Hanyang Medical Reviews

Hanyang Med Rev 2017;37:93-98
https://doi.org/10.7599/hmr.2017.37.2.93
pISSN 1738-429X eISSN 2234-4446

Review

게놈 의학에 기계학습을 적용한 딥지노믹스 연구사례 리뷰

김태형

테라젠이텍스 바이오 연구소

Genomic medicine is to determine how an individual's DNA alteration can affect the risk of various diseases and to understand mechanisms and design targeted treatments. Here, we focus on how machine learning helps model the relationship between DNA and molecular phenotypes in a cell. Modern biology enables high throughput measurements of many cellular variables that can be handled as a training target for predictable models, such as gene expression, splicing, and protein binding to DNA or mRNA. With the increasing availability of large datasets and advanced computer skills such as deep learning, researchers have opened a new era in effective genomic medicine.

Key words: deep learning; machine learning; genomic data; genome biology; genome medicine

Corresponding Author: Tae Hyung Kim
443-270, 경기도 수원시 영통구 이의동
차세대융합기술원 테라젠이텍스 바이오 연구소
Theragen Bio Institute, TheragenEtex,
Suwon, Korea.
E-mail: thkim@therabio.kr

Running title: Robotic Thyroidectomy

Received 2 Aug 2017
Revised 19 Oct 2017
Accepted 9 Nov 2017

서 론

본 논문은 게놈 의학과 게놈 생물학 분야의 중요한 문제들을 기계학습 기술을 적용해 다루고 있는 딥지노믹스(Deep Genomics)의 최근 연구 결과들을 리뷰하여 정리한 것이다.

1950년대 초, DNA 분자 구조가 밝혀짐으로써 DNA가 유전정보가 담겨 있는 물리적 저장 매체라는 것이 처음 밝혀졌다[1]. 이후 50년이 지난 2001년에는 인간 게놈 프로젝트를 통해 표준 인간 게놈 (Reference Human Genome)이 완성되었다[2]. 하지만 이를 해석하는 것은 또 다른 난관이었다. 게놈은 생명체를 만들기 위한 설계도임에도 불구하고 우리가 읽어낸 인간 게놈 지도는 문자열로 이뤄진 단순한 게놈 서열의 정보일 뿐 이를 완벽히 해석할 수 없었던 것이다. 비로소 오늘날에서야 게놈 데이터 생산 기술의 비약적인 발전과 함께 게놈 빅데이터를 활용할 수 있는 상황이 되었으며 이러한 데이터를 기반으로 게놈 의학 분야에 기계학습의 적용이 가능해지고 있다. 예방 조치를 위한 잠재적 질병의 발병을 예측하고 적합한 타깃 치료제를 찾기 위해 이들 데이터는 활용될 예정이며 이러한 게놈 의학을 구현하기 위해서는 세포 내부 현상을 모니터링 하고 게놈 서열을 정확하게 해석할 수 있는 컴퓨터 모델 및 시스템이 개발되어야 한다. 이런 컴퓨터 시스템의 구축은 실험실 수준에서의 모델 생물 실험 보다 더 효과적으로 유전 변이에 기반을 둔 잠재적 치료제의 신속하고 효율적인 발굴을 가능케 한다.

현재까지는 게놈 내 단백질 코딩 엑손 영역이 가장 해석이 잘 되어 있는 영역이다. 인간 게놈은 전체의 2%에 해당하는 20,000여개의 단백질을 코딩하는 유전자[3]와 25,000개 이상의 비 단백질 코딩 유전자[4]를 가지고 있으며 이외의 영역 또는 일부 다른 영역은 생명에 아주 결정적으로 작용하거나, 혹은 없어지더라도 전혀 문제가 되지 않기도 한다[5]. 그리고 질병의 원인이 되는 돌연변이들은 단백질 코딩 영역 외의 다른 영역에서도 빈번히 발견된다. 특히 기능적 비 단백질 코딩 영역은 대부분 게놈을 조절하는 조절서열영역(regulatory sequence region)이다. 이들은

엑손/인트론을 구분해 다양한 전사체를 어떻게 선택해 만들지를 결정하여 유전자의 발현에 관여하고 이를 통해 세포의 복잡성에 가장 크게 기여한다. 이러한 복잡한 영역까지 해석하기 위해서는 살아 있는 세포가 게놈을 읽고 해석하는 것처럼 초인간적인 분석 능력(super-human analytical ability)이 요구되며 이를 구현할 시스템이 필히 수반되어야 한다.

이러한 이유로, 게놈 생물학 전문가들은 게놈을 해석할 수 있는 기계학습 기술을 개발하기 시작했다. 마침 게놈 해독 기술이 저렴해지면서 엄청난 데이터들이 대거 쏟아졌으며 특히 엄청난 규모의 암 유전체 데이터를 생산하는 TCGA 같은 암 발생 원인을 연구할 수 있는 데이터들이 축적되었다. 최근에는 유전자 편집 기술을 이용해 특정 유전자에 새로운 서열을 도입하거나 제거 할 수 있게 되었으며 이를 통해 대규모로 유전자 각각의 기능을 예측할 수가 있게 되었다. 즉, 자유자재로 게놈을 읽고(decoding), 쓰고(encoding) 할 수 있게 되었음으로 이를 기반으로 본격적으로 세포 내 현상 및 게놈 특정 영역의 기능 해석, 특정 질병과의 연관성을 학습(기계학습)을 통해 이해하려는 시도가 이루어지고 있는 것이다.

본 론

1. 기계 학습을 이용한 게놈 해석

인류는 오랜 기간 동안 유전형으로부터 표현형을 예측하기 위한 많은 시도를 했으나 쉽게 해결되지 않았다. 물론 기계학습 기술로도 쉽지 않은 문제이며 복잡하고 해결하기 어려운 이유는 다음과 같다. 먼저 유전형과 표현형 사이에는 우리가 알지 못하는 복잡하고 다양한 세포 내 환경변수들이 다수 존재하고 있다. 그래서 단순히 유전형 데이터만 가지고 표현형을 정확하게 예측하는 것은 불가능하다. 그래서 유전형-표현형 사이에 존재하는 다양한 형태의 세포 내 현상을 대변하는 관련 데이터들(유전자 발현, 선택적 스플라이싱, 단백질/RNA 바인딩 DNA 등)을 잘 정의하고 기계학습 데이터 세트로 잘 활용해야만 한다. 또 다른 이유로는 우리가 질병 위험을 예측하는 모델을 추론 할 수 있다 하더라도 이 모델이 세포 메커니즘을 반영하는 숨겨진 변수들을 전혀 반영하지 못할 가능성이 높고 우리가 추론한 대부분 모델이 전혀 일치 되지 않을 가능성이 높다는 것이다. 치료제를 개발하기 위해서는 질병 메커니즘에 관한 통찰력을 가지는 것이 매우 중요하며 이러한 통찰력이 갖추어진다면 정확한 표적에 대한 지식이 없다 하더라도 우리가 원하는 타깃을 찾거나 표현형 스크리닝에 대한 아주 중요한 정보를 발견할 수가 있다[6]. 이러한 한계를 극복하는 방법으로써 분자표현형(molecular phenotype)이라고 하는 측정 가능한 중간 단계의 세포 변수 예측은 최적의 컴퓨터 모델을 만듦에 있어 매우 강력한 접근법이 될 것이다.

우리가 추론하고 이를 통해 기계학습해야 할 핵심적인 분자표현형, 즉 세포 변수들을 언급 해보자면, 유전자로부터 단백질까지 코딩 과정에 엑손은 어떻게 달라지며 어떤 조절 단백질이 특정 어떤 위치에 결합하여 발현에 기여하는지, 전사체 카피수는 얼마나 되는지, 전사체가 단백질로 발현되는 비율과 그 단백질의 농도가 분자표현형이 될 수 있을 것이다. 이와 같은 분자표현형을 모델로 해서 추론하는 것이 매우 중요하며 이들 세포 변수들은 표현형보다 게놈 서열과 더 밀접하게 관련되어 있어 더 쉽게 결정될 수 있다[7]. 이들 세포 변수들은 유전자-전사체, 전사체-단백체, 단백체-구조체 등과 같이 중간 세포 활성물질이 될 수 있으므로 이들 모두는 치료제 개발을 위한 매우 좋은 타깃이 될 가능성이 매우 높다.

2. 게놈 생물학/게놈 의학 분야 기계학습 적용

생명과학자들은 살아 있는 세포들의 특정 생명 현상과 숨어 있는 세포 변수들을 이해하기 위해 게놈 데이터와 상호작용에 관여하는 데이터를 실험을 통해 대량으로 생산해 내며 데이터 과학자들은 이들 데이터를 이용해 목적한 특정 생명현상을 위한 계산 모델(computational model)을 만들어 학습하게 한다. 이때 다양한 세포 변수들을 이용해 특정 계산 모델을 잘 만들기 위해서는 생물학적으로 이들을 정확하게 측정하는 분석법이 있어야 하며 여러 다양한 조건에서의 훈련 데이터를 수집할 수 있어야 한다. 최근 게놈 데이터들을 대량 생산하고 분석할 수 있는 기술들이 상용화 됨으로써 기존보다 수만~수십만 배 비용을 적게 들이고 효율적으로 실험이 가능해져 특정 세포 변수에 해당하는 수만 개에서 수백만 개의 측정 데이터들을 손쉽게 얻을 수 있게 되었다.

우리는 게놈 시퀀싱을 통해 흥미 있는 영역 또는 전체 게놈 데이터를 확보해 다양한 종의 유전적 변이 의 프로파일링을 하거나 의학적 목적으로 사람 개개인의 게놈에서 발생하는 변이의 프로파일링을 할 수 있게 되었다. 사실 암과 신경질환에서 발생하는 체세포 변이들은 변화가 있지만 개인 게놈은 상대적으로 변화하지 않고 안정적이므로 다른 변수에 비해 훨씬 쉽게 가능한 때문이다.

하지만 전사체의 경우는 다르다. 개개인의 세포마다 그 주위 환경에 따라서 영향을 받아 발현 양상이 매우 다르며 조직 타입에 따라서도 모두 다를 수 있다. 하지만 RNA-Seq이라는 대용량 시퀀싱 기술이 적용되어 이들 데이터를 어떤 기술보다도 효율적으로 확보할 수가 있게 되었으며 또 다른 애플리케이션으로는 특정 DNA 영역에 결합하는 단백질들이 어떤 것이 있는지 대규모로 프로파일링하는 것도 가

능해졌다[8]. 이러한 세포/조직 별 발현 전사체와 단백질 결합과 관련된 데이터들을 이전보다는 손쉽게 확보할 수 있게 됨으로써 기계학습 기술을 이용해 세포 변수들의 예측 모델을 추론하려는 데이터 과학자들에게는 큰 기회가 열렸다.

게놈 의학 분야에서 기계학습 또는 인공지능 기술이 효율적으로 적용되기 위해서는 입력 데이터의 충분한 확보가 가능해야 한다. 최근 즉시 이용 가능한 게놈 데이터의 공급이 급속도로 많이 늘어나고 있어 환자의 질병을 예측 시 게놈 데이터가 표준으로 활용될 가능성이 점점 커지고 있다. 그래서 결국 모델의 일반화가 잘 이루어진다면 병변 세포의 실험적 측정 없이 궁극적으로는 게놈 서열 내 존재하는 변이의 분석만으로도 환자의 질병 상태를 정확하게 예측하는 것을 기대하며 실제로 딥러닝 기술을 이용해 참조 게놈과 건강한 조직들을 사용하여 훈련한 모델을 통해 척수성 근위축증, 유전성 비용종성 대장암, 그리고 자폐증을 일으키는 키 돌연변이를 정확하게 예측하는 데 성공했다[9].

3. 유전형–표현형 사이에 존재하는 다양한 세포 변수들

다양한 세포 내 변수들로는 유전자 전사, DNA 메틸화, 폴리아데닐화, 염색질 구조, RNA 결합, DNA 결합 등이 존재한다. 인간 게놈은 매우 복잡한 조직과 기관을 만들 수 있는 정보를 포함하고 있음에도 불구하고 단지 20,000여개 유전자로 구성되어 있으며 이들의 기능을 조절하고 보완하기 위한 microRNA와 전사 조절 인자와 같은 주요 기능과 상호 기능하는 영역을 추가로 가지고 있다. 이러한 복잡성이 가능한 이유는 하나의 유전자가 다양한 방법으로 엑손/인트론을 선택적으로 스플라이싱하여 조직 특이적인 세포 상황에 따라 전사체의 구조를 변형시키는 스플라이싱 조절(splicing regulation) 과정이나 DNA와 pre-mRNA와 같은 수많은 게놈 인자 사이의 복잡한 상호작용 때문이다. 이들 조절 영역에는 트랜스 조절 단백질들이 특정 DNA 서열과 상호작용하여 유전자의 발현을 조절한다 [10]. 이로 인해 특정 유전자가 특정 단백질 하나만을 만들어 내는 것에 그치지 않고 다양한 선택적 스플라이싱을 통해 엑손이 재구성되어 다양한 형태의 단백질을 만들어 낼 수 있다. 이를 통해 제한된 20,000여개의 유전자를 가지고 있지만 아주 다양한 단백질 레퍼토리(protein repertoire)로 확장할 수 있게 된다. 통계적으로 단백질 코딩 유전자 마다 평균 4개 이상의 전사체를 만들어 낸다고 알려져 있다[3]. 2개 이상의 엑손을 가진 유전자들의 95%는 선택적 스플라이싱을 한다는 증거들이 발견되기도 했다[11]. 뿐만 아니라 많은 질병에서 유전적 변이에 의한 스플라이싱 조절이 이루어진다는 것이 발견되었다. 이러한 현상은 척추동물들에서 더 자주 발견되고 있으며 복잡한 세포 타입일수록 더 흔히, 특히 인간의 뇌에서 가장 복잡한 선택적 스플라이싱 현상이 존재한다는 것이 발견되었다[12]. 심각한 정신질환 환자의 뇌에서 특히 비정상 스플라이싱 현상이 발견되고 있으며 자폐증 환자들의 대뇌피질에서 선택적 스플라이싱 패턴이 일관성 있게 발견돼 일부 자폐증 메커니즘으로 스플라이싱 조절의 실패로 일어나는 현상이라는 것을 알게 되었다 [13]. 또한 질병 유전적 변이 발생으로 인한 유전적 질환의 60%는 엑손/인트론 스플라이싱 과정의 결함과 관련되어 있다고 예상한다[14].

특히 이러한 분석은 RNA-Seq 데이터를 사용해 수십만 개의 엑손 들이 특정 세포 타입에서 어떻게 발현되는지 측정할 수 있게 됨으로써 이들 데이터들은 스플라이싱을 조절하는 모델을 정확하게 예측하기 위한 계산 모델을 훈련하기 위한 목적으로 사용될 수 있다. 최근 딥지노믹스와 같은 회사의 경우는 엑손/인트론 스플라이싱 메커니즘 관련해 몇 가지 통찰을 기반으로 스플라이싱 추론 모델을 디자인하고 데이터 세트로 트레이닝함으로써 게놈에서 발견되는 변이들의 질병에 대한 기여도를 정확하게 예측할 수 있게 되었다고 한다. 스플라이싱의 계산 모델은 스플라이싱이 일어나는 게놈 영역 근처의 특징들을 추출하고 mRNA 데이터로부터는 엑손이 유지되거나 배제되는 빈도를 예측함으로써 가능해졌다[9]. 추론 모델을 트레이닝하기 위해, 여러 다양한 조직/세포 내 존재하는 다양한 전사체의 패턴을 RNA-Seq 기술을 이용해 생산된 데이터를 기반으로 분석하였다. 엑손들의 경계를 포함한 접합부위(junction site)에 맵핑된 리드들을 카운팅함으로써 mRNA에서 특정 엑손이 얼만큼의 빈도로 포함/배제 되었는지 정량적으로 측정 가능해졌고 특이적인 유전 변이를 분석하기 위해, 정상 DNA 서열과 변이가 발생한 DNA 서열들은 계산 모델의 입력 값으로 사용해 스플라이싱에 있어 정량적으로 변화를 주는 변이들을 찾아내는 방식이다. 이 접근법은 척수성 근위축증과 유전적 비용종성 대장암과 자폐증 환자들을 포함한 다양한 질환에 있어 임상적으로 잘 알려진 변이들과 기존 연구를 통해 알려지지 않은 새로운 변이들을 정확하게 예측할 수 있게 되었다. 실험검증으로 이들 예측이 꽤 잘 맞아떨어진다는 것이 확인하였으며 척수성 근위축증 환자들의 아주 드문 변이들을 분석해 중추신경계(central nervous system)발달, 시냅스 전달에 관련된 19개의 비정상 스플라이싱된 유전자들을 모두 발견하여 보고하였다[9].

또 다른 변수로는 단백질 서열의 상호작용에 의한 DNA 또는 RNA와 단백질 사이의 화학적 결합이 있다. 이들 상호작용은 세포 내 핵심이 되는 많은 과정에 영향을 주기 때문에 이들을 정확하게 모델링 하는 것은 매우 중요하다. 단백질 서열에 결합하는 모델은 유전체를 해석하고 유전변이의 효과를 예측하기 위해 매우 필수적이다. 기계학습은 이러한 생명현상을 이해하는 역할을 할 수가 있을 것으로 보인다.

인간 게놈은 적어도 1,400개 이상의 DNA에 결합하는 단백질들과 1,500개 이상의 RNA에 결합하는 단백질들을 코딩하고 있고 이러한 결합 단백질은 가장 큰 단백질 카테고리를 형성하고 있다. DNA에 결합하는 단백질들은 그들이 특정 유전자에 결합하여 RNA가 전사되어 생성되는 속도에 영향을 주며 그들을 전사인자(transcription factors, TFs)라고 부른다. 생물학자들은 개별 단백질들의 서열 특이성을 측정하는 대용량 실험 방법을 개발했으며 시퀀싱 기술을 통해 실제로 생체 내에서 결합하는 패턴을 분석할 수 있게 되었다. 또한 마이크로 어레이를 통해 미리 합성해 놓은 DNA 조각(40,000~250,000 프로브)에 결합하는 단백질 시그널 데이터셋들은 이미 공개되어 있어 이들을 딥러닝 기술을 이용해 패턴을 분석해 낼 수가 있다[15]. 초창기에는 하나 또는 두 개 정도는 다르더라도 허용하는 형태로 "공통염기서열(consensus sequence)"이라는 개념으로 하나의 패턴으로 지정해 여기에 단백질이 결합한다고 가정하고 결합 사이트를 모델링 하는 방법이 있었으나 현재는 서열의 여러 다양성을 허용하는 position frequency matrix(PFM)을 적용하고 발견한 특정 패턴들을 모아 서열 하나하나에 가중치를 줌으로써 그들 패턴을 프로파일링하고 "sequence logo"화 시켜 우리가 찾고자 하는 단백질 결합 영역을 게놈 내에서 스크리닝해 후보 결합 영역을 발견하는 방식이 주로 사용된다[16].

4. 전산 생물학 분야 기계학습 적용

최근 몇 년 사이 기계학습 연구자들은 음성인식과 시각처리에 집중하고 있다. 시각화는 그 자체가 직관적이기 때문에 기계학습 역사상 가장 오랜 기간 연구되어 온 분야이다. 인류는 오랜 시간 동안 시각처리에서 매우 만족스러운 결과를 만들어 왔고 그렇지 않은 경우 오랜 기간 노력해 새로운 통찰로 이를 극복해 왔다. 이처럼 기계학습 과학자들이 흔하게 접할 수 있는 음성인식 자료와 시각 자료를 통해서 경험을 많이 할 수 있었던 반면 생물학 데이터의 경우 쉽게 접근할 수가 없으며 단순한 효모와 같은 단일 세포 생명체에 대해서도 메커니즘을 전혀 이해할 수가 없었다. 유전자형-표현형 관계는 아마도 "ImageNet"과 같은 고감도 시각화 대회에서 수행하는 것보다도 더 복잡한 문제를 가지고 있는 것처럼 보이며 이는 세포 내 많은 상호작용, 정량 그리고 세포 형성 과정들과 같은 복잡한 현상이 존재하고 이들은 드러나지 않은 채 "숨겨진 변수(hidden variable)"로 존재하고 있기 때문일 것이다. 또한 이를 시스테믹하게 분석하는 기술이 존재하지 않았기 때문에 우리가 이들을 관찰할 수가 없었다. 그러나 최근 생물학에서도 대량 실험 기술이 다양하게 적용됨으로 인해 도출된 데이터를 이용한 기계학습 기술과 딥러닝 기술이 적용되어 주요 성과를 낼 수 있을 것으로 기대된다.

유전체 의학에 있어 세포 변수 접근법은 실질적으로 가능한 "in silico" 예측으로써 매우 중요하게 질병 메커니즘을 이해하는 통찰을 제공할 것으로 보인다. 전사체의 구조를 변화시키는 스플라이싱의 경우를 다시 예로 들자면, 스플라이싱과 단백질 서열 결합의 주요 세포 변수들은 유전적 변이로부터 어떤 질병의 위험을 확인하는데 있어 매우 유용하다.

이 때 세포 변수들은 개별이 아닌 동시에 적용돼야 한다. 예를 들면 mRNA 전사는 스플라이싱과 함께 움직이고 있어 하나의 세포 변수는 또 다른 변수의 예측을 개선하는데 적용될 수가 있을 것으로 보인다.

기계학습 추론 모델로 고려해야 할 다양한 세포 변수들
- 유전체내 기능 영역
- 전사 조절을 위한 결합 영역
- 스플라이싱 패턴
- 절단 위치(cleavage site) 및 폴리아데닐레이션(polyadenylation) 사이트
- RNA 구조
- 단백질 구조

5. 기계 학습에 사용할 수 있는 유전형-표현형 모델링 공개 데이터베이스들

전 세계 생명과학 연구자들의 노력으로 대규모 유전형 데이터에서부터 암과 같은 특정 표현형에 이르기까지 다양한 수준의 생물학적 시스템을 측정하기 위한 대규모 데이터가 축적되었으며 특히 유전체, 전사체, 후성 유전체, 그리고 단백질체와 같은 오믹스 데이터들이 임상 정보를 포함해 데이터베이스화되어 공개되었다. 이러한 데이터 리소스를 잘 사용하면 단일 데이터 리소스에서는 확인할 수 없는 정보를 상호 보완할 수 있으며 유전자형과 표현형 간의 엄청난 격차를 다소 해소하여 보다 정확한 생물학적 모델들을 만들 수 있을 것이다.

이용 가능한 오믹스 데이터베이스들[7]
- GTEx (Genotype Tissue Expression): 유전형(SNP칩, 전장엑솜, 전장게놈), 전사체(RNA-Seq), 표현형(포괄적인 표현형 정보와 임상 정보)
- NCI-60 (National Cancer Institute Anticancer Drug Screen): 유전형(전장엑솜), 전사체(mRNA칩, miRNA칩), 단백질체(SWATH 프로파일), 표현형(암세포주 및 약물처리 정보)
- ENCODE (Encyclopedia of DNA Elements): 유전형(세포주의 전장게놈), 전사체(RNA-Seq), 후성유전체(ChIP-seq, DNase-seq, 5C:Chromatin Conformation Capture Carbon Copy)

- ICGC (International Cancer Genome Consortium): 유전형(암 전장게놈), 표현형(병리 및 임상 정보)
- TCGA (The Cancer Genome Atlas): 유전형(암 전장 게놈 및 전장 엑솜), 전사체(RNA-Seq, miRNA-Seq), 후성유전체(methyl-Seq), 단백질체(역상 단백질칩: reverse phase protein array), 표현형(병리/기초/임상 정보, 약물 정보)
- 1000 Genome Project: 유전형(전장게놈), 전사체(RNA-Seq), 표현형(가계 및 조상정보)
- NIH Roadmap Epigenomics Proejct: 유전형(전장게놈), 전사체(RNA-Seq, small RNA-Seq), 후성유전체(ChIP-Seq), 표현형(수십 종류의 세포주 및 다양한 배양세포)
- GIANT (Genetic Investigation of Anthropometric Traits): 유전형(SNP칩), 표현형(신체크기 및 비만측정)

결 론

게놈 생물학, 게놈 의학 그리고 정밀 의학에 있어 기계학습의 역할은 앞으로 몇 년 이내 더 급속히 성장할 것으로 예상된다. 복잡한 대규모 데이터 세트에서도 효과적으로 학습할 수 있는 기술을 개발하는 것을 목표로 딥러닝 커뮤니티에서도 노력하고 있다. 유전형과 표현형 사이에는 수많은 생물물리학적 과정의 계층과 상호작용이 연관되어 있고 우리는 아직 이들의 대부분을 완벽하게 이해 하지 못한 상황이지만 강력한 컴퓨터 기술의 발달과 함께 곧 이러한 복잡한 생물학적 상호작용들을 모델링 할 수 있기를 기대하고 있다. 또한 기계 학습과 함께 최근 급속히 발전하는 딥러닝 기술은 이미지인식, 음성인식 그리고 자연어처리에 있어 인간 수준의 성능을 발휘하거나 넘어서고 있지만 게놈은 인간의 인지능력으로 해석 가능한 범위를 넘어서 게놈 생물학에서는 이 정도 수준의 접근으로는 이해가 쉽지 않은 실정이다.

즉, 인간은 오랜 진화적 선택압을 통해 어떤 생명체 보다 강력한 인식력, 해석력 그리고 반응 능력을 보유하게 되었지만, 게놈을 해석할 수 있는 능력을 개발하기 위한 선택압을 받은 적이 없기에 이 복잡한 게놈을 해석할 수 있는 능력을 획득하지는 못했다. 결과적으로 가장 최근의 생물학적 지식과 학습 알고리즘을 적용해 사람의 눈으로도 확인할 수 없는 이런 문제들을 기존과는 다른 방식으로 조심스럽게 그리고 빠르게 검증해 나가야 한다. 어떤 질병과 게놈과의 연관성은 매우 복잡해 실제 숫자들을 입력한다고 해서 모델링 되어 해결될 가능성은 매우 낮으므로 이미지와 음성 인식과는 대조적으로 주어진 입력 값들이 어떤 것을 예측할 수가 있는지 정확하게 하나하나 알아가야 한다.

지금 당장은 이런 전산적 접근 방법이 기존의 실험 그리고 임상진단을 완전히 대체할 수는 없겠지만, 특정 가설을 확인하기 위해 필요한 방법들을 검증하는데 걸리는 시간을 대폭 줄여 나갈 수 있다.

또한 실제 세포 변수들을 측정하는 것은 환자의 표현형을 관찰하는 것보다 훨씬 힘든 일이다. 많은 수의 환자로부터 개인 별 몇 개의 변수들을 측정하는 것 보다 소수의 환자 그룹으로부터 수 십 만개의 세포 변수를 측정하는 것이 훨씬 더 좋다. 이를 처리하기 위한 알고리즘은 1,000개 이상의 노드(16,000개 이상의 CPU)를 가진 분산형 병렬 처리 컴퓨터에서 가동 가능했으나 대규모 기계 학습에 적합한 GPU 클러스터가 나와 대규모 기계 학습을 빠르고 경제적으로 수행할 수 있게 되었다.

앞으로의 게놈 빅데이터 시대에서는 기계학습 연구 프로젝트는 초기 단계에서부터 확장성, 이동성, 재현성을 염두에 두고 설계되어야 할 것이다. 또한 게놈과 질병 위험도 데이터에 앞으로는 세포 수준의 분자표현형 데이터를 더하여 게놈의 영향을 명시적으로 모델링 해야 한다. 즉, 궁극적으로 게놈 시퀀싱과 같이 싸고/빠르고/비침습적인 측정 기술을 이용해 이들 변수를 어떻게든 정확하게 측정해 실험적 방법을 통한 비싸고/느리고/침습적인 변수들을 모두 대체하여 정확하게 예측할 수 있어야 할 것이다.

REFERENCES

1. Watson JD and Crick FH. Molecular structure of nucleic aids: A structure for deoxyribose nucleic acid. Nature 1953;171:737–8.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
3. de Klerk E, 't Hoen PA. Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing. Trends Gen 2015;31:128–39.
4. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res 2012;22:1760–74.
5. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 2011;478:476–82.
6. Moffat JG, Rudolph J, and Bailey D. Phenotypic screening in cancer drug discovery-past, present and future. Nature Rev Drug Discov 2014;13:588–602.
7. Leung MK, Delong A, Alipanahi B, and Frey BJ. Machine learning in

genomic medicine: A review of computational problems and data sets. Proc IEEE 2016;104:176–97.
8. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012;22:1813–31.
9. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 2014;347:Epub 2014 Dec 18
10. Singh RK and Cooper TA. Pre-mRNA splicing in disease and therapeutics. Trends Mol Med 2012;18:472–82.
11. Pan Q, Shai O, Lee LJ, Frey BJ, and Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40:1413–5.
12. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science 2012;338:1587–93.
13. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature 2011;474:380–4.
14. López-Bigas N, Audit B, Ouzounis C, Parra G, and Guigó R. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett 2005;579:1900-3.
15. Levo M and Segal E. In pursuit of design principles of regulatory sequences. Nature Rev Genetics 2014;15:453–68.
16. Stormo GD. DNA binding sites: Representation and discovery. Bioinformatics 2000;16:16–23.