonical splice sites, resulting in much higher splice site alignment precision. The remainder of the analysis in the section uses the results of this version of the pipeline.

## 4.3 Pipeline Statistics

A total of 76.1% and 65.1% of ESTs were unambiguously aligned against the genomes of *Arabidopsis* and *Chlamydomonas*, respectively, using GMAP after alignments had been filtered for canonical splice sites. The sequences that could not be unambiguously aligned were either short, had unacceptably low sequence identity, or aligned to the chloroplast or mitochondiral genomic sequences.

For *Arabidopsis*, 1,218,069 ESTs were mapped to 28,251 annotated genes, pseudogenes, or transposable element genes in the TAIR9 genome annotation [22] and to 37,344 unique clusters that did not map to known annotated genes. The majority of these unmappable clusters lay in close proximity to known genes, but some clearly indicate previously unannotated transcriptional activity. A more detailed analysis of these clusters is later in this section. After filtering and editing, a total of 903,061 alignments were used to conduct alternative splicing analysis.

For *Chlamydomonas*, 164,433 ESTs were unambiguously aligned to 9,454 known genes and 4,497 unannotated regions. As with *Arabidopsis*, many of the clusters are associated with

known genes. However, as the *Chlamydomonas* draft genome annotation is much less mature than that of *Arabidopsis* at the time of this writing, there is a significant number of unannotated alignments that likely correspond to undiscovered genes. Table 4.2 contains summary statistics for the *Arabidopsis* and *Chlamydomonas* pipelines.

| **Pipeline Step** | *Arabidopsis* | *Chlamydomonas* |
|---|---|---|
| Initial ESTs | 1,599,105 | 252,484 |
| Initial Alignments | 1,824,113 | 244,329 |
| Filtered by % ID | 1,309,119 | 182,417 |
| Unique EST Alignments | 1,218,069 | 164,433 |
| Edited Alignments | 1,235,590 | 157,808 |
| Strand-Adjusted Alignments | 903,061 | 114,023 |
| Gene Clusters | 28,251 | 9,454 |
| Unannotated Clusters | 37,344 | 4,497 |
| All Clusters | 65,824 | 13,951 |

Table 4.2: Pipeline statistics for *Arabidopsis* and *Chlamydomonas* pipelines. The strand-adjusted alignments figures are the alignments used in the final AS analysis.

There is a number of clusters that do not map to any known annotation for both organisms. The majority of these unannotated clusters lie within 1000 base pairs of an annotated gene. However, there are a number of clusters that are very far from any known annotation with significant alignment support. For *Chlamydomonas*, there were a total of 107 unannotated clusters with more than ten high quality EST alignments where 76 and 31 were less than and greater than 1000 bases away from the nearest gene, respectively. For *Arabidopsis*, a total of 414 clusters with more than ten EST alignments and without a gene mapping were found, where 295 and 119 were less than and greater than 1000 bases away from the nearest gene, respectively. It is likely that the clusters greater than 1000 bases away from any known gene represent previously undiscovered genes or untranslated RNAs that have some other cellular function. Table 4.3 contains statistics on the unannotated clusters.

## 4.4 Alternative Splicing Analysis

Like in previous studies in *Arabidopsis* and *Chlamydomonas*, IR emerges as the most prevalent type of AS, followed by Alt3', Alt5', ES, and lastly AltB. In *Arabidopsis*, the proportions of AS events generally agree with Wang's analysis but deviate in three respects. First, while IR

| **Organism** | **$<$ 1000 bp (Genes)** | **$\geq$ 1000 bp (Genes)** | **Total (Genes)** |
|---|---|---|---|
| *Arabidopsis* | 295 (285) | 119 (115) | 414 (399) |
| *Chlamydomonas* | 76 (71) | 31 (26) | 107 (97) |

Table 4.3: Statistics on clusters with more than 10 ESTs that could not be mapped to known genes. Clusters are split into those nearer than 1000 base pairs from the nearest gene and those farther. In parentheses we indicate the distinct number of genes the clusters are nearest, as some clusters are nearest to the same gene.

| | *Arabidopsis* | | *Chlamydomonas* | |
|---|---|---|---|---|
| AS type | Events (%) | Genes (All) | Events (%) | Genes (All) |
| IR | 2,506 (43.7) | 1748 (1757) | 262 (45.5) | 231 (231) |
| ES | 635 (11.1) | 358 (358) | 83 (14.4) | 56 (56) |
| Alt3' | 1,683 (29.4) | 1526 (1527) | 140 (24.3) | 133 (135) |
| Alt5' | 815 (14.2) | 769 (771) | 83 (14.4) | 80 (81) |
| AltB | 84 (1.5) | 82 (82) | 8 (1.4) | 8 (8) |
| **Total** | 5,723 | 3428 (3430) | 576 | 445 (448) |

Table 4.4: Alternative Splicing Statistics. In parentheses of the Events columns are the proportions of the event type. In parentheses of Genes columns are the total number of clusters where the AS events were found, including unannotated clusters. Sum of Genes is less than Total figures because some genes have more than one AS type.

is the most prevalent form of AS, the numer of events and proportion (44.0%) are significantly smaller in this analysis compared to Wang who reported 4,635 events comprising 56.1% IR. An explanation for this might be the short intron filtering step this pipeline undergoes, whereby some EST deletions may have been previously detected as IR events. Also, filtering all alignments with any non-canonical splice site junctions also will have an effect on this proportion. This is because it is often the case that the IR splice form of a gene is the least prevalent form of an intron. Of the 9,219 *Arabidopsis* transcripts involved in IR, 6,588 (71.4%) of the transcripts had the retained intron as the non-prevalent form. Since the retained intron involved in IR is usually non-prevalent, and given that there are several ways an alignment can be filtered out of the analysis in our pipeline, it is possible that alignments containing retained introns are filtered before AS detection. This pipeline is generally more conservative than Wang's in the sense that an entire EST alignment may be eliminated if only part of the alignment fails a filter (*e.g.* short exon filtering edit, existence of a non-canonical splice site, etc.). The proportions of Alt3' and Alt5' events in this study, 29.6% and 14.6%, are both elevated when compared to the previous study of 21.9% and 10.5%. However, the number of Alt3' and Alt5' events detected in this

pipeline are nearly identical to the 1,810 and 845, respectively, reported by Wang. Thus, the relative proportions for these types of events are increased only because there were fewer IR events detected.

This analysis finds only 72% of the number of genes that exhibit AS reported by Wang. From Table 4.4, there are 3,428 annotated genes that exhibit evidence of AS from this analysis. Wang reports 4,704 annotated genes exhibit AS. However, the genes found by the two analyses are different as shown in Figure 4.1. Between the two analyses, a total of 5,809 genes are found to

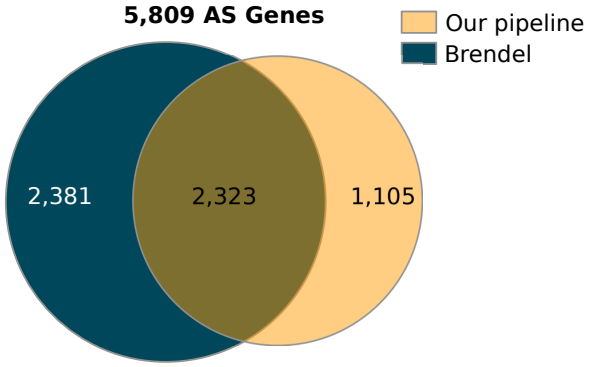


Figure 4.1: Comparison of numbers of AS genes found between this analysis and Wang *et al*.

exhibit AS, constituting 20.2% of the 28,691 known *Arabidopsis* genes according to the TAIR9 annotation. This is slightly lower than the proportion of 21.8% reported by Wang.

In *Chlamydomonas*, the AS event proportions follow that of the results in our previous study. As with *Arabidopsis* and other land plants, IR is the most prevalent form of AS followed by Alt3', Alt5', ES, and lastly AltB. The numbers of genes found to exhibit AS as well as the numbers of events are also similar, which is not surprising considering the same dataset and much of the same software pipeline was used. The only notable differences we observe are reduced event counts for IR and Alt3'. Our previous work found 305 IR and 158 Alt3', differences of 50 and 18 events, respectively. This pipeline has additional filtering steps implemented (*e.g.* filtering of short exon alignments) that might easily explain this disparity.

The distance, or *offset*, between alternative splice sites in an Alt3' or Alt5' event can help to determine the effect an AS event might have on a resultant protein. Since three consecutive bases of mRNA code for an amino acid, an addition of bases due to an AS event that is not divisible by three often drastically changes the codon sequence encoded by an mRNA molecule. These

changes in an mRNA's codon sequence, called a *frameshift*, occur frequently in conjunction with AS events in *Arabidopsis* and other organisms [26]. There is evidence that a high proportion of mRNAs affected by AS are degraded even before translation into protein can occur on account of frameshifts [1, 27]. Figure 4.2 contains offset distributions for Alt3' and Alt5' AS events in both organisms.

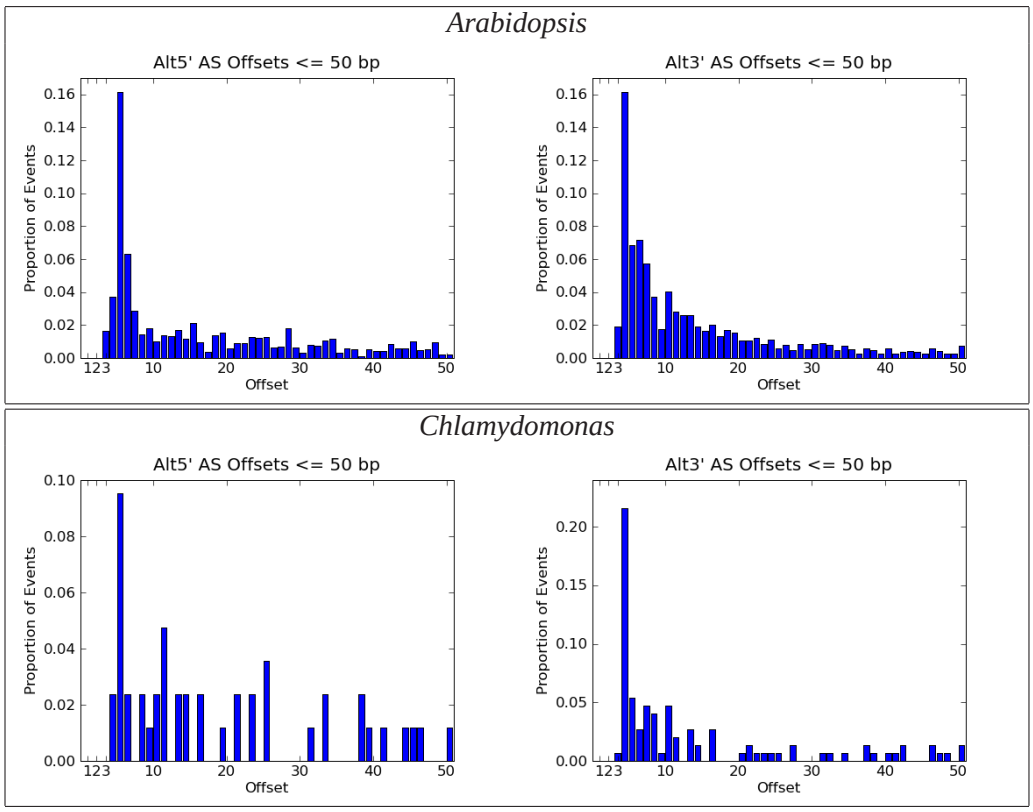


Figure 4.2: Offset distributions for Alt3' and Alt5' AS events for both organisms. Only offsets less than 50 base pairs are shown. There were no offsets smaller than 3 bases and only a small number larger than 50.

The clear peaks in both organisms for Alt5' AS is an offset of 5 base pairs. This is consistent with the findings of previous studies [26] and suggests that a large number of Alt5' AS events result in a frameshift. For Alt3', the clear peak for both organisms is an offset of 4 base pairs, which also supports the result that many Alt3' events alter the codon sequence encoded by mRNA.

## 4.5 Splice Site Strength

The cellular machinery responsible for splicing introns out of RNA does so by recognizing signals in the nucleotide sequence of a gene. This signal most commonly has the canonical dinucleotide pairs of GT/AG or GC/AG at the beginning and end of introns. The bases surrounding these canonical dinucleotides are less conserved but are still of importance to the splicing mechanism. By quanitfying the level of conservation of splice site sequences the relative strength of different types of splice sites can be compared. As shown in our previous work on *Chlamydomonas*, splice sites involved in AS are weaker than those that are not alternatively spliced. It may therefore be possible to use splice site strength as a feature when predicting whether a putative splice junction may exhibit AS.

Sequences corresponding to AS events identified by the pipeline were analyzed for splice site strength following the protocol in [28]. Splice sites for each type of AS event (IR, Alt3', Alt5', ES) as well as splice sites where we found no evidence of AS were used to construct motifs using the TAMO package [29]. Splice sites were considered to be 3 bases of exonic sequence and 10 bases of intronic sequence, resulting in motifs of length 13. Splice site instances were scored according to:

$$\textit{score} = \sum_{j=1}^{N} \log\left(\frac{p_m(s(j), j)}{p_{bg}(s(j), j)}\right), \tag{4.1}$$

where $N$ is the length of the motif, $s(j)$ is the nucleotide at position $j$, $p_m(i, j)$ is the probability of seeing nucleotide $i$ at position $j$ of the motif, and $p_{bg}(i, j)$ is the background distribution for nucleotide $i$ at position $j$ of the background motif. To construct the background motif we used a background based on exon sequences for the exonic part of the motif, and a background based on intronic sequences for the intronic part of the motif. Two sets of motif instances were scored using Eqn. (4.1), and the significance of the difference between the scores was determined using the Wilcoxon Rank Sum test. $z$-scores were calculated using the normal approximation of the test and converted to $p$-values.

The splice site analysis protocol described above was conducted for both *Arabidopsis* and *Chlamydomonas*. As reported for other organisms, the splice sites involved in AS are weaker than those associated with constitutive splicing with high statistical significance. Table 4.5 contains splice site strength statistics.

| | *Arabidopsis* | | | | *Chlamydomonas* | | | |
|---|---|---|---|---|---|---|---|---|
| | **5' site** | | **3' site** | | **5' site** | | **3' site** | |
| Event | Score | *p*-value | Score | *p*-value | Score | *p*-value | Score | *p*-value |
| IR | 5.38 | 4.57e-318 | 5.27 | 1.06e-227 | 7.58 | 2.46e-55 | 7.37 | 5.41e-9 |
| ES | 6.58 | 8.40e-6 | 6.05 | 2.45e-32 | 8.94 | 2.25e-5 | 8.03 | 6.09e-13 |
| Alt3' | 6.24 | 1.26e-11 | 4.81 | 0 | 8.89 | 0.0073 | 6.47 | 1.30e-67 |
| Alt5' | 4.41 | 0 | 5.24 | 1.29e-108 | 6.67 | 1.80e-68 | 7.27 | 4.60e-5 |
| Constitutive | 6.72 | NA | 6.40 | NA | 9.26 | NA | 8.10 | NA |

Table 4.5: Splice site strength statistics

The prevalent and non-prevalent splice site strengths for 3' and Alt5' AS events were identified and tested against constitutive splice sites. The prevalent form of an AS event is the one supported with the largest number of EST alignments. For both *Arabidopsis* and *Chlamydomonas* the non-prevalent splice sites are weaker than the prevalent forms, and the prevalent splice sites are weaker than constitutive forms. Table 4.6 and Figure 4.3 contain the splice site strength statistics and WebLogo images [30], respectively, for this analysis.

| | *Arabidopsis* | | | | |
|---|---|---|---|---|---|
| Event | Non-Prevalent | *p*-value | Prevalent | *p*-value | Constitutive |
| Alt3' | 4.96 | 1.492e-35 | 5.74 | 9.48e-53 | 6.40 |
| Alt5' | 4.62 | 2.42e-34 | 5.81 | 2.34e-22 | 6.72 |
| | *Chlamydomonas* | | | | |
| Event | Non-Prevalent Score | *p*-value | Prevalent Score | *p*-value | Constitutive Score |
| Alt3' | 6.09 | 3.24e-10 | 7.82 | 0.06 | 8.10 |
| Alt5' | 6.31 | 8.10e-09 | 8.38 | 0.25 | 9.26 |

Table 4.6: Prevalent vs. Non-Prevalent vs. Constitutive splice site motif statistics. Scores are the average of all individual splice site instance scores computed against the background described above. The first *p*-value column is the significance of testing the Non-Prevalent motif against the Prevalent motif. The second *p*-value column is the significance between the Prevalent and Constitutive motifs.

From Table 4.6, we notice that the non-prevalent motif scores are lower than both the prevalent and constitutive scores for both organisms. This makes biological sense in that a splice site with a weaker signal is less likely to be spliced than one with a stronger signal, supporting the evidence that these sites are non-prevalent.

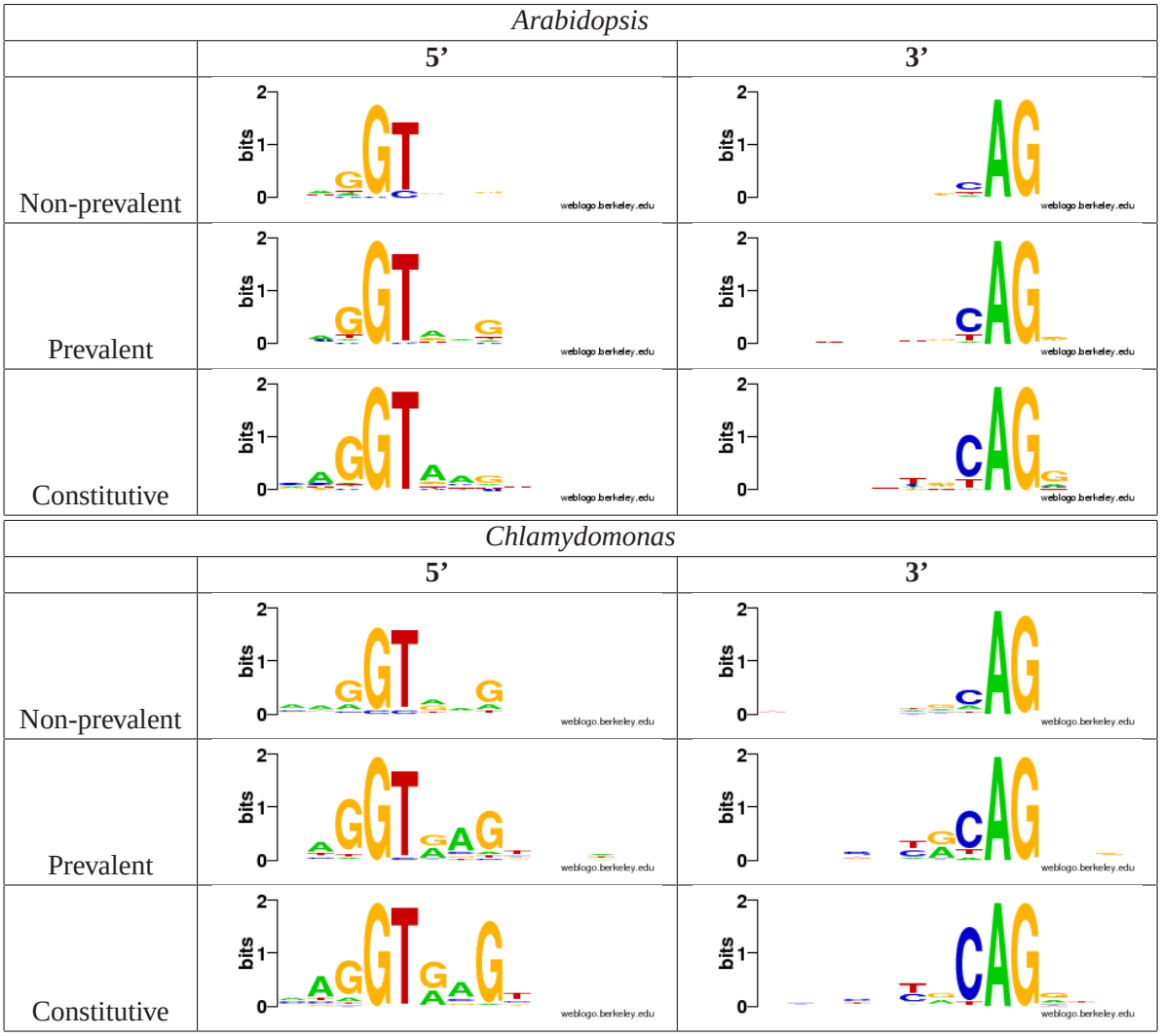


Figure 4.3: WebLogo images [30] of prevalent, non-prevalent, and constitutive splice sites.

## 4.6 Conclusions and Future Work

In this work we presented an computational pipeline for AS analysis. We applied the pipeline to the largest available EST datasets of the model organisms *Arabidopsis* and *Chlamydomonas*. The results generally confirm previous work on both of these organisms, most significantly that IR is the most prevalent form of AS and splice sites involved in AS are weaker than those of constitutive splicing. A significant number of potentially undiscovered genes were also identified in both organisms. The pipeline used to conduct this analysis includes novel editing techniques for improving our confidence in the detected AS events. Our approach of identifying only distinct biological events from a splice graph ensures the number of AS events has clear biological

relevance.

A large set of AS events were discovered in this work. These events can now be used to construct datasets that will allow further analysis of AS. The genetic sequences up- and down-stream of AS events can be extracted and analyzed for signals implicated in AS. The splice site strength analysis can also be used as one of a set of features to discover putative AS sites along the genome.

The pipeline described here has also been applied to a cross-organism AS analysis of serine-rich (SR) genes in collaboration with Dale Richardson of the Department of Bioinformatics and Population Genetics, University of Cologne for his PhD thesis.

# REFERENCES


[1] Sergei A. Filichkin, Henry D. Priest, Scott A. Givan, Rongkun Shen, Douglas W. Bryant, Samuel E. Fox, Weng-Keen Wong, and Todd C. Mockler. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research*, 20(1):45–58, 2010.

[2] Adam Labadorf, Alicia Link, Mark Rogers, Julie Thomas, Anireddy Reddy, and Asa Ben-Hur. Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii. *BMC Genomics*, 11(1):114, 2010.

[3] Douglas L. Black. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annual Review of Biochemistry*, 72(1):291–336, 2003.

[4] National Institute of Health. Deoxyribonucleic Acid Fact Sheet. `http://www.genome.gov/25520880`, 2009. [Online; accessed 14-March-2010].

[5] National Center for Biotechnology Information. EST Fact Sheet. `http://www.ncbi.nlm.nih.gov/About/primer/est.html`, 2004. [Online; accessed 14-March-2010].

[6] Colin A. Graham and Alison J. Hill. *Introduction to DNA Sequencing*, volume 167 of *Methods in Molecular Biology*. January 2001.

[7] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453, 1970.

[8] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.

[9] Uta Schulze, Bettina Hepp, Cheng Soon Ong, and Gunnar Ratsch. PALMA: mRNA to genome alignments using large margin algorithms. *Bioinformatics*, 23(15):1892–1900, 2007.

[10] Volker Brendel, Liqun Xing, and Wei Zhu. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics*, 20(7):1157–1169, 2004.

[11] SR Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[12] S.F. Altschul, W. Gish, W. Miller, E.W. Meyers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410(8), October 1990.

[13] W. James Kent. BLAT: The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, 2002.

[14] Davide Campagna, Alessandro Albiero, Alessandra Bilardi, Elisa Caniato, Claudio Forcato, Svetlin Manavski, Nicola Vitulo, and Giorgio Valle. PASS: a program to align short sequences. *Bioinformatics*, 25(7):967–968, 2009.

[15] Thomas D. Wu and Colin K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21:1859–1875(17), 1 May 2005.

[16] Mikhail S. Gelfand, Andrey A. Mironov, and Pavel A. Pevzner. Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 93(17):9061–9066, 1996.

[17] S. Heber, M. Alekseyev, S Sze, H. Tang, and P. Pevzner. Splicing graphs and EST assembly problem. *Bioinformatics*, 18(Suppl 1):S181–S188.

[18] Kei Iida, Motoaki Seki, Tetsuya Sakurai, Masakazu Satou, Kenji Akiyama, Tetsuro Toyoda, Akihiko Konagaya, and Kazuo Shinozaki. Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. *Nucl. Acids Res.*, 32(17):5096–5103, 2004.

[19] Bing-Bing Wang and Volker Brendel. Genomewide comparative analysis of alternative splicing in plants. 103(18):7175–7180, 2006.

[20] Chlamy Center (Duke University). About Chlamydomonas. `http://www.chlamy.org/info.html`, 2010. [Online; accessed 14-March-2010].

[21] Eoghan D. Harrington and Peer Bork. Sircah: a tool for the detection and visualization of alternative transcripts. *Bioinformatics*, 24(17):1959–1960, 2008.

[22] The Arabidopsis Information Archive. `http://www.arabidopsis.org`, 2010. [Online; accessed 14-March-2010].

[23] Sabeeha S. et al Merchant. The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. *Science*, 318(5848):245–250, 2007.

[24] National Center for Biotechnology Information. dbEST: Expressed Sequence Tag Database, 2010. [Online; accessed 14-March-2010].

[25] C. Liang, Y. Liu, L. Liu, A.C. Davis, Y. Shen, and Q.Q. Li. Expressed sequence tags with cDNA termini: previously overlooked resources for gene annotation and transcriptome exploration in Chlamydomonas reinhardtii. *Genetics*, 179(1):83, 2008.

[26] Matthew Campbell, Brian Haas, John Hamilton, Stephen Mount, and C Robin Buell. Comprehensive analysis of alternative splicing in rice and comparative analyses with arabidopsis. *BMC Genomics*, 7(1):327, 2006.

[27] Benjamin P. Lewis, Richard E. Green, and Steven E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 100(1):189–192, 2003.

[28] Christina L. Zheng, Xiang-Dong Fu, and Michael Gribskov. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA*, 11(12):1777–1787, 2005.

[29] D. Benjamin Gordon, Lena Nekludova, Scott McCallum, and Ernest Fraenkel. Tamo: a flexible, object-oriented framework for analyzing transcriptional regulation using dna-sequence motifs. *Bioinformatics*, 21(14):3164–3165, 2005.

[30] Gavin E. Crooks, Gary Hon, John-Marc Chandonia, and Steven E. Brenner. WebLogo: A Sequence Logo Generator. *Genome Research*, 14(6):1188–1190, 2004.

# Glossary

**alternative splicing** the phenomenon where a gene has more than one splice form, 3

**amino acids** the 22 molecules that are the building blocks of proteins, 3

**base complementarity** the matching of A ↔ T and C ↔ G in DNA, 2

**bases** see **nucleotide**, 2

**canonical splice site** a splice site marked by the first two and last two intronic bases having GT/AG or GC/AG, 3

**codon** three consecutive mRNA bases that are translated into an amino acid, 3

**exon** the coding portions of a gene that are translated into protein, 3

**Expressed Sequence Tag (EST)** a sequence originating from an expressed gene, 3

**gene** region of a genome encodes information about how a cell functions, 2

**indel** any insertion, deletion, or mutation of a base in a sequence, 11

**intron** a non-coding sequence of a gene spliced out of pre-mRNA before protein translation, 3

**messenger RNA** RNA that has had introns spliced out of it, 3

**mRNA** shorthand for messenger RNA, 3

**nucleotide** one of four DNA molecules adenine (A), cytosine (C), guanine (G), or thymine (T), 2

**percent identity** synonym for sequence identity, 6

**pre-mRNA** shorthand for precursor messenger RNA, 2

**precursor messenger RNA** RNA that has been transcribed from a gene but still contains introns, 2

**sequence alignment** the task of finding the most compatible subsequence between two sequences of characters, 6

**sequence identity** the proportion of bases in an alignment that match between two sequences out of all paired bases, 6

**splice form** a pattern of exons and introns of a gene, 3

**splice graph** graph constructed out of a set of alignments that compactly represents splicing information, 7

**splice junction** synonym for splice site, 3

**splice site** location within a gene defining the boundary between an exon and an intron, 3

**spliced alignment** sequence alignment strategy that allows long consecutive gaps in one of the sequences, 6

**splicing** the process of excising introns from pre-mRNA, 3

**transcription** the process of copying a gene's nucleotides to pre-mRNA, 2

**translation** the process of creating a protein out of mRNA, 3