Systematic splicing pattern analysis using splicing graph approach

Durgaprasad Bollina[1], and Shoba Ranganathan[1,2]
[1] *Department of Chemistry and Biomolecular Sciences & Biotechnology Research Institute, Macquarie University, Sydney, NSW 2109, Australia*
[2] *Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, 119260*
{dbollina@cbms.mq.edu.au, shoba.ranganathan@mq.edu.au}

Abstract

Alternative splicing is a ubiquitous mechanism that generates complexity in higher eukaryotes, generating several transcripts and proteins from a single gene. Although several alternative splicing databases exist, they do provide neither visualization nor standardized classification of alternative splicing events. We have used a splicing graph approach to represent all transcripts from a single gene pattern approach, providing compact visualization as well as automated rule-based event classification. To facilitate common subgraph identification and the ability to search through splicing databases using a user-defined pattern, we now present the deconvolution of the splicing graph and splicing events into a minimum splicing pattern set, composed of four classes. An application of the proposed algorithm for graph matching in alternative splicing graph analysis in human tumour marker genes is presented.

1. Introduction

All Alternative splicing (AS) is a ubiquitous mechanism that generates transcript diversity in higher eukaryotes. In a eukaryotic genome, a gene comprises coding regions called exons, separated by non-coding regions called introns. A single gene can therefore generate a number of unique transcripts by combining exons and introns in different ways, leading to the phenomenon of AS. The realization that alternative splicing is an important way of controlling gene regulation has spawned several large-scale efforts to create bioinformatics resources on alternate transcripts and protein isoforms [1-5]. AS databases provide information on alternative splicing on a gene-by-gene basis but they lack visual representation and the systematic classification of AS events.

A graph G = (V,E) in its basic form is composed of vertices and edges. V is the set of vertices (also called nodes or points) and E is the set of edges (also known as arcs, connections, paths or lines) of graph G. Graph vertices and edges containing information as simple labels (i.e. a name or number) are found in labelled graphs. When the vertices and edges contain additional information, called vertex and edge attributes, the resultant is an attributed graph [6-8]. This concept can be further distinguished as vertex-attributed (or weighted graphs) and edge-attributed graphs.

While dealing with graphs, some of the questions that are relevant are: How can we align them and extract common or shared segments? How can we search through a dataset, to look for a specific segment? This is a fundamental question in many areas of computational biology from matching models of objects to microarray image analysis, to searching for clusters of similar patterns in large databases and fusing information held in biological pathways.

A digraph in which the relationships among vertices are asymmetrical [9-10], is representative of gene architecture, in the biological context. These asymmetrical relationships are indicated by the arrows of the digraph.

1.1. Graph theory and bioinformatics

The earliest paper on graph theory [8] dates back to 1786 by Leonhard Euler. Euler discussed whether or not it is possible to stroll around Konigsberg (later called Kaliningrad) crossing each of its bridges across the river Pregel, exactly once and gave the conditions which are necessary to permit such a stroll. Graph theory was developed further into directed graphs in 1856 by Kirkman and Hamilton [8], who studied trips for visiting defined sites exactly once. Current examples of directed graphs include the World Wide Web, where files are the vertices or nodes and a link from one file to another is a directed edge.

Many fields such as computer vision, scene analysis, chemistry and molecular biology have applications in which images have to be processed and some regions have to be searched and identified. When this processing is performed by a computer automatically without the intervention of a human expert, a useful way of representing the knowledge is by using graphs [6]. When using graphs to represent objects or images, vertices usually represent regions (or features) of the object or images, and edges between them represent the relations between these regions. Similar graphs can be used for representing objects or general knowledge, and they can be either directed or undirected. When edges are undirected, they simply indicate the existence of a relation between two vertices. On the other hand, directed edges are used when relations between vertices are considered in an asymmetric manner. One such example is the use of splicing graphs [11-14] for the visual representation of alternative transcript diversity of a single gene.

1.2 Subgraphs and supergraphs

Definition: A graph whose vertices and edges are subsets of another graph.

Formal Definition: Let G = (V,E) be a graph. Let V1 be a subset of V and E1 be a subset of E such that G1 = (V1, E1) is a graph. Then G1 is called a subgraph of G. A graph G'=(V', E') is a subgraph of another graph G=(V, E), which is the supergraph of G'.

For a graph G, a subgraph is a graph whose vertices and edge sets are subsets of G. A supergraph of G is a graph that contains G as its subgraph[9]. In our splicing graph analysis, a splicing pattern (shown in Figure 1) is a subgraph of the supergraph of splicing events (shown in Table 1), which is in turn, a subgraph of the splicing graph. In other words, the splicing graph is the supergraph of splicing events, which are themselves supergraphs of splicing patterns. This relationship is shown in Figure 2.

1.3 Bioinformatics analysis of alternative splicing using graphs

As eukaryotic genomes are sequenced and annotated, several databases dedicated to AS are now available [11-15], leading to genome-wide computational analysis, reviewed by Lee and Wang [16]. Although AS databases give an insight into the amount of alternative splicing, they do not provide any visual representation and classification of the types of alternative splicing events occurring [17]. As the number of transcripts per gene increases, it has become increasingly difficult to identify branch points and systematically analyze and classify AS events. Directed acyclic graphs were used by Modrek and Lee [18] for EST analysis, with the genomic DNA sequence as reference. Pevzner and coworkers [14] first used de Bruijn graphs to depict the transcripts alone, without referring to the genomic DNA sequence, where the maximum common sub-sequences between transcripts were condensed into nodes and the variable regions connected by edges. Such an approach has been used to generate the Alternative Splicing Gallery (ASG) resource [11].This representation, however, does not have a biological basis and is therefore difficult to interpret in terms of biological experiments and findings.

Our approach has been to use directed acyclic splicing graphs, without a genomic DNA sequence as reference and defining exons as nodes, interconnected by introns as edges or paths through the splicing graph, representing individual transcripts. Such a schema was applied to the *Drosophila melanogaster* genome [12], to generate the DEDB data resource. Here, the first transcript served as a reference sequence to generate splicing graphs, with automatic rule-based classification of splicing events. The use of exons and introns as nodes and edges, respectively, has the intuitive advantage of biological interpretation. We have now standardized the classification of exons into distinct and variable, based on their conservation or otherwise, in a set of transcripts. Further, we developed a robust, java-based method to depict a set of transcripts as a compact splicing graph, which is available freely through the Alternative Splicing Graph server (ASGS) [19], supplemented by automatic ruled-based classification of AS events, to facilitate transcriptome analysis.

In this study, we have extended our earlier method to classify component subgraphs as splicing patterns, using which we can generate AS events as supergraphs. An ensemble of AS events constitutes the complete splicing graph. Splicing patterns are a novel deconvolution approach to the graph bases analysis of alternative splicing, enabling both pattern-based searching through AS data resources as well as detecting subgraphs across genomes for comparative transcriptome analysis.

2 Splicing patterns

The splicing graph representation provides an intuitive approach to alternative splicing pattern analysis, where gene architecture can be classified using a maximum of four novel splicing patterns, from which the nine commonly-occurring AS events can be generated.

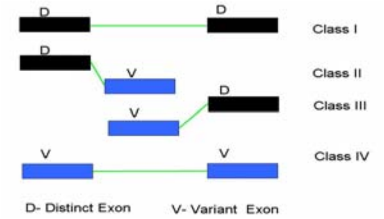

Figure. 1. Classification of inter-exonic connections as splicing patterns.

2.1 Decomposition of splicing graphs into a set of novel splicing patterns

Through the analysis of splicing graphs, we can find distinct reference and associated variant exons. The relationship of each exon to its successor is designated as a splicing pattern. The minimum number of splicing patterns required to exhaustively recreate each AS event and splicing graph is four. These inter-exonic categories are labeled as class I (Distinct-Distinct), class II (Distinct-Variant), class III (Variant-Distinct) and class IV (Variant-Variant), shown in Figure 1. Using the unique set of splicing patterns (Fig. 1), we can depict all commonly occurring alternative splicing events as combination of classes I-IV.

2.2 Algorithm for systematic splicing pattern detection

Given a set of transcripts for any eukaryotic gene, in terms of the genomic coordinates of the introns and exons of each transcript, their step-wise processing to generate the minimum set of component splicing patterns, is set out below.

1. For each transcript or isoform of a given gene, extract the genomic coordinates of exons and introns.
2. Place all exons into a new list and sort of genomic position and size. Note that all exons are represented in the standardized sense direction (+ or 5’ to 3’), even if the original transcripts are antisense (- or 3’ to 5’).
3. For each pair of overlapping exons, the one with well-determined boundaries, occurring in several transcripts, is retained as a distinct exon, while the other is classified as variable exon.
4. If an exon contains two or more exons completely, retain the smaller separate exons as distinct, with the large one classified as variable. This is because the large exon contains intronic regions and cannot be called an “exon” in the true sense.
5. Repeat steps 3 and 4 above till exons are sorted into distinct and variable, after which they are sequentially numbered.
6. Connect distinct and variant exons, using the intervening intronic regions.
7. Classify each sequential exon pair of every transcript using the splicing patterns defined in section 2.1 above.
8. Each original transcript of the gene should comprise distinct and variable exons from the list in step 2. This checking step ensures completeness of data.
9. Generate the exon table, the splicing pattern table and the splicing event table for each alternatively spliced gene.

2.3 Subgraph-supergraph relationships in AS

The splicing graph provides a new dimension to the analysis of AS, with splicing patterns as the elements of splicing events, and splicing events as components of the splicing graph. In an analytical sense, splicing patterns are subgraphs of splicing events, which are themselves subgraphs of the splicing graph. In an integrative sense, the splicing graph is the supergraph of splicing events, which are supergraphs of splicing patterns.This interplay between splicing patterns, splicing events and splicing graphs is shown in Figure. 2.

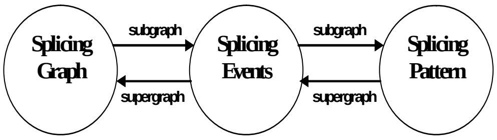

Figure. 2. Classification of inter-exonic connections as splicing patterns.

3 Comprehensive AS analysis of a tumor marker gene

The splicing graph representation provides an intuitive approach to alternative splicing pattern analysis, where gene architecture can be classified using a maximum of four novel splicing patterns, from which the nine commonly-occurring AS events can be generated.

Table 1. Analysis of splicing events in the human gene HE4 (shown in Figure 3).

Transcripts diversity as common exon-list	
Transcript no.	Component nodes (or exons)
1.	1+2+7
1.	1+2+6+3+4
3	1+3+4
4	8+4
5	5+2+3+9

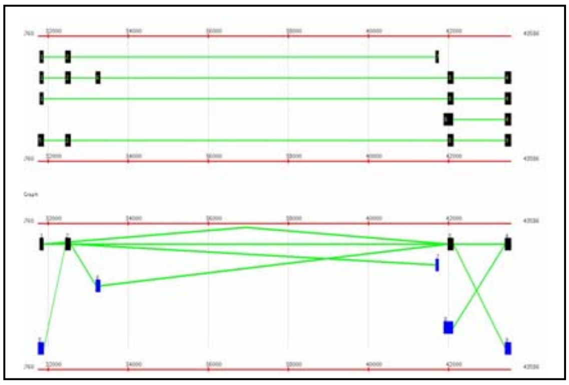

Figure. 3. Splicing graph analysis of the transctipts of HE4 gene using the splicing pattern approach. A) Transcript summary. B) Splicing Graph for HE4. Distinct exons are shown in black and varable exons in blue. All exons (nodes in the splicing graph) are numbered sequentially, starting with distinct exons and genomic positions, followed by variable exons and their genomic locations.

Table 2. Splicing pattern summary for HE4 (shown in Figure 3).

Class	Number	Nodes
Class I	6	1+2; 1+2; 3+4; 1+3; 3+4; 2+3;
Class II	3	2+7; 2+6; 3+9;
Class III	3	6+3; 8+4; 5+2;
Class IV	0	

Table 3. Splicing events for HE4, in terms of nodes.

Alternative Splicing	Nodes involved	
Alternative Transcriptional Start Site	1.	(1,5)+2
	2.	(3,8)+4
Alternative Transcriptional Termination Site	1.	2+(7;4);
	2.	3+4;
	3.	3+(4,9);
Cassette exon	1.	1+{2+6}+3

4 Discussions

In our earlier approach, the first transcript served as a reference sequence to generate splicing graphs, with automatic rule-based classification of splicing events. The use of exons and introns as nodes and edges, respectively, has the intuitive advantage of biological interpretation. Such a schema was applied to the Drosophila melanogaster genome [12], to the DEDB data resource. Further we developed java based robust method to ease of representing a set of transcripts as a compact graphic is provided by the Alternative Splicing Graph server (ASGS), a web service for generating the splicing graph[19], with automatic ruled-based classification of AS events, to facilitate transcriptome analysis. However to do the comparative analysis of same splicing graph (gene itself) or with respect other splicing graphs(gene to gene) , what we need deconvolution of splicing graph into splicing patterns.

We have enhanced our earlier method as systematic graph-pattern detection-analysis approach, to identify the splicing pattern and subgraph to enable automated splicing pattern based splicing event search across different genomes. For example, In HE4 [20-21], if we search for cassette exon, follow by splicing pattern class I, will result in to 1+{2+6}+3 and 3+4. Using this novel approach user can search for splicing pattern and splicing events in splicing graphs of different genomes.

6. Conclusions

This novel way of decomposing and representing the transcript diversity is adding visual knowledge to alternative splicing pattern analysis as splicing graph, sub-graphs and super graphs. the first is on novel sub-graph identification as splice pattern for specific splicing events like intron retention, and the second on connection subgraphs. For the latter, the goal is to find the splicing path, given a set of nodes and connections as directed graph. Nodes that capture specific relationship between them are responsible for most of the splicing path is the best ones to report as biological markers. A systematic graph-pattern detection algorithm is applied to alternative splicing analysis in human tumour marker genes is presented.

7. References

1. M. Baldonado, C.-C.K.Chang, L. Gravano, A. Paepcke,” The Stanford Digital Library Metadata Architecture. Int”. J. Digit. Libr. 1: 108–121.1997
2. K.B.Bruce, L. Cardelli, B.C. Pierce: Comparing Object Encodings. In: Abadi, M., Ito, T. (eds.),”Theoretical Aspects of Computer Software”. Lecture Notes in Computer Science, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York 415–438. 1997
3. J. van Leeuwen,(ed.)”Computer Science Today. Recent Trends and Developments ”. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York ,1995.
4. Z. Michalewicz,.”Genetic Algorithms + Data Structures = Evolution Programs”. 3rd edn. Springer-Verlag, Berlin Heidelberg New York .1996
5. B. R. Graveley, ”Alternative splicing: increasing diversity in the proteomic world”. Trends Genet., 17 (2001) 100–107.
6. M. A. Eshera, K. S. Fu,”An image understanding system using attributed symbolic representation and inexact graph-matching”. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 8, 604-619. 1986
7. T. R. Messmer, ”Efficient Graph Matching Algorithms for Preprocessed Model Graphs”. Ph.D. Thesis, Inst. of Comp. Science and Appl. Mathematics, University of Bern, 1996.
8. J.R. Ullmann, .” An Algorithm for Subgraph Isomorphism” Journal of the Association for Computing Machinery, vol. 23, 31-42,1976.
9. C. Foggia, M.V.Sansone, M.V.” A Database of Graphs for Isomorphism and Subgraph Isomorphism Benchmarking”, Proc. of the 3rd IAPR TC-15 International Workshop on Graph-based Representations, Italy, 2001.
10. X. Huang, J. Lai, S.F. Jennings,” Maximum common subgraph: some upper bound and lower bound results”. BMC Bioinformatics. 7 S6, 2006
11. J. Leipzig, P. Pevzner, and S. Heber. “The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome” Nucleic Acids Res., 32, 3977-3983, 2004
12. B. T. Lee, , T.W. Tan, S.Ranganathan,” DEDB: a database of Drosophila melanogaster exons in splicing graph form”. BMC Bioinformatics, 5, 189, 2004
13. P. Kim, N.Kim, Y. Lee, Y., B. Kim, Y. Shin, Y.,S. Lee,” ECgene: genome annotation for alternative splicing” Nucleic Acids Res., 33, D75-D79, 2005.
14. S.Stamm, J.J Riethoven, V.Le Texier, C. Gopalakrishnan, V. Kumanduri, Y. Tang, N.L, Barbosa-Morais, T.A, Thanaraj,”ASD: a bioinformatics resource on alternative splicing”. Nucleic Acids Res., 34, D46-D55, 2006.
15. D. Holste, G. Huo, V. Tung, and C.B Burge,” HOLLYWOOD: a comparative relational database of alternative splicing”. Nucleic Acids Res., 34 , D56-D62. 2006.
16. C. Lee, Q. Wang,” Bioinformatics analysis of alternative splicing”. Brief. Bioinform. 6, 23-33,2005
17. S. Heber, M. Alekseyev, S.H. Sze, H., Tang, Pevzner, P.A.: “Splicing graphs and EST assembly problem”. Bioinformatics 18:S181-8. 2002.
18. B**.** Modrek, C. Lee,” A genomic view of alternative splicing”. Nature Genet. 30,13-19.2002.
19. D. Bollina, B.T.K. Lee, T.W. Tan, ”Ranganathan, S.: ASGS: an alternative splicing graph web service”. Nucleic Acids Res. 34: W444–W447. 2006.
20. S.Ranganathan, K.J. Simpson, K.J.,D.C. Shaw, K.R. Nicholas, “The whey acidic protein family: a new signature motif and three-dimensional structure by comparative modeling”.J Mol Graph Model. 17 ,134-6. 1999.
21. R. Drapkin, R., H.H. Von Horsten, Y. Lin, S.C. Mok, C. P. Crum, W. R. Welch, J.L. Hecht,”Human epididymis protein 4 (HE4) is a secreted glycoprotein that is overexpressed by serous and endometrioid ovarian carcinomas”. Cancer Res. 2005 Mar 15;65(6):2162-9. 2005.