생명의료 빅데이터 분석에 기반한 자동 정밀 진단 시스템의 현재와 미래

부산대학교 ▌송길태*

1. 서 론

컴퓨터 소프트웨어 분야는 생명 의학 기술을 비롯하여, 차세대 로봇 등 여러 첨단 기술 및 컴퓨터 시스템과의 융복합에 더욱 박차를 가하고 있다. 특히 인간 건강과 수명에 직결되는 생명 의학 분야와 관련된 컴퓨터 소프트웨어 융합 기술 진보는 우리의 예상 보다 훨씬 빠르게 진행되고 있다. 환자 맞춤형 자동 정밀 진단을 실현시킬 차세대 디지털 병원 시스템은 지금 현재 진행형으로 구축되고 있으며, 머지 않아 상용화 단계에 이를 것으로 전망된다.

2. 자동 정밀 진단에 기반한 디지털 병원 시스템

자동 정밀 진단에 기반한 디지털 병원 시스템은 전자적으로 진료 기록을 저장하여 과거 축적된 정보를 통계학적으로 유미의한 차원에서 분석하는 것에서 머무르지 않고, 진료 시 개인의 환경, 생활 패턴, 그리고 유전적 특성을 자동으로 분석하여 진단 결정을 내리는 시스템이다 [1]. 보다 진화된 컴퓨터소프트웨어를 응용한 디지털 병원 시스템은 환자 개인의 유전 정보, 진료 기록, 이미지 데이터 및 환자의 진료 노트 등을 포함한 각종 다양한 형태의 대용량 자료들과 클라우딩 컴퓨팅을 기반으로 한 빅데이터 분석기술을 통해 각 자료들간의 연관성을 정교하게 분석해낼 것이다. 아울러 애플 왓치, 스마트 컨택트 렌즈, 자폐아의 감정을 읽는 구글 안경과 같은 새로운 IT 모바일 기기의 등장은 미래형 디지털 병원 시스템 구현이라는 혁신을 더욱 앞당길 것이다. 애플왓치를 통해서 실시간으로 개인의 건강 정도나 운동량을 측정할 수 있고, 스마트 컨택트 렌즈를 통해서 실시간으로 각 개인의 건강 상태와 관련한 이미지나 비디오 영상을 저장할 수 있다 [2]. 그리고 구글 안경은 사람의 감정을 읽어내는 기술에까지 도달하여, 이미 자폐아 등의 감정을 알아낼 수 있게 되었다 [3]. 이러한 IT 기기와 각 개인으로부터의 저장된 다양한 데이터들은 네트워크를 통해 대용량 데이터들과 연결된다. 이렇게 인터넷을 통해 연결된 수많은 이질적 형태의 빅데이터들은 환자 맞춤형 자동 진단을 내리는데 중요한 단서가 된다. 따라서 이질적 형태의 대용량 데이터를 분석하기 위한 머신러닝 알고리즘 설계 및 소프트웨어 개발은 미래형 차세대 디지털 병원 시스템을 구축 하는데 있어서 핵심적인 문제이다.

이 문제를 해결하기 위하여 세계 각지의 기업과 대학 연구소에서는 각종 생명의학 정보 데이터의 분석 자동화를 위한 알고리즘 설계 및 소프트웨어 개발, 데이터 베이스 구축 그리고 시각화 문제 등의 연구를 활발하게 진행 중에 있다.

3. 주요 의료 빅데이터

3.1 유전체 염기 서열 빅데이터

기술의 발달과 함께 기하급수적으로 축적되고 있는 생명의학 빅데이터의 대표적인 예로 텍스트 형태의 유전체 빅데이터를 들 수 있다. 유전체란 사람의 다양성을 표현하는 유전 정보를 담고 있는 DNA 내의 염기 서열 전체를 말한다. DNA 내의 염기는 A, C, G, T 네 종류가 존재하는데 유전체 염기 서열 데이터는 이 네 종류의 문자만으로 구성된 문자열 순서라고 생각하면 그 개념을 쉽게 이해할 수 있다. 사람 유전체 데이터의 경우 약 30억 개의 염기 서열로 구성 되는 데 각 염기 서열 영역에 대한 기능을 육안으로 일일이 해독해내는 것은 천문학적인 시간과 노력이 소요되는 등 현실적으로 거의 불가능하다고 할 수 있다 [4]. 수작업의 한계를 뛰어 넘기 위해서는 컴퓨터 알고리즘,

* 정회원

머신 러닝 및 데이터 마이닝 기법의 개발이 필수적이라고 할 수 있다.

3.2 유전체 빅데이터 분석을 위한 세계적인 컨소시엄 프로젝트

유전체 대용량 데이터의 각 염기 서열 영역에 대한 기능을 해독해내기 위해서 DNA 염기 서열 데이터는 각종 다양한 형태의 실험 및 진료 데이터에 연결시켜 임상적으로 검증하는 작업이 요구된다. 이러한 연구를 위해서 IT와 생명 의료 분야 여러 연구실이 함께 참여하는 많은 세계적인 컨소시엄 프로젝트가 활발히 진행 중에 있다.

3.2.1 ENCODE 프로젝트

ENCODE 프로젝트는 2003년에 발족되어 현재는 미국과 유럽의 30여개의 주요 대학과 연구소가 참여하고 있다. 이 프로젝트에서는 유전체 빅데이터의 변이 패턴을 분석하여 암과 같은 주요 유전 질병과의 연관성을 밝혀내는 것을 궁극적인 목표로 하고 있다 [5]. 이 컨소시엄 프로젝트에서는 다양한 인간 유전체 빅데이터 분석 자동화를 위하여 150여개의 소프트웨어 툴이 개발되었다. 현재는 사용자가 이러한 툴을 클라우드 환경에서 사용할 수 있도록 하는 시스템을 구축 중에 있다. 이 컨소시엄에서 생성되고 분석된 데이터는 ENCODE 데이터 포털 (https://www.encodeproject.org)에서 체계적으로 관리되고 있고 웹 인터페이스를 통해 사용자들이 손쉽게 데이터를 이용할 수 있다.

3.2.2 유전체 10K 프로젝트

인간 유전체 변이 패턴과 그 기능은 쥐나 원숭이와 같은 포유류 유전체 빅데이터 분석을 통해서도 밝혀낼 수 있다. 예를 들어 HIV 연구를 위해서 마카크원숭이 (Pig-tail Rhesus)가 널리 쓰인다. 이러한 연구를 위해 10만개의 척추동물 유전체 수집 및 분석을 목표로 하는 “Genome 10K Project”가 구성이 되어 진행중이다 (https://genome10k.soe.ucsc.edu). 특히 이 프로젝트에서는 수많은 염기 서열 조각들이 랜덤하게 구성된 염기 서열 원데이터를 하나의 완성된 유전체 염기 서열로 복원해 내기 위해 개발된 여러 어셈블리 알고리즘을 검증하기 위한 Assemblathon 프로젝트도 함께 진행되고 있다 [6]. Assemblathon과 같은 소프트웨어 검증을 위한 컨소시엄이 구성되는 이유는 소프트웨어를 이용한 의료 빅데이터 분석 결과에 대한 정확성을 평가하기가 난해한 경우가 많기 때문이다.

3.2.3 암 유전체 분석 컨소시엄

정상 세포내의 DNA 염기 서열에서 비정상변이가 일어나게 되면 많은 복잡한 염기 서열의 변형을 가져오는 암세포로 발전하기도 하는데 이러한 암 유전체 염기서열을 연구하기 위한 컨소시엄 프로젝트도 활발히 진행 중에 있다(예: TCGA-https://cancergenome.nih.gov, ICGC-http://www.icgc.org). 특히 암과 같은 유전적 질병 환자에 대한 유전체 염기서열 데이터 분석과 유전체 빅데이터에 기반한 의료 서비스가 조만간 우리나라 건강보험 적용을 받을 것으로 전망된다. 또한 최근 제 3세대 유전체 데이터 생성 기법 (3rd generation sequencing) [7]의 출현으로 유전체 데이터 품질이 점차 개선되고 있다. 이에 따른 개인 및 암 유전체 데이터 분석 서비스 수요 증가로 유전체 빅데이터 볼륨은 기하급수적으로 증가할 전망이다.

3.3 유전체 빅데이터 분석

암 유전체 및 환자 개인 유전체 빅데이터를 분석하기 위한 머신러닝 알고리즘과 자동화 소프트웨어를 개발이 활발히 진행 중에 있다. 현재 머신 러닝은 학습하는 기법 자체를 연구하는 단계에서 벗어나 학습해 낸 결과들을 실제 응용 도메인에서 의미있는 특징(features)으로 추출해 내어서 사용자에게 어떻게 의미있는 정보로 전달할 수 있을 것인가에 초점을 맞추고 있다. 예를 들면 유전체 빅데이터 분석을 위해 뉴럴 네트워크 방식으로 학습해 낸 결과의 정보의 정확성과 우수성을 수치화하는 문제가 머신러닝 분야의 주요 이슈 중 하나이다.

또한 유전체 빅데이터들은 세계 여러 연구소에서 다양한 데이터베이스 및 웹 리소스로 관리되고 있다. 세계 컨소시엄을 통해 정해진 표준에 따라 여러 정보들이 수집되고 연결되는데, 급격한 기술의 발달로 이 표준을 넘어서는 새로운 빅데이터들이 양산되고 있다. 이러한 새로운 빅데이터 분석을 자동화하고, 이를 반영하는 새로운 표준을 자동으로 구축하는 툴을 구현하는 문제도 주요 이슈 중 하나이다. 최근 여러 연구실에서 개발되고 유지되어 오고 있는 다양한 유전체 데이터베이스를 하나의 표준으로 통합하는 프로젝트가 발족되어 진행되고 있다.

여기에 그치지 않고 각종 생명 의학 빅데이터와 소셜 네트워크 정보와의 융합을 통하여 개인이나 특정 집단의 인간관계 활동 및 생활 패턴과 질병과의 연관성을 찾아내는 알고리즘 및 소프트웨어 개발도 진행중에 있다. 2016년 봄에 University of California, Berkeley의 Simons Institute에서 주관한 “Algorithmic Challenges in Genomics”라는 주제의 워크샵이 열렸는데, 이 워크샵에서는 유전체 빅데이터와 네트워크 정보를 연결하는 최신 알고리즘이 활발히 논의 되었다.

4. 의료 빅데이터 기술에 대한 앞으로의 전망

의료 빅데이터 분야를 이끌어가고 있는 미국 스탠포드 대학교에서는 의료, IT, 데이터 과학 분야의 세계적인 전문가들이 모여서 유전체 대용량 데이터와 같은 의료 빅데이터 분석을 통한 자동 정밀 진단 시스템 실현이라는 주제로 "Big Data in Biomedicine"이라는 컨퍼런스를 매년 5월 개최하고 있다. 이러한 네트워크를 통해 구글, IBM, 애플 등 세계 유수의 IT 기업 뿐만 아니라, 미국 국립 보건원 (NIH), 스탠포드 대학 병원 등 첨단 의료 기술을 선도해 가는 의료 기관에서는 클라우드 컴퓨팅과 같은 새로운 IT 기술과 유전체 염기서열 데이터와 같은 새로운 형태의 의료 진단 컨텐츠 융합을 통해 미래형 의료 혁신을 이루어내고자 하는 노력을 다각도로 벌이고 있다. 또한 여러 독립적인 시스템에 분산되어 관리되고 있는 대부분의 의료 관련 빅데이터들을 체계적으로 통합하여 관리하는 시스템 구축을 위한 연구가 활발하게 진행되고 있다(예: ClinGen 프로젝트, https://www.clinicalgenome.org). 대부분 의료 빅데이터는 임상적으로 밀접히 연관되어 있기 때문에 이러한 통합 시스템 구축을 통해서 연결된 각종 의료 빅데이터는 데이터 마이닝 및 머신 러닝 기법을 적용하여 암과 같은 희귀한 유전적 난치병으로 고통 받고 있는 환자들에 대한 획기적인 진단 기법을 설계하는데 기여할 수 있을 것이다. 이러한 개인 유전체 및 질병 유전체 빅데이터 해독에 기반한 자동 정밀 진단 시스템은 미래창조과학부, 보건복지부 관련 기관 및 정부 연구소, 주요 의과 대학을 비롯한 주요 국내 병원 및 관련 산업체의 큰 숙원 사업이다. 빅데이터 분석 머신 러닝을 실행하는데 요구되는 고성능 컴퓨팅 시스템 및 네트워크 인프라연구, 이미지와 비디오 데이터 프로세싱 등과 같은 분야와의 융합은 빅데이터 분석 자동화 및 스마트 진단 결정시스템이 차세대 미래형 디지털 병원 시스템 구축이라는 큰 그림으로 완성되어 질 수 있으리라 기대된다.

참고문헌

[1] IBM Global Business Services, "The digital hospital evolution", IBM Corporation White Paper, 2013.

[2] N. M. Farandos and et al., "Contact Lens Sensors in Ocular Diagnostics", *Advanced Healthcare Materials*, vol 4, pp. 792-810, 2014.

[3] J. C. Kraft, "Can this Google glass app help kids with autism?", KQED, September 2015.

[4] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome", *Nature*, vol 409 (6822), pp. 860-921, 2001.

[5] The ENCODE Project Consortium, "The ENCODE (ENCyclopedia Of DNA Elements) Project", *Science,* Vol. 306 (5696), pp. 636-640, 2004.

[6] D. Earl and et al., "Assemblathon 1: A competitive assessment of *de novo* short read assembly methods", *Genome Research*, vol 21 (12), pp. 2224-2241, 2011.

[7] D. J. Munroe and T. J. Harris, "Third-generation sequencing fireworks at Marco Island", *Nature Biotechnology*, Vol 28, pp. 426-428, 2010.

약 력

송 길 태

1999 서울대학교 전산과학과 졸업(학사)
2001 서울대학교 컴퓨터공학부 졸업(석사)
2001~2004 해군사관학교 전산과학과(전임강사)
2011 Computer Science and Engineering, Pennsylvania State University 졸업 (박사)
2012~2016 Post-doctoral scholar at Stanford University
2016~현재 부산대학교 정보컴퓨터공학부 조교수
관심분야 : 데이터마이닝, 생명의료정보학
Email : gsong@pusan.ac.kr