다중 전사체 서열의 시각화와 불리언 질의를 이용한 비교

박태원*, 조환규*, 이도훈*
*부산대학교 컴퓨터공학과
e-mail: {darkptw, hgcho, dhlee}@pusan.ac.kr

Visualization of Multiple Transcript Sequences and Comparison using Boolean Query

TaeWon Park*, Hwan-Gue Cho*, DoHoon Lee*
*Dept of Computer Engineering, Pusan National University

요 약

생물정보학 데이터를 분석하는 과정에서 서열 데이터의 시각화는 연구자에게 방대한 서열 데이터의 특성을 눈으로 쉽게 이해하기 위한 필수 과정이다. 대조 실험 데이터나 다중 서열 데이터를 시각화해 주는 많은 도구들이 있지만 방대한 유전체 서열에서 사용자가 원하는 다중 데이터간의 비교 영역을 찾아서 시각화해주는 기능이 부족한 것이 현 상황이다. 본 논문은 불리언 질의를 통해서 다중 전사체 서열을 효율적으로 비교하고 그 결과를 시각화해주는 방법을 제안한다.

1. 서론

시각화는 주어진 문제에 내재되어 있는 특성이나 관계를 규명하는데 직관적 통찰력을 가져다주는 방법이다. 생물정보학 분야에서 방대한 서열정보를 분석하고 서열간의 관계, 특성을 비교하기 위한 방법으로 시각화에 대한 연구가 꾸준히 진행되었고 현재에도 그에 대한 다양한 연구결과가 발표되고 있다. RNA-Seq 연구가 DNA microarray[1]연구를 대체할 만큼 급속도로 확산되면서 서열 분석 및 시각화 도구들이 RNA-Seq 연구에 맞추어 응용되거나 새로운 방법들이 연구되고 있다. 단편서열 매핑 단계, RNA 전사체 조립 단계 등의 결과를 보기 위해서 시각화 도구가 사용되고 있다.

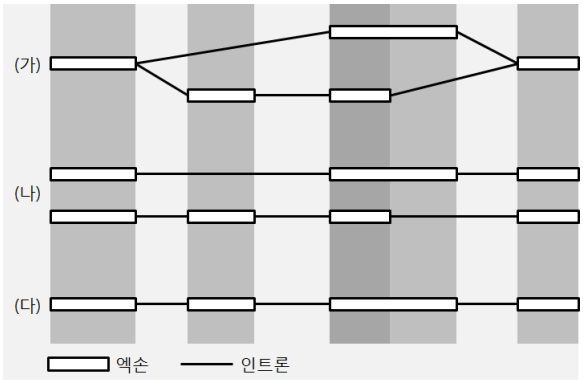
대표적으로 널리 사용되는 생물정보학 서열 데이터 시각화 도구 중에는 ENSEMBL browser[1]와 UCSC genome browser[2]가 있다. 이 두 도구는 웹 기반의 시각화 도구로 알려진 서열 데이터베이스나 사용자가 업로드한 서열 데이터를 웹브라우저에서 보여주는 기능을 가진다. Savant genome browser[3]는 로컬 응용 프로그램으로 사용자가 시각화하기를 원하는 서열 데이터를 인덱싱하는 방식을 통해 웹 기반의 시각화 도구에 비해서 빠르게 원하는 영역을 보여주는 것이 특징이다.

이러한 기존의 도구들은 다중 서열 데이터를 트랙별로 시각화 해주어 한 화면에 대조 실험 데이터를 보는 것이 가능하다. 하지만 다중 서열 위치 정보를 직접적으로 비교할 수 있는 효율적인 방법을 제공해주고 있지 않아 사용자가 직접 시각화 영역을 이동시키면서 서열을 비교해야 하는 한계점을 가지고 있다. 본 논문이 제안하는 불리언 질의를 이용한 시각화 및 비교 방법은 새로운 유전자(novel gene)나 차이가 보이는 서열 구간에 대한 정보를 빠르게 파악할 뿐 아니라 여러 도구들 간의 결과 차이, 질병에 대한 대조 실험 결과간의 차이를 보다 쉽게 알 수 있는 통찰력 있는 정보를 제공한다.

본 연구는 질병관리본부 학술연구용역과제 (2012-E72006-00) 연구비를 지원받아 수행되었습니다.

2. 시각화 및 비교 도구 설계

전사체 서열은 일반적으로 시작과 끝으로 구성된 단순한 구간 서열과는 다른 몇 가지 특성을 가지고 있다. 첫째, 전사체 서열은 RNA 염기서열로써 단백질로 코딩되는 엑손 구간과 코딩되지 않는 인트론 구간이 번갈아가면서 이어지는 구조를 가진다. 둘째, 동일한 유전자에서도 선택접합으로 인해 다양한 종류의 전사체가 하나의 집합으로 발현된다. 이와 같은 전사체 서열의 특성으로 인해 효과적으로 전사체 집합을 시각화할 수 있는 방법이 필요하다.



(그림 1) 전사체 집합을 시각화하는 3가지 방식 (가) 그래프 구조 방식 (나) 개별 전사체 나열 방식 (다) 전사체 집합 병합 방식

SpliceGrapher[4]는 (그림 1)의 (가)와 같이 하나의 유전자에서 발현되는 전사체 집합을 단일 그래프 구조로 만들어 보여주는 방식을 사용한다. 반면에 UCSC Genome Browser는 (다)와 같이 전사체 집합을 병합하여 하나의 전사체로 보여주다가 사용자가 자세하게 보기를 요청하면 (나)와 같이 각각의 전사체를 나열하여 보여주는 방식을 취한다. 전사체 집합을 나열하여 모두 보여주는 방식이나 그래프 구조를 보여주는 방식은 단일 전사체 서열을 자세히 보여주기에 좋은 방법이다. 하지만 두 방식은 다중 전사체 서열을 한꺼번에 보여주기에는 복잡하여 효과적인 시각화에 방해가 되므로 (다)와 같이 전사체 집합을 병합하여 보여주는 방식이 사용자가 한 눈에 다중 데이터를 비교해서 보는데 유리함으로 (다) 방식으로 전사체 집합을 시각화하도록 구현하였다.

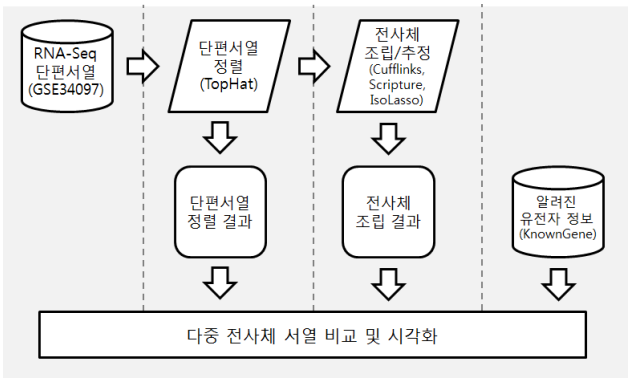
<표 1> 다중 전사체 서열 비교 시 불리언 연산별 생물정보학적인 의미

불리언 연산	전사체 서열 비교 시 의미
AND	공통적인 발현 구간
OR	서열 병합
XOR	한 서열에서만 발현되는 구간
NOT	발현되지 않은 구간

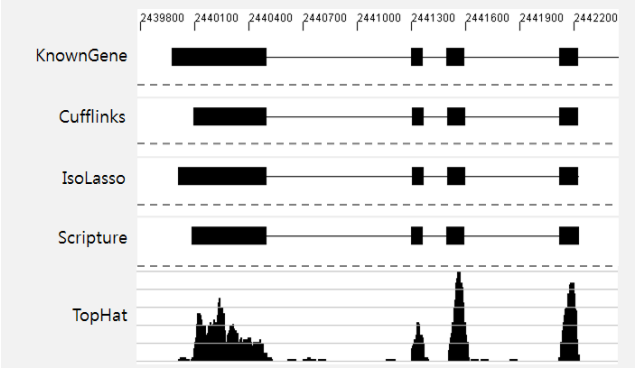
다중 전사체 서열을 비교하는 효과적인 방법으로 불리언 질의를 사용해서 서열간의 비교 기능을 구현했다. 불리언 질의는 정보 검색 시스템이나 인터넷 검색 엔진에서도 많이 사용되는 방식으로 AND, OR, XOR, NOT 연산을 사용해 검색어와 관련된 정보를 찾을 수 있게 해준다. 다중 서열 데이터에서 위치 정보를 활용해 특정 데이터에서만 나타나는 독특한 발현 구간을 찾는 경우와 같이 불리언 질의는 사용자의 검색의도를 정확히 표현할 수 있는 장점을 가진다. <표 1>과 같이 4가지 불리언 연산을 기본 연산이라고 하면 기본 연산을 조합해서 사용함으로써 사용자는 다중 전사체 서열 데이터에서 복잡한 조건을 만족하는 영역을 정확히 질의할 수 있다.

3. 실험 및 토의

실제 RNA-Seq 데이터를 통한 실험을 위해서 (그림 2)와 같이 GEO(Gene Expression Omnibus)에 공개된 RNA-Seq 단편서열 데이터(GSE34097)를 사용해 전사체를 조립했다. 전사체 조립 중간 과정인 단편서열 정렬을 위한 도구로는 TopHat[5]을 사용했다. TopHat에서 나온 정렬 결과를 3가지 조립 도구(Cufflinks[6], Scripture[7], IsoLasso[8])를 사용해 전사체의 구조를 조립/추정했다. 효과적인 시각화를 위해 (그림 3)과 같이 알려진 유전자 서열(UCSC KnownGene) 데이터베이스와 단편서열 정렬 결과를 다중 전사체 서열과 함께 보여주도록 했다.

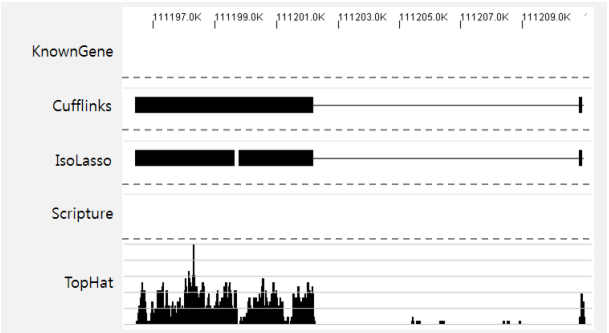


(그림 2) RNA-Seq 전사체 서열 시각화 및 비교를 위한 실험 과정



(그림 3) 알려진 유전자 정보와 3가지 종류의 전사체 조립 도구의 결과 및 단편서열 정렬 결과를 보여주는 화면

불리언 질의를 활용하여 3가지 도구의 결과를 병합한 서열 중에 기존에 알려져 있지 않은 유전자에서 발현된 전사체를 질의하기 위해서 “NOT(KnownGene) AND (Cufflinks OR IsoLasso OR Scripture)”로 불리언 질의가 가능하다. 질의의 결과는 많지만 구간의 길이가 가장 긴 영역을 시각화하면 (그림 4)로 나타난다. Cufflinks와 IsoLasso가 추정한 전사체가 참조 유전체의 1번 염색체 상의 111,196,432-111,210,968 영역에서 발현되는 것을 확인할 수 있다.



(그림 4) 불리언 질의 “NOT(KnownGene) AND (Cufflinks OR IsoLasso OR Scripture)”의 결과를 보여주는 화면

4. 결론

기존 시각화 도구들이 단순히 전사체 서열을 보여주며 사용자가 직접 눈으로 다중 서열을 비교하는 불편한 과정을 줄이기 위해서 다중 전사체 서열을 간단하게 시각화하고 불리언 질의를 사용해 사용자가 원하는 방식으로 서열을 비교하여 시각화할 수 있는 도구를 구현하였다.

제안하는 방식은 전사체 서열의 위치 정보를 기반으로 시각화 및 비교를 구현한 방식이다. 향후에 전사체 서열의 위치 정보뿐만 아니라 각각의 발현량을 고려하여 보다 개선된 시각화 및 비교를 구현하는 문제와 질의 결과의 순위를 매기는 방식에 대한 연구가 필요하다.

참고문헌

[1] T. Hubbard, et al. "Ensembl 2005" Nucl. Acids Res. doi:10.1093/nar/gki138

[2] W. James Kent, et al. "The Human Genome Browser at UCSC" Genome Res. doi:10.1101/gr.229102

[3] Marc Fiume, et al. "Savant: Genome Browser for High Throughput Sequencing Data" Bioinformatics. doi:10.1093/bioinformatics/btq332

[4] Mark F Rogers, et al. "SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data" Genome Biology. doi:10.1186/gb-2012-13-1-r4

[5] Trapnell C, et al. "TopHat: discovering splice junctions with RNA-Seq". Bioinformatics. doi:10.1093/bioinformatics/btp120

[6] Trapnell C, et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation" Nature Biotechnology. doi:10.1038/nbt.1621

[7] Mitchell Guttman, et al. "Ab initio reconstruction of cell type - specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs" Nature Biotechnology. doi:10.1038/nbt.1633

[8] Wei Li, et al. "IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly" Journal of Computational Biology. doi:10.1089/cmb.2011.0171.