전사체 시각화 프레임워크 개발

황혜련, 김소라, 조환규
부산대학교 컴퓨터공학과
e-mail:haelyen@pusan.ac.kr

Transcriptome visualization framework development

HyeRyeon Hwang, Sora Kim, Hwan-Gue Cho
Dept of Computer Engineering, Pusan National University

요 약

정보의 시각화는 추상적 정보를 직관적으로 이해하기 쉽도록 시각적으로 명확하게 표현하는 방법을 말한다. 대용량의 바이오 데이터를 다루는 생물정보학(bioinformatics) 분야에서는 컴퓨터의 높은 성능을 활용하여 수많은 유전학적 데이터들을 분석하고 있다. 다양한 생물정보학 실험에서 전사체는 특정한 조건에서 발현된 RNA의 총합을 말한다. 분석된 전사체 정보는 텍스트형태로 제공이 되는데 이를 사용자가 수작업으로 비교하는 데에는 한계가 있다. 따라서 분석된 전사체 정보를 효과적으로 인지할 수 있도록 시각화하는 연구들이 진행되고 있다. 본 논문에서는 그래프 라이브러리인 yFile을 활용하여 추정된 전사체를 실시간으로 시각화하여 제공하는 방법을 제안한다. GTF파일을 입력받아서 데이터베이스에 저장하고 이 정보를 이용하여 그래프를 생성한다. 실험 결과는 전사체를 시각화 하는 방법을 통하여 다양한 전사체 정보를 알아 낼 수 있고, 최종적으로는 novel gene을 찾는 것이 가능할 것으로 기대한다.

1. 서론

차세대 시퀀싱(Next Generation Sequencing)기술은 DNA를 읽어 들여 유전체서열상의 변이를 밝혀내는 방법이다[1,2]. A, G, T, C와 같은 염기서열의 조각을 빠르고 정확하게 읽어냄으로써 유전자의 발현과 유전체 해독에 이용되고 있다. 현재 기술의 발달로 데이터의 처리 비용과 시간을 효과적으로 감소할 수 있게 됨으로써 저렴한 비용으로 높은 처리량을 내는 것이 가능해졌고, 방대한 양의 시퀀싱 데이터들을 처리하고 분석하는 프로그램에 대한 연구도 계속 되고 있다. 따라서 복잡한 생물학적 실험이나 대용량의 바이오 데이터를 다루는 생물정보학(bioinformatics) 분야 에서는 많은 유전정보의 유전학적 데이터들의 결과를 분석하는데 있어서 사용자들이 한눈에 인지하기 쉽게 할 수 있도록 도와주는 시각화 기능이 더욱 중요해지고 있다.

정보의 시각화는 추상적 정보를 직관적으로 이해하기 쉽도록 시각적으로 명확하게 표현하는 방법을 말한다. 생물정보학에서는 분석된 전사체 정보는 텍스트형태로 제공이 되는데 이를 사용자가 수작업으로 비교하는 데에는 한계가 있다. 따라서 분석된 전사체 정보를 효과적으로 인지할 수 있도록 시각화하는 연구들이 진행되고 있다. 본 논문에서는 사용자들이 좀 더 쉽게 데이터를 분석할 수 있고, 시각화 정보를 통해 더 많은 정보를 얻을 수 있도록 그래프로 가시화하고자 한다. GTF파일을 입력받아서 데이터베이스에 저장하고 이 정보를 YFile[3]을 이용하여 시각화해서 그래프를 생성한다. 또한 이것을 기반으로 시각화 도구를 개발하는데 있어서 어떠한 기능들이 적합하고 그러한 장점들을 잘 이용할 수 있을 것인지에 대하여 논의하고자 한다.

2. 관련연구

SpliceGrapher[4]는 실험하는 종의 기존에 존재하고 있는 gene model을 기본으로 사용하여 predicted splice graph를 그려주는 도구이다[3]. 다른 기존의 도구들과 달리 exon, intron, CDS 등의 정보가 들어있는 유전자 모델을 사용함으로써 처음부터 모델을 설계할 필요 없이 SpliceGrapher의 기본적인 뼈대를 잡고 시작할 수 있기 때문에 다른 도구에 비해 계산할 것이 적어지고 정확도의 측면에서 장점을 가진다. 최종적으로는 총 10개의 출력 파일이 생성되며 추정된 전사체를 *.GFF(General Feature Format)형식으로 나타낸다. 그렇게 출력된 파일 중 *_predicted.GFF 파일은 도구 내의 가시화 기능을 이용해 전사체를 표현해주고 *.PDF 파일로 출력된

다. 그렇지만 단순히 SpliceGrapher의 출력 결과만 보았을 때 사용자는 전사체가 어디서 발현 되었고, 어디서 나온 결과 파일을 사용했는지 등의 자세한 세부정보까지 알 수 없다. 또한 자체 도구에서 나온 결과가 아니라면 출력결과를 시각화 하는 것이 쉽지 않다. 자체 도구에서 나온 출력 결과를 시각화 한다 하더라도 *.EST 파일도 사용해야 하고 시각화하기 위한 세부적인 옵션 설정을 따로 해주어야 하기 때문에 번거롭다. 현재 개발 중인 도구는 SpliceGrapher와 같은 시각화 도구이지만 좀 더 사용자측면에서 도움이 될 수 있고 많은 정보들을 얻을 수 있도록 하는 사용자에게 적합한 도구이다. 특히 데이터가 미리 DB에 저장되어 있기 때문에 사용자가 원하는 종과 gene을 선택한다면 번거로운 과정을 거치지 않고도 짧은 시간 내에 시각화 결과를 얻을 수 있다.

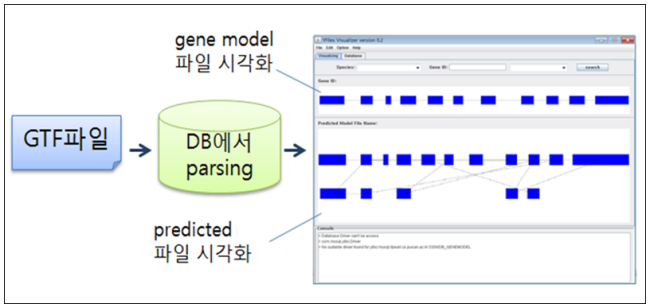
3. YFile을 이용한 그래프 가시화

yFiles는 독일의 yWorks에서 개발한 extensive Java class library로써 graph, diagram, network등을 그려주고 분석, 가시화하는데 필요한 컴포넌트와 알고리즘을 제공한다[5].

입력된 데이터를 이용해 그래프를 자동적 지정해 주는 기능을 가지고 있다. 그래프를 그릴 때 노드의 위치를 매번 설정해 주지 않아도 자동으로 노드의 위치를 지정해 주면서 그래프를 완성시켜 준다. 이러한 yFile의 기능을 이용해 전사체를 가시화 한다.

3.1 시스템 구조도

개발 중인 시각화 프로그램의 시스템 구조도는 [그림1]과 같다. *.GTF 파일을 DB에서 parsing한 뒤 저장하고, parsing된 데이터를 이용해서 정보를 표시해주고 그래프를 그린다. 최종 적으로는 gene model 파일을 시각화한 그래프와 predicted 파일을 시각화 한 그래프가 출력된다. gene model 그래프를 기반으로 predicted 그래프와 비교해서 발현된 전사체를 비교하고 결과를 통해 novel gene까지도 찾아낼 수 있을 것이다.



(그림 1) 시각화 프로그램의 시스템 구조도.

3.2 파일 format

본 연구에서는 여러 가지 도구에서 나온 *.GFF, *.BED, *.GTF format을 모두 GTF format으로 변환해주는 개발 중 인 내부 프로그램의 변환기를 이용해 한가지의 통일된 GTF format을 이용해 그래프를 가시화 한다. GTF format은 seqname, source, feature, end, score, strand, frame, attribute fields로 구성되어 있다.

3.3 GL000213.1_model.gtf 파일 가시화

그래프로 가시화 할 데이터는 미리 연동되어 있는 DB를 통해서 가져온다. *.GTF format으로 변환된 파일을 parsing해서 DB에 저장해 놓고 필요한 데이터를 불러와서 가시화한다. [그림2]는 입력 받는 파일의 형식이며, 총 4개의 정보를 포함하고 있다. ID와 ID_TRANSCRIPT는 CDS의 ID와 ID_TRANSCRIPT 정보를 나타내며, Start와 End는 CDS의 시작 위치와 끝 위치를 나타내 줌으로써 이 정보를 이용해서 node를 그린다.

ID	ID_TRANSCRIPT	Start	End
GL000213.1	ENST00000327822	138767	139287
GL000213.1	ENST00000327822	134276	134390
GL000213.1	ENST00000327822	133943	134116
GL000213.1	ENST00000327822	131064	131170
GL000213.1	ENST00000327822	129228	129365
GL000213.1	ENST00000327822	126648	126718
GL000213.1	ENST00000327822	121073	121143
GL000213.1	ENST00000327822	119629	119673
GL000213.1	ENST00000327822	118422	118588
GL000213.1	ENST00000327822	109884	110007
GL000213.1	ENST00000327822	108029	108247

(그림 2) 입력 받는 GTF 파일의 형식

[그림 2]와 같은 GL000213.1_model.gtf 파일을 입력 받아서 그래프로 가시화 하면 [그림3]과 같이 나

타난다. GL000213.1_model.gtf 파일에서 start와 end 정보를 이용해서 node의 길이를 구해서 그려준다. 노드의 높이나 색과 같은 요소는 간단한 명령어를 사용해 변경해 줄 수 있다. 또한 mouse over 기능을 이용해 노드 위에 마우스를 올리면 정보가 [그림 4]와 같이 출력된다. 지금은 단순히 start, end, width 정보를 표시하고 있지만 다른 정보들을 이용해 여러 사항들을 추가해 나갈 것이다. 다른 도구에는 이러한 정보를 표시해 주는 기능이 없으므로 이 부분에서도 다른 도구들의 취약점을 보안할 수 있다. 상단의 toolbar에 있는 Fit content 기능은 그래프가 화면의 크기에 맞게 조절해 준다. 사용자가 화면의 크기를 늘리거나 할 때, 화면 크기에 맞게 그래프의 크기가 조절되는 것이다. Zoom in Zoom out 기능은 버튼을 누를 경우 그래프의 확대, 축소가 가능하게 해준다. 전사체나 CDS를 특정 부분만 확대해서 보기를 원한다면 Zoom in 기능을 사용하면 된다.

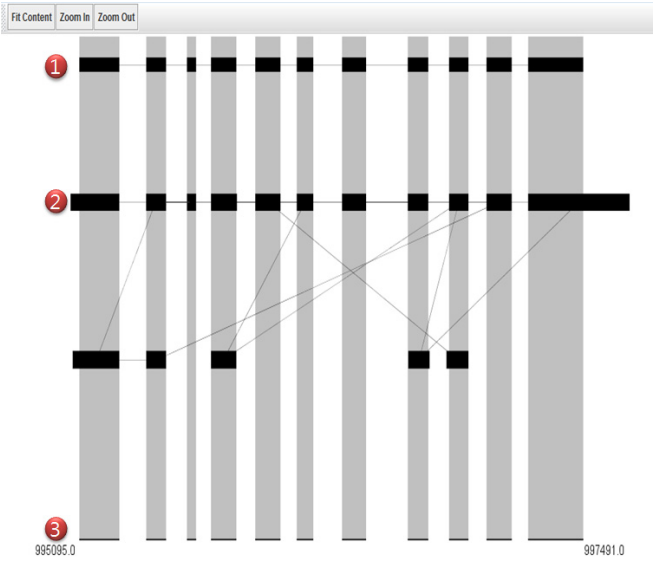
(그림 3) gene GL000213.1의 gene model 파일 가시화.

(그림 4) mouse over 기능을 이용해 정보 출력.

3.4 AT5G03770_model.gtf, AT5G03770_predicted.gtf 파일 가시화

AT5G03770 파일은 SpliceGrapher에서 테스트 데이터로 실험에 사용된 파일이다. tanscript ID와 Protein ID는 AT5G03770.1이고 Chromosome 5번에 속한다. [그림 5]는 gene AT5G03770의 gene model 파일인 AT5G03770_model.gtf와 최종 결과로 나온 predicted 파일인 AT5G03770_predicted.gtf파일을 yfile을 이용해 가시화 한 그래프이다. (1)번이 gene model 파일을 가시화 한 것이고, (2)번이 predicted 파일을 가시화 한 것이다. (3)번은 (1)번 gene model의 depth를 구해서 표시해준다. (1)번에서 CDS가 같은 위치에 겹치는 경우에 depth에 차례로 쌓이게 된다. 또한 gene model을 기준으로 시작과 끝 위치를 depth와 함께 레이블로 표시해줌으로써 범위를 한눈에 파악할 수 있다. 역시 gene model을 기준으로 CDS의 시작과 끝에 색을 넣어줌으로써 predicted 파일과 비교를 하는 것이 쉽도록 했다.

개발 중인 시각화 도구를 이용해 gene model 파일과 결과 파일인 predicted 파일을 비교해 최종적으로 novel gene을 찾는 것이 목표이다. 파일을 그래프로 가시화함으로써 한눈에 CDS의 위치를 파악하고 비교할 수 있고, CDS의 정보까지 손쉽게 알아낼 수 있다. 또한 다른 시각화 도구를 제공해주는 SpliceGrapher와 달리 그래프를 보았을 때 추정된 전사체도 알아보기 쉽게 그려진다.



(그림 5) gene AT5G03770의 gene model 파일과 predicted 파일을 가시화.

4. 결론

본 논문에서는 전사체를 시각화 하는 도구를 개발하는 방법을 제안했다. 개발 도구에서는 다양한 도구의 결과 데이터를 시각화 해주기 위해서 *.GFF, *.BED, *.GTF format을 모두 GTF format으로 변환해주는 내부 프로그램의 변환기를 이용한다. 또한 시각화 기능에 취약한 기존의 도구들과 달리 간단한 사용자의 입력만으로 데이터를 시각화 할 수 있도록

만들었다. 개발한 시스템을 통하여 파일을 그래프로 가시화함으로써 한눈에 CDS의 위치를 파악하고 비교할 수 있었으며, CDS의 정보까지 손쉽게 알 수 있었다. 또한 다른 시각화 도구인 SpliceGrapher에 비해 실시간으로 시각화가 가능하였고 추정된 전사체를 알아보기 쉬웠다. 개발된 시각화 도구를 이용해 최종적으로는 novel gene을 찾아냄으로서 생물학을 연구하는 사람들이 유용하게 사용할 수 있을 것으로 기대된다.

감사의 글

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

참고문헌

[1] Roger S. Pressman "Software Engineering A Practltiners' Approach" 3rd Ed. McGraw Hill

[2] Metzker, M. L. "Sequencing technologies – the next generation." Nature Rev. Genet. 11, pp. 31-46, 2010.

[3] "yworks", http://www.yworks.com/

[4] Rogers, Mark and Thomas, Julie and Reddy, Anireddy and Ben-Hur, Asa, "SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data," Genome Biology, 13, R4, 2012.

[5] Kim SeonYeong, "Graph visualization of th-mirna regulation network using yfiles," Technical Report, pp.SYS-miRNA-003, 2010.