

**BIOINFORMATICS ANALYSIS OF ALTERNATIVE SPLICE
VARIANTS**

LEE TECK KWONG BERNETT

(B. Sc. (Hons.), NUS)

A thesis submitted
for the Degree of Doctoral of Philosophy
Department of Biochemistry
National University of Singapore
2005

Acknowledgement

I would like to extend my heartfelt appreciation and gratitude to my supervisors Professor Shoba Ranganathan, Associate Professor Tan Tin Wee and Associate Professor Too Heng Phon for their support throughout my candidature. I would also like to thank my fellow PhD candidates Kong Lesheng, Vivek Gopalan, Justin Choo, Paul Tan and Victor Tong for their encouragement and wonderful discussions, Mark de Silva and Lim Kuan Siong for systems support and Madeleine Koh for administrative assistance. Thanks also go out to the people in the Department of Biochemistry and the Dean's office of the Faculty of Medicine for their assistance in administrative matters. Lastly, I would also like to thank the Agency for Science, Technology and Research for the scholarship provided to the Bioinformatics Centre, NUS that enabled this work.

Table of Contents

Acknowledgement.....	2
Table of Contents.....	3
Summary	7
List of Tables.....	9
List of Figures	13
Chapter 1: Introduction and literature survey	20
Chapter 1: Introduction and literature survey	20
1.1 RNA splicing.....	20
1.2 Exon and intron definition model	27
1.3 Alternative splicing	27
1.3.1 Splice site motifs.....	27
1.3.2 Splicing enhancers and suppressors	28
1.3.4 Interaction with the transcriptional machinery	29
1.4 Impact of alternative splicing	30
1.4.1 Protein diversity	30
1.4.2 Relevance to pharmacogenomics.....	31
1.5 Bioinformatics resources for alternative splicing.....	32
1.5.1 General bioinformatics databases used in the study of alternative splicing.....	32
1.5.2 Alternative splicing databases	44
1.5.3 Bioinformatics applications used in the study of alternative splicing ...	49
1.5.4 Computational methods for detection of alternative splicing	55
1.5.5 Objectives.....	64

Chapter 2: Data representation and visualization of alternative splicing	67
2.1 Introduction.....	67
2.2 Implementation.....	76
2.2.1 Construction of splicing graph.....	76
2.2.2 Detection and classification of alternative splicing events in the splicing graphs.....	80
2.2.3 Splicing Graph Module (SGM)	81
2.2.4 SGM web service.....	89
2.3 Conclusion.....	97
Chapter 3: Drosophila melanogaster Exon Database (DEDB).....	98
3.1 Introduction.....	98
3.2 Implementation.....	99
3.2.1 Splicing graph source data	99
3.2.2 Splicing graphs construction	101
3.2.3 Domain searches using Pfam	104
3.2.4 Types of alternative splicing events used in the classification	111
3.2.5 Rules used for the classification of alternative splicing events in DEDB	111
3.2.6 Classification of alternative splicing events in DEDB	123
3.2.7 Web interface.....	126
3.3 Comparisons with other splicing graph databases/services	136
3.4 Conclusion.....	140
Chapter 4: Genome-wide analysis of alternative splicing in <i>Drosophila melanogaster</i>	141
4.1 Introduction.....	141

4.2 Material and Methods	143
4.2.1 General statistics	143
4.2.2 Exon and intron length analysis	145
4.2.3 Exon number analysis.....	147
4.2.4 Nucleotide composition analysis.....	147
4.2.5 Splicing motif analysis.....	148
4.2.6 Domain boundary analysis.....	150
4.2.7 Number of splicing graphs partitioned using GO terms.....	152
4.2.8 Effects of alternative splicing on the coding sequence	152
4.2.9 Effects of alternative splicing on the domains	153
4.3 Results and Discussion	153
4.3.1 General statistics	153
4.3.2 Exon and intron length analysis	161
4.3.3 Exon number analysis.....	181
4.3.4 Nucleotide composition analysis.....	185
4.3.5 Splicing motif analysis.....	188
4.3.6 Domain boundary analysis.....	211
4.3.7 Number of splicing graphs partitioned using GO terms.....	217
4.3.8 Effects of alternative splicing on the coding sequence	228
4.3.9 Effects of alternative splicing on the domains	231
4.4 Conclusion.....	234
Overall conclusion.....	237
Future directions	242
Bibliography	244
Appendix A: Alternative splicing bioinformatics resources	249

Appendix B: Oral and poster presentations	250
Oral presentations	250
Poster presentations	250
Appendix C: Publications	251

Summary

The phenomenon of alternative splicing has in recent years been transformed from one where it is thought to occur fortuitously to one where it is known to be tightly controlled and essential to the functional of the cell. Advances in sequencing technologies and bioinformatics tools have enabled the generation and annotation of a large amount of genomic and transcriptional sequence information that could be used as a basis for the study of alternative splicing. The aim of this thesis seeks to use this information to study the characteristics and effects of alternative splicing in a well-studied organism such as *Drosophila melanogaster*.

At the outset, the phenomenon of alternative splicing is introduced, along with a literature survey in Chapter 1. In addition to this, background information is also presented on the bioinformatics resources and methods employed in the study of alternative splicing.

Chapter 2 illustrates the work done on the concept of splicing. Splicing graphs are an excellent way to represent the various splice isoforms of a gene. It allows one to capture all the information inherited in the various splice isoforms in a manner that is compact and amiable for the analysis of alternative splicing. Codes were developed that allowed the construction of splicing graphs from mRNA sequence information. These codes have also been provided were as a free web service that allows users to generate their own splicing graphs.

The splicing graph methodology developed in Chapter 2 serves as the backbone for the genome-wide analysis of alternative splicing, leading to the development of a database of *Drosophila melanogaster* exons entitled DEDB

(*Drosophila melanogaster* Exon Database) (Chapter 3). Splicing graphs were extended via the creation of rules that allow the detection and classification of alternative splicing events. These rules allow the detection of various types of alternative splicing events within a single gene. The ability to detect multiple forms of alternative splicing within a single gene model offers finer granularity of classification than currently available. This classification is put to effective use in data analysis based on individual types of alternative splicing events. The DEDB database thus serves as the basis for further analysis in Chapter 4.

Chapter 4 presents the genome-wide analysis done on the alternative splicing events for *Drosophila melanogaster*. Analysis was focused on two aspects, characterization of alternative splicing and the effects of alternative splicing. Analysis on the characteristics of alternative splicing revealed that alternatively spliced genes exhibits variation in the exon lengths, intron lengths, exon number per gene and splice site information content as compared to constitutively spliced genes. In addition to uncovering the characteristics of alternatively spliced genes, analysis done on the effects of alternative splicing also bore fruit. Alternative splicing occurs more frequently in genes that are involved in signal transduction thereby providing speculations that the various splice isoforms are implicated in gene regulation. Surprisingly alternative splicing appears to have a tendency to affect the non-critical regions of the proteins thereby providing more of a modulating or regulatory effect than that of a gain/loss of function effect.

List of Tables

Table 1. GenBank divisions and their description.	35
Table 2. Breakdown of the number of records in UniGene by organism. Statistics as of 21 st August 2004.	42
Table 3. Statistics of AltExtron as of 20th August 2004.	46
Table 4. List of programs and its short description in the HMMER package.	54
Table 5. Text files generated by gameXMLParser.py. The various fields in the text files are provided together with the description of the field. The scaffolds.txt file contains information about the scaffolds (large genomic fragments of the chromosomes). The annotations.txt file contains information about genome annotations, each genome annotation being located on a scaffold (hence the link to scaffolds). A single genome annotation contains information about a single gene. The genome annotation contains GO information (go.txt), FlyBase links (fbg.txt and fba.txt) and synonyms (synonyms.txt). Due to alternative splicing, each gene can potentially be composed of several transcripts, each being represented by a single entry in the file sequences.txt. A single transcripts also contains at least one exon (exons.txt) and possibly introns (introns.txt).	100
Table 6. Text files generated by clusterGenes.py. The various fields in the text files are provided together with the description of the field.	105
Table 7. Contents of text files generated by Python script protein.py.	106
Table 8. Contents of the text file intron_phase.txt generated by the script intronPhase.py.	106
Table 9. Contents of the text files generated by the script protein2.py.	106
Table 10. Contents of the text files generated by the script nodeConnectionCalculations.py.	107
Table 11. The text files generated by the script parseHmmPfam.py.	108
Table 12. Text files generated by the script hmm_positions.py.	109
Table 13. Text files generated by the script parsePfam.py.	110
Table 14. Text file generated by the script parseInterpro.py.	110
Table 15. MySQL tables used to store the information generated by the script getAlternativeSplicing.py.	124
Table 16. Table showing the various types of alternative splicing events and its associated number of events and splicing graphs.	125

Table 17. Types of exons and introns length analyzed.	146
Table 18. Statistics generated by the script totalNumbers.py.	157
Table 19. Number of splicing graphs containing specific number of transcripts.	158
Table 20. Table showing the number of splicing variant and the corresponding number of potential and actual splicing graph. The number of potential splicing graph contains the number of possible paths through the graph while the number of actual splicing graph contains the number of actual transcripts known.	160
Table 21. Basic statistical measures of the exons and introns lengths. The number of exons and introns involved is listed together with the minimum, 1 st quartile, median, mean, 3 rd quartile, maximum and inter-quartile range. ...	165
Table 22. Basic statistical measures for the number of exons for alternatively spliced genes and constitutively spliced genes.....	182
Table 23. Table showing the nucleotide composition of various types of exons and introns. The overall nucleotide composition as well as the nucleotide composition broken down by length into four quartiles are provided.	187
Table 24. Percentage of the various types of splicing motifs found in DEDB....	189
Table 25. Information content of multi constitutive internal exon acceptor and donor sites. The number of sites, the information content for all the sites as well as the information content of individual quartiles based on exon length are displayed.	192
Table 26. Table showing the basic statistical measures of the individual information content of various splice site motifs. The number of splice site motifs involved (No.), the minimum (Min.), the first quartile (1 st qu.), the median (Median), the mean (Mean), the third quartile (3 rd qu.), the maximum (Max.) and the inter-quartile range (IQR) are shown.	196
Table 27. Chi Square Goodness-of-Fit test results for domain boundary analysis. The Type column indicates whether the numbers represent introns lying within or outside the amino acid position. The Exp column shows the expected number of introns while the Obs column shows the actual observed number of introns. The Chi column shows the Chi Square value.	213
Table 28. Results for the Chi Square Goodness-of-Fit test for symmetrical introns. The p-value is 1.176×10^{-4} which indicates that there is a tendency for introns flanking exons to have symmetrical phases.....	216
Table 29. Molecular function GO terms and the corresponding number of splicing graphs. All counts provide an absolute number of all splicing graphs that contain the GO term. Con. counts is the number of constitutive splicing graphs that contain the GO term. Alt. counts is the number of alternative splicing graphs that contain the GO term. All per. is the percentage of	

splicing graph containing this GO term in relation to all splicing graphs having molecular function GO terms. Con. per. is the percentage of constitutive splicing graphs having this GO term in relation to all splicing graphs having molecular function GO terms. Alt. per. is the percentage of alternative splicing graphs having this GO term in relation to all splicing graphs having molecular function GO terms. The ratio of Con. per. over Alt. per. is shown in the ratio column. A value greater than 1 is indicative of more a higher amount of constitutive splicing graph over alternative splicing graph. Rows having ratios higher than one are in yellow and rows having ratios less than one are in blue. The indentation of the GO terms indicates its relative position in the GO hierarchy. 220

Table 30. Biological process GO terms and the corresponding number of splicing graphs. All counts provide an absolute number of all splicing graphs that contain the GO term. Con. counts is the number of constitutive splicing graphs that contain the GO term. Alt. counts is the number of alternative splicing graphs that contain the GO term. All per. is the percentage of splicing graph containing this GO term in relation to all splicing graphs having biological process GO terms. Con. per. is the percentage of constitutive splicing graphs having this GO term in relation to all splicing graphs having biological process GO terms. Alt. per. is the percentage of alternative splicing graphs having this GO term in relation to all splicing graphs having biological process GO terms. The ratio of Con. per. over Alt. per. is shown in the ratio column. A value greater than 1 is indicative of more a higher amount of constitutive splicing graph over alternative splicing graph. Rows having ratios higher than one are in yellow and rows having ratios less than one are in blue. The indentation of the GO terms indicates its relative position in the GO hierarchy. 223

Table 31. Cellular process GO terms and the corresponding number of splicing graphs. All counts provide an absolute number of all splicing graphs that contain the GO term. Con. counts is the number of constitutive splicing graphs that contain the GO term. Alt. counts is the number of alternative splicing graphs that contain the GO term. All per. is the percentage of splicing graph containing this GO term in relation to all splicing graphs having cellular process GO terms. Con. per. is the percentage of constitutive splicing graphs having this GO term in relation to all splicing graphs having cellular process GO terms. Alt. per. is the percentage of alternative splicing graphs having this GO term in relation to all splicing graphs having cellular process GO terms. The ratio of Con. per. over Alt. per. is shown in the ratio column. A value greater than 1 is indicative of more a higher amount of constitutive splicing graph over alternative splicing graph. Rows having ratios higher than one are in yellow and rows having ratios less than one are in blue. The indentation of the GO terms indicates its relative position in the GO hierarchy. 227

Table 32. Chi Square Goodness-of-Fit test for all forms of alternative splicing. The P-value is 0. 229

Table 33. Chi Square Goodness-of-Fit test for alternative acceptor sites. The P-value is 5.015×10^{-4} 229

Table 34. Chi Square Goodness-of-Fit test for alternative donor sites. The P-value is 2.761×10^{-82} .	229
Table 35. Chi Square Goodness-of-Fit test for alternative initiation exons. The P-value is 0.	229
Table 36. Chi Square Goodness-of-Fit test for alternative termination exons. The P-value is 4.603×10^{-2} .	230
Table 37. Chi Square Goodness-of-Fit test for cassette exon. The P-value is 2.637×10^{-7} .	230
Table 38. Chi Square Goodness-of-Fit test for intron retentions. The P-value is 1.013×10^{-72} .	230
Table 39. Chi Square Goodness-of-Fit test for all forms of alternative splicing. The P-value is 1.509×10^{-48} .	232
Table 40. Chi Square Goodness-of-Fit test for alternative acceptor sites. The P-value is 7.260×10^{-4} .	232
Table 41. Chi Square Goodness-of-Fit test for alternative donor sites. The P-value is 2.524×10^{-5} .	232
Table 42. Chi Square Goodness-of-Fit test for alternative initiation exons. The P-value is 2.836×10^{-21} .	232
Table 43. Chi Square Goodness-of-Fit test for alternative termination exons. The P-value is 2.767×10^{-10} .	233
Table 44. Chi Square Goodness-of-Fit test for cassette exons. The P-value is 7.646×10^{-9} .	233
Table 45. Chi Square Goodness-of-Fit test for intron retention. The P-value is 1.683×10^{-7} .	233

List of Figures

- Figure 1. The splicing process of group I introns. 22
- Figure 2. The splicing process of group II introns. 23
- Figure 3. The splicing process of nuclear spliceosomal introns. 26
- Figure 4. A sample of a GenBank record. Each line contains a number of fields which are column delimited..... 35
- Figure 5. Growth of GenBank in terms of the number of base pairs and records. Data shown as of 16 August 2004. 36
- Figure 6. An example of a Swiss-Prot record..... 39
- Figure 7. Schematic diagram of the transcripts produced by a constitutive gene. Experimentally sequence transcript sequences will align perfectly barring any sequencing errors with each other in regions of overlap..... 59
- Figure 8. Schematic diagram of the transcripts produced by an alternatively spliced gene, which in this case exhibits a cassette exon event (exon 2 is omitted in splicing variant 1). Transcripts from splicing variant 1 will not align perfectly with transcripts from splicing variant 2..... 60
- Figure 9. Schematic diagram of the effects of clustering of transcript sequences from splicing variant 1 and splicing variant 2 shown in Figure 8. Depending on which is the reference sequence, a deletion or insertion will occur in the alignment. 61
- Figure 10. Two dimension array representation of a set of exons that make up a transcript. A transcript can be simply represented as a series of exons, each exon having a start and end position. The first dimension in the array has two elements, the start and end position respectively while the second dimension represents the number of exons that make up the transcript. 70
- Figure 11. Alternative splicing information represented by a set of transcript. Splice variants can be individually represented as two dimension arrays. A set of these arrays can then be used to determine the type of alternative splicing present. Splice variant has an addition exon having the start and end position of 900 and 1020 respectively. This is missing in splice variant 2. Therefore the conclusion is that the exon starting at position 900 to 1020 is a cassette exon..... 70
- Figure 12. Visual representation of transcripts that contain alternative splicing events. (A) Multiple sequence alignment visual representation provides sequence information that is otherwise unavailable in the schematic view. Two transcripts are shown with splice variant 2 missing exon 2. This is clearly a case of a cassette exon. (B) The same information in the multiple sequence alignment represented as a schematic visual representation.

Sequence information is lost but better overall view of the forms of alternative splicing is possible due to a lack of distraction by the sequence information. (C) When the number of transcripts available increases, it becomes more difficult to determine the overall view of the types of alternative splicing. The visual representation here also shows a case of a cassette exon. However, the number of different transcripts that constitutes this makes the determination more difficult..... 71

Figure 13. Splicing graph representation of alternative splicing events. The numerous splice variants are used in the graph construction process to produce a splicing graph that is a condensed view of all the splice variants. The splicing graph is clearly easier to interpret as any bifurcation in the graph is suggestive of alternative splicing..... 74

Figure 14. Our implementation of the visual splicing graph representation. The representation proposed by Heber *et al.* is shown on the left with a blue background while our implementation is shown on the right with a yellow background. The main difference is that each unique exon is represented making the alternative splicing events more visible. 75

Figure 15. Construction of splicing graphs. Initial groups of transcripts are found by comparing the start and end positions of the transcripts, this resulted in two groups in this figure. Overlapping exons (in green) and shared exons (in red) are defined and used to perform the final grouping. Once the final grouping is completed, overlapping and shared exons are collapsed into a single unit forming the final splicing graph. 79

Figure 16. UML class diagram for the classes in the file Genome.py. 85

Figure 17. UML class diagram for classes in the file DrawGenome.py. 86

Figure 18. UML class diagrams for the data representation related classes in the file Graph.py. 87

Figure 19. UML class diagrams for the visualization related classes in the file Graph.py..... 88

Figure 20. SGM web service homepage..... 92

Figure 21. SGM web service input form. The only required input is a GFF file that describes the transcripts required for the generation of splicing graphs..... 93

Figure 22. The results of the GFF file submission. A list of all the splicing graphs generated from the transcript is shown (the list is partly reproduced here due to its length). Each splicing graph entry shows the identifiers of the transcripts used to generate the splicing graph, the number of nodes (vertices) in the splicing graph as well as the number of paths in the splicing graph. Any number of paths greater than 1 indicates that there is alternative splicing. By clicking on the link provided for each entry, one can get a detailed look at the splicing graph..... 94

Figure 23. Details of the resulting splicing graph generated by the SGM web service. Links are provided that allows users to download graphic files of the splicing graph. Details of the nodes and the transcripts are also shown.....	95
Figure 24. An example of the splicing graph visual representation produced by SGM. The splicing graph is shown at the top with each exon as a black colored bar labeled with a number. Introns are shown as green colored lines connecting the exons. Immediately below the splicing graph are the visual representation of the transcripts used to construct the splicing graph. Each exon like in the splicing graph visual representation is shown as a black bar connected to each other by introns as green colored lines. The identifier of each transcript is shown just on top of each transcript visual representation.	96
Figure 28. Rules used in the classification of the alternative splicing events in DEDB.....	112
Figure 29. Alternative transcriptional start sites.	115
Figure 30. Alternative transcriptional termination sites.....	116
Figure 31. Alternative initiation exons.	117
Figure 32. Alternative termination exons.....	118
Figure 33. Alternative acceptor sites.....	119
Figure 34. Alternative donor sites.	120
Figure 35. Cassette exons.	121
Figure 36. Intron retention.....	122
Figure 37. Screenshot of the splicing graph viewer in DEDB. The splicing graph viewer consists of three frames. The top frame shows the navigation aid that allows for rapid navigation between splicing graphs. The middle frame shows graphical representations of the splicing graph together with the transcripts that were used to construct the splicing graph. The bottom frame provides a place for detailed textual information.	127
Figure 38. Top navigational frame. The first button on the menu links the user back to the query page allowing them to search for specific splicing graphs fulfilling certain criteria. The next two buttons allow the user to navigate to the next and previous splicing graph while the next two buttons transverse through all the splicing graphs exhibiting alternative splicing events. There is also a drop down selector (Figure 36) that allows users to jump to splicing graphs showing specific type of alternative splicing event.....	131
Figure 39. Navigational drop down selector. This drop down selector allows users to select for splicing graphs having specific types of alternative splicing. ...	131

- Figure 40. Schematic diagram of the splicing graph shown in the middle frame. Nodes corresponding to exons are shown as black bars connected via green lines representing connections (which are introns). The Node ID is displayed as white numbers on each of the nodes. The start and stop codons are shown respectively as green and red lines on the nodes. Users can click on the nodes to get more information in the detailed information area. 132
- Figure 41. Schematic representation of the transcripts that make up the splicing graph. Each transcript is shown as a series of black bars (exons) connected together via green lines (introns). The graphical elements representing the exons and introns can be clicked to show more detailed information in the detailed information area. The FlyBase Gene ID for each transcript is shown at the top of each transcript. The start and stop codons for each transcript is shown respectively as a green and red line. The intron phase of each intron is shown as a number in the middle of the green line. Colored bars if present below the transcripts represents Pfam domains that has been detected on the transcripts. Users can click on the domains to display more information in the detailed information area. 132
- Figure 42. Screenshot of DEDB query page. Users can query DEDB via a BLAST search, by FlyBase Gene Name, FlyBase Gene Symbol, Pfam Accession Number and Pfam Identifier. 133
- Figure 43. Screenshot of Alternative splicing event classification page in DEDB. 134
- Figure 44. Screenshot of a page listing the splicing graphs that contain a particular type of alternative splicing event. Each item in the list is a link to the Splicing Graph Viewer. 135
- Figure 42. Global View of human dscam gene using MAASE. Deletions in some isoforms are depicted by the dark yellow lines in a manner similar to the original splicing graph visualization. The global view shows the introns with a size that is relative to that of the exons which leads to very long introns. This differs from SGM where the intron size is visually reduce to depict more of the exons. This visual representation makes it difficult to see that there is actually two cassette exons. The alternative initiation exons appear to be quite oblivious. 137
- Figure 43. Exon Region Alignment view of MAASE. This is quite similar to the transcript view produced by SGM. 138
- Figure 44. Splicing graph produced by ASG. The splicing graph is shown at the top and is very similar to the original splicing graph visual representation. The transcripts supporting the splicing graph are depicted below. An advantage that ASG has over SGM is that alternative splicing events are marked in colors in the splicing graph. 139
- Figure 45. Schematic representation of the amino acid positions used in the domain boundary analysis. Each colored block represents a single amino acid residue. Amino acid residues in purple are part of the domain while

residues in pink are outside the domain. Introns lying exactly at the boundary of the domains are designed as being in position 0. Introns lying in the amino acids within the domains are in increasing negative integers beginning from the domain boundary. Likewise, introns lying in the amino acids outside the domains are in increasing positive integers beginning from the domain boundary. A range of 10 amino acids flanking the domain boundary was used for the analysis..... 151

Figure 46. Histogram of single constitutive exons. Histogram is limited to exons of length 5000 nucleotides and below..... 166

Figure 47. Histogram of multiple constitutive exons. Histogram is limited to exons of length 5000 nucleotides and below..... 167

Figure 48. Histogram of multiple constitutive initiation exons. Histogram is limited to exons of length 5000 nucleotides and below. 168

Figure 49. Histogram of multiple constitutive internal exons. Histogram is limited to exons of length 5000 nucleotides and below. 169

Figure 50. Histogram of multiple constitutive termination exons. Histogram is limited to exons of length 5000 nucleotides and below..... 170

Figure 51. Histogram of constitutive introns. Histogram is limited to introns of length 1000 nucleotides and below..... 171

Figure 52. Histogram of cassette exons. Histogram is limited to exons of length 5000 nucleotides and below. 172

Figure 53. Histogram of lengths of upstream introns of cassette exons. Histogram is limited to exons of length 5000 nucleotides and below. 173

Figure 54. Histogram of lengths of downstream introns of cassette exons. Histogram is limited to exons of length 5000 nucleotides and below. 174

Figure 55. Histogram of lengths of intron retention introns. Histogram is limited to exons of length 1000 nucleotides and below. 175

Figure 56. Histogram of lengths of upstream exons of intron retention. Histogram is limited to exons of length 1000 nucleotides and below. 176

Figure 57. Histogram of lengths of downstream exons of intron retention. Histogram is limited to exons of length 1000 nucleotides and below. 177

Figure 58. Histogram of lengths of intron retention containing exons. Histogram is limited to exons of length 5000 nucleotides and below..... 178

Figure 59. Histogram of lengths of alternative donor exons. Histogram is limited to exons of length 5000 nucleotides and below. 179

Figure 60. Histogram of lengths of alternative acceptor exons. Histogram is limited to exons of length 5000 nucleotides and below..... 180

Figure 61. Histogram of the number of exons of constitutively spliced transcripts. The histogram is limited to transcripts having 30 or less exons.....	183
Figure 62. Histogram of the number of exons for alternatively spliced transcripts. The histogram is limited to transcripts having 30 or less exons.....	184
Figure 63. Sequence logo of multi constitutive internal exon acceptor sites. The region -30 to +2 is shown. The information content is 9.745 bits.	193
Figure 64. Sequence logo of multi constitutive internal exon acceptor sites. Only the exons in the exon length first quartile are used. The region -30 to +2 is shown. The information content is 9.814 bits.....	193
Figure 65. Sequence logo of multi constitutive internal exon acceptor sites. Only the exons in the exon length second quartile are used. The region -30 to +2 is shown. The information content is 9.544 bits.	193
Figure 66. Sequence logo of multi constitutive internal exon acceptor sites. Only the exons in the exon length third quartile are used. The region -30 to +2 is shown. The information content is 9.571 bits.....	193
Figure 67. Sequence logo of multi constitutive internal exon acceptor sites. Only the exons in the exon length fourth quartile are used. The region -30 to +2 is shown. The information content is 10.090 bits.....	194
Figure 68. Sequence logo of multi constitutive internal exon donor sites. The region +5 to -10 is shown. The information content is 8.525 bits.	194
Figure 69. Sequence logo of multi constitutive internal exon donor sites. Only the exons in the exon length first quartile are used. The region +5 to -10 is shown. The information content is 8.644 bits.....	194
Figure 70. Sequence logo of multi constitutive internal exon donor sites. Only the exons in the exon length second quartile are used. The region +5 to -10 is shown. The information content is 8.364 bits.....	194
Figure 71. Sequence logo of multi constitutive internal exon donor sites. Only the exons in the exon length third quartile are used. The region +5 to -10 is shown. The information content is 8.396 bits.....	195
Figure 72. Sequence logo of multi constitutive internal exon donor sites. Only the exons in the exon length fourth quartile are used. The region +5 to -10 is shown. The information content is 8.728 bits.....	195
Figure 73. Histogram of the individual information content of multi constitutive internal exon acceptor splice motif.....	197
Figure 74. Histogram of the individual information content of multi constitutive internal exon donor splice motif.	198
Figure 75. Histogram of the individual information content of cassette exon acceptor splice motif.	200

Figure 76. Histogram of the individual information content of cassette exon donor splice motif.....	201
Figure 77. Histogram of the individual information content of intron retention acceptor splice motif.....	203
Figure 78. Histogram of the individual information content of intron retention donor splice motif.....	204
Figure 79. Histogram of the individual information content of alternative acceptor splice motif.....	206
Figure 80. Histogram of the individual information content of intron retention acceptor splice motif.....	207
Figure 81. Histogram of the individual information content of multi constitutive initiation exon donor splice motif.....	209
Figure 82. Histogram of the individual information content of multi constitutive termination exon acceptor splice motif.....	210
Figure 83. Plot of $\log(1/p\text{-value})$ against amino acid positions. The plot shows that there is a very high statistical significance in the number of introns lying at amino acid positions 0 to + 2.	214

Chapter 1: Introduction and literature survey

1.1 RNA splicing

The discovery of RNA splicing was initiated independently in 1977 by Richard J. Roberts and Phillip A. Sharp (Sharp, 1994) where they observed that genes were separated by non-coding DNA. This discovery was first conceived when biochemistry assays indicated that one end of the adenovirus mRNA did not conform to the then current notion of a single linear piece of coding mRNA as found in bacteria. Using electron microscopy, they observed three RNA loops when the mRNA was hybridized to the original DNA. This led to the conclusion that the mRNA was composed of discontinuous segments of coding DNA/RNA termed exons interrupted by non-coding DNA/RNA termed introns and not a single linear piece as that observed in the bacteria. Since the mRNA is transcribed from the DNA, it too contains coding and non-coding segments and necessitates the need for a process that removes the non-coding segments to produce a single linear coding strand suitable for translation. This process is now known to be RNA splicing where the introns are removed to retain the exons allowing for translation. The regions where the introns meet the exons are named as the 5' splice site (or donor site) and 3' splice site (or acceptor site).

Today, four different types of introns namely, group I introns, group II introns, nuclear tRNA introns and nuclear spliceosomal introns are known. Group I introns were first discovered in the protozoan *Tetrahymena thermophila* where the pre-rRNA contains a group I introns that is self-splicing (Cech, 1990). The splicing process (see Figure 1) of group I introns like group II and nuclear spliceosomal introns, involves two transesterification reactions resulting in the

removal of the introns and the joining of the adjacent exons. A guanosine co-factor bounded to a specific active site then takes part in the first transesterification reaction resulting in the release of the 5' exon. The 3' hydroxyl group of the 5' exon then participates in the second transesterification reaction that results in the ligation of the 5' exon to the 3' exon and the release of the introns. The splicing process requires no protein components and demonstrates the capabilities of RNA to be catalytic (one evidence pointing to a RNA world).

Group II introns (Lehmann & Schmidt, 2003) are also spliced out using two transesterification processes (Figure 2). Unlike group I introns, which require the use of a guanosine co-factor, a 2'-hydroxyl group on an adenosine residue is used in the first transesterification process. The subsequent process is similar to that of group I introns where a second transesterification process is carried out involving the 3'-hydroxyl group of the 5' exon and a phosphate group at the acceptor site. This leads to the joining of the two neighboring exons and the release of a lariat intron. Like group I introns, group II introns have been demonstrated to be self-splicing.

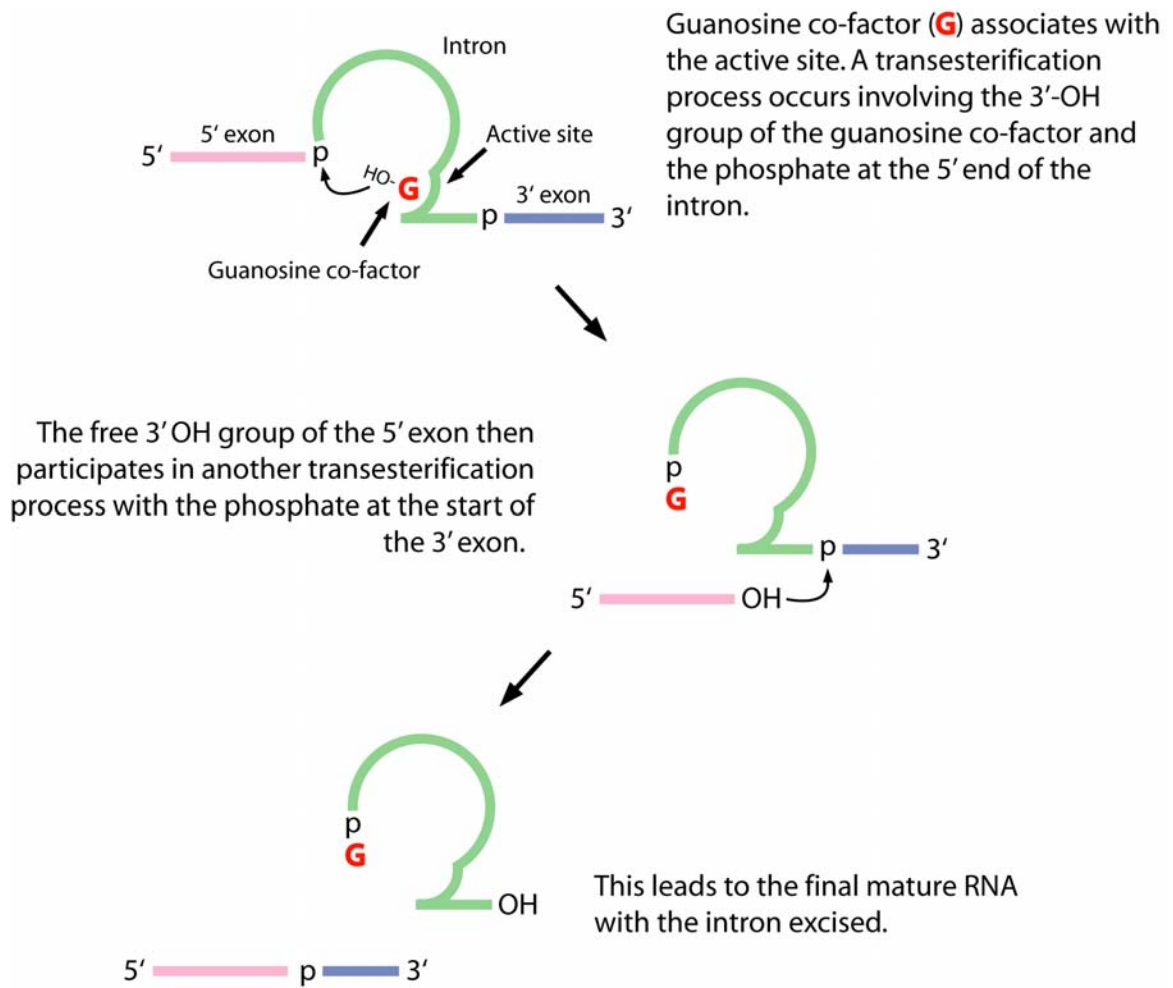


Figure 1. The splicing process of group I introns.

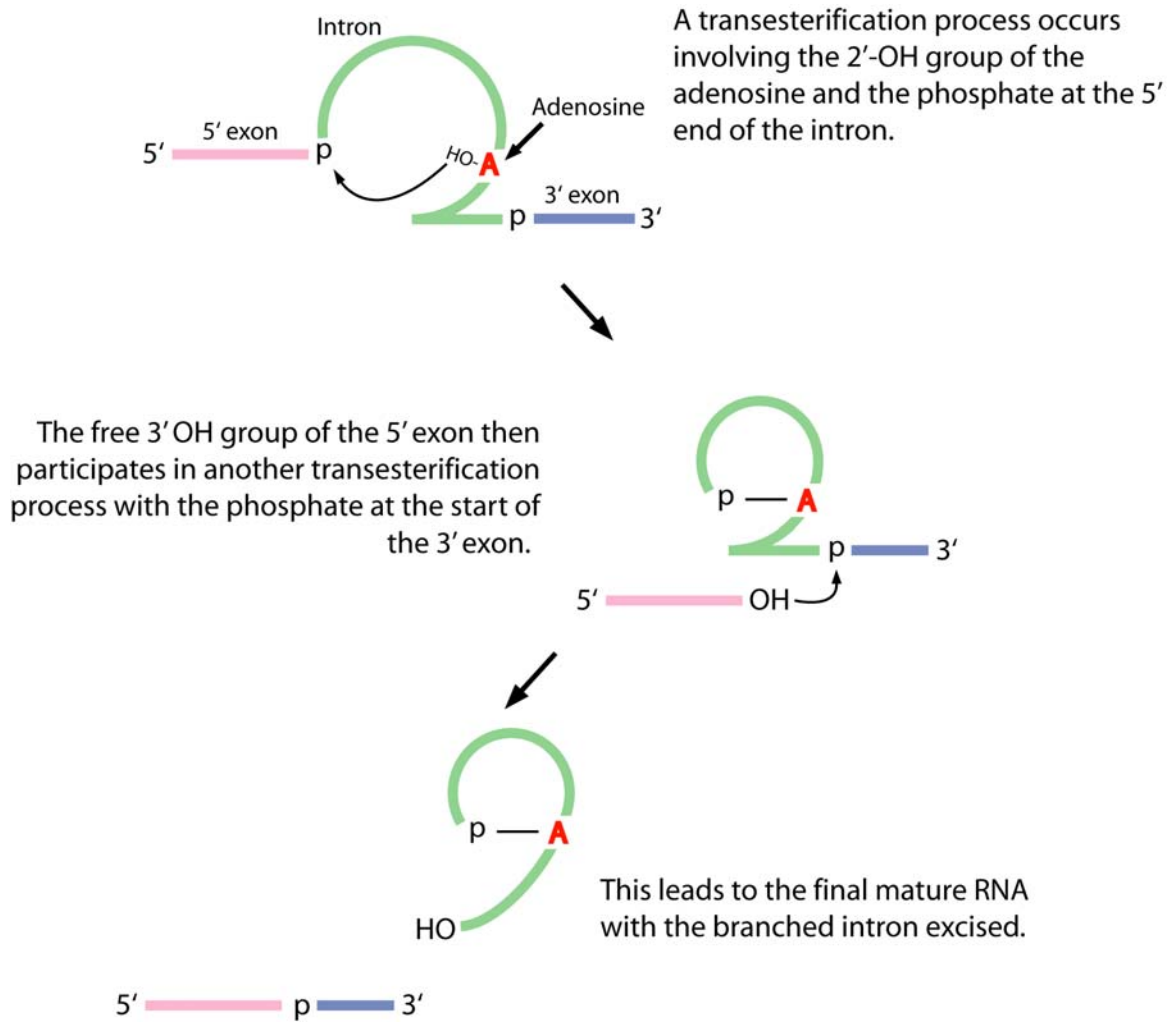


Figure 2. The splicing process of group II introns.

Nuclear spliceosomal introns (Black, 2003; Nilsen, 2002; Staley & Guthrie, 1998) are spliced out in a manner similar to that of group II introns (Figure 3). Like group II introns, the splicing process involves an intronic adenosine residue that through its 2'-hydroxyl group partake in the first transesterification reaction. The result of the splicing process is also similar resulting in a mature RNA as well as a lariat intron. However unlike group I or group II introns where there is a need for internal base pairing that results in RNA secondary structures, nuclear spliceosomal introns requires the assistance of snRNPs (small nuclear ribonucleoprotein particles). These snRNPs are composed of both RNA and protein components. The RNA components are short uridine-rich RNA, which are abundant in the mammalian nuclei. There are six different types of snRNAs named U1 to U6. These snRNAs are involved in base pairing with the intronic and exonic sequences thereby are thought to play the role of the internal base pairing found in group I and group II introns. In addition to base pairings, numerous protein-to-protein interactions taken place in the spliceosome. The mRNA sequence itself consists of specific signals that snRNPs binds. The signals consist of a donor site found at the junction of the 5' exon and the intron, an acceptor site found at the junction of the intron and the 3' exon, a branchpoint containing the adenosine responsible for the transesterification process and a pyrimidine-rich region which is bound by U2AF65 (65kDa subunit of U2 auxiliary factor). U2AF (U2 auxiliary factor) is a dimer consisting of a 65kDa and a 35kDa subunit. Splicing begins with the binding of U1 snRNP to the donor site, SF1 to the branchpoint, U2AF65 to the pyrimidine-rich region, U2AF35 to the AG dinucleotide at the acceptor site. U1 snRNP binds to the donor site via base pairing using the U1 snRNA. This entire complex is termed the E (early) complex.

U2 snRNP is also bounded in this complex but not to the branchpoint. Once U2 snRNP is bound to the branchpoint, the A complex is formed. A complex consisting of U4/U5/U6 snRNPs then binds to form the B complex. Rearrangements of the various components then occur with the U6 snRNP replacing U1 snRNP and U1/U4 snRNP dissociating from the complex. The resulting C complex then catalyzes the two transesterification processes that lead to the loss of the lariat intron and the joining of the adjacent exons.

Nuclear tRNA introns are spliced in a manner quite different from the rest of the intron types. Instead of two transesterification processes, cleavage and ligation reactions are used to remove the intron. An endonuclease is used to excise the intron on both ends and ligation then occurs with the expenditure of energy in the form of GTP and ATP.

The work presented here is concerned with only nuclear spliceosomal introns as they are the major type of splicing and there is the involvement of alternative splicing. Therefore, the use of splicing for the rest of the text is restricted to nuclear spliceosomal type.

The nascent mRNA sequence contains both exons and introns. The intron contains a branchpoint that includes the adenosine used in the transesterification reaction and a pyrimidine-rich region. The donor and acceptor splice sites are found at either ends of the introns. Consensus sequences are shown below the labels.

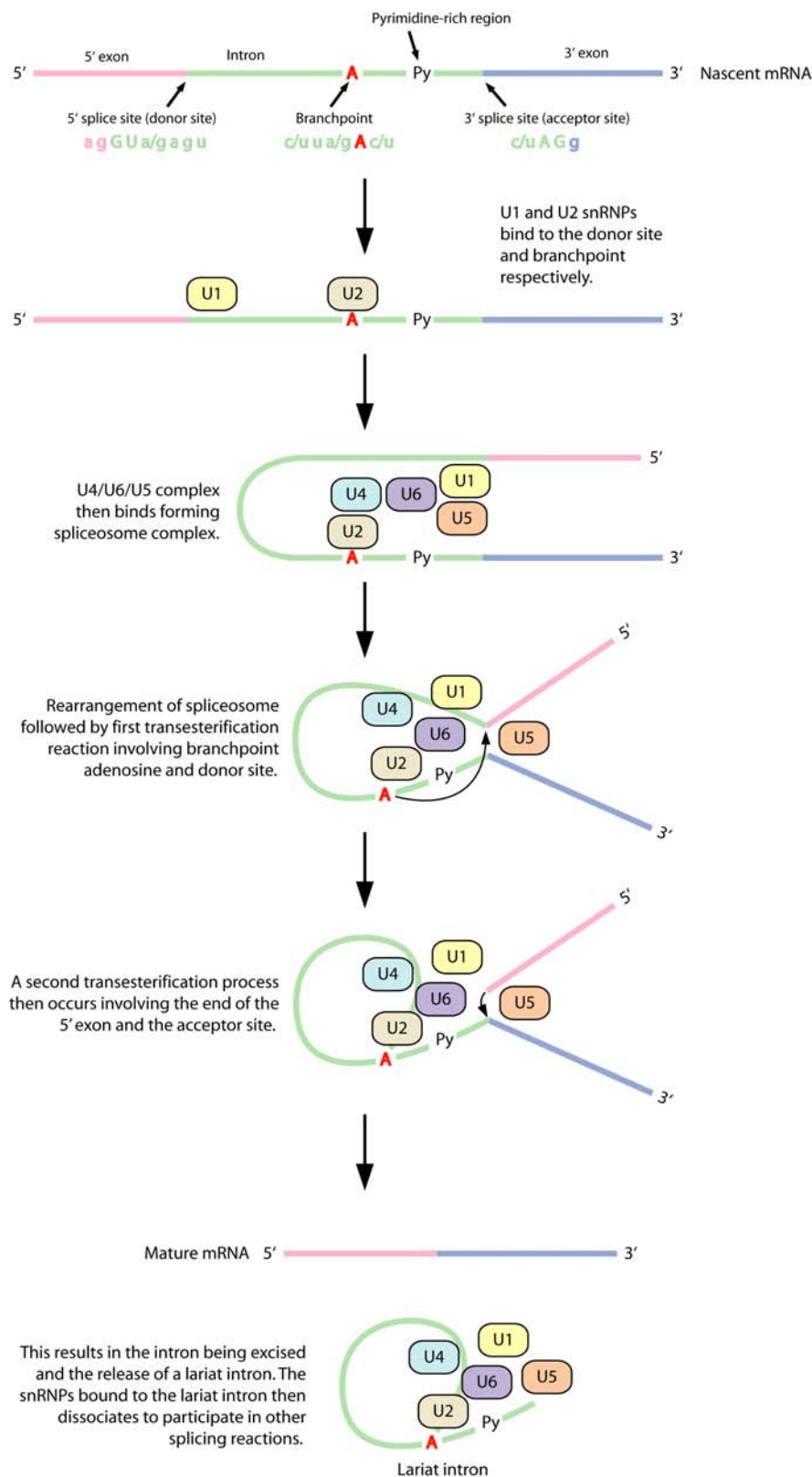


Figure 3. The splicing process of nuclear spliceosomal introns.

1.2 Exon and intron definition model

Most introns of higher eukaryotes are quite long in length reaching hundreds of thousands of nucleotides while most exons are in comparison much shorter and exhibit a narrow distribution between 50-300 nucleotides (Hawkins, 1988). This has led to the conception of exon definition where the splicing machinery across exons interact to define the exon (Hoffman & Grabowski, 1992). In fact, exons that exhibit lengths longer than 300bp shows decreased efficiency in splicing (Robberson, Cote, & Berget, 1990). In lower eukaryotes where the intron length is much shorter and shows a narrow distribution, the intron definition is thought to occur (Talerico & Berget, 1994). However not all the intron and exons in the same organism fall into either one of these two forms of definition, exon definition and intron definition may take place in the same organism depending on the lengths of the exons and introns (Talerico & Berget, 1994).

1.3 Alternative splicing

There are several elements that can influence RNA splicing. These elements allow for the selective inclusion and exclusion of specific exons allowing for the generation of multiple transcripts from the same transcript. This phenomenon has been entitled alternative splicing.

1.3.1 Splice site motifs

The splice site motifs themselves influence RNA splicing by their ability to interact effectively with the spliceosome. This interaction is determined by the level of conservation of the splice site motif to the canonical one. Splice site motifs that are more similar to the canonical splice site motifs are more likely to be incorporated into the final transcripts. These are shown by the fact that mutations

in the splice site motif can prevent the inclusion of exons. Weak splice site motifs or motifs that exhibit a weak level of conservation to the canonical splice site motif can be enhanced by other protein factors that compensates for the weakness of the conservation. This can allow for the differential inclusion and exclusion of exons leading to alternative splicing.

1.3.2 Splicing enhancers and suppressors

In addition to the basic splicing machinery, there are *cis* elements on the mRNA molecule that modulates splicing. These *cis* elements consisting of motifs that reside in the exonic or intronic sequences and may inhibit or promote splicing of specific exons or introns. Thus, four types of *cis* elements are found, ESE (exonic splicing enhancers), ISE (intronic splicing enhancers), ESS (exonic splicing suppressors) and ISS (intronic splicing suppressors). These *cis* elements are the binding targets for *trans* acting protein elements. These *trans* acting protein elements typically contain a RNA binding motif and another domain responsible for protein-protein interactions. SR proteins and hnRNP (heterogeneous nuclear ribonucleoprotein) proteins are known members of these *trans* acting elements. SR proteins typically contain one or two RNA recognition motif (RRM) as well as an arginine/serine (RS) rich domain at the carboxyl end. The RRM are responsible for their binding to the RNA molecule while the RS domain participates in protein-protein interactions. HnRNP proteins do not contain a RS domain but some members contain an arginine/glycine rich domain that is responsible for protein-protein interaction as well as RNA binding.

1.3.4 Interaction with the transcriptional machinery

Recent studies have implicated the transcriptional machinery in alternative splicing. Both transcription and RNA processing take place in the nucleus and can occur together. This has raised the possibility that the splicing machinery may interact with the transcriptional machinery. Furthermore, as RNA splicing takes place even while the transcript is being synthesized, upstream *cis* elements like the 5' and 3' splice site motifs, ESE, ESS, ISE and ISS that can influence splicing are presented to the splicing machinery on a temporary basis. For example, a 10kbp intron would take approximately 5 to 10 minutes to synthesize (Goldstrohm, Greenleaf, & Garcia-Blanco, 2001). This means that the 5' splice site would be available for 5-10 minutes prior to the 3' splice site. Studies have shown that only RNA polymerase II synthesized mRNA are spliced while RNA polymerase I and III synthesized mRNA are not spliced (Sisodia, Sollner-Webb, & Cleveland, 1987; Smale & Tjian, 1985).

Studies have discovered that the CTD (carboxyl terminal domain) of RNA polymerase II is a potential binding site for many proteins some of which influence splicing. This was shown in experiments where a truncated CTD resulted in inhibition of splicing (McCracken et al., 1997; Misteli & Spector, 1999). The CTD consists of a long series of heptapeptide repeats that could serve as a binding site for multiple proteins. Direct evidence has been obtained for the binding of Prp40p, a yeast U1 sRNP-associated protein to phosphorylated CTD (Morris & Greenleaf, 2000).

Although some studies have indicated that RNA polymerase II is not absolutely required for splicing (Green, Maniatis, & Melton, 1983; Padgett, Hardy,

& Sharp, 1983; Tani & Ohshima, 1991), it is likely that RNA polymerase II acts as an enhancer to the splicing process.

1.4 Impact of alternative splicing

Alternative splicing in recent years have been implicated in numerous cell function (Stamm et al., 2005). No attempts will be made to duplicate the excellent review article cited, but two impacts of alternative splicing which is of relevance in this work is shown below.

1.4.1 Protein diversity

The completion of the first draft of the human genome project (Lander et al., 2001; Venter et al., 2001) has brought about questions as to the cause of the complexity of the human being. Estimates of the number of genes contained in the human genome is a mere 35,000, only double of that of the 18,000 or so genes in *Drosophila melanogaster* (Adams et al., 2000). Therefore, the complexity of the human being cannot be simply accounted for by the number of genes. There has to exist other mechanisms for increasing the protein diversity of the organism and alternative splicing is one such mechanism. Alternative splicing is able to allow for the production of a number of different transcriptions from a single gene. This has altered the central dogma of a single gene to a single protein to one where a single gene can potential produce several protein products.

An example of the potential for alternative splicing to produce a large number of protein products from a single gene is the CD44 gene (Bajorath, 2000; Borland, Ross, & Guy, 1998). The CD44 genes consist of 21 exons which includes 12 which are alternatively spliced giving rise to the potential to encode

for 1024 isoforms. Even more telling is the Dscam gene in *Drosophila melanogaster* which consists of 115 exons of which 95 are alternatively spliced. This means that the Dscam gene can potentially produce 38,106 isoforms (Celotto & Graveley, 2001; Schmucker et al., 2000), more than double of the number of genes in *Drosophila melanogaster* and equal to the number of genes in human.

In addition to these examples, there are numerous studies (Brett et al., 2000; Clark & Thanaraj, 2002; Croft et al., 2000; Hanke et al., 1999; Kan, Rouchka, Gish, & States, 2001; Lander et al., 2001; Mironov, Fickett, & Gelfand, 1999; Modrek, Resch, Grasso, & Lee, 2001) that a significant portion of the human genes are alternatively spliced (40-60%). The actual percentage of genes being alternatively spliced is likely to be higher as some isoforms may be available in small amounts or which are differentially regulated thus escaping detections by current methods.

1.4.2 Relevance to pharmacogenomics

The development of drugs is hindered by the lack of drug effectiveness and occurrences of unwanted side effects. Both of these problems are affected by RNA splicing. RNA splicing being a commonplace cellular mechanism plays an important role in the development of new drugs. The tightly controlled alternative splicing process can be altered in disease state. Known examples include the MKK3/6-p38 pathway that has been found to induce localization of splicing factors from the nucleus to the cytoplasm (Misteli, 2000). These splicing factors include hnRNPA1 which influences the alternative splicing of genes, which include the fibroblast growth factor receptor 2 (FGFR2) (Lopez, 1998). Changes in the ratio of the various isoforms of FGFR2 are known to lead to severe

developmental defects in mice (De Moerlooze et al., 2000). More telling is the fact that of the Human Gene Mutation Database (Krawczak et al., 2000) shows that 4771 out of 50139 records contained in their database as of January 2005 affect the splicing of the genes involved.

1.5 Bioinformatics resources for alternative splicing

1.5.1 General bioinformatics databases used in the study of alternative splicing

Other than specialized alternative splicing databases, several other databases are also used in the study of alternative splicing. These databases will be discussed below together with some of the implications they have for the study of alternative splicing.

1.5.1.1 GenBank

GenBank (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2004) is one of the primary depositories for nucleotide sequences apart from DDBJ (DNA DataBank of Japan) and EMBL-Bank by EMBL (European Molecular Biology Laboratory). All three databases are part of the International Nucleotide Sequence Database Collaboration and exchange information on a daily basis such that the information contained in either one of these databases is found in the other two. Therefore, any of the three databases contains essentially the same information.

Biologists can deposit the nucleotide sequences that they sequenced to GenBank who will then make it available to the entire research community. As of 16 August 2004, the most recent release of GenBank being release 142 contains

35,532,003 records totally 40,325,321,348 bp. The entire GenBank is made available as a series of text files (numbering 634 files), each containing a number of records in GenBank format. The GenBank format shown in Figure 4 is a human readable text format that is column delimited. Other than sequence information, general information about the entire sequence as well as specific information about certain portions of the sequence is present in the GenBank format. The GenBank record is basically broken up into three parts, the top header portion provides annotation about the record itself, the center portion contains the Feature Table section that provides detailed annotation about specific portions of the sequence and lastly the bottom portion provides the nucleotide sequence. Gene structure information important for the study of alternative splicing like the exon/intron arrangement are found in the feature table section that provides annotation for specific portions of the sequence.

To facilitate handling of GenBank which occupies 153 GB of data storage, the files comprising GenBank is broken into divisions shown in Table 1. This allows users to obtain specific records without downloading the entire GenBank. Of interest is the EST division otherwise known as dbEST that contains large numbers of ESTs, which are suitable for clustering or alignment against the genome to locate for alternative splicing.

The growth of GenBank in recent years is staggering as shown in Figure 5. This is especially so for the EST division which has 21,939,770 records comprising of 11,453,940,185 bp. This has good implications for the study of alternative splicing as it means that there is a large pool of transcript information present in the current GenBank useful for the detection of alternative splicing.

Furthermore, this source of information is going to increase substantially judging from the growth of GenBank.

```

LOCUS      AAURRA          118 bp ss-rRNA          RNA          16-JUN-1986
DEFINITION A.auricula-judae (mushroom) 5S ribosomal RNA.
ACCESSION  K03160
VERSION    K03160.1  GI:173593
KEYWORDS   5S ribosomal RNA; ribosomal RNA.
SOURCE     A.auricula-judae (mushroom) ribosomal RNA.
  ORGANISM  Auricularia auricula-judae
            Eukaryota; Fungi; Eumycota; Basidiomycotina; Phragmobasidiomycetes;
            Heterobasidiomycetidae; Auriculariales; Auriculariaceae.
REFERENCE  1 (bases 1 to 118)
  AUTHORS  Huysmans,E., Dams,E., Vandenberghe,A. and De Wachter,R.
  TITLE    The nucleotide sequences of the 5S rRNAs of four mushrooms and
            their use in studying the phylogenetic position of basidiomycetes
            among the eukaryotes
  JOURNAL  Nucleic Acids Res. 11, 2871-2880 (1983)
FEATURES   Location/Qualifiers
  rRNA     1..118
            /note="5S ribosomal RNA"
BASE COUNT 27 a      34 c      34 g      23 t
ORIGIN     5' end of mature rRNA.
            1 atccacggcc ataggactct gaaagcactg catcccgctc gatctgcaaa gttaaccaga
            61 gtaccgcccc gttagtacca cggtggggga ccacgcggga atcctgggtg ctgtggtt
//

```

Figure 4. A sample of a GenBank record. Each line contains a number of fields which are column delimited.

Division	Description
PRI	Primate sequences.
ROD	Rodent sequences.
MAM	Other mammalian sequences.
VRT	Other vertebrate sequences.
INV	Invertebrate sequences.
PLN	Plant, fungi and algae sequences.
BCT	Bacterial sequences.
VRL	Viral sequences.
PHG	Phage sequences.
SYN	Synthetic and chimeric sequences.
UNA	Unannotated sequences.
EST	EST sequences.
PAT	Patent sequences.
STS	STS (sequence tagged site) sequences.
GSS	GSS (genome survey sequence) sequences.
HTG	HTGS (high throughput genomic sequence) sequences).
CON	Complex records.

Table 1. GenBank divisions and their description.

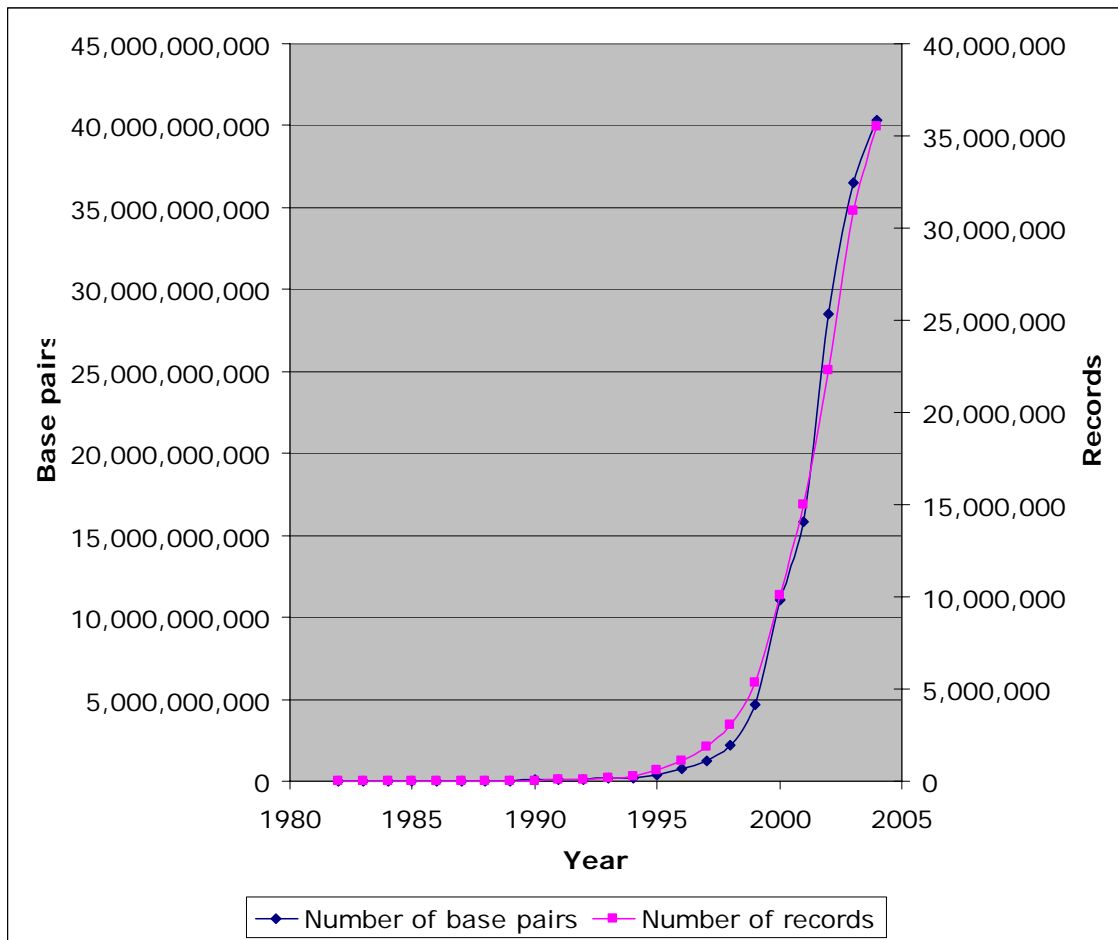


Figure 5. Growth of GenBank in terms of the number of base pairs and records. Data shown as of 16 August 2004.

1.5.1.2 Entrez Genome

Entrez Genome is part of Entrez (Wheeler et al., 2003) developed at NCBI (National Center for Biotechnology Information). Entrez is a system that integrates the resources (including GenBank) at NCBI and exposes a text-based search utility that allows users to search all the resources integrated. Entrez Genomes includes genomes of bacteria, archaea and eukaryotes. These genomic sequences are especially useful in revealing the gene structure by alignment of transcript sequences against genomic sequences. Gene structure information can then be clustered to reveal alternative splicing events. In recent years, due to the maturation of the sequencing technology, several genomes have been completely sequenced. As the technology gets increasing accessible, there will be a marked increase in the number of genomes available for the determining gene structure information via alignment of transcripts. This is further facilitated by the large increase in the number of transcript data in GenBank.

1.5.1.3 Swiss-Prot

Swiss-Prot (Bairoch, Boeckmann, Ferro, & Gasteiger, 2004) is a primary database that deals with protein sequences. As of 19th August 2004, Swiss-Prot contains 156,998 records totally 57,769,334 amino acids. Swiss-Prot aims to provide quality annotation through manual curation with minimum redundancy. An example of a Swiss-Prot record is shown in Figure 6. The Swiss-Prot is very rich having numerous detailed annotations about the protein sequence. Much like the GenBank format, the Swiss-Prot record is broken up into three parts. The top header portion contains information about the protein sequence in general. Another distinguishing feature of Swiss-Prot is the emphasis on cross-linking to other databases; as such Swiss-Prot records typically contain numerous links to

other databases like PDB (Berman et al., 2000), Pfam (Bateman et al., 2004) and InterPro (Mulder et al., 2003). The center portion contains the Feature Table that provides detailed annotation about specific regions of the protein sequence. The bottom portion contains the protein sequence. Of interest in the study of alternative splicing is the availability of the VARSPLIC in the Feature Table, which provides information on the effects of alternative splicing on the protein sequence. This has been exploited for several studies on alternative splicing (Kersey, Hermjakob, & Apweiler, 2000; Kondrashov & Koonin, 2003; Kriventseva et al., 2003). As in GenBank records, Swiss-Prot records are deposited by biologist, however experts on the particular protein further check Swiss-Prot records. This leads to increased accuracy as well as richness of the annotation. However due to the manual curation aspect, the coverage of Swiss-Prot is not as wide as one would hope for. Therefore to counter this limitation, a computer-annotated portion entitled TrEMBL is provided as a supplement to Swiss-Prot. This supplement consists of all translation of EMBL nucleotide sequences that are not available in Swiss-Prot. As of 19th August 2004, TrEMBL consist of 1,379,120 records encompassing 431,121,293 amino acids. This is significantly larger than the number of records in Swiss-Prot (about ten times larger), reflecting the time consuming nature of manual curation, which is unable to keep up with the explosive increase in the number of nucleotide sequences.

```

ID VEGA_CHICK STANDARD; PRT; 216 AA.
AC P52582; Q91420;
DT 01-OCT-1996 (Rel. 34, Created)
DT 15-JUL-1998 (Rel. 36, Last sequence update)
DT 05-JUL-2004 (Rel. 44, Last annotation update)
DE Vascular endothelial growth factor A precursor (VEGF-A) (Vascular
DE permeability factor) (VEGF).
GN Name=VEGF; Synonyms=VEGFA;
OS Gallus gallus (Chicken), and
OS Coturnix coturnix japonica (Japanese quail).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Archosauria; Aves; Neognathae; Galliformes; Phasianidae; Phasianinae;
OC Gallus.
OX NCBI_TaxID=9031, 93934;
RN [1]
RP SEQUENCE FROM N.A.
RC SPECIES=Chicken; TISSUE=Heart;
RA Takahashi T.;
RT "Chick embryonic ventricular myocytes VEGF.";
RL Submitted (FEB-1998) to the EMBL/GenBank/DBJ databases.
RN [2]
RP SEQUENCE FROM N.A. (ISOFORMS VEGF-190; VEGF-166 AND VEGF-146).
RC SPECIES=C.c.japonica; TISSUE=Embryo;
RX MEDLINE=96005007; PubMed=7556923;
RA Flamme I., von Reutern M., Drexler H.C., Syed-Ali S., Risau W.;
RT "Overexpression of vascular endothelial growth factor in the avian
RT embryo induces hypervascularization and increased vascular
RT permeability without alterations of embryonic pattern formation.";
RL Dev. Biol. 171:399-414(1995).
RN [3]
RP SEQUENCE OF 60-187 FROM N.A. (ISOFORMS VEGF-190 AND VEGF-166).
RC SPECIES=C.c.japonica;
RX MEDLINE=95301109; PubMed=7781909;
RA Flamme I., Breier G., Risau W.;
RT "Vascular endothelial growth factor (VEGF) and VEGF receptor 2 (flk-1)
RT are expressed during vasculogenesis and vascular differentiation in
RT the quail embryo.";
RL Dev. Biol. 169:699-712(1995).
CC -!- FUNCTION: Growth factor active in angiogenesis, vasculogenesis and
CC endothelial cell growth. Induces endothelial cell proliferation,
CC promotes cell migration, inhibits apoptosis and induces
CC permeabilization of blood vessels. Binds to the VEGFR1/Flt-1 and
CC VEGFR2/Kdr receptors, heparan sulfate and heparin (By similarity).
CC -!- SUBUNIT: Homodimer; disulfide-linked. Also found as heterodimer
CC with PlGF (By similarity).
CC -!- ALTERNATIVE PRODUCTS:
CC Event=Alternative splicing; Named isoforms=3;
CC Comment=Additional isoforms seem to exist;
CC Name=VEGF-190;
CC IsoId=P52582-1; Sequence=Displayed;
CC Name=VEGF-166;
CC IsoId=P52582-2; Sequence=VSP_004633, VSP_004634;
CC Note=Has been shown to exist only in quail so far;
CC Name=VEGF-146;
CC IsoId=P52582-3; Sequence=VSP_004635, VSP_004636;
CC Note=Has been shown to exist only in quail so far;
CC -!- TISSUE SPECIFICITY: Abundantly and equally expressed in heart and
CC liver. In kidney glomeruli, brain and yolk sac, VEGF-166 is 5- to
CC 10-times more abundant than VEGF-190.
CC -!- DEVELOPMENTAL STAGE: VEGF-166 is expressed early at day 1 and is
CC upgraded during gastrulation. Expression of VEGF-190 is detectable
CC only from day 2.
CC -!- DOMAIN: VEGF-190 contains a basic insert which acts as a cell
CC retention signal.
CC -!- SIMILARITY: Belongs to the PDGF/VEGF growth factor family.
CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration
CC between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC the European Bioinformatics Institute. There are no restrictions on its
CC use by non-profit institutions as long as its content is in no way
CC modified and this statement is not removed. Usage by and for commercial
CC entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC or send an email to license@isb-sib.ch).
CC -----
DR EMBL; AB011078; BAA24925.1; -.
DR EMBL; S79680; AAB35371.1; -.
DR HSSP; P49763; 1FZV.
DR InterPro; IPR000072; PD_growth_factor.
DR Pfam; PF00341; PDGF; 1.
DR ProDom; PD001629; PD_growth_factor; 1.
DR SMART; SM00141; PDGF; 1.
DR PROSITE; PS00249; PDGF_1; 1.
DR PROSITE; PS50278; PDGF_2; 1.
KW Alternative splicing; Angiogenesis; Glycoprotein; Growth factor;
KW Heparin-binding; Mitogen; Multigene family; Signal.
FT SIGNAL 1 26 By similarity.
FT CHAIN 27 216 Vascular endothelial growth factor A.
FT DISULFID 52 94 By similarity.
FT DISULFID 83 128 By similarity.
FT DISULFID 87 130 By similarity.
FT DISULFID 77 77 Interchain (By similarity).
FT DISULFID 86 86 Interchain (By similarity).
FT CARBOHYD 101 101 N-linked (GlcNAc...) (Potential).
FT VARSPLIC 142 142 K -> N (in isoform VEGF-166).
FT /FTId=VSP_004633.
FT VARSPLIC 143 166 Missing (in isoform VEGF-166).
FT /FTId=VSP_004634.
FT VARSPLIC 166 166 F -> L (in isoform VEGF-146).
FT /FTId=VSP_004635.
FT VARSPLIC 167 210 Missing (in isoform VEGF-146).
FT /FTId=VSP_004636.
SQ SEQUENCE 216 AA; 25203 MW; 82E669C2F6FC6DA7 CRC64;
MNFLLTWHHW GLAALLYLQSL AELSKAAPAL GDGERKPNVEV IKFLEVYERS FCRTIETLVLD
IFQEYFDEVE YIFRPSCVPL MRCAGCCGDE GLECVVDVYV NVTMEIARIK PHQSQHIAMH
SFLQHSKDCD RPKKDKNKQ EKSKRKGK GKRRKRRKGR YKPPSPHCEP CSERRKHLFV
QDPQCKCSC KFTDSRCKSR QLELNERTCR CEKFRR
//

```

Figure 6. An example of a Swiss-Prot record.

1.5.1.4 UniGene

The large number of EST sequence in the public databases like GenBank represents a rich source of transcript information. However there is a high level of redundancy in these sequences. UniGene (Wheeler et al., 2003) is a system that clusters the sequences in GenBank on the basis of sequence similarity to form clusters of sequences believed originated from a single gene thereby reducing the redundancy resulting in more useful data for further work. Therefore the sequences that make up a UniGene cluster can be considered to be derived from a single gene. Each UniGene record is linked to relevant NCBI resources as well as external databases such as the Mouse Genome Informatics (MGI). The genomic location of the cluster is also provided if available. The number of record in UniGene is represented in Table 2.

Since alternative splicing occurs within the gene, each UniGene cluster ought to contain the various splicing variants if present. This makes UniGene a valuable source of information for alternative splicing studies. The information contained in each UniGene cluster can be used in two ways to extract alternative splicing information. The first way uses the sequences contained within each cluster for alignment against the corresponding genomic sequence resulting in the determination of the gene structure. Variations in the gene structure of the different sequences in the cluster are evidence for alternative splicing. The second way directly infers the possibility of alternative splicing by perform a comparison between the sequences in the cluster. In the event that there is no alternative splicing, the sequences in the cluster should overlap each other perfectly. In the event that alternative splicing occurs, large insertions and

deletions would appear in the alignments. The insertion and deletions are due to the differences in the selection of the exons.

Organism	Number of UniGene records
Bos taurus	24,195
Canis familiaris	15,665
Homo sapiens	107,014
Mus musculus	76,876
Ovis aries	3,160
Rattus norvegicus	40,380
Sus scrofa	21,947
Gallus gallus	21,035
Xenopus laevis	23,754
Xenopus tropicalis	14,632
Danio rerio	21,088
Oncorhynchus mykiss	14,241
Oryzias latipes	8,133
Salmo salar	1,076
Ciona intestinalis	14,098
Strongylocentrotus purpuratus	2,614
Anopheles gambiae	14,653
Apis mellifera	5,900
Bombyx mori	2,055
Drosophila melanogaster	14,616
Caenorhabditis elegans	15,942
Schistosoma mansoni	1,192
Physcomitrella patens	6,962
Pinus taeda	12,695
Arabidopsis thaliana	20,783
Glycine max	12,629
Helianthus annuus	1,909
Lactuca sativa	9,656
Lotus corniculatus	8,187
Lycopersicon esculentum	5,061
Malus x domestica	5,049
Medicago truncatula	5,297
Populus tremula	2,709
Solanum tuberosum	5,713
Vitis vinifera	12,518
Hordeum vulgare	12,029
Oryza sativa	33,722
Saccharum officinarum	4,773
Sorghum bicolor	8,146
Triticum aestivum	24,728
Zea mays	14,179
Chlamydomonas reinhardtii	6,077
Dictyostelium discoideum	3,865
Toxoplasma gondii	7,110
Magnaporthe grisea	5,721
Neurospora crassa	3,217
Hydra magnipapillata	5,789
Total	702,790

Table 2. Breakdown of the number of records in UniGene by organism. Statistics as of 21st August 2004.

1.5.1.5 Pfam

Pfam is a protein domain family database. It contains multiple sequence alignments as well as profile HMM (hidden Markov models) of the domains. Similar to Swiss-Prot, there are two parts to Pfam, there is Pfam-A which is the curated portion and Pfam-B which is automatically computed to provide larger coverage. The source data for Pfam is a non-redundant set of protein sequences derived from both Swiss-Prot and TrEMBL. Pfam-A families are first created manually using a seed alignment involving sequences known in that particular family. Using this seed alignment, a profile HMM is created and searched against the non-redundant set of protein sequences to identify new members on the basis of a good match (being above a certain threshold). These newly detected members are then incorporated into the alignment to form the full alignment. This full alignment is then used to create a new profile HMM that can be used to detect for new members for that particular Pfam family. Pfam-B differs from Pfam-A on the fact that the seed alignments are derived from ProDom (Corpet, Gouzy, & Kahn, 1998) instead. ProDom is a database of protein domain families generated by sequence comparison and the families in ProDom are used only in the event that the sequences within the ProDom family do not intersect substantially with any Pfam-A family. As of 19th August 2004, 74% of proteins have at least one match to Pfam-A and 21% of proteins have at least one match to Pfam-B. Therefore, used together, Pfam provides protein domain information on 95% of all protein sequences. The main interest in Pfam for the study of alternative splicing is the determination of the effects of alternative splicing on the protein domain organization. The profile HMM provided by Pfam can be used in conjunction with the HMMER package to locate for protein domains on gene

models. The impact of alternative splicing can then be known by correlating the protein domains and the various different splicing variants generated by alternative splicing.

1.5.2 Alternative splicing databases

Many specialized alternative splicing databases have been created for the purpose of studying the phenomenon of alternative splicing. The following is a selection of some of these databases.

1.5.2.1 Alternative Splicing Database (ASD)

ASD (Thanaraj et al., 2004a) is the project of the ASD consortium and it consists of three distinct databases, namely AltSplice, AltExtron and AEDB (Alternative exon database). AltSplice and AltExtron are both created using a computational pipeline that generates the alternative splicing dataset. AltExtron differs from AltSplice being the development database while AltSplice is the production database. AltExtron also differs from AltSplice in the dataset that it uses. AltExtron uses data from EMBL/GenBank while AltSplice uses data from the Ensembl genome annotation project. AltExtron and AltSplice both generate their data by first aligning genome and transcript sequences. These alignments are then checked and analyzed for alternative splicing events. After the alternative splicing dataset is obtained, it is then further annotated. Currently, these annotations include characterization of splice signals (being weak or strong), characterization of intron/exon forms (for example U2 or U12 type intron), the expression states of the splicing variants, conservation of splicing events across species, and incorporation of mutation and SNP data. Currently as of 20th August 2004, AltSplice consists of 15,384 multi exonic gene structure with 11,844 of

these having more than one splicing variant and a further 8,400 with delineated alternative splicing events. The contents of AltExtron as of 20th August 2004 is shown in Table 3.

AEDB on the other hand contains alternative splicing information extracted from literature. Each record in AEDB consists of the alternatively spliced exons' nucleotide sequences as well as biological annotation including developmental regulation, tissue specificity, alternative splicing isoform function and its association with diseases. A web server is available as with AltSplice and AltExtron that provides access to the database. As of 20th August 2004, the database consists of 1555 records.

ASD is unique due to the fact that it consists of information that is generated using 2 different approaches, one using computational methods and the other being extraction from literature. The computational approach allows one to generate large amounts of data as seen from the statistics of AltSplice and AltExtron, however the data is likely to be less complete than that obtained via extraction from literature. Extraction from literature however is time consuming and thus the coverage is less than ideal. The data obtained from both processes can therefore effectively compensate for each of the approaches' drawbacks.

Organism	Number of multi exonic genes	Number of alternatively spliced genes
Human	5,584	2,581
<i>Drosophila melanogaster</i>	7,881	1,302
Mouse	1,766	658
Caenorhabditis elegans	9,552	598
Arabidopsis thaliana	13,302	758
Rat	228	42
Chicken	177	20
Cow	339	34
Zebra fish	257	21

Table 3. Statistics of AltExtron as of 20th August 2004.

1.5.2.2 ASDB

ASDB (Dralyuk, Brudno, Gelfand, Zorn, & Dubchak, 2000) like ASD is a database that contains information on alternatively spliced genes. However the approach taken by ASDB is different from that adopted by ASD. ASDB obtains its information by parsing Swiss-Prot and GenBank records for alternative splicing features. In the case of Swiss-Prot records, the feature “VARSPPLIC” in the feature table and the words “alternative splicing” in the comments fields are used to identify records having alternative splicing data. The identified records are then clustered to pool proteins that are of the same gene together to provide a more accurate view of the number of splicing variants of that gene. Since Swiss-Prot records are only concerned with protein sequences, the data generated using Swiss-Prot records form the protein section of ASDB entitled ASDB (proteins). For the nucleotide section of ASDB named ASDB (nucleotides), GenBank records are parsed for the presence of the words “alternative splicing”, after which only records containing complete gene sequences are retained to generate the final ASDB (nucleotides). ASDB as of August 20th 2004 contains 5024 records clustered into 366 gene models.

1.5.2.3 Human Alternative Splicing Database (HASDB)

Much like AltSplice and AltExtron, HASDB (Modrek et al., 2001) uses transcript sequences aligned against the genome to derive alternative splicing information. The source of the transcript sequences is UniGene (Wheeler et al., 2003) and the genomic sequence is the assembly of the human genome. The alignment is first performed using BLAST with consensus sequences of UniGene clusters against the human genomic assembly. After the genomic location of the gene is determined, that portion of the genomic sequence together with all the transcript

sequences in the UniGene cluster are aligned using dynamic programming to form a multiple sequence alignment. Alternative splicing events are then detected from the multiple sequence alignment.

HASDB contains 6201 alternative splicing events using approximately 2.1 million transcript sequences as the starting point of the procedure. The data is available and searchable on the web server.

1.5.2.4 Alternative Splicing Annotation Project (ASAP)

ASAP (C. Lee, Atanelov, Modrek, & Xing, 2003) is a database that grew out of HASDB. The methodology for ASAP is similar to that of HASDB but the data contained in it has been substantially increased. As of August 2002, the amount of data in ASAP is four times that of HASDB and ASAP not only includes human genes but also mouse genes.

ASAP has a very intuitive graphical interface that not only shows the various splicing variants individually, but also an overview of all the splicing variants in a single linear representation (much like the output of a splicing graph). The protein coding region is also drawn allowing users to appreciate the effects of alternative splicing on the protein sequence. Furthermore, tissue specificity information is also incorporated allowing users to search by tissue specificity.

1.5.2.5 Putative Alternative Splicing database (PALS db)

The approach taken by PALS db (Huang, Chen, Lai, Yang, & Yang, 2002) is different from the rest of the databases that we have discuss till now. Instead of aligning transcript to the genome, extracting information from literature or parsing existing public database records, PALS db generates alternative splicing data

using alignments of transcript sequences. Specifically PALS db uses human and mouse UniGene clusters as the source data. From each UniGene cluster, the longest sequence is taken as the reference sequence, and all the other sequences in the cluster are aligned against this reference sequence. Any deviations from a perfect alignment aside from those due to sequencing errors are potential evidence for alternative splicing. Although evidence for alternative splicing is obtained via this approach the exact nature of the alternative splicing is unknown. The latest release being release 6 dated 21st April 2003 consists of 33,111 human and 18,942 mouse UniGene clusters. This resulted in 14,232 human clusters (43%) and 8,794 mouse clusters (46.4%) having evidence of alternative splicing.

1.5.3 Bioinformatics applications used in the study of alternative splicing

Numerous bioinformatics applications have been used in the study of alternative splicing. Some of these applications were developed for the purpose of generating information on the gene structure which is the source data required for the detection of alternative splicing while others were developed for other purposes and have been adapted for use in the study of alternative splicing. The following section describes some of these applications and their contribution in the study of alternative splicing.

1.5.3.1 *BLAST (Basic Local Alignment Search Tool)*

BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) is one of the most widely used bioinformatics tool and one, which most biologist are familiar with. BLAST is a heuristic local alignment search tool that allows users to quickly search a

database of biological sequences with a query sequence in the bid to locate for similar sequences. The heuristic nature of BLAST enables it to search rapidly through the entire database of sequence to identify similar sequences. Instead of perform a computational intensive optimal local alignment using dynamic programming against each and every sequence in the database, BLAST first breaks up the query sequence into small fragments and uses these fragments to search the sequences within the database. Only in the event that the fragment matches perfectly to a region with the sequence is the sequence considered for further alignment. By incorporation of this step, BLAST can quickly eliminate any sequence that is unlikely to generate a good alignment on the assumption that any sequence that does not share a small stretch of identical sequence with the query sequence is unlikely to be similar. This initial match between the database sequence and the query sequence serves as the initial alignment (consisting of perfect matches). Once a list of database sequences sharing a match to one of the query fragment is determined, an extension of the initial alignment is carried out to lengthen the alignment. This extension process proceeds via dynamic programming, thus ensuring that the alignment is capable of incorporating insertion, deletions and mismatches. These are the final alignments that are shown to the users together with statistical measures like the expectancy value, which provides biologists with a measure on how likely is the alignment due to chance. The fact that the BLAST is a local alignment method means that as long as there is a stretch of similar sequence between the query and the database sequence, it will be picked up.

In the context of alternative splicing, BLAST has been used to align transcript sequences to the genomic sequences to reveal the gene structure

(Modrek et al., 2001; Thanaraj et al., 2004a). Although BLAST is a very popular and fast search tool, it is not ideally suited for the determination of the gene structure needed for the analysis of alternative splicing. Firstly, since it is a local alignment method, each exon will appear as a local alignment when aligned against the genome. However, BLAST does not correctly delineate the boundaries of the exon as the extension process will likely lead to an alignment that exceeds the boundaries of the exon. Furthermore, BLAST does not order the local alignments into a complete gene structure. In view of these limitations, other alignment methods have been designed to specifically generate gene structure information.

1.5.3.2 *sim4*

sim4 (Florea, Hartzell, Zhang, Rubin, & Miller, 1998) is a dedicated program that uses the local alignment methodology to align transcript sequences to genomic sequences. Like BLAST, it first breaks up the transcript sequence into many small fragments and uses each fragment to search the genomic sequence. Fragments having identical matches to the genomic sequence initiate the extension process (similar to BLAST). This results in a list of local alignments much like BLAST. Each local alignment at this point of time represents a probable exon. However, instead of stopping at this point as in BLAST searches, *sim4* uses dynamic programming to select a series of local alignments that best covers the entire original transcript. After this process is completed, the series of local alignments may overlap, in this event, the ends of the overlapping local alignments are trimmed such that it produces a canonical splice site if possible. Should there be a gap between consecutive local alignments, then the local alignments flanking the gap are first extended to close the gap. If this is not

possible, then another local alignment is searched within this gap. The result of these processes is the final global alignment that is reported.

1.5.3.3 Spidey

Like *sim4*, *Spidey* (Wheelan, Church, & Ostell, 2001) is a specialized program meant for the alignment of transcript sequences to genomic sequences. Like *sim4*, *Spidey* relies on several local alignments to generate the final global alignment. In fact, *Spidey* uses BLAST to detect the local alignments. The main difference between *sim4* and *Spidey* is that *Spidey* implements the concept of a window that prevents interference of nearby paralogs and pseudogenes. A window is basically a section of the genomic sequence on which the alignment is restricted to. This allows paralogs to be restricted to a window and not interfere with the alignment of the actual gene.

1.5.3.4 MGAlign

MGAlign (B. T. Lee, Tan, & Ranganathan, 2003) attempts to further improve on methods like *sim4* and *Spidey* by first locating the ends of the transcript which forms a window much like the one defined by *Spidey*. This window then serves as the boundaries where the transcript will be aligned to the genomic sequence. By forming a window as the first step of the algorithm, *MGAlign* seeks to reduce the computational cost in locating for local alignments on the genomic sequence.

1.5.3.5 Blat

BLAT (BLAST-like alignment tool) is a search tool that can be used to align transcript sequences to genomic sequences to determine the gene structure (Kent, 2002). Like *Spidey*, *sim4* and *MGAlign*, BLAT utilizes the local alignment methodology to delineate the gene structure. However, BLAT implements an

indexing procedure that consists of non-overlapping fragments of the genomic sequence that allows for very fast determination of the genomic location of any transcript fragment. This allows BLAT to be much faster than sim4, Spidey and MGAlign as the time consuming process of scanning the genomic sequence repeatedly for matches against the transcript fragments is averted. The main drawback to this is the large memory requirement necessary for the index.

1.5.3.6 HMMER

The HMMER (Eddy, 1998) package of software deals with the creation of use of profile HMM (Hidden Markov Models). Profile HMMs are statistical models of multiple sequence alignments that retain the level of conservation of each amino acid in each position of the multiple sequence alignment. This is unlike simple patterns like those provided by Prosite where only the prominent amino acid at each position is retained in the pattern. This allows for a much more sensitive search to be carried out. The HMMER package of software includes the programs listed in Table 4. The package contains all the programs required for the creation and usage of profile HMM. The program hmmpfam in the HMMER package has been used to locate for Pfam domains in gene structures. This allows for the study of the effects of alternative splicing on the domain organization of the gene.

Program name	Description
hmmalign	Allows for the alignment of sequences to the existing model.
hmmbuild	Uses a multiple sequence alignment to build a profile HMM.
hmmcalibrate	Calculates more accurate expectancy values allowing for better parameters that makes searches more sensitive.
hmmconvert	Allows for the conversion of the HMMER HMM format to other formats.
hmmemit	Creates sequences using the probabilities from the profile HMM.
hmmfetch	Used to retrieve a specific model from the HMM database.
hmmindex	Indexes a HMM database.
hmmpfam	Searches a sequence against the HMM database.
hmmsearch	Searches a HMM against a sequence database.

Table 4. List of programs and its short description in the HMMER package.

1.5.4 Computational methods for detection of alternative splicing

Several computational methods have been used to detect for alternative splicing events. This section provides an overview of these methods, their strengths and their weakness.

1.5.4.1 Public database annotations

Public databases like GenBank and Swiss-Prot provide rich annotations in their sequence records. These annotations can be exploited to generate alternative splicing information for analysis.

GenBank sequence records contain gene structure information in their feature table section of the record. The availability of gene structure information allows for the detection of alternative splicing. Should a single gene have variation in the exon and intron arrangement (obtained from the gene structure), then there is evidence of alternative splicing. The sequences in the databases are redundant therefore a single gene could be represented by several sequence records. By clustering the sequences with gene structure information on the basis of their sequence similarity to form clusters of genes (each cluster believed to be representative of a single gene), any exon and intron arrangement variation in the sequences within the cluster is evidence for alternative splicing. Therefore, a straight forward approach to generating alternative splicing information would be to parse the public databases for sequences having gene structure information (namely the exon and intron arrangement) and then clustering the sequences to form gene clusters. Should there be variations in the exon and intron arrangement, then alternative splicing is likely. These variations can be analyzed to determine the exact nature of the alternative splicing.

Other public databases like Swiss-Prot have alternative splicing data directly incorporated into the sequence record. In this case, the task is simply to parse the records for this alternative splicing information.

This method of generating alternative splicing information has the advantage that the method is relatively simple. The databases are easily accessible through the internet and the design of parsers to parse the records is not technically difficult and many such parsers are freely available for such use. As the coverage of the public databases continue to increase (the increase has been explosive in recent years due to the maturation of sequencing technology), the amount of information that can be gathered will correspondingly increase. There is no need for human intervention as the process can be automated and thus the method is scalable (avoiding the manual curation bottleneck) as the size of the databases increases.

However, as the method relies on the annotation inherit in the records, any errors in the annotation will leads to erroneous alternative splicing information. Computational checks can be utilized but they will only be able to correct for the oblivious mistakes like overlapping exons and lack of consistency between the protein sequence it's corresponding translated coding sequence. Other errors like the incorrect assignment of exon boundaries are not possible to be corrected.

1.5.4.2 Extraction from literature

Another straight forward method of determining alternative splicing information is to simply extract them from literature. The literature represents a large pool of information that has been experimentally tested and thus of high reliability. Furthermore the data may be richer than that obtained from other methods in that

the data may include the mechanism of the alternative splicing, tissue specificity of certain splicing variants and any associated protein regulation factors.

However, this method relies heavily on manual intervention, requiring large numbers of man hours to extract these information. Computation methods like text searches may help to lower the amount of manual involvement but it will still require substantial amounts of manual labor.

1.5.4.3 Clustering of transcripts

Alternative splicing information can also be obtained solely through the use of transcript sequences. For constitutively spliced genes, since there is only one possible transcript that is made, all experimentally sequenced transcript should overlap each other perfectly as shown in Figure 7. In the case of alternatively spliced genes, each splicing variant will produce its own pool of experimentally sequenced transcripts as shown in Figure 8. The effect of alignment transcript of different pool is that they do not match perfectly in some cases due to the differential usage of exons. By clustering of the transcripts sequences produced by the same gene as shown in Figure 9, large deletions or insertions will occur depending on which is the reference transcript sequence. This is the phenotype exhibited by alternatively spliced genes and can be computational detected by checking for large deletions and insertions.

A big advantage of this method is that there is no need for any genomic sequences, thus this method can be applied to any organism where no genomic sequence is available. Transcript information is usually easier to come by and as evidenced by the growth of the EST division of GenBank, there is a large pool of such data that is growing at a phenomenal rate.

The main disadvantage of this method is that although evidence of alternative splicing is available, the exact nature of the alternative splicing is unclear. Alternative acceptor site, alternative donor site, intron retention and cassette exons can all exhibit the insertion and deletion phenotype. These four types of alternative splicing events cover all known types of alternative splicing. Therefore, this method does not allow us to determine the type of alternative splicing at work, it merely provides support for alternative splicing. Furthermore, prior to using this method, the transcript sequences have to be clustered to form clusters of transcript sequences that are derived from a single gene. This is sometimes difficult to do as transcripts belong to different splicing variants may be clustered individually instead of being pooled together. This is due to the fact that clustering methods rely on sequence similarity and in the case of splicing variants, they only share some degree of similarity due to their differential exon usage. Should this degree of similarity be lower than that used as the threshold for clustering the sequences, then the transcripts of the various splicing variants will cluster individually. Therefore, the frequency of alternative splicing detected using this method may be underestimated.

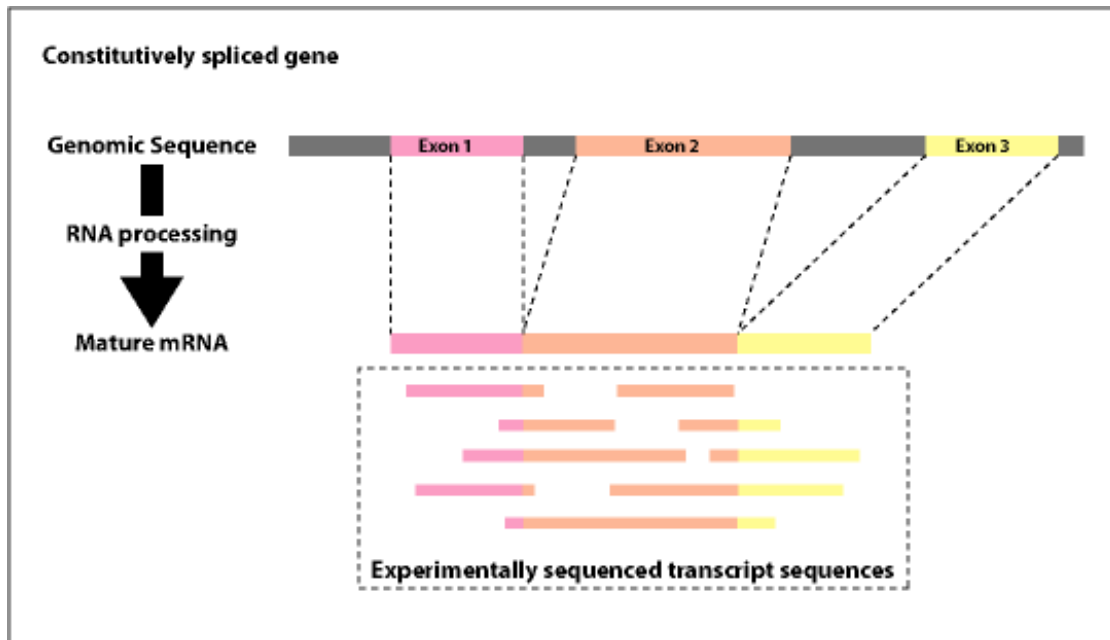


Figure 7. Schematic diagram of the transcripts produced by a constitutive gene. Experimentally sequence transcript sequences will align perfectly barring any sequencing errors with each other in regions of overlap.

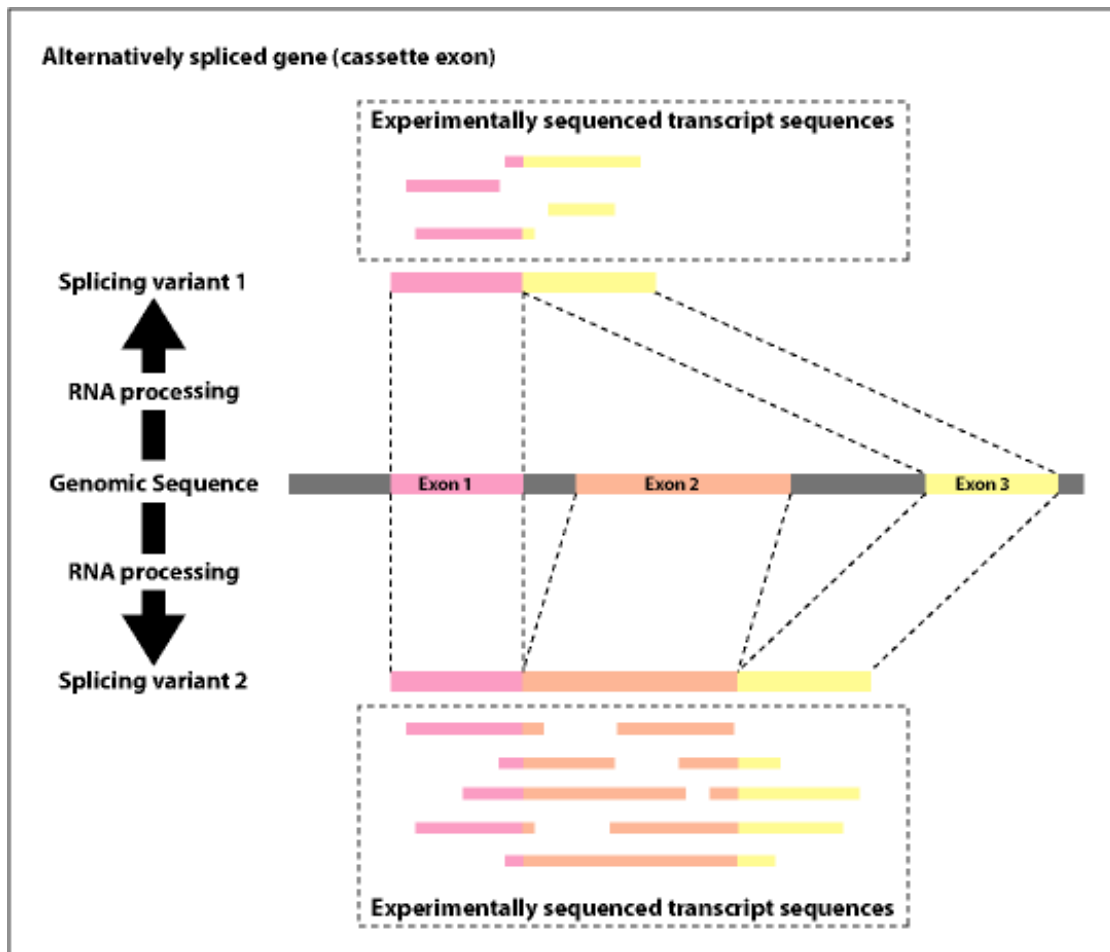


Figure 8. Schematic diagram of the transcripts produced by an alternatively spliced gene, which in this case exhibits a cassette exon event (exon 2 is omitted in splicing variant 1). Transcripts from splicing variant 1 will not align perfectly with transcripts from splicing variant 2.

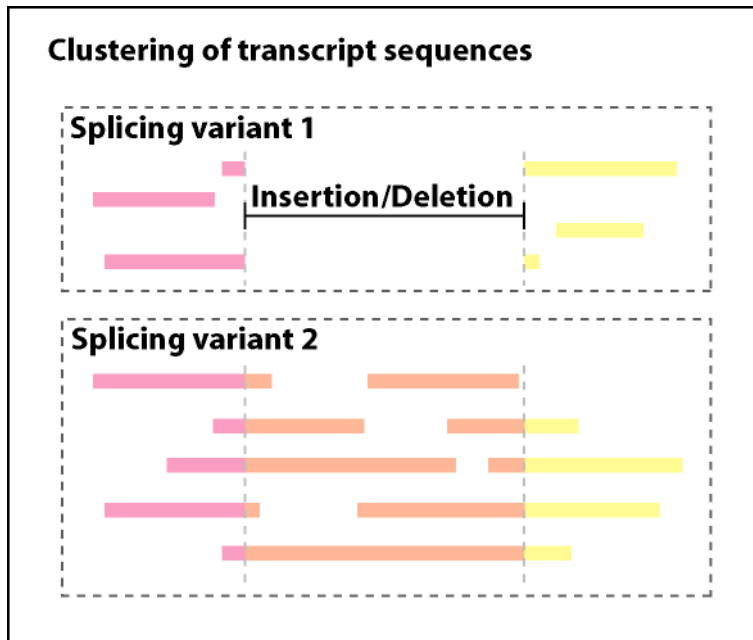


Figure 9. Schematic diagram of the effects of clustering of transcript sequences from splicing variant 1 and splicing variant 2 shown in Figure 8. Depending on which is the reference sequence, a deletion or insertion will occur in the alignment.

1.5.4.4 Alignment of transcripts to genome

Alignment of transcript sequences to the genomic sequences is yet another method of obtaining alternative splicing information. The effect of aligning the transcript sequence to the genomic sequence is that gene structure information is revealed, specifically the exon and intron arrangement. Knowledge of the exon and intron arrangement allows for the detection of alternative splicing. Should a single gene have transcripts producing different exon and intron arrangement, then this can be taken as evidence of alternative splicing. The fact that the transcript sequences align to a particular position on the genomic sequence can be used as proof that they originate from the same gene. Thus, the task is simply to identify at each genomic position for variants in the gene structures produced by transcript to genomic sequence alignments. Both global and local alignment methods can be utilized to perform the alignments that determine the exon and intron arrangement.

Should global alignment methods be employed, these methods should be able to account for the large gaps caused by introns. In view of this, the standard gap penalties have to be modified. Furthermore, additional constraints have to be placed on the alignment procedure to delineate exons such that the flanking splice site motifs are canonical. This is due to the fact that there could be multiple possible exon boundaries with identical alignment scores and biologically the exons are flanked by splice site motifs that are usually canonical.

Local alignment methods benefit from the fact that there exist heuristic local alignment methods like BLAST that allow for rapid local alignments. For local alignment methods, each exon would form a local alignment with the genomic sequence as the exons are interrupted by intronic sequences preventing

an alignment that stretches from the start to the end of the transcript. Due to the nature of local alignment, each subsequence of the transcript sequence could potentially produce several local alignments and thus there exist a need to select a series of local alignments that covers the length of the transcript. Furthermore, the series of local alignments have an additional constrain in that the region of the genomic sequence covered by these local alignments should be ordered from 5' to 3' due to the co-linearity of both transcript and genomic sequence. In practice, dynamic programming is used to select for such a series of local alignments. Should the coverage of the transcript be incomplete due to gaps between the local alignments, then these gaps need to be filled, either through extension of the flanking exons or the identification of a new local alignment between the flanking exons. Incomplete coverage of the transcripts at both the 5' and 3' end is usually acceptable as the ends of the transcripts are usually of low quality and the presence of natural elements such as the poly(A) tail prevents alignment. Once the series of local alignments covering the transcript is obtained, an overlapping of the local alignment has to be resolved to correctly delineate the boundaries of the exons. This is facilitated by the fact that exons are usually flanked by canonical splice site motifs. Any program that determines the gene structure through the use of local alignments will have to delineate the boundaries satisfying both the need to have a good alignment score and a canonical splice site. The final alignment at this point should provide the necessary gene structure information required for alternative splicing detection.

The use of sequence alignment to determine the gene structure is hampered by the presence of paralogs and pseudogenes. In the case of paralogs, sufficient sequence similarity might cause the paralog transcripts to be

aligned to another paralog's genomic location and vice versa. In the case of pseudogenes, transcripts from the real function gene might be aligned to the genomic location of the pseudogene. In many cases, these problems can be resolved by selected for the best alignment.

The use of experimentally derived genomic and transcript sequence information (cDNA and ESTs) to extract alternative splicing information means that the results are relatively reliable, hampered by the quality of the sequences. Furthermore, as the sequencing efforts continue, there would be more data that will allow the detected of rare splice isoforms.

1.5.5 Objectives

In view of the increasing amounts of primary sequence information and genome annotation, this thesis attempts to use this information to understand the process and effects of alternative splicing on a genomic level. However, prior to any actual work on understanding the nature of alternative splicing, there is a need to obtain a set of clean data that is represented in a form suitable for analysis. Therefore, the objectives of this work are as follows:

- Selection and improvements on an alternative splicing data representation
- Creation of a clean set of alternative splicing data suitable for analysis
- Analysis of a clean set of alternative splicing data to gain insights into the process and effects of alternative splicing on a genomic level
 - General statistics
 - Exon and intron length
 - Exon number
 - Nucleotide composition

- Splicing motif analysis
- Domain boundary analysis
- Gene Ontology partitioning
- Effects of alternative splicing on coding sequences and protein domains

The second chapter addresses the alternative splicing data representation problem by selecting and utilizing a data representation that is appropriate for analysis. Most databases on alternative splicing represent the various splice isoforms as a form of multiple sequence alignment, which is not ideal for further manipulation and analysis. The main problems associated with this form of representation is the redundancy involved in representing multiple splice isoforms, each containing a subset of the sequences that make up the gene and the additional computational requirement required in the comparison of the splice isoforms. Therefore, there is a need to use a representation that is capable of reducing the redundancy in the representation that will then serves also to reduce the computational needs. The data representation will also have implications in the visual representation of alternative splicing which is critical in allowing users to make sense of the types and implications of alternative splicing. After the problem of data representation is resolved, the selected data representation can then be used as a basis to classify the various types of alternative splicing. This will allow for a finer granularity in the analysis to follow.

With the selection and completion of a classification system for alternative splicing, the focus is now on the creation of a set of clean high quality data for analysis (third chapter). Since most of the current alternative splicing databases provides data on higher eukaryotes such as human and mouse, this thesis will

focus its attention on lower eukaryotes such as yeast, worm and fruitfly that still exhibit alternative splicing. Lower eukaryotes do exhibit alternative splicing but there are known characteristics like the size of the introns in relation to the exons that differs from higher eukaryotes. Preference will also be given to organism with a sequenced genome that contains high quality annotations. This will serve as the basis for a genomic level set of high quality alternative splicing data.

Finally, with the attainment of a set of high quality data, analysis of alternative splicing can then proceed in chapter four. This introductory chapter has examined some of the evidences for the mechanism of alternative splicing from wet laboratory studies indicating the well-regulated and controlled nature of alternative splicing. This therefore implicates that there are signals within the genes that guide alternative splicing and analysis will thus be done to examine the characteristics of the genes involved in alternative splicing. The elucidation of these characteristics holds promise for the creation of *in silico* methods aimed at detecting alternative splicing. Another aspect of alternative splicing covered in this chapter is the impact of alternative splicing on the cell. The fourth chapter will also seek to understand these impacts thereby allowing us to understand the significance of alternative splicing in the cellular context.

Chapter 2: Data representation and visualization of alternative splicing

2.1 Introduction

The data representation used for studying alternative splicing is fundamentally important as it impacts the type of analysis and the ease by which it is carried out. Therefore the selection of a suitable data representation is critical to this work and efforts have been taken to ensure that the data representation chosen would not restrict but aid in the analysis to be undertaken.

Traditionally, data structures used to represent transcript information were used for alternative splicing. A set of these data structures can be used to represent the various transcript isoforms. By comparing the various transcripts, various types of alternative splicing events can be determined. This approach has the advantage that it is quite simple to represent transcripts using a data structure. For example, a two dimension array as shown in Figure 10 can easily be used to represent the start and end genomic positions of a set of exons that make up the transcript. The minimum amount of information required is the start and end genomic position of each exon (however in most cases, there will be far more information available). The use of the genomic positions or nascent transcript positions is important, as the introns are not present in the mature transcript. The use of the mature transcript positions would make comparisons between the various transcripts impossible.

A set of these data structures can then be used for alternative splicing as shown in Figure 11. Alternative splicing results in multiple transcripts or splice

variants each being different from each other in some ways. Each splice variant can be represented by a single two-dimension array. A set of these representations can then be used to determine the type of alternative splicing present in the gene. The differences between the various splice variants can be located and classified into various types of alternative splicing. The example provided in Figure 11 shows a case of a cassette exon, which can be picked up by determining whether an exon is only present in some splice variants and not others.

The visualization of such a data representation is also quite simple. One can easily represent this data representation visually in the form of a multiple sequence alignment or a schematic diagram as shown in Figure 12. There is need to represent the nascent transcript instead of mature transcript as the intron information is lost in the mature transcript. Typically, the genomic sequence is also drawn to inform users that the transcript representation is that of the nascent transcript. The most complete set of information is present in the multiple sequence alignment. However, most of the information is not necessary or useful in the visualization of alternative splicing events. For example, sequence information is generally not required for the entire length of the sequence. Most users are concerned with only the splice motifs. Furthermore, the presence of the sequence information makes the representation far more complicated looking and distracts one from the determining the overall picture of the types and occurrences of alternative splicing events. Therefore, by removing the sequence information, one gets the schematic representation, which is a lot easier to visualize for alternative splicing events. This allows one to quickly determine the types and frequency of the alternative splicing events. The drawback to this

approach is that when the number of transcripts gets large, the ability to infer the various alternative splicing events becomes more difficult as illustrated in Figure 12C. The alternative splicing event namely a cassette exon is the same in both Figure 12B and Figure 12C, however it is much more difficult to see the event when large numbers of transcripts are present. This is further compounded by the fact that some of the transcripts are present in a partial form due to the limitation of sequencing technology. This can be alleviated in part by the selection of a representative transcript that indicative of a single splice isoform as performed by ASD (Stamm et al., 2006; Thanaraj et al., 2004b). However, for genes with many splice isoforms, this is still insufficient.

Not only does the visualization get more difficult, but the computation to determine the forms of alternative splicing gets increasing more intensive as there are far more data for comparison. The presence of partial transcripts as would be present due to the limitations of sequencing technology means that not only does the computation gets more intensive but more complex as the algorithm has to account for the partial sequences. The complex nature of the algorithm also means that the likelihood of errors in the assignment of alternative splicing events also increases.

	1st dimension	
2nd dimension	1	100
	150	330
	450	710
	900	1020
	1500	1710

Figure 10. Two dimension array representation of a set of exons that make up a transcript. A transcript can be simply represented as a series of exons, each exon having a start and end position. The first dimension in the array has two elements, the start and end position respectively while the second dimension represents the number of exons that make up the transcript.

Splice variant 1			Splice variant 2		
	1st dimension			1st dimension	
2nd dimension	1	100	2nd dimension	1	100
	150	330		150	330
	450	710		450	710
	900	1020		1500	1710
	1500	1710			

Figure 11. Alternative splicing information represented by a set of transcript. Splice variants can be individually represented as two dimension arrays. A set of these arrays can then be used to determine the type of alternative splicing present. Splice variant has an addition exon having the start and end position of 900 and 1020 respectively. This is missing in splice variant 2. Therefore the conclusion is that the exon starting at position 900 to 1020 is a cassette exon.

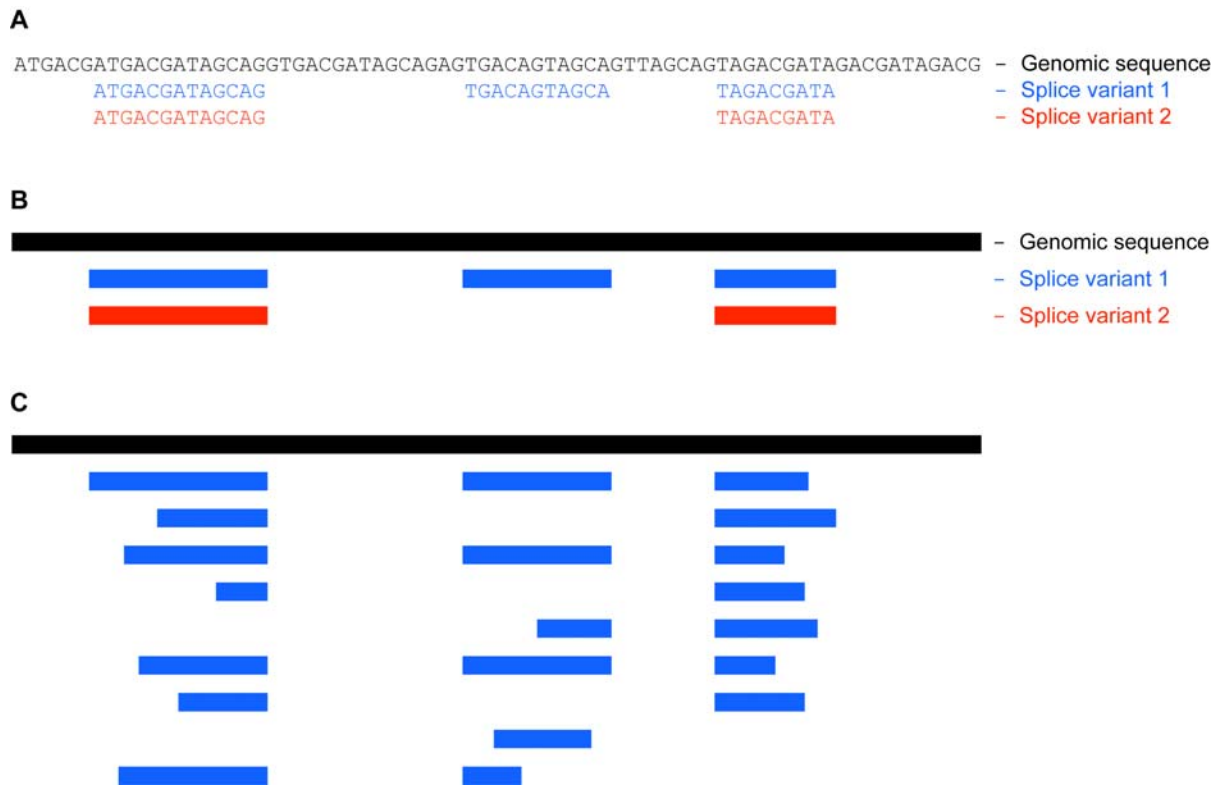


Figure 12. Visual representation of transcripts that contain alternative splicing events. (A) Multiple sequence alignment visual representation provides sequence information that is otherwise unavailable in the schematic view. Two transcripts are shown with splice variant 2 missing exon 2. This is clearly a case of a cassette exon. (B) The same information in the multiple sequence alignment represented as a schematic visual representation. Sequence information is lost but better overall view of the forms of alternative splicing is possible due to a lack of distraction by the sequence information. (C) When the number of transcripts available increases, it becomes more difficult to determine the overall view of the types of alternative splicing. The visual representation here also shows a case of a cassette exon. However, the number of different transcripts that constitutes this makes the determination more difficult.

In view of these limitations, splicing graphs were proposed by Heber *et al.* in 2002 (Heber, Alekseyev, Sze, Tang, & Pevzner, 2002). He propose that the splice variants of a particular gene be denoted by $\{t_1, \dots, t_n\}$ and that each splice variant be a set of genomic positions v_i where $v_i \neq v_j$ for $i \neq j$. Therefore, a set of all the transcribed genomic positions V is simply a union of all v_i as in $V = \bigcup_{i=1}^n v_i$. The splicing graph G is then a directed graph on the set of transcribed genomic positions V that contains a set of edges (v, w) where v and w are consecutive positions in any of the splice variants t_i . Each of the splice variants t_i is then a path in the splicing graph G . Therefore a splicing graph that has only one path does not exhibit any alternative splicing. To reduce the complexity of the graph, consecutive vertices having 1 indegree and 1 outdegree is merged together. This makes sense as large numbers of the genomic positions are not involved in alternative splicing and thus this greatly reduces the complexity of the graph. Any genomic position that is involved in alternative splicing will have either >1 indegree and/or >1 outdegree. This data representation is capable of storing all the information inherited in the original set of splice variants while providing a basis for further analysis. As the genomic positions are represented only once and not multiple number of times as in the previous traditional approach, the computation intensity as well as the complexity of detecting and classifying alternative splicing is far simpler.

The splicing graph representation has been adopted in this work due to the ease by which the determining of alternative splicing can be carried out. In addition, further analysis can be carried out easily on such a data structure. However in view of the fact that most transcript information is readily available in the form of start and end genomic positions, we have redefined the vertices as

unique exons having a unique set of start and end genomic positions instead of a single/merged genomic position. This reduces the computation required for graph construction by having far less initial vertices as well as eliminates the merging of vertices step.

The splicing graph provides a condensed view of the various splice variants of a gene and this also makes it suitable for visual representation. Unlike the traditional approach of visually representing individual splice variants, the splicing graph collapse all the splice variants into a single directed graph that is visualized using a single linear representation as shown in Figure 13. Any bifurcation in the splicing graph is indicative of alternative splicing. By reducing the number of elements on the visual representation, the various alternative splicing events are more pronounced and easier to pick up. Furthermore the effects of the alternative splicing are also clearly known from visual inspection of the graph.

Although the splicing graph approach has been adopted in this work, the visual representation like the graph construction has been altered mainly to make the effects of the alternative splicing more visible as shown in Figure 14. The visual representation differences are inherited from the differences in the underlying data representation. Therefore, all the unique exons are represented visually and the lines connecting them are the introns. By making all the exons visible, alternative splicing events are more prominent. Furthermore, the presence of all the unique exons provides opportunity for the design of user interfaces where each of these exons may trigger events most likely leading to details about the exon in question.

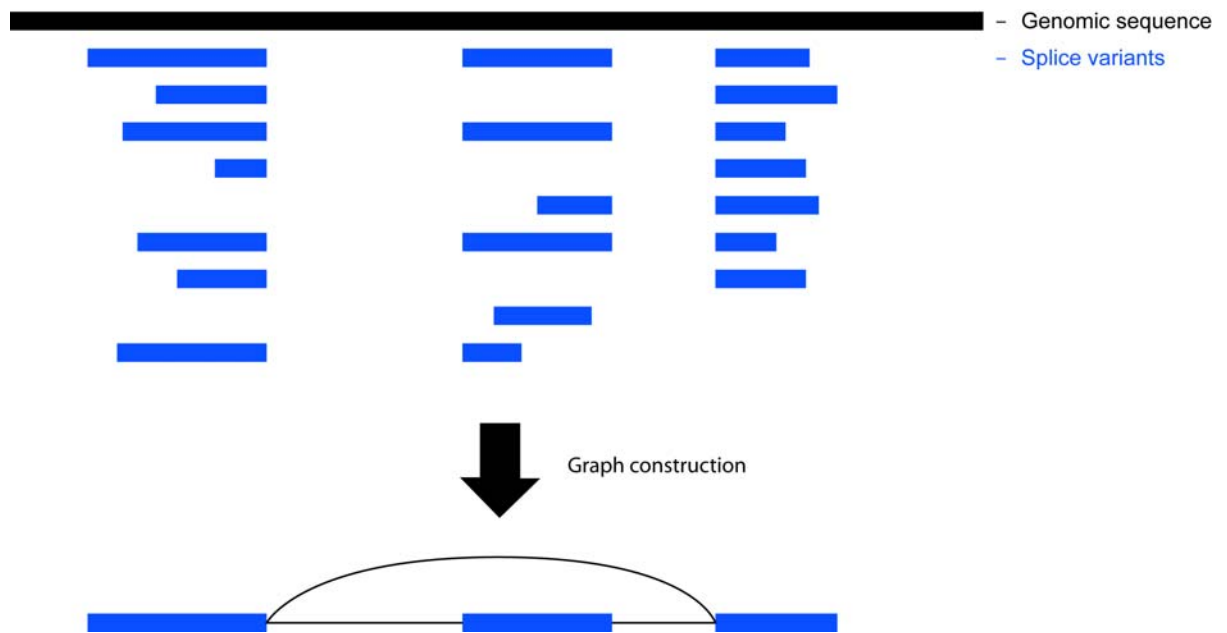


Figure 13. Splicing graph representation of alternative splicing events. The numerous splice variants are used in the graph construction process to produce a splicing graph that is a condensed view of all the splice variants. The splicing graph is clearly easier to interpret as any bifurcation in the graph is suggestive of alternative splicing.

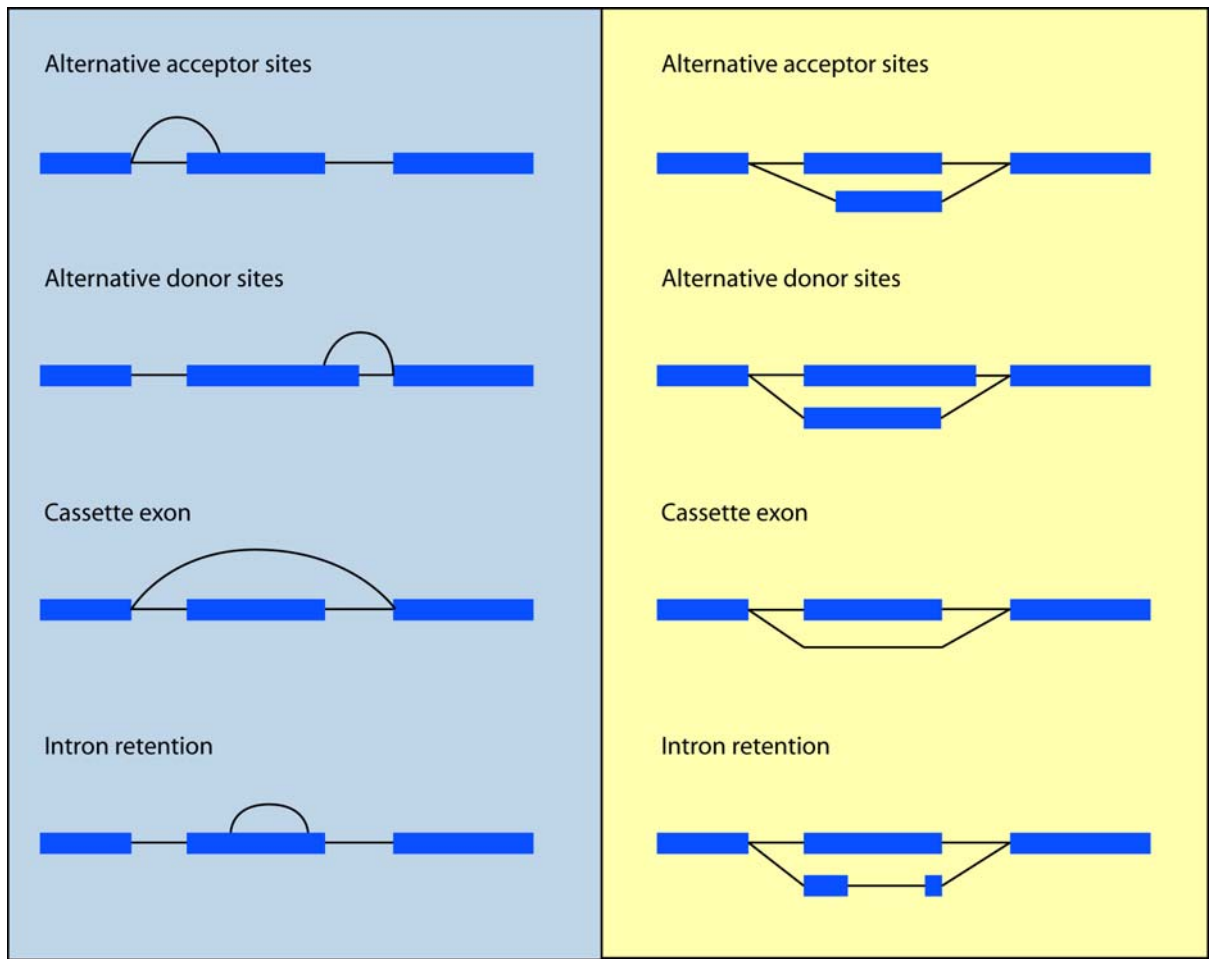


Figure 14. Our implementation of the visual splicing graph representation. The representation proposed by Heber *et al.* is shown on the left with a blue background while our implementation is shown on the right with a yellow background. The main difference is that each unique exon is represented making the alternative splicing events more visible.

One of the strengths of the splicing graph representation is that the detection and classification of alternative splicing events is quite simple. Simple yet robust rules can be created that allows for the detection of multiple types of alternative splicing events within the same splicing graph.

2.2 Implementation

2.2.1 Construction of splicing graph

The construction of the splicing graphs requires at the minimum a set of transcripts with known exon start and end genomic positions. This is usually obtained by alignment of transcript sequences into the genomic sequence thereby revealing the genomic positions of individual exons. Transcripts can have either a single exon or multiple exons. Therefore, exons can be classified into four types, single exons, initiation exons, internal exons and termination exons. Single exons are exons where the transcript contains just a single exon and therefore, both start and end positions are not reliable in the sense that they may not have been completely sequenced. Initiation exons are the first exon of a transcript and its start position is unreliable. Internal exons are exons that are not the first or the last of a transcript. Its start and end position is reliable as the limitations of sequencing do not fall on them. Termination exons are the last exons of a transcript and its end position is unreliable. From the set of transcripts, one or several splicing graphs will be generated. Each splicing graph will consist of a set of transcripts that share at least one exon. Much like the exons, vertices in the splicing graph can be classified into three forms, initiation, internal and termination. Initiation vertices have no indegree as they are corresponds to the initiation exons. Internal vertices have both indegree and outdegree as they are

flanked by other exons. Termination vertices have no outdegree as they are the termination exons. These vertices have the same level of reliability regarding their start and end positions as their exon equivalents.

The first step involves reading in of the transcript information, the minimum being a set of exons having the start and end genomic positions. These transcripts are then divided into two sets one for the sense strand and one for the anti-sense strand. These two sets are processed separated using the same processes. The transcripts in each set is first grouped together on the basis that they overlap one another based on the start and end genomic positions of the transcript. Then splicing graphs are generated within each group by checking for transcripts that share common exons. The purpose of the initial groups by the start and end transcript genomic position is to reduce the number of comparisons as comparisons of all the transcripts for common exons would be computational expensive. By placing transcripts into smaller groups, fewer comparisons are required within the group and this translates to fewer comparisons in total saving computational time.

Comparisons with the groups starts by making one of the transcript the initial splicing graph. Each of the remaining transcripts are then in turn compared to the splicing graphs generated within the group to determine if they belong in any particular splicing graph. This involves checking whether any of the exon in the transcript is the same as any of the vertices in the splicing graph. This check takes into account that the initiation and termination exons of the transcripts may be incomplete due to the limitations of sequencing being unable to generate full-length transcript sequences. These types of exons are termed overlapping exons and are defined as having one of its positions undefined or unclear (for initiation

exons, the undefined position is the start position and for termination exon, the undefined position is the end position). These exons merely require defined position to be matched to be part of a splicing graph. Exons that are internal to the transcripts are termed shared exons and require both start and end positions to be matched to be part of a splicing graph. Should the transcript not be part of any splicing graph, then a new splicing graph is created with the transcript's exons as the initial vertices.

The final set of splicing graphs generated from the set of transcripts is the summation of the various splicing graphs in all the initial groups. Different types of alternative splicing events can then be detected and classified in each of these splicing graphs. This construction process is visually depicted in Figure 15.

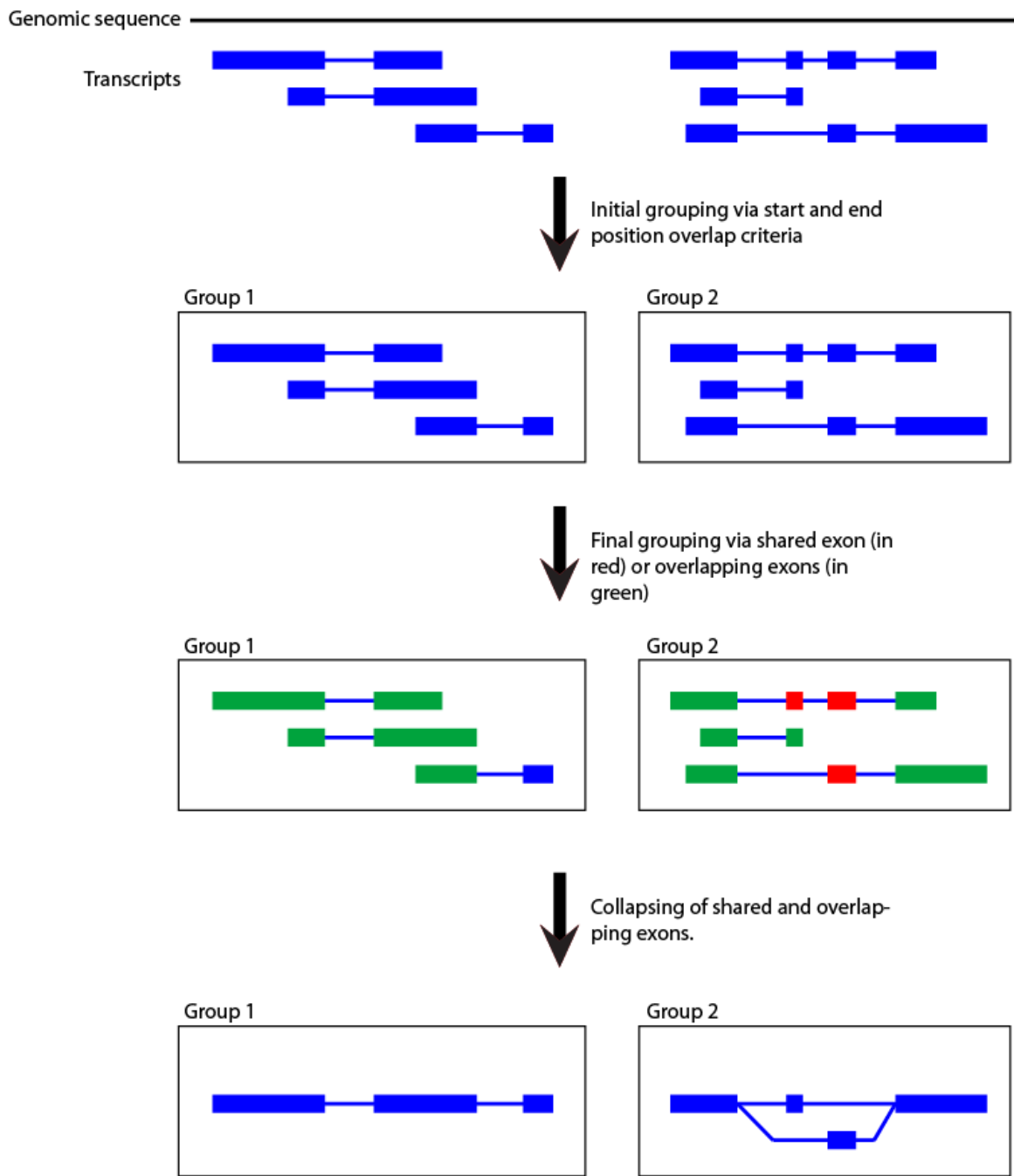


Figure 15. Construction of splicing graphs. Initial groups of transcripts are found by comparing the start and end positions of the transcripts, this resulted in two groups in this figure. Overlapping exons (in green) and shared exons (in red) are defined and used to perform the final grouping. Once the final grouping is completed, overlapping and shared exons are collapsed into a single unit forming the final splicing graph.

2.2.2 Detection and classification of alternative splicing events in the splicing graphs.

Alternative transcriptional start sites (TSS) exists when more than one node is found that has no previous connection (indicating that it is an initiation node) and which contains a unique start position. The start position of initiation nodes is the assumed to be the transcriptional start site and multiple TSS implies alternative transcriptional start sites. However due to the nature of sequencing, the ends of the sequenced transcripts may not reflect the true biological ends, therefore care must be take in the interpretation of any results on alternative transcriptional start site.

Alternative transcriptional termination sites (TTS) occur when more then one node is found that has no next connection (indicating that it is a termination node) and which contains a unique end position. The end position of the termination node is the transcriptional termination site and the multiple TTS indicates alternative transcriptional termination sites. Much like the condition for alternative TSS, the ends denoting the TTS should be interpreted with care.

Alternative initiation exons occur when multiple initiation nodes (having no previous connections) are found to that have unique end positions. The rationale is that the start positions of initiation nodes are frequently incorrect as the 5' UTR (untranslated region) is rarely completely sequenced. Therefore, initiation nodes differing just in the start position cannot be easily determined to be different. We have thus also used the end position as the criteria. Furthermore, the 5' end of the initiation node is not recognized by the splicing machinery, only the 3' end (donor site) is recognized. The same reasoning goes for alternative termination exons except that the positions are reversed.

As for alternative acceptor sites, these are found in a set of overlapping nodes (>1 node) that have differing start positions linked to a common node. The set of nodes should be overlapping else they would be classified as cassette exons. The same goes for alternative donor sites.

Cassette exons by definition are internal exons that are differentially included in the various splicing isoforms of a gene. The rule as far as splicing graphs are concerned requires a cassette exon to be an internal node whose start and end position falls within a connection (an intron). The fact that the node occurs as part of a connection in some other splicing isoform implies that it is skipped hence fulfilling the definition.

Intron retention on the other hand are introns which are not spliced out resulting in it being retained forming part of an exon. The rule based on splicing graphs requires a connection whose start and end position falls within a node for a positive intron retention event. The definition is fulfilled as the connection (intron) is found as being part of a node (exon).

2.2.3 Splicing Graph Module (SGM)

A Python module was written to allow for the creation of splicing graphs. The module also implements the classification rules described in the previous section. The module also allows for the visualization of splicing graph via a graphics file output.

The Splicing Graph Module (SGM) was written in an OOP (Object Oriented Programming) approach. The various UML (Unified Modeling Language) class diagrams of the various classes in the module are illustrated in Figure 16, Figure 17, Figure 18 and Figure 19. The module is organized in such a manner that there exist classes that describe the basic objects required for the

splicing graph data representation. Additional visualization oriented classes are then sub-classed from the basic classes in order to reduce code redundancy and aid code reusability.

The classes in the file `Genome.py` (see Figure 16) provide classes used to represent a transcript that is made up of a number of exons and introns. The most primitive class is `GEntity` which describes a class that contains a start and end position. The classes `GExon` and `GIntron` then inherit this class to include an additional sequence attribute. In order to allow for annotations of the transcript sequence, an additional class entitled `GFeature` is coded to represent any annotation on the transcript. The main attribute of this class is an array (or list in Python grammar) of `GEntity` used to denote the range of the annotation. These three classes namely, `GExon`, `GIntron` and `GFeature` are then used in the `GTranscript` class to present a transcript. All the above mentioned three classes are represented in `GTranscript` as arrays of objects. In addition to these three classes, the `GTranscript` class contains the start and end of the CDS (coding sequence) as well as the transcript orientation. An ID attribute is also provided as an identifier for the transcript.

The classes that capitalize on the above mentioned classes using inheritance for the visualization of a transcript lies in the Python file `DrawGenome.py` (see Figure 17). Visualization is achieved using 2D (two dimension) graphical files that are the final result of the classes contained in this Python file. 2D graphics uses a coordinate system similar to the 2D Cartesian coordinate system and the Python graphics module used (`ReportLab`) has the origin of the coordinate system at the upper left hand corner of the image. As most of the elements used in the visualization of the transcript as well as the

splicing graph can be simplified as lines of a certain thickness, a class entitled GPosition was created to hold the information required to describe one point of the line. A line can then be described using two GPosition instances. Each GPosition instance contains the x-axis position as well as the top and bottom y-axis positions used to describe the thickness of the line. Arrays of GPosition instances are then used by both DrawGExon and DrawGItron classes to correspond to the visual representation. The DrawGExon and DrawGItron classes inherit from the GExon and GItron classes and in addition to the arrays of GPosition instances contain color codes for drawing. A new specialization of the GTranscript class as DrawGTranscript uses the new subclasses DrawGExon and DrawGItron to represent the information required for visualization as well as containing new methods dedicated for visualization. In order to transform the genomic positions to the graphics coordinate system positions, a new class entitled GMapping is created that allows for the mapping of genomic positions to the graphics coordinate system positions. Each GMapping specifies one mapping and a GMappings class is created to encapsulate an array of GMapping that can be used to mapping any genomic position to a graphics coordinate system position.

Since splicing graphs are built using transcript information. The splicing graph related classes in the Python file Graph.py (see Figure 18 and Figure 19) uses the classes defined in Genome.py and DrawGenome.py to represent and produce visual representations of splicing graphs. The splicing graph data representation is achieved by the Graph class (see Figure 18). The Graph class contains arrays of GTranscript, Node and Connection instances. The Node class represents a vertex in the splicing graph and would be equivalent to a unique

exon. The Connection class represents an edge in the splicing graph and would be equivalent to a unique intron. The Graph class is implemented as a double linked list with alternating Node and Connection instances in the list. This means that a Node instance knows the previous and next Connection instance and a Connection instance knows the previous and next Node instances. In certain cases, it is more useful for the Node instance to be able to determine the next and previous Node instances and therefore the Node class also provides for this. The pool of GExon and GIntron contains the GTranscript instances that make up the Graph instance may not be unique and therefore Node and Connection instances may contain multiple instances of GExon and GIntron implemented as list of objects. Both the Node and Connection classes inherit from the GEntity class as they both require the presence of a start and end position, which is provided by the GEntity class. All the alternative splicing event classification rules are implemented in the Graph class as a series of methods, each detecting for one type of alternative splicing event.

Visualization of the splicing graphs is then attained via specialization of the Node, Connection and Graph classes as DrawNode, DrawConnection and DrawGraph classes (see Figure 19). Much like the DrawExon and DrawIntron classes, the DrawNode and DrawConnection classes contain arrays of GPosition as well as color code information to represent the visual aspect. In a similar fashion as that of the GTranscript class, the DrawGraph class contains specialized methods for visualization such as the calculation of the graphic coordinate system positions from the genomic positions.

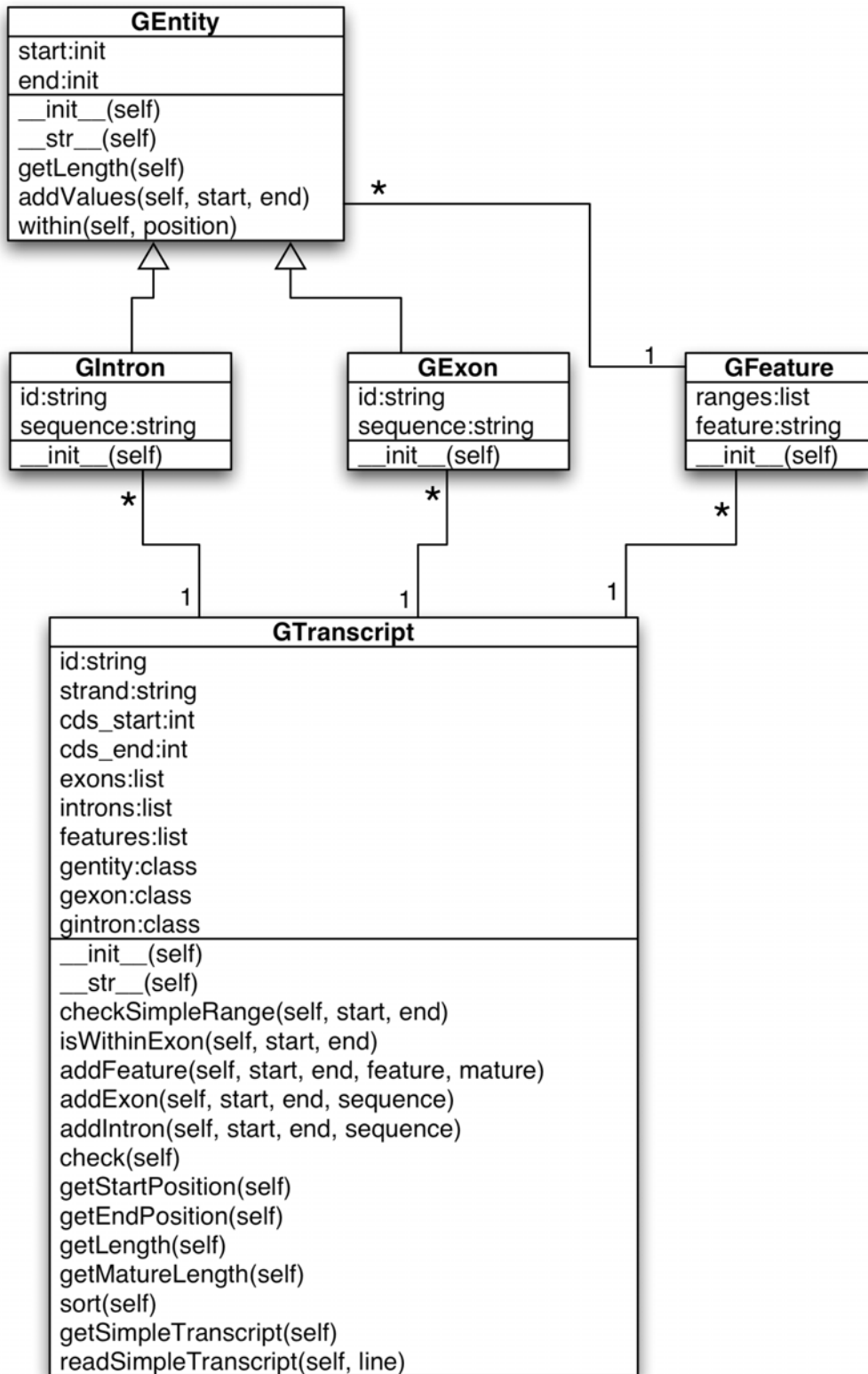


Figure 16. UML class diagram for the classes in the file Genome.py.

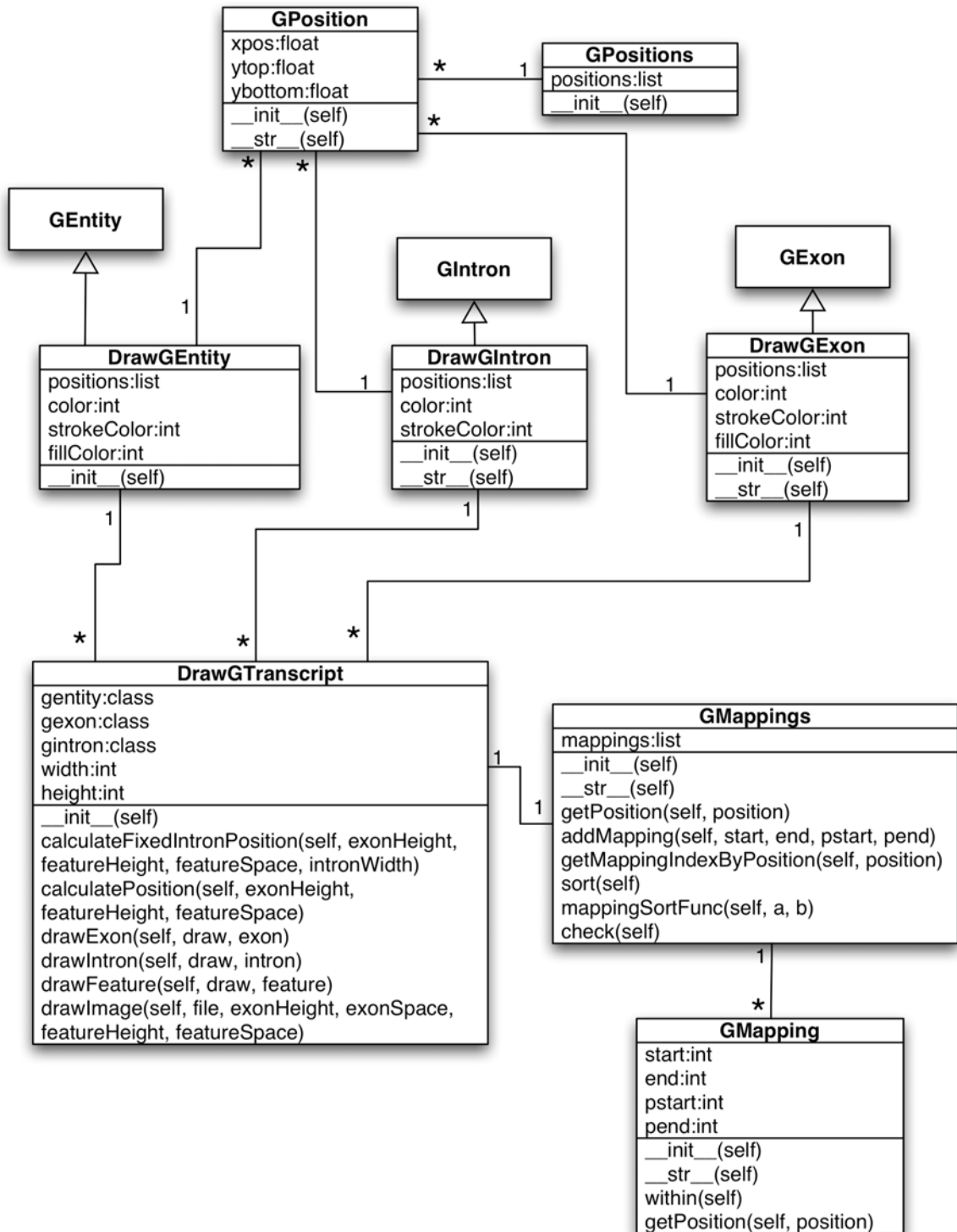


Figure 17. UML class diagram for classes in the file DrawGenome.py.

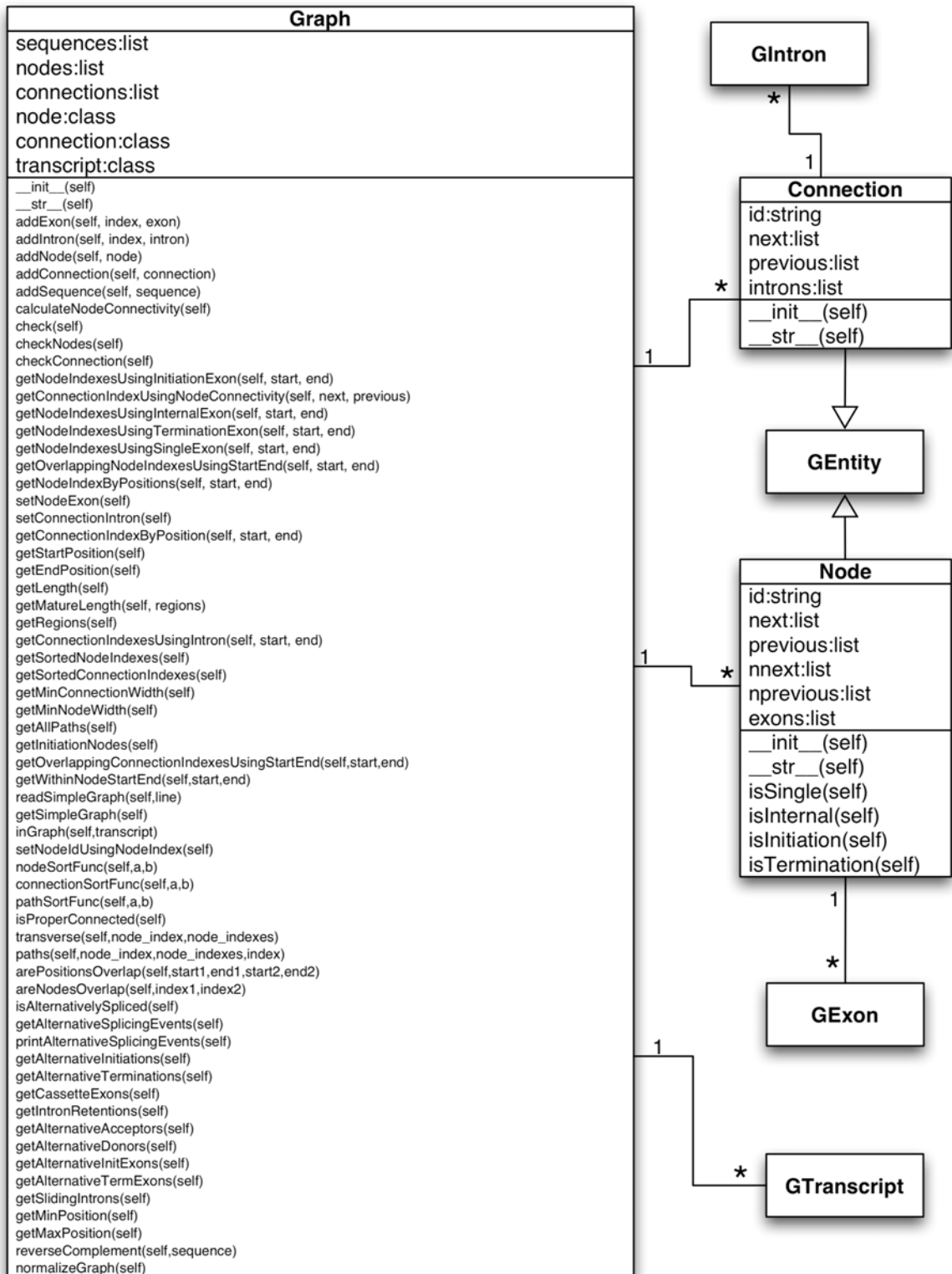


Figure 18. UML class diagrams for the data representation related classes in the file Graph.py.

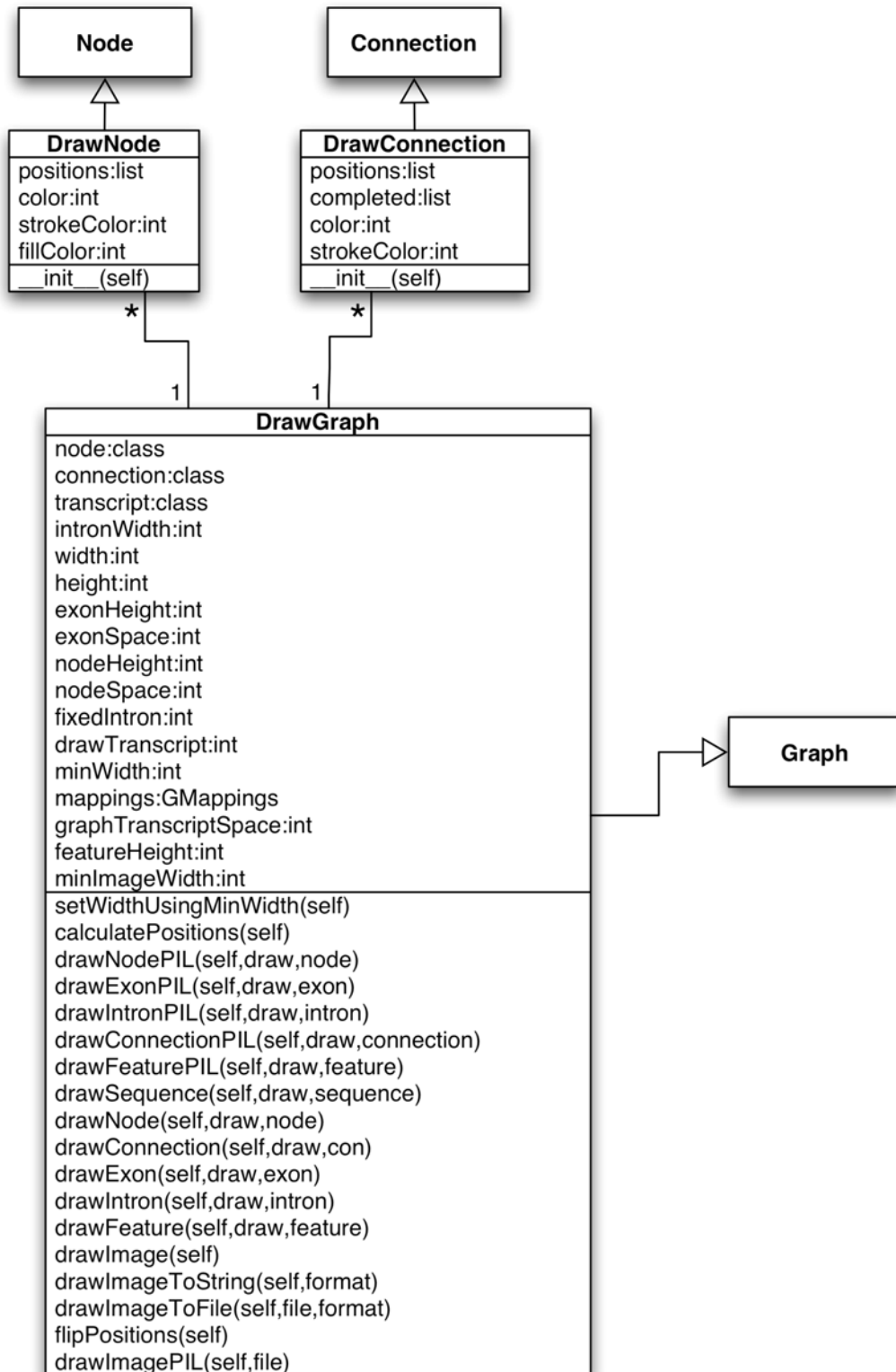


Figure 19. UML class diagrams for the visualization related classes in the file Graph.py.

The generation of splicing graphs using SGM requires as input transcript information that can be provided in GFF format or in a tab delimited format. Functions are available in Graph.py to read these formats and return a list of GTranscript instances. The list of GTranscript instances are then be provided as input to the function clusterSequences(sequences, graph) which will generate a list of splicing graphs as Graph instances.

The availability of these classes contained in SGM allows for the ease of creation of any application that requires the use of splicing graphs. Included in the distribution of SGM is a script entitled draw_graph.py capitalizing of the OOP nature of SGM to generate splicing graphs and its visual representation as graphics file from a GFF (Gene Finding Format or General Feature Format) file containing transcript information. Another script entitled example.py shows how to subclass the classes in SGM to provide for specialized behaviors.

SGM is distributed as a Python module packages using distutils to provide for ease of installation. SGM requires in addition to a recent version of Python (version 2.3 was used for development), ReportLab 1.19, Python Imaging Library 1.1.4 and RenderPM. These libraries are freely available. The module can be downloaded from <http://proline.bic.nus.edu.sg/sgm/download.html>.

2.2.4 SGM web service

In addition to the Python module SGM, work was done to create a web service also entitled SGM (Figure 20) that would allow users to easily create splicing graphs without the need to perform any programming. This web service is freely available at <http://proline.bic.nus.edu.sg/sgm/index.html>. The web service requires as input a GFF file that provides transcript information shown in Figure 21. The transcript information is read from the GFF file provided by the user and

the classes in SGM are then used to generate the splicing graphs from the GFF file. Immediately after submitting the GFF file, the user is provided with a results ID that can be used to retrieve the results at a later date. The results provided by SGM consist of a list of splicing graph entries as shown in Figure 22. Each splicing graph entry shows the transcripts that were used to construct the splicing graph. The number of nodes (unique exons) in the splicing graph is also shown together with the number of possible paths through the splicing graph. Any number of paths greater than one indicates the presence of alternative splicing events. By clicking on the link provided in each entry, the user will get a new HTML page showing detailed information about the splicing graph in question. This information shown in this page (Figure 23) includes details of each node in the splicing graph as well as details of each transcript used to construct the splicing graph. Links are also provided to graphic files containing the visual representation of the splicing graph as shown in Figure 24. The visual representation of the splicing graph consists of the splicing graph shown at the top with the transcripts used to construct the splicing graph shown below the splicing graph. The black colored bars in both the transcript and splicing graph visual representation show the exons while the green colored lines show the introns. The splicing graph view together with the transcripts shown below provides the users with a very complete view of the alternative splicing events. Two different formats of graphics files are provided. The first being the PNG (Portable Network Graphics) format which is directly viewable in most web browsers. Many applications on various platforms have no problem using PNG files and thus compatibility is not an issue. The PNG format is raster in nature and thus does not scale well. To compensate for this weakness, PDF (Portable

Document Format) files are also available for download. In addition to being scalable to any size without lost in resolution, PDF is also a vector-based format that allows for easily manipulation of the splicing graph visual representation. This can be easily done in a software package like Adobe Illustrator or Macromedia Freehand. This enables users to customize the splicing graph for their specific purpose without the need to do any programming.

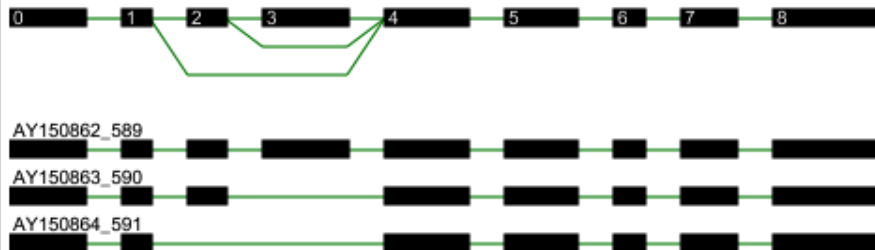
SGM : Splicing Graph Modules



[Home](#) | [Create](#) | [Results](#) | [Download](#) | [Help](#)

Introduction

SGM (Splicing Graph Modules) is a collection of Python modules housed within a Python package to draw splicing graphs (a form of visual representation of several transcript sequences which may exhibit alternative splicing). An example of the splicing graph that is drawn by SGM (which includes schematic representation of the transcripts that make up the splicing graph) is shown below.



The use of splicing graphs as a visualization tool allows users to quickly and easily make sense of the alternative splicing. Traditional approach represents the transcripts schematically in a multiple sequence alignment like layout (much like the lower half of the example above). This traditional approach requires much effort on the users to make sense of the alternative splicing taking place. In the event where there are large numbers of transcripts, this is nearly impossible.

The Python package allows users to produce splicing graph representation of their data. The package consist of several object oriented classes which users can override or specialize to produce custom splicing graphs.

For users who do not want to bother with any programming, a Python script within the package is available that allows images of splicing graphs to be generated using a series of transcript information as input.

A web service is also available for users to generate images of splicing graphs. This is especially useful for users who do not want to install the package.

The images generated can either be raster (PNG format) or vector (PDF). The vector format is especially useful as it can be scaled to any size and is easily editable using a vector graphics program like Adobe Illustrator.

Contacts

[Prof. Shoba Ranganathan](#)
[Prof. Tan Tin Wee](#)
[Lee Teck Kwong Burnett](#)

Copyright © 2004 Department of Biochemistry, National University of Singapore

Figure 20. SGM web service homepage.



Create splicing graphs

To create images of splicing graphs, you need to first create a GFF file containing sequence information organized as a series of exon position information. An example showing two sequences (the first having 2 exons and the second having 3 exons) is shown below:

```
AA567111_1_e1 Blat_PSL exon 230650 230989 100 - 0 AA567111_1
AA567111_1_e2 Blat_PSL exon 231342 231455 100 - 0 AA567111_1
AA694710_2_e1 Blat_PSL exon 228267 228336 100 + 0 AA694710_2
AA694710_2_e2 Blat_PSL exon 228397 228731 100 + 0 AA694710_2
AA694710_2_e3 Blat_PSL exon 228819 229038 100 + 0 AA694710_2
```

Each line consisting of several fields separated by tabs representing a single exon. Descriptions of the various fields are as follows:

Field Number	Description
1	Exon ID. Not used in the module.
2	Source of information. Not used in the module.
3	Type of entry. Only lines contain "exon" will be used. All other types are ignored. Therefore it is important that the GFF file contains this field.
4	Start position. The start position of the exon on the genome. This is used in the module and hence has to be correct.
5	End position. The end position of the exon on the genome. This is used in the module and hence has to be correct.
6	Score. Not used in the module.
7	Strand. Used in the module to determine the strand of the sequence. Only the strand of the first exon of the sequence is used. The rest is ignored.
8	Frame. Not used in the module.
9	Comments. The sequence ID is written here. This is very important as the exons of a particular sequence is grouped together on the basis of this ID. Please ensure that this ID is unique for each sequence (I would recommend that you append a unique number to the accession as shown in the example above) and that each exon of a sequence has the same sequence ID.

Click [here](#) for a example GFF file. Visit <http://www.sanger.ac.uk/Software/formats/GFF/> for more information on the GFF format.

After the GFF file has been created, use the form below to submit the sequences for the creation of splicing graphs.

Select your input GFF file:

no file selected

Figure 21. SGM web service input form. The only required input is a GFF file that describes the transcripts required for the generation of splicing graphs.



Results ID: Kd26nYqDiyyc8P1DqYJe

Splicing Graph 1


Sequences : AI405834_22, AY051614_108, BF498840_50, CK658723_103, BF488079_43, BT003800_115, BF489483_45, BF496742_49, AI063542_2, BG633911_63, BF490262_46, BI352641_81, AI944454_27, BI355695_84, BI567480_86, BI576395_87, BI582679_88, BI585989_89, BI586276_90

Number of nodes : 13**Number of paths** : 6Click [here](#) to see details of this splicing graph.**Splicing Graph 2**

Sequences : AY060776_109, BF495471_48, BF499662_51, AI405479_21, AI404321_20, AI295658_15, BF500767_53, BF502816_55, AI388080_19, BF506667_60, BF488727_44, BI240172_78

Number of nodes : 8**Number of paths** : 4Click [here](#) to see details of this splicing graph.

Figure 22. The results of the GFF file submission. A list of all the splicing graphs generated from the transcript is shown (the list is partly reproduced here due to its length). Each splicing graph entry shows the identifiers of the transcripts used to generate the splicing graph, the number of nodes (vertices) in the splicing graph as well as the number of paths in the splicing graph. Any number of paths greater than 1 indicates that there is alternative splicing. By clicking on the link provided for each entry, one can get a detailed look at the splicing graph.

SGM : Splicing Graph Modules 

Home | Create | Results | Download | Help

Results ID: Kd26nYqDiyyc8P1DqYJe

Click [here](#) to return to the results summary.

Splicing Graph Images

Format	In this window	New window
PNG	In this window	New window
PDF	In this window	New window

Node Details

Node Index	Start	End	Next	Previous
0	15665	15857	1	
10	16407	16598	2	
1	16549	16598	2	0
2	17136	17282	3	1 10
3	17341	17572	4	2
4	17787	18023	5	3
12	18775	18985	5	
5	19048	19208	6, 11	4 12
6	19274	19285	7	5
11	19274	19452	8	5
7	19340	19452	8	6
8	19521	20714	9	7 11
9	20787	22324		8

Sequence Details

Sequences ID	Exons
AI405834_22	15665-15857, 16549-16598, 17136-17282, 17341-17572, 17787-17800
AY051614_108	15665-15857, 16549-16598, 17136-17282, 17341-17572, 17787-18023, 19048-19208, 19274-19285, 19340-19452, 19521-20714, 20787-21772
BF498840_50	16407-16598, 17136-17282, 17341-17572, 17787-17922
CK658723_103	17833-18023, 19048-19208, 19274-19452, 19521-19636
BF488079_43	18775-18985, 19048-19208, 19274-19452, 19521-19663
BT003800_115	18775-18985, 19048-19208, 19274-19452, 19521-20714, 20787-21134
BF489483_45	18785-18985, 19048-19208, 19274-19452, 19521-19641
BF496742_49	19048-19208, 19274-19452, 19521-19776
AI063542_2	19134-19208, 19274-19452, 19521-19654
BG633911_63	19189-19208, 19274-19285, 19340-19452, 19521-20077
BF490262_46	19419-19452, 19521-20060
BI352641_81	19580-20181
AI944454_27	20527-20714, 20787-20931
BI355695_84	20846-21182
BI567480_86	21603-22244
BI576395_87	21603-22324
BI582679_88	21603-22176
BI585989_89	21603-22273
BI586276_90	21603-22256

Copyright © 2004 Department of Biochemistry, National University of Singapore

Figure 23. Details of the resulting splicing graph generated by the SGM web service. Links are provided that allows users to download graphic files of the splicing graph. Details of the nodes and the transcripts are also shown.



Figure 24. An example of the splicing graph visual representation produced by SGM. The splicing graph is shown at the top with each exon as a black colored bar labeled with a number. Introns are shown as green colored lines connecting the exons. Immediately below the splicing graph are the visual representation of the transcripts used to construct the splicing graph. Each exon like in the splicing graph visual representation is shown as a black bar connected to each other by introns as green colored lines. The identifier of each transcript is shown just on top of each transcript visual representation.

2.3 Conclusion

The splicing graph data representation was chosen as the data representation as it provides an easy means to present a set of transcript that contains alternative splicing events. The computational requirement for the transformation of a set of transcripts into a splicing graph is quite trivial and the splicing graph representation provides a logical way of organizing the data such that classification and analysis of alternative splicing events is made much simpler.

A Python module entitled SGM was written to take a set of transcript and generate a series of splicing graphs. In addition to the generation of splicing graphs, the module implements the alternative splicing event classification rules that allows for the classification of the various alternative splicing events in the splicing graphs. Furthermore, the module also allows for the visualization of the splicing graph as graphics files. This allows users to make sense of the alternative splicing in an intuitive manner. The packaging of the codes as a module implies that the module can be used by other to generate splicing graph and therefore the module is make available for free to the public.

For non-programmers, a website was created that allows for the creation of splicing graphs using a web interface. The only input required is a GFF file that contains transcript information. With the aid of the SGM module, a series of splicing graph is generated. Graphic files that contain visual representation of the splicing graph can then be downloaded by users for their own use.

The work done in this chapter is then put to use in the next chapter where a clean dataset of alternative splicing events was created using the data representation work done and illustrated in this chapter.

Chapter 3: *Drosophila melanogaster* Exon Database (DEDDB)

3.1 Introduction

The completion of the draft sequence of the *Drosophila melanogaster* genome in March 2000 (Adams et al., 2000; Hoskins et al., 2002) and the availability of quality annotations by FlyBase in 2002 (Misra et al., 2002) presents an excellent opportunity for the study of alternative splicing. This is especially so since most alternative splicing databases focuses on higher eukaryotes. Therefore, the use of *Drosophila melanogaster* represents an opportunity to understand alternative splicing in a lower eukaryote. Although the annotations themselves provide an insight to the amount of alternative splicing, they do not provide any classification of the types of alternative splicing events present. Different forms of alternative splicing have different biological bases and the classification of alternative splicing events is critical for further work in deciphering the regulatory controls that govern these processes. To this end, we transformed all known gene structure information obtained from the genome annotations into splicing graphs based on the approach first proposed by Heber et al. in 2002 (Heber et al., 2002). We then created simple but robust rules for classifying the splicing graphs into various alternative splicing events. The rules created allows for the detection of multiple forms of alternative splicing within the same gene. To facilitate the assessment of the impact of alternative splicing on the protein product in particular with respect to the domain organization of the protein, Pfam (Bateman et al., 2004) domains were mapped onto the transcripts using HMMER (Eddy,

1998). All these data were then loaded into a database entitled DEDB (Drosophila melanogaster Exon Database) that is housed in MySQL (B. T. Lee, Tan, & Ranganathan, 2004). To aid in visualizing these splicing graphs, a web-based splicing graph viewer was also developed. The splicing graph viewer integrates gene structure, transcript, protein and domain information into an easily understandable interface that is viewable with any current web browser. The splicing graphs as well as the alternative splicing event classifications are available for download as XML files. A XML schema is available for parsing and validation of the XML files.

3.2 Implementation

3.2.1 Splicing graph source data

Annotations (release 3.2) were obtained from FlyBase in GAME XML format. The XML files are organized by scaffolds, each being a single large piece of genomic sequence. A number of these scaffolds constitute the Drosophila melanogaster genome. Within each scaffold are a number of annotations, each having one or more transcripts with multiple transcripts indicative of an alternatively spliced gene. A parser written in Python (gameXMLParser.py) was used to parse out the transcript data in tab delimited text files. The contents of the various text files generated are given in Table 5.

Filename	Fields	Description
scaffolds.txt	scaffold_id	Primary key.
	scaffold	Scaffold name.
	sequence	Scaffold nucleotide sequence.
	chromosome	Chromosome of which scaffold is part of.
	start	Start position of the scaffold.
	end	End position of the scaffold.
annotations.txt	annotation_id	Primary key.
	scaffold_id	Link to scaffolds.
	name	Name of annotation.
	gene_name	Name of the gene that the annotation represents.
go.txt	annotation_id	Link to annotations.
	go	GO id.
fba.txt	annotation_id	Link to annotations.
	flybase_id	FlyBase annotation id.
fbg.txt	annotation_id	Link to annotations.
	flybase_id	FlyBase gene id.
synonyms.txt	annotation_id	Link to annotations.
	synonym	Alternative gene name.
sequences.txt	sequence_id	Primary key.
	annotation_id	Link to annotations.
	name	Name of the sequence.
	strand	Orientation of the sequence.
	translation_start	Translation start position.
	translation_end	Translation end position.
	dna	Nucleotide sequence.
	protein	Protein sequence.
exons.txt	exon_id	Primary key.
	sequence_id	Link to sequences.
	start	Start position of the exon.
	end	End position of the exon.
	sequence	Nucleotide sequence of the exon.
introns.txt	intron_id	Primary key.
	sequence_id	Link to sequences.
	start	Start position of the intron.
	end	End position of the intron.
	sequence	Nucleotide sequence of the intron.

Table 5. Text files generated by gameXMLParser.py. The various fields in the text files are provided together with the description of the field. The scaffolds.txt file contains information about the scaffolds (large genomic fragments of the chromosomes). The annotations.txt file contains information about genome annotations, each genome annotation being located on a scaffold (hence the link to scaffolds). A single genome annotation contains information about a single gene. The genome annotation contains GO information (go.txt), FlyBase links (fbg.txt and fba.txt) and synonyms (synonyms.txt). Due to alternative splicing, each gene can potentially be composed of several transcripts, each being represented by a single entry in the file sequences.txt. A single transcripts also contains at least one exon (exons.txt) and possibly introns (introns.txt)

The various text files extracted from the annotations are then loaded into MySQL as individual tables for ease of retrieval and analysis. Of importance is that each annotation record represents a set of transcripts that make up a single gene. This is important as the splicing graph construction uses this fact.

3.2.2 Splicing graphs construction

The splicing graph module described in Chapter 2 was used to construct and visualize the splicing graphs contained in DEDB. The splicing graph representation describes exons and introns as vertices and edges respectively. Internally the vertices and edges are known as nodes and connections as they these labels provide a better description of these entities and these labels will be used for the rest of the text. Each node in the splicing graph is unique in its start and end position. Similarly each connection is unique in its start and end position. Nodes are linked together by connections as individual exons are also separated by introns. The splicing graph in this case is a directed graph as the underlying biological data (exons and introns) has directionality (5' to 3'). Every connection has a previous and next node as all introns are flanked by exons. However, a node is not required to have any previous or next connection due to the existence of single exonic genes. Nodes in multiple exonic genes will have at least one previous or next connection. A node having no previous connection is due to an initiation exon and thus labeled as an initiation node. Nodes having no next connection are due to termination exons and are thus termination nodes. All internal nodes will therefore have at least one next and one previous connection. A transcript is therefore merely a path in this graph starting at an initiation node and stopping at a termination node. As any path starting at an initiation node and ending at a termination node is a potential transcript, the splicing graph is

capable of determining the number of potential splicing variants in a gene. Furthermore, as a splicing variant is merely a path in the graph, any bifurcation in the graph is indicative of alternative splicing as the bifurcation is suggestive of alternative paths, hence alternative splicing variants. Lastly, transcripts contained with the splicing graph would share common nodes as splicing variants are unique combinations of exons and not unique sets of exons indicative that they share at least one common exon.

To construct the splicing graphs, we first retrieve genome annotation information from MySQL using the tables previously loaded with annotation data. Each annotation record is a collection of transcripts, which represents alternative splicing variants of a single gene. Each transcript is modeled as a GTranscript object (see Figure 16, Figure 17, Figure 18 and Figure 19 for the UML class diagrams). To construct the splicing graphs in DEDB, an algorithm implemented in the script clusterGenes.py was applied to each annotation record to obtain a splicing graph. The algorithm first takes a transcript, which is represented as a GTranscript object and makes that the first transcript to be contained in a Graph object. Each of the GExon and GItron instance contained in the GTranscript object is represented in the Graph as a Node and Connection object. At this point, there is only one path through the Graph that being the one found in the GTranscript object. Each subsequent transcript represented as a GTranscript Object is checked against this Graph object. Should the GTranscript object contain any GExon object that has a perfect match in terms of the start and end position to a Node object in the Graph, the GTranscript object is a splicing variant and its GExon and GItron objects added to the Graph. Should the GTranscript object not contain any such GExon object, a new Graph object is created with

this GTranscript object serving as its first transcript. Subsequent transcripts will be checked against these two Graph objects. At the end of the exercise, a series of Graph objects will be obtained, each being a cluster of splicing variants with each transcripts being represented as a path through the graph. To ease later analysis of the splicing graphs, the splicing graph is reverse complemented if the transcripts used to construct the splicing graph are in the negative orientation. Furthermore, the numeric positions of the nodes and connections are offset such that the smallest position is one.

The resulting splicing graphs are then converted into text files for loading back into MySQL for further work. The text files in Table 6 are created.

The resulting splicing graphs contained in MySQL are then checked by the Python script `createGoodClusters.py`. This script checks that the nodes in the splicing graphs are all correctly linked. The translation start and stop as well as coding exons were calculated by the Python script `protein.py`. The script generates text files listed in Table 7. This provides the capabilities to determine whether an exon or node is a coding exon and whether it contains a translation start or stop codon. The intron phase information is calculated by the Python script `intronPhase.py`. The text file generated by this script is presented in Table 8. The text files generated by the script `protein.py` and `intronPhase.py` are then loaded into MySQL.

The information obtained by the scripts `protein.py` and `intronPhase.py` are then used by the Python script `protein2.py` to determine the protein start and end position of each exon as well as the offset prior to the start of the codon. The contents of the text files are listed in Table 9. Much like the rest of the text files, these text files were also loaded into MySQL.

Additional information on the composition and length of the nodes and connections were computed for later analysis by the Python script `nodeConnectionCalculations.py`. The data generated by the script are found in the form of text files that are described in Table 10 and subsequently loaded into MySQL.

3.2.3 Domain searches using Pfam

Domain searches were done using the program `hmmpfam` from the HMMER package obtained from <http://hmmer.wustl.edu/>. Pfam (Bateman et al., 2004) release 12.0 fragment HMM models obtained from <ftp://ftp.sanger.ac.uk/pub/databases/Pfam> were used as the source of the HMM models used to search the protein sequences. The resulting output file was parsed by the Python script `parseHmmPfam.py` to produce two text files entitled `hmm_families.txt` and `hmm_domains.txt` whose contents are listed in Table 11. The resulting text files were then loaded into MySQL.

To facilitate further analysis on the domains with relation to the exons, the position of the detected domains were mapped onto exons. This was done by the script `hmm_positions.py`, which generated the text files listed in Table 12.

Pfam HMM model annotations stored in the file `pfamseq.gz` was obtained from <ftp://ftp.sanger.ac.uk/pub/databases/Pfam>. A parser was written in Python entitled `parsePfam.py` was written to parse the Pfam annotations into various text files listed in Table 13 for loading into MySQL.

To provide more complete domain descriptions, InterPro records were obtained from <ftp://ftp.ebi.ac.uk/pub/databases/interpro/> in the form of XML records. These XML based records were then parsed by `parseInterpro.py` to generate the text file described in Table 14.

Filename	Fields	Description
clusters.txt	cluster_id	Primary key.
	basePos	The offset amount.
cluster_ids.txt	cluster_id	Link to clusters.
	sequence_id	Link to sequences.
nodes.txt	node_id	Primary key.
	cluster_id	Link to clusters.
	start	Start position of the node (after offset).
	end	End position of the node (after offset).
	sequence	Nucleotide sequence of the node.
node_ids.txt	node_id	Link to nodes.
	exon_id	Link to exons.
node_next.txt	node_id	Link to nodes.
	next	Next connection.
	next_id	Next connection_id.
node_previous.txt	node_id	Link to nodes.
	previous	Previous connection.
	previous_id	Previous connection_id.
connections.txt	con_id	Primary key.
	cluster_id	Link to clusters.
	start	Start position of the node (after offset).
	end	End position of the node(after offset).
	sequence	Nucleotide sequence of the connection.
Connection_ids.txt	con_id	Link to connections.
	intron_id	Link to introns.
Connection_next.txt	con_id	Link to connections.
	next	Next node.
	next_id	Next node_id.
Connection_previous.txt	con_id	Link to connections.
	previous	Previous node.
	previous_id	Previous node_id.

Table 6. Text files generated by clusterGenes.py. The various fields in the text files are provided together with the description of the field.

Filename	Fields	Description
translation_start.txt	node_id	Link to nodes.
	exon_id	Link to exons.
	position	Translation start position.
	nposition	Translation start position corrected for strand.
translation_stop.txt	node_id	Link to nodes.
	exon_id	Link to exons.
	position	Translation stop position.
	nposition	Translation stop position corrected for strand.
coding_exon.txt	node_id	Link to nodes.
	exon_id	Link to exons.

Table 7. Contents of text files generated by Python script protein.py.

Filename	Fields	Description
intron_phase.txt	con_id	Link to connections.
	intron_id	Link to introns.
	phase	The intron phase of this intron/connection/

Table 8. Contents of the text file intron_phase.txt generated by the script intronPhase.py

Filename	Fields	Description
protein_positions.txt	node_id	Link to nodes.
	exon_id	Link to exons.
	start	Protein start position of the exon.
	end	Protein end position of the exon.
coding_offset.txt	node_id	Link to nodes.
	exon_id	Link to exons.
	offset	The offset prior to the first codon.

Table 9. Contents of the text files generated by the script protein2.py.

Filename	Fields	Description
node_cal.txt	node_id	Link to nodes.
	length	Length of the nodes.
	comp_a	Number of adenosine triphosphate.
	comp_t	Number of thymidine triphosphate.
	comp_g	Number of guanosine triphosphate.
	comp_c	Number of cytidine triphosphate.
	comp_n	Number of unknown nucleotide.
connection_cal.txt	con_id	Link to connections.
	length	Length of the connections.
	comp_a	Number of adenosine triphosphate.
	comp_t	Number of thymidine triphosphate.
	comp_g	Number of guanosine triphosphate.
	comp_c	Number of cytidine triphosphate.
	comp_n	Number of unknown nucleotide.

Table 10. Contents of the text files generated by the script nodeConnectionCalculations.py.

Filename	Fields	Description
hmm_families.txt	family_id	Primary key.
	sequence_id	Link to sequences.
	model	Name of the model.
	description	Description of the model.
	score	Score of the match.
	e_value	E-value of the match.
hmm_domains.txt	family_id	Link to hmm_families.
	domain_id	Primary key.
	sequence_id	Link to sequences.
	model	Name of the model.
	domain_number	Domain number.
	domain_total	Total number of domains.
	seq_start	The start position of the match on the sequence.
	seq_end	The end position of the match on the sequence.
	seq_start_state	The start state of the match on the sequence.
	seq_end_state	The end state of the match on the sequence.
	hmm_start	The start position of the match on the HMM model.
	hmm_end	The end position of the match on the HMM model.
	hmm_start_state	The start state of the match on the HMM model.
	hmm_end_state	The end state of the match on the HMM model.
	score	The score of the match.
	e_value	The E-value of the match.
	hmm_alignment	The HMM portion of the alignment.
	matchline	The matchline of the alignment.
seq_alignment	The sequence portion of the alignment.	

Table 11. The text files generated by the script parseHmmpfam.py.

Filename	Fields	Description
Domain_start.txt	domain_id	Link to hmm_domains.
	exon_id	Link to exons.
	position	The start position of the domain.
Domain_end.txt	domain_id	Link to hmm_domains.
	exon_id	Link to exons.
	position	The end position of the domain.
Domain_within.txt	domain_id	Link to hmm_domains.
	exon_id	Link to exons.

Table 12. Text files generated by the script `hmm_positions.py`.

Filename	Fields	Description
pfam.txt	id	Primary key.
	accession	Pfam accession.
	description	Pfam model description.
	author	Pfam author.
	seed_alignment_method	Method used to align the seed.
	gathering method	Search threshold used to build the full alignment.
	trusted_cutoff	Lowest score of a match in the full alignment.
	noise_cutoff	Highest score of a match not in the full alignment.
	type	Type of family.
	sequence_number	Number of sequences in the alignment.
	alignment_method	The method of alignment.
	build_method	Build method used.
	comments	Comments.
pfam_reference	id	Link to pfam.
	title	Title of the reference.
	journal	The reference journal.
	medline	The medline ID.
pfam_links	id	Link to pfam
	dbase	Cross-referenced database name.
	accession	Accession number.
pfam_sequences	id	Link to pfam.
	name	Name of the sequence.
	dbase	Database source of the sequence.
	accession	Sequence accession.
pfam_alignments	id	Link to pfam.
	name	Name of the alignment.
	alignment	The alignment itself.

Table 13. Text files generated by the script parsePfam.py.

Filename	Fields	Description
interpro.txt	id	Link to pfam.
	abstract	Domain abstract.

Table 14. Text file generated by the script parseInterpro.py.

3.2.4 Types of alternative splicing events used in the classification

To provide for meaningful subsets of splicing graph for analysis, the splicing graphs were classified. The following types of alternative splicing events are used for the classification:

1. Alternative donor site
2. Alternative acceptor site
3. Intron retention
4. Cassette exon

We have also included within the database alternative transcription start and termination site events although the choice of the start and end of transcription is under the control of the transcriptional machinery rather than the splicing machinery. We provide for the category of alternative initiation exons and alternative termination exons, as the 5' and 3' end of the transcripts is not usually accurate due to the technical difficulties in their determination.

3.2.5 Rules used for the classification of alternative splicing events in DEDB

The use of splicing graphs as the data representation for the alternative splicing data allows for easy creation of classification rules shown in Figure 25. These rules are independent of each other and thus each splicing graph potentially contains none to many alternative splicing events.


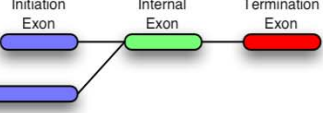
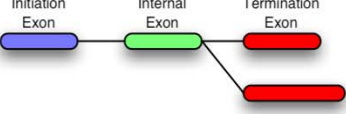
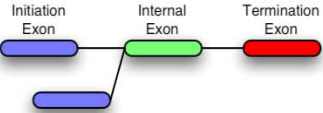
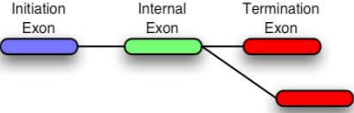
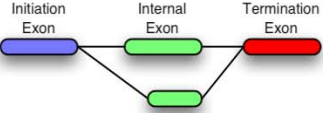
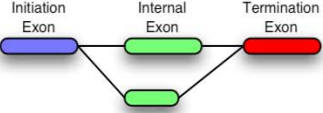
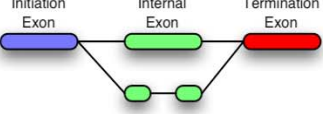
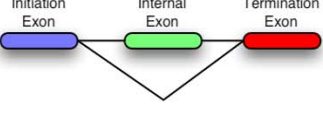
<p>Gene Structure</p> 	<p>A typical gene structure having three exons. The first exon is termed the initiation exon and is colored blue and the last exon is termed the termination exon and is colored red. All other exons are termed internal exons and are colored green.</p>
<p>Alternative transcriptional start site</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having no previous connections with a unique start position.
<p>Alternative transcriptional termination site</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having no next connections with a unique end position.
<p>Alternative initiation exon</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having no previous connections with a unique end position.
<p>Alternative termination exon</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having no next connections with a unique start position.
<p>Alternative acceptor site</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A set of overlapping nodes that are connected to a common upstream node. 2. The set of overlapping nodes should have unique start positions.
<p>Alternative donor site</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A set of overlapping nodes that are connected to a common downstream node. 2. The set of overlapping nodes should have unique end positions.
<p>Intron retention</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having a connection whose start and end position falls within itself.
<p>Cassette exon</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node whose start and end position falls within a single connection. 2. The node should have at least one next and previous connection.

Figure 25. Rules used in the classification of the alternative splicing events in DEDB.

Alternative transcriptional start sites (TSS) exists when more than one node is found that has no previous connection (indicating that it is an initiation node) and which contains a unique start position. The start position of initiation nodes is the transcriptional start site and multiple TSS implies alternative transcriptional start sites. Nodes 3, 5, 6, 7 and 9 in Figure 26 contain alternative transcriptional start sites.

Alternative transcriptional termination sites (TTS) occur when more than one node is found that has no next connection (indicating that it is a termination node) and which contains a unique end position. The end position of the termination node is the transcriptional termination site and the multiple TTS indicates alternative transcriptional termination sites. Nodes 7, 10 and 13 in Figure 27 have alternative transcriptional termination sites.

Alternative initiation exons occur when multiple initiation nodes (having no previous connections) are found to that have unique end positions. The rationale is that the start positions of initiation nodes are frequently incorrect as the 5' UTR (untranslated region) is rarely completely sequenced. Therefore, initiation nodes differing just in the start position cannot be easily determined to be different. We have thus also used the end position as the criteria. Furthermore, the 5' end of the initiation node is not recognized by the splicing machinery, only the 3' end (donor site) is recognized. Nodes 1 and 2 in Figure 28 are alternative initiation exons.

The same reasoning goes for alternative termination exons except that the positions are reversed. Nodes 7, 10 and 15 in Figure 29 are alternative termination exons.

As for alternative acceptor sites, these are found in a set of overlapping nodes (>1 node) that have differing start positions linked to a common node. The set of nodes should be overlapping else, they would be classified as cassette exons. Nodes 1 and 3 in Figure 30 contain alternative acceptor sites. They are both linked to node 2.

The same goes for alternative donor sites. Nodes 5 and 10 in Figure 31 contain alternative donor sites. They are linked to node 4. Take note that node 11 illustrates the point that the set of nodes have to be overlapping. From the graph, it is quite clear that node 11 is a cassette exon.

Cassette exons by definition are internal exons, which are differentially included in the various splicing isoforms of a gene. The rule as far as splicing graphs are concerned requires a cassette exon to be a internal node whose start and end position falls within a connection (an intron). The fact that the node occurs as part of a connection in some other splicing isoform implies that it is skipped hence fulfilling the definition. Node 1 in Figure 32 is a classic example of a cassette exon.

Intron retention on the other hand are introns which are not spliced out resulting in it being retained forming part of an exon. The rule based on splicing graphs requires a connection whose start and end position falls within a node for a positive intron retention event. The definition is fulfilled as the connection (intron) is found as being part of a node (exon). The connection between nodes 4 and 5 in Figure 33 is clearly retained in node 8.

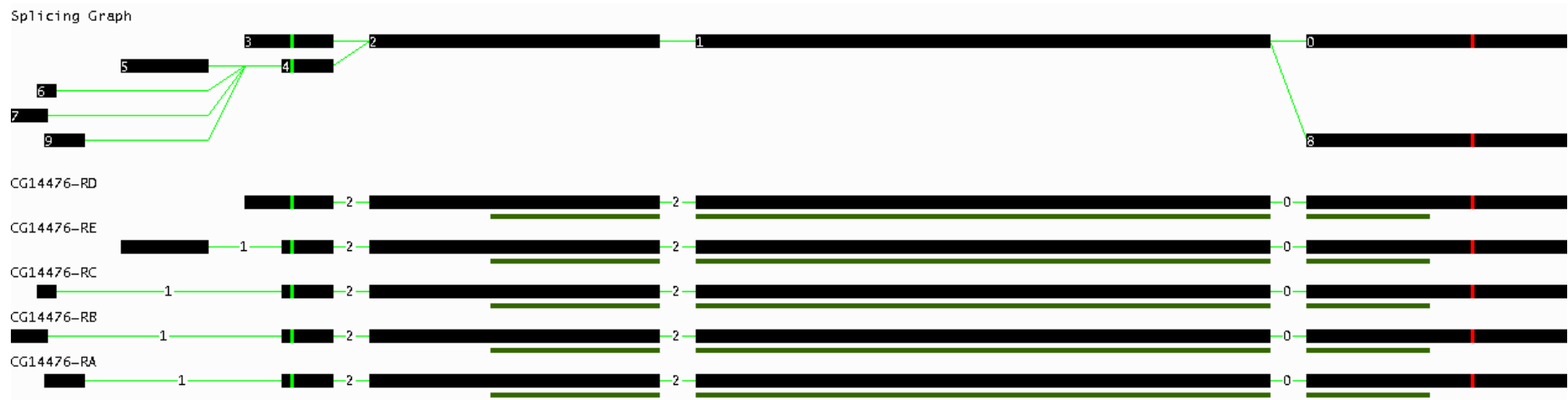


Figure 26. Alternative transcriptional start sites.

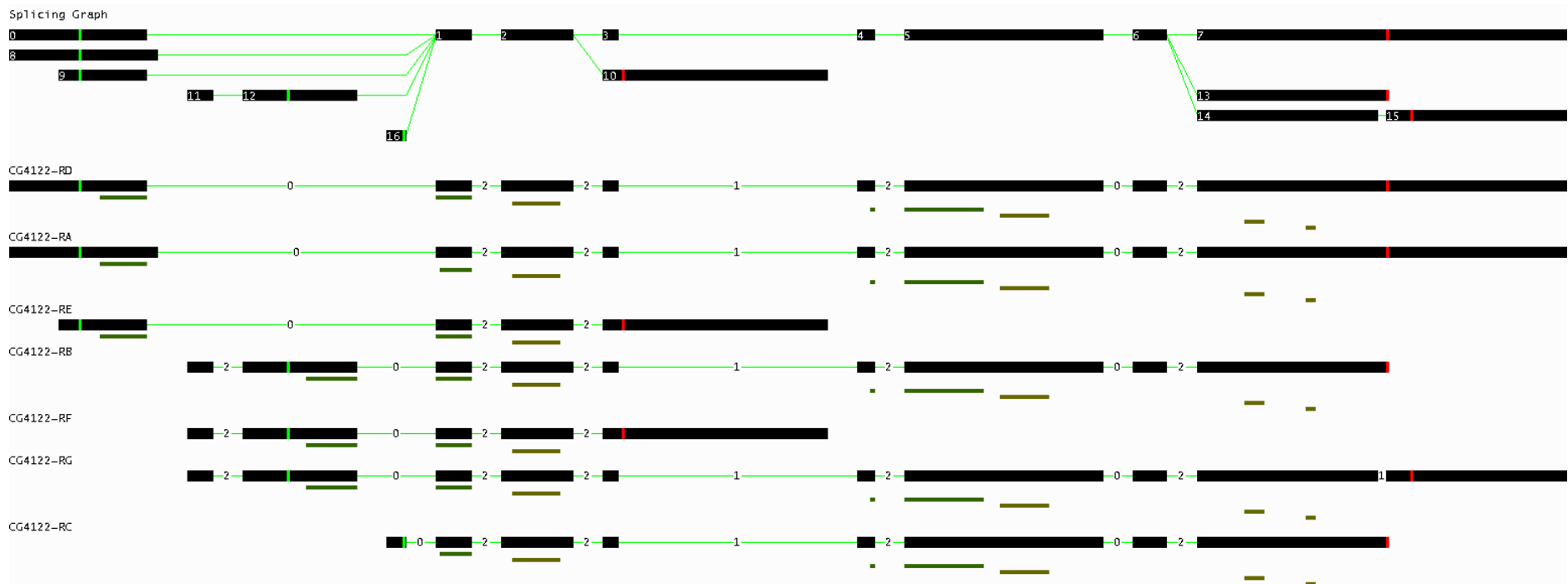


Figure 27. Alternative transcriptional termination sites.



Figure 28. Alternative initiation exons.

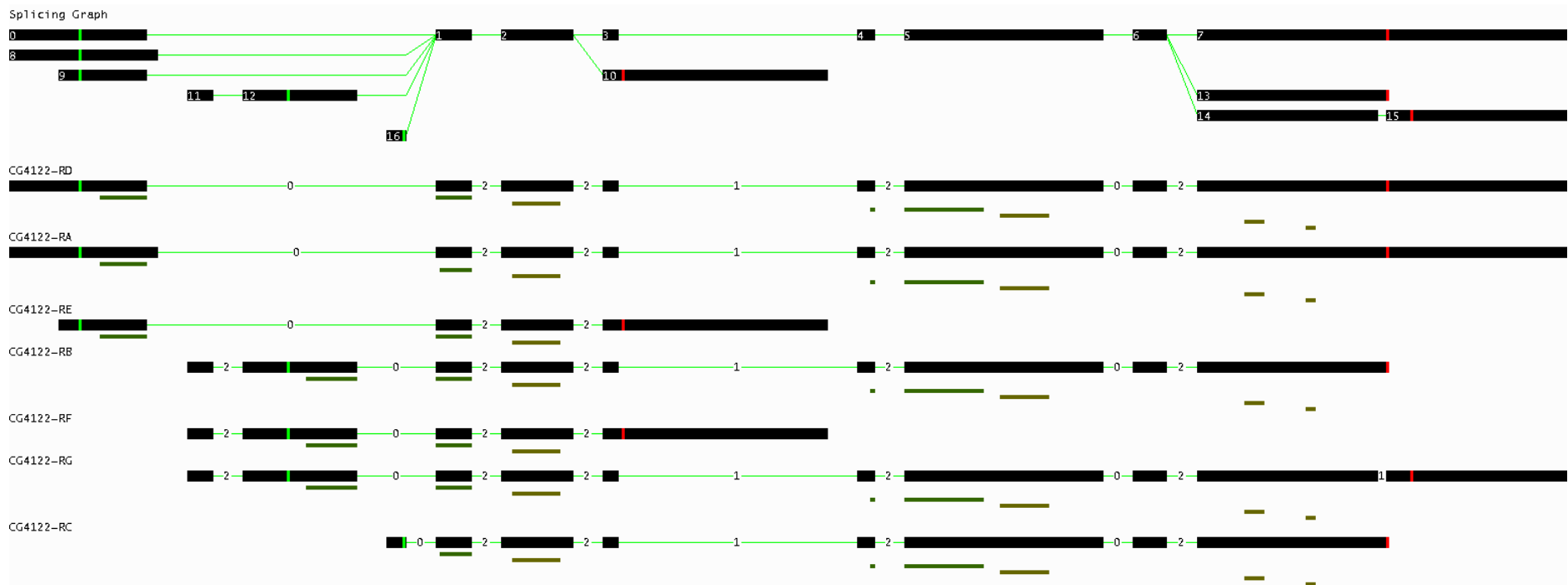


Figure 29. Alternative termination exons.

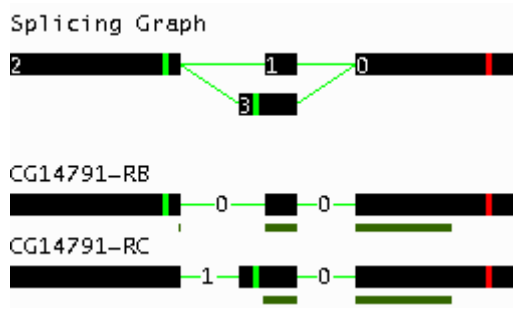


Figure 30. Alternative acceptor sites.

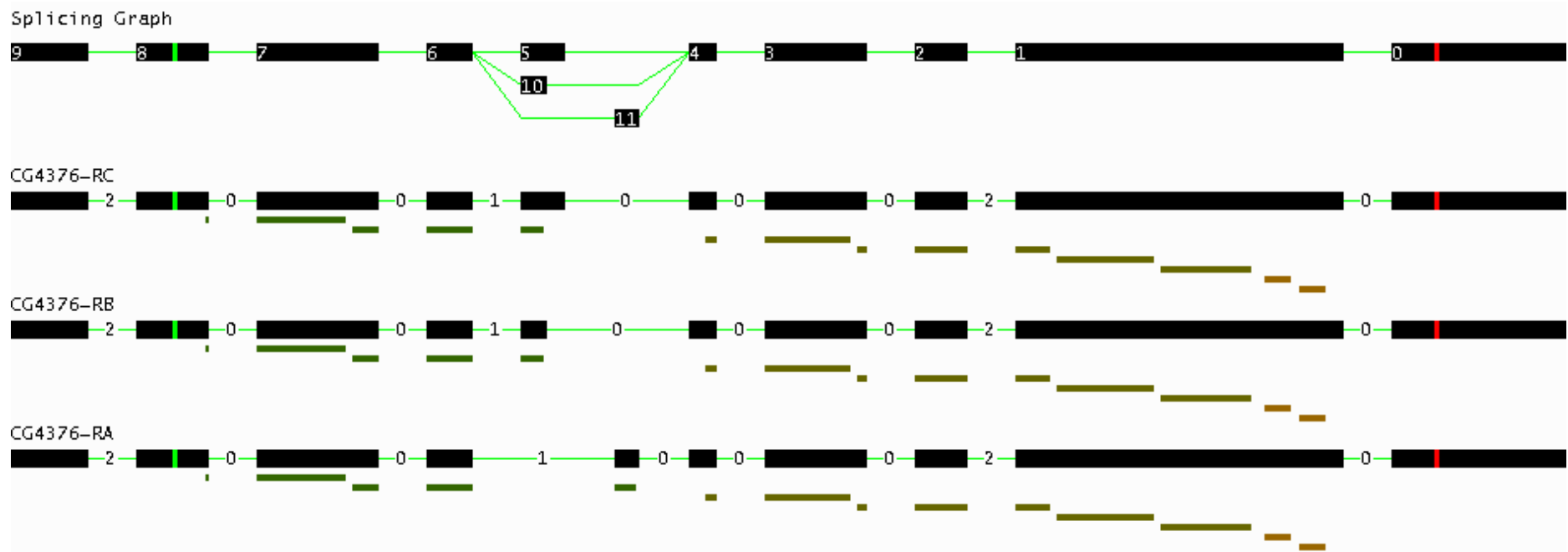


Figure 31. Alternative donor sites.

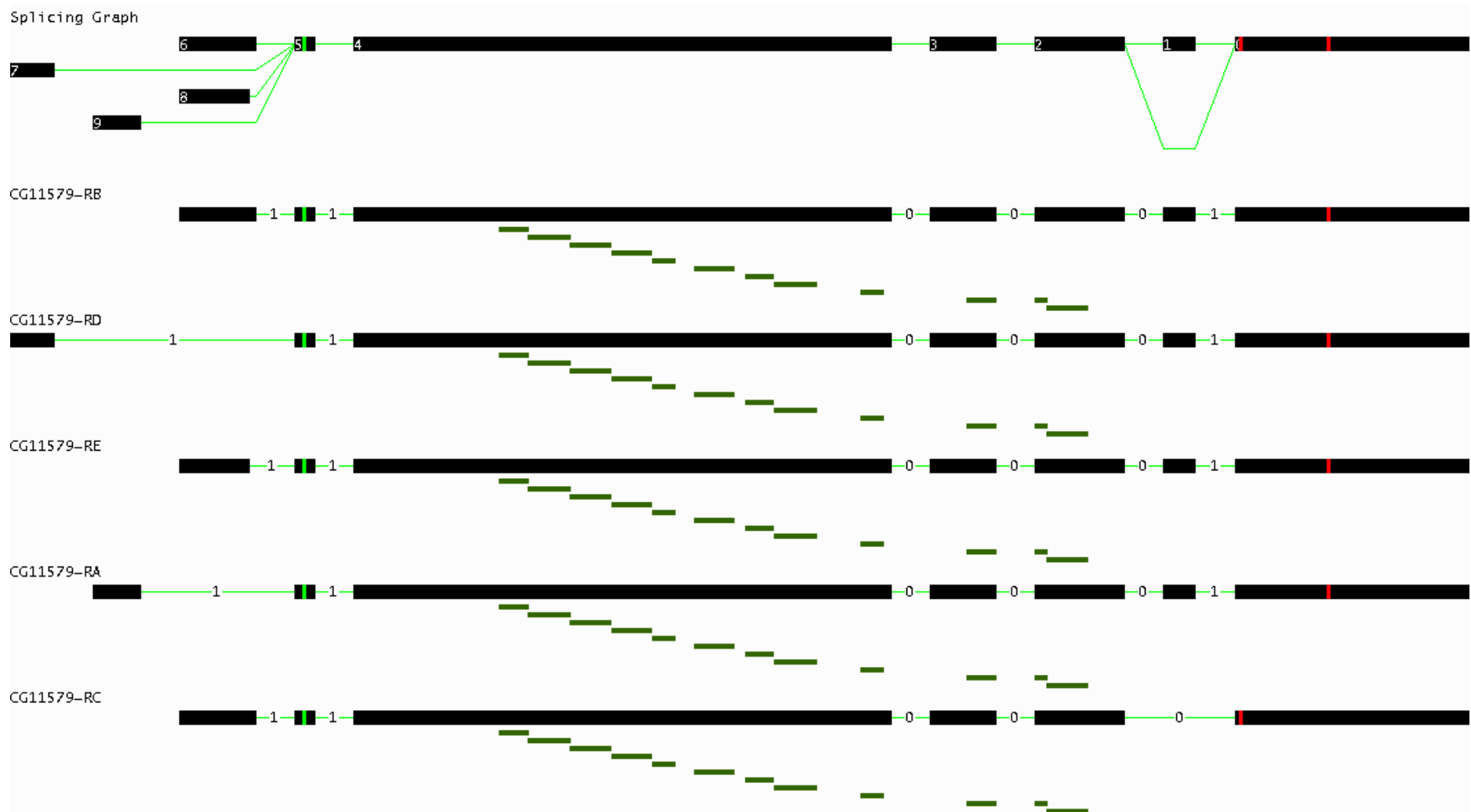


Figure 32. Cassette exons.

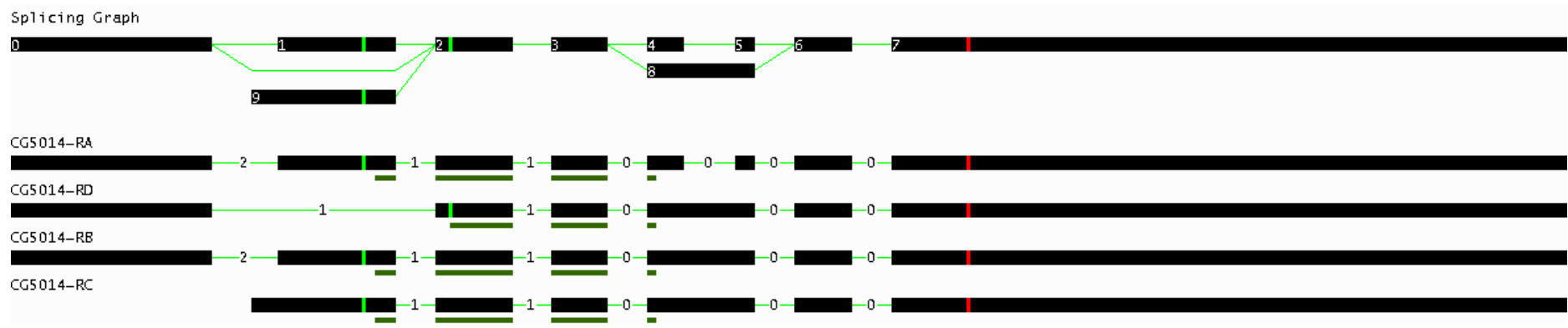


Figure 33. Intron retention.

3.2.6 Classification of alternative splicing events in DEEB

The splicing graphs stored in DEEB were classified by the Python script `getAlternativeSplicing.py` using the rules described above and the resulting data stored back into the database in the tables listed in Table 15. The results of the classification are shown in Table 16.

Table name	Field name	Description
single_exonic_gene	cluster_id	Link to clusters.
multi_exonic_gene	cluster_id	Link to clusters.
alternative_spliced	cluster_id	Link to clusters.
constitutive_exon	node_id	Link to nodes.
constitutive_intron	con_id	Link to connections.
alternative_sequence	sequence_id	Link to sequences.
constitutive_sequence	sequence_id	Link to sequences.
acceptors	node_id	Link to nodes.
	motif	The acceptor motif.
donors	node_id	Link to nodes.
	motif	The donor motif.
alt_acceptor	node_id	Link to nodes.
	common_id	The common node_id.
alt_donor	node_id	Link to nodes.
	common_id	The common node_id.
alt_init	node_id	Link to nodes.
alt_term	node_id	Link to nodes.
alt_init_exon	node_id	Link to nodes.
alt_term_exon	node_id	Link to nodes.
intron_retention	exon	Link to node that contains the retained intron.
	intron_5	Link to the 5' node flanking the intron retention.
	intron_3	Link to the 3' node flanking the intron retention.
	length	Length of the intron retention.
	connection	Link to the connection that is retained.
cassette_exon	node_id	Link to nodes.
internal_exon	node_id	Link to nodes.
initiation_exon	node_id	Link to nodes.
termination_exon	node_id	Link to nodes.

Table 15. MySQL tables used to store the information generated by the script getAlternativeSplicing.py.

Type of alternative splicing event	Number of events	Number of splicing graphs having the event
Alternative initiation exon	4723	1858
Alternative termination exon	1104	504
Alternative transcriptional start site	4211	1696
Alternative transcriptional termination site	1511	691
Alternative acceptor	905	384
Alternative donor	1399	587
Cassette exon	1228	644
Intron retention	983	700

Table 16. Table showing the various types of alternative splicing events and its associated number of events and splicing graphs.

3.2.7 Web interface

To allow for easy access to the database, a web interface was created using Apache as the web server on the server <http://proline.bic.nus.edu.sg>. A Splicing Graph Viewer consisting of a series of CGI (Common Gateway Interface) scripts were written to display the splicing graph and its associated information on the web shown as a screenshot in Figure 34. All modern web browsers are supported, for example Internet Explorer, Mozilla Firefox and Apple Safari.

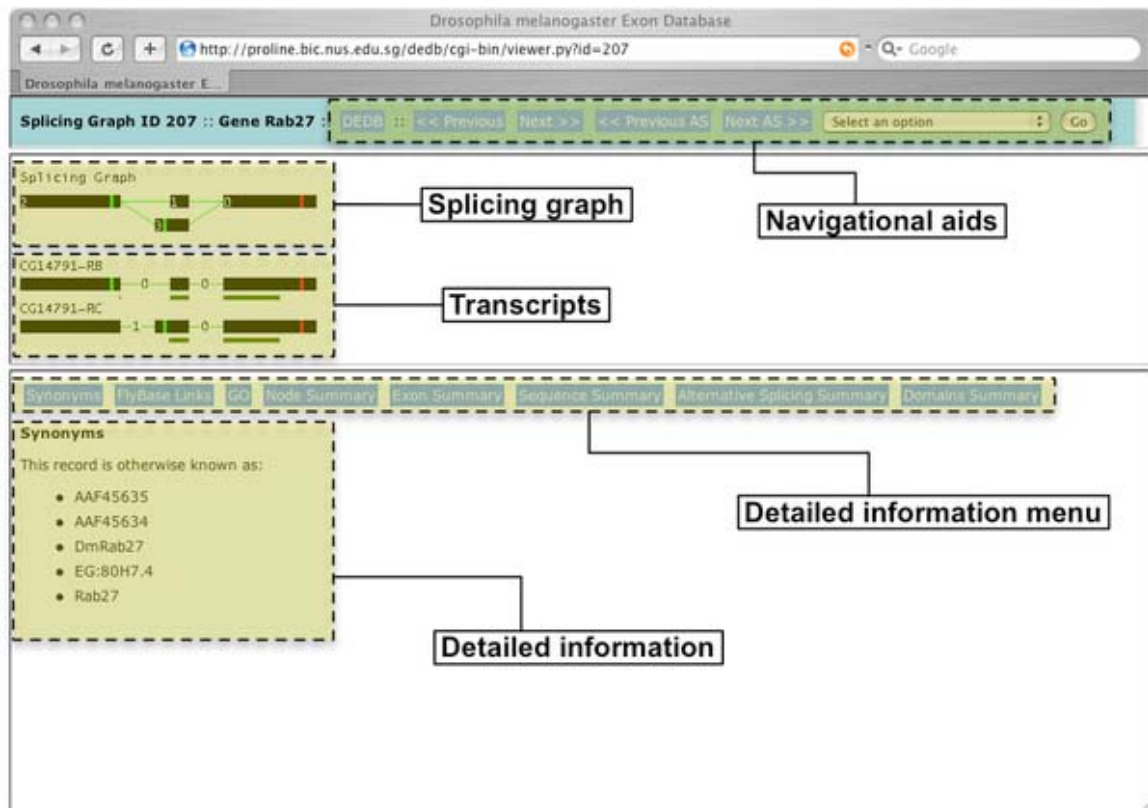


Figure 34. Screenshot of the splicing graph viewer in DEDB. The splicing graph viewer consists of three frames. The top frame shows the navigation aid that allows for rapid navigation between splicing graphs. The middle frame shows graphical representations of the splicing graph together with the transcripts that were used to construct the splicing graph. The bottom frame provides a place for detailed textual information.

The Splicing Graph Viewer consists of three frames shown in Figure 34. The navigational scheme is designed to be used in a top down manner, with navigational elements at the top providing a means of navigation between entities while the elements at the bottom allowing for navigation within the entity. The top frame shown in Figure 35 is the navigational frame that contains the navigational aids to allow users to navigate through the splicing graphs contained in DEDB quickly. Users can transverse through all the splicing graphs as well as through all the splicing graphs exhibiting alternative splicing events. Furthermore, users can also directly jump to splicing graphs exhibiting specific types of alternative splicing (Figure 36). The middle frame shows schematic representation of the currently selected splicing graph (Figure 37) as well as the transcripts (Figure 38) that make up the splicing graph. Users can elect to display more detailed information about the nodes, exons, introns and domains by clicking on its schematic representation in the middle frame. The information is displayed in the detail information area. The bottom frame contains a menu having a series of buttons that when clicked displays the selected information in the detailed information area. The interface provides sufficient information for users to quickly make sense of the types of alternative splicing present in the splicing graph. Any alternative splicing can be quickly determined by looking for any bifurcation in the splicing graphs and the type of alternative splicing can be quickly determined by the alternative path taken by the splicing variants. The number of potential paths through the splicing graph is likely to be higher than the number of splicing variants observed, the interface therefore provides schematic representations of the transcripts that make up the splicing graph for users to visualize the observed transcripts. The interface also allows the user to understand the implication of

alternative splicing on the protein sequence. This is achieved by displaying the start and stop codons on both the splicing graph and transcript schematic representations as well as the intron phase between the introns. This allows users to quickly determine if the alternative splicing affects the coding region and to determine if the exons altered by the alternative splicing affects the reading frame (by inspection of the intron phase). The effects of alternative splicing on the domain organization can also be determined visually as Pfam HMM domains have been mapped onto the transcripts. Extensive help is also available at the website for users on the usage of the Splicing Graph Viewer.

A search page is available on the website for users to query the database for specific splicing graphs shown in Figure 39. Users can search for specific splicing graphs using a BLAST search, the FlyBase Gene Name, the FlyBase Gene Symbol, the Pfam Accession Number and the Pfam Identifier. The BLAST search facilities allow users to search for orthologues of their gene of interest in *Drosophila melanogaster*. This might provide insights into the splicing variants in their gene of interest, as there exist a certain degree of conservation in alternative splicing between orthologues.

The results of the alternative splicing event classification are available on the website as a HTML page entitled "Alternative splicing event classification". This page shown in Figure 40 provides a table listing the type of alternative splicing event and the number of such events in DEDB. The type of alternative splicing event is linked to pages (an example is shown in Figure 41) that provide a listing of the splicing graphs that contain that particular alternative splicing event.

For users who require large portions of the DEDB data, download facilities are provided on the website. Users can download the data onto their local system for future analysis. To assist users in parsing the data, DEDB data is available in the form of XML files. XML parsers are widely and freely available for most popular programming languages like Perl, Python and Java. This allows the XML files to be parsed quickly and easily. Furthermore, a XML schema is provided that describes the XML file to further assist the parsing process.

In a bid to provide for better integration with other databases, DEDB provides links to FlyBase records that are the source data for DEDB and Pfam for more extensive domain information. Efforts were also made together with FlyBase to provide links to DEDB from FlyBase. FlyBase gene records now contain links to DEDB. This allows users navigating FlyBase to easily link to DEDB for visualization of the splicing variant information in the form of a splicing graph.

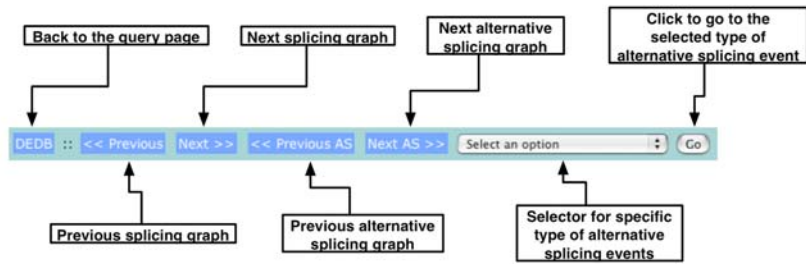


Figure 35. Top navigational frame. The first button on the menu links the user back to the query page allowing them to search for specific splicing graphs fulfilling certain criteria. The next two buttons allow the user to navigate to the next and previous splicing graph while the next two buttons transverse through all the splicing graphs exhibiting alternative splicing events. There is also a drop down selector (Figure 36) that allows users to jump to splicing graphs showing specific type of alternative splicing event.

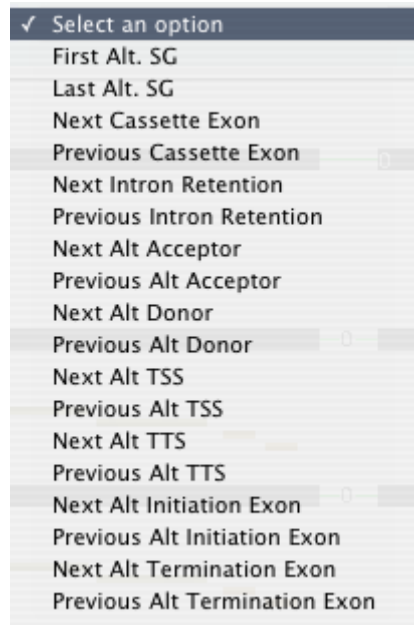


Figure 36. Navigational drop down selector. This drop down selector allows users to select for splicing graphs having specific types of alternative splicing.

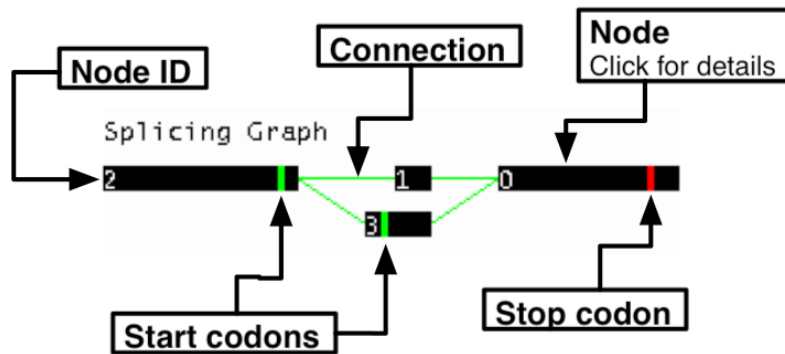


Figure 37. Schematic diagram of the splicing graph shown in the middle frame. Nodes corresponding to exons are shown as black bars connected via green lines representing connections (which are introns). The Node ID is displayed as white numbers on each of the nodes. The start and stop codons are shown respectively as green and red lines on the nodes. Users can click on the nodes to get more information in the detailed information area.

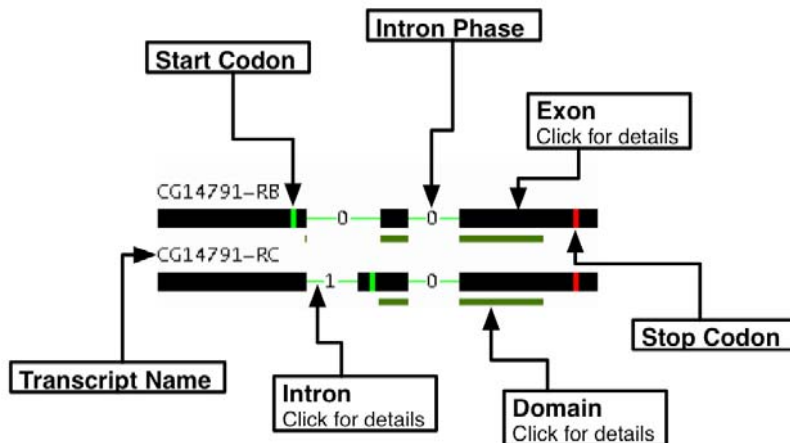


Figure 38. Schematic representation of the transcripts that make up the splicing graph. Each transcript is shown as a series of black bars (exons) connected together via green lines (introns). The graphical elements representing the exons and introns can be clicked to show more detailed information in the detailed information area. The FlyBase Gene ID for each transcript is shown at the top of each transcript. The start and stop codons for each transcript is shown respectively as a green and red line. The intron phase of each intron is shown as a number in the middle of the green line. Colored bars if present below the transcripts represents Pfam domains that has been detected on the transcripts. Users can click on the domains to display more information in the detailed information area.

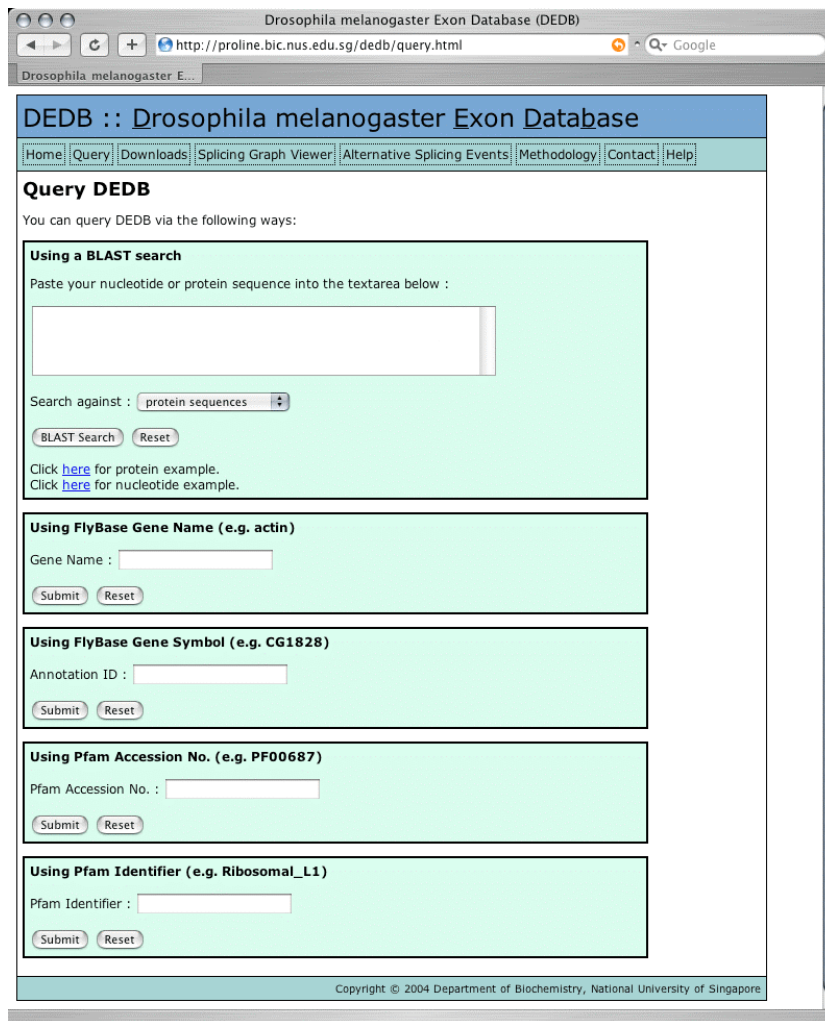


Figure 39. Screenshot of DEDB query page. Users can query DEDB via a BLAST search, by FlyBase Gene Name, FlyBase Gene Symbol, Pfam Accession Number and Pfam Identifier.

Drosophila melanogaster Exon Database (DEDB)

http://proline/dedb/cgi-bin/altsplice.py

DEDB :: Drosophila melanogaster Exon Database

Home Query Downloads Splicing Graph Viewer Alternative Splicing Events Methodology Contact Help

Alternative splicing event classification

The following table shows the type and number of alternative splicing events classified using splicing graphs. Details of the rules used for the classification are available in the [Methodology](#) page. Click on the alternative splicing event to view a list of splicing graph IDs that contain the particular alternative splicing event.

Type of alternative splicing event	Number
Alternative initiation exon	4723
Alternative termination exon	1104
Alternative transcriptional start site	4211
Alternative transcriptional termination site	1511
Alternative acceptor	905
Alternative donor	1399
Cassette exon	1228
Intron retention	983

Copyright © 2004 Department of Biochemistry, National University of Singapore

Figure 40. Screenshot of Alternative splicing event classification page in DEDB.

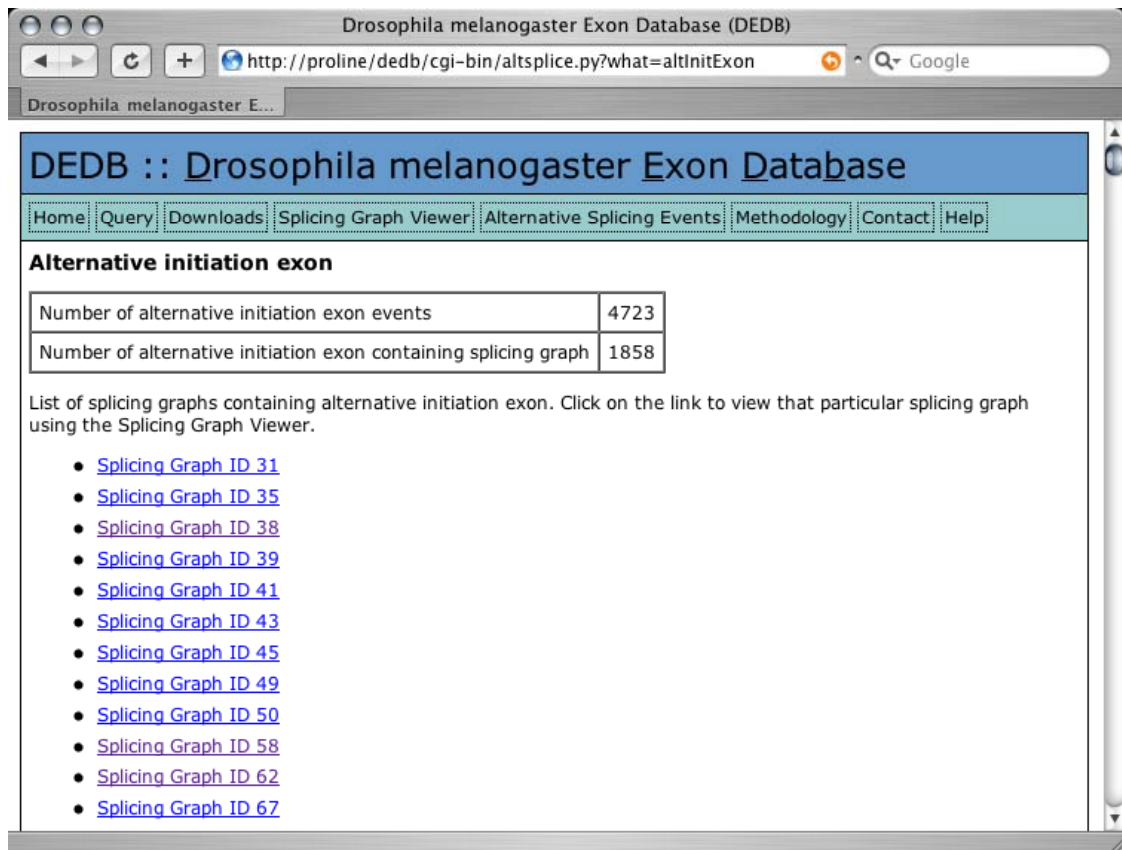


Figure 41. Screenshot of a page listing the splicing graphs that contain a particular type of alternative splicing event. Each item in the list is a link to the Splicing Graph Viewer.

3.3 Comparisons with other splicing graph databases/services

Other databases/services like ASG (Alternative Splicing Gallery) and MAASE (The Manually Annotated Alternative Spliced Events Database) depict the splicing graph as shown in Figure 42, Figure 43 and Figure 44 (Leipzig, Pevzner, & Heber, 2004; Zheng et al., 2005). Both of these databases/services show the splicing graph in the original manner. DEDB modifies the representation somewhat by including every form of the exons in the representation thereby making it more obvious that there are changes to the exons. This was meant to make the visual representation more intuitive to the biologist who can now see all the various forms of the exons. All the various databases/services including DEDB have both the splicing graph as well as the transcript view. One feature of ASG that DEDB currently does not have is the coloring of the various alternative splicing events. However DEDB provides additional information in the form of the coding sequence start and end position, the intron phases and Pfam domains mapped by hmmpfam (Eddy, 1998).

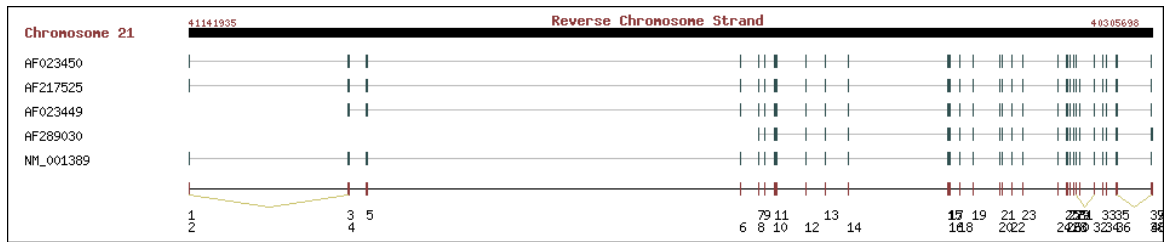


Figure 42. Global View of human dscam gene using MAASE. Deletions in some isoforms are depicted by the dark yellow lines in a manner similar to the original splicing graph visualization. The global view shows the introns with a size that is relative to that of the exons which leads to very long introns. This differs from SGM where the intron size is visually reduce to depict more of the exons. This visual representation makes it difficult to see that there is actually two cassette exons. The alternative initiation exons appear to be quite oblivious.

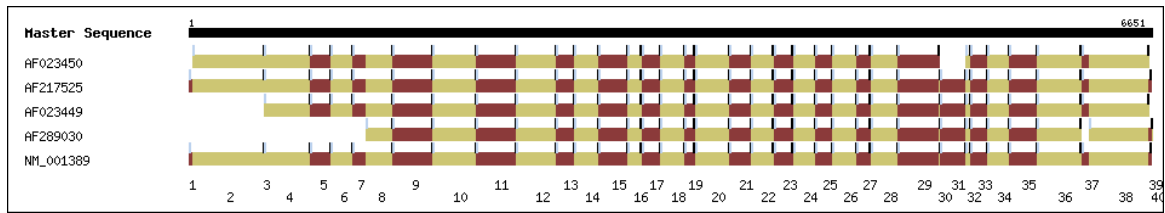


Figure 43. Exon Region Alignment view of MAASE. This is quite similar to the transcript view produced by SGM.

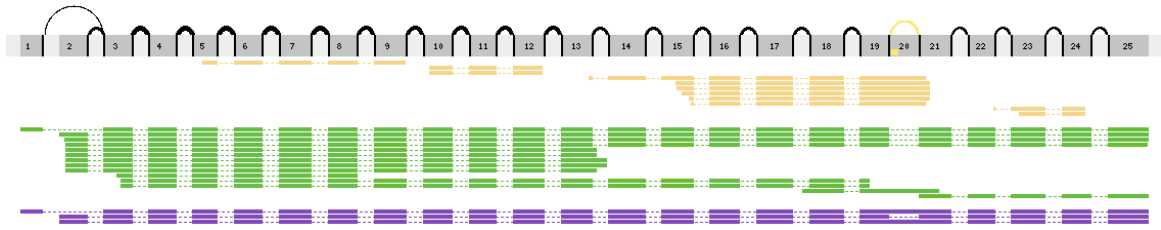


Figure 44. Splicing graph produced by ASG. The splicing graph is shown at the top and is very similar to the original splicing graph visual representation. The transcripts supporting the splicing graph are depicted below. An advantage that ASG has over SGM is that alternative splicing events are marked in colors in the splicing graph.

3.4 Conclusion

DEDB was successfully constructed using FlyBase gene annotations as its source data. The database provides for the visualization of splicing variant information in the form of splicing graphs that allow users to quickly make sense of the alternative splicing as well as its implications. DEDB also attempts to enrich the information by providing additional information on the protein coding region, intron phase and domain organization. The splicing graphs are also classified based on the alternative splicing events. Access to the database is facilitated by a query system that allows users to query via a number of criteria including gene names, domain identifiers and BLAST searches. To aid other bioinformaticians in large-scale analysis, the data in DEDB is also available as XML files allowing for ease of information parsing.

The data in DEDB represents a significant genomic level dataset in a form that is amenable to further analysis that is fully described in the next chapter.

Chapter 4: Genome-wide analysis of alternative splicing in *Drosophila melanogaster*

4.1 Introduction

Drosophila melanogaster is an invertebrate where splicing is a known phenomenon. Surveys of the exon and intron lengths of *Drosophila melanogaster* began in 1988 with a study by John Hawkins (Hawkins, 1988). The study showed that introns and exons in different organism exhibit differences. This study was followed by others where the differences were further elaborated (Guo, Lo, & Mount, 1993; Mount et al., 1992). These studies have indicated that there are differences in the splicing machinery of *Drosophila melanogaster* as compared to other organisms, making *Drosophila melanogaster* an interesting subject for the study of alternative splicing (Graveley, 2005; Nagengast & Salz, 2001; C. Schneider, Will, Brosius, Frilander, & Luhrmann, 2004). In view of the availability of large amounts of genomic (Adams et al., 2000) and transcript sequence information in conjunction with high quality manual curation (Misra et al., 2002), a genome level analysis of alternative splicing was carried out in *Drosophila melanogaster*.

The first part of the analysis seeks to understand the characteristics of alternative splicing in *Drosophila melanogaster*. To this end, simple statistics about the frequency of alternative splicing, the length distribution of exons and introns and the number of exons and introns per gene were analyzed. The results of these analyses would be important in devising methods for the detection of alternatively spliced exons and introns as demonstrated in a recent paper using

SVM (Support Vector Machine) as the machine learning method (Zhang, Heller, Hefter, Leslie, & Chasin, 2003). The results also have implications on the model of splicing found in *Drosophila melanogaster*, being either exon or intron definition (Berget, 1995; Robberson et al., 1990; Talerico & Berget, 1994). To better understand the contribution of the splice site motifs to alternative splicing, we have adopted the use of information theory to describe the level of conservation of the splice site motifs (T. D. Schneider, 1997; T. D. Schneider & Stephens, 1990; T. D. Schneider, Stormo, Gold, & Ehrenfeucht, 1986). Previous studies (Itoh, Washio, & Tomita, 2004; Rogan, Faux, & Schneider, 1998) have indicated that the information content of alternatively spliced splice site motifs have less conservation than constitutively spliced splice site motifs. In addition to these analyses, the frequency of introns lying close to protein domain boundaries was also analyzed. Having introns close to the protein domain boundaries is thought to allow for the selective inclusion of functions into the splice isoforms.

In addition to characterizing alternative splicing, analyses were also undertaken to understand the effects of alternative splicing. GO (Gene Ontology) terms assigned to each of the genes were analyzed for concentrations of alternative splicing in certain categories of genes. This is in view of a recent study (Modrek et al., 2001) that showed that alternative splicing is observed in certain types of genes. Another recent study (Kriventseva et al., 2003) on the effects of alternative splicing indicated that most alternative splicing events tend to disrupt or include protein domains, thus affecting the function of the protein to a large extent. A similar analysis was also carried out that analyzes the effects of alternative splicing on both the coding sequence as well as the protein domains in hopes of verifying the results.

4.2 Material and Methods

The source of data for the analysis is DEDB that was shown in the previous chapter. The relevant portions of DEDB that was used in each of the analysis are described fully in this section.

4.2.1 General statistics

Data housed in the following MySQL tables were extracted for analysis:

1. sequences
2. cluster_ids
3. good_clusters
4. single_exonic_gene
5. multi_exonic_gene
6. exons
7. introns
8. nodes
9. connections
10. alternative_spliced
11. alt_init
12. alt_term
13. alt_init_exon
14. alt_term_exon
15. alt_acceptor
16. alt_donor
17. cassette_exon
18. intron_retention

The data extracted was used to generate the following statistics using a Python script entitled totalNumbers.py:

1. Total number of transcripts
2. Total number of single exonic genes
3. Total number of multi exonic genes
4. Total number of splicing graphs
5. Total number of exons
6. Total number of introns
7. Total number of nodes
8. Total number of connections
9. Total number of splicing graphs having alternative splicing events
10. Total number of splicing graphs having alternative TSS events
11. Total number of splicing graphs having alternative TTS events
12. Total number of splicing graphs having alternative initiation exon events
13. Total number of splicing graphs having alternative termination exon events
14. Total number of splicing graphs having alternative acceptor events
15. Total number of splicing graphs having alternative donor events
16. Total number of splicing graphs having cassette exon events
17. Total number of splicing graphs having intron retention events
18. Total number alternative TSS events
19. Total number alternative TTS events
20. Total number alternative initiation exon events
21. Total number alternative termination exon events
22. Total number alternative acceptor events
23. Total number alternative donor events

- 24. Total number alternative cassette exon events
- 25. Total number alternative intron retention events
- 26. Number of transcripts per splicing graph
- 27. Number of possible and actual splicing variants

These statistics were then loaded back into MySQL for later retrieval. The number of unique paths in a splicing graph was determined by transversing the splicing graph and enumerating the number of distinct paths.

4.2.2 Exon and intron length analysis

Exon and intron length data from MySQL was extracted by the script `exonLengthDistribution.py` for analysis. The following types of exons and introns in Table 17 were analyzed.

Type of exon or intron
Single constitutive exons
Multi constitutive exons
Multi constitutive initiation exons
Multi constitutive internal exons
Multi constitutive termination exons
Cassette exons
5' introns flanking cassette exons
3' introns flanking cassette exons
Alternative acceptor exons
Alternative donor exons
Intron retained exons
5' exons flanking retained intron
3' exons flanking retained intron
Retained introns
Constitutive introns

Table 17. Types of exons and introns length analyzed.

Basic statistical measures like the minimum, maximum, mean, median, 1st quartile, 3rd quartile and inter-quartile range were calculated for all the various types of exons and introns listed in Table 17. In addition, a histogram was drawn for each type of exons and introns. The various statistical measures as well as the histograms were saved back to MySQL for later retrieval.

4.2.3 Exon number analysis

The number of exons for alternatively spliced genes and constitutively spliced genes were analyzed using the data from MySQL tables exons, alternative_sequence and constitutive_sequence. Basic statistical measures like the minimum, maximum, first quartile, median, mean and third quartile were computed. Histograms were also plotted. Again, as with the rest of the analysis, the statistical measures together with the histograms were stored back to MySQL for later use.

4.2.4 Nucleotide composition analysis

The nucleotide composition of the following types of exons and introns were analyzed for any patterns:

- alternative acceptor exons
- alternative donor exons
- cassette exons
- constitutive introns
- retained introns
- multi constitutive exons
- multi constitutive initiation exons
- multi constitutive internal exons

- multi constitutive termination exons
- single constitutive exons

The data was derived from the tables in MySQL. In addition to splitting the exons and introns by type, each type of exon and intron is further divided into 4 different quartiles based on the length (using the data obtained during the analysis of the exon and intron lengths). Like the rest of the analysis, the data generated was stored into MySQL for later use.

4.2.5 Splicing motif analysis

The splice site motifs were extracted from the relevant tables in MySQL and the frequencies of the various types of splice motifs were computed. The frequencies were calculated for GT-AG, GC-AG, AT-AC type splice motifs. All types of splice sites were lumped into a category entitled others.

To describe the level of conservation of the splice site motifs, information theory (T. D. Schneider, 1997) was used. Multi constitutive internal exon (MCIE) splice site motifs were considered to have sufficient signal for splicing and used for comparison against various other splice site motifs found in various alternative splicing events. An individual information weight matrix was used to determine the individual information content of every splice site motif. The individual information weight matrix was constructed using the following equation:

$$R_{iw}(b,l) = 2 - (-\log f(b,l) + e)$$

where $f(b,l)$ is the frequency of the base (b) at position l and e is the correction factor (T. D. Schneider et al., 1986) approximated using the following equation:

$$e = \frac{(s-1)}{(2 \times \ln 2 \times n)}$$

where s is the number of states (4 in the case of nucleotides) and n is the number of sequences used. The individual information content for each motif can then be computed using:

$$R_i(j) = \sum_l \sum_{b=a}^t s(b,l,j) R_{iw}(b,l)$$

where $s(b,l,j)$ is 1 where the nucleotide in the motif matches b and 0 otherwise.

The information content of each type of alternative splicing events can then be calculated as the mean of all the individual information content of all the motifs that involved in the specific alternative splicing event. This allows for the assessment of the level of conservation of the splice site motif in various types of alternative splicing events by comparison of the mean individual information content. In addition, this allows for the creation of the sequence logo (T. D. Schneider & Stephens, 1990) for visualization of the donor and acceptor splice site motifs. As usual, all the information for the computation of the information content were obtained from the relevant tables in MySQL. Information content as well as the corresponding sequence logo was determined for MCIE donor and acceptor sites. The information content was also calculated for every exon length quartile to determine if there are any significant differences between exon length and conservation. Individual information content matrices were constructed for MCIE acceptor and donor sites. The weight matrices were then used to determine the individual information content of the splice site motifs of the following types of alternative splicing events:

1. Cassette exon acceptor splice site
2. Intron retention acceptor splice site
3. Alternative acceptor splice site

4. Constitutive internal exon acceptor splice site
5. Constitutive termination exon acceptor splice site
6. Cassette exon donor splice site
7. Intron retention donor splice site
8. Alternative donor splice site
9. Constitutive internal exon donor splice site
10. Constitutive initiation exon donor splice site

Histograms of the individual information content were also constructed.

4.2.6 Domain boundary analysis

The domain boundaries were also analyzed to determine if there is any tendency for introns to lie near the domain boundaries. Having introns near the domain boundaries are thought to allow for ease of exon shuffling. The analysis counted the number of introns lying within 10 amino acids of domain boundaries as shown in Figure 45 as well as the number of introns falling outside this 10 amino acid region. The expected number of introns to fall within and outside the regions were then computed and used for a Chi Square Goodness-of-Fit test to determine if the number of introns lying in the domain boundaries deviate from the expected. A plot was also drawn showing the p-value at specific amino acid positions.

In addition to analyzing whether introns have a tendency to lie near the domain boundaries, for domains that containing flanking introns in the -30 to $+30$ region, a Chi Square Goodness-of-Fit test was done to determine if there is a tendency for these flanking introns to be symmetrical. Symmetrical introns are a pair of introns where the intron phase is the same. This allows the exon flanked by the introns to shuffle much easier as the coding frame is retained.

single amino acid residue

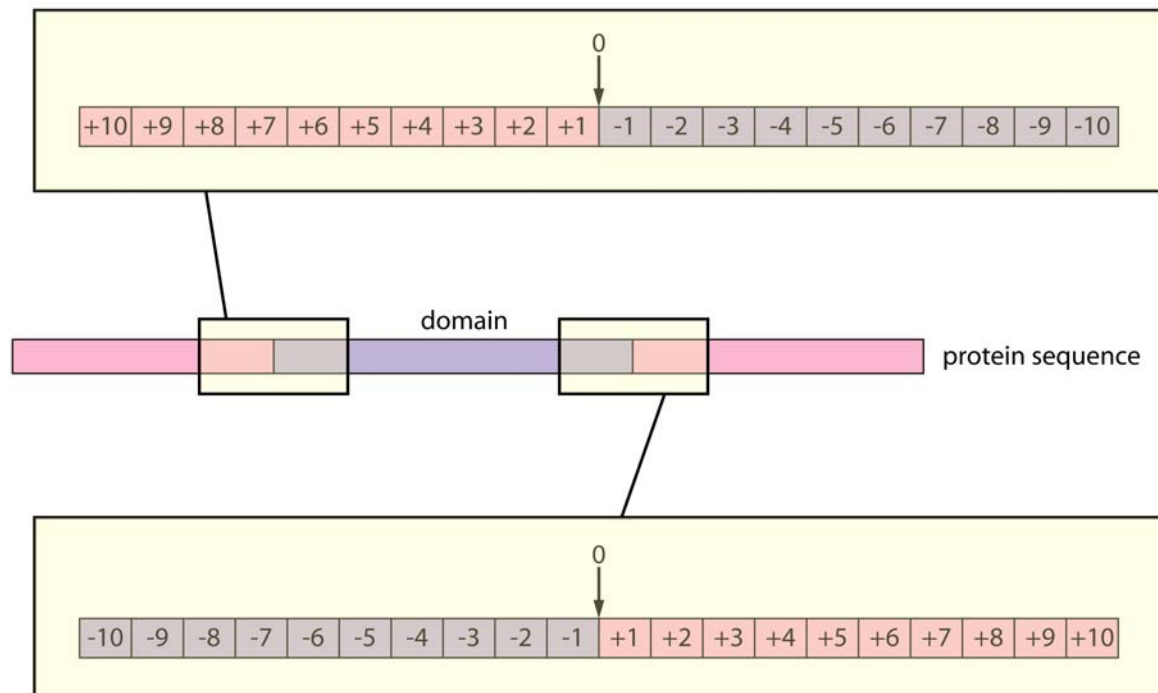


Figure 45. Schematic representation of the amino acid positions used in the domain boundary analysis. Each colored block represents a single amino acid residue. Amino acid residues in purple are part of the domain while residues in pink are outside the domain. Introns lying exactly at the boundary of the domains are designed as being in position 0. Introns lying in the amino acids within the domains are in increasing negative integers beginning from the domain boundary. Likewise, introns lying in the amino acids outside the domains are in increasing positive integers beginning from the domain boundary. A range of 10 amino acids flanking the domain boundary was used for the analysis.

4.2.7 Number of splicing graphs partitioned using GO terms

The slim versions of the Gene Ontology dated 31st of August 2004 available from ftp://ftp.geneontology.org/pub/go/GO_slims/ was used to provide a high level GO view of the splicing graph data in DEDB. Corresponding splicing graphs having the respective GO terms were calculated and then partitioned into two sets, one set being splicing graphs without any alternative splicing events termed constitutive splicing graphs and another set being splicing graphs having alternative splicing events termed alternative splicing graphs. Therefore each GO term is associated with three numbers, the total number of splicing graphs having the GO term, the number of constitutive splicing graph having the GO term and the number of alternative splicing graphs having the GO term. As usual all the data were extracted from MySQL for the computation. From these absolute counts, the percentage of splicing graph for each GO term is then computed as well as a ratio of constitutive splicing graph against alternative splicing graph. A positive ratio indicates that there is a higher percentage of constitutive splicing graphs as opposed to alternative splicing graphs and vice versa.

4.2.8 Effects of alternative splicing on the coding sequence

To determine the effects of alternative splicing on the coding sequence, we have adopted the metrology used by Kriventseva *et al.* (Kriventseva et al., 2003). This analysis seeks to determine if alternative splicing events have a preference or greater likelihood of affecting the coding sequence. The analysis used a Chi Square Goodness-of-Fit test in an attempt to answer this question. To determine the expected frequency of any alternative splicing event overlapping the coding sequence, a sliding window matching its length was used to scan the splicing

graph. At each position, the sliding window will be in one of the three different state, either not overlap, partially overlap or complete overlap the coding sequence. These states are counts and then converted into the expected frequencies. The observed frequency is then simply the state of the alternative splicing event in relation to the coding sequence. These values are then used for the Chi Square Goodness-of-Fit test to determine whether the effects of the type of alternative splicing event in question on the coding sequence deviates from the expected. The data in the relevant MySQL tables was utilized for this analysis.

4.2.9 Effects of alternative splicing on the domains

This analysis is similar to the previous analysis other than the fact that instead of the coding sequence, it is the domains that is analyzed. Due to the fact that the not all the domains would be mapped onto the transcripts and that there could still be domains which are unknown, instead of using the alternative splicing event as the sliding window, the domain is used as the sliding window. This avoids the above mentioned complications and answers the question whether the domains have a higher tendency to include alternative splicing events and thus be affected by it.

4.3 Results and Discussion

4.3.1 General statistics

The statistics generated by the script totalNumbers.py are shown in Table 18, Table 19 and Table 20. DEEB currently comprises of 18,156 transcripts, which clusters into 13,222 genes of which 2374 are single exonic and 10848 are multi exonic. The fact that the majority of the genes are multi exonic indicates that the

fruitfly is an excellent organism for studies of alternative splicing due to the large potential for alternative splicing (possible only with multi exonic genes). The genome wide coverage of DEEB allows conclusions to be made on a genomic level, not possible with other datasets.

Approximately 20% of genes in fruitfly (2646 of 13222) are alternatively spliced. If only multi exonic genes are considered, then this percentage increases to 24%, almost a quarter of all multi exonic genes. This is much less than the amount of alternative splicing in humans, however the estimates provided here are likely to be an underestimation as the curation process would have missed some alternative splicing events. The 2646 alternatively spliced genes are responsible for an additional 4934 transcripts demonstrating the potential for increased protein diversity from the same gene pool. Alternative splicing event classification via splicing graphs reveals a large difference in the number of alternative initiation exon events (4723) as compared to the number of alternative termination exon events (1104), with alternative initiation exons being the more predominant event. Biologically, this could mean that there are multiple promoter regions for the genes, allowing for differential gene control. It could also mean that there are differences in the targeting signals in the 5' region of genes that allows for different cellular localization and the use of different initiation exons allows for this. However, note the fact that the transcripts used to generate the splicing graphs consist of far more 5' sequencing reads as compared to 3' sequencing reads. This may be the contributing factor to the large number of variations at the 5' end as opposed to the 3' end. Therefore, there may be far more variations not detected in the 3' end.

There are also a greater number of alternative donor site events (1399) as compared to alternative acceptor site events (905). This seems to show that the regulation of alternative splicing favors the control of the donor site even though there is much higher conservation at the donor site as compared to the acceptor site.

The number of intron retention events (983) is outnumbered by the number of cassette exon events (1228). This appears to suggest that exon definition is the primary splicing recognition model with defects in the recognition leading to cassette exons or it could mean increased likelihood of abnormalities in exon definition. Either way, it indicates that both exon and intron definition process occurs in fruitfly.

The number of transcripts per gene shown in Table 19 indicates that most genes contain less than 10 splicing variants. The largest number of splicing variant is 25. This data seem to contradict known biological knowledge, for example, the *dscam* gene is known to contain larger numbers of splicing variants than 25. This is most likely due to the process of curation, which is limited to the number of ESTs and cDNA known for the genes. In the case of *dscam*, this gene model was constructed from only 49 ESTs and 2 cDNA sequences, which is likely to be incomplete. As the EST databases increase in size, we would expect a more complete and accurate picture of the amount of alternative splicing and hence the figures reported are likely to be underestimates.

The data shown in Table 20 shows that the number of possible splicing variants as determined by the number of possible unique paths through the splicing graph is far less than the number of actual splicing variants observed. This could be due to low abundance of splicing variants making detection difficult

or it could be due to the presence of alternative splicing regulators leading to specific splicing variants. Although we cannot discount the difficulties in splicing variant detection, the presence of a minority of the possible splicing variants seems indicative that in many cases, there is a strict control over the particular splicing variants that is expressed. This is supplemented by experiments indicating strict control over the exact splicing variant expressed.

Statistic	Number
Total number of transcripts	18156
Total number of single exonic genes	2374
Total number of multi exonic genes	10848
Total number of splicing graphs	13222
Total number of exons	88403
Total number of introns	70247
Total number of nodes	60744
Total number of connections	46090
Total number of splicing graphs having alternative splicing events	2646
Total number of splicing graphs having alternative TSS events	1696
Total number of splicing graphs having alternative TTS events	691
Total number of splicing graphs having alternative initiation exon events	1858
Total number of splicing graphs having alternative termination exon events	504
Total number of splicing graphs having alternative acceptor events	384
Total number of splicing graphs having alternative donor events	587
Total number of splicing graphs having cassette exon events	644
Total number of splicing graphs having intron retention events	700
Total number alternative TSS events	4211
Total number alternative TTS events	1511
Total number alternative initiation exon events	4723
Total number alternative termination exon events	1104
Total number alternative acceptor events	905
Total number alternative donor events	1399
Total number alternative cassette exon events	1228
Total number alternative intron retention events	983

Table 18. Statistics generated by the script totalNumbers.py.

Number of transcripts	Number of splicing graphs
2	1569
3	553
4	265
5	103
6	67
7	35
8	24
9	7
10	6
11	8
12	1
13	4
14	1
15	1
17	1
25	1
Total	2646

Table 19. Number of splicing graphs containing specific number of transcripts.

Number of splicing variants	Potential number of splicing graphs having this number of variants	Actual number of splicing graphs having this number of variants
1	10576	10576
2	1382	1592
3	330	537
4	399	262
5	63	102
6	161	68
7	14	31
8	99	24
9	22	7
10	22	6
11	3	8
12	46	1
13	1	4
14	7	1
15	10	1
16	18	0
17	2	1
18	13	0
20	7	0
21	3	0
22	2	0
24	7	0
25	1	1
26	1	0
28	2	0
30	3	0
32	3	0
33	1	0
34	2	0
36	1	0
39	1	0
40	4	0
45	1	0
48	3	0
54	1	0
72	1	0
82	1	0
106	2	0

108	1	0
112	1	0
115	1	0
132	1	0
166	1	0
336	1	0
360	1	0
Total	13222	13222

Table 20. Table showing the number of splicing variant and the corresponding number of potential and actual splicing graph. The number of potential splicing graph contains the number of possible paths through the graph while the number of actual splicing graph contains the number of actual transcripts known.

4.3.2 Exon and intron length analysis

The exon and intron length analysis results seem to agree well with existing data in Xpro (Gopalan, Tan, Lee, & Ranganathan, 2004) which was derived from GenBank. Table 21 clearly shows that there is a difference in the length of single constitutive exons (897 median) as opposed to multi constitutive exons (270 median). Single constitutive exons are much longer (about 3 times using the median measure) than multi constitutive exons. The distribution for single constitutive exons (Figure 46) is quite spread out as compared to the distribution for multi constitutive exons (Figure 47). Although there is a peak for the single constitutive exons distribution, it is nowhere as well defined as that for multi constitutive exons. This probably means that there is a length restriction on exon lengths in multi exonic genes due to the requirement of splicing machinery as would be expected for exon definition. The single exonic genes do not seem to exhibit this restriction, as there is no interaction required between the start and end of transcription. Each of these involves separate recognition process, the former require the recognition of the promoter and the transcriptional start site while the later requires the recognition of the transcriptional termination site. Post-transcriptional modification of these two types of genes differs in the absence of splicing in the case of the single exonic genes, hence the lack of any restriction. The sharp peak observed in multi exonic genes could also be indicative of an optimal exon length for the splicing machinery in the exon definition model.

A breakdown of multi constitutive exons reveals that initiation and internal exons have similar distributions (Figure 48 and Figure 49 respectively) that are quite different from termination exons (Figure 50). The distributions of multi

constitutive initiation and internal exons (medians being 230 and 219 respectively) are quite similarly both in the shape and the peak with the main difference being in a more defined distribution in the case of internal exons. Termination exons (567 median) on the other hand exhibit a much broader distribution with a far less defined peak. Should exon definition be commonplace in fruitfly, this is likely to manifest in a bell shaped distribution of internal exon lengths due to the spatial limitations of protein-protein interaction across the exons. This is exactly what is observed in multi constitutive internal exons. Therefore leading to the biological conclusion that exon definition is present and is relatively common in fruitfly. Initiation exons on the other hand suffers limitations due to interact between the mRNA 5' capping and splicing apparatus and termination exons are defined via interactions between the splicing machinery and the polyadenylation machinery (Goldstrohm et al., 2001). The fact that the distribution of initiation exons is a bell shaped distribution seems to indicate that such interaction between the mRNA 5' capping and splicing machinery does take place and that there are spatial constrains on these interaction. On the other hand, termination exons do not seem to exhibit a distinct bell shaped curve. This leads to the hypothesis that either the spatial constrains for this interact is more relaxed or that this interaction is less commonplace which could mean that intron definition is used in the intron preceding the termination exon.

Constitutive introns (Figure 51) have a very narrow left skewed bell shaped curve with a median of 69bp. This is highly indicative of intron definition in fruitfly, although as the data in the exon length indicates, exon definition is still a likely mechanism. Therefore, fruitfly is likely to employ both models of splicing.

Cassette exons are thought to be the result of abnormalities in exon definition (Berget, 1995). The median for cassette exons (150) is lower than that of multi constitutive internal exons. The distribution of the lengths of cassette exons shown in Figure 52 is a normal-like distribution skewed to the left. Multi constitutive internal exons are a mixture of exons resulting from exon and intron definition and the median reflects this fact being a composite of the optimal length of exons being defined via exon definition and the distribution due to intron definition exons. Intron definition exons can take any length as they are not bounded by spatial constraints on the exon lengths. Studies done indicate that the decision to use exon or intron definition is largely due to the nature of the exon and its flanking introns, with the shorter entity being the focus of the definition (for example a short exon flanked by long introns would be defined via exon definition) (Hawkins, 1988; Hoffman & Grabowski, 1992; Robberson et al., 1990; Talerico & Berget, 1994). This implies that intron definition exons are likely to be longer than exon definition exons and this would increase the median of multi constitutive internal exons. Therefore, the median cassette exons may be a better measure for the optimal exon length for exon definition. This conclusion agrees well with existing data contained in Xpro (Gopalan et al., 2004) where the mouse and human exons have similar lengths (exon definition is thought to be the major method in human and mouse due to their long intron lengths). Supporting the idea that definition occurs across the shorter entity is the fact that the median for the 5' and 3' flanking introns (medians being 653 and 639 respectively) are far greater than that of the flanked exon (150 median) and of the median of constitutive introns. The distributions for the 5' and 3' flanking introns (Figure 53

and Figure 54 respectively) show a peak with short intron lengths followed by a uniform like distribution unlike that observed for the lengths of cassette exons.

In the case of intron retention, they are likely to be the result of intron definition. As such, they would have lengths similar to that of constitutive introns, however this is not the case as observed in the table (134 median as opposed to 69 median for constitutive introns). A glance at the histogram shown in Figure 55 reveals the reason, being that the histogram has a rather odd shape likely due to insufficient data. In most cases, the intron length tends to be short in agreement with the intron definition model, however in some cases the intron is rather long, which has slewed the median quite a fair bit. A look at the flanking exons (medians being 154 and 237 respectively) distribution shown in Figure 56 and Figure 57 indicate that the retained intron is still shorter indicating intron definition is still at work. The recognition of longer introns could be facilitated by intronic splicing enhancers (ISE) that span the introns allowing for intron definition across long introns. The exon resulting from intron retention tends to be far longer (median of 720) than regular exons and that is to be expected since it is the combination of two exons with a single intron. The distribution of the lengths of the exon containing the intron retention is shown Figure 58, being the composite of the 5' and 3' exon flanking the retained intron and the retained intron itself.

Interesting enough, both alternative acceptor and donor exons differ from multi constitutive exons. Alternative acceptor exons has a longer median (306) while alternative donor exons has a shorter median (172). The distribution for the alternative donor exons (Figure 59) seems to be narrower than that of alternative acceptor exons (Figure 60) too. The biological significance of this is not clear.

Type	Number	Min	1 st Qu	Median	Mean	3 rd Qu	Max	IQR
Single constitutive exons	2374	144	582	897	1140	1461	27730	879
Multi constitutive exons	461001	3	154	270	464	561	27510	407
Multi constitutive initiation exons	8895	3	129	230	362	443	27510	314
Multi constitutive internal exons	27178	7	143	219	387	434	14580	291
Multi constitutive termination exons	10028	5	312	567	763	987	8229	675
Cassette exons	1228	3	90	150	279	262	10500	172
5' introns flanking cassette exons	1368	40	207	653	2274	1957	86250	1750
3' introns flanking cassette exons	1268	40	166	639	2678	2423	60920	2257
Alternative acceptor exons	905	26	156	306	494	615	4429	459
Alternative donor exons	1399	13	96	172	275	308	10500	212
Intron retained exons	983	92	446	720	971	1255	5769	809
5' exons flanking retained intron	983	15	85	163	271	304	3639	219
3' exons flanking retained intron	983	18	138	261	429	534	5288	396
Retained introns	983	32	64	101	207	232	2503	168
Constitutive introns	36589	40	60	69	690	237	108400	177

Table 21. Basic statistical measures of the exons and introns lengths. The number of exons and introns involved is listed together with the minimum, 1st quartile, median, mean, 3rd quartile, maximum and inter-quartile range.

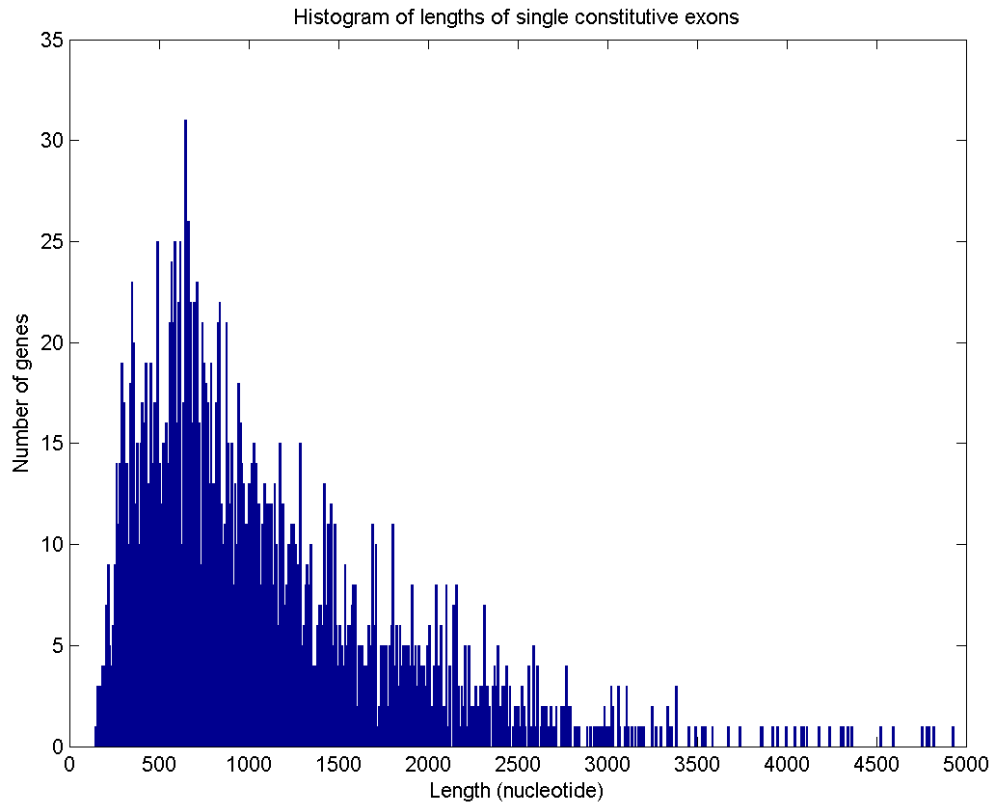


Figure 46. Histogram of single constitutive exons. Histogram is limited to exons of length 5000 nucleotides and below.

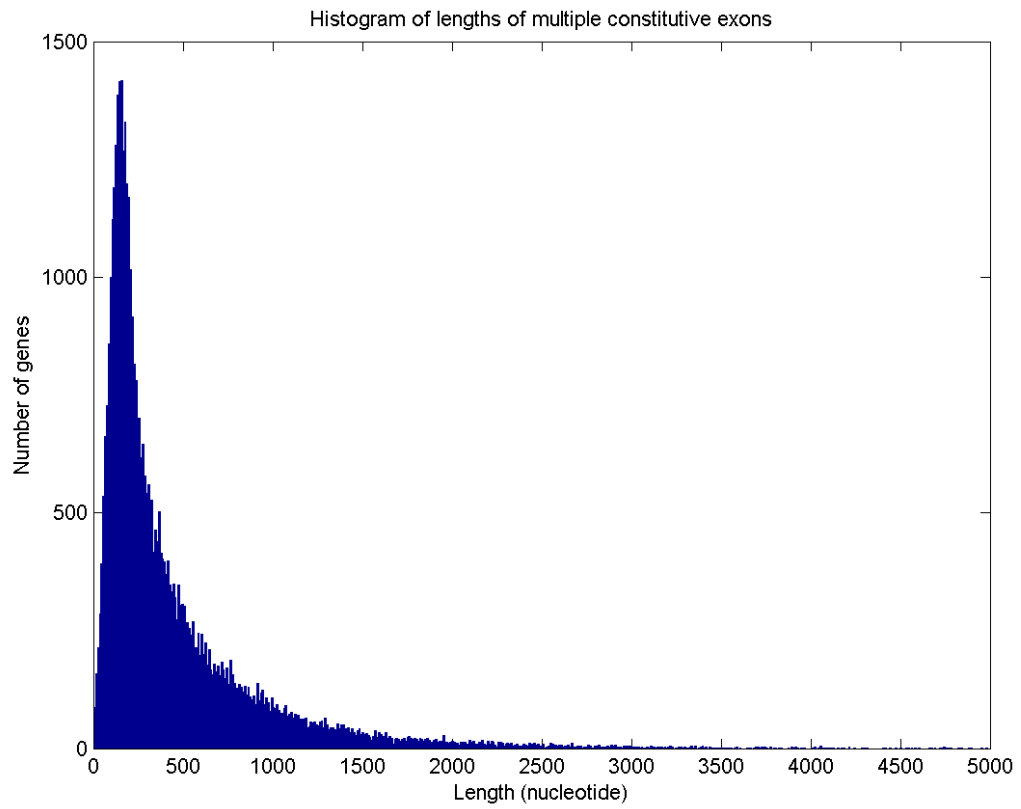


Figure 47. Histogram of multiple constitutive exons. Histogram is limited to exons of length 5000 nucleotides and below.

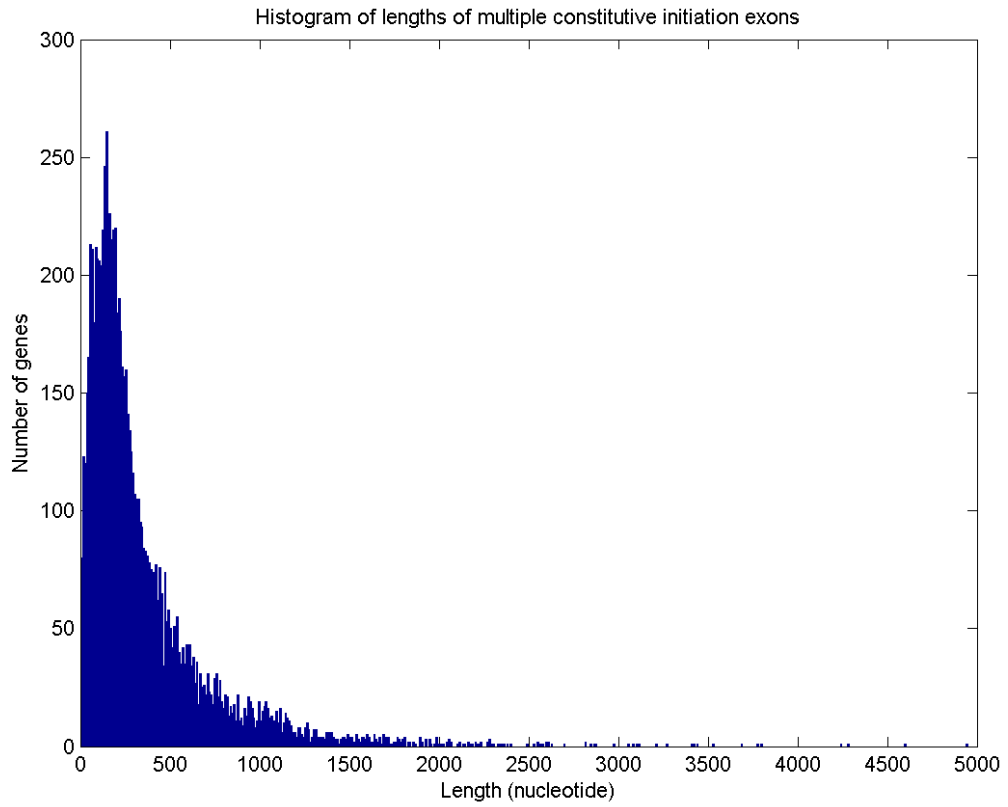


Figure 48. Histogram of multiple constitutive initiation exons. Histogram is limited to exons of length 5000 nucleotides and below.

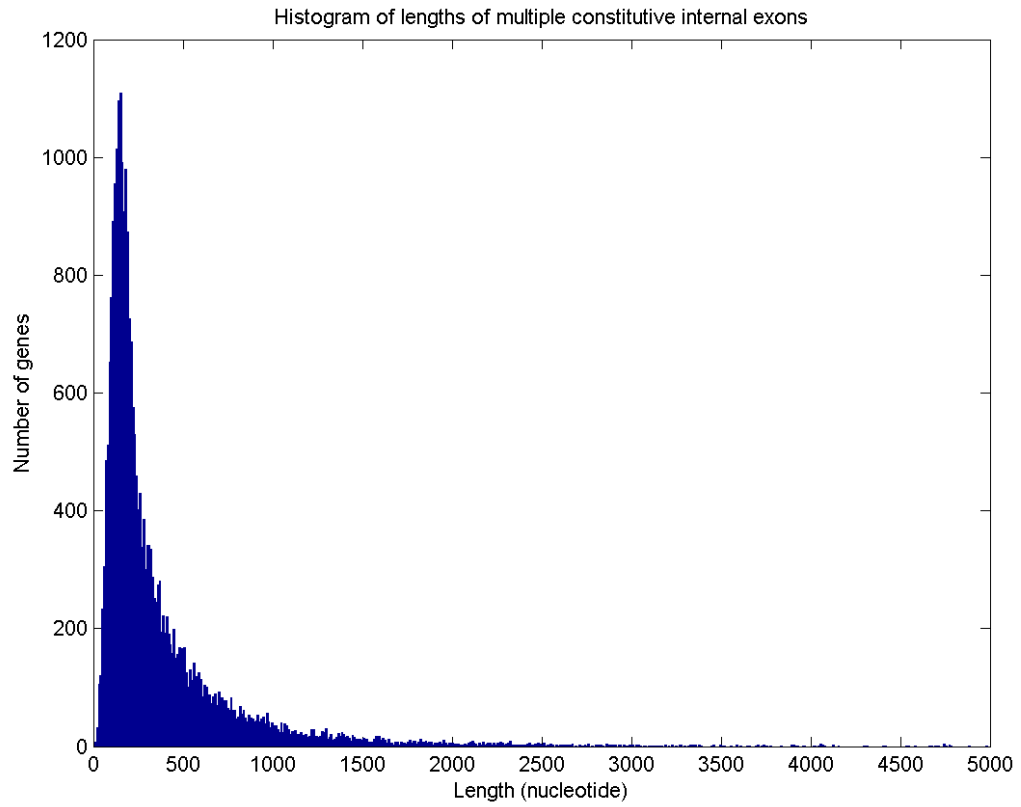


Figure 49. Histogram of multiple constitutive internal exons. Histogram is limited to exons of length 5000 nucleotides and below.

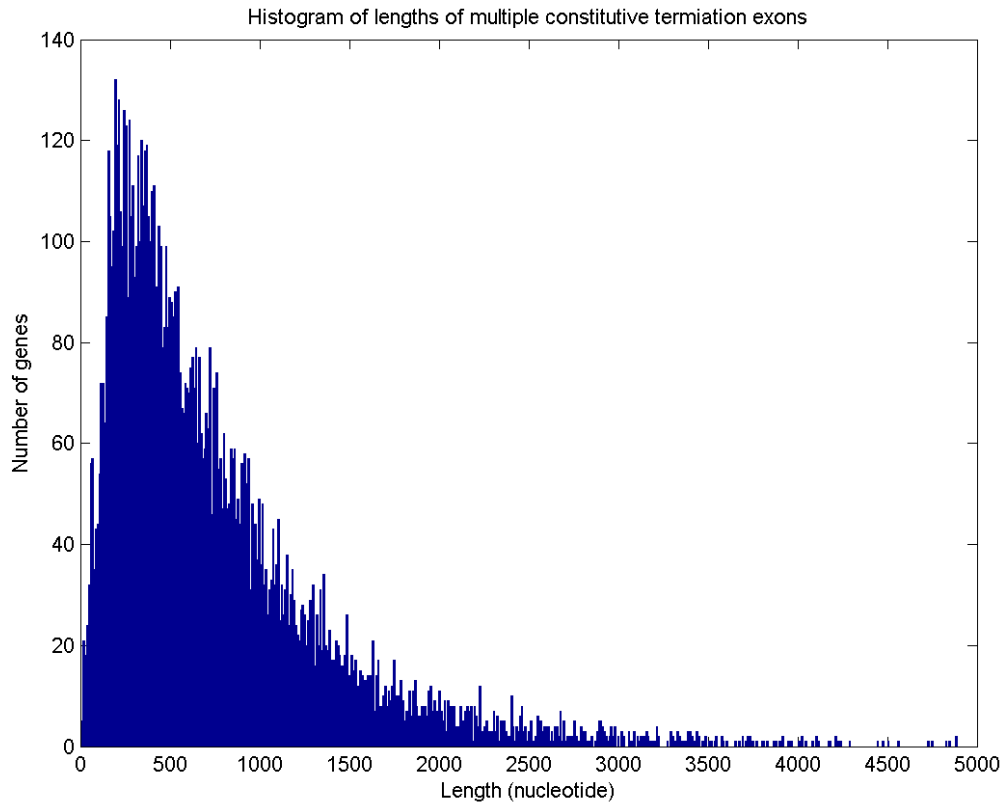


Figure 50. Histogram of multiple constitutive termination exons. Histogram is limited to exons of length 5000 nucleotides and below.

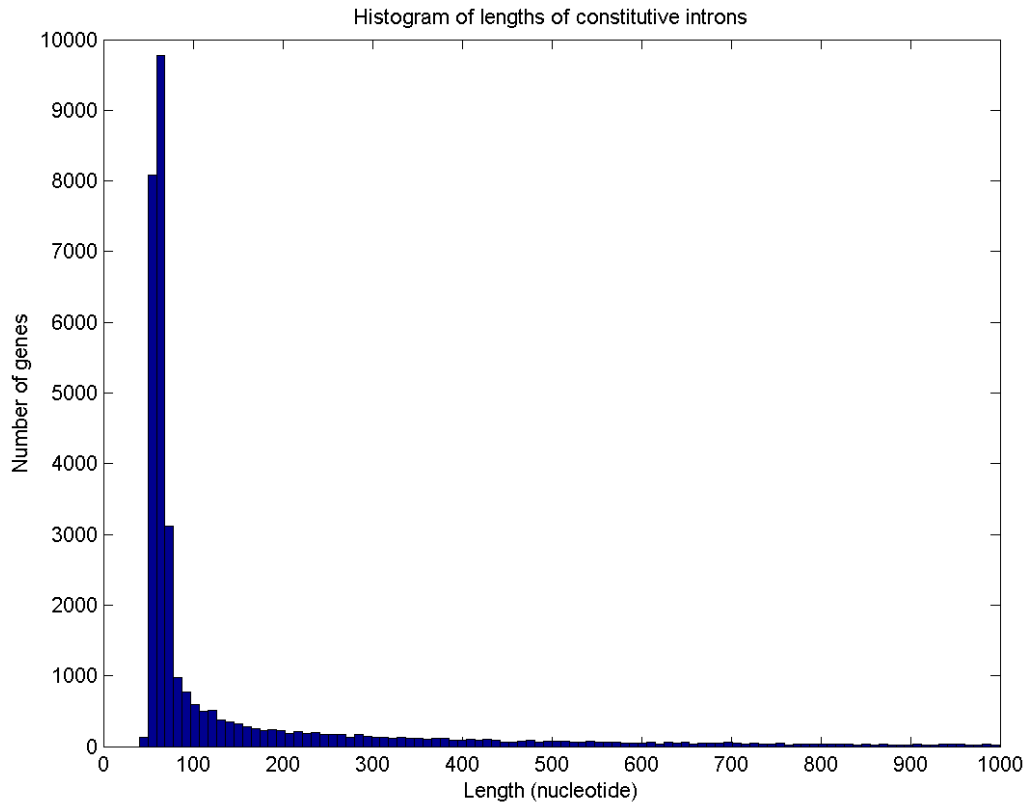


Figure 51. Histogram of constitutive introns. Histogram is limited to introns of length 1000 nucleotides and below.

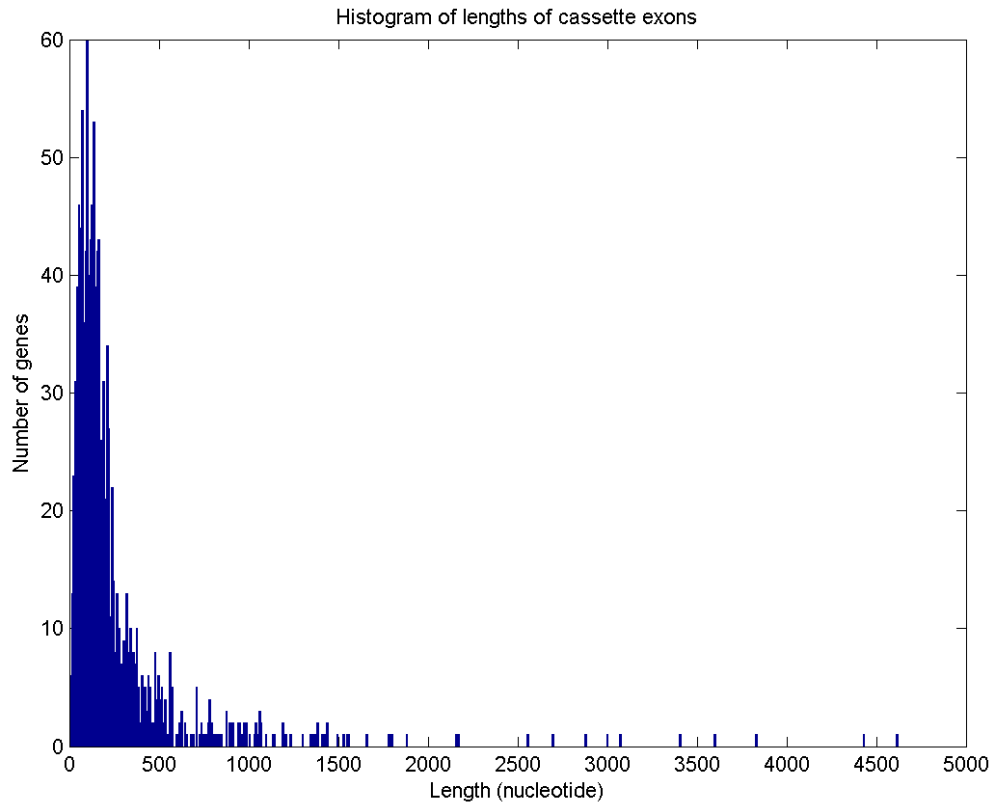


Figure 52. Histogram of cassette exons. Histogram is limited to exons of length 5000 nucleotides and below.

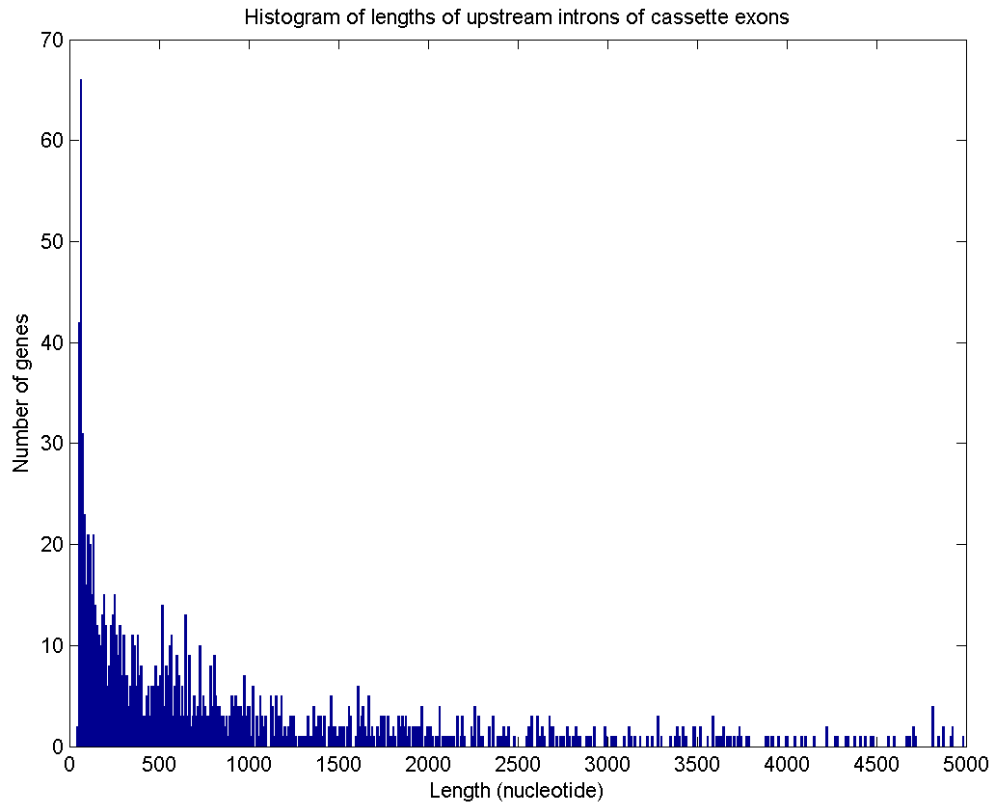


Figure 53. Histogram of lengths of upstream introns of cassette exons. Histogram is limited to exons of length 5000 nucleotides and below.

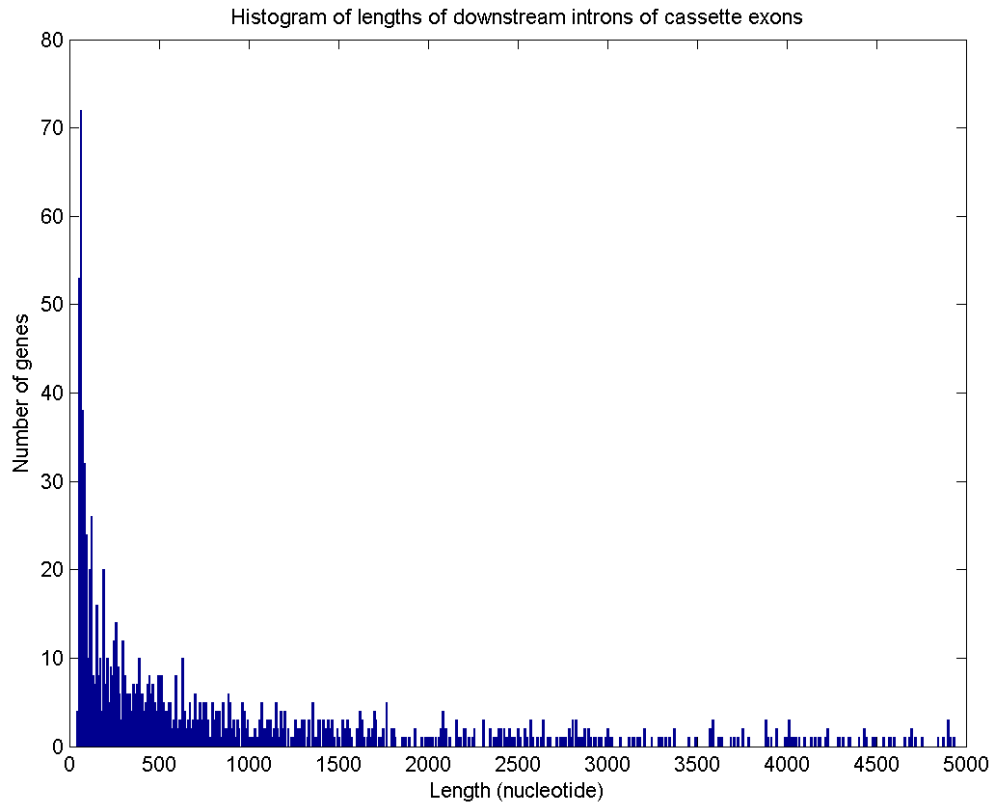


Figure 54. Histogram of lengths of downstream introns of cassette exons. Histogram is limited to exons of length 5000 nucleotides and below.

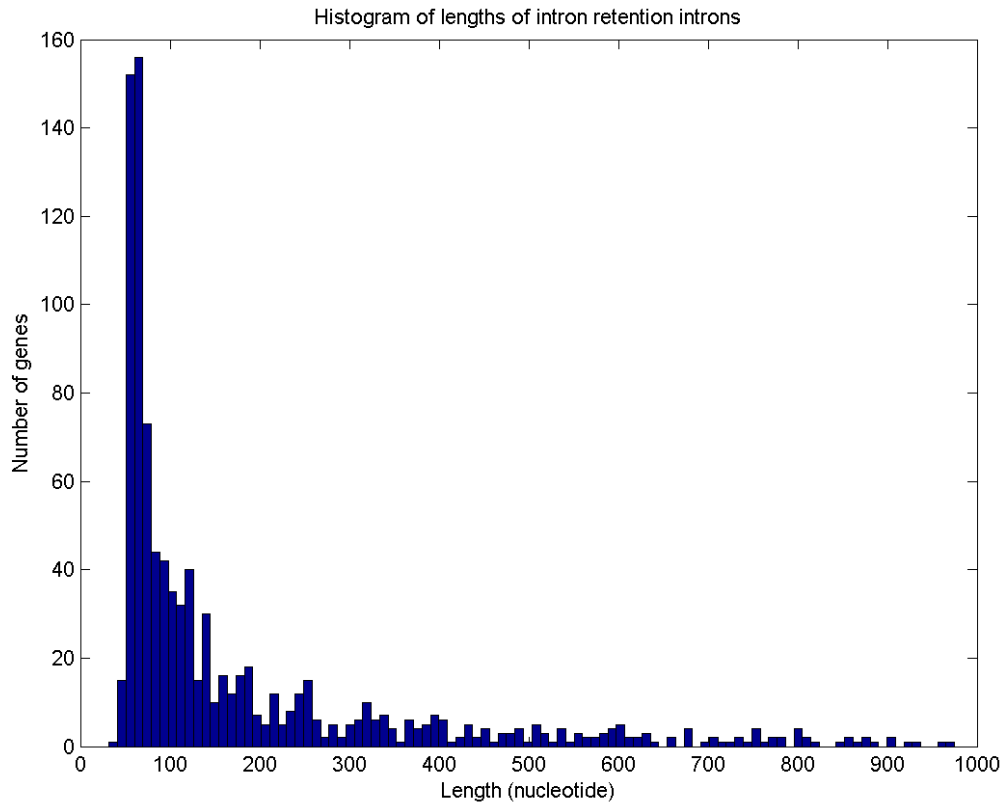


Figure 55. Histogram of lengths of intron retention introns. Histogram is limited to exons of length 1000 nucleotides and below.

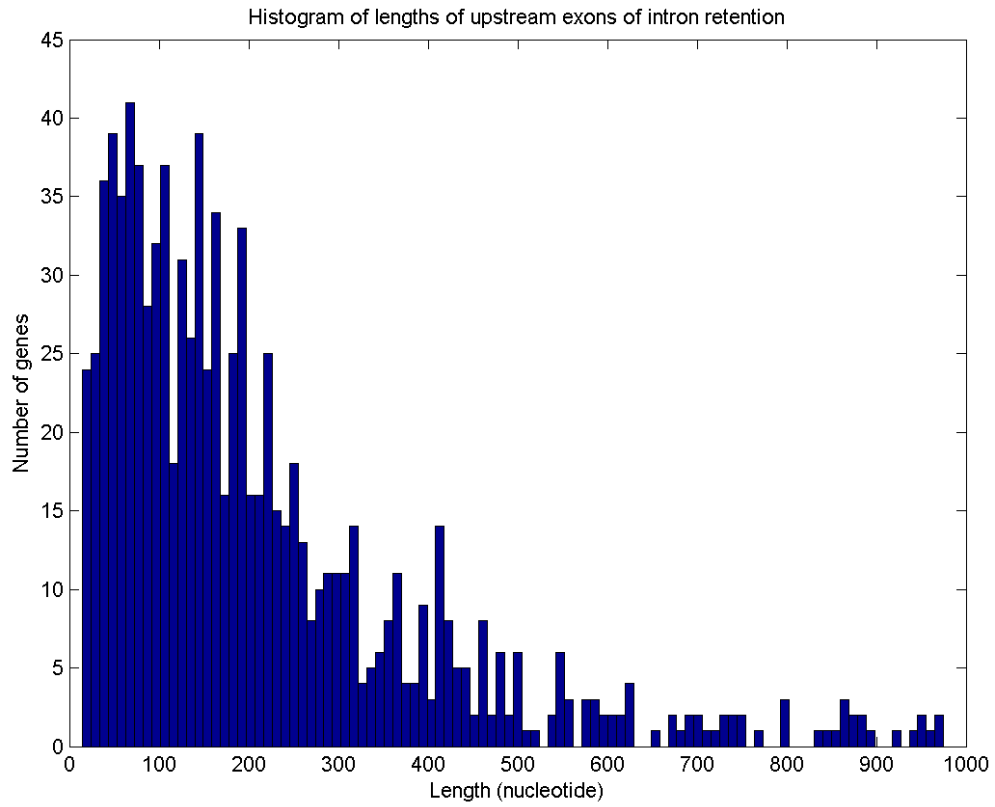


Figure 56. Histogram of lengths of upstream exons of intron retention. Histogram is limited to exons of length 1000 nucleotides and below.

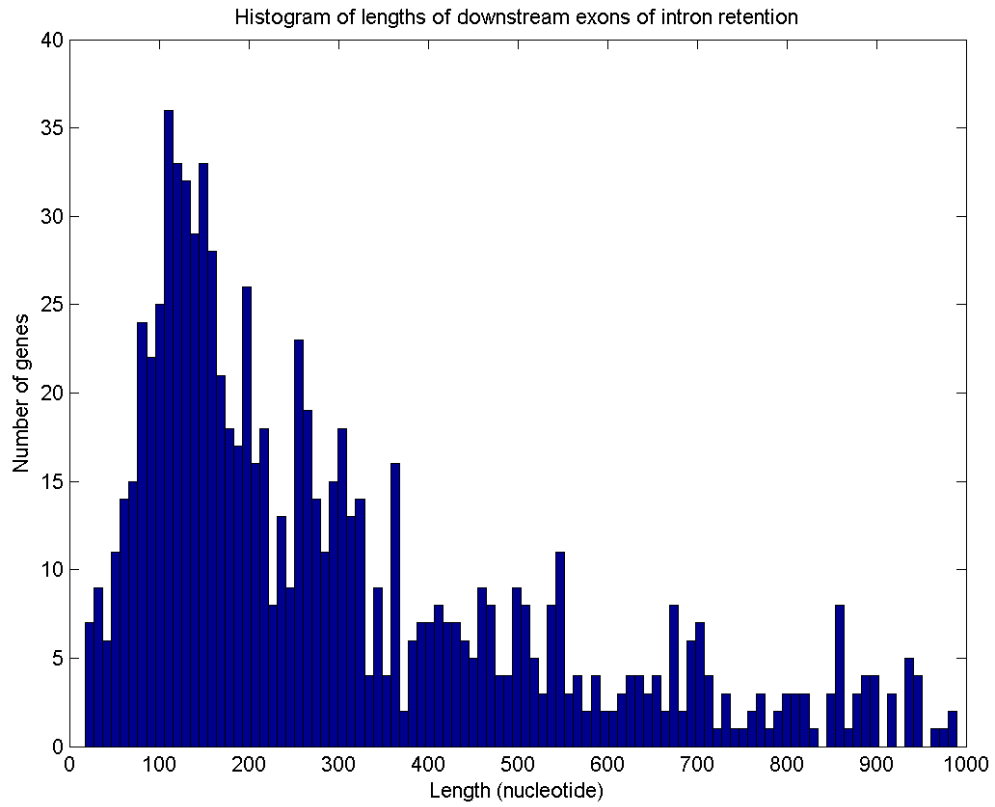


Figure 57. Histogram of lengths of downstream exons of intron retention. Histogram is limited to exons of length 1000 nucleotides and below.

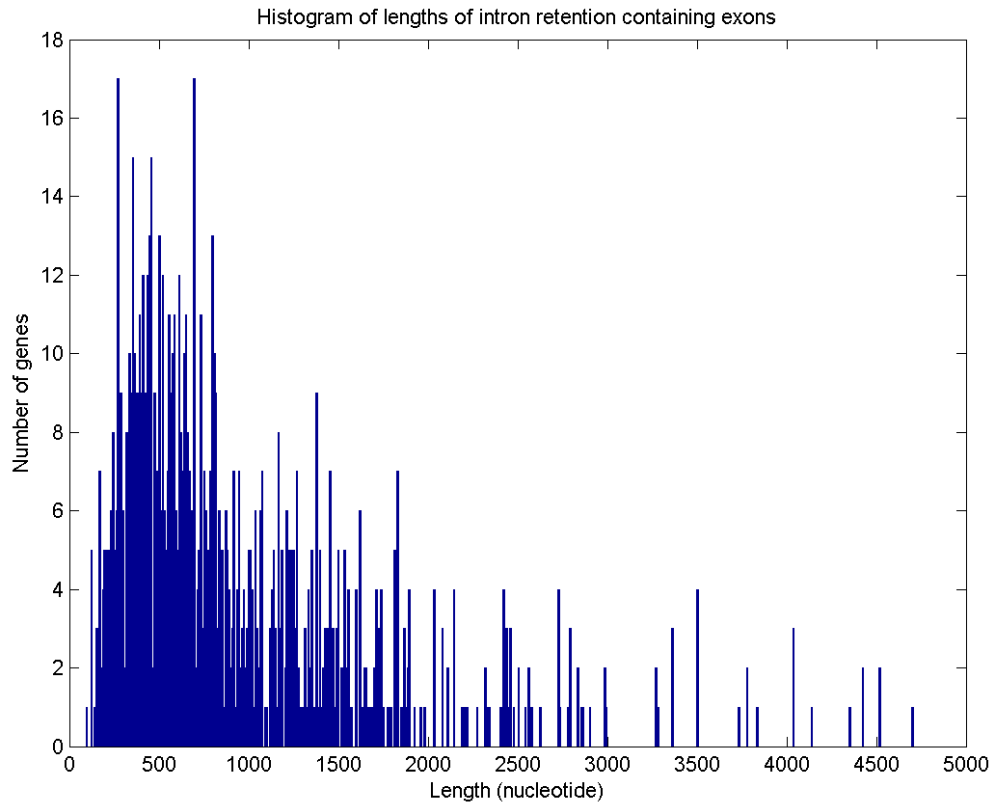


Figure 58. Histogram of lengths of intron retention containing exons. Histogram is limited to exons of length 5000 nucleotides and below.

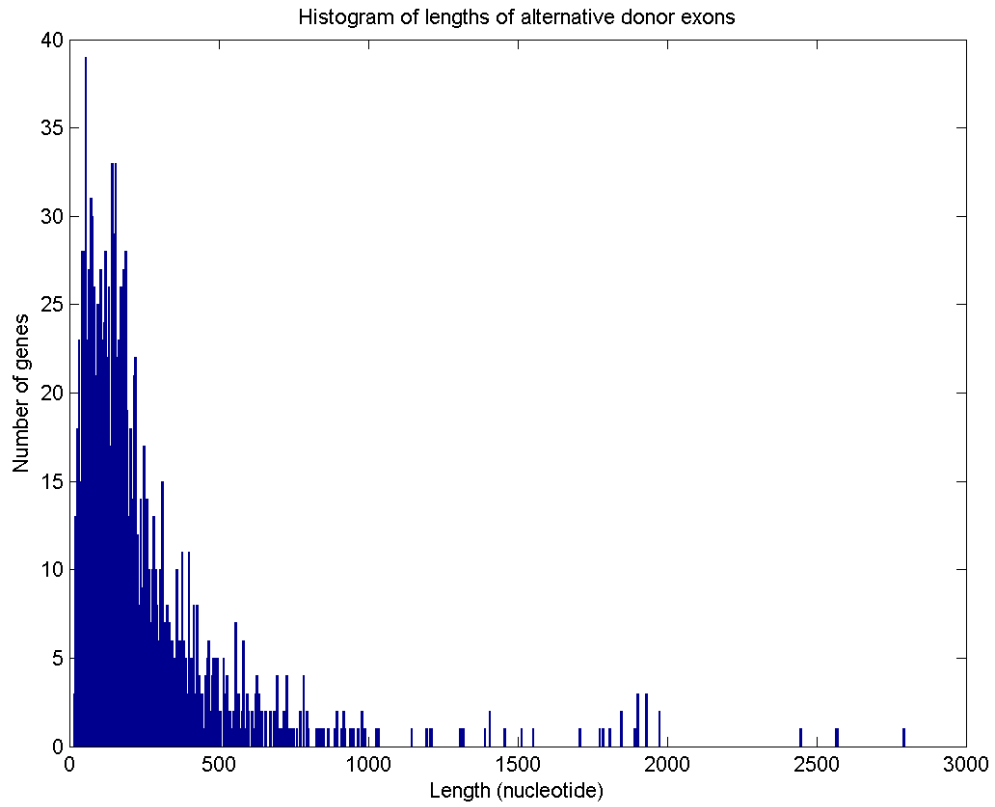


Figure 59. Histogram of lengths of alternative donor exons. Histogram is limited to exons of length 5000 nucleotides and below.

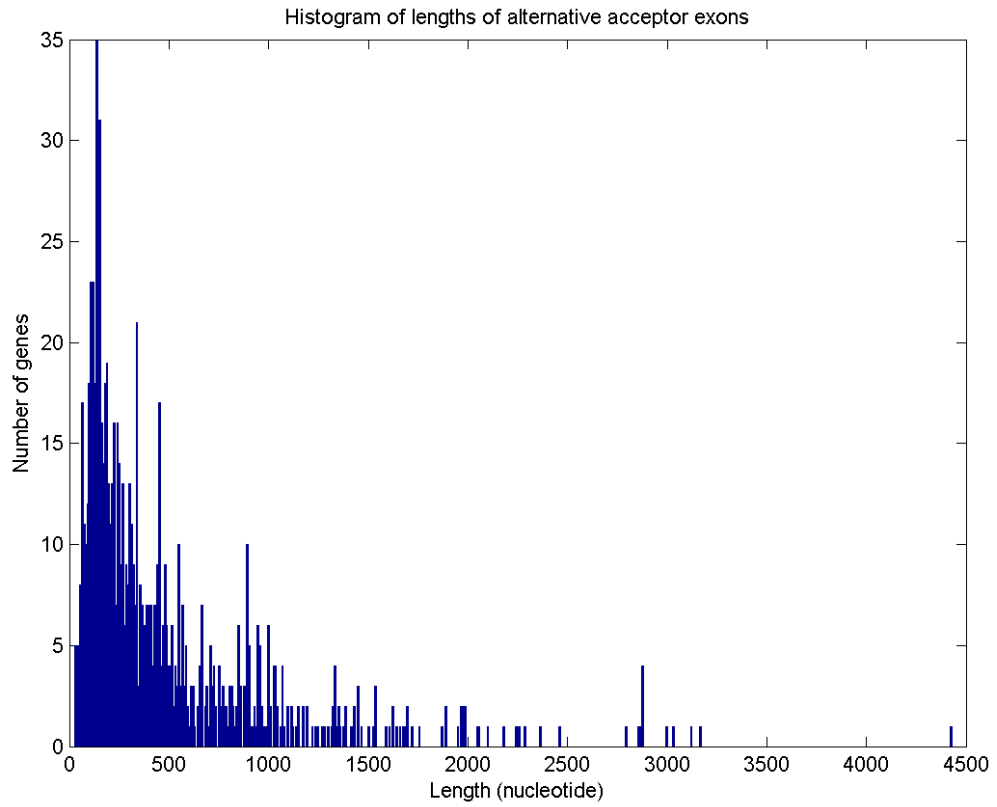


Figure 60. Histogram of lengths of alternative acceptor exons. Histogram is limited to exons of length 5000 nucleotides and below.

4.3.3 Exon number analysis

The data contained in Table 22 simply shows that alternative splicing largely (> 75%) occurs in genes with more than four exons while genes with less than 4 exons exhibits no alternative splicing 75% of the time. This is also reflected in the histograms. The histogram for constitutively spliced genes (Figure 61) shows that the majority of the genes have few exons, while the histogram for the alternatively spliced genes (Figure 62) show the opposite, that these genes have larger numbers of exons. This is logical as alternative splicing is not possible with two or less exons and the amount of alternative splicing variants increases dramatically with the number of exons. This also means that an alternative splicing prediction program can predict with 75% accuracy that a gene in *Drosophila melanogaster* with four or more exons is alternatively spliced. The converse is also true, that 75% of the time the prediction that a gene with four or less gene is not alternatively spliced is accurate. This means that the number of exons could be a useful feature for the design of alternative splicing prediction algorithms.

Type	Number	Min	1st Qu	Median	Mean	3rd Qu	Max
Constitutive spliced genes	10599	1	2	3	4	4	80
Alternatively spliced genes	7557	1	4	6	7	8	46

Table 22. Basic statistical measures for the number of exons for alternatively spliced genes and constitutively spliced genes.

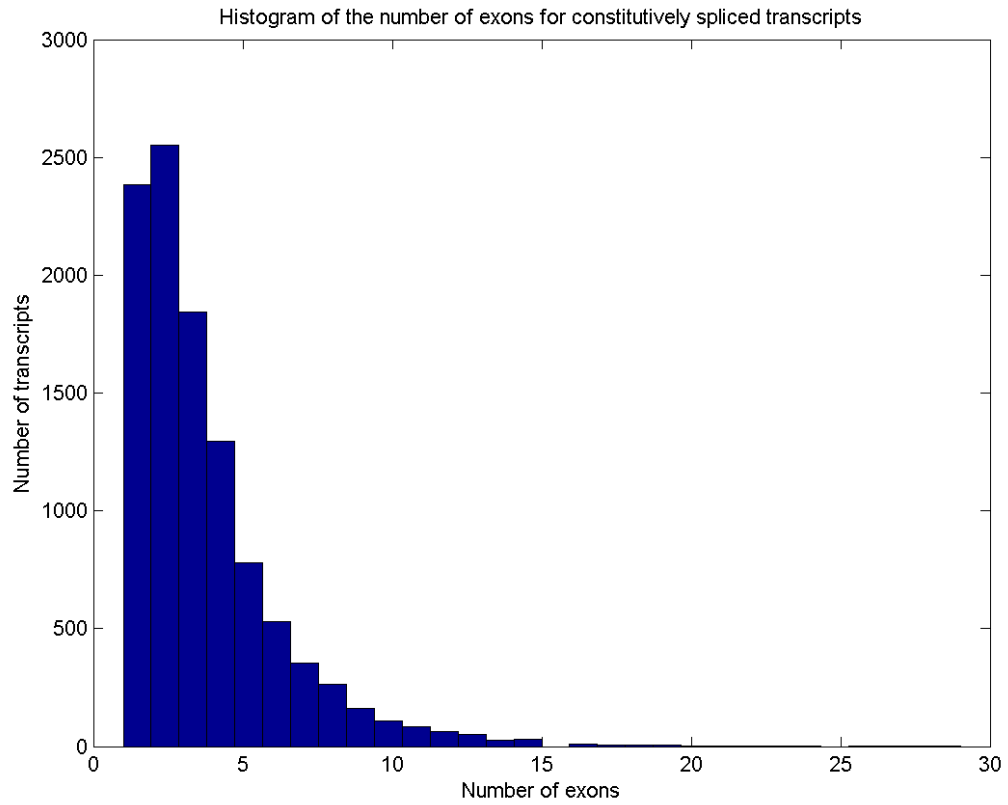


Figure 61. Histogram of the number of exons of constitutively spliced transcripts. The histogram is limited to transcripts having 30 or less exons.

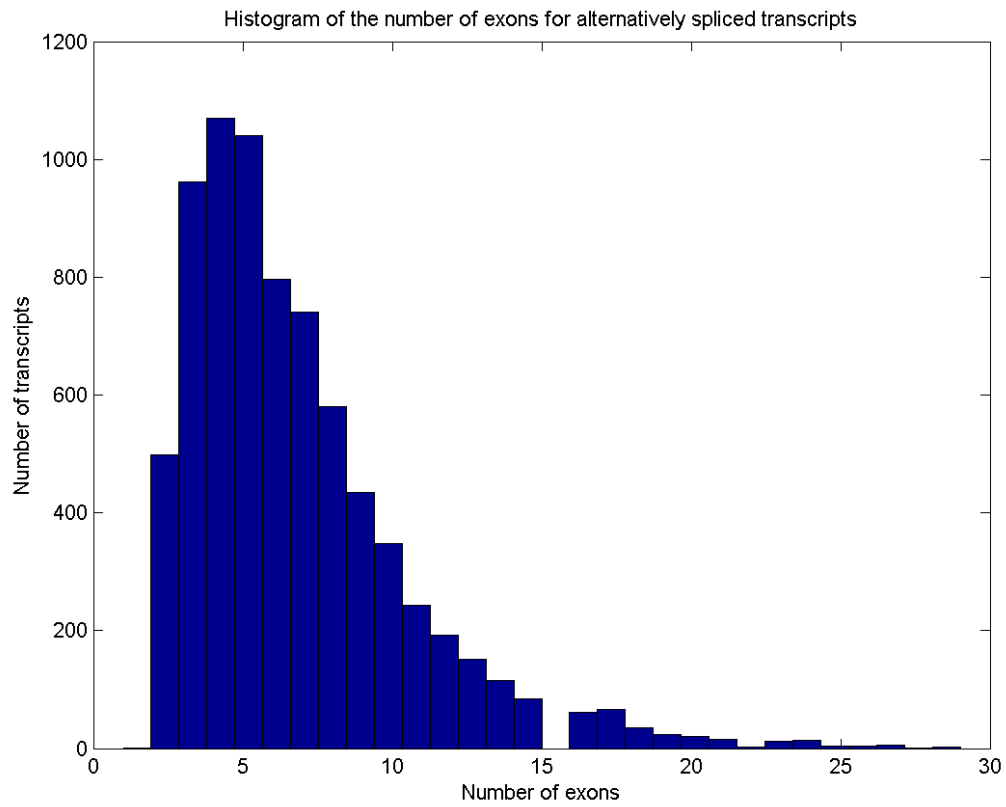


Figure 62. Histogram of the number of exons for alternatively spliced transcripts. The histogram is limited to transcripts having 30 or less exons.

4.3.4 Nucleotide composition analysis

The data for composition shown in Table 23 shows quite a fair bit of fluctuations. The most predominant trend observed is that introns have high thymine percentage whereas exons have a low thymine percentage. There is also a high abundance of adenine in both intron and exons. Introns also exist low GC content in comparison with exons. The exception to this is alternative donor site, which does not have an obvious low thymine percentage. Other variations of the various types of exons and introns observed could be due to splicing enhancers and suppressors. The nucleotide motifs of these protein factors could have altered the nucleotide composition or the variation observed could be due to noise.

Type	Nuc	All	1st qu.	2nd qu.	3rd qu.	4th qu.
Alternative acceptor exons	a	1.129	1.146	1.141	1.135	1.122
	c	1.050	1.021	1.038	1.065	1.049
	g	0.969	0.992	0.998	0.954	0.968
	t	0.852	0.841	0.823	0.846	0.861
	n	0.000	0.000	0.000	0.000	0.000
Alternative donor exons	a	1.172	1.143	1.186	1.172	1.172
	c	0.921	0.854	0.856	0.880	0.953
	g	0.939	0.885	0.922	0.932	0.949
	t	0.969	1.118	1.036	1.016	0.926
	n	0.000	0.000	0.000	0.000	0.000
Cassette exons	a	1.158	1.203	1.150	1.139	1.161
	c	1.044	0.921	0.975	1.012	1.072
	g	0.962	0.912	0.951	0.946	0.972
	t	0.836	0.964	0.923	0.904	0.795
	n	0.000	0.000	0.000	0.000	0.000
Constitutive introns	a	1.178	1.215	1.199	1.202	1.175
	c	0.805	0.688	0.721	0.756	0.812
	g	0.787	0.687	0.678	0.683	0.797
	t	1.229	1.410	1.402	1.360	1.214
	n	0.000	0.000	0.000	0.000	0.000
Retained introns	a	1.172	1.215	1.151	1.131	1.182
	c	0.866	0.824	0.848	0.884	0.868
	g	0.796	0.760	0.797	0.803	0.797
	t	1.166	1.201	1.204	1.181	1.152
	n	0.000	0.000	0.000	0.000	0.000
Multi constitutive exons	a	1.080	1.085	1.050	1.061	1.091
	c	1.019	0.976	1.014	1.023	1.022
	g	1.005	0.995	1.013	1.014	1.001
	t	0.897	0.944	0.923	0.902	0.887
	n	0.000	0.000	0.000	0.000	0.000
Multi constitutive initiation exons	a	1.102	1.159	1.153	1.115	1.083
	c	0.994	0.872	0.913	0.968	1.029
	g	0.996	0.911	0.928	0.977	1.024
	t	0.908	1.058	1.007	0.939	0.864
	n	0.000	0.000	0.000	0.000	0.000
Multi constitutive internal exons	a	1.025	1.062	1.013	1.006	1.030
	c	1.067	1.002	1.046	1.071	1.077
	g	1.061	1.023	1.041	1.067	1.067
	t	0.847	0.913	0.900	0.857	0.827

	n	0.000	0.000	0.000	0.000	0.000
Multi constitutive termination exons	a	1.145	1.154	1.138	1.130	1.153
	c	0.963	0.925	0.948	0.960	0.972
	g	0.930	0.911	0.919	0.933	0.934
	t	0.962	1.010	0.995	0.978	0.940
	n	0.000	0.000	0.000	0.000	0.000
Single constitutive exons	a	1.087	1.035	1.059	1.069	1.115
	c	1.010	1.059	1.024	1.015	0.994
	g	1.001	1.042	1.000	1.010	0.989
	t	0.902	0.864	0.917	0.906	0.902
	n	0.000	0.000	0.000	0.000	0.000

Table 23. Table showing the nucleotide composition of various types of exons and introns. The overall nucleotide composition as well as the nucleotide composition broken down by length into four quartiles are provided.

4.3.5 Splicing motif analysis

Most of the splicing motifs found in DEDB are of the GT-AG type making up 99.264% of all splicing motifs (Table 24). GT-AG type splicing motifs may be spliced either by the U2 type or by U12 type spliceosome. However, the U2 type spliceosome is more abundant and therefore most of the GT-AG type splicing motifs will be spliced by this type of spliceosome. The U2 type spliceosome is also responsible for the splicing of GC-AG type splicing motifs while the U12 type spliceosome in addition to the GT-AG splicing motif also recognize the AT-AC type splicing motifs. The percentages of the various types of splicing motif found in fruitfly are quite similar to most other eukaryotes being predominantly GT-AG type.

Motif	Number	Percentage
GT-AG	45751	99.264%
GC-AG	278	0.603%
AT-AC	9	0.020%
Others	52	0.113%

Table 24. Percentage of the various types of splicing motifs found in DEDB.

To further analyze the acceptor and donor splicing site motifs, we have opted to use information content as a measure of the level of sequence conservation found in these sites. This analysis only includes GT-AG type splice motifs. By measuring the information content of splice site motifs which are assumed to contain all the necessary information required for splicing, we can then assess the difference in the information content of the splice site motifs which are involved in alternative splicing. We have conducted the analysis on the assumption that the multi constitutive internal exon acceptor and donor splice site motifs contain all the information required for normal splicing. MCIE contains mean information content of 9.745 bits and 8.525 bits for the acceptor and donor site respectively (Table 25). There is some variation in the information content with respect to the length of the exon. Exons in the first and last quartiles show more conservation than the second and third quartiles. A possible explanation is that short and long exons require more signals from the splice motifs to compensate for their extreme lengths so that they can be spliced effectively. However the splicing motifs itself does not change as reflected by Figure 63, Figure 64, Figure 65, Figure 66, Figure 67, Figure 68, Figure 69, Figure 70, Figure 71 and Figure 72. The MCIE acceptor motif consists of a highly conserved AG dinucleotide at the -2 and -1 position. The -3 position consist of either a cytosine or thymine. Interestingly, there is a total lack of conservation in the -4 position. Positions -5 to -12 shows weak conservation of cytosine and thymine while positions -13 to -30 shows weak conservation in adenine and thymine. Positions -5 to -30 shows a overall weak preference for the thymine. There is absolutely no conservation in the exonic region. This means that the exonic sequences are not under any selection pressure due to splicing. The motifs for

the four quartiles show the same pattern. MCIE donor sites too exhibit a relatively normal distribution although it is slightly skewed to the right with a mean of 8.525 bits. This is lower than that of the acceptor site. Like the acceptor site, there is an increase in the information content for the first and fourth quartile. Unlike the acceptor site, there is some level of conservation in the exonic sequence at positions +1 and +2 with more conservation at +1. There is high conservation in the -1 and -2 positions, in the form of a GT dinucleotide. This is followed by conservation in positions -3 to -6. Position -3 shows a preference for either adenine or guanine. Adenine is the preferred nucleotide at position -4. Position -5 shows a strong conservation for guanine, being present more than half the time while position -6 shows a dominance of thymine. Therefore, unlike the acceptor site, which has the information content spread over a longer region, the donor site has several well-conserved residues instead.

Motif type	Number	All	1st qu.	2nd qu.	3rd qu.	4th qu.
Multi constitutive internal exon acceptor site	27163	9.745	9.814	9.544	9.571	10.090
Multi constitutive internal exon donor site	27039	8.525	8.644	8.364	8.396	8.728

Table 25. Information content of multi constitutive internal exon acceptor and donor sites. The number of sites, the information content for all the sites as well as the information content of individual quartiles based on exon length are displayed.

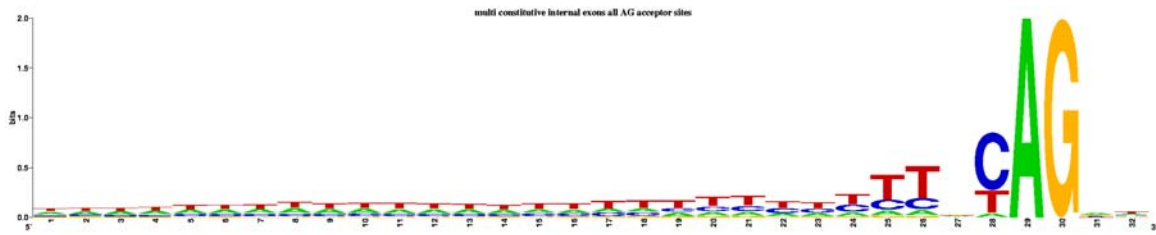


Figure 63. Sequence logo of multi constitutive internal exon acceptor sites. The region -30 to +2 is shown. The information content is 9.745 bits.

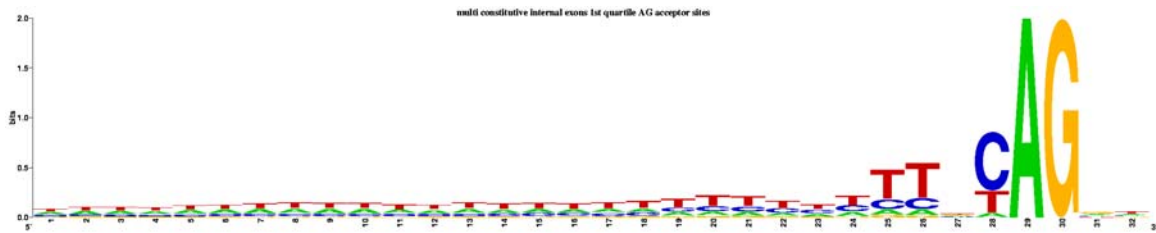


Figure 64. Sequence logo of multi constitutive internal exon acceptor sites. Only the exons in the exon length first quartile are used. The region -30 to +2 is shown. The information content is 9.814 bits.

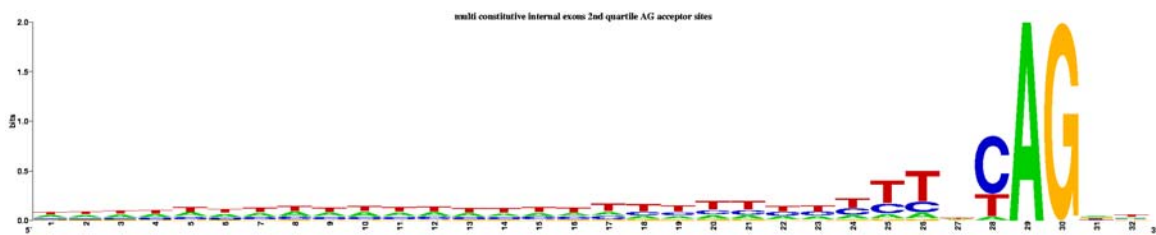


Figure 65. Sequence logo of multi constitutive internal exon acceptor sites. Only the exons in the exon length second quartile are used. The region -30 to +2 is shown. The information content is 9.544 bits.

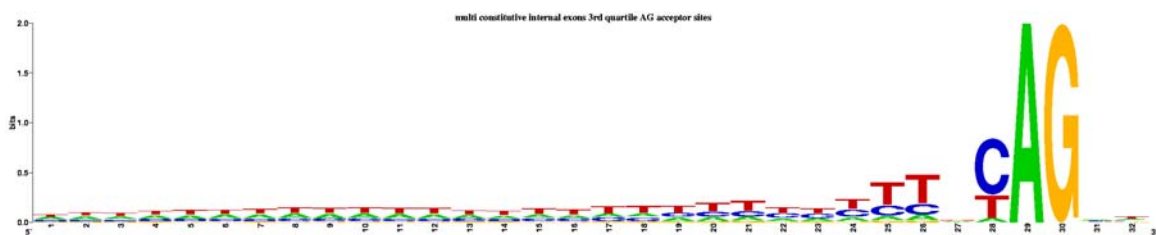


Figure 66. Sequence logo of multi constitutive internal exon acceptor sites. Only the exons in the exon length third quartile are used. The region -30 to +2 is shown. The information content is 9.571 bits.

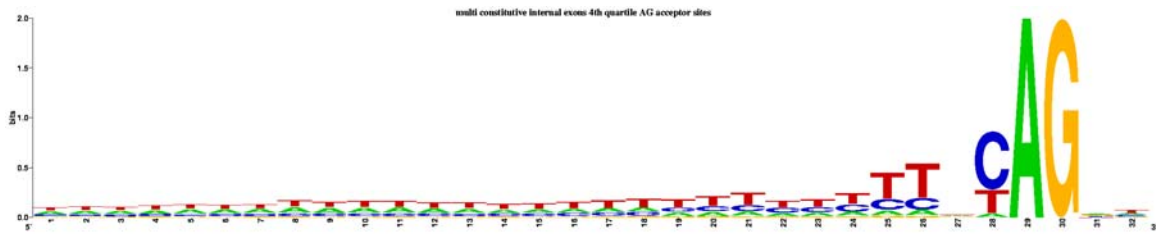


Figure 67. Sequence logo of multi constitutive internal exon acceptor sites. Only the exons in the exon length fourth quartile are used. The region -30 to +2 is shown. The information content is 10.090 bits.



Figure 68. Sequence logo of multi constitutive internal exon donor sites. The region +5 to -10 is shown. The information content is 8.525 bits.

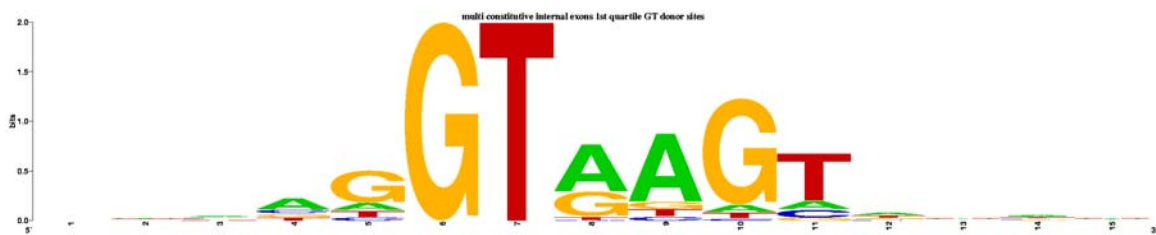


Figure 69. Sequence logo of multi constitutive internal exon donor sites. Only the exons in the exon length first quartile are used. The region +5 to -10 is shown. The information content is 8.644 bits.

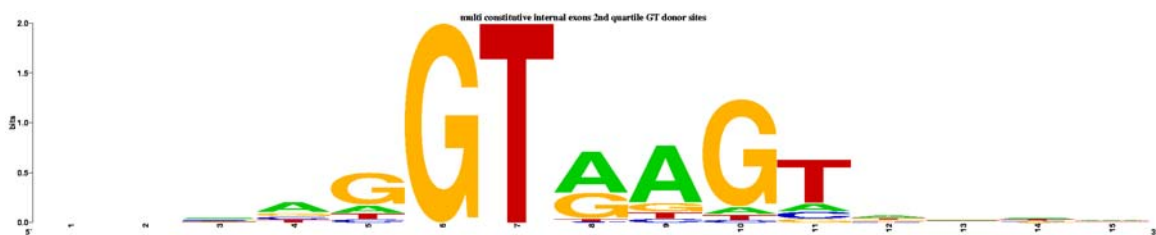


Figure 70. Sequence logo of multi constitutive internal exon donor sites. Only the exons in the exon length second quartile are used. The region +5 to -10 is shown. The information content is 8.364 bits.

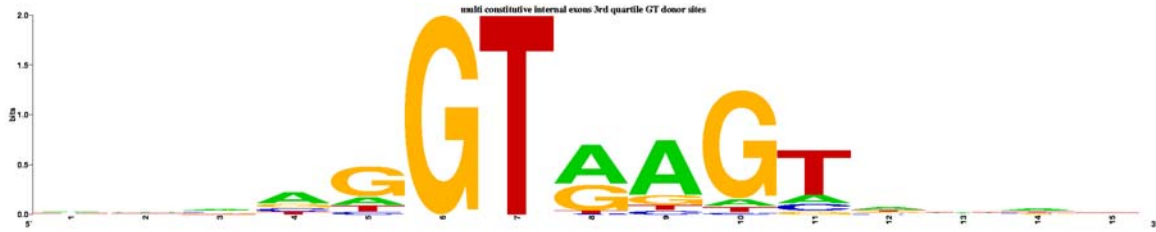


Figure 71. Sequence logo of multi constitutive internal exon donor sites. Only the exons in the exon length third quartile are used. The region +5 to -10 is shown. The information content is 8.396 bits.

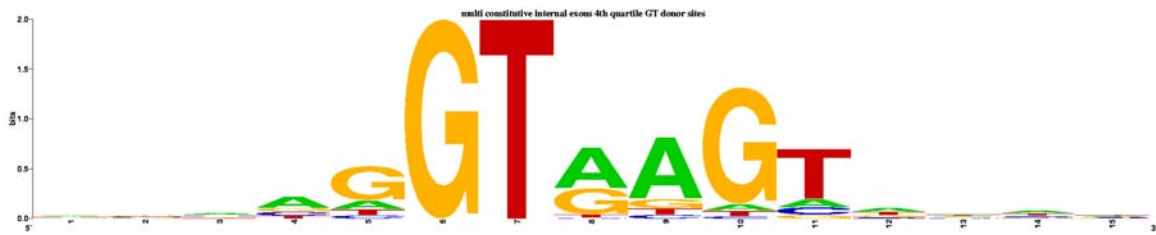


Figure 72. Sequence logo of multi constitutive internal exon donor sites. Only the exons in the exon length fourth quartile are used. The region +5 to -10 is shown. The information content is 8.728 bits.

MCIE acceptor and donor site motifs were used to construct matrices that allows for the computation of individual information content (IIC). The mean of the IIC of the acceptor and donor site should be equal to the information content of the acceptor and donor site calculated above. This is the case as observed from Table 25 and Table 26 with minor differences due to a lost of precision of floating point calculation. The histograms of MCIE acceptor and donor sites IIC (Figure 73 and Figure 74) looks similar to a normal distribution that is skewed slightly to the right.

Type	No.	Min.	1 st qu.	Median	Mean	3 rd qu.	Max.	IQR
Cassette exon acceptor	1227	-10.840	3.978	7.258	7.079	10.400	22.240	6.422
Retained intron acceptor	980	-18.270	3.414	6.874	6.188	9.703	19.000	6.289
Alternative acceptor	897	-15.770	3.283	5.964	5.951	9.420	19.100	6.137
Multi constitutive internal exon acceptor	27163	-17.160	7.358	9.940	9.744	12.410	23.250	5.052
Multi constitutive termination exon acceptor	10023	-14.500	7.663	10.290	9.974	12.650	22.600	4.987
Cassette exon donor	1201	-9.238	4.791	7.151	6.762	9.229	13.880	4.438
Retained intron donor	953	-8.471	3.243	6.112	5.389	8.091	14.600	4.848
Alternative donor	1368	-8.360	3.693	5.969	5.608	7.594	13.680	4.261
Multi constitutive internal exon donor	27039	-10.240	7.139	8.915	8.517	10.380	15.180	3.241
Multi constitutive initiation exon donor	8848	-10.520	6.518	8.465	7.949	10.050	14.980	3.532

Table 26. Table showing the basic statistical measures of the individual information content of various splice site motifs. The number of splice site motifs involved (No.), the minimum (Min.), the first quartile (1st qu.), the median (Median), the mean (Mean), the third quartile (3rd qu.), the maximum (Max.) and the inter-quartile range (IQR) are shown.

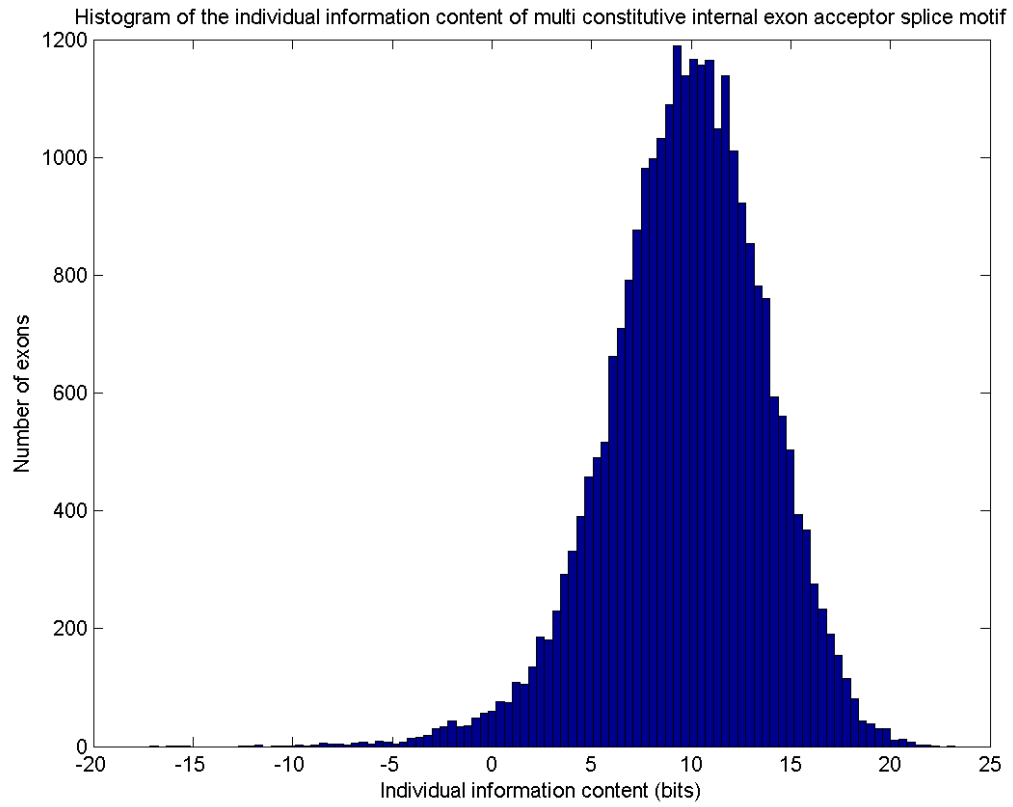


Figure 73. Histogram of the individual information content of multi constitutive internal exon acceptor splice motif.

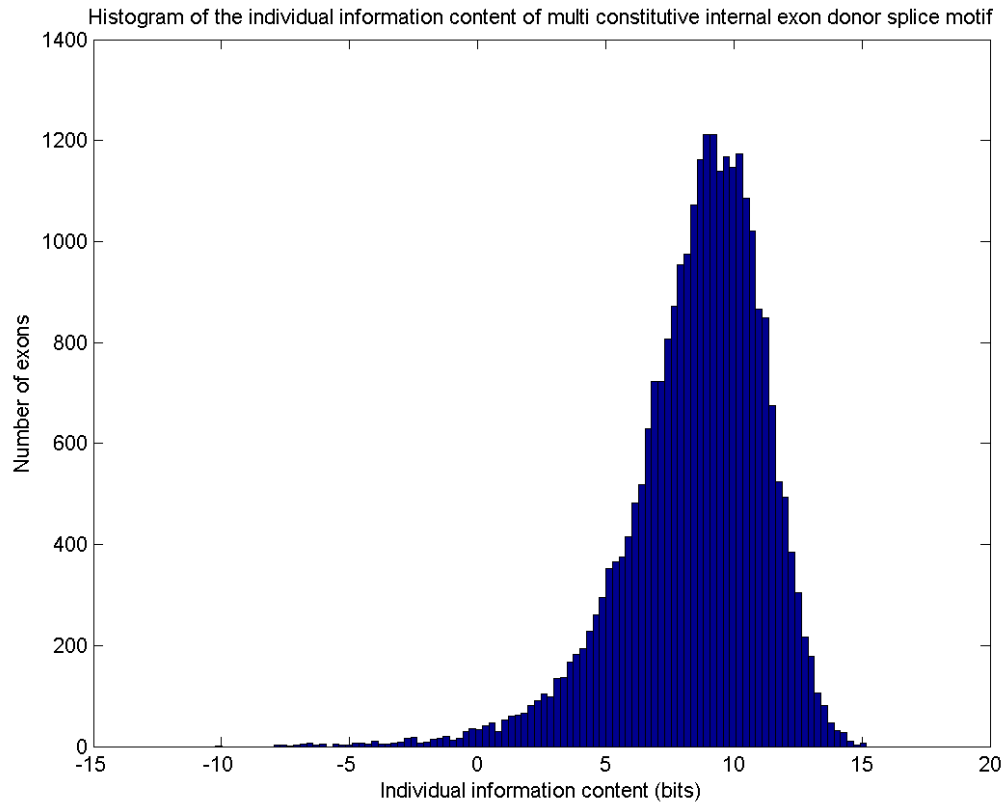


Figure 74. Histogram of the individual information content of multi constitutive internal exon donor splice motif.

Cassette exons have lower mean IIC as compared to MCIE for both acceptor (7.079 bits) and donor (6.762 bits) sites. This is to be expected, as cassette exons are the result of a failure of recognition of either or both splicing motifs. However, the distribution of the IIC shown in Figure 75 and Figure 76 for the acceptor and donor splice site respectively is such that a portion of the splice sites motifs of cassette exons does exhibit IIC higher than that of MCIE. In these cases, action by ISS or ESS could be at work resulting in the lost of recognition. For cassette exons with low IIC, ESE and ISE could be at play with allows for the recognition of the splicing sites. The inter-quartile range (IQR) of cassette exons is also larger than that of MCIE.

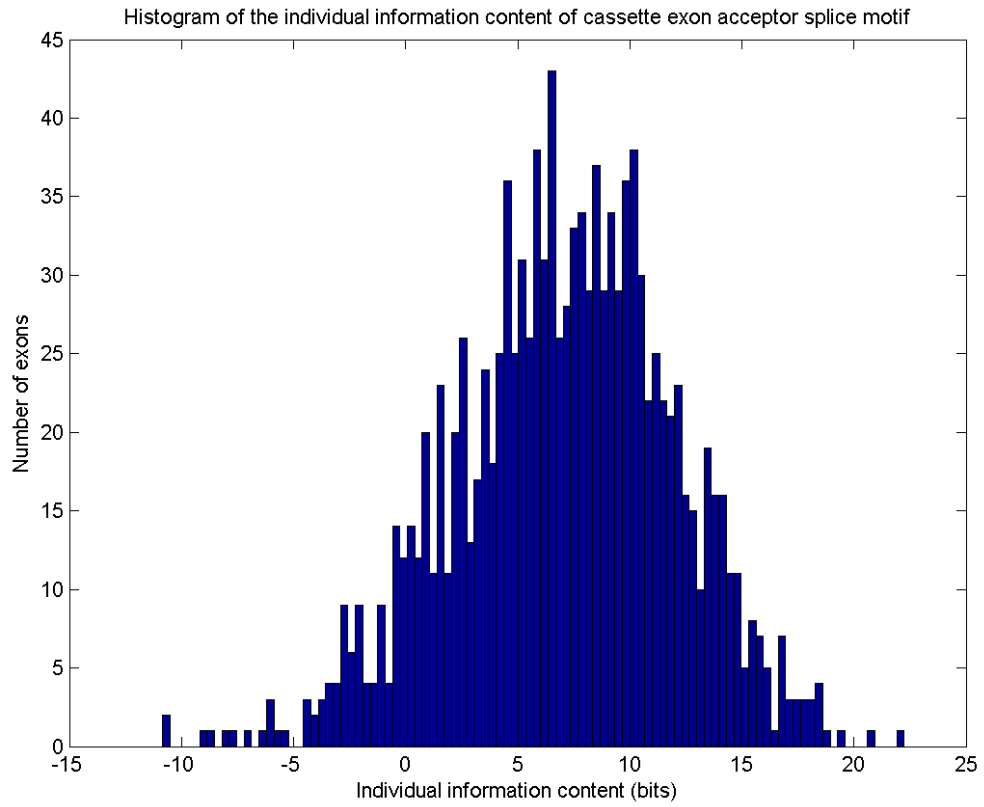


Figure 75. Histogram of the individual information content of cassette exon acceptor splice motif.

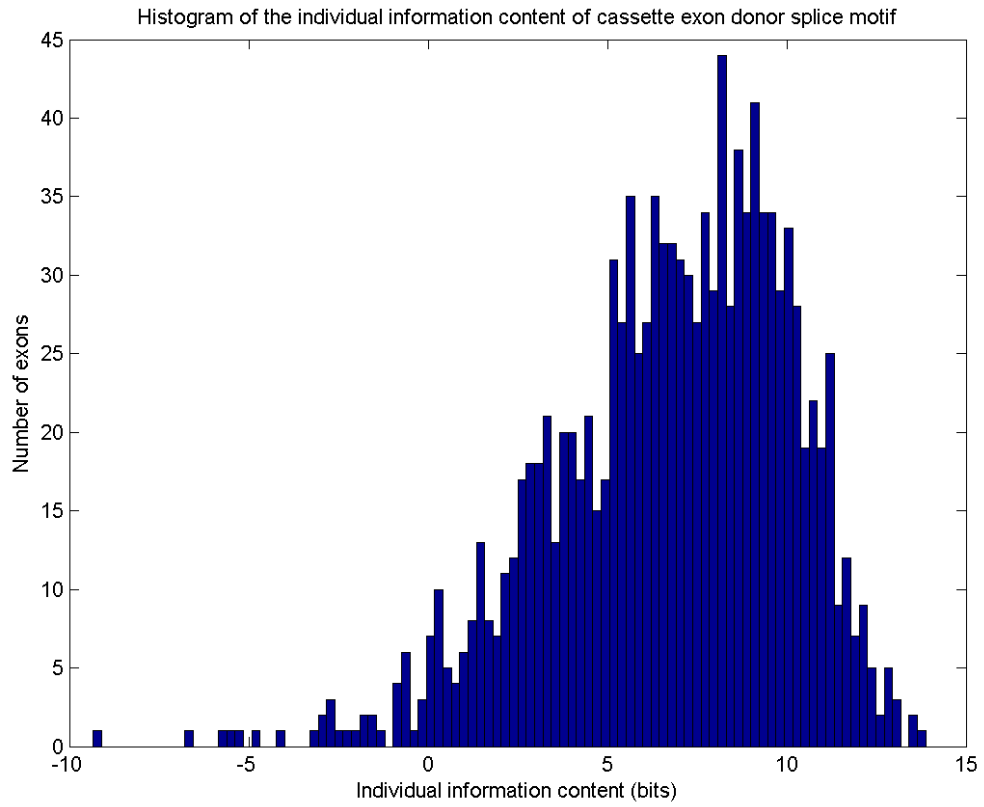


Figure 76. Histogram of the individual information content of cassette exon donor splice motif.

Intron retention introns like cassette exons exhibit low mean IIC values for both acceptor (6.188) and donor (5.389) sites (see Figure 77 and Figure 78 for histograms). While the third quartile IIC values for cassette exon splice site motifs are higher than that of MCIE, the third quartile IIC for intron retention intron splice site motifs are lower. This means that intron retention introns splice sites have far lower IIC than cassette exons splice sites. Therefore, action by ESE and ISE are going to be far more commonplace in intron retention. Like cassette exons, intron retention introns show a larger IQR than MCIE.

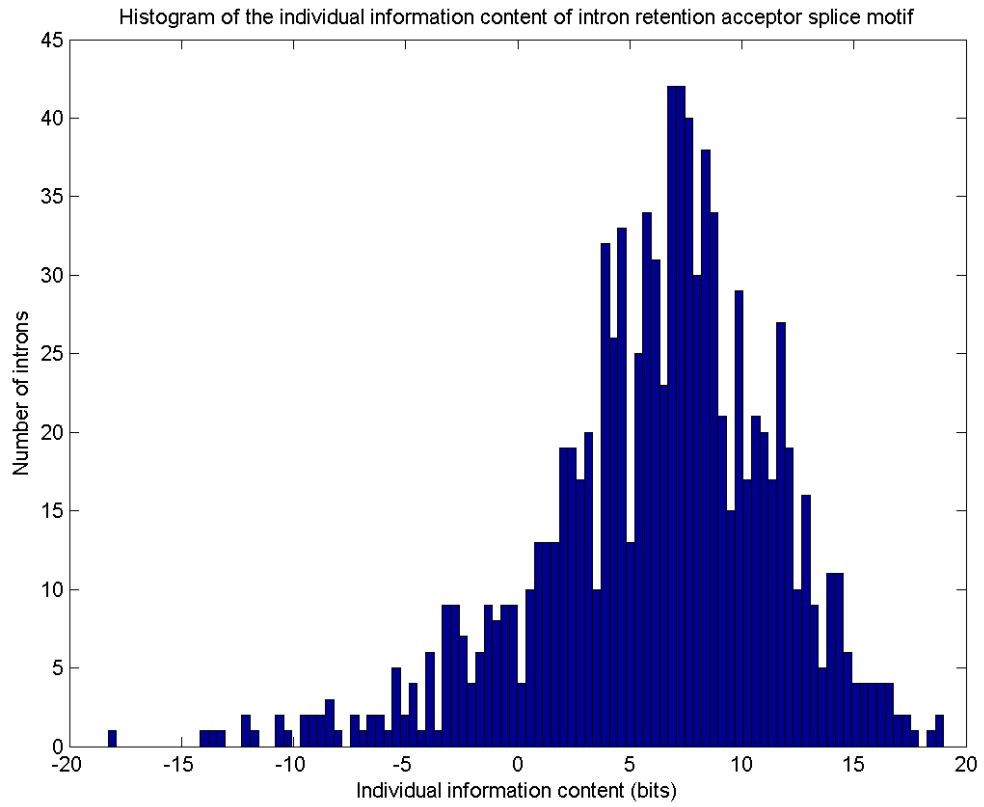


Figure 77. Histogram of the individual information content of intron retention acceptor splice motif.

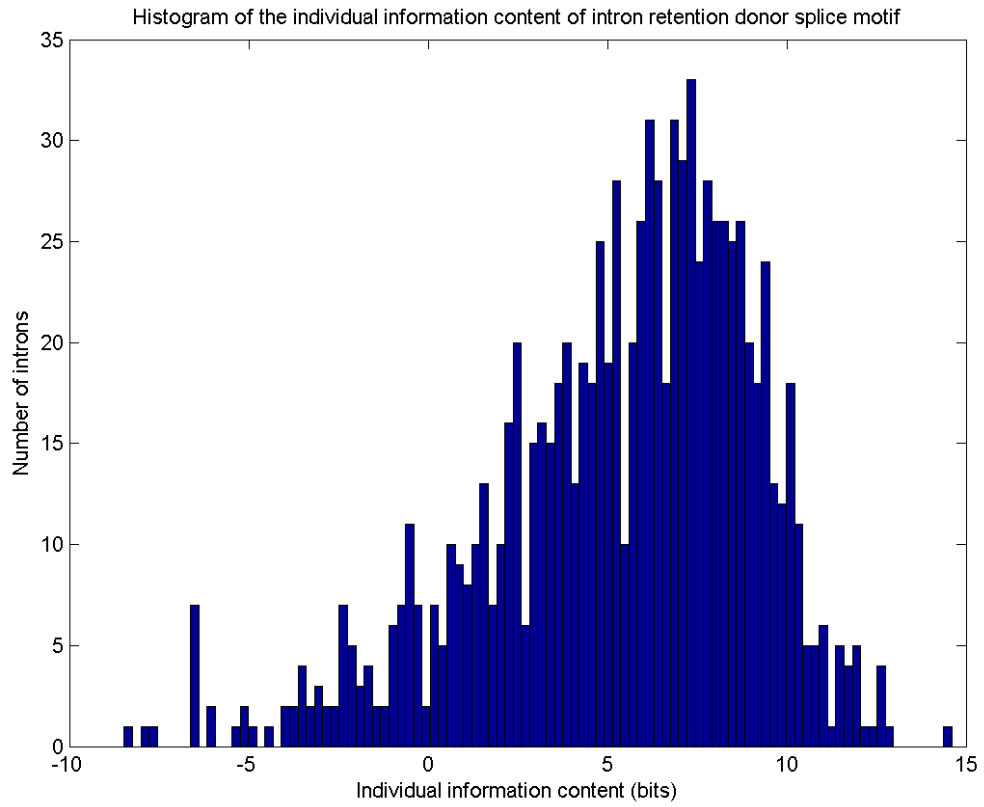


Figure 78. Histogram of the individual information content of intron retention donor splice motif.

Alternative acceptor and donor splice motifs show very low information content (see Figure 79 and Figure 80 for histograms). Both show third quartile IIC below the mean MCIE IIC. Much like intron retention introns, this indicates that action by ESE and ISE in selecting the correct acceptor or donor site is more commonplace. There could be multiple possible donor and acceptor sites with flanking ESE and ISE that allows for differential activation of these splice sites via RNA-protein and protein-protein interaction with the spliceosome.

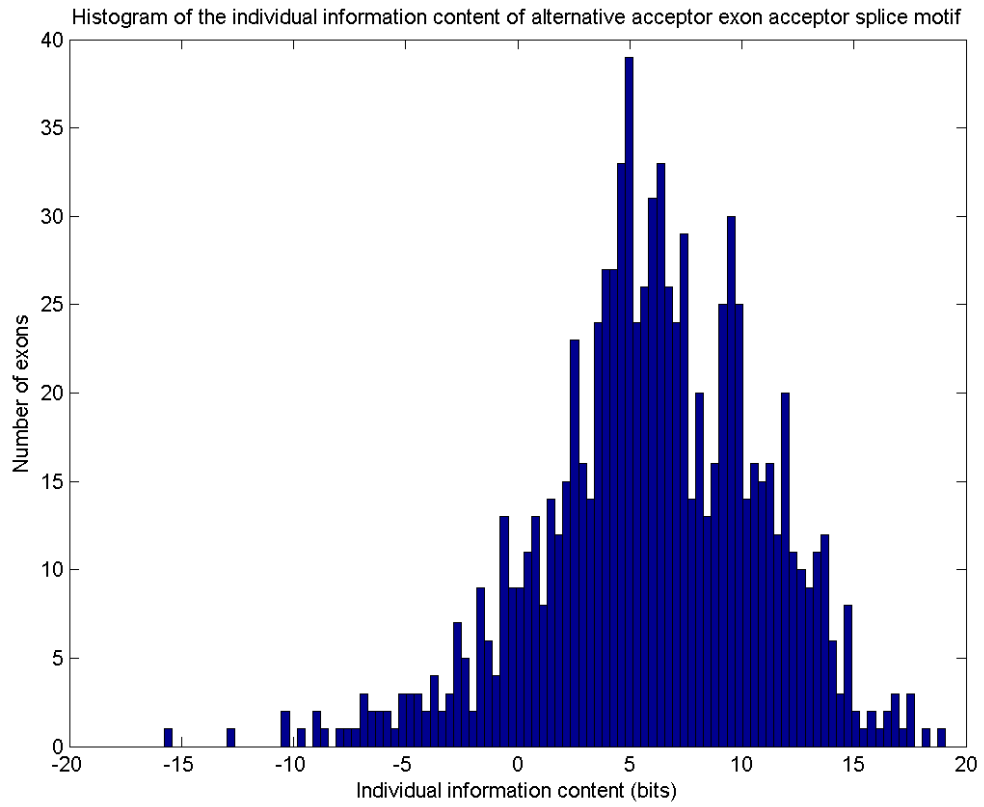


Figure 79. Histogram of the individual information content of alternative acceptor splice motif.

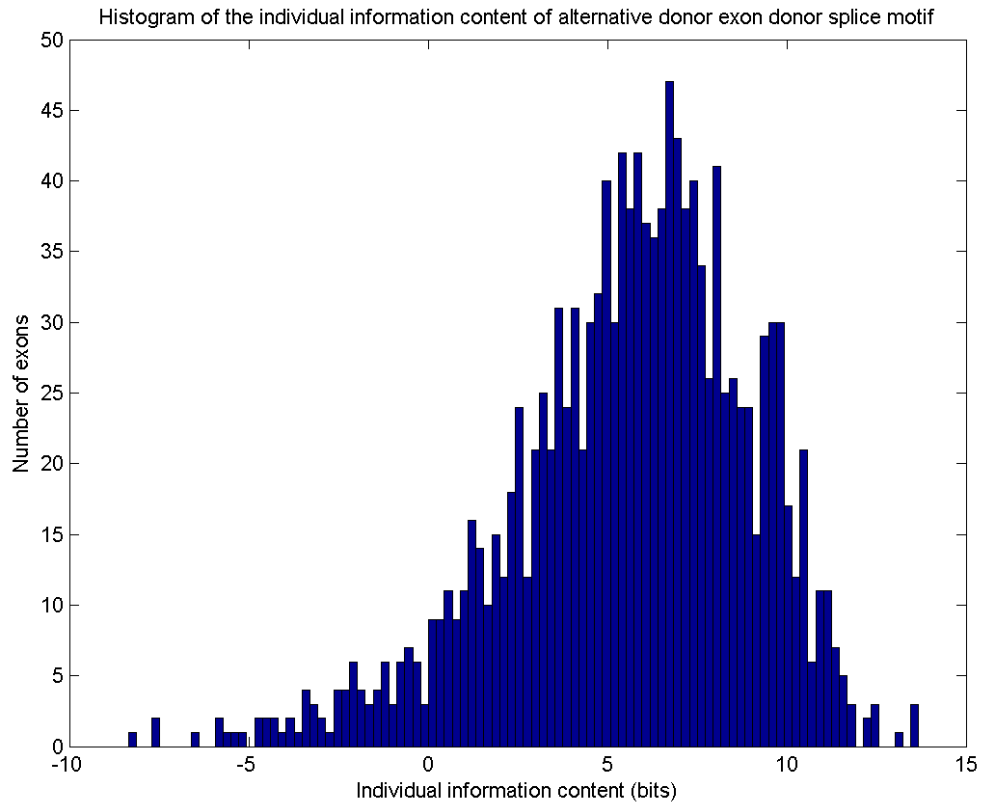


Figure 80. Histogram of the individual information content of intron retention acceptor splice motif.

Multi constitutive initiation exon donor site mean IIC (7.949 bits) is lower than that of MCIE donor site mean IIC (8.517 bits). The histogram of the IIC of multi constitutive initiation exon donor sites as shown in Figure 81 has a similar shape as that of MCIE except for a lower mean IIC. This could reflect the special needs of the initiation exon in its interaction with the spliceosome. This situation is reversed in the case of multi constitutive termination exon acceptor sites, where the mean IIC (9.974 bits) is higher than that of MCIE (9.744 bits). The histogram of the IIC of multi constitutive termination exon acceptor sites shown in Figure 82 is similar to that of MCIE with the sole exception of a higher mean IIC.

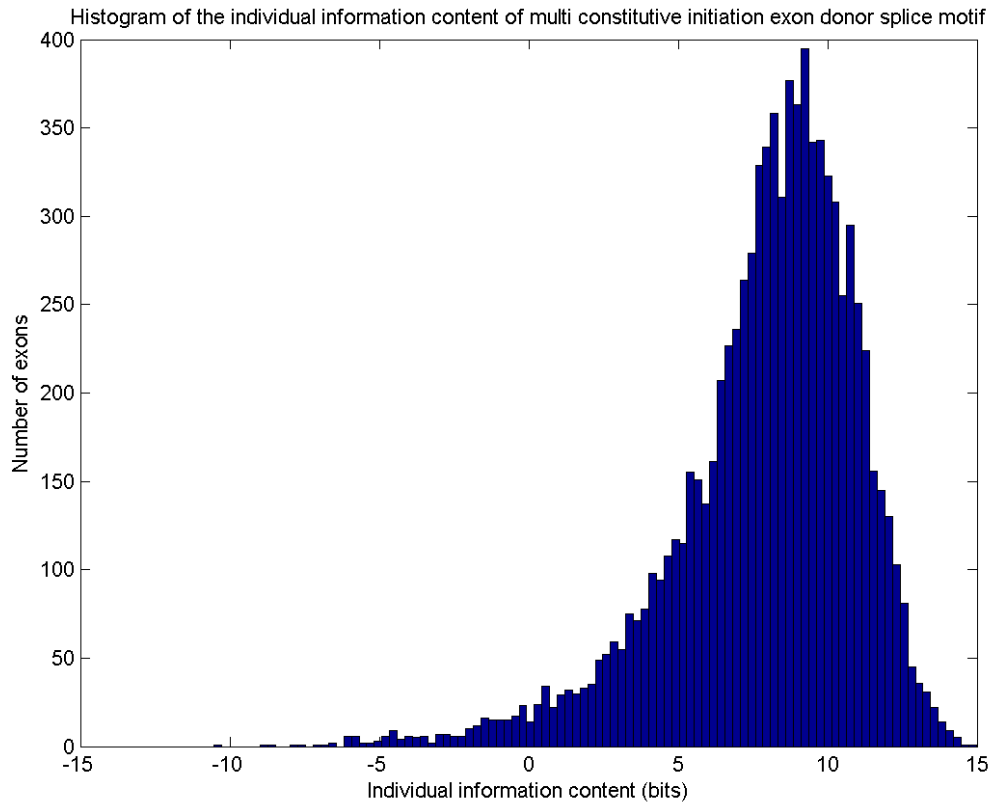


Figure 81. Histogram of the individual information content of multi constitutive initiation exon donor splice motif.

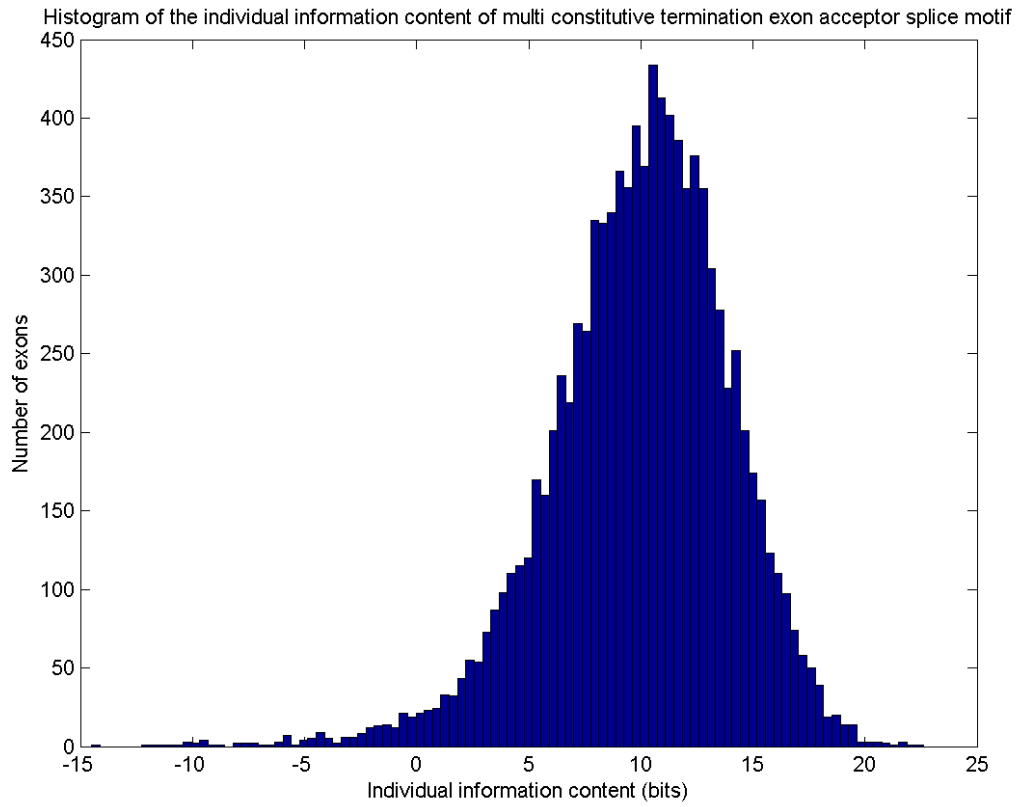


Figure 82. Histogram of the individual information content of multi constitutive termination exon acceptor splice motif.

4.3.6 Domain boundary analysis

The Chi Square Goodness-of-Fit test results are shown in Table 27. A graphical plot of the data shown in Table 27 is shown in Figure 83. Both the table and the graphical plot show that there is a tendency for introns to lie at amino acids positions 0 to +2. This would facilitate exon shuffling by ensuring that the shuffled exon contains a functional domain. This would allow genes to get new functionality through the gain of a new domain.

AA Position	Type	Exp	Obs	Chi	P-value
-10	Within	137.739	136.000	0.022	4.945e-01
-10	Outside	25885.261	25887.000	0.000	
-9	Within	138.284	185.000	15.782	1.793e-04
-9	Outside	25884.716	25838.000	0.084	
-8	Within	138.551	167.000	5.842	2.653e-02
-8	Outside	25884.449	25856.000	0.031	
-7	Within	138.781	158.000	2.662	1.312e-01
-7	Outside	25884.219	25865.000	0.014	
-6	Within	139.068	184.000	14.518	3.385e-04
-6	Outside	25883.932	25839.000	0.078	
-5	Within	139.316	190.000	18.439	4.715e-05
-5	Outside	25883.684	25833.000	0.099	
-4	Within	139.752	199.000	25.119	1.641e-06
-4	Outside	25883.248	25824.000	0.136	
-3	Within	139.754	165.000	4.561	5.050e-02
-3	Outside	25883.246	25858.000	0.025	
-2	Within	139.754	185.000	14.649	3.168e-04
-2	Outside	25883.246	25838.000	0.079	
-1	Within	139.754	150.000	0.751	3.427e-01
-1	Outside	25883.246	25873.000	0.004	
0	Within	45.032	105.000	79.860	2.126e-18
0	Outside	25977.968	25918.000	0.138	
1	Within	128.169	211.000	53.530	1.041e-12
1	Outside	25894.831	25812.000	0.265	
2	Within	123.806	242.000	112.837	1.201e-25
2	Outside	25899.194	25781.000	0.539	
3	Within	115.664	174.000	29.422	1.912e-07
3	Outside	25907.336	25849.000	0.131	

4	Within	110.125	128.000	2.902	1.165e-01
4	Outside	25912.875	25895.000	0.012	
5	Within	107.433	125.000	2.873	1.182e-01
5	Outside	25915.567	25898.000	0.012	
6	Within	105.442	153.000	21.450	1.052e-05
6	Outside	25917.558	25870.000	0.087	
7	Within	103.673	147.000	18.107	5.641e-05
7	Outside	25919.327	25876.000	0.072	
8	Within	101.907	137.000	12.085	1.160e-03
8	Outside	25921.093	25886.000	0.048	
9	Within	100.559	135.000	11.796	1.341e-03
9	Outside	25922.441	25888.000	0.046	
10	Within	99.091	129.000	9.028	5.384e-03
10	Outside	25923.909	25894.000	0.035	

Table 27. Chi Square Goodness-of-Fit test results for domain boundary analysis. The Type column indicates whether the numbers represent introns lying within or outside the amino acid position. The Exp column shows the expected number of introns while the Obs column shows the actual observed number of introns. The Chi column shows the Chi Square value.

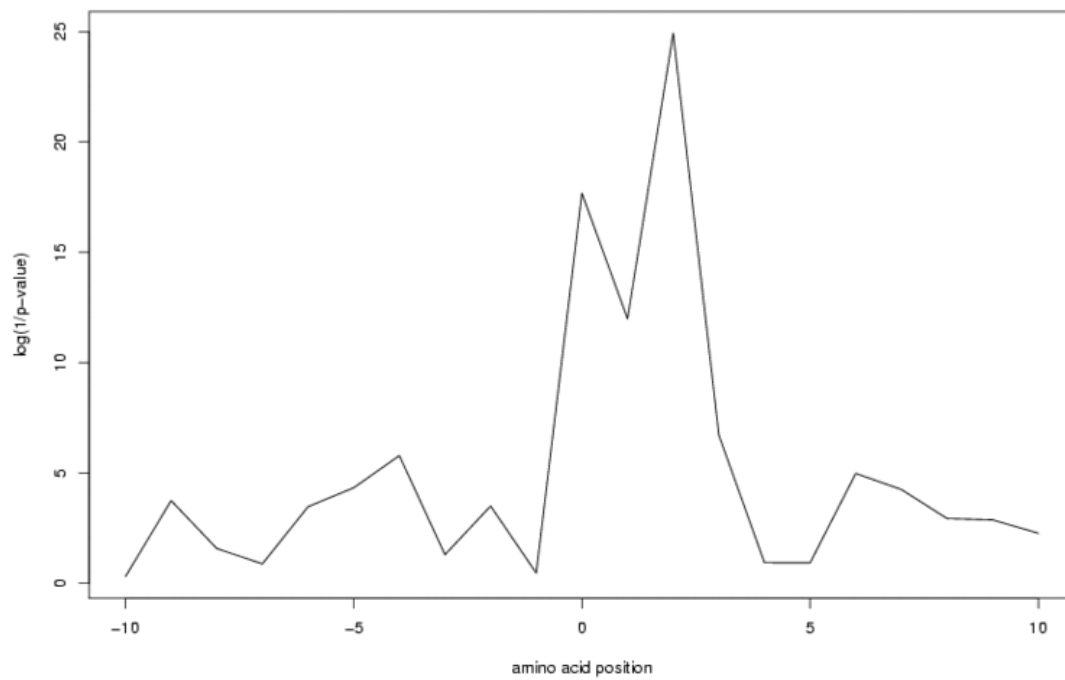


Figure 83. Plot of $\log(1/p\text{-value})$ against amino acid positions. The plot shows that there is a very high statistical significance in the number of introns lying at amino acid positions 0 to + 2.

The results for the Chi Square Goodness-of-Fit test for symmetrical introns are shown in Table 28. A p-value of 1.176×10^{-4} was obtained that showed that the exons have a higher than expected chance of having flanking introns of symmetrical phases. This would greatly facilitate the process of exon shuffling, as the shuffled exon would retain the reading frame. Taken together with the results that there is a greater than expected likelihood of introns lying in the domain boundaries, this shows strong evidence for exon shuffling.

	Obs	Exp	Chi
Symmetrical	119	90.333	9.097
Non-symmetrical	152	180.667	4.549
Total	271	271.000	

Table 28. Results for the Chi Square Goodness-of-Fit test for symmetrical introns. The p-value is 1.176×10^{-4} which indicates that there is a tendency for introns flanking exons to have symmetrical phases.

4.3.7 Number of splicing graphs partitioned using GO terms

GO terms were used to partition the splicing graphs contained in DEDB. The results for the three root terms are shown in Table 29, Table 30 and Table 31.

The molecular function GO term data shown in Table 29 shows that alternative splicing does not preferentially affect proteins involved in catalytic activities (ratio of 1.14), however upon closer inspection, alternative splicing does affect preferentially proteins having kinase (ratio of 0.52), transferase (especially protein kinase activities having a ratio of 0.52) and phosphoprotein phosphatase functions (0.50). These functions are typically involved in cell signaling and thereby imply that alternative splicing may be an important contributor in cell signaling. This is also supported by the fact that the proteins having signal transducer activity have a higher incidence of alternative splicing (ratio of 0.79). Signal transduction involves large numbers of protein-protein interactions and this is evident in the preference of alternative splicing events in proteins that are involved in binding (ratio of 0.76). This is especially the case for protein binding proteins, which has a low ratio of 0.51. Of interest is that proteins involved in cytoskeletal protein binding show even greater likening for alternative splicing events (ratio of 0.32). Although proteins having transporter activity do not have a significant preference for alternative splicing, one of the subcategory consisting of proteins having neurotransmitter transporter activity does have a strong fondness for alternative splicing (ratio being 0.31). This result seems to parallel other studies, which have indicated that proteins involved in neurological functional have a higher incidence of having alternative splicing events (Modrek et al., 2001). Alternative splicing events are also found in transcriptional (ratio of 0.82)

and translational control (0.67) indicating that alternative splicing may be an important contributor to protein diversity and production.

GO Term	All counts	Con. counts	Alt. counts	All per.	Con. Per.	Alt. Per.	Ratio
molecular_function	6488	4785	1703	100.00	100.00	100.00	1.00
motor activity	61	41	20	0.94	0.86	1.17	0.73
catalytic activity	3100	2363	737	47.78	49.38	43.28	1.14
kinase activity	346	205	141	5.33	4.28	8.28	0.52
protein kinase activity	0	0	0	0.00	0.00	0.00	0.00
transferase activity	812	562	250	12.52	11.74	14.68	0.80
protein kinase activity	257	153	104	3.96	3.20	6.11	0.52
hydrolase activity	1440	1126	314	22.19	23.53	18.44	1.28
nuclease activity	63	53	10	0.97	1.11	0.59	1.89
Phosphoprotein Phosphatase activity	84	49	35	1.29	1.02	2.06	0.50
peptidase activity	544	460	84	8.38	9.61	4.93	1.95
signal transducer activity	917	631	286	14.13	13.19	16.79	0.79
receptor activity	505	375	130	7.78	7.84	7.63	1.03
receptor binding	0	0	0	0.00	0.00	0.00	0.00
structural molecule activity	634	452	182	9.77	9.45	10.69	0.88
transporter activity	905	645	260	13.95	13.48	15.27	0.88
ion channel activity	144	95	49	2.22	1.99	2.88	0.69
neurotransmitter transporter activity	15	7	8	0.23	0.15	0.47	0.31
electron transporter activity	138	115	23	2.13	2.40	1.35	1.78
binding	2147	1462	685	33.09	30.55	40.22	0.76
nucleotide binding	85	57	28	1.31	1.19	1.64	0.72
nucleic acid binding	1142	799	343	17.60	16.70	20.14	0.83
DNA binding	498	336	162	7.68	7.02	9.51	0.74
chromatin binding	76	55	21	1.17	1.15	1.23	0.93
transcription factor activity	0	0	0	0.00	0.00	0.00	0.00
RNA binding	312	212	100	4.81	4.43	5.87	0.75
translation factor activity, nucleic acid binding	0	0	0	0.00	0.00	0.00	0.00
receptor binding	239	164	75	3.68	3.43	4.40	0.78
calcium ion binding	88	57	31	1.36	1.19	1.82	0.65
protein binding	528	310	218	8.14	6.48	12.80	0.51
cytoskeletal protein binding	204	97	107	3.14	2.03	6.28	0.32

actin binding	108	45	63	1.66	0.94	3.70	0.25
lipid binding	60	42	18	0.92	0.88	1.06	0.83
oxygen binding	0	0	0	0.00	0.00	0.00	0.00
carbohydrate binding	25	23	2	0.39	0.48	0.12	4.09
antioxidant activity	26	20	6	0.40	0.42	0.35	1.19
chaperone regulator activity	1	0	1	0.02	0.00	0.06	0.00
enzyme regulator activity	322	222	100	4.96	4.64	5.87	0.79
transcription regulator activity	739	516	223	11.39	10.78	13.09	0.82
transcription factor activity	260	178	82	4.01	3.72	4.82	0.77
triplet codon-amino acid adaptor activity	0	0	0	0.00	0.00	0.00	0.00
translation regulator activity	75	49	26	1.16	1.02	1.53	0.67
translation factor activity, nucleic acid binding	72	49	23	1.11	1.02	1.35	0.76
nutrient reservoir activity	5	5	0	0.08	0.10	0.00	0.00

Table 29. Molecular function GO terms and the corresponding number of splicing graphs. All counts provide an absolute number of all splicing graphs that contain the GO term. Con. counts is the number of constitutive splicing graphs that contain the GO term. Alt. counts is the number of alternative splicing graphs that contain the GO term. All per. is the percentage of splicing graph containing this GO term in relation to all splicing graphs having molecular function GO terms. Con. per. is the percentage of constitutive splicing graphs having this GO term in relation to all splicing graphs having molecular function GO terms. Alt. per. is the percentage of alternative splicing graphs having this GO term in relation to all splicing graphs having molecular function GO terms. The ratio of Con. per. over Alt. per. is shown in the ratio column. A value greater than 1 is indicative of more a higher amount of constitutive splicing graph over alternative splicing graph. Rows having ratios higher than one are in yellow and rows having ratios less than one are in blue. The indentation of the GO terms indicates its relative position in the GO hierarchy.

As far as biological processes are concerned (Table 30), most of the alternative splicing events are found preferentially in cell communication (ratio of 0.67), development (ratio of 0.59), death (ratio 0.46) and behavior (ratio of 0.31). All these biological process involve large amounts of signal transduction and alternative splicing appears to be an important player in this. These processes all involve decision making presumably involving the regulation of specific protein isoforms made possible by alternative splicing.

GO Term	All counts	Con. counts	Alt. counts	All per.	Con. per.	Alt. per.	Ratio
biological_process	6045	4407	1638	100.00	100.00	100.00	1.00
cell communication	1341	863	478	22.18	19.58	29.18	0.67
signal transduction	1081	707	374	17.88	16.04	22.83	0.70
cell-cell signaling	417	272	145	6.90	6.17	8.85	0.70
cell recognition	17	10	7	0.28	0.23	0.43	0.53
host-pathogen interaction	0	0	0	0.00	0.00	0.00	0.00
development	1321	813	508	21.85	18.45	31.01	0.59
reproduction	364	217	147	6.02	4.92	8.97	0.55
morphogenesis	826	484	342	13.66	10.98	20.88	0.53
cell growth	0	0	0	0.00	0.00	0.00	0.00
embryonic development	250	144	106	4.14	3.27	6.47	0.50
cell differentiation	230	142	88	3.80	3.22	5.37	0.60
growth	23	14	9	0.38	0.32	0.55	0.58
cell growth	0	0	0	0.00	0.00	0.00	0.00
regulation of gene expression, epigenetic	40	26	14	0.66	0.59	0.85	0.69
physiological process	5586	4073	1513	92.41	92.42	92.37	1.00
response to stress	273	198	75	4.52	4.49	4.58	0.98
cell growth and/or maintenance	2194	1490	704	36.29	33.81	42.98	0.79
transport	1128	771	357	18.66	17.49	21.79	0.80
ion transport	368	244	124	6.09	5.54	7.57	0.73
protein transport	458	328	130	7.58	7.44	7.94	0.94
cell proliferation	739	519	220	12.22	11.78	13.43	0.88
cell cycle	489	346	143	8.09	7.85	8.73	0.90
cell organization and biogenesis	651	405	246	10.77	9.19	15.02	0.61
cytoplasm organization and biogenesis	498	312	186	8.24	7.08	11.36	0.62
organelle organization and biogenesis	476	293	183	7.87	6.65	11.17	0.60
mitochondrion organization and biogenesis	6	4	2	0.10	0.09	0.12	0.74
cytoskeleton organization and biogenesis	414	249	165	6.85	5.65	10.07	0.56
cell growth	24	12	12	0.40	0.27	0.73	0.37
cell homeostasis	0	0	0	0.00	0.00	0.00	0.00

metabolism	4029	2954	1075	66.65	67.03	65.63	1.02
carbohydrate metabolism	349	264	85	5.77	5.99	5.19	1.15
energy pathways	95	65	30	1.57	1.47	1.83	0.81
electron transport	96	69	27	1.59	1.57	1.65	0.95
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	1422	1012	410	23.52	22.96	25.03	0.92
DNA metabolism	282	209	73	4.67	4.74	4.46	1.06
transcription	735	513	222	12.16	11.64	13.55	0.86
amino acid and derivative metabolism	225	160	65	3.72	3.63	3.97	0.91
lipid metabolism	441	332	109	7.30	7.53	6.65	1.13
coenzyme and prosthetic group metabolism	130	107	23	2.15	2.43	1.40	1.73
catabolism	675	528	147	11.17	11.98	8.97	1.34
biosynthesis	593	413	180	9.81	9.37	10.99	0.85
protein biosynthesis	0	0	0	0.00	0.00	0.00	0.00
protein metabolism	1692	1239	453	27.99	28.11	27.66	1.02
protein biosynthesis	380	269	111	6.29	6.10	6.78	0.90
protein modification	591	412	179	9.78	9.35	10.93	0.86
secondary metabolism	0	0	0	0.00	0.00	0.00	0.00
cell death	0	0	0	0.00	0.00	0.00	0.00
response to external stimulus	753	582	171	12.46	13.21	10.44	1.27
response to abiotic stimulus	241	174	67	3.99	3.95	4.09	0.97
response to biotic stimulus	487	377	110	8.06	8.55	6.72	1.27
response to endogenous stimulus	118	86	32	1.95	1.95	1.95	1.00
death	251	139	112	4.15	3.15	6.84	0.46
cell death	220	121	99	3.64	2.75	6.04	0.45
cell homeostasis	58	39	19	0.96	0.88	1.16	0.76
behavior	130	59	71	2.15	1.34	4.33	0.31
viral life cycle	1	1	0	0.02	0.02	0.00	0.00

Table 30. Biological process GO terms and the corresponding number of splicing graphs. All counts provide an absolute number of all splicing graphs that contain the GO term. Con. counts is the number of constitutive splicing graphs that contain the GO term. Alt. counts is the number of alternative splicing graphs that contain the GO term. All per. is the percentage of splicing graph containing this GO term in relation to all splicing graphs having biological process GO terms. Con. per. is the percentage of constitutive splicing

graphs having this GO term in relation to all splicing graphs having biological process GO terms. Alt. per. is the percentage of alternative splicing graphs having this GO term in relation to all splicing graphs having biological process GO terms. The ratio of Con. per. over Alt. per. is shown in the ratio column. A value greater than 1 is indicative of more a higher amount of constitutive splicing graph over alternative splicing graph. Rows having ratios higher than one are in yellow and rows having ratios less than one are in blue. The indentation of the GO terms indicates its relative position in the GO hierarchy.

Alternative splicing does not seem to have an inclination for any specific cellular location (Table 31). The only significant location is the plasma membrane with a ratio of 0.53. This sort of ties in with the results from the molecular function and biological processes, which shows alternative splicing being found in proteins involved in cell signaling. The plasma membrane is a location where large numbers of cell signaling receptors reside.

GO Term	All counts	Con. counts	Alt. counts	All per.	Con. per.	Alt. per.	Ratio
cellular_component	3090	2175	915	100.00	100.00	100.00	1.00
extracellular	227	172	55	7.35	7.91	6.01	1.32
extracellular matrix	46	24	22	1.49	1.10	2.40	0.46
extracellular space	11	8	3	0.36	0.37	0.33	1.12
cell	2924	2037	887	94.63	93.66	96.94	0.97
intracellular	2216	1551	665	71.72	71.31	72.68	0.98
nucleus	1032	740	292	33.40	34.02	31.91	1.07
nuclear chromosome	0	0	0	0.00	0.00	0.00	0.00
nuclear membrane	0	0	0	0.00	0.00	0.00	0.00
nucleoplasm	176	137	39	5.70	6.30	4.26	1.48
nucleolus	29	27	2	0.94	1.24	0.22	5.68
chromosome	112	74	38	3.62	3.40	4.15	0.82
nuclear chromosome	25	18	7	0.81	0.83	0.77	1.08
cytoplasmic chromosome	0	0	0	0.00	0.00	0.00	0.00
cytoplasm	1291	884	407	41.78	40.64	44.48	0.91
cytoplasmic chromosome	0	0	0	0.00	0.00	0.00	0.00
mitochondrion	403	305	98	13.04	14.02	10.71	1.31
endosome	8	5	3	0.26	0.23	0.33	0.70
vacuole	42	29	13	1.36	1.33	1.42	0.94
lysosome	14	9	5	0.45	0.41	0.55	0.76
peroxisome	27	25	2	0.87	1.15	0.22	5.26
endoplasmic reticulum	122	93	29	3.95	4.28	3.17	1.35
Golgi apparatus	46	30	16	1.49	1.38	1.75	0.79
lipid particle	1	1	0	0.03	0.05	0.00	0.00
microtubule organizing center	32	21	11	1.04	0.97	1.20	0.80
cytosol	204	128	76	6.60	5.89	8.31	0.71
ribosome	0	0	0	0.00	0.00	0.00	0.00
cytoskeleton	200	136	64	6.47	6.25	6.99	0.89
plastid	0	0	0	0.00	0.00	0.00	0.00
cytoplasmic vesicle	72	39	33	2.33	1.79	3.61	0.50
ribosome	159	116	43	5.15	5.33	4.70	1.13
cilium	0	0	0	0.00	0.00	0.00	0.00
thylakoid	0	0	0	0.00	0.00	0.00	0.00

nuclear membrane	38	29	9	1.23	1.33	0.98	1.36
plasma membrane	452	252	200	14.63	11.59	21.86	0.53
cilium	0	0	0	0.00	0.00	0.00	0.00
external encapsulating structure	0	0	0	0.00	0.00	0.00	0.00
cell wall	0	0	0	0.00	0.00	0.00	0.00
cell envelope	0	0	0	0.00	0.00	0.00	0.00

Table 31. Cellular process GO terms and the corresponding number of splicing graphs. All counts provide an absolute number of all splicing graphs that contain the GO term. Con. counts is the number of constitutive splicing graphs that contain the GO term. Alt. counts is the number of alternative splicing graphs that contain the GO term. All per. is the percentage of splicing graph containing this GO term in relation to all splicing graphs having cellular process GO terms. Con. per. is the percentage of constitutive splicing graphs having this GO term in relation to all splicing graphs having cellular process GO terms. Alt. per. is the percentage of alternative splicing graphs having this GO term in relation to all splicing graphs having cellular process GO terms. The ratio of Con. per. over Alt. per. is shown in the ratio column. A value greater than 1 is indicative of more a higher amount of constitutive splicing graph over alternative splicing graph. Rows having ratios higher than one are in yellow and rows having ratios less than one are in blue. The indentation of the GO terms indicates its relative position in the GO hierarchy.

4.3.8 Effects of alternative splicing on the coding sequence

To determine the significance of the effects of the various types of alternative splicing events on the coding sequence, each type of alternative splicing events was tested using the Chi Square Goodness-of-Fit test to determine if any of the types of alternative splicing events are significantly enriched in the coding sequence. Table 32 shows that in general alternative splicing events tend not to affect the coding sequence. This observation holds true for all the various forms of alternative splicing events (Table 33, Table 34, Table 35, Table 36 and Table 38) except for cassette exons (Table 37) where there is a higher incidence rate in the coding sequence. This appears to suggest that alternative splicing events are selected to be outside of the coding sequence. This means that they are unlikely to lead to non-functional proteins. It would seem that cassette exons on the other hand are favored.

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	3885	1930.006	1980.305
Partial overlap	6281	8224.419	459.227
Complete overlap	1	12.575	10.654
Total	10167	10167	2450

Table 32. Chi Square Goodness-of-Fit test for all forms of alternative splicing. The P-value is 0.

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	149	112.225	12.050
Partial overlap	294	330.775	4.088
Complete overlap	0	0.000	0.000
Total	443	443	16

Table 33. Chi Square Goodness-of-Fit test for alternative acceptor sites. The P-value is 5.015×10^{-4} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	363	153.473	286.054
Partial overlap	260	469.075	93.189
Complete overlap	0	0.452	0.452
Total	623	622	379

Table 34. Chi Square Goodness-of-Fit test for alternative donor sites. The P-value is 2.761×10^{-82} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	2521	1046.793	2076.140
Partial overlap	2692	4162.196	519.312
Complete overlap	0	4.011	4.011
Total	5213	5213	2599

Table 35. Chi Square Goodness-of-Fit test for alternative initiation exons. The P-value is 0.

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	122	112.883	0.736

Partial overlap	1285	1287.865	0.006
Complete overlap	1	7.252	5.390
Total	1408	1407	6

Table 36. Chi Square Goodness-of-Fit test for alternative termination exons. The P-value is 4.603×10^{-2} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	155	232.446	25.804
Partial overlap	1073	995.472	6.038
Complete overlap	0	0.081	0.081
Total	1228	1228	31

Table 37. Chi Square Goodness-of-Fit test for cassette exon. The P-value is 2.637×10^{-7} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	406	183.821	268.541
Partial overlap	519	740.400	66.205
Complete overlap	0	0.779	0.779
Total	925	924	335

Table 38. Chi Square Goodness-of-Fit test for intron retentions. The P-value is 1.013×10^{-72} .

4.3.9 Effects of alternative splicing on the domains

The phenomenon seen in the effects of alternative splicing on the coding sequence is echoed in the effects on the protein domains (Table 39, Table 40, Table 41, Table 42, Table 43, Table 44 and Table 45) with the sole exception that even cassette exons are found preferentially outside protein domains. This means that most of the alternative splicing will modulate rather than abolish protein function. Therefore, in general, alternative splicing is selected against in protein domains. This means that any alternative splicing occurring in protein domains have greater significance, probably being part of its regulation.

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	7330	6768.562	46.570
Partial overlap	806	1168.956	112.696
Complete overlap	412	610.482	64.531
Total	8548	8547	223

Table 39. Chi Square Goodness-of-Fit test for all forms of alternative splicing. The P-value is 1.509×10^{-48} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	732	700.041	1.459
Partial overlap	65	90.025	6.957
Complete overlap	0	6.934	6.934
Total	797	797	15

Table 40. Chi Square Goodness-of-Fit test for alternative acceptor sites. The P-value is 7.260×10^{-4} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	1032	989.731	1.805
Partial overlap	38	68.126	13.322
Complete overlap	8	20.142	7.320
Total	1078	1077	22

Table 41. Chi Square Goodness-of-Fit test for alternative donor sites. The P-value is 2.524×10^{-5} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	2406	2198.091	19.665
Partial overlap	210	352.856	57.836
Complete overlap	148	213.053	19.863
Total	2764	2764	97

Table 42. Chi Square Goodness-of-Fit test for alternative initiation exons. The P-value is 2.836×10^{-21} .

Type of overlap	Observed counts	Expected counts	Chi Square value
-----------------	-----------------	-----------------	------------------

No overlap	644	538.376	20.722
Partial overlap	158	196.391	7.505
Complete overlap	187	254.232	17.780
Total	989	989	46

Table 43. Chi Square Goodness-of-Fit test for alternative termination exons. The P-value is 2.767×10^{-10} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	1302	1194.297	9.713
Partial overlap	218	297.970	21.463
Complete overlap	68	95.733	8.034
Total	1588	1587	39

Table 44. Chi Square Goodness-of-Fit test for cassette exons. The P-value is 7.646×10^{-9} .

Type of overlap	Observed counts	Expected counts	Chi Square value
No overlap	974	914.010	3.937
Partial overlap	96	140.379	14.030
Complete overlap	1	16.611	14.672
Total	1071	1070	32

Table 45. Chi Square Goodness-of-Fit test for intron retention. The P-value is 1.683×10^{-7} .

4.4 Conclusion

An analysis of alternative splicing on the genomic level in *Drosophila melanogaster* has revealed several interesting findings. Firstly, the amount of alternative splicing in this analysis was in the region of 20 percent. This is lower than that predicted for higher eukaryotes like human but still it means that a significant portion of the genes in *Drosophila melanogaster* is alternatively spliced meaning that a fair portion of the proteome is a result of alternative splicing. The nature of the data in that it is manually curated is likely to make this estimate a conservative one. This is supported by the fact that well-known genes having numerous alternative spliced isoforms like Dscam are only represented in this dataset having far fewer isoforms. This limitation can be eased by computational methods that detect for alternative splicing which will result in more splicing isoform information.

It is interesting to note that alternatively spliced genes exhibit characteristics, which are different from constitutively spliced genes. One of these characteristics is the length of the exons within the genes. Alternative splicing events like cassette exons, intron retentions, alternative donor sites and alternative acceptor sites involve exons or introns with different lengths as compared to multi constitutive internal exons. The number of exons is also generally higher in alternatively spliced genes. The splice sites themselves also show differing phenotypes in constitutive and alternative spliced genes, with alternatively spliced genes exhibiting less consensus (lower information content) in the splice site motifs. This is good news for the design of predictive methods to detect for alternative splicing as these characteristics can be utilized in the design

of such methods. It also indicates that the cell contains machinery that through these differing signals identifies and effect alternative splicing.

In addition to detecting for characteristics that differ between constitutively and alternatively spliced genes, analysis was also carried out on the effects of alternative splicing. The gene ontology (GO) breakdown of alternatively spliced genes indicates that genes involved in certain molecular function and biological processes have a higher rate of being alternatively spliced. A large number of genes, which are alternatively spliced, are involved in signal transduction. This is logical since signal transduction involves the need for differential control that can be afforded by alternative splicing. However, the effects of alternative splicing seem to be targeted on the portions of the gene, which do not affect its protein function. This finding comes as a surprise as a previous study (Kriventseva et al., 2003) indicates that alternative splicing prefers affecting the function of the protein. The most probably reason for this difference is the nature of the data analyzed. The study by Kriventseva *et al.* uses data contained in Swiss-Prot that includes proteins from several organisms. The data in Swiss-Prot are also likely to include splice variant information that is of significance to biologist (most likely meaning that the function of the protein is affected). In contrast, the data used for this analysis is restricted to only fruitfly and the data is more likely to be impartial with regards to the type of alternative splicing annotated. However, the data analyzed suffers from incompleteness as mentioned earlier and thus the true effects of alternative splicing on the protein maybe in between this analysis and the previous study.

This analysis also shows that the exon and intron definition model for splicing are likely to co-exist in *Drosophila melanogaster*. This concurs with a

previous study that indicate the same conclusion (Talerico & Berget, 1994). There is also a tendency for introns to flank the domain boundaries that could have implications in exon shuffling.

Overall conclusion

The first chapter has illustrated the nature of alternative splicing and its implications. No longer can we attribute the phenomenon of alternative splicing as mistakes in the splicing of transcripts. Alternative splicing is a highly regulated one that involves multiple protein and nucleic acid elements in a tightly orchestrated process. It represents not only an additional layer of control in the highly complex process of gene regulation but a means that enables the cell to produce far larger numbers of proteins that could contribute to the complexity of the organism.

In addition to details of the splicing machinery and its implications, the first chapter also describes current bioinformatic tools and databases in use in the analysis of alternative splicing. The trend observed here indicates that there will continue to be increases in the amount of sequence information available for use in understanding of alternative splicing. This increase comes about as a result of maturation in sequencing technology that has enabled the complete sequencing of entire genomes. This is indeed good news for the understanding of alternative splicing as it indicates that genome level analysis of alternative splicing is now within reach.

In view of these observations, this thesis has set forth to understand the nature of alternative splicing on a genomic level. The first step to this understanding involves the transformation of the information that is currently available into a form that is suitable for analysis. This has been tackled in Chapter 2 where the splicing graph data representation model was taken and extended for our purpose. We have modified the splicing graph creation model to

one which is simpler and more suitable for our purpose taking into consideration the nature of the starting material (a series of positions denoting the exon boundaries). This allows for the merger of several splice isoforms into a single splicing graph allowing for manipulation of the various splice isoforms of the gene as a single entity. This is in contrast to more conventional means of manipulation, which involves handling the various splice isoforms as individual entities. The splicing graph representation thus allows for easier and quicker computation and the conventional means. Not only is it easier to handle, the splicing graph form lends itself readily to the creation of classification rules (discussed in Chapter 2 and elaborated in Chapter 3). These rules make it possible to detect for individual alternative splicing events allowing analysis to be carried on each form of alternative splicing events. This gives more granularity to the analysis of alternative splicing (discussed in Chapter 4). For the benefit of the research community at large, the codes developed were used for the creation of a web service (SGM) that allows users to create their own splicing graphs. In addition, the codes developed are placed on the web site allowing other researchers to use them for their own purpose.

Using the splicing graph data model extended in Chapter 2, Chapter 3 deals with the creation of an accurate and clean dataset housed in the database DEDB for use in the analysis of alternative splicing. The organism selected for the analysis is *Drosophila melanogaster* that not only has its complete genome sequenced but which a comprehensive set of manual annotation is available. This allows for highly accurate and clean information that is subsequently analyzed. *Drosophila melanogaster* is also a eukaryote where alternative splicing has been demonstrated conclusively for a number of genes such as DSCAM and

as data in Chapter 4 indicates, a fair portion of the genes are alternatively spliced. This provides for support in that alternative splicing does contribute to the proteome of the organism. Furthermore, most work on alternative splicing on a genomic level has been targeted at higher eukaryotes such as mouse and human, therefore the work on a lower eukaryote like *Drosophila melanogaster* will shed some light on the extent that the characteristics of alternative splicing is conserved among the various organisms. In addition to merging all the annotation available for *Drosophila melanogaster* into splicing graphs, DEDB provides additional information by mapping Pfam domains and coding sequences onto the splicing graph as well as making splicing specific information like intron phase and splice site motifs. DEDB goes one further step by making all these information accessible via an intuitive web interface that should be familiar to most biologists. The web interface also demonstrate another strength of splicing graphs, in that a visual representation of it makes it easy to understand the nature and types of alternative splicing available. For computational inclined researchers, DEDB makes available its data in an XML format that captures all the information available via the web interface. This allows researchers to carry out analysis or other work on a set of genes that has its alternative splicing events delineated and classified.

The availability of the data in DEDB serves as the starting material for the analysis of alternative splicing in Chapter 4. The analysis conducted is focused on characterizing the nature of alternative splicing and its effects. Typical analysis such as the frequency of the various types of alternative splicing, the length of the exons and introns involved in the various types of alternative splicing and the frequency of the splicing motifs were done. A fair number of both intron retention

and cassette exon events seems to indicate that both intron and exon definition occur in *Drosophila melanogaster* in agreeable with other studies. This seems to be in contrast to higher eukaryotes where cassette exons are more dominant in part due to the shortness of the exon in comparison to the introns. *Drosophila melanogaster* on the other hand contains a fair number of short introns, in fact the introns in *Drosophila melanogaster* appears to follow a narrow distribution. Analysis of the lengths of exons and introns also reveals that the exons and introns involved in the various types of alternative splicing exhibit distinct characteristics that could be exploited in methods that seek to predict alternative splicing events. The frequencies observed are pretty much typical with the canonical splice site motif (GT..AG) being the most common form. Information content analysis of the various splice site motifs in the various alternative splicing events produced a general consensus in that the splice site motifs involved in alternative splicing are less conserved than that of constitutive exons. Again, this is in agreement with other studies that indicate the alternative splicing splice site motifs are generally weaker. This again is good news for alternative splicing detection methods.

As for the effects of alternative splicing, analysis using GO terms indicate that most of the gene having alternative splicing are involved in signal transduction. This makes sense, as alternative splicing will then serve as another additional layer of control providing more control in gene regulation. A recent study (Modrek et al., 2001) indicates that human genes involved in nervous systems are more likely to be alternatively spliced. The results here seem to parallel this study and demonstrate that lower eukaryotes also exhibit preferences for alternative splicing in certain categories of genes.

What was surprising in the analysis was that alternative splicing events appear to be located outside of coding sequences and protein domains. This is in contrast to one study (Kriventseva et al., 2003) done on Swiss-Prot data. As discussed in Chapter 4, this is likely to be due to the nature of the dataset used in that study and our dataset.

The work done in this thesis appears to indicate that lower eukaryotes such as *Drosophila melanogaster* do show evidence of alternative splicing events and that a significant portion of the genes is alternatively spliced. Furthermore, exons and introns involved in alternative splicing exhibit certain distinct characteristics. Alternative splicing also appears to target specific categories of genes. All this evidence points to the fact that alternative splicing is a highly regulated process that contains its own distinct signals and interactors and that this process serves a physiology role in certain categories of genes.

Future directions

The work done in Chapter 2 paves the way for further development that will allow for the assembly of ESTs to splicing graphs. Although the current methodology is capable of EST assembly, there are certain issues involved in EST assembly that are different from merging high quality annotation. Firstly, any assembly method has to account for the inaccuracy of EST data; therefore, the existing method has to be extended to allow for a measure of the reliability of any exon in the splicing graph. Users can then remove any exon that does not fulfill certain criteria. Another issue is the determination of the coding sequence, which has to be computed for each possible variant and checked. The orientation of the gene has also to be determined. These problems can be alleviated by the use of annotated sequences within the assembly to serve as a foundation for the determination of both the coding sequence and the orientation of the gene.

The current dataset contained in DEDB are derived from high quality genome annotation data. The main issue with this data is its completeness as far as the coverage of alternative splicing is concerned, which has been discussed in Chapter 4. To overcome this limitation, the data in DEDB could be complemented by incorporating EST data. This will involve the work mentioned in the preceding paragraph. The result of this will be a dataset, which will be a more accurate measure of amount of alternative splicing in *Drosophila melanogaster*.

The analysis covered in Chapter 4 has revealed a series of features that could potentially be used in some form of machine learning method to detect for alternative splicing events. The development of such methods will prove useful as the not all genomes are sequenced currently and some organisms are lacking in

transcript sequence information. This will also provide insights into whether the characteristics discovered through the analysis are sufficient for the recognition of alternative spliced exons and introns.

With the advent of more compute power and sequencing projects, it would be possible to perform the analysis carried out in Chapter 4 to other organisms. In fact, all of the referred works mentioned in Chapter 4 were done on organism other than *Drosophila melanogaster*.

Incorporation of polymorphism and 3D structure information into the analysis of the effects of alternative splicing are also a distinct possibility. By mapping polymorphism information onto the splicing graphs, information on the effects of natural polymorphism and mutations on alternative splicing can be visualized and quantified. 3D structure information once mapped on the splicing graph could also prove interesting and valuable in understand the effects of alternative splicing on the 3D conformation of the protein. Preliminary work in this area has begun with the design of automated methods for solving 3D structures of protein via homology modeling (Kong, Lee, Tong, Tan, & Ranganathan, 2004). The solving of 3D structure by such automated methods is likely to be required in view of the fact that PDB, the major protein structure database currently does not provide sufficient coverage of the proteome to be statistically significant.

Bibliography

- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461), 2185-2195.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3), 403-410.
- Bairoch, A., Boeckmann, B., Ferro, S., & Gasteiger, E. (2004). Swiss-Prot: juggling between evolution and stability. *Brief Bioinform*, 5(1), 39-55.
- Bajorath, J. (2000). Molecular organization, structural features, and ligand binding characteristics of CD44, a highly variable cell surface glycoprotein with multiple functions. *Proteins*, 39(2), 103-111.
- Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res*, 32 Database issue, D138-141.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2004). GenBank: update. *Nucleic Acids Res*, 32 Database issue, D23-26.
- Berget, S. M. (1995). Exon recognition in vertebrate splicing. *J Biol Chem*, 270(6), 2411-2414.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28(1), 235-242.
- Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, 72, 291-336.
- Borland, G., Ross, J. A., & Guy, K. (1998). Forms and functions of CD44. *Immunology*, 93(2), 139-148.
- Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., et al. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett*, 474(1), 83-86.
- Cech, T. R. (1990). Nobel lecture. Self-splicing and enzymatic activity of an intervening sequence RNA from *Tetrahymena*. *Biosci Rep*, 10(3), 239-261.
- Celotto, A. M., & Graveley, B. R. (2001). Alternative splicing of the *Drosophila* Dscam pre-mRNA is both temporally and spatially regulated. *Genetics*, 159(2), 599-608.
- 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(4), 451-464.
- Corpet, F., Gouzy, J., & Kahn, D. (1998). The ProDom database of protein domain families. *Nucleic Acids Res*, 26(1), 323-326.
- Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., & Mattick, J. S. (2000). ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet*, 24(4), 340-341.
- De Moerlooze, L., Spencer-Dene, B., Revest, J., Hajihosseini, M., Rosewell, I., & Dickson, C. (2000). An important role for the IIIb isoform of fibroblast growth factor receptor 2 (FGFR2) in mesenchymal-epithelial signalling during mouse organogenesis. *Development*, 127(3), 483-492.
- Dralyuk, I., Brudno, M., Gelfand, M. S., Zorn, M., & Dubchak, I. (2000). ASDB: database of alternatively spliced genes. *Nucleic Acids Res*, 28(1), 296-297.

- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), 755-763.
- Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*, 8(9), 967-974.
- Goldstrohm, A. C., Greenleaf, A. L., & Garcia-Blanco, M. A. (2001). Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing. *Gene*, 277(1-2), 31-47.
- Gopalan, V., Tan, T. W., Lee, B. T., & Ranganathan, S. (2004). Xpro: database of eukaryotic protein-encoding genes. *Nucleic Acids Res*, 32(Database issue), D59-63.
- Graveley, B. R. (2005). Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell*, 123(1), 65-73.
- Green, M. R., Maniatis, T., & Melton, D. A. (1983). Human beta-globin pre-mRNA synthesized in vitro is accurately spliced in *Xenopus* oocyte nuclei. *Cell*, 32(3), 681-694.
- Guo, M., Lo, P. C., & Mount, S. M. (1993). Species-specific signals for the splicing of a short *Drosophila* intron in vitro. *Mol Cell Biol*, 13(2), 1104-1118.
- Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., et al. (1999). Alternative splicing of human genes: more the rule than the exception? *Trends Genet*, 15(10), 389-390.
- Hawkins, J. D. (1988). A survey on intron and exon lengths. *Nucleic Acids Res*, 16(21), 9893-9908.
- 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.
- Hoffman, B. E., & Grabowski, P. J. (1992). U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon. *Genes Dev*, 6(12B), 2554-2568.
- Hoskins, R. A., Smith, C. D., Carlson, J. W., Carvalho, A. B., Halpern, A., Kaminker, J. S., et al. (2002). Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biol*, 3(12), RESEARCH0085.
- 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(1), 186-190.
- Itoh, H., Washio, T., & Tomita, M. (2004). Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *Rna*, 10(7), 1005-1018.
- 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(5), 889-900.
- Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res*, 12(4), 656-664.
- Kersey, P., Hermjakob, H., & Apweiler, R. (2000). VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics*, 16(11), 1048-1049.
- Kondrashov, F. A., & Koonin, E. V. (2003). Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet*, 19(3), 115-119.

- Kong, L., Lee, B. T., Tong, J. C., Tan, T. W., & Ranganathan, S. (2004). SDPMD: an automated comparative modeling server for small disulfide-bonded proteins. *Nucleic Acids Res*, 32(Web Server issue), W356-359.
- Krawczak, M., Ball, E. V., Fenton, I., Stenson, P. D., Abeyasinghe, S., Thomas, N., et al. (2000). Human gene mutation database—a biomedical information and research resource. *Hum Mutat*, 15(1), 45-51.
- Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S., et al. (2003). Increase of functional diversity by alternative splicing. *Trends Genet*, 19(3), 124-128.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921.
- Lee, B. T., Tan, T. W., & Ranganathan, S. (2003). MGAlignIt: A web service for the alignment of mRNA/EST and genomic sequences. *Nucleic Acids Res*, 31(13), 3533-3536.
- Lee, B. T., Tan, T. W., & Ranganathan, S. (2004). DEDB: a database of *Drosophila melanogaster* exons in splicing graph form. *BMC Bioinformatics*, 5(1), 189.
- Lee, C., Atanelov, L., Modrek, B., & Xing, Y. (2003). ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res*, 31(1), 101-105.
- Lehmann, K., & Schmidt, U. (2003). Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit Rev Biochem Mol Biol*, 38(3), 249-303.
- Leipzig, J., Pevzner, P., & Heber, S. (2004). The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res*, 32(13), 3977-3983.
- Lopez, A. J. (1998). Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet*, 32, 279-305.
- McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., et al. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, 385(6614), 357-361.
- Mironov, A. A., Fickett, J. W., & Gelfand, M. S. (1999). Frequent alternative splicing of human genes. *Genome Res*, 9(12), 1288-1293.
- Misra, S., Crosby, M. A., Mungall, C. J., Matthews, B. B., Campbell, K. S., Hradecky, P., et al. (2002). Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol*, 3(12), RESEARCH0083.
- Misteli, T. (2000). Different site, different splice. *Nat Cell Biol*, 2(6), E98-E100.
- Misteli, T., & Spector, D. L. (1999). RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. *Mol Cell*, 3(6), 697-705.
- 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(13), 2850-2859.
- Morris, D. P., & Greenleaf, A. L. (2000). The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. *J Biol Chem*, 275(51), 39935-39943.
- Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O., & Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res*, 20(16), 4255-4262.

- Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., et al. (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, *31*(1), 315-318.
- Nagengast, A. A., & Salz, H. K. (2001). The Drosophila U2 snRNP protein U2A' has an essential function that is SNF/U2B" independent. *Nucleic Acids Res*, *29*(18), 3841-3847.
- Nilsen, T. W. (2002). The spliceosome: no assembly required? *Mol Cell*, *9*(1), 8-9.
- Padgett, R. A., Hardy, S. F., & Sharp, P. A. (1983). Splicing of adenovirus RNA in a cell-free transcription system. *Proc Natl Acad Sci U S A*, *80*(17), 5230-5234.
- Robberson, B. L., Cote, G. J., & Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol*, *10*(1), 84-94.
- Rogan, P. K., Faux, B. M., & Schneider, T. D. (1998). Information analysis of human splice site mutations. *Hum Mutat*, *12*(3), 153-171.
- Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., et al. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, *101*(6), 671-684.
- Schneider, C., Will, C. L., Brosius, J., Frilander, M. J., & Luhrmann, R. (2004). Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein particle in Drosophila. *Proc Natl Acad Sci U S A*, *101*(26), 9584-9589.
- Schneider, T. D. (1997). Information content of individual genetic sequences. *J Theor Biol*, *189*(4), 427-441.
- Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, *18*(20), 6097-6100.
- Schneider, T. D., Stormo, G. D., Gold, L., & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J Mol Biol*, *188*(3), 415-431.
- Sharp, P. A. (1994). Split genes and RNA splicing. *Cell*, *77*(6), 805-815.
- Sisodia, S. S., Sollner-Webb, B., & Cleveland, D. W. (1987). Specificity of RNA maturation pathways: RNAs transcribed by RNA polymerase III are not substrates for splicing or polyadenylation. *Mol Cell Biol*, *7*(10), 3602-3612.
- Smale, S. T., & Tjian, R. (1985). Transcription of herpes simplex virus tk sequences under the control of wild-type and mutant human RNA polymerase I promoters. *Mol Cell Biol*, *5*(2), 352-362.
- Staley, J. P., & Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, *92*(3), 315-326.
- Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., et al. (2005). Function of alternative splicing. *Gene*, *344*, 1-20.
- Stamm, S., Riethoven, J. J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., et al. (2006). ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res*, *34*(Database issue), D46-55.
- Talerico, M., & Berget, S. M. (1994). Intron definition in splicing of small Drosophila introns. *Mol Cell Biol*, *14*(5), 3434-3445.
- Tani, T., & Ohshima, Y. (1991). mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing. *Genes Dev*, *5*(6), 1022-1031.

- Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J. J., Le Texier, V., & Muilu, J. (2004a). ASD: the Alternative Splicing Database. *Nucleic Acids Res*, *32* Database issue, D64-69.
- Thanaraj, T. A., Stamm, S., Clark, F., Riethoven, J. J., Le Texier, V., & Muilu, J. (2004b). ASD: the Alternative Splicing Database. *Nucleic Acids Res*, *32*(Database issue), D64-69.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304-1351.
- Wheelan, S. J., Church, D. M., & Ostell, J. M. (2001). Spidey: a tool for mRNA-to-genomic alignments. *Genome Res*, *11*(11), 1952-1957.
- Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., et al. (2003). Database resources of the National Center for Biotechnology. *Nucleic Acids Res*, *31*(1), 28-33.
- Zhang, X. H., Heller, K. A., Hefter, I., Leslie, C. S., & Chasin, L. A. (2003). Sequence information for the splicing of human pre-mRNA identified by support vector machine classification. *Genome Res*, *13*(12), 2637-2650.
- Zheng, C. L., Kwon, Y. S., Li, H. R., Zhang, K., Coutinho-Mansfield, G., Yang, C., et al. (2005). MAASE: an alternative splicing database designed for supporting splicing microarray applications. *Rna*, *11*(12), 1767-1776.

Appendix A: Alternative splicing bioinformatics resources

Resource name	Type	URL
GenBank	Nucleotide database	http://www.ncbi.nlm.nih.gov/Genbank/index.html
Entrez Genome	Genome database	http://www.ncbi.nlm.nih.gov/Genomes/
Swiss-Prot	Protein database	http://www.expasy.org/
UniGene	Clustered EST database	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=unigene
Pfam	Protein domain database	http://www.sanger.ac.uk/Software/Pfam/
Alternative Splicing Database (ASD)	Alternative splicing database	http://www.ebi.ac.uk/asd/
ASDB	Alternative splicing database	http://hazelton.lbl.gov/~teplitski/alt/
Human Alternative Splicing Database (HASDB)	Alternative splicing database	http://www.bioinformatics.ucla.edu/~splice/HASDB/
Alternative Splicing Annotation Project (ASAP)	Alternative splicing database	http://www.bioinformatics.ucla.edu/ASAP/
Putative Alternative Splicing database (PALS db)	Alternative splicing database	http://palsdb.ym.edu.tw/
Basic Local Alignment Search Tool (BLAST)	Sequence alignment	http://www.ncbi.nlm.nih.gov/blast
sim4	Sequence alignment	http://globin.cse.psu.edu/globin/html/docs/sim4.html
Spidey	Sequence alignment	http://www.ncbi.nlm.nih.gov/spidey/
MGAlign	Sequence alignment	http://proline.bic.nus.edu.sg/mgalign/
Blat	Sequence alignment	http://www.genomeblat.com/genomeblat/index.asp
HMMER	HMM sequence analysis	http://hmmer.wustl.edu/

Appendix B: Oral and poster presentations

Oral presentations

1. MGAlign and Alternative Splicing

Singapore Bioinformatics Symposium 2003, Singapore, 15 August 2003.

2. MGAlign, a tool for aligning mRNA sequences to genomic sequences

Institute of Mathematical Sciences: Post-Genome Knowledge Discovery Workshops, Singapore, 2002.

Poster presentations

1. DEDB: a splicing graph database

Alternate Transcript Diversity Symposium, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK, 2004.

2. Drosophila melanogaster Exon Database (DEDB)

12th International Conference on Intelligent Systems for Molecular Biology (ISMB), Glasgow, Scotland, 2004.

3. MGAlign.

1st Bilateral Symposium on Advances in Molecular Biotechnology and Biomedicine, Singapore, 2002.

4. MGAlign, a tool for aligning mRNA sequences to genomic sequences

10th International Conference on Intelligent Systems for Molecular Biology (ISMB), Edmonton, Canada, 2002.

Appendix C: Publications

MGAlignIt: a web service for the alignment of mRNA/EST and genomic sequences

Bernett T. K. Lee¹, Tin Wee Tan¹ and Shoba Ranganathan^{1,2,*}

¹Department of Biochemistry and ²Department of Biological Science, National University of Singapore, 8 Medical Drive, Singapore 117597, Singapore

Received February 14, 2003; Revised and Accepted April 1, 2003

ABSTRACT

Splicing is a biological phenomenon that removes the non-coding sequence from the transcripts to produce a mature transcript suitable for translation. To study this phenomenon, information on the intron–exon arrangement of a gene is essential, usually obtained by aligning mRNA/EST sequences to their cognate genomic sequences. MGAlign is a novel, rapid, memory efficient and practical method for aligning mRNA/EST and genome sequences. We present here a freely available web service, MGAlignIt (<http://origin.bic.nus.edu.sg/mgalign/mgalignit>), based on MGAlign. Besides the alignment itself, this web service allows users to effectively visualize the alignment in a graphical manner and to perform limited analysis on the alignment output. The server also permits the alignment to be saved in several forms, both graphical and text, suitable for further processing and analysis by other programs.

INTRODUCTION

The completion of a draft human genome (1,2) in February 2001 revealed that the human genome has only ~30 000–40 000 genes, compared to the initial estimate of 100 000 genes, to account for the complexity in the species. Drafts of numerous other genomes (3–8) have been completed and once again, it is observed that the number of genes do not increase greatly with the complexity of the organism. However, the number of proteins in the organism is far in excess of the number of genes, as there exist numerous cellular mechanisms that lead to multiple gene products from a single gene. Alternative splicing, the differential joining of exons, is one of these mechanisms and perhaps the most important one (9–11).

To understand and study splicing and its implications, we need to determine the intron–exon arrangement in a gene, typically unraveled by performing a sequence alignment of the mRNA sequence of the gene with its cognate genomic sequence (12). A database of human and mouse genes showing alternate exon arrangements is an example of such a

study (Alternate Exon Database, <http://www.ebi.ac.uk/asd/altextron/access.html>). This approach produces a global alignment consisting of several local alignments, with each local alignment being a single exon. The usefulness of this approach in obtaining intron–exon arrangement information is enhanced by the fact that complete genomic sequences of several organisms are now available, thus providing part of the data for such alignments. Other than genome sequencing projects, there are several projects seeking to generate full-length cDNA sequences (13,14) of organisms under way. Besides full-length cDNA sequences, there exist a huge amount of ESTs (expressed sequence tags) in public databases like dbEST (15). The availability of these resources means that obtaining intron–exon arrangement information of genes from sequence alignments of mRNA/EST sequences to genomic sequences is very practical nowadays.

MGAlign (Lee, B.T.K., Tan, T.W. and Ranganathan, S., unpublished results) is a new method that aligns mRNA/EST sequences to genomic sequences using a rapid heuristic method. A web interface at <http://origin.bic.nus.edu.sg/mgalign/mgalignit.html> that utilizes MGAlign as the alignment technique is available for users to easily visualize the intron–exon arrangements and to perform limited analysis on the alignments. In this paper, we will describe the web service, MGAlignIt, as an alignment service.

OVERVIEW OF THE WEB SERVICE

The MGAlignIt web service provides a dynamic means of aligning pairs of mRNA/EST and genomic sequences at the same website, based on MGAlign, a method that rapidly aligns an mRNA/EST sequence to its cognate genomic sequence using a heuristic approach. The software is available in binary form for several platforms (Microsoft Windows, Sun Solaris, Linux, Mac OS X and SGI IRIX) at <http://origin.bic.nus.edu.sg/mgalign>. The web server aims to provide the global research community with a free alignment portal for mRNA/EST and genomic sequences, using MGAlign, with output options enabling simple yet intuitive visualization of the intron–exon arrangement graphically, a rapid means of determining the effects of alternative splicing on the alignment, the ability to perform limited analysis on the alignment and the facility to

*To whom correspondence should be addressed. Tel: +65 68743566; Fax: +65 67782466; Email: shoba@bic.nus.edu.sg

save the alignment in various formats for integration with other tools.

MGAlign fulfils these collective aims using a three-step process. The first step tackles the alignment itself, by basically aligning the mRNA/EST and genomic sequences. This is followed by a visualization step where the user can visualize and perform limited analysis on the alignment. Lastly, the user is able to request outputs of the alignment in various forms, for local analysis, publication and presentation.

ALIGNMENT

A schematic diagram of the algorithm used by MGAlign is shown in Figure 1. The distinguishing feature of the algorithm is in the heuristics used to reduce the search space. MGAlign achieves this reduction in search space by using two search phases instead of one. The first search phase aims to locate for a pair of matches on the genomic sequence such that the rest of the alignment lies between this pair of matches as shown in Figure 1B. Thus the algorithm has only to search within this segment of genomic sequence bounded by the pair of matches for the rest of the alignment. In the event that more than one pair of matches is found, the process given in Figure 1 is repeated for each pair of matches and the one with the best score reported. After the pair of matches is found, sub-sequences of the mRNA sequence is then used to locate for local alignments within the segment of genomic sequence bounded by the pair of matches. There is a possibility that the same sub-sequence will result in more than one local alignment. Therefore the algorithm has to select a subset of the local alignments that maximizes the amount of aligned mRNA sequence as shown in Figure 1C. The next two steps involve filling in of gaps using shorter sub-sequences and trimming of overlapping exons using information on splice site motifs as illustrated in Figure 1D and E. More details of the algorithm are described elsewhere (Lee, B.T.K., Tan, T.W. and Ranganathan, S., unpublished results), with benchmarking results indicating that MGAlign is more accurate and faster than sim4 and Spidey, with only limited memory requirements (available from <http://origin.bic.nus.edu.sg/mgalign/comparison.html>).

VISUALIZATION

Once the alignment is finished, the output of the algorithm is parsed into a web page (Fig. 2) consisting of three frames to visualize and analyze the alignment. The top frame entitled 'Top Menu' includes a status bar that provides feedback and instructions to users as well as a link to bring the users back to the homepage. The middle frame named 'Graphical View' provides a graphical representation of the alignment [a Portable Network Graphics (PNG) image], which can be enlarged or reduced in size via the zoom in/out buttons for better display. This is especially useful for users as it provides an overall view of the alignment, giving them a sense of the length and distribution of the exons and introns. The genome sequence is represented by a black bar, followed by the mRNA/EST sequence, represented in two ways. The top series of colored bars (each bar corresponding to an individual exon) have their dimensions and positions scaled relative to the length of the genomic sequence. This gives the users a feel of the size and

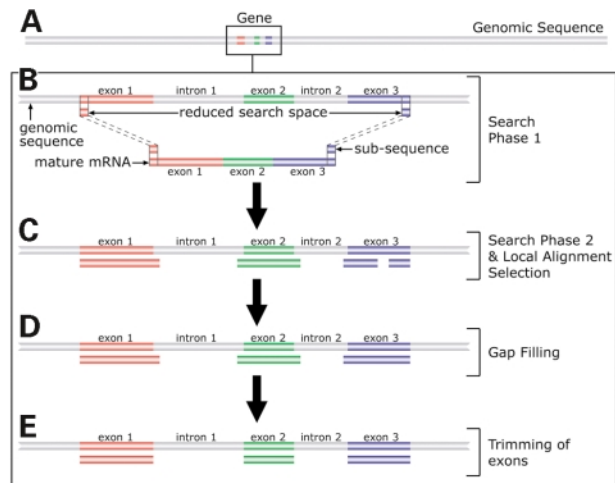


Figure 1. Schematic diagram of MGAlign's approach. (A) The box represents the region occupied by the mRNA/EST sequence on the genomic sequence (grey bar). (B) Search Phase 1 locates matches on the genomic sequence using short sub-sequences from the ends of the mRNA sequence, to reduce the alignment space. (C) Search Phase 2 locates regions of local alignment use by searching within the reduced search space. (D) Alignment gaps are filled by searching with smaller sub-sequences. (E) Lastly overlaps between the exons are trimmed based on splice site motifs.

distribution of each exon as compared to the genomic sequence. The use of alternating colors helps in differentiating exons, which could be a problem when they are closely packed. The bottom series of colored bars have their sizes scaled to the length of the mRNA/EST sequence. This provides users a means of estimating the relative size and location of each exon and to aid in correlating the exons from the top series to the bottom series, connected by grey lines. The splice site motifs, intron phases and the start and end positions of the exons are indicated on the bottom series of colored bars. This allows users to determine at a glance if there are any non-canonical splice sites and the effects of any alternative splicing. For example, Figure 2B shows the 5' portion of an alignment, where a frame shift in the coding sequence will result if exon 2 is spliced out, since exon 2 is flanked by introns of different phases. However, if exon 5 were to be spliced out, there would not be any frame shift as the introns bordering exon 5 are in phase. To further aid in determining the effects of alternative splicing, the longest open reading frame (ORF) of the mRNA sequence is drawn below the second series of colored bars as a light blue colored bar, allowing easy identification of the coding region and rapid determination of the effects of any frame shifts. For example, insertion of a frame shift-inducing exon between exons 1 and 2 would not affect the coding sequence as the start of the coding sequence lies within exon 2. As the mRNA/EST sequence could potentially have several ORFs, there is an option to select which ORF is to be used in the image. The amino acid sequence corresponding to each ORF is also linked to the BLAST (16) search page at NCBI (National Center for Biotechnology Information) for fast protein identification. The image of the alignment is itself an intuitive navigational tool. Users can get detailed information about specific parts of the alignment by clicking on elements in the image. Selecting an exon by clicking on it, provides detailed information about that exon in the bottom frame.

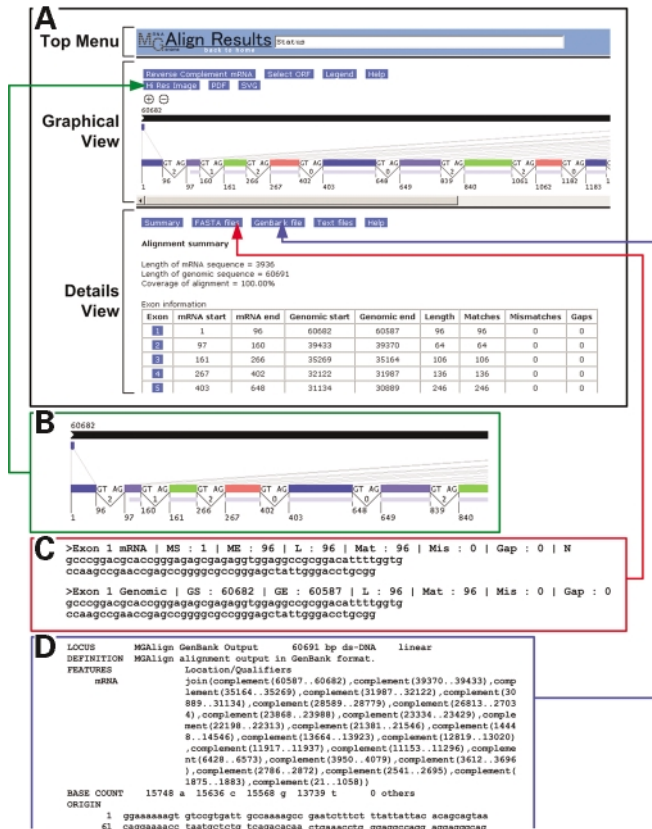


Figure 2. Results from the MAlignIt web server. (A) The main results page is divided into three frames entitled 'Top Menu', 'Graphical View' and 'Details View'. (B) Clicking the 'Hi Res Image' button leads to the high-resolution image, part of which is shown here. (C) Alignments of individual exons, introns and ORFs can be obtained by clicking on the 'FASTA files' button. The FASTA alignment for exon 1 is shown here. (D) A GenBank formatted record of the alignment is presented by clicking the 'GenBank file' button.

The bottom frame entitled 'Details View' provides additional information about specific portions of the alignment, such as the pair-wise alignment view of an exon with its cognate genomic sequence. The information to be displayed in this frame is controlled by selecting specific elements in the image found in the 'Graphical View' and the buttons found in the 'Details View'. The interface is designed such that users can focus on the information they require. An instance of this flexible display option would be to view the alignment near a specific splice site. This is achieved by selecting the specific intron in the image in the 'Graphical View' and then clicking on the button entitled 'Splice site details' in the 'Details View'. The desired information is then displayed in the 'Details View'. Users can also do limited analysis within this frame. One such analysis is searching for patterns in both exonic and intronic sequences. This is especially useful for determining the existence of specific signals like branch sites in intronic sequences or exonic splicing enhancers in exonic sequences. Users can opt to use their own patterns or they may select one of the predefined ones available from the MAlignIt website. The syntax followed is the one prescribed by PROSITE (17) allowing rapid and flexible searches. Besides pattern searches, exonic sequences and ORFs are

linked to the NCBI's BLAST search page, permitting immediate access to homolog detection, by searching the GenBank (18) database. Links to online help is provided on every MAlignIt page. Selecting these links provide users with context specific help. In addition, there is an online tutorial on using the web services and as well as comprehensive help pages for all aspects of the output. An email helpline is also available for MAlignIt users.

OUTPUT OPTIONS

Besides a simple visual representation of the alignment on the web page, the web service also provides many forms of output that users can save on their local computers. The graphical view of the alignment can be saved in three different formats. The first of these formats is a high resolution PNG raster image of the alignment, which appears in a new window when users click on 'Hi Res Image' in the 'Graphical View' frame. This provides users with a commonly used image format that can be incorporated into many applications. To facilitate scaling the image size and high quality printing, two additional formats are available. In the PDF (Portable Document Format) format, the alignment is shown as an image in an A4-sized PDF file, which users can save onto their computers. This format is compact and excellent for printing, as the image has already been resized to fill the paper and can be viewed and printed from any computer platform via a PDF viewer. The PDF file may also be imported and edited using common graphic applications such as Adobe Illustrator, specially for adding custom annotations to the alignment. The last graphical format is SVG (Scalable Vector Graphics), which is also a vector-based image format recommended as the standard by the W3C (the World Wide Web Consortium) that provides similar advantages as the PDF format. Applications such as Adobe Illustrator can be used to edit and annotate SVG images.

Other than graphical representations, the alignment can also be saved in two commonly used sequence formats, FASTA and GenBank. Users can obtain FASTA files of individual exons or introns as well as composite files containing all introns or all exons. Useful information about the alignment is stored in the header line of the FASTA files in a format that is easy for computers to parse. Each exon is stored as two FASTA sequences, one being the exonic portion of the alignment and the other being the genomic portion of the alignment as seen in Figure 2C. The FASTA format is a widely used format and making the alignment available in the FASTA format means that many other software can read the alignment generated. This permits incorporation of the alignment into other programs for further analysis.

The GenBank formatted record does not have the pairwise alignment information of each exon but it does allow for annotation of the genomic sequence in the Feature Table portion of the record. The alignment is annotated as an mRNA feature in the Feature Table as shown in Figure 2D. This facilitates the alignment to be used in other programs that read GenBank formatted records such as Artemis (19). Artemis is a standalone DNA sequence visualization and annotation tool.

The users can also save summaries of the exon, intron and ORF information. These files are provided as comma delimited text files, such that they can be viewed and formatted by several spreadsheet programs such as Microsoft Excel. This text output format is a very useful companion to the graphical output as the graphical output does not provide the full details of the alignment.

CONCLUSION

The MGAlignIt web service provides a platform for generating alignments of mRNA/EST sequences to their cognate genomic sequences in a freely accessible manner. In addition to providing a means for alignment, the web service makes possible easy visualization of the alignment, allowing users to quickly determine the effects of any changes in the splicing pattern. Furthermore, the web service allows further analysis such as BLAST searches and pattern searching. Lastly users can obtain soft copies of the alignment in several formats, including high resolution graphics, for further analysis, publication and printing.

ACKNOWLEDGEMENTS

We would like to thank our colleagues at the Department of Biochemistry, National University of Singapore for their helpful comments and discussions. B.T.K.L. would also like to thank the National University of Singapore for the award of an Agency for Science, Technology and Research, Singapore (ASTAR) scholarship.

REFERENCES

- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
- Blattner, F.R., Plunkett, G., III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
- Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R.K. *et al.* (1997) Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, **387**, 67–73.
- Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.
- The Arabidopsis Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium. *Science*, **282**, 2012–2018.
- Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.
- Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
- Black, D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.
- Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
- Clark, F. and 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.
- Strausberg, R.L., Feingold, E.A., Klausner, R.D. and Collins, F.S. (1999) The mammalian gene collection. *Science*, **286**, 455–457.
- Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
- Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST—database for ‘expressed sequence tags’. *Nature Genet.*, **4**, 332–333.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K. and Bairoch, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.

Database

Open Access

DEDB: a database of *Drosophila melanogaster* exons in splicing graph form

Bernett TK Lee¹, Tin Wee Tan¹ and Shoba Ranganathan*^{1,2}

Address: ¹Department of Biochemistry, National University of Singapore, Singapore and ²Biotechnology Research Institute, Macquarie University, Sydney, Australia

Email: Bernett TK Lee - bernett@bic.nus.edu.sg; Tin Wee Tan - tinwee@bic.nus.edu.sg; Shoba Ranganathan* - shoba@els.mq.edu.au

* Corresponding author

Published: 07 December 2004

Received: 31 August 2004

BMC Bioinformatics 2004, **5**:189 doi:10.1186/1471-2105-5-189

Accepted: 07 December 2004

This article is available from: <http://www.biomedcentral.com/1471-2105/5/189>

© 2004 Lee et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: A wealth of quality genomic and mRNA/EST sequences in recent years has provided the data required for large-scale genome-wide analysis of alternative splicing. We have capitalized on this by constructing a database that contains alternative splicing information organized as splicing graphs, where all transcripts arising from a single gene are collected, organized and classified. The splicing graph then serves as the basis for the classification of the various types of alternative splicing events.

Description: DEDB <http://proline.bic.nus.edu.sg/dedb/index.html> is a database of *Drosophila melanogaster* exons obtained from FlyBase arranged in a splicing graph form that permits the creation of simple rules allowing for the classification of alternative splicing events. Pfam domains were also mapped onto the protein sequences allowing users to access the impact of alternative splicing events on domain organization.

Conclusions: DEDB's catalogue of splicing graphs facilitates genome-wide classification of alternative splicing events for genome analysis. The splicing graph viewer brings together genome, transcript, protein and domain information to facilitate biologists in understanding the implications of alternative splicing.

Background

The completion of the draft sequence of the *Drosophila melanogaster* genome in March 2000 [1,2] and the availability of quality annotations by FlyBase in 2002 [3] presents an excellent opportunity for the study of alternative splicing. Although the annotations themselves provide an insight to the amount of alternative splicing, they do not provide any classification of the types of alternative splicing events present. Different forms of alternative splicing have different biological bases and the classification of alternative splicing events is critical for further work in deciphering the regulatory controls that govern

these processes. To this end, we transformed all known gene structure information obtained from the genome annotations into splicing graphs based on the approach first proposed by Heber et al. in 2002 [4]. We then created simple but robust rules for classifying the splicing graphs into various alternative splicing events. The rules created allows for the detection of multiple forms of alternative splicing within the same gene. To facilitate the assessment of the impact of alternative splicing on the protein product in particular with respect to the domain organization of the protein, Pfam [5] domains were mapped onto the transcripts using HMMER [6]. All these data were then

loaded into DEDB (*Drosophila melanogaster* Exon Database) [7]. To aid in visualizing these splicing graphs, a web-based splicing graph viewer was also developed. The splicing graph viewer integrates gene structure, transcript, protein and domain information into an easily understandable interface that is viewable with any current web browser. The splicing graphs as well as the alternative splicing event classifications are available for download as XML files. A XML schema is available for parsing and validation of the XML files.

Construction and content

Data source

Drosophila melanogaster genome annotations (release 3.2) were obtained from FlyBase [8] as Game XML files. Gene structure information including the location of the transcript, the start and end positions of each exon that make up the transcript and the protein coding region were parsed out, checked for consistency and then loaded into a relational database (MySQL). Pfam HMM models were retrieved from Pfam release 12 and used as the database for the hmmpfam program (part of HMMER) to search the transcript protein sequences for structural domains, with an expectancy values of less than 0.001. The results of the search were parsed, mapped onto the protein sequence and imported into the database.

Construction of the splicing graphs

The transcripts contained in the database were retrieved and clustered on the basis that they occupy overlapping genomic positions. Splicing graphs are then constructed using these clusters of transcripts. In each cluster, exons and introns having identical start and end positions are merged into nodes and connections respectively. The nodes are then linked via connections to form the complete splicing graph that is loaded into the database. In cases where the transcripts are located on the negative strand, the entire splicing graph is reverse complemented so that all the splicing graphs contained in the database have sense strand nucleotide sequences. These steps are graphically depicted in Figure 1. The result of this process generated 13,222 splicing graphs of which 2,646 are alternatively spliced. Details of the contents of the database are found in Table 1 and on the website via the "Stats" link.

Classification of alternative splicing

Rules are then derived to detect specific alternative splicing events as shown in Figure 2 (details and examples of the rules are available on the website). Apart from the classical alternative splicing events like cassette exons, intron retention, alternative donor sites and alternative acceptor sites, we have also elected to classify other gene structure events like alternative transcriptional start/termination sites as well as alternative initiation/termination exons. The reason for the existence of the alternative initiation/

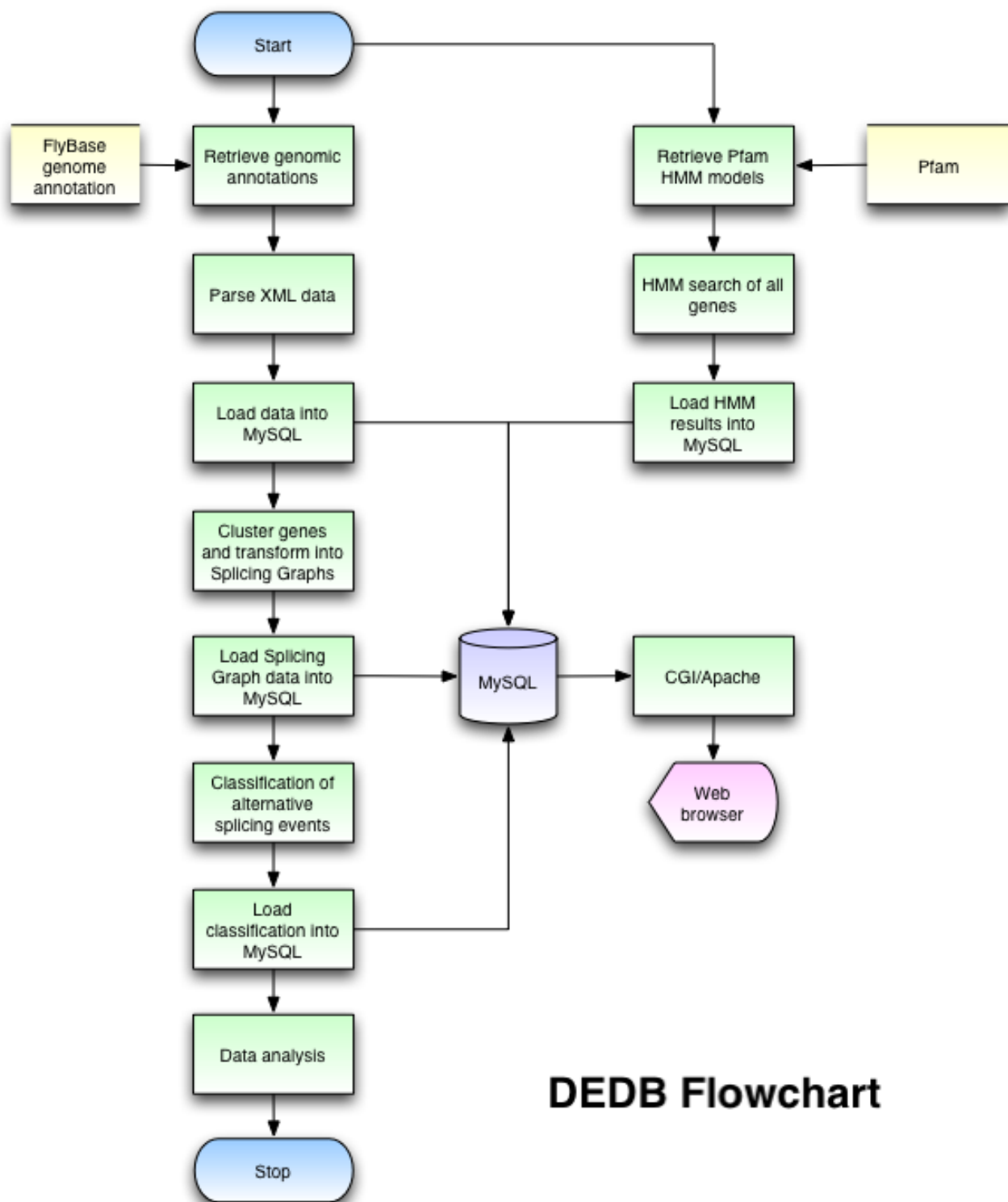
termination exon categories is due to the fact that the 5' and 3' ends of the transcripts are usually difficult to determine experimentally and are thus less accurate. Therefore, any differences in the start and end of the transcripts could be simply due to the sequencing difficulties. The inclusion of the alternative initiation/termination exons category is an attempt to circumvent this problem as alternative initiation/termination exons (which are classified based on the end position of initiation exons and the start position of termination exons) are unaffected by the sequencing difficulties and thus represent true alternative exons. Alternative transcriptional start/termination sites, however, are dependent on sequencing results and provide a means of classifying gene segments with differences in the start positions of initiation exons and the end positions of termination exons, with a view to updating entries in this category, when the 5' and 3' ends of these transcripts are determined accurately. These rules were then used on all the splicing graphs and the detected alternative splicing events loaded into the database.

Access

The database together with the splicing graph viewer is freely available at <http://proline.bic.nus.edu.sg/dedb/index.html>. Users can query the database using FlyBase gene names, FlyBase Gene Symbols, Pfam Accession Numbers or Pfam Identifiers via the query page. Users can also query the database using BLAST [9] searches. This is particularly useful if one wishes to know the *Drosophila melanogaster* homology together with alternative splicing information of a particular gene. Lists of splicing graphs for the various types of alternative splicing events are also provided on the website for users who are interested in a certain type of alternative splicing. For users who wish to use large subsets of the data, they can download the XML files available from the same site. To aid parsing and validation of the XML file, a XML schema is available. DEDB can also be accessed via links on Flybase gene records, under the external database links section. Correspondingly, the DEDB Splicing Graph Viewer provides links back to FlyBase Gene and Annotation records, where experimental evidence for the gene structure has also been provided. Basic statistical analysis of the database can be found at the DEDB website <http://proline.bic.nus.edu.sg/dedb/stats.html>.

Splicing graph viewer

The splicing graph viewer consists of HTML pages created using a series of Python CGI (common gateway interface) scripts served by the Apache web server. The splicing graph viewer (Figure 3) is a three frame HTML page that shows the splicing graph in the center frame with detailed textual information in the bottom frame and navigation aids in the top frame (details elaborated on the website). The content is organized such that generalized



DEDB Flowchart

Figure 1

Flowchart depicting the process used to generate DEDB. Processes and data sources are coloured green and yellow respectively. The main data store is a MySQL server and the data housed is exposed to users using HTML pages and CGI (Common Gateway Interface) served by an Apache web server.

Table 1: Contents of DEDB. Table showing a breakdown of the contents in the database.

Item	Number
Total number of transcripts	18,156
Total number of single exonic genes	2,374
Total number of multi exonic genes	10,848
Total number of splicing graphs	13,222
Total number of exons	88,403
Total number of introns	70,247
Total number of nodes	60,744
Total number of connections	46,090
Total number of splicing graphs having alternative splicing events	2,646
Total number of splicing graphs having alternative TSS events	1,696
Total number of splicing graphs having alternative TTS events	691
Total number of splicing graphs having alternative initiation exon events	1,858
Total number of splicing graphs having alternative termination exon events	504
Total number of splicing graphs having alternative acceptor events	384
Total number of splicing graphs having alternative donor events	587
Total number of splicing graphs having cassette exon events	644
Total number of splicing graphs having intron retention events	700
Total number alternative TSS events	4,211
Total number alternative TTS events	1,511
Total number alternative initiation exon events	4,723
Total number alternative termination exon events	1,104
Total number alternative acceptor events	905
Total number alternative donor events	1,399
Total number alternative cassette exon events	1,228
Total number alternative intron retention events	983

information is provided to the users initially, permitting users to quickly zoom in on the information they need by clicking on an item of interest. The splicing graph shown in Figure 3 has alternative acceptor sites, where nodes 1 and 3 are alternatively used. Each node here can be selected and the corresponding node information, with sequence details is then dynamically displayed in the bottom frame. The transcripts leading to this splicing graph are depicted below the splicing graph. The connections in each transcript represent introns, which can also be selected to obtain intron-specific information. The rationale for providing the transcripts is because not all the paths in the splicing graphs are expressed transcripts, so the connections depicted in the splicing graph view are transcript-specific and thus not selectable by DEDB users. The provision of schematic diagrams of the transcripts along with the splicing graph provides the user with knowledge of which transcripts are expressed/detected. Below each transcript, links to structural domains, are shown as thin lines wherever available, linked to detailed information derived from Pfam and viewable on the bottom frame. The sequences displayed by clicking on the nodes in the splicing graph are always shown in the sense orientation to facilitate translation of coding sequences and Pfam mapping, while the exon and intron sequences shown by selecting genomic segments on the transcripts

will retain their original orientation identified by chromosomal mapping.

Utility and Discussion

Visualization of alternative splicing

By condensing all the various splicing variants into a single graph, where each splicing variant is a path through the graph, users can quickly establish the types and effects of various alternative splicing events present in the gene. Users can quickly pick up bifurcations which denote alternative splicing events far quicker than in the case of the traditional approach of presenting separate schematic representations of each splice variant, where the user has to correlate the splicing patterns from the transcript diagrams, to determine the impact and type of alternative splicing. The classical approach is particularly tedious in cases where the number of splice variants are numerous (for example the *Drosophila moe* gene, splicing graph 916, on the DEDB methodology link) resulting in the user having to correlate large amounts of data to comprehend all the alternative splicing events taking place. The DEDB schematic representation of the splicing graph is different from the one proposed by Heber et al. [4] and implemented in Alternative Splicing Gallery (ASG) [10]. The original representation used single linear block representation of exons connected by lines representing introns

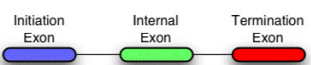
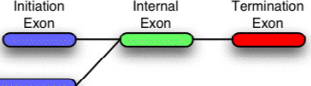
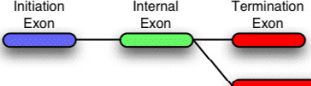
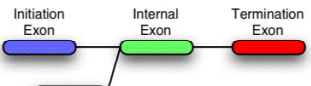
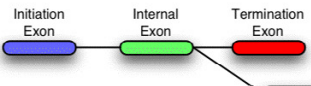
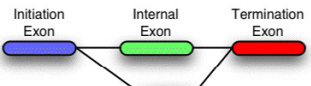

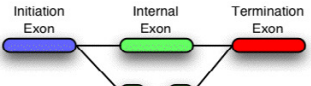
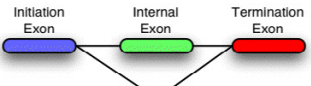
<p>Gene Structure</p> 	<p>A typical gene structure having three exons. The first exon is termed the initiation exon and is colored blue and the last exon is termed the termination exon and is colored red. All other exons are termed internal exons and are colored green.</p>
<p>Alternative transcriptional start site</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having no previous connections with a unique start position.
<p>Alternative transcriptional termination site</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having no next connections with a unique end position.
<p>Alternative initiation exon</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having no previous connections with a unique end position.
<p>Alternative termination exon</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having no next connections with a unique start position.
<p>Alternative acceptor site</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A set of overlapping nodes that are connected to a common upstream node. 2. The set of overlapping nodes should have unique start positions.
<p>Alternative donor site</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A set of overlapping nodes that are connected to a common downstream node. 2. The set of overlapping nodes should have unique end positions.
<p>Intron retention</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node having a connection whose start and end position falls within itself.
<p>Cassette exon</p> 	<p>Rule/s</p> <ol style="list-style-type: none"> 1. A node whose start and end position falls within a single connection. 2. The node should have at least one next and previous connection.

Figure 2
Rules used to classify splicing graphs in DEDB. Initiation, internal and termination exons are colored blue, green and red respectively. Each of the type of alternative splicing event and gene structure event is depicted as a single boxed row. Within each boxed row, a graphical illustration of the gene structure is shown on the left together with the textual rules on the right.

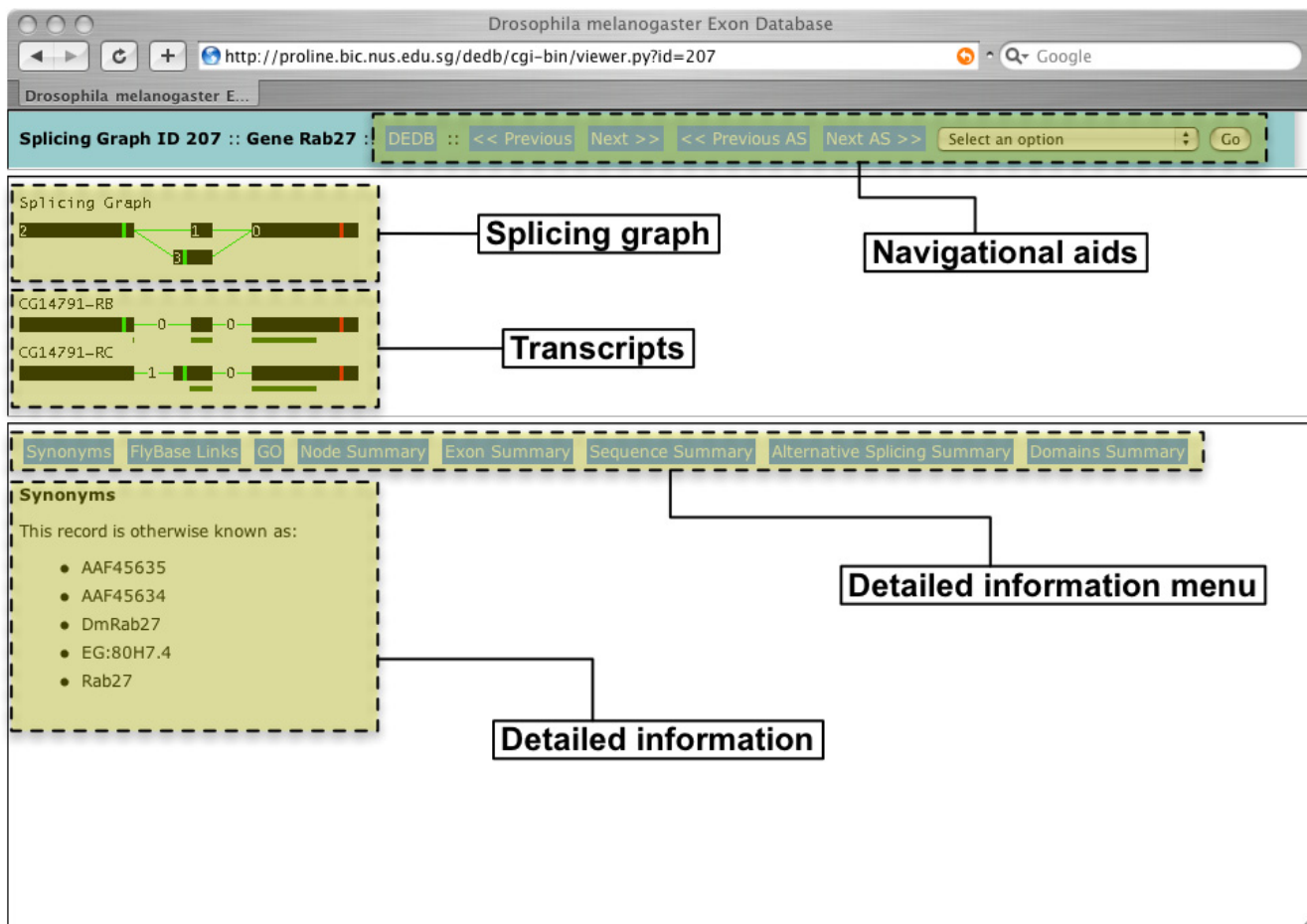


Figure 3
Screenshot of DEDB Splicing Graph Viewer. The viewer is divided into three frames. The top frame contains the navigational aids to allow users to quickly locate the splicing graph of interest. The middle frame shows the splicing graph together with all the transcripts used to generate the splicing graph. The splicing graph and the transcripts are interactive allowing users to click on them for more information in the bottom frame. The bottom frame shows textual information about the splicing graph, which are selectively displayed by either clicking on interactive elements in the middle frame or by clicking on the row of buttons in the bottom frame.

and alternative donor and acceptor sites as well as intron retention represented by single blocks. Instead, we have chosen to depict all the exons individually as we felt that this is more intuitive for biologists by making the impact of the alternative splicing more pronounced. Protein sequence details like the start and end of translation as well as detected Pfam domains are presented by the splicing graph viewer. This allows biologists to infer the impact of alternative splicing on the corresponding protein sequences as well as the domain organization. Users can also download FASTA sequences of specific entities like introns and exons for other analysis. Biologists may also be interested in the *Drosophila melanogaster* homology of their gene of interest, which is made possible through a

BLAST search on the DEDB query page. The splicing graph of the *Drosophila melanogaster* homology may provide insights into the possible splice variants in the gene of interest. It could also provide information on the level of conservation of alternative splicing between orthologs.

Classified splicing dataset

The use of splicing graphs allows the creation of simple but robust rules that can detect multiple distinct alternative splicing events within the same gene. Traditional approaches usually require the construction of more complex rules. For example, the detection of a cassette exon in the tradition approach requires that an internal exon be checked against all the introns in all the splicing variants

to detect instances where the exon falls within an intron. This process has to be repeated for each exon against all the introns resulting in a long and complex computation. Furthermore as the exon could be found in several splicing variants, the detected cassette exon could be redundant and additional steps have to be taken to remove this redundancy. All of this is avoided by the splicing graph representation, as it is a condensed view of all the various splicing variants arising from a single gene. Classification of the alternative splicing types in *Drosophila melanogaster* would allow users to target specific types of alternative splicing events for analysis. This is useful as the various types of alternative splicing have different biological bases and therefore exhibit different phenotypes. The analysis of these phenotypes will be greatly aided by a set of data that is specific to one form of alternative splicing as provided by DEDB. The availability of a clean datasets of alternative splicing events [11,12] has proved to be useful in providing insights into the phenomenon of alternative splicing [13]. The data available from DEDB would no doubt be useful to many users studying alternative splicing as a major factor leading to complexity in higher eukaryotes.

Initial analysis

A summary of the alternative splicing events in DEDB is presented in Table 1. Detailed statistical information (general statistics, exon and intron length statistics and motif analysis) are available from the "Stats" page of the website (Lee, Tan and Ranganathan, unpublished results). Note however that the genes models are constructed with far more 5' ESTs than 3' ESTs [3] and the results must be viewed in the light of available experimental EST data.

Of the total of 13,222 genes in DEDB, 2,646 (20%) are alternatively splicing. This is significantly less than the amount of alternative splicing found in higher eukaryotes like humans [14], but sufficient to indicate that alternative splicing is a common phenomenon in *Drosophila melanogaster*. The amount of alternative splicing increases to 24.4% if we consider transcript diversity in the 10,848 multi-exonic genes alone.

Failure of intron definition is more likely to result in intron retention as opposed to exon definition in which case, failure leads to cassette exons. Initial analysis of the DEDB data indicates a bias towards cassette exons (1,228) over intron retention (983) events, so that exon definition is less stringent than intron definition. The short introns in *Drosophila melanogaster* are also thought to result in greater intron definition. The data observed could be due to the splicing machinery adopting a definition model dependent on the length of the intron or exon in question [15]. This is supported by the fact that cassette exons tend to be flanked by introns far longer than the mean value (exon and intron length statistics available via the "Stats"

link). The median for the cassette exon length is 150 bp in contrast with the flanking 5' and 3' introns, which are 653 bp and 639 bp respectively. The reverse is also true for intron retention where the median for the intron being retained is 101 bp while the flanking 5' and 3' exons are 163 bp and 261 bp respectively.

Information content analysis indicates that alternative donor and acceptor sites (with mean values of 5.95 and 5.61 bits) possess less information than constitutive sites (9.74 and 8.52 bits respectively; additional data available on website). This observation is consistent with the general notion that alternatively spliced exons exhibit splicing motifs deviating more from the consensus motifs [16]. Cassette exons (CE) and intron retentions (IR) also show lower mean individual information content on both donor (CE: 6.76 and IR: 5.39 bits) and acceptor sites (CE:7.08 and IR: 6.19 bits) as compared to constitutive exons.

The addition of Pfam domain information allows users to assess the impact of alternative splicing events on the proteins generated, enabling correlations not possible with the genome annotations alone.

Future work

Future work would focus on integrating other relevant information onto the splicing graphs, such as three-dimensional structural information as well as DEDB analysis results. Expansion of the splicing graph representation available in DEDB to other organisms is also underway.

Conclusions

The data housed in DEDB is organized as splicing graphs, which allows for ease of alternative splicing classifications. This has allowed DEDB to provide clean sets of data containing specific types of alternative splicing events. These specific sets of data could prove useful in understanding the biological basis of alternative splicing because different forms of alternative splicing have different biological basis. The splicing graph viewer provided allows biologists to quickly and intuitively understand the effects of alternative splicing on a gene of interest, thus aiding their research.

Availability and requirements

The database is available at <http://proline.bic.nus.edu.sg/dedb/index.html> suitable for most graphical web browser. XML files of the data contained in the database are also available together with an XML schema to aid parsing.

Authors' contributions

BTKL carried out the construction of the database as well as the splicing graph viewer. TTW and SR are responsible

for the database concept and participated in its design and construction. All authors have read and approved the final manuscript.

Acknowledgements

We would like to thank Dr Donald Gilbert for his help in the creating links to DEDB from the FlyBase gene records. We would also like to thank the bioinformatics team at the Department of Biochemistry, National University of Singapore and the anonymous reviewers for their helpful comments and discussions. Bennett Lee is grateful to the National University of Singapore for the award of an Agency for Science, Technology and Research, Singapore (A-STAR) scholarship.

References

- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidenkiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, Woodage T, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.
- Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, Kennedy C, Mungall CJ, Sullivan BA, Sutton GG, Yasuhara JC, Wakimoto BT, Myers EW, Celniker SE, Rubin GM, Karpen GH: **Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly**. *Genome Biol* 2002, **3**:RESEARCH0085.
- Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, Smith CD, Tupy JL, Whitfield EJ, Bayraktaroglu L, Berman BP, Bettencourt BR, Celniker SE, de Grey AD, Drysdale RA, Harris NL, Richter J, Russo S, Schroeder AJ, Shu SQ, Stapleton M, Yamada C, Ashburner M, Gelbart WM, Rubin GM, Lewis SE: **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review**. *Genome Biol* 2002, **3**:RESEARCH0083.
- Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA: **Splicing graphs and EST assembly problem**. *Bioinformatics* 2002, **18 Suppl 1**:S181-8.
- Bateman A, Birney E, Cerruti L, Durbin R, Etmiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database**. *Nucleic Acids Res* 2002, **30**:276-280.
- Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**:755-763.
- DEDB: *Drosophila melanogaster* Exon Database** [<http://prc.line.bic.nus.edu.sg/dedb/index.html>]
- The FlyBase Consortium: **The FlyBase database of the *Drosophila* genome projects and community literature**. *Nucleic Acids Res* 2003, **31**:172-175.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-410.
- Leipzig J, Pevzner P, Heber S: **The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome**. *Nucleic Acids Res* 2004, **32**:3977-3983.
- Lee C, Atanelov L, Modrek B, Xing Y: **ASAP: the Alternative Splicing Annotation Project**. *Nucleic Acids Res* 2003, **31**:101-105.
- Thanaraj TA, Stamm S, Clark F, Riethoven JJ, Le Texier V, Muilu J: **ASD: the Alternative Splicing Database**. *Nucleic Acids Res* 2004, **32 Database issue**:D64-9.
- Roca X, Sachidanandam R, Krainer AR: **Intrinsic differences between authentic and cryptic 5' splice sites**. *Nucleic Acids Res* 2003, **31**:6321-6333.
- Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms**. *FEBS Lett* 2000, **474**:83-86.
- Berget SM: **Exon recognition in vertebrate splicing**. *J Biol Chem* 1995, **270**:2411-2414.
- Itoh H, Washio T, Tomita M: **Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes**. *Rna* 2004, **10**:1005-1018.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



SDPMOD: an automated comparative modeling server for small disulfide-bonded proteins

Lesheng Kong¹, Bernett Teck Kwong Lee¹, Joo Chuan Tong¹, Tin Wee Tan¹ and Shoba Ranganathan^{1,2,*}

¹Department of Biochemistry, National University of Singapore, 8 Medical Drive, 117597, Singapore and
²Biotechnology Research Institute, Macquarie University, NSW 2109, Australia

Received February 15, 2004; Revised and Accepted March 24, 2004

ABSTRACT

Small disulfide-bonded proteins (SDPs) are rich sources for therapeutic drugs. Designing drugs from these proteins requires three-dimensional structural information, which is only available for a subset of these proteins. SDPMOD addresses this deficit in structural information by providing a freely available automated comparative modeling service to the research community. For expert users, SDPMOD offers a manual mode that permits the selection of a desired template as well as a semi-automated mode that allows users to select the template from a suggested list. Besides the selection of templates, expert users can edit the target–template alignment, thus allowing further customization of the modeling process. Furthermore, the web service provides model stereochemical quality evaluation using PROCHECK. SDPMOD is freely accessible to academic users via the web interface at <http://proline.bic.nus.edu.sg/sdpmod>.

INTRODUCTION

Small disulfide-bonded proteins (SDPs) are a special class of proteins that are relatively small in size (length ≤ 100 residues) and have disulfide bonds within their three-dimensional (3D) structures (1). SDPs include many secretory proteins which serve predatory, defensive or regulatory roles (such as toxins, inhibitors and hormones), and they are rich source for therapeutic drugs (2) and pesticides (3). The 3D structures of SDPs are essential for understanding the functions of SDPs and for drug design. However, 3D structure determination through experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are still both time-consuming and expensive. This results in a gap between the number of known 3D structures and the number of primary sequences that could be narrowed using large-scale automated protein structure prediction.

Among current structure prediction methods, comparative modeling is the most reliable method for generating 3D models. Comparative modeling of protein structures often requires expert knowledge and proficiency in specialized methods. In the mid-1990s, Peitsch and coworkers developed the first automated modeling server SWISS-MODEL (4), which is currently the most widely used server of this genre. Recently, several other automated comparative modeling servers have also been developed, such as CPHmodels (5), 3D-JIGSAW (6), ModWeb (7) and ESyPred3D (8).

Although so many automated comparative modeling servers are available, most of them do not work well on small SDPs for two reasons. Most of the automated servers are primarily designed for globular protein domains, making it difficult to discriminate small-sized SDPs from background noise. Taking as an example the sequence of α -conotoxin PnIA (9) (PDB id: 1PEN; 16 residues; 2 disulfide bridges in its structure), we note that both SWISS-MODEL and ModWeb report that they do not cover the modeling of sequences <25 or ≤ 30 amino acid residues in length, respectively, while the other three servers state that no suitable templates can be identified for this sequence.

The second reason is that SDPs have distinct characteristics from medium-sized and large globular proteins. They usually do not have a compact hydrophobic core, which is a major factor in stabilizing protein structure. Their side chains are more likely to be exposed to solvent and their conformations are more flexible. The 3D structures of small proteins are usually dominated by disulfide bridges, metal or ligand (according to SCOP classification) (10) and tend to bind or interact with large molecules. In small disulfide-rich proteins, the effects of disulfide bridges and constrained residues such as prolines are more significant than sequence similarity. As such, the comparative modeling rules for such proteins are highly specific and different from those adopted for large globular proteins. These distinct features require specific methods and datasets to be developed for the comparative modeling of SDPs.

To address these problems, we have first developed special strategies and rules for large-scale automated comparative modeling of the entire family of conotoxins (L. Kong and

*To whom correspondence should be addressed. Tel: +61 2 9850 6262; Fax: +61 2 9850 8313; Email: shoba@els.mq.edu.au

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

S. Ranganathan, unpublished data). Subsequently these rules were extended to other SDPs. Here, we present SDPMOD, a comprehensive comparative modeling server that is designed specifically for SDPs with specialized rules and datasets.

MATERIALS AND METHODS

Non-redundant SDP structure dataset

Before the modeling can proceed, a non-redundant dataset for SDPs needs to be created to serve as the template repository. Structures containing protein chains of length <100 amino acids with at least two cysteines were retrieved from the Protein Data Bank (PDB) (11) and loaded into MySQL, a relational database management system for flexible query and manipulation. The redundancy in SDP structures was removed at two levels. First, for NMR structures which have multiple monomer models, the representative monomers were selected using NMRCCLUS (12). Second, when multiple structures exist for the same sequence, the representative structure was chosen according to its structural qualities. The structural qualities are ranked by the following criteria (adopted from PDB): (i) X-ray structures over NMR structures, (ii) higher-quality factor ($1/\text{resolution}-R\text{-value}$) for X-ray structures and higher restraint per residue for NMR, (iii) better geometry, (iv) fewer missing atoms and non-standard residues and (v) later deposition date. Based on the above strategy, a non-redundant structure database for SDPs was generated. Currently it contains >1300 non-redundant protein chains and their coordinates. The database will be automatically updated once a month.

Modeling procedure

The SDPMOD server performs comparative modeling in four steps: (i) template selection, (ii) target-template alignment, (iii) model building and (iv) model evaluation (13). Figure 1 shows the detailed modeling procedure for automated modeling. The non-redundant dataset is first filtered using the number of cysteine residues, and the resulting template sequences are globally aligned to the target sequence using a modified scoring matrix derived from the non-redundant SDP dataset. The best templates are then selected based on the alignment scores. Target-template alignment and model building are achieved by MODELLER (14) (<http://salilab.org/modeller/modeller.html>), using a customized matrix to ensure that all the cysteine residues are well aligned. The final models are chosen according to the MODELLER objective function score, which reflects low energy and least stereochemical violations. Finally, the overall structural quality of the generated models is evaluated against stereochemical parameters derived from high-quality experimental structures by PROCHECK (15) (<http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html>).

Benchmarking

A large-scale benchmarking exercise was completed using the fully automated mode of the SDPMOD server. A control set of 664 sequences (a subset of our non-redundant SDP dataset) with known structures was used to evaluate the reliability of the server. The C α root mean square deviation (RMSD) values between models and their actual experimental

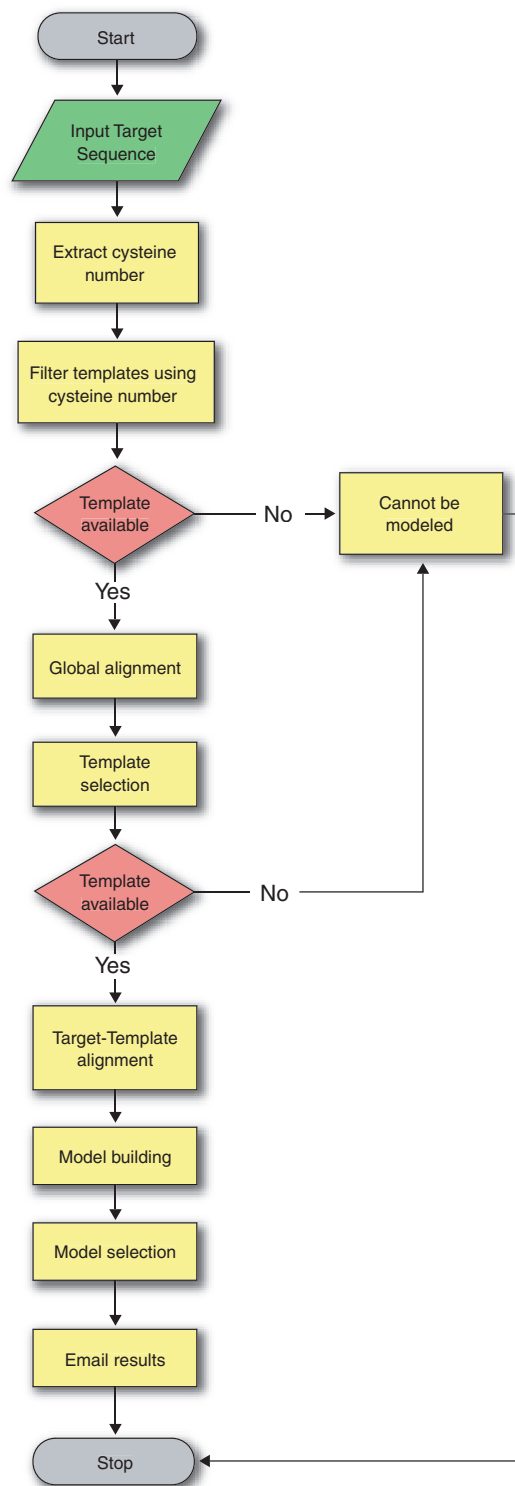


Figure 1. The SDPMOD methodology for automatic comparative modeling of small disulfide-bonded proteins.

structures were calculated. The benchmarking results show SDPMOD can predict 3D models with a reasonable accuracy. For example, in the 40–70% sequence identity range, 64% of models have C α RMSD values <1.5 Å. The detailed analysis of the accuracy of our modeling protocol is available from <http://proline.bic.nus.edu.sg/sdpmmod/accuracy.html>.

Menu Bar
Use this to navigate around the site

Email Address
Provide your email address to receive results

Sequence Input
Paste or input your sequence into this box. Use only plain sequence (length between 6 to 100 and cysteine residues greater than 2)

Submit Button
Click the 'Submit' button to do the prediction

Clear Button
Click the 'Clear' button to re set the email address, MODELLER key and sequence

MODELLER Key
Provide the MODELLER license key to run the server

Example Input Sequence
Click the 'raw format' link to obtain an example input sequence

Sali Lab
Register at Sali Lab to obtain a copy of MODELLER key to run the server

SDPMOD
HOME MODELING METHODOLOGY ACCURACY CONTACT US HELP

The web service for automatic comparative modeling of disulphide-bonded proteins

Please provide your email address and MODELLER key, the modeling results will be sent to you after the job is done.

Your email address*:

MODELLER Key*:

Sequence Name:

Paste your sequence (length<=100 and cysteine number>=2) here* (in raw format)

HPDAKXQCRSNKANTFFVCFLAALAGLLFOLDICVIAAGALPPIADEFOITSHTOEWWWSS

Submit Clear

Required fields. To obtain MODELLER key, go to [Sali Lab](#).

Copyright 2004. National University of Singapore. Department of Biochemistry

Figure 2. Example of the SDPMOD input page.

WEB SERVICE

SDPMOD is freely accessible to academic or non-profit users via a web interface (shown in Figure 2) at <http://proline.bic.nus.edu.sg/sdpmod>. SDPMOD is primarily designed as a fully automated procedure for ease of use. However, due to the complexity of comparative modeling, human intervention and expert knowledge may be required for optimal modeling of some proteins at two critical stages, namely template selection and target–template alignment (6). To allow for human intervention, the current version of the SDPMOD server provides three modes of modeling (fully automated, semi-automated and manual) to meet the different needs of the expert users.

The ‘fully automated’ mode presents an easy-to-use interface. Users can simply submit a target sequence with their email address and their MODELLER license key, obtained from the MODELLER registration page <http://salilab.org/modeller/registration.shtml>, and the modeling will be carried out automatically according to the procedure described in Figure 1. In the ‘semi-automated’ mode, a ranked list of potential templates will be returned after the target sequence is submitted. Users can then choose the best template and adjust the target–template alignment using expert knowledge. In the ‘manual’ mode, users are allowed to propose a template from our non-redundant SDP structure dataset and modify the target–template alignment where necessary.

After the modeling process is completed, a link with the prediction results will be returned via email. Users can refer to

the link to view the prediction result and download the models. The prediction results consist of (i) a summary of the selected template(s), (ii) the predicted model based on each template in PDB format and (iii) a brief report for each modeling attempt that includes the target–template alignment used in model building, a comparison of the model against the template by means of RMSD and a PROCHECK report on the stereochemical quality of the models.

ACKNOWLEDGEMENTS

We would like to thank our colleagues at the Department of Biochemistry, National University of Singapore for their helpful comments and discussions. We are especially grateful to Professor Andrej Sali for permitting us to use MODELLER as a part of the server and Dr Ben Webb for useful suggestions. L.K. and B.L. would also like to thank the National University of Singapore for the award of Agency for Science, Technology and Research, Singapore (ASTAR) scholarships that made this work possible.

REFERENCES

- Harrison, P.M. and Sternberg, M.J. (1996) The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J. Mol. Biol.*, **264**, 603–623.
- Shen, G.S., Layer, R.T. and McCabe, R.T. (2000) Conopeptides: from deadly venoms to novel therapeutics. *Drug Discov. Today*, **5**, 98–106.

3. Richardson, M. (1977) The proteinase inhibitors of plants and micro-organisms. *Phytochemistry*, **16**, 159–169.
4. Peitsch, M.C. (1996) ProMod and Swiss-Model: internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.*, **24**, 274–279.
5. Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J. and Brunak, S. (1997) Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.*, **10**, 1241–1248.
6. Bates, P.A., Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, **Suppl 5**, 39–46.
7. Pieper, U., Eswar, N., Stuart, A.C., Ilyin, V.A. and Sali, A. (2002) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.*, **30**, 255–259.
8. Lambert, C., Leonard, N., De Bolle, X. and Depiereux, E. (2002) ESyPred3D: prediction of proteins 3D structures. *Bioinformatics*, **18**, 1250–1256.
9. Hu, S.H., Gehrmann, J., Guddat, L.W., Alewood, P.F., Craik, D.J. and Martin, J.L. (1996) The 1.1 Å crystal structure of the neuronal acetylcholine receptor antagonist, alpha-conotoxin PnIA from *Conus pennaceus*. *Structure*, **4**, 417–423.
10. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
11. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
12. Kelley, L.A., Gardner, S.P. and Sutcliffe, M.J. (1996) An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.*, **9**, 1063–1065.
13. Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
14. Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
15. Laskowski, R.A., Moss, D.S. and Thornton, J.M. (1993) Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.*, **231**, 1049–1067.

Xpro: database of eukaryotic protein-encoding genes

Vivek Gopalan¹, Tin Wee Tan¹, Bernett T. K. Lee¹ and Shoba Ranganathan^{1,2,*}

¹Department of Biochemistry and ²Department of Biological Sciences, National University of Singapore, Singapore 119260

Received August 15, 2003; Revised and Accepted September 23, 2003

ABSTRACT

Xpro is a relational database that contains all the eukaryotic protein-encoding DNA sequences contained in GenBank with associated data required for the analysis of eukaryotic gene architecture. In addition to the information found in the GenBank records, which includes properties such as sequence, position, length and description about introns, exons and protein-coding regions, Xpro provides annotations on the splice sites and intron phases. Furthermore, Xpro validates intron positions using alignment information between the record's sequence and EST sequences found in dbEST. In the process of validation, alternative splicing information is also obtained and can be found in the database. The intron-containing genes in the Xpro are also classified as experimental or predicted based on the intron position validation and specific keywords in the GenBank records that are present in predicted genes. An Entrez-like query system, which is familiar to most biologists, is provided for accessing the information present in the database system. A non-redundant set of Xpro database contents is also obtained by cross-referencing to the Swiss-Prot/TrEMBL and Pfam databases. The database currently contains information for 493 983 genes—351 918 intron-containing genes and 142 065 intron-less genes. Xpro is updated for each new GenBank release and is freely available via the internet at <http://origin.bic.nus.edu.sg/xpro>.

INTRODUCTION

Analysis of the intron/exon features in eukaryotic genes forms the basis for investigating the origin and evolution of genes (1–11). Due to the ever-increasing number of genome and EST sequencing projects, GenBank, the primary repository of nucleotide sequences, is growing at an unprecedented rate. This growth of data and the poor annotation of exon/intron details required for molecular evolution studies in the primary nucleotide database have made specialized databases that provide insight into genomic features specific to eukaryotic genes a necessity.

To facilitate such studies, we have developed Xpro, a relational database system that contains all the eukaryotic protein-coding genes from the GenBank (12) database (release 136). It provides a set of specific data such as exon sequences, the corresponding protein sequences, intron sequences, intron positions and phases, and splice sites, in a well-organized way for analyzing key features of eukaryotic gene architecture. In addition to these, each gene entry in the Xpro database is also cross-referenced to Swiss-Prot/TrEMBL (13) records and the associated Pfam (14) protein families. All these features along with the advanced query system and visual representation of gene structure make this database unique from the currently available databases on eukaryotic gene architecture (15–20).

VALIDATION OF INTRON POSITIONS

The intron positions defined in the GenBank records have to be validated before they are used for any evolutionary studies. This has to be done because GenBank includes not only intron positions derived from experiments but also those based on gene prediction programs such as GeneWise, GENSCAN, etc. Two methods are used for the validation of intron positions. The first one involves text searches of the GenBank header information for the keywords that are commonly used for gene prediction. Those records that contain these keywords are classified as predicted while the rest of the records are classified as experimental as described by Saxonov *et al.* (16). The second method of intron validation is based on the analysis of pairwise alignment of intron-containing mRNA sequences in the Xpro to experimentally derived EST sequences in the dbEST (21) database using BLAT (22). In this method, the intron positions are considered valid only if there is a high level of similarity in the aligned EST sequences in the exon–exon boundaries of the query sequence. Although this method of validation of intron positions is based on experimental EST data and thus is far more accurate than the first method, it does not cover the full data set present in GenBank due to a lack of EST information. These methods are used in tandem and thus provide a higher level of validation than that provided by current databases.

ALTERNATIVE SPLICE VARIANTS

Alternative splicing is a characteristic feature of eukaryotic organisms. It not only increases the complexity and diversity of gene products in the eukaryotes but also forms the basis for analyzing various evolutionary events such as protein domain duplication and exon shuffling. Hence, the Xpro data, which

*To whom correspondence should be addressed. Tel: +65 6874 3566; Fax: +65 6778 2466; Email: shoba@bic.nus.edu.sg

represent the gene features in the eukaryotes, are annotated for various types of alternative splicing. The alternative splicing types are assigned based on the gaps that occur in the alignment between the intron-containing GenBank mRNA sequences (only exon sequences) and the EST sequences in dbEST.

INTRON POSITION MAPPING IN PROTEIN HOMOLOGUES

The protein sequences in the Xpro database are redundant as they are derived from GenBank, which does not provide a non-redundant set of protein data. This redundancy is removed by cross-referencing the Xpro protein sequences to the Swiss-Prot/TrEMBL database, which contains only non-redundant protein sequences. This results in each Swiss-Prot/TrEMBL entry having multiple Xpro protein sequences. Single representative Xpro protein sequences from each of these clusters thus constitute a set of non-redundant protein sequences. The protein sequences in Xpro database are further classified based on Pfam, into functionally non-redundant protein families. This allows Xpro to aid in the analysis of the conservation and evolution of introns in protein homologues. A graphical display of intron positions mapped to the multiple sequence alignment of Pfam protein family sequences gives further insight to this analysis. In this way, Xpro provides a clean and validated data set targeted for the analysis of eukaryotic gene architecture.

DATABASE CONSTRUCTION AND IMPLEMENTATION

Data source

The data for the Xpro database are obtained from GenBank's invertebrate, plant, primate, rodent and mammalian divisions, which represent all the eukaryotic gene entries. The sequence, length and position data of the introns/exons were obtained by parsing the header, CDS feature and sequence fields in the GenBank records. The intron/exon features and the protein sequence from the GenBank records are extracted respectively from segment details and the translation qualifier present in the CDS field of the feature table. Intron-containing genes are specifically identified based on the keyword 'join' in the CDS field. Exon and intron sequences are derived from the DNA sequence based on the location specified in the CDS join features. In the case of intron-containing genes, if the 3' and 5' ends of the flanking exons for an intron are available in different GenBank records—as in the segmented genes—then the intron fragments are derived from DNA sequences corresponding to the reference GenBank locus entries. Further on, phases of the introns are deduced from the exon lengths and the splice sites are extracted from the surrounding exon sequences. Partial sequences are identified based on the '<' or '>' symbol in the CDS field and categorized as 5' and 3' deletions accordingly.

Xpro focuses only on the gene architecture of protein-coding regions in eukaryotes. Hence, exons and introns corresponding to 3' and 5' UTRs are not included in the database. The pseudogenes, with no protein translations (16), are also not added to the database.

Swiss-Prot/TrEMBL records contain cross-references to GenBank/EMBL protein records and Pfam records. These cross-references are extracted and stored in the relational database for use in the generation of a non-redundant subset of Xpro as described by Schisler and Palmer (17). Pfam accession number, Swiss-Prot/TrEMBL cross-references, domain ranges and multiple sequence alignment information are extracted from the Pfam flat files (version 10.0, July 2003). Each Xpro entry is linked to a Swiss-Prot/TrEMBL entry via the Swiss-Prot/TrEMBL cross-references and using the Swiss-Prot/TrEMBL cross-references extracted from Pfam entries, each Xpro entry is also linked to Pfam entries. This allows individual Xpro records to be classified into protein families and such families can be analyzed for intron distribution within the family.

Data processing

Intron position validation and the capability of intron/exon structures to form alternative splice variants are obtained by alignment of all the mRNA (exon sequences only) of the intron-containing genes in the Xpro database with the EST sequences in the dbEST database using BLAT alignment. A sequence identity cut-off of 90% and the 'fastsearch' option are chosen as BLAT alignment parameters. The BLAT outputs are then stored as tables in the Xpro database for efficient extraction and querying. The intron positions in the Xpro data sets are considered validated if they are covered by at least one EST sequence. The alternative splicing phenomena are analyzed based on gaps and percentage identity in the alignment of the mRNA sequence with EST sequences in dbEST.

Data representation

The relational model of data representation is based on the relationships between the locus, GenBank accession, protein accession and intron/exon features present in the GenBank records. The relationship between the GenBank protein accession numbers and accession numbers of the Swiss-Prot/TrEMBL, Pfam and dbEST databases are also considered when building the relational model. The database schema representing the relationship between various tables used in Xpro is available from the Xpro website. The data in Xpro are normalized to remove redundancies and to improve the query speed. The data for the tables in the Xpro database are obtained by parsing the various data sources using PERL scripts.

Implementation

Xpro is housed in a MySQL database (version 3.23.29) (23) running on a UNIX server (SGI ORIGIN 3200, Apache version 1.3). The database is freely available via the internet at <http://origin.bic.nus.edu.sg/xpro>.

WEB INTERFACE

An Entrez-like query system, familiar to biologists, is provided for efficient search of the data. Thus Xpro can be searched with valid accession numbers, GI numbers, Swiss-Prot/TrEMBL accession numbers, Pfam accession numbers, locus names or keywords from the definition field present in GenBank gene entries.



Figure 2. Graphical representation of various alternative splicing variants based on the BLAT alignment of the protein-coding region of mRNA sequence derived from the Xpro database with the EST sequences in the dbEST database. Exon sequences in the alignment are represented in different colors so that the type of alternative splicing can be identified.

appropriate links for better understanding and visualization of the gene features.

A unique feature of Xpro is the validation of exon positions based on EST alignments, available from the link marked 'EST View.' An interactive graphical interface for visualizing the BLAT alignment of available ESTs to the query sequence is provided (Fig. 2). By selecting each EST in the alignment, the pairwise alignment is also analyzed for the type of alternative splicing and the sequence alignment itself is displayed in text and graphical formats.

In addition, a BLAST (24) query page is also provided for searching an input sequence against the Xpro database. The

output of the BLAST query is displayed graphically with intron positions and alignment gaps mapped on them, allowing better visualization of intron distributions in homologous genes.

AVAILABILITY

The entire Xpro data is freely available for download as FASTA formatted flat files or mysqldump files. The database is freely available via the internet at <http://origin.bic.nus.edu.sg/xpro>.

Each FASTA formatted sequence in the flat file download is represented by the GenBank definition, protein ID, GI number and accession number in addition to specific gene-structure details such as intron positions, phases, intron-exon size and number and splice sites.

New releases of Xpro will be made available for each new release of GenBank, which is approximately once every 2 months.

XPRO DATA STATISTICS

Detailed statistics of characteristic features of eukaryotic genes in Xpro are provided in the 'statistics' link in the Xpro home page. It includes the distributions of intron lengths, number of introns, phase, splice sites, exon lengths, number of exon and coding sequence lengths for common model organisms such as *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*. These distributions are represented as graphs. Other essential features such as GC content, intron density (number of introns/kb coding sequence) and intron penetration (percentage of genes with introns) (17) are also calculated for the model organisms and are presented as tables.

APPLICATIONS

The origin and evolution of introns and their relationship to gene evolution have been analyzed based on eukaryotic gene structure and the properties of genomic elements like introns and exons (1–9). Hence, Xpro can be used as a main data source for gene evolution studies, integrating eukaryotic-specific data present in various diverse data resources.

The validated data sets, the user-friendly web interface and elaborate data statistics make Xpro a unique and valuable database system for analyzing eukaryotic genes.

ACKNOWLEDGEMENTS

The authors thank their colleagues at the Department of Biochemistry, National University of Singapore for their helpful comments, discussions and support. V.G. gratefully acknowledges the National University of Singapore for the award of a scholarship from the Agency for Science, Technology and Research, Singapore (A*STAR).

REFERENCES

- Gilbert,W. and Glynias,M. (1993) On the ancient nature of introns. *Gene*, **135**, 137–144.
- Gilbert,W. (1987) The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 901–905.
- Kriventseva,E.V. and Gelfand,M.S. (1999) Statistical analysis of the exon–intron structure of higher and lower eukaryote genes. *J. Biomol. Struct. Dyn.*, **17**, 281–288.
- Roy,S.W., Fedorov,A. and Gilbert,W. (2002) The signal of ancient introns is obscured by intron density and homolog number. *Proc. Natl Acad. Sci. USA*, **99**, 15513–15517.
- Roy,S.W., Fedorov,A. and Gilbert,W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl Acad. Sci. USA*, **100**, 7158–7162.
- Roy,S.W., Lewis,B.P., Fedorov,A. and Gilbert,W. (2001) Footprints of primordial introns on the eukaryotic genome. *Trends Genet.*, **17**, 496–501.
- Long,M., Rosenberg,C. and Gilbert,W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl Acad. Sci. USA*, **92**, 12495–12499.
- Vivek,G., Tan,T.W. and Ranganathan,S. (2003) XdomView: protein domain and exon position visualization. *Bioinformatics*, **19**, 159–160.
- Fedorova,L. and Fedorov,A. (2003) Introns in gene evolution. *Genetica*, **118**, 123–131.
- Stoltzfus,A., Spencer,D.F., Zuker,M., Logsdon,J.M., Jr and Doolittle,W.F. (1994) Testing the exon theory of genes: the evidence from protein structure. *Science*, **265**, 202–207.
- Stoltzfus,A., Spencer,D.F. and Doolittle,W.F. (1995) Methods for evaluating exon–protein correspondences. *Comput. Appl. Biosci.*, **11**, 509–515.
- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
- Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
- Bateman,A., Birney,E., Cerruti,L., Durbin,R., Ewinger,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. et al. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
- Sakharkar,M.K., Kanguane,P., Petrov,D.A., Kolaskar,A.S. and Subbiah,S. (2002) SEGE: A database on 'intron less/single exonic' genes from eukaryotes. *Bioinformatics*, **18**, 1266–1267.
- Saxonov,S., Daizadeh,I., Fedorov,A. and Gilbert,W. (2000) EID: the Exon-Intron Database—an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.
- Schisler,N.J. and Palmer,J.D. (2000) The IDB and IEDB: intron sequence and evolution databases. *Nucleic Acids Res.*, **28**, 181–184.
- Sakharkar,M., Passetti,F., de Souza,J.E., Long,M. and de Souza,S.J. (2002) ExInt: an Exon Intron database. *Nucleic Acids Res.*, **30**, 191–194.
- Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P. and Mattick,J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet.*, **24**, 340–341.
- Lopez,P.J. and Seraphin,B. (2000) YIDB: the Yeast Intron DataBase. *Nucleic Acids Res.*, **28**, 85–86.
- Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
- Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
- Dubois,P. (2003) *MySQL*. New Riders Press, Indianapolis, IN, USA.
- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.