일반 논문( ) / 학부생 논문(○)

염기서열 데이터의 일반적인 다형성 분석 및 시각화 어플리케이션

이일섭, 이건명*
충북대학교 소프트웨어학과

Generic Polymorphism Analysis and Visualization Application of Nucleotide Sequences

Il Seop Lee, Keon Myung Lee
Department of Computer Science, ChungBuk National University

요 약

유전체는 생명체가 가지고 있는 모든 유전적 정보를 담고 있다. 특정 종 내에서는 개체별로 고유의 특성이 나타나며, 이 특성은 유전체의 염기서열 분석을 통해 확인할 수 있다. 종 내에서 개체의 다형성을 파악하기 위해, 각 염기서열의 대립유전자형 출현 빈도의 변화를 관찰하는 많은 연구가 진행되고 있다. 이 논문에서는 다양한 종류의 염기서열에서 대립유전자형빈도의 변화량을 쉽게 파악할 수 있는 분석 및 시각화 어플리케이션을 제안하고, 수두대상포진바이러스의 염기서열 데이터를 이용해 실험한 결과를 소개한다.

1. 서론

최근 유전체 연구의 성장과 더불어 염기서열 분석 기술의 발달에 따라 생명체의 복잡한 염기서열의 정보를 빠르게 얻을 수 있다. 염기서열에서 염기분포의 변화는 개체의 대립형질을 결정해 특정 종 내의 개체의 다형성을 결정짓는데 중요한 역할을 한다. 그로인해, 종 내의 염기서열에서 단일 염기의 차이가 어떤 유전적 기능의 차이로 이어지는지 분석하는 단일염기다형성(Single Necleotide Polymorphism, SNP)에 관한 많은 연구가 이루어지고 있다[1].

이 논문에서는 염기서열간의 염기분포의 차이를 쉽게 파악할 수 있는 분석 및 시각화 어플리케이션을 제안하고, 수두대상포진바이러스(VZV)의 염기서열 데이터를 통해 실험한 결과를 소개한다.

2. 단일염기다형성(SNP)

개체간의 염기서열 차이는 수많은 단일 염기의 차이로 구성된다. 하나의 염기가 삭제되거나 더해지며, 또는 다른 염기로 변화한다. 염기서열분석을 통해 얻어진 염기서열에서 각각의 위치에 대해 대부분을 차지하는 염기를 다수염기(major), 두 번째로 큰 염기를 소수염기(minor)로 분류하며 소수염기의 비율이 5% 이상인 경우만 일반적인 단일염기다형성 대상으로 간주한다. 그 외에는 염기분석에러(Sequencing Error)로 간주하여 분석 대상에서 제외된다[2]. 최근 SNP 연구는 질병의 유전적 관계를 규명하거나 맞춤약물의 개발을 위해 응용된다.

3. 단일염기다형성 분석 및 시각화

3.1 단일염기다형성 분석기

염기서열 분석 데이터의 신뢰성을 위해 동일위치에서 단일염기의 합계 35 이상 및 소수 염기의 비율 5% 이상인 염기위치만을 분석 및 시각화 대상으로 한다.

*This research was supported by the MSIT, Korea, under the Seoul Accord Vitalization Program(IITP-2018-2012-1-00598) supervised by the IITP.

*Corresponding Author : Keon Myung Lee(kmlee@cbnu.ac.kr)

분석기의 그래픽 인터페이스의 모습은 Fig. 1과 같으며, 파일 불러오기 기능과 분석에 필요한 정보를 얻기 위한 컬럼 명 설정 및 분석할 유전적 정보 종류를 입력받는 부분으로 구성된다. 분석은 각 염기서열 위치의 게놈 구조와 반복되는 지역, 유전자 발현지역과 발현되지 않는 지역의 4가지 기준으로 소수염기의 비율 변화를 집계한 분석결과를 엑셀문서로 제공한다

Fig. 1. 분석기의 유저 인터페이스

3.2 염기분포 변화의 시각화

시각화는 염기서열간의 염기분포의 변화를 시각화한다. 기존의 소수염기의 비율만을 표기하던 산점도 시각화를 개선해, 다수염기와 소수염기의 종류와 소수염기 비율의 변화를 파악가능하다. 또한, 사용자 상호작용을 통해 염기종류 및 유전적 정보, 변화율 그리고 염기위치를 필터링할 수 있다. 시각화의 모습은 Fig 2.와 같다. 소수염기의 비율변화에 따라 밝기가 진해지는 방향으로 변화함을 표현하며, 다수 염기는 소수 염기의 변화하는 방향의 머리 부분에 표현된다.

3.3 수두대상포진바이러스 분석 및 시각화

수두대상포진바이러스는 바이러스 특성에 따라 배양을 거듭할수록 소수염기 비율의 변화가 발생한다. 실험에 사용한 데이터는 pOka strain으로, 배양 횟수에 따라 p10, p60, p110의 세 개의 염기서열로 구성된다. [Fig. 2]는 pOka 데이터로 시각화 한 모습이며, 하단에서부터 p10에서 p60으로의 변화, p60에서 p110, p10에서 p110으로의 변화를 표현한다. 계대가 지남에 따라 변이율이 높고, T/c, A/g 염기조합의 변화가 많이 나타남을 쉽게 확인할 수 있다.

4. 결론

본 논문에서는 염기서열 데이터의 분석 및 시각화 어플리케이션을 제안하고, VZV 염기서열 데이터를 통해 실험한 결과를 소개하였다. VZV 염기서열 데이터의 염기종류별 변화비율 및 위치와 같은 변화특징을 쉽게 파악할 수 있었다. 향후 연구에서는 각 염기위치간의 유전적 관계를 파악할 수 있는 분석 및 시각화가 이루어진다면 더 나은 분석효과를 기대할 수 있을 것이다[3].

REFERENCES

[1] The International SNP Map Working Group. (2001). A map of human genome sequence variation containing 1.42 million single nucelotide polymorphisms, *Nature 409,* 928-933.

[2] The International HapMap Consortium. (2005. 10. 27.). A haplotype map of the human genome. *Nature 437,* 1299-1320.
DOI : 10.1038/nature04226

[3] Manolio TA. (2010). How to interpret a genome-wide association study. *JAMA. 363(2),* 166-76.
DOI : 10.1001/jama.299.11.1335

Fig. 2. 염기분포 변화량 시각화