염기서열 데이터 분포 변화량의 시각화 방법

이일섭, 이건명
충북대학교 소프트웨어학과
e-mail:lis123kr@cbnu.ac.kr

Distribution Evolution Visualization of Nucleotide Sequence Data

Il Seob Lee, Keon Myung Lee
Dept of Computer Science, ChungBuk National University

요 약

유전체는 생명체의 구성에 관련된 모든 정보를 포함하고 있다. 특정 종을 하나의 유전체로 표현하지만, 해당 종에 속하는 개체의 염색체는 조금씩 차이가 있어 개체별로 고유의 특성이 나타난다. 바이러스와 같이 개체 변이가 많이 일어나는 종에서는 종내에서 변이가 심할 수 있다. 종내에서 변이에 따른 특성을 파악하기 위해, 각 염기 위치별로 염기분포를 관찰하는 연구들이 진행되고 있다. 이 논문에서는 염기분포의 변화를 쉽게 분석할 수 있도록, 각 염기 위치에서의 분포변화를 시각화하는 방법을 제안하고 구현 결과를 소개한다.

1. 서론

유전체 연구의 성장에 따라 사람의 복잡한 염기서열 정보를 빠르게 얻는 일이 가능하다. 염기서열에서 염기 분포의 변화는 유전자 이상과 관련된 질환의 원인을 규명하거나 복합질병의 유전자 결함을 찾는데 중요한 역할을 한다. 차세대 염기서열분석(NGS)기술이 발달됨에 따라 대용량의 염기서열을 보다 빠르고 신속하게 저비용으로 분석할 수 있다[1]. NGS는 짧은 길이의 시퀀스 조각인 리드(read)들을 대량으로 생성하여 이를 바탕으로 전체 염기서열을 밝혀낸다. 대량의 리드들로부터 같은 염기 위치에 대해 다수로 염기 결정을 하게 된다. 염기서열 데이터를 분석함에 있어 시간에 따라 같은 염기서열에 대해 염기분포가 어떻게 변화하는지 파악하는 시각화 방법이 필요하다.

본 연구에서는 동일종인 염색체의 염기 위치별 분포 변화를 분석할 수 있는 시각화 방법을 제안한다. 각 염기 위치마다 염기 분포를 나타내기 위해 바이러스학 등에서 사용되는 분포에 대한 여러 척도의 값을 이용하여 분포의 변화를 시각화하고 다양한 인간 상호작용기법을 통해 데이터 표현의 한계점을 보완한다. 또한, 사용자로부터 특정 척도 변화의 특징을 입력받아 해당 척도에서 유사한 특징을 지닌 염기서열을 추출하는 기능을 통해 분석의 효율을 높인 시각화 방법을 제안한다.

2. 관련 연구

2.1 분포에 대한 척도

염기서열에서 염기 분포의 작은 변화가 유전질병의 원인이 될 수 있다. 따라서 염기 분포의 변화를 파악하는 것이 중요하며 분포의 변화를 자세히 파악하기 위한 척도로 8가지를 사용한다[2]. 사용한 척도는 Shannon entropy (H_s), Hill numbers(^qD, q=1, 2, ∞, -∞), Functional Attribute Diversity(FAD), Mutation frequency at the entity level(Mfe), Nucleotide diversity(π)이다. 위 척도들은 바이러스학 등에서 유전자내에 생긴 많은 돌연변이점의 분산을 나타내는 돌연변이 스펙트럼의 분포를 파악하기 위해 연구되었다[3]. 이 8가지 척도를 이용하여 염기서열에서 염기들의 분포를 파악한다.

2.2 데이터 시각화

데이터 시각화는 데이터 분석 결과를 쉽게 이해할 수 있도록 시각적 수단을 통해 정보를 효과적으로 전달하는 것을 말한다. 데이터 시각화를 위한 수단으로 D3.js[4], Google Chart Tools, R을 활용한 프로그래밍 등이 있다. 본 논문에서는 D3.js를 활용하여 제안하는 시각화 방법을 구현한 모습을 보여준다. D3.js는 사용자와 상호작용을 가능케 하여 동적인 효과를 주어 효율적이고 다채로운 데이터 시각화를 구현해준다[5]. 다른 시각화 도구와 달리 그래프의 기본 형태를 제공하지 않지만 직접 그릴 수 있도록 하여 활용 수준에 따라 폭넓게 구현 가능하다.

3. 제안하는 시각화 방법

3.1 시각화 인터페이스 설계

제안하는 시각화 방법의 기본 형태는 [그림 1]과 같은 누적 세로 막대형 그래프이며, 스택형 막대그래프라고도 부른다. 하나의 누적 세로 막대형 그래프는 염기 분포를 표현하는 8가지 척도 중 하나의 변화량을 표현한다. 유전

정보는 염기들의 순서를 통해 표현되기 때문에 염기서열의 순서를 그래프의 X축 상에 나타낸다. 결과적으로 8개의 척도를 표현하기 위해 총 8개의 누적 세로 막대형 그래프가 필요로 하며 X축을 일치시켜 염기서열에 대해 각 척도의 비교를 용이하게 한다.

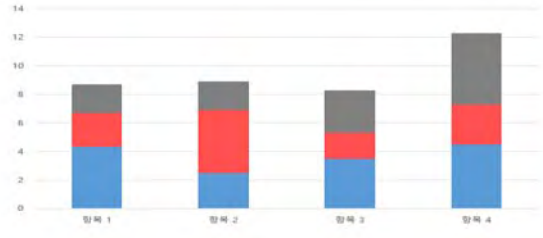

그림 1. 일반적인 누적 세로 막대형 그래프

NGS를 통해 결정된 같은 범위의 여러 염기서열이 추출된 시간에 따라 염기의 분포의 변화가 어떻게 진행되는지에 초점을 둔다. 누적 세로 막대형 그래프의 막대에서 누적되는 층은 아래층부터 순차적인 값의 변화량을 의미한다. 즉, 각 염기 위치에서 막대의 같은 스택 층은 같은 시간대의 변화량을 의미하게 된다. 값의 증가 및 감소를 각각 파란색과 빨간색으로 구분하며 값의 변화가 없다면 어두운 회색으로 표현한다.

염기서열의 길이로 인해 X축에 모든 염기 위치를 표현하지 못하는 한계점과 시각적인 효과로 분석의 효율성을 위해 다양한 인간 상호작용기법을 사용한다. 누적되는 막대에서 순차적인 값의 변화를 효율적으로 관찰하기 위해 마우스 이벤트를 이용해 8가지 척도의 같은 스택 층의 막대를 부각시키고 정렬하여 비교를 돕는다. 또한, 염기서열의 범위를 조절시킬 수 있는 브러시기능을 추가하여 염기서열 표현의 한계점을 보완한다.

전체 그래프의 상단부분에 염기 위치마다 염기들의 종류와 수로 표현된 염기서열을 나타내어 분포에 대한 척도와 함께 염기 데이터의 이해를 돕는다.

3.2 시각화 결과

본 논문에서 제안하는 시각화의 시스템 구현은 Python의 웹 프레임워크인 Django와 시각화 라이브러리인 D3.js를 이용한다. Django에 포함된 데이터베이스에 염기서열 데이터를 저장하고 D3.js의 data함수를 이용해 웹 환경의 그래픽을 표현하는 SVG(Scalable Vector Graphics)요소와 결합한다. 또한, D3.js의 마우스 이벤트를 활용하여 동적인 효과를 준다.

[그림 2]는 시각화 시스템의 구조를 도식화한 것이다. D3.js에 불러온 데이터와 결합된 SVG요소들은 데이터에 따라 값을 표현하며 시간에 따라 누적 막대의 아래층부터 순차적인 값의 변화량이 표현된다. 순차적인 값의 변화량을 효율적으로 파악하기 위해 [그림 3]와 같이 D3.js의 Mouseover 이벤트를 이용해 8가지 척도의 같은 스택 층의 막대를 부각시키며 정렬한다. 이러한 시각적인 효과를 통해 Hill number(q=2), Hill number(q=−∞), FAD를 제외한 나머지의 척도가 대체적으로 증가함을 쉽게 알 수 있다. [그림 3]에서 다섯 번째에 표현된 염기 위치에 대해 Hill number(q=2)는 감소하며 Hill number(q=−∞), FAD의 값은 변화가 없고 나머지 척도는 모두 증가함을 파악할 수 있다. 또한, 이 염기 위치에서는 A염기로만 이루어져 있는 것을 상단 부분에 표현된 염기서열을 통해 확인할 수 있다. 이처럼 구현된 시각화 시스템을 이용하여 척도 값의 증가 및 감소를 통해 염기의 분포가 다양해졌는지 그리고 변이 빈도가 어떻게 변화하였는지 쉽게 파악할 수 있다.

3.3 분포 변화의 특징 추출 기능

염기서열의 염기 위치별 비슷한 분포의 변화가 일어나는 특징을 추출하는 기능을 통해 분석을 돕는다. 사용자로

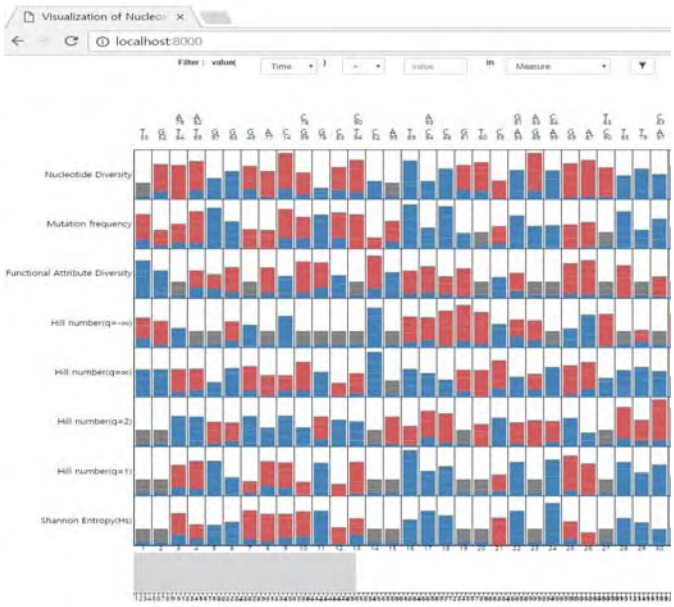

그림 2. 구현된 시각화 시스템

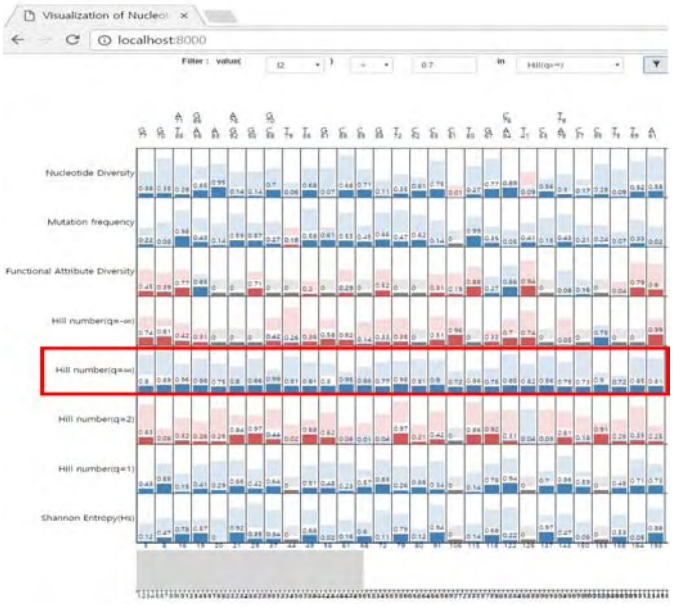

그림 3. 시각적 효과 및 특징 추출된 모습

부터 특정 척도와 조건 값을 입력받아 해당 척도와 그 값이 조건에 부합하는 염기 위치들을 뽑아 정렬한다. [그림 3]의 빨간색 영역은 Hill number(q=∞)의 값의 변화가 0.7 이상인 특징을 갖는 염기 위치들을 뽑아 정렬한 모습이다. 이러한 기능을 통해 염기 위치별 염기 분포가 집중되는 특징을 파악할 수 있다. 또한, 염기 분포가 분산되는 염기 위치, 변이가 빈번하게 발생하는 염기 위치 등 다양한 척도의 조건 값에 따라 염기서열의 위치별 특징을 파악할 수 있다.

4. 결론

NGS를 통해 얻어진 동일종인 염색체의 염기서열에서 염기 분포의 변화량을 분석할 수 있는 시각화 방법에 대해 제안하였다. 염기 위치별 분포를 파악하기 위해 바이러스학 등에서 사용되는 여러 척도를 사용하였고, D3.js를 이용해 구현 결과를 보였다. 또한, 다양한 인간 상호작용 기법과 특징 추출 기능을 추가하여 분석의 효율을 높일 수 있었다. 향후 연구에서는 특징 추출에 대해서 머신러닝 기법과 같은 다양한 알고리즘을 통해 이루어진다면 더 나은 분석효과를 기대할 수 있을 것이다.

Acknowledgement

본 논문은 교육부가 지원하고 충북대학교가 수행하는 지역선도대학육성사업의 지원을 받아서 수행되었습니다.

참고문헌

[1] 배세은, 김하연, 이지혜, 장진화, 손현석, “Bioinformatics의 새로운 기술-NGS의 현재 그리고 미래”, 보건학논집, Vol. 48, No. 1, pp. 12-22, 2011.

[2] Gregori J., Perales C., Rodriguez-Frias F., Esteban JI., Quer J., Domingo E., “Viral quasispecies complexity measures”, Vol. 493, pp. 227-237, Elsevier, 2016.

[3] Gregori J., Salicrú M., Domingo E., Sanchez A., “Inference with viral quasispecies diversity indices: clonal and NGS approaches”, Vol. 30, No. 8, pp. 1104-1111, Bioinformatics, 2014.

[4] 파블로 나바로 카스틸로, “D3.js 실시간 데이터 시각화”, 에이콘, 2015.

[5] 스캇 머레이, “D3.js: 쉽고 빠른 인터렉티브 데이터 시각화”, 인사이트, 2014.