DOI : 10.5392/JKCA.10.10.102

One-leaf One-node 트리를 이용한 선택 스플라이싱 탐지 및 예측

Detection and Prediction of Alternative Splicing with One-leaf One-node Tree

박민서
메사추세츠 대학교 컴퓨터과학과

Minseo Park(mpark@cs.uml.edu)

요약

선택 스플라이싱은 유전자 발현의 중요한 과정 중 하나이다. 선택 스플라이싱이 발생함에 따라, 돌연변가 발생하여, 질병을 일으킬 수 있다. 대부분의 선택 스플라이싱 연구는 EST(Expressed Sequence Tag)를 이용한다. 그러나, EST를 이용하여 선택 스플라싱을 예측하는 데는 몇 가지 단점이 있다. EST가 저장되어 있는 라이브러리가 잘 정돈되어 있지 않거나, 잘못 열거되어 있을 경우, 실험 시 EST를 잘못 선택할 수 있다. 또한, EST가 아직 발견되지 않은 유전 서열에서는 선택 스플라이싱을 찾을 방법이 없다. 이 논문에서는 이러한 EST 기반 연구의 약점을 개선하고, 선택 스플라이싱의 탐지 및 예측의 질을 높이기 위해서, pre-mRNA에서 One-leaf One-node Tree 알고리즘을 제안한다. 이 트리는 Arabidopsis thaliana의 각 염색체에 대해서 실험되었다. 실험 결과, 모든 염색체에서 codons에 따라 일반 스플라싱과 선택 스플라싱이 다른 패턴을 가지는 것으로 나타났다. 트리 알고리즘에서 도출된 패턴으로 부터, 아직 발견되지 않은 선택 스플라싱도 예측할 수 있다.

■ 중심어 : | 선택 스플라이싱 | One-leaf One-node 트리 알고리즘 | pre-mRNA |

Abstract

Alternative splicing is an important process in gene expression. Alternative Splicing can lead to mutations and diseases. Most studies detect alternatively spliced genes with ESTs (Expressed Sequence Tags). However, reliance on ESTs might have some weaknesses in predicting alternative splicing. ESTs have been stored in the libraries. The EST libraries are often not clearly organized and annotated. We can pick erroneous ESTs. It is also difficult to predict whether or not alternative splicing exists for those genes where ESTs are not available. To address these issues and to improve the quality of detection and prediction for alternative splicing, we propose the *One-leaf One-node Tree Algorithm* that uses pre-mRNAs. It is achieved by codons, three nucleotides, as attributes for each chromosome in Arabidopsis thaliana. The proposed decision tree shows that alternative and normal splicing have different splicing patterns according to triplet nucleotides in each chromosome. Based on the patterns, alternative splicing of unlabeled genes can also be predicted.

■ keyword : | Alternative Splicing | One-leaf One-node Tree Algorithm | pre-mRNA |

I. 서 론

유전자 서열 중에서 단백질로 코딩되지 않는 인트론 (Intron) 을 잘라 내는 과정을 스플라이싱이라고 한다. 이 스플라이싱은 Gene Expression 과정 중 pre-mRNA 에서 mRNA (mature RNA)로 변형될 때 발생된다. 따

접수번호 : #100823-012
접수일자 : 2010년 08월 23일
심사완료일 : 2010년 10월 25일
교신저자 : 박민서, e-mail : mpark@cs.uml.edu

라서, mRNA는 단백질 기능을 제공하지 못하는 Intron을 제외한 exon으로만 구성된 RNA 이다. 이 서열에 의해, 단백질 서열이 결정되고, 이에 따라, 생물학적 기능도 결정된다. 그러나, 여러 요인에 의해서, 특정 인트론이 잘라내어지지 않거나, 단백질로 코딩되는 엑손(Exon)이 특이하게 잘라질 수 있다. 이러한 과정을 선택 스플라이싱 이라고 한다. 선택 스플라이싱이 일어나면, 하나의 pre-mRNA가 여러 개의 다른 mRNA로 전이 될 수 있으며, 더 나아가 여러 종류의 단백질로 변형될 수 있다 [1-3]. 그 결과, 유전적 변이나 질병을 일으킬 수 있다[4].

선택 스플라이싱은 크게 4가지로 나눌 수 있다: 'Alternative Acceptor Site,' 'Cassette Exon,' 'Alternative Donor Site,' 'Retained Intron[7].' 이 논문은 Acceptor와 Donor sites에서 일어나는 선택 스플라이싱에 대해서 다룬다. 그 이유는 실험 대상으로 식물에서 대표적인 연구 대상인 Arabidopsis thaliana을 다루기 때문이다. 이 식물의 선택 스플라이싱 중 과반수 이상이 두 영역 (Acceptor Site와 Donor Site)에서 일어난다[5][6][10].

현재 선택 스플라이싱에 관한 연구들이 많이 진행 되고 있으나, 유전 서열정보 만을 이용해서 선택 스플라이싱을 탐지 및 예측하기는 쉽지 않다 [2][3]. 대부분의 연구가 실험 데이터인 EST(Expressed Sequence Tag) 정보에 의존한다. EST는 mRNA로 부터 형성되기 때문에, 단백질로 코딩되는 엑손(Exon)중 일부분이라고 할 수 있다. 이것을 이용해서 엑손과 인트론의 경계값을 인식할 수 있고, 그에 따라 스플라이싱 영역 또한 탐지할 수 있다. 그러나, EST는 실험데이터이기에 발현 정도에 따른 정확성 문제, EST가 존재하지 않은 영역에 대한 탐지나 아직 발견되지 않은 영역에서 발생하는 스플라이싱에 대한 이슈가 남아 있다.

이 논문은 이러한 EST사용에 따른 문제점을 개선하기 위해, EST를 직접적으로 사용하지 않는 방법을 제안 한다. EST의 실험데이터 대신 유전 서열인 pre-mRNA에 기반하여 선택 스플라이싱을 찾는 방법을 제안한다. 3개의 nucleotides(A,T,C,G) 의 묶음(Codon)을 속성으로 하는 One-leaf One-node Tree Algorithm을 제안한다. 이 알고리즘을 검증하기 위해서, TAIR(The Arabidopsis Information Resource)[8]와 TIGR (The Institute for Genomic Research)[9]에 저장 되어 있는 유전 정보와 선택 스플라이싱 정보가 이용된다. 또한 검증을 위해 메사추세츠 대학교 생물학과 실험데이터가 함께 사용된다. 검증 결과, 모든 염색체 에서 선택 스플라이싱과 일반 스플라이싱이 서로 다른 패턴을 가지는 것으로 나타났다.

2장에서는 현재 선택 스플라이싱 연구에서 주로 사용되고 있는 EST 와 Microarray 를 이용한 접근을 기술하고, 데이타 마이닝에서의 분류 방법을 기술한다. 3장에서는 이 논문에서 제안하고 있는 pre-mRNA에서 선택 스플라이싱 을 탐지 할 수 있는 One-leaf One-node Tree Algroithm에 대해서 기술한다. 4장에서는 성능 평가 및 결과를 보여주며, 5장에서는 결론 및 향후 연구에 대해서 기술한다.

II. 관련연구

1. EST에 의한 연구

대부분의 선택 스플라이싱에 관한 연구는 EST (Expressed Sequence Tag)에 의존한다[11-14]. EST는 실험 데이터의 일종으로 mRNA에서 만들어지는 작은 유전서열 조각이다[4]. 따라서, 인트론을 포함하고 있지 않다. 이 조각들을 전체 유전 서열에 매핑해 봄으로써, 유전자 중 발현되는 영역이 어디인지, 단백질로 코딩되는 영역이 어디인지를 알 수 있다. 실제로, 엑손과 인트론 의 경계를 찾는 데 유용하게 사용되고 있다.

EST를 이용한 선택 스플라이싱 연구는 2가지 과정 (Gap-patching과 Redundancy-Removing)을 통해 완성된다[1]. Gap-patching 은 유전 서열과 EST를 매칭할 때, 매칭되지 않는 오차 영역이나 잘 조합되지 않는 부분을 수정하는 과정이다. Redundancy-removing은 불필요한 EST들로 인해 중복 발생되는 스플라이싱 영역을 필터링하는 과정이다.

EST 라이브러리에 저장되어 있다. 라이브러리를 통해 EST 정보에 쉽게 접근할 수 있다[4]. 그러나, EST에

기반한 선택 스플라이싱 연구에는 한계점이 있다. EST가 실험데이터 이므로, 만약, 특정 유전 서열에서 EST가 아직 발견되지 않았 거나, 존재하지 않는다면, 그 유전 서열에서 일어나는 선택 스플라이싱은 감지할 수 없다. 또한, EST 라이브러리 안에서의 한계점이 있다. EST 라이브러리에는 상대적으로 낮은 질의 EST가 존재할 수 있고[13][15][16], 불필요 하게 중복되거나 오류를 가지는 EST들이 존재할 수 있다[16][17]. 따라서, 잘못된 정보로 인해, 잘못된 스플 라이싱정보와 선택 스플라싱정보를 구별하기가 쉽지 않다[7][17]. "Genbank" 데이타 베이스는 이런 EST의 문제를 다소 개선시켰지만, 여전히 다수의 에러와 EST 질의 문제가 이슈가 되고 있다[16].

2. Microarray에 의한 연구

Alternative Splicing 연구의 또다른 접근은 Microarray를 이용하는 방법이다. Microarray를 이용한 연구는 base-calling에 기반한다. Microarray를 이용해서 유전자 발현 정도를 탐지하는 과정은 다음과 같다[18]. 세포에서 RNA(Target) 을 추출한다. 이 추출된 Target은 증폭되어지고 형광 물질에 의해서 라벨링 되어진다. 그 후, 라벨링 되어진 Target이 Microarry위의 DNA probe와 base-pair 룰에 의해 교배되어 진다. 교배되어진 Target들은 형광색으로 변하게 된다. 마지막으로, Microarray 스캐너는 형광색을 감지해서 컴퓨터에 입력한다. 형광색으로 변한 probe의 위치가 발현 되는 유전 서열을 의미한다.

선택 스플라이싱을 발견하는데 Oligonucleotide microarray가 주로 사용된다. 가장 대표적인 oligo nucleotide는 Affymetrix Gene ChipTM 를 꼽을 수 있다[19]. Oligonucleotide microarray는 일반적인 DNA microarray와 달리, probe로 긴 DNA 서열을 사용하는 대신 다수의 짧은 oligonucleotide probe들을 사용한다. 이 array는 Whole genome을 한 개의 microarray 에서 한꺼번에 실험할 수 있는 장점이 있다. 칩에서 일어나는 신호은 발현정도를 나타 내며 그것을 통해 splice variant들를 알 수 있다[20]. 그러나, 이 방법 역시 여러 가지 한계점이 있다. 실험적 접근으로 인한 데이터 사용에 대한 한계가 있다. 실험 데이타를 모으기가 쉽지 않다. 데이터의 정확성을 높이기 위해서는 반복적인 실험이 필요하나, 비용 면에서 제약이 있다[19][20].

3. 데이터마이닝에서 분류기법 소개

데이타 마이닝의 분류기법이 생물 데이타를 분석 하는데 이용될 수 있다: 'Nearest Centroid Classifier,' 'Nearest Neighbor Classifier,"Discriminant Analysis,' 'Support Vector Machine,' '의사 결정 트리 (Decision Tree)[21].' Nearest Centroid Classifier는 각 그룹별 샘플 데이터의 중심 값들을 계산한 후, 새로운 데이타 할당 시, 그 중심 값들과 비교한 후, 가장 가까운 그룹에 할당한다. Nearest Neighbor Classifier는 샘플 데이타 그룹에서 가장 가까운 데이타를 찾은 후, 새로운 데이터를 그 데이터와 같은 그룹에 할당하는 방법이다[22]. Discriminant Analysis 는 여러 가지 속성을 가지고 최적의 함수를 만들고, 그것을 이용해서 데이터를 분류 한다[23]. Support Vector Machine은 High -dimensional Feature Space 에서 Separating Geometric Hyper plane을 이용해서 데이터를 분류하는 것이다[24]. 마지막으로, 의사결정 트리(Decision Tree)는 데이타가 더 이상 분류 되지 않을 때까지 트리 모양으로 나누어 나가는 방법이다[21][25]. Nearest Centroid Classi fier, Nearest Neighbor Classifier, Discriminant Analysis, Support Vector Machine들은 Geometric 데이터에서 효과적이다. Discriminant Analysis가 Non-geometric 데이터에서도 사용되고 있지만, 복잡한 분류 함수가 필요하다는 단점이 있다.

선택 스플라이싱 연구는 Non-geometric 데이터를 기반으로 하며, 스플라이싱의 패턴을 알기 위한 연구로, 데이터의 분류 결과 뿐만 아니라 과정이 중요한 연구이다. 따라서, 분류하는 과정을 보기에 효과적인 의사 결정트리를 응용 하여 새로운 트리(One-leaf One-node Tree)를 제안한다. 의사결정트리는 두 가지 과정으로 구성된다: 분리와 가지치기. 분리는 데이타를 분류기준(Classification Criteria)에 따라 서브 트리로 나누는 과정이다. 그 분류기준은 속성 (Attribute)에 의해서 정의되어 진다. 트리는 노드와 리프로 이루어 진다. 노드는 특정 속성에 대해 테스트하는 과정이며, 리프는 더 이

상 분리 되지 않는 데이터들의 그룹이다. [그림 1]은 의사결정 트리의 구조를 보여준다.

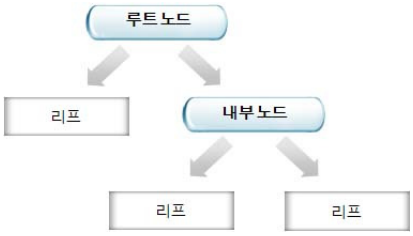

그림 1. 의사결정 트리 구조

의사결정트리는 과적합 문제 (overfitting problem)가 발생할 수 있다[26]. 문제를 해결하기 위해 트리 깊이가 불필요하게 깊어지는 것이다 [27]. [그림 2]는 과적합 문제를 보여 준다. 과적합 문제로 지나치게 많은 노드를 생길 경우, 새로운 데이터가 할당될 때, 예측 오차가 생길 수 있다. 이것은 성능 저하를 초래할 수 있다 [27]. 필요 이상으로 많은 노드가 생성된 트리는 적당한 깊이를 갖는 트리 모형으로 만들 필요가 있다. 이런 과정을 가지치기(pruning)라고 한다. 이 과정을 통해 과적합 문제의 발생을 줄여 줄 수 있다. 요약하자면, 의사결정트리는 과적합 문제가 발생할 수는 있지만, Non-geometic 데이타를 분류 하는데 효율적이다 [28][29]. 특히, 분류 결과 뿐만 아니라 과정을 시각적으로 보여 줌으로써, 데이타들 간의 관계나 패턴을 찾는데 효과적이다[21]. 실제로, Bioinformatics에서 복잡한 단백질 구조나 유전자 발현 데이터 분석[30][31]에 사용되고 있으며, 스플라이싱을 예측하는 연구에서도 Markov Modeling과 함께 이진 의사결정트리(Binary Decision Tree)가 사용되었다[12].

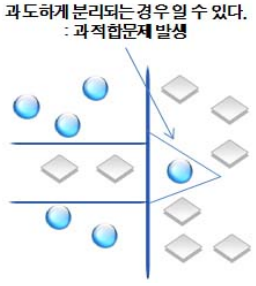

그림 2. 과적합 문제

III. pre-mRNA에서 선택 스플라이싱 탐지 및 예측 기법

이 장에서는 실험적 데이타를 직접적으로 사용하지 않고, pre-mRNA에서 유전 서열을 분석해서, 선택 스플라이싱을 탐지 및 예측하는 기법을 제안 한다. 1절은 이 논문에서 제안하는 기법의 특징에 대해서 간략히 소개 하고, 2절은 제안하고 있는 알고리즘(One-leaf One-node Tree)에 대해서 자세히 다룬다. 3절은 실험 데이타를 설명한다.

1. 제안하는 기법의 특징

이 논문에서 제안하는 기법은 EST와 같은 실험 데이타의 의존도는 낮추면서, 유전 서열 정보를 이용하여, 선택 스플라이싱를 탐지 및 예측하기 위함이다. 이 목적을 달성하기 위해,

첫째, 식물 중 Arabidopsis thaliana를 실험 종으로 선택 한다. Arabidopsis의 유전 서열은 TAIR 웹 사이트 공개 되어 있으며, 주석도 함께 제공하기 때문에 쉽게 데이타에 접근할 수 있다. 또 다른 이유는 포유류나 다른 동물들에 비해 식물에 관한 연구가 덜 진행되어 있는 편이기 때문 이다[13]. 종은 다르지만, 식물의 유전 서열 연구가 포유류나 더 나아가 사람의 유전 서열 연구에도 도움을 줄 것이라 기대한다. 둘째, 2.1과 2.2에서 언급했던 EST와 Microarray에 기반한 연구의 제약점을 피하기 위해, EST나 Microarray의 실험 데이타 대신 pre-mRNA의 서열을 기반으로 한다. 셋째, Computational 기법으로 One-leaf One-node Tree를 제안한다. 트리의 분리과정을 시각적으로 제공할 수 있는 장점(II-3 절 참조)을 응용 발전한 One-leaf One-node Tree를 제안한다.

2. One-leaf One-node Tree

One-leaf One-node Tree를 만들기 위해서는 속성(Attribute)과 기준(Criterion)이 필요 하다. 이 논문은 pre-mRNA의 스플라이싱을 선택 스플라이싱과 일반 스플라이싱으로 분리 하기 위해 codon (triplet nucleotide)을 속성으로 하는 두 가지 기준을 제안한다.

기준 1: 해당 유전 서열에 일반 스플라이싱 에만 존

재하는 codons이 존재하는가?

기준 2: 해당 유전 서열에 선택 스플라이싱 에만 존재하는 codons이 존재하는가?

이 두 개의 기준들을 순차적으로 번갈아 가면서 적용된다. 이 때, 기준 1을 먼저 적용한다. 일반 스플라이싱이 선택 스플라이싱보다 많기 때문이다. 많은 데이타가 분류될 수 있는 기준을 먼저 적용해야 됨은 규칙기반분류기법에서 이미 증명 되었다[26]. One-leaf One-node Tree는 기준이 적용 될 때마다 한 개의 리프가 만들어진다. 리프에 있는 데이타(유전 서열)는 더 이상 관찰 대상으로 다루어 지지 않는다. 따라서 한 개의 기준이 적용될 때마다, 많은 수의 관찰 데이타가 선택 스플라이싱인지 일반 스플라이싱인지 결정되어짐에 따라, 트리가 한 레벨씩 내려갈 때 마다 관찰 데이타가 줄어 들게 된다. 이 과정을 통해, 일반 의사결정트리의 불필요한 레벨 수를 줄이기 위한 방법으로 도입된 가지치기(Pruning)역할을 따로 수행하지 않고 트리 생성과정에서 함께 수행할 수 있다. 즉, 가지를 치면서 트리를 만드는 것이다. 따라서, 과적합(overfitting)을 피하 면서 데이타 분류도 할 수 있어 트리의 성능을 향상 시킬 수 있다. [그림 3]은 One-leaf One-node Tree를 보여준다.

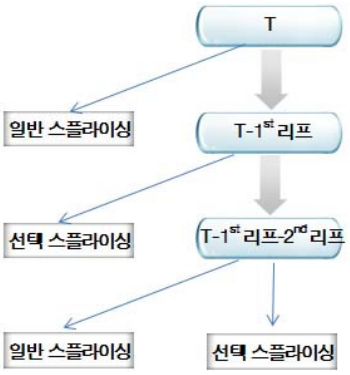

그림 3. One-leaf One-node Tree:기준 1과 기준2에 의해 일반 스플라이싱과 선택 스플라이싱으로 나누어 진다.T 는 Total Data를 의미한다. 1st 리프: 1st 일반 스플라이싱, 2nd 리프: 2nd 선택 스플라이싱이다.

서론에서 언급했듯이, 이 논문에서는 Arabidopsis thaliana의 두 선택스플라이싱 영역인 Acceptor/Donor Sites를 다룬다. One-leaf One-node Tree는 각 염색체에서 Acceptor/donor site가 선택 스플라이싱과 일반 스플라이싱으로 분류되거나 더 이상 분리가 일어나지 않을 때까지 트리를 만든다.

3. 실험 데이터 집합

Arabidopsis thaliana의 유전 서열은 이미 TAIR(The Arabidopsis thaliana Resource)[8] 웹 사이트에 공개되어 있어서 쉽게 접근할 수 있다. 공개된 유전 서열들은 One-lead One-node Tree 알고리즘의 훈련 데이타와 테스트 데이타로 사용되어 진다. Arabidopsis thaliana의 선택 스플라이싱 정보를 얻기 위해 TIGR 데이타 베이스[9]가 이용된다. TIGR 데이타 베이스는 Acceptor와 Donor Sites에서 발생하는 선택 스플라이싱 유전자를 저장하고 있다. [표 1]은 One-lead One-node Tree를 테스트하기 위해 사용 되어진 데이타 집합을 보여 준다. 3485개의 Donor Sites와 3471개의 Acceptor Sites가 테스트 되어 진다.

표 1. 실험데이타 집합

	유전자	테스트 유전자	테스트 Acceptor	테스트 Donor
염색체1	119	107	838	838
염색체2	99	86	533	533
염색체3	96	87	612	612
염색체4	94	85	557	557
염색체5	139	118	931	945

Ⅳ. 결과

1. 속성:Triplet Nucleotides(Codon)

두 개의 기준(III-2절 참조)들을 기반으로 만들어진 One-leaf One-node Tree는 Codons(Triplet Nucleotides)이 선택 스플라이싱과 일반 스플 라이싱을 분리하는 중요한 요소임을 입증시켰다. 엑손 (exon)에서는 코돈(Codons)이 선택 스플라이싱 과 일반 스플라이싱을 분류하는 중요한 요소임을 쉽게 이해할 수 있다. 모든 유전 서열이 생물학적 기능을 하기 위해서는 아미노

산으로 바뀌어야 되고, 아미노산이 단백질로 변형되어야 생물학적 기능을 만든다. 이 때, 모든 아미노산들은 코돈으로부터 만들어 지기 때문이다. 그러나 단백질로 변형되지 않는 인트론(Intron)에서의 속성을 찾기 위해서는 여러 가지 조합을 테스트 해봐야 한다. 한개, 두개, 세개, 네개 nucleotides를 속성으로 하여 트리를 만들어 보았다. 그 결과, 하나 또는 두개 nucleotides을 가지고 만든 트리는 선택 스플라이싱과 일반 스플 라이싱간의 차이점을 발견할 수 없었다. 두 스플라이싱을 분류해 낼 수 있는 서열을 발견할 수 없었다. 4개 nucleotides를 속성으로 하는 트리는 두 스플 라이싱간의 차이점은 발견할 수 있었으나, 너무 많은 속성으로 인해 스플라이싱에 대한 패턴을 찾기 어려웠다. 반면, 세 개의 nucleotides을 속성으로 한 트리는 좋은 결과 값과 함께, 스플라이싱 패턴을 찾을 수 있었다.

2. One-leaf One-node Tree의 깊이와 리프 수

이 절에는 Arabidopsis의 acceptor/donor sites에서 One-leaf One-node Tree의 결과를 보여준다. 각 염색체마다 고유의 유전 특성을 가지고 있으므로, 각 염색체마다 트리를 만든다. [표 2]에서 볼 수 있듯이 트리의 깊이가 깊지 않으며, 많지 않은 수의 리프를 갖는다. [표 2]는 각 염색체에 대한 트리 깊이(레벨)와 리프 수를 보여준다. [그림 4]는 염색체 2의 Acceptor sites에 대한 실제 결과 트리이다.

[표 2]에서 보면, 염색체 5의 acceptor sites에 대한 트리는 다소 긴 트리를 갖는다. 그 이유는 세번째와 여섯번째 레벨에서 기준 1과 기준 2의 적용만으로는 일반 스플라싱과 선택 스플라이싱을 구별할 수 없었기 때문이다. 염색체 5에서 두 스플라이싱에 대한 패턴을 찾기 위해, 좀 더 보완된 기준을 제안한다. 이 기준은 염색체 5의 세번째와 여섯번째 레벨에서만 적용된다.

표 2. 트리의 레벨과 리프 수

Site	염색체 (레벨/리프 수)				
	1	2	3	4	5
Acceptor	6/7	3/4	3/4	5/6	11/12
Donor	2/3	2/3	2/3	1/2	1/2

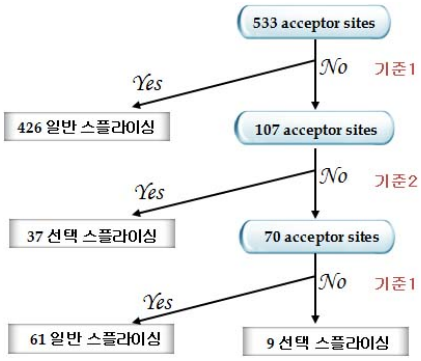

그림 4. 염색체 2의 Acceptor Sites의 One-leaf One-node Tree

1. **선택(Selection)**: pre-mRNA에서 기준 1과 기준 2를 적용한 결과, 현재까지 속성으로 사용된 codons 수가 가장 많이 발견된 지점(K0)을 선택한다. 이 접근은 consensus를 이용하는 연구에서 착안하였다[12].

2. **필터링(Filtering)**: 1단계에서 선택되어진 지점(K0)에서 가장 많이 나타나는 codon을 기준으로 acceptor sites를 필터링한다.

3. **비교(Comparison)**: 일반 스플라이싱과 선택 스플라이싱을 특정 지점 K0에 나타나는 codons에 따라 비교 분석한다.

$$h(i, k) = f(i, k) + f(i - 1, k) \quad (1)$$

$$k0 = \arg \max_{k=[intron10,exon10]} h(i, k). \quad (2)$$

f(i, k) = 레벨 i (i≥2)의 지점 k 에서 적용된 기준에서 고려된 codons수를 나타낸다.

이 발전된 기준을 적용한 결과, 염색체 5에서도 acceptor sites가 완벽하게 선택 스플라이싱과 일반 스플라이싱으로 분리되었다. 다음 절에서 자세히 서술한다.

3. 선택 스플라이싱 탐지 및 예측 결과

One-leaf One-node Tree로부터 만들어진 패턴을 평가하기 위해, TIGR 데이터베이스[9]의 선택 스플라이싱 결과들과 비교 한다. TIGR 데이타베이스는 여러 시스템들과 비교하여 정확성이 검증된 데이타베이스이다[12]. 평가와 검증은 false positive와 false negative ratio 값으로 측정한다. 이 연구에서 false positive는 선

택 스플라이싱이 일반 스플라이싱으로 잘못 인식되는 경우를 의미 하며, false negative는 일반 스플라이싱이 선택 스플라이싱으로 잘못 인식되는 경우를 의미한다. 모든 염색체에서, false positive와 false negative rates은 0으로 나타났다. 염색체 2의 donor sites의 트리에서만 false positive rate이 0.03으로 나타났다.

이 논문에서 제안된 One-leaf One-node Tree를 이용하면, EST가 발견되지 않은 유전자나, 다른 데이타베이스에서 아직 발견되지 않은 선택 스플라이싱도 예측할 수 있다. 더 나아가, One-leaf One-node Tree로 부터 도출된 패턴으로부터 선택 스플라이싱의 원인이 되는 지점과 codons도 함께 예측할 수 있다. 메사추세츠 대학교 생물학과 실험실의 데이타(유전자 At1g27450와 At1g30460)를 사용해서 예측에 관한 검증을 한다([그림 5] 참조).

One-leaf One-node Tree로부터 만들어진 패턴을 통해 테스트한 결과, 유전자 At1g27450 의 acceptor sites (엑손과 인트론의 경계에서 양쪽으로 30 nucleotides까지) 중 네번째 엑손 에서 선택 스플라이싱이 탐지되었다. 엑손과 인트론의 경계로 부터 왼쪽으로 10~12 떨어진 지점의 'tgg' 가 선택 스플라이싱의 발생 요인이 된다. 또한 2번째 엑손에서 왼쪽으로 28~30 떨어진 지점의 'aat'가 선택 스플라이싱을 일으키는 것으로 나타났다([그림 5] 참조). 그러나 실제 실험에서는 2번째 엑손은 일반 스플라이싱으로 나타났다. 비록, 2번째 엑손에서 실험과는 다른 결과를 보였지만, 'aat'가 엑손에서 멀리 떨어진 (28~30)지점에서 탐지되었기 때문에, 예측의 정확도가 떨어질 수 있다. 나머지 acceptor /donor

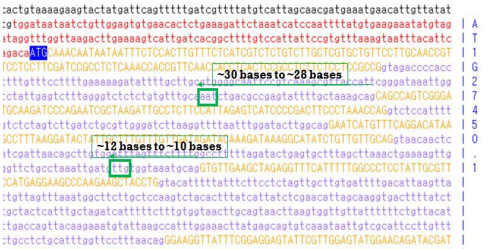

그림 5. 유전자 At1g27450 : 노랑색이 엑손을 의미하며, 보라색이 인트론을 의미한다.

sites들에서의 테스트는 실험과 같은 결과를 보였다. 유전자 At1g30460의 acceptor sites들은 트리를 이용한 예측과 물리적 실험 결과 모두에서 선택 스플라이싱을 발생시키지 않았으며, Donor sites 에서는 트리를 이용한 예측과 실험 결과가 일치했다.

V. 결론 및 향후 연구

TIGR 데이타베이스의 연구 결과와 실제 실험데이터와의 비교로 부터 이 논문에서 제안된 One-leaf One-node Tree가 Arabidopsis thaliana에서 선택 스플라이싱을 탐지 및 예측 하는데 효과적임을 보여 주었다. 기존의 다른 시스템과 비교하여 볼때, 이 트리의 장점은 EST 등과 같은 실험 데이타를 직접 이용하지 않고, pre-mRNA에서 유전 서열을 직접 이용하여 선택 스플라이싱을 탐지 및 예측 할 수 있다는 것이다. 그 결과 실험 데이타로부터 발생할 수 있는 에러를 피할 수 있었다. 두 번째 장점은, 선택 스플라이싱 탐지 영역에 대한 제한이 없다. EST가 없는 영역을 비롯, 아직 발견되지 않은 영역 에서도 선택 스플라이싱을 탐지 및 예측할 수 있다. 또한, One-leaf One-node Tree 알고리즘을 이용함으로써, pre-mRNA에서 선택 스플라이싱과 일반 스플라이싱을 복잡한 계산 과정 없이 쉽게 분리할 수 있다. 마지막으로, 가장 큰 장점은 아직 다른 시스템들에서 레이블 되지 않았거나 발견되지 않은 유전자 서열에서도 선택 스플라이싱을 발견할 수 있다는 것이다.

향후, 이 기법은 acceptor/donor sites뿐만 아니라 다른 선택 스플라이싱 패턴의 탐지 및 예측에도 적용될 수 있을 것으로 기대된다. 각 종마다 그들 나름의 고유의 특징을 갖고 있지만, 많은 공통된 서열이 존재한다. 따라서, 식물뿐만 아니라 동물, 휴먼 유전자 분석에서 이 기법이 적용될 수 있을 것으로 기대 된다. 특히, 이 기법은 질병을 다루고 예측하는데 적용될 수 있을 것으로 본다. 환자와 일반인의 유전 서열을 비교함으로써, 침해, 당뇨병, 암등 유전적 요인에 의해서 발생할 수 있는 질병의 패턴을 밝힐 수 있을 것으로 기대된다.

참 고 문 헌

[1] T. Chuang, F. Chen, and M. Chou, "A compareative method for identification of gene structures and alternatively spliced variant," Bioinformatics, Vol.20, pp.3064-3079, 2004.

[2] R. Sorek, R. Shemesh, Y. Cohen, O. Basechess, G. Ast, and R. Shamir, "A Non-EST-Based Method for Exon-Skipping Prediction," Genome Research, Vol.14, pp.1617-1623, 2004.

[3] S. Stamm, J. Riethoven, V. Le Texier, C. Gopalakrishnan, V. Kumanduri, Y. Tang, N. Barbosa-Morais, and T. Thanaraj, "ASD: a bioinformatics resource on alternative splicing," Nucleic Acids Research, Vol.34, pp.D46 - D55, 2006.

[4] http://www.ncbi.nlm.nih.gov.

[5] B. Haas, A. Delcher, S. Mount, J. Wortman, R. Smith Jr, L. Hannick, R. Maiti, C. Ronning, D. Rusch, C. Town, S. Salzberg, and O. White, "Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies," Nucleic Acids Research, Vol.31, pp.5654-5666, 2003.

[6] M. Campbell, B. Haas, J. Hamilton, S. Mount, and C. Buell, "Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis," BMC Genomics, Vol.7, p.327, 2006.

[7] R. Nurtdinov, I. Artamonova, A. Mironov, and M. Gelfand, "Low conservation of alternative splicing patterns in the human and mouse genomes," Human Molecular Genetic, Vol.12, pp.1313-1320, 2003.

[8] http://www.arabidopsis.org.

[9] http://www.tigr.org

[10] D. Black, "Mechanisms of alternative pre-messenger RNA splicing," Annual Review of Biochemistry, Vol.72, pp.291-336, 2003.

[11] K. Iida, M. Seki, T. Sakurai, M. Satou, K. Akiyama, T. Toyoda, A. Konagaya, and K. Shinozaki, "Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis Thaliana based on full-length cDNA sequences," Nucleic Acids Re-search, Vol.32, pp.5096-5103, 2004.

[12] M. Pertea, X. Lin, and S. Salzberg, "GeneSplicer: a new computational method for splice site prediction," Nucleic Acids Research, Vol.29, pp.1185-1190, 2001.

[13] B. Wang and V. Brendel, "Genomewide comparative analysis of alternative splicing in plants," in Proceedings of the National Academy of Science of the United States of America, pp.7175-7180, 2006.

[14] W. Zhu, S. Schlueter, and V. Brendel, "Refined annotation of the Arabidopsis Thaliana genome by complete EST mapping," Plant Physiology, Vol.132, pp.469-484, 2003.

[15] C. Iseli, V. Jongeneel, and P. Bucher, "ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences," in Proceedings of the Seventh ISMB, pp.138-148, 1999.

[16] C. Jongeneel, "Searching the expressed sequence tag (EST) databases: panning for genes," Briefings in Bioinformatics, Vol.1, pp.76-92, 2000.

[17] J. Collins, M. Goward, C. Cole, L. Smink, E. Huckle, S. Knowles, J. M. Bye, D. Beare, and I. Dunham, "Reevaluating human gene annotation: a second-generation analysis of chromosome 22," Genome Research, Vol.13, pp.27-36, 2003.

[18] D. Raghunandan, L. Guglielmo, D. K., and A. Animesh, "Clinical applications of DNA microarray analysis," Journal of Experimental Therapeutics and Oncology, Vol.3, pp.297-304,

2003.

[19] S. Mehta, "DNA Microarrays in Health Care & Drug Discovery," http://plasticdog.cheme.columbia.edu/.

[20] G. Hu, S. Madore, B. Moldever, T. Jatkoe, D. Balaban, J. Thomas, and Y. Want, "Predicting Splice Variant from DNA Chip Expression Data," Genome Research, Vol.11, pp.1237-1245, 2001.

[21] E. Garrett-Mayer and G. Parmigiani, "Clustering and Classification Methods for Gene Expression Data Analysis," Johns Hopkins University, Dept. of Biostatistics Working Papers, Vol.70, 2004.

[22] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," in Proceedings of IEEE Transaction on Information Theory, pp.21-27, 1967.

[23] R. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, Vol.7, pp.178-188, 1936.

[24] V. Vapnik, Statistical Learning Theory. New York, NY: John Wiley & Sons, 1998.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[26] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations*. Academic Press, 2000.

[27] A. Nabhan and A. Rafea, "Tuning statistical machine translation parameters using perplexity," in Proceedings of the 2005 IEEE International Con-ference on Information Reuse and Integration, pp.338-343, 2005.

[28] E. Brand and R. Gerritsen, "Decision Trees," DBMS Online, 1988, http://www.dbmsmag.com/-9807m05.html.

[29] K. Delisle, "Decision Trees and Evolutionary Programming," Artificial Intelligence Depot., Tech. Report, http://aidepot.com/Tutorial/DecisionTrees .html.

[30] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic DNA," Journal of Molecular Biology, Vol.268, pp.78-94, 1997.

[31] H. Zhang and C. Yu, "Tree-based analysis of microarray data for classifying breast cancer," Fron-tiers in Bioscience, Vol.7, pp.C63-C67, 2002.

저 자 소 개

박 민 서(Minseo Park) 정회원

- 2003년 2월 : 연세대학교 컴퓨터과학과(공학석사)
- 2009년 10월 : 메사추세츠 대학교 컴퓨터과학과(이학박사. Ph.D.)
- 2010년 1월 ~ 현재 : Samsung SDS, Helathcare Service Center of Excellence, Senior Researcher.
- 2010년 9월 ~ 현재 : 성균관대학교 융합의과학원 초빙 연구자

<관심분야> : 바이오인포메틱스, 유전자 시퀀스 분석 알고리즘, 선택 스플라이싱, 사용자 인터페이스. 유전자 브라우저