Weighted Alternative Splicing Graphs

Hsun-Chang Chang * Tze-Wei Huang † Po-Shun Yu ‡ Yaw-Ling Lin §¶

Abstract

Alternative splicing of a single pre-mRNA can give rise to different mRNA transcripts. Consequently, alternative splicing is an important mechanism for generating protein diversity from a single gene. Although alternative splicing is an important biological process, standard molecular biology techniques have only identified several hundred alternative splicing variants and create a bottleneck in terms of experimental validation.

In this paper, we propose methods of obtaining models of weighted alternative splicing graphs and ways of generating all alternative splicing forms from a weighted alternative splicing graph. Basically, the method uses the UniGene clusters of human Expressed Sequence Tags (ESTs) to identify alternative splicing sites. Furthermore, we propose linear time algorithms that correctly produce all possible alternative splicing variants with their corresponding probabilities. Using these methods, we infer several sets of putative alternative splicing forms; these results are then compared with methods proposed by others.

Keywords: splicing graph, alternative splicing, expressed sequence tag (EST), weighted alternative splicing graph, algorithm, EST assembly.

1 Introduction

RNA splicing is an essential, precisely regulated post-transcriptional process that occurs prior to mRNA translation. A gene is first transcribed into a pre-messenger RNA (pre-mRNA), which is a copy of the genomic DNA containing intronic regions destined to be removed during pre-mRNA processing (RNA splicing), as well as exonic sequences that are retained within the mature mRNA. Alternative splicing of eukaryotic pre-mRNAs is a mechanism for generating potentially many transcript isoforms from a single gene. It serves versatile regulatory functions in controlling major developmental decisions and fine-tuning of gene function [13]. The majority of alternative splicing events give rise to different protein products. At least 35% of human genes undergo alternative splicing [11, 10]. The physiological functions of these splice variants may be similar, opposite, or unrelated. In addition, they may differ in other properties, such as stability, tissue and cellular localization, temporal expression pattern, and responses to agonists or antagonists. The presence or level of specific splice variants may cause and/or indicate pathological or normal conditions. A class example is the Prostate Specific Anitgene, the most important maker available today for diagnosing and monitoring patients with prostate cancer [8].

Expressed sequence tags(ESTs) are single sequencing reads from cDNA clones. Even if ESTs resources are plagued by problems such as poor sequence quality and intronic contamination, they are still important tools for exon finding [3] and detection of alternative splicing [10] at the present time, biologists assemble them into consensus sequence in order to form EST contigs and analyse alternative splicing variants [3, 4, 15]. The problem of using expressed sequence tags (ESTs)

*Department of Comput. Sci. and Info. Management, Providence University, 200 Chung Chi Road, Shalu, Taichung County, Taiwan 433. e-mail: `hcchang@cs.pu.edu.tw`

†Department of Comput. Sci. and Info. Management, Providence University, 200 Chung Chi Road, Shalu, Taichung County, Taiwan 433. e-mail: `g9134012@cs.pu.edu.tw`

‡Department of Comput. Sci. and Info. Management, Providence University, 200 Chung Chi Road, Shalu, Taichung County, Taiwan 433. e-mail: `peteryu@cs.pu.edu.tw`

§Department of Comput. Sci. and Info. Management, Providence University, 200 Chung Chi Road, Shalu, Taichung County, Taiwan 433. e-mail: `yllin@pu.edu.tw`

¶The work is supported in part by the National Science Council, Taiwan, R.O.C, grant NSC-92-2213-E-126-011.

for genomic DNA annotation and prediction of exon-intron structure is not trivial, even when all splicing sites are correctly defined. One of the main difficulties is that a considerable number of ESTs map to intronic regions, or could be products of aberrant or incomplete splicing [2, 14, 15]. Moreover, the computational challenges of finding all alternative splicing variants can be understood if one considers a gene with 20 exons with one alternative splicing site per exon. In this case, the number of potential splicing variants at least 2^{20}. We can take advantage of the notion of the *splicing graph* [7] built from available EST and cDNA data. The graph conveniently encodes all potential splicing variants and shows the relationships between different transcripts implying the overall structure of transcripts. Here the idea is to abandon the linear sequence representation of each transcript and replace it with a directed graph called *splicing graph* representation where each transcript corresponds to a path in the graph. Splicing graphs may be rather complicated. As an example, the gene model of the *Drosophila Dscam* gene implies roughly $38,000$ potential transcripts [6]. The benefit of the *splicing graph* approach is that it takes into account all ESTs derived from different transcripts which cover a given position(vertex) rather than only ESTs derived from a single transcript as in the conventional approach.

In this paper, we use the idea of splicing graphs and present methods of performing quantitative analysis for all possible alternative splicing forms. Our graph-theoretical approach basically exploit all possible directed paths starting from the source of the (weighted) splicing graph, and correctly report all variants of splicing forms as well as their corresponding probabilities. The method we proposed is also closely related to the notation of Eulerian superpaths problem [16, 17, 19, 18].

2 Method

Given a directed acyclic graph $G = (V, E)$ with a source vertex $s \in V$, the vertex set V represents the state and E represents the transition probability from one state to another. For a (connected) path of G, starting from s, we use $N(p)$ to denote the set of all immediate vertices following the *last* vertex of the path p; when p is just a single vertex $v \in V$, it follows that $N(v)$ is the neighboring set of v. Let $p_1, \ldots, p_n$ be the set of all RNA transcripts for a given gene of interest. Each transcript p_i corresponds to a set of genomic positions (also called *exons*) $V_i \subset V$; note that $V_i \neq V_j$ for $i \neq j$. It follows that the set of all transcribed position $V = V_1 \cup V_2 \cup \ldots \cup V_n$ is the union of all sets V_i. The *splicing graph* G is the directed graph on the set of transcribed sites V that contains a directed edge (u, v) if and only if u and v are consecutive positions in one of the transcripts. Every transcript can be viewed as a path in the splicing graph G, and the whole graph G is the union of all paths. Each transcribed site has at least one emission probability from one transcribed site to next.

Here in our model, we assume the transcribing process is a probabilistically independent random process. It follows that all possible paths variants and their corresponding probabilities can be deduced from the random model. A higher probability in one of these paths implies a more probable putative alternative splicing forms can be produced by them, resulting a quantitative analysis of different ASF's. These possible alternative splicing forms by following steps. First, a depth first search can be performed on G to obtain the topological sorted ordering of these transcribed sites. Following the topological sorted ordering, we check whether the endpoint of a vector is sink. Here a sink vertex is one without descendants. Associated each vector with an predescendant-list to store its visited predescendants, if the vector is sink, output its corresponding visited list; otherwise, the predecessors of this vector's output-list should be appended to its children's visited list. The detailed description of our algorithm is shown in Figure 3.

Definition 1 ***weighted alternative splicing graph***
An edge-weighted directed acyclic splicing graph $G = (V, E)$ *is a* weighted alternative splicing graph *if* G *has a single starting (source) vertex* $s \in V$*, and each edge* e *of* G *is associated with a probability* $0 \leq Prob(e) \leq 1$*, such that* $\forall u \in V$*, we must have* $\sum_{v \in N(u)} Prob(uv) = 1$

Definition 2 ***weighted alternative splicing forms problem***
Given a weighted alternative splicing graph $G = (V, E)$*, the* alternative splicing forms problem *is to find all possible alternative splicing graph forms withe their associated probabilities.*

Let $\pi = \langle s, v_1, \cdots, v_k \rangle$ be a path of a weighted alternative splicing graph G starting from source s. Denote the probability associated with the path π by $Prob(\pi) = Prob(sv_1) \cdot Prob(v_1 v2) \cdots Prob(v_{k-1} v_k)$. It follows that

Lemma 1 *Given a weighted alternative splicing graph G with the source s, let $P = \{\pi_1, \pi_2, \ldots, \pi_p\}$ be all distinct paths from the source s of equal length n. It follows that $\sum_{i=1}^{p} Prob(\pi_i) = 1$ for every $n \geq 1$.*

Proof. The lemma can be easily proved by induction on the paths of length k starting from s. The condition obviously holds if $n = 1$ by definition. Assuming the property holds for length k. Now consider the case of all paths of length $k+1$. Let the set of all paths of length k are $\{a_1, \ldots, a_x\}$, and we have $Prob(a_1) + \ldots + Prob(a_x) = 1$ by inductive hypothesis. Note that the set of all paths of length $k+1$ will be $\{a_1 \circ N(a_1)\} \cup \cdots \cup \{a_k \circ N(a_x)\}$; the situation is illustrated at Figure 1. Let $q(\pi)$ denote the last vertex of a path π. By independent property, we have the sum of probabilities for paths of length $k+1$ being $\sum_{v \in N(a_1)} Prob(a_1 \circ v) + \cdots + \sum_{v \in N(a_x)} Prob(a_x \circ v) = Prob(a_1) \cdot \sum_{v \in N(a_1)} Prob(q(a_1)v) + \cdots + Prob(a_x) \cdot \sum_{v \in N(a_x)} Prob(q(a_x)v) = Prob(a_1) + \cdots + Prob(a_x) = 1$ since $\sum_{v \in N(u)} Prob(uv) = 1$ for any vertex $u \in V$ by the definition of weighted alternative splicing graphs. □

Theorem 1 *Let $U \subset V$, $U = \{u_1, u_2, \ldots, u_m\}$, be the set of all sinks within a weighted alternative splicing graph G. Let $P = \{\pi_1, \pi_2, \ldots, \pi_n\}$ be all distinct paths from the source s to sinks, then $\sum_{i=1}^{n} Prob(\pi_i) = 1$.*

Proof. Given a weighted alternative splicing graph G, without loss of generality, we can add self-loops to all its sink vertices as shown in Figure 2, and obtain an augmented graph G'. Let ℓ denote the length of the longest path of P. Note that the set of all paths of length ℓ, P', is just the set of paths P such that some shorter paths of P are lengthen by appending repeated sinks. However, by lemma 1, we have $\sum_{\pi \in P'} Prob(\pi) = 1$, it follows that $\sum_{\pi \in P} Prob(\pi) = 1$. □

Theorem 2 *Give a alternative splicing graph. We can correctly compute all possible alternative splicing forms in time linearly propositional to the size of its alternative splicing forms.*

Proof. We prove the correctness of this theorem using the theorem 1 and lemma 1, Consider the algorithm shown in Figure 3. The graph $G = (V, E)$ is represented using adjacency lists. The color of each vertex $u \in V$ is stored in the variable $color[u]$,and the predecessor of u is stored in the variable $\pi[u]$. If u has no predecessor ,the $\pi[u] =$ NIL. Each vertex v has two timestamps, the first timestamp $d[v]$ records when v is first discovered (and grayed), and the second timestamp $f[v]$ records when the search finishes examining v's adjacency list (and blackens v) . During an execution of DFS, the loops on lines 1-2 and lines 5-7 of DFS take time $\Theta(V)$,exclusive of the calls to DFS-VISIT(v). The procedure DFS-VISIT(v) is called exactly once for each vertex $v \in V$,since DFS-VISIT(v) is invoked on white vertices and the first thing it dons is paint the vertex gray. During an execution of DFS-VISIT(v),the loop on lines 3-6 is executed $\Theta(E)$ times.

Finally, the procedure ASF-FIND takes $\Theta(V+E)$ for the topological sort, and it takes time linearly proportional to the output size of the alternative splicing forms since it can be seen that each element of the output uses constant time in appending the out-lists of predecessor vectors onto the out-lists of its children vector, performed on line 11. □

3 Preliminary Experiments and Result

To validate our approach we applied it to get the human adenylosuccinate lyase (ADSL) gene [9, 21, 12]. Adenylosuccinate lyase (ADSL) is a bi-functional enzyme acting in *de novo* purine nucleotide recycling. To date, about 50 patients have been diagnosed world-wide and reports on about half of them have been published. The disease usually appears within the first months of life with neurological involcement. Our input data come from UniGene clusters of ESTs [1, 20, 5, 24]. It contains 13 exons of overall length about 2kb. In order to compare the accuracy of our approach with *splicing graphs* approach that were studied by Kmoch *et al.*(2002). We store our input data into our database and find all possible alternative splicing sites. There are two examples as following. Then we compute all possible alternative splicing forms at our bioinformatic's workstation [23].

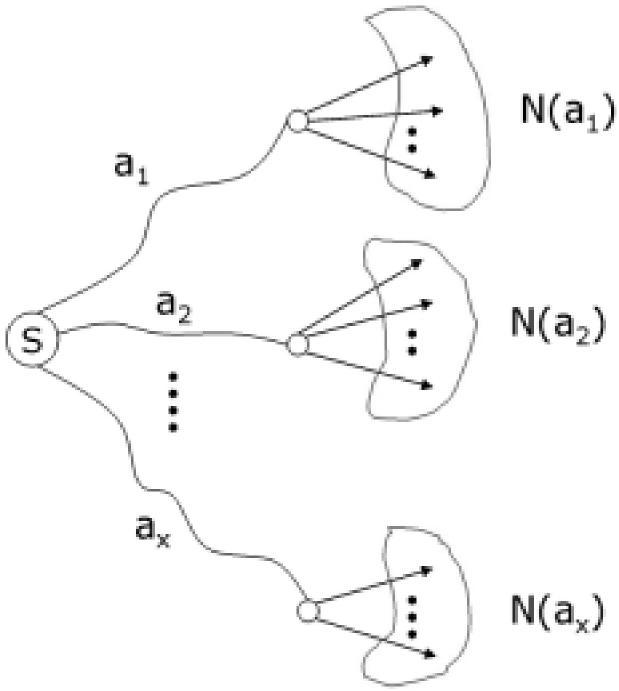

Figure 1: All possible paths of a DAG.

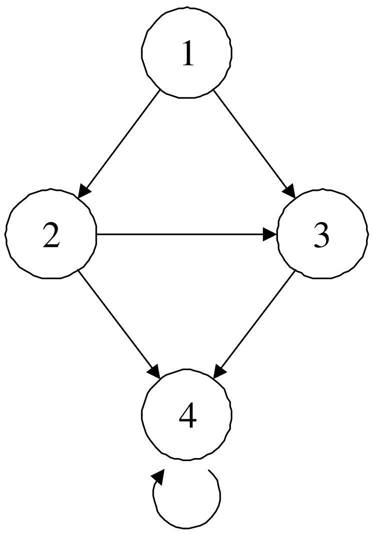

Figure 2: Adding self-loops to sinks of a DAG.

```
Input: Alternative splicing sites V = ⟨v1, v2, ..., vn⟩.
Output: All putative alternative splicing forms ASFs.
DFS(G)
for each vertex u ∈ V[G]
     do color[u] ← WHITE; π[u] ← NIL;
time ← 0;
for each u ∈ V[G]
    do if color[u] = WHITE
          then DFS-Visit(u)
DFS-Visit(u)
color[u] ← GRAY;    ▷ While vertex u has just been discovered (u, v)
d[u] ← time ← time + 1;
for each vertex v ∈ Adj[u]    ▷ Explore edge (u, v).
    do if color[v] ← WHITE
          then π[v] ← u
               DFS-Visit(u)
color[u] ← BLACK;
f[u] ← time ← time + 1;

Asf-Find(G)
call DFS(G) to compute finishing times f[v] for each vertex v.
as each vertex is finished, insert it onto the front of a linked list.
for each v ∈ V do out-list(v) ← NIL;
return the linked list of vertices.   ▷ Topological sort order (u, v).
out-list(s) ← ⟨s⟩
for each vertex u in the topological sorted ordering do
    if u is sink
        then Output out-list(u)
    else
        for each vertex v ∈ Adj[u] do   ▷ u is predecessor of v.
            Append(out-list(u) ∘ v, out-list(v))   ▷ Append u's out-list to v's out-list.
```

Figure 3: Finding all putative alternative splicing forms.

Table 1: All possible alternative splicing sites (ASS) within Adenylosuccinate lyase (ADSL) gene.

starting position of ASS	ending position of ASS	Length of ASS(bp)	Number of count
19960620	19960827	208	72
19963801	19963905	45	66
19965036	19965119	84	63
19969660	19969834	175	50
19970052	19970094	43	47
19971190	19971280	91	33
19972061	19972134	74	28
19972275	19972423	149	23
19973769	19973866	98	50
19975074	19975158	85	63
19975673	19975848	176	54
19977230	19977669	440	76

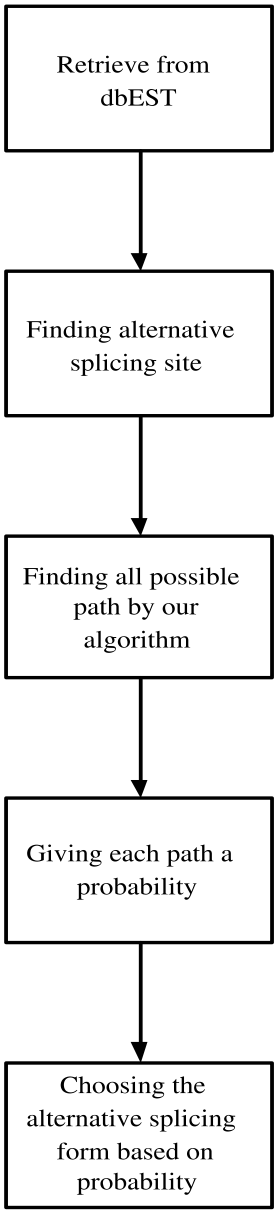

Figure 4: The flow chart for finding all possible alternative splicing forms.

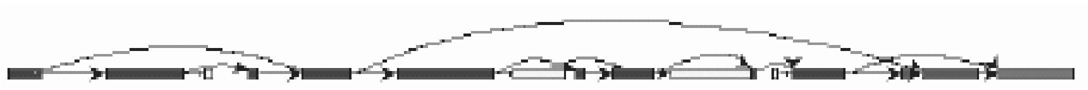

Figure 5: A splicing graph.

Table 2: Another possible alternative splicing sites (ASS).

Segment Start	Segment End	Segment length	Count of EST
58155044	58155338	294	131
58156132	58156308	176	73
58157585	58157839	254	216
58159221	58159411	190	221

4 Discussion

In contrast to other traditional methods, our approach does not need large sets of training data to construct species-specific models of genes or assemble ESTs into linear sequences, we take advantage of our algorithm and alternative splicing sites acquiring from UniGene [22] clusters of ESTs to calculate all possible paths and their probabilities. Table 1 represent all possible alternative splicing site of the Adenylosuccinate lyase (ADSL) gene. Theoretically two types of alternative splicing events might exists, one generated randomly and one generated through regulated process. Spurious events are expected to occur at lower frequencies than regulated events because biological processes have inherent error rates that are difficult to quantify and could be highly variable. In order to avoid the danger of eliminating biologically meaningful information, we will conserve all possible alternative splicing variant and its probability. In the future, we plan to implement our algorithm and calculate all probabilities of alternative splicing variant. Our final destination is providing the program for biologists on our web site [23].

Acknowledgements

The authors are sincerely grateful to Dr. Yin-Te Tasi for constructive comments and valuable suggestions. We thank Dr. Fang-Rong Hsu for providing the experimental data in their EST database. Most of all, we thank Dr. Hwan-You Chang, Dr. Hwei-Ling Peng and her PhD student Ying-Tsong Chen for biological knowledge about EST and alternative splicing.

References

[1] Mark S. Boguski and Gregory D. Schuler. Establishing a human transcript map. *Nature Genetics*, 10:369–371, 1995.

[2] D. Brett, J.Hanke, G.Lehmann, S. Haase, S.Delbruck, S. Krueger, J. Reich, and P. Bork. Est comparison indicates 38alternative splice forms. *FEBS Lett.*, 474:83–86, 2000.

[3] J. Burke, H. Wang, W. Hide, and D.B. Davison. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.*, 8:276–290, 1998.

[4] E. Coward, S.A. Haas, and M. Vingron. Splicenest: visualizing gene structure and alternative splicing based on est clusters. *Trends Genet.*, 18:53–55, 2002.

[5] et al G. D. Schuler. A gene map of the human genome. *Science*, 274:540–546, 1996.

[6] BR. Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, 17:2001, 2001.

[7] Steffen Heber, Max Alekseyev, Sing-Hoi Sze, Haixu Tang, and Pavel A. Pevzner. Splicing graphs and est assembly problem. *Bioinformatics*, pages 181–188, 2002.

[8] Nathalie Heuze, Sophie Olayat, Ninette Gutman, Marie-Louise Zani, and Yves Courty. Molecular cloning and expression of an alternative hklk3 transcript coding for a variant protein of prostate-specific antigen. *Cancer Res.*, 59:2820–2824, 1999.

[9] Jaeken J and van den Berghe G. An infantile autistic syndrome characterised by the presence of succinylpurines in body fluids. *Lancet*, ii:1058–1061, 1984.

[10] Zhengyan Kan, Eric C. Rouchka, Warren R. Gish, and David J. States. Gene structure prediction and alternative splicing analysis using genomically aligned ests. *Genome Res.*, 11:875–888, 2001.

[11] Zhengyan Kan, David States, and Warren Gish. Selecting for functional alternative splices in ests. *Genome Res.*, 12:1815–1826, 2002.

[12] Stanislav Kmoch, Hana Hartmannova, Blanka Stiburkova, Jakub Krijt, Marie Zikanova, and Ivan Sebesta. Human adenylosuccinate lyase (adsl), cloning and characterization of full-length cdna and its isoform, gene structure and molecular basis for adsl deficiency in six patients. *Hum. Mol. Genet.*, 9:1501–1513, 2002.

[13] A. J. Lopez. Alternative splicing of pre-mrna: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, 32:279–305, 1998.

[14] A.A. Mironov and J.W. Fickettand M.S. Gelfand. Frequent alternative splicing of human genes. *Genome Research*, 9:1288–1293, 1999.

[15] Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nature Genet.*, 30:13–19, 2001.

[16] Z. Mulyukov and P. A. Pevzner. Euler-pcr: finishing experiments for repeat resolution. *Proceedings of the Pacific Symposium on Biocomputing.*, pages 199–210, 2002.

[17] P. A. Pevzner and H. Tang. Fragment assembly with double-barreled data. *Bioinformatics.*, 1:S225–33, 2001.

[18] P. A. Pevzner, H. Tang, and M. S. Waterman. A new approach to fragment assembly in dna sequencing. *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology*, pages 256–267, 2001.

[19] P.A. Pevzner, H. Tang, and M.S. Waterman. An eulerian path approach to dna fragment assembly. *Proc. Natl. Acad. Sci. USA*, 98:9748–9753, 2001.

[20] Gregory D. Schuler. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med*, 75:694–698, 1997.

[21] I. Sebesta, J. Krijt, S. Kmoch, H. Hartmannova, M. Wojda, and J. Zeman. Adenylosuccinase deficiency: clinical and biochemical findings in 5 czech patients. *J. Inherit. Metab. Dis.*, 20:343–344, 1997.

[22] UniGene. `http://www.ncbi.nlm.nih.gov/UniGene/`.

[23] Providence University Bioinfo Web. `http://bioinfo.cs.pu.edu.tw/`.

[24] David L. Wheeler, Colombe Chappey, Alex E. Lash, Detlef D. Leipe, Thomas L. Madden, Gregory D. Schuler, Tatiana A. Tatusova, and Barbara A. Rapp. Database resources of the national center for biotechnology. *Nucl Acids Res*, 31:28–33, 2003.