딥러닝: 유전체학을 위한 새로운 컴퓨터 모델링 기술

이 주 성

DGIST

E-mail: jsleescgl@dgist.ac.kr

요약문

데이터 주도 과학으로서 유전체학 분야에서는 기계학습을 데이터 특성 파악 및 새로운 생물학적 가설 유도를 위하여 이용되고 있다. 하지만, 급격하게 증가하고 있는 유전체 데이터에서 새로운 의미들을 얻기 위해서는 보다 심층적인 기계학습 모델을 필요로 하는데, 컴퓨터 비젼과 자연어 처리 분야에서 주로 사용되는 딥러닝 기술이 이제는 생물학 분야 중에서도 DNA accessibility와 Splicing과 같은 유전자 조절 기전에 있어서 유전적 변이의 영향을 예측하는 등의 수많은 유전체학 데이터 모델링을 위한 방법으로서 사용되고 있다.

Key Words: Bioinformatics, Genomics, Systems Biology, Deep learning

본 자료는 Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20, 389–403 (2019).의 논문을 한글로 번역, 요약한 자료입니다.

목 차

1. 서론

기능적 혹은 광범위한 개념의 유전체학은 유기체의 유전체적 요소들의 기능을 규명하는데, genome 시퀀싱, 전사체 분석, 단백체학과 같은 유전체 스케일에서의 실험을 통하여 이루어진다. 유전체학은 데이터 주도하는 과학으로서 발전하고 있는데, 이는 기존의 지식을 기반으로 하는 모델이나 가설에 대한 연구보다는 새로운 정보들을 발굴하는 것을 시사한다. 대표적인 예로는 유전자형과 표현형 사이의 관련성, 환자들의 분류를 위한 바이오마커 발굴, 전사 강화인자(enhancer)와 같은 생화학 활성이 이루어진 genome 상의 위치, 유전자들의 기능성 예측 등이 있다.

유전체 데이터는 상당히 규모가 크고 구조가 복잡하기 때문에, 새로운 발견과 밝혀지지 않은 상관성 및 연계성을 찾고 새로운 가설 및 모델을 제시하여 예측하기 위해서는 전문적인 분석 기술 도구가 필요하다. 여러 분석 알고리즘들 중, 가정과 특정 도메인들을 사용자가 직접 설정해야 하는 기법들과는 달리, 기계학습에서는 데이터 내에서의 특징 및 패턴을 자동적으로 발견하기에, 유전체학과 같이 데이터가 주도하는 분야에서 사용하기에 적합하다. 하지만, 기계학습 방법의 단점은 보여지는 데이터 자체에 매우 의존도가 높으며, 특징 추출 과정에서 어떻게 특징들을 추출하는지, 어떤 특징들이 기계학습을 이용한 예측에 사용되는지에 따라 성능이 감소될 수 있다. 위와 같은 한계점을 딥러닝에서는 DNN (Deep Neural Network)의 발전을 통하여 해결하였다. DNN은 수행과정 중에 얻게 되는 중간결과들을 입력값으로 다시 사용하여 이를 컴퓨터가 스스로 반복하여 계산하는 방식인 end-to-end 학습으로 복잡한 특징들을 계산하고 추출하는데 성공적인 결과를 보여주었다.

DNN의 구축과 훈련은 데이터의 폭발적인 증가세와 알고리즘의 발전 그리고 GPU와 같은 컴퓨터 하드웨어 요소의 발전이 기반이 되어 가능하였다. 지난 7년간, DNN은 컴퓨터 비전, 음성 인식 등 컴퓨터 과학 분야에서 주로 사용되었는데, 2015년 DNA 염기서열 데이터에 처음으로 적용된 연구가 발표되면서부터, DNN 및 딥러닝 기술을 유전체학 분야에 적용한 수많은 논문들이 게재되었으며, 동시에 딥러닝 분야의 연구자들은 위의 기술들의 성능 향상과 모델링 기술 레퍼토리의 확대에 힘쓰고 있어, 이들 중 일부 기술들은 이미 유전체 연구분야에서 영향력을 보이고 있다.

본 리뷰에서 우리는 딥러닝을 이용한 모델링 기술과 유전체 데이터로의 적용에 대하여 우선하여, 지도적 학습을 위한 4가지의 주요 모델들과 함께 이 방법들이 어떻게 유전체 데이터 내에서 패턴을 찾아내는지를 설명할 것이다. 그 다음으로 여러 가지 데이터셋과 데이터 유형들의 통합을 위한 multitask/ multimodal 학습 기술과 기존에 존재하는 모델을 기반으로 빠르게 변형 및 발전 시켜 사용하는 전이 학습 기술, 마지막으로 비지도적 학습을 위하여 사용되는 autoencoder와 단세포 유전체학 분야에서 처음으로 적용된 generative adversarial networks (GANs) 알고리즘에 대하여 소개할 것이다. 유전체학 분야를 넘어서 딥러닝에 대한 보다 깊고 자세한 배경지식들을 위해서는 이 논문과 더불어 컴퓨터 생물학자 및 생물정보학자들을 대상으로 하는 딥러닝 관련 다른 여러 논문들을 함께 읽어보는 것을 권장한다.

2. 지도 학습(Supervised learning)

지도학습의 목적은 특징들을 입력값으로 받아 타겟 변수에 대한 예측치를 결과로 반환하는 모델을 얻는 것이다. 예를 들어, splicing 현상에서, RNA 염기서열들을 특징으로서 사용하고 이를 입력 값으로 하여 특정 intron이 spliced out인지 아닌지를 설계한 모델을 훈련하여 예측하는 것이 지도학습을 통하여 해결할 수 있는 문제 중 하나이다 **(그림 1)**. 여기서 기계학습 모델을 훈련시킨다는 것은 모델을 구성하고 있는 여러 가지 파라미터(옵션값)들을 학습하면서 손실함수(또는 비용함수)를 최소화 시키는 과정을 의미한다

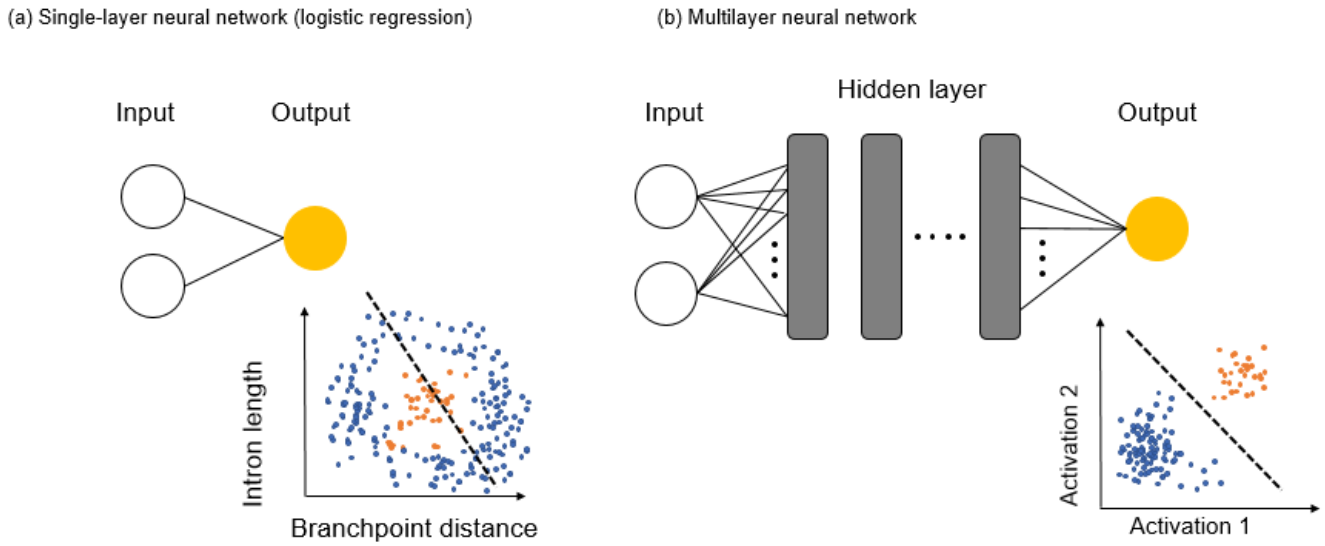

그림 1. 은닉층을 포함된 인공신경망을 사용한 비선형 모델링.

2.1. DNN을 이용한 복잡도 의존성 모델링

컴퓨터 생물학 분야에서의 많은 지도적 학습 문제들의 경우, 입력 데이터를 행렬 형식으로 받는데, 데이터가 생성될 때부터 행렬 형태로 주어지는 경우도 있지만, 전처리 과정을 통하여 행렬화시켜야 하는 경우도 있다(DNA 염기서열의 K-mer 개수로의 형태 변환). Intron-Splicing 예측 문제에서는, intron 길이와 splicing branchpoint 위치가 전처리된 특징들로써 행렬 데이터에 수집된다. 이와 같이 행렬(표) 형식의 데이터는 단순한 선형 모델(예: 로지스틱 회귀)부터 보다 일반적이고 유연한 모델인 비선형 모델들(예: 신경망, 비선형 회귀)에서 표준적으로 사용되는 데이터 포맷이다. 위와 같은 입력 데이터 형태(표 형식)로 Intron-Splicing 문제에 대하여 단순 선형 회귀 분석법인 로지스틱 회귀 분석(이진 분류기)과 DNN (비선형 지도학습법)을 적용하여 예측 결과를 통한 성능을 비교해본 결과, 단순히 입력 값에 있는 특징들의 가중치를 기반으로 한 가중치 합만으로는 intron spliced out 여부에 대한 예측을 정확히 할 수 없는데 반하여, DNN (다중층 신경망)을 이용한 경우 명료하게 분류해낼 수 있음 보여준다 **(그림 1)**. **그림 1B**에서 보인 신경망을 이용한 예측분석은 여러 은닉층

(hidden layer)들을 사용하여 자동적인 비선형적 특성들을 학습하여 단순 선형회귀법(로지스틱 회귀)보다 좋은 성능을 보였는데, 이때 설계된 여러 은닉층들을 다수의 선형 모델들이라고 생각할 수 있으며, 이들의 결과값들이 비선형 활성함수(예, Sigmoid, ReLu 함수)에 의하여 변형된다. 즉, 이러한 은닉층들은 입력 특성들을 통해 데이터의 복잡한 패턴들을 파악해 분류 문제를 해결하는데 활용된다. DNN에서는 많은 은닉층들을 활용하고 은닉층의 각 뉴런이 이전 은닉층의 모든 뉴런으로부터 입력을 받는 구조를 "fully connected"라고 한다. 신경망은 일반적으로 "stochastic gradient descent"라는 알고리즘을 모델 트레이닝에 사용하는데, 이는 데이터셋의 규모가 충분히 큰 경우에 많이 파라미터 학습법이다. Fully connected 신경망(Fully connected neural network; FCNN)은 이미 유전체학 분야에서 많이 사용되어지고 있는데, splicing 현상에서 특정 exon이 spliced-in 될 확률, 주어진 질병 유발 유전적 변이의 우선순위 예측, 주어진 genomics region 내에서의 cis 조절인자 예측이 대표적인 적용사례이다. 물론 FCNN 방법이 항상 좋은 성능을 보이는 것은 아니다, 전통적인 기계학습 알고리즘인 random forest 혹은 선형회귀보다 성능이 좋다는 보고가 많이 따르지만, 어떠한 경우는 gradient-boosted decision tree가 FCNN의 성능을 앞선다는 결과가 Kaggle 대회에서 보인 바 있다. 그럼에도 FCNN은 딥러닝 툴박스로서 매우 중요한 요소이며, 다른 신경망 층(예: convolutional layer)과도 효율적으로 병합되어 사용되기도 한다.

2.2. 컨볼루션을 통한 Sequential 데이터 내 국소 패턴 발굴(CNN in genomics)

데이터 유형 특이적 패턴 파악(데이터 특이적 국소 패턴 분석)은 효율적인 예측을 위해서 중요한데, 단일 유형이 아닌 다양한 형태의 데이터가 존재하는데 이들의 데이터가 무작위로 섞인 상태로 행렬화되어 FCNN을 적용한다면 데이터 내의 유의미한 패턴을 파악하기 어려울 것이다. 유전체학 분야의 한 가지 예를 들자면, 특정 genomic region이 주어질 때, 해당 영역이 전사인자가 붙는 구간인지 아닌지를 예측하는 문제를 들 수 있다. 데이터는 ChIP-seq을 통해 얻어진 높은 신뢰도를 가진 특정 genomic region에 대한 binding event 데이터 한 가지와 전사인자가 붙는 구간의 시퀀스(binding motif)이다. Binding event는 숫자형 데이터이고, Binding motif는 시퀀스 데이터이다. K-mer instance의 개수를 세거나, 해당 시퀀스 내에서 position weight matrix (PWM)을 사용한다면 가능할 수도 있겠지만, 전사인자가 여러 개의 motif 조합에 따른 영향을 받는다는 문제와 과적합(overfitting)이라는 한계점이 있다.

위의 문제의 해결법 중 한 가지가 컨볼루션을 이용하는 것이다. **그림 2**에서처럼 컨볼루션 층은 데이터의 국소영역에 적용되어 패턴을 분석한다. 이러한 접근법은 다수의 PWM을 사용하여 염기서열을 스캐닝하는 것으로 보여지기도 한다. **그림 2B**에서처럼 컨볼루션 층에서는 전체 염기서열에 걸쳐 여러 개의 필터를 이용하여 스캐닝을 수행하고, 비선형 활성 함수(예: ReLU)를 필터링한 결과값에 적용시킨다. 활성함수를 통과한 결과값들은 다음으로 "풀링(Pooling)" 연산을 거치게 된다. 풀링 연산은 보통 최대값 혹은 평균값을 이용하여 진행하는데 이 과정을 통하여 데이터 사이즈를 효과적으로 감소시킬 수 있다 (**그림 2D**). 위의 과정을 통하여, 특정 거리범위 내에서 어떠한 전사인자의 binding motif가 존재하는지를 예측할 수 있다. 이와 같이, 컨볼루션 층의 결과값은 FCNN의 입력값으로 다시 반환되어 사용되어 최종 예측 과정을 수행해낼 수 있다 (**그림 2 F-H**). 따라서, 한 개의 딥러닝 모델 안에서 서로 다른 유형의 신경망층(예, 컨볼루션 층 + Fully connected 층)이 혼재되어 좋은 성능을 보일 수 있는데, 유전체학 분야에서는 특히 염색체 구조 관련 분석과 같이 공간적 특성이 반영된 염기서열 데이터 분석 관련하여 적용할 수 있다.

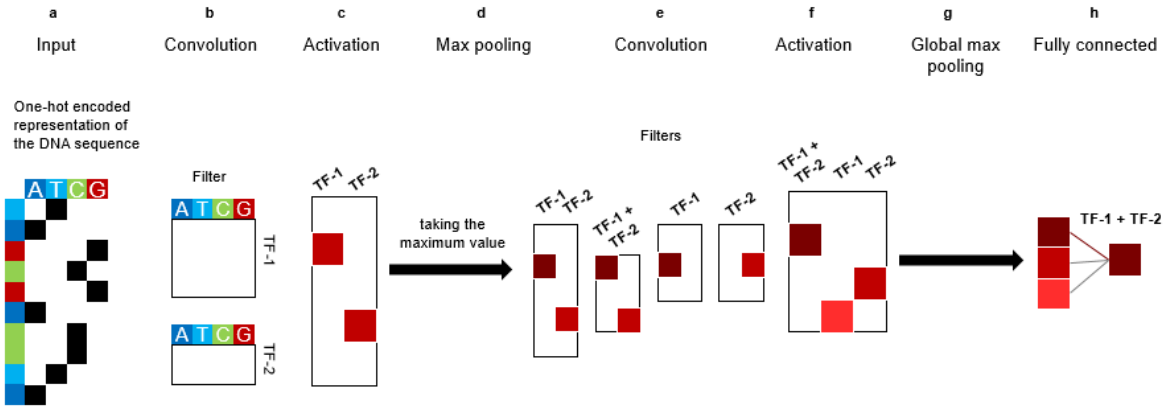

그림 2. 컨볼루션 신경망(CNN)을 이용한 전사인자의 타겟 모델링.

2.3. 순환신경망을 이용한 장거리 염기서열 데이터 모델링(Recurrent Neural Netowkr (RNN) in genomics)

순환신경망은 Sequential 데이터(예, DNA 염기서열, 시계열 데이터)분석을 위해 CNN 외에 사용되는 또 다른 방법이다. 순환신경망 알고리즘은 같은 연산을 각각의 시퀀스 요소에 적용한다 (**그림 3C**). 위의 연산은 두 가지 입력을 받는데, 새로운 입력값과 함께 바로 이전 시퀀스 요소에 대한 "메모리"를 함께 입력 받는다. 여기서 이전의 결과에 대한 기억을 사용한다는 점이 순환신경망이 다른 신경망과 차별성을 보이는 부분이다. 순환신경망에서는 연산과정에서 메모리를 업데이트하고 선택적으로 결과를 방출하여 이어지는 층으로 통과시키거나 바로 모델 예측에 활용될 수 있다. 이론적으로, CNN과 가장 차이를 보이는 메모리 보유라는 기능 때문에, 순환신경망은 긴 시퀀스 데이터 처리에 용이하며, 유전체학 분야에서 적용될 수 있는 사례는, start/ stop codon을 모두 포함하고 있는 염기서열에서 "Open-reading-frame" 구간을 발견해내는 문제에 적용할 수 있다. 이전 시퀀스에 대한 메모리를 갖는 부분이 CNN에 비하여 장점이 될 수도 있지만, 이로 인하여 병렬연산이 용이하지 않다는 점에서 CNN보다 연산 수행속도면에서 느리다는 점이 단점 중 하나이다.

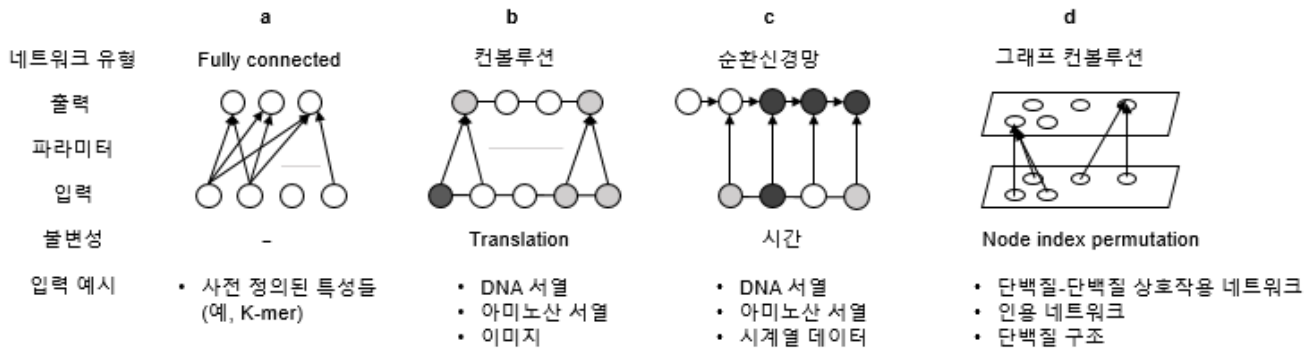

그림 3. 여러 종류의 인공신경망에 대한 구조 및 특징.

유전체학 분야에서는 앞선 FCNN+CNN 사례와 같이 RNN도 다른 구조의 신경망과 병합하여 사용이 가능한데, 예를 들어, 단세포 DNA 메틸레이션 위치 분석, RBP binding 전사인자 binding과 DNA 접근성 분석과 같은 염색체 구조체 분석 등에 순환신경망은 CNN과 병합된 구조의 신경망으로서 사용될 수 있으며, 순환신경망 자체적으로도 miRNA 생물학 분야에서 타켓을 찾은 연구에 적용된 사례가 있다. 또한 DNA 염기서열 데이터 중 raw 데이터에서 base calling 하는 과정에서도 순환신경망에 적용되어지고 있다. 이렇게 많은 사례에 순환신경망 구조가 적용되어지고 있음에도 불구하고, 공통 시퀀스 모델링 수행에 대해서 CNN과의 전반적인 성능 비교에 대한 연구는 여전히 부족한 상황이다.

2.4. 그래프-CNN을 이용한 그래프 구조 데이터 모델링

단백질-단백질 상호작용, 유전자 조절 기전 네트워크와 같은 그래프 구조의 데이터들은 유전체학 분야에서 어느 곳에서나 존재한다. 그래프-CNN(GCN) 네트워크는 각각의 특징들을 그래프 상에서 노드로 사용하고, 기계학습 문제를 풀기 위하여 노드 사이의 연관성을 활용한다. GCN은 연속적으로 각 층에서 이웃하는 여러 노드들의 통합으로 계속해서 새로운 특성들을 만들어내 층마다 그래프 구조를 변화 시켜 나간다. GCN은 훈련과정을 거쳐 노드 분류 및 비지도학습적 노드 임베딩(차원 축소에 활용), 노드 사이 연결선 및 그래프 분류에 활용될 수 있으며, 이와 같은 활용성은 유전체학 분야에서 단백질-단백질 상호작용 예측, 화학물질 혹은 약물후보군의 물리 화학적 성상 파악 등 지도/비지도 학습 모두에 적용이 가능하다. 위와 같은 특성을 기반으로 GCN은 유전체학 분야 내 그래프 구조 관련 데이터 분석에 있어 활용성이 크기 때문에, 추후에 보다 많은 적용 사례가 나올 것으로 기대된다.

2.5. 연산 수행 사이의 정보 공유 및 다중 데이터 통합

유전체 데이터는 종종 생물학적 작용과 연관된 상관성이 있는 정보를 포함하고 있다. 상관성이 있는 측정값들은 단일 데이터와 서로 다른 데이터 사이에서 발생할 수 있다. 한 예로서, 다수의 전사인자들의 binding affinity를 예측하는 문제에 대해서 생각해본다면, 단일 모델에 비하여, 다중 수행 모델을 사용할 때, 여러 전사인자들의 결합을 종합적으로 예측할 수 있다. 이처럼 다중 수행 모델을 사용할 경우, 신경망을 구성하는 대부분의 층들이 공유되며, 마지막에 각 수행 작업별로 층이 분기된다. 모델링된 전사인자의 공동-결합 및 공통 단백질 도메인으로 인하여, 복잡한 시퀀스 특성들이 다수의 전사인자들에 걸쳐서 추출될 수 있으며, 이는 예측 성능을 높이며, 전산인자당 요구되는 데이터의 양이 줄어드는 장점이 있다. 추가적으로, 다중 수행 모델은 연산 과정을 공유하기 때문에 단일 수행 모델보다 예측을 위한 계산과정을 보다 빠르게 마칠 수 있다.

다중 수행 모델에서의 손실 함수는 단순히 각각의 수행과정에서의 손실값의 합을 사용한다. 수행과정마다 손실값이 매우 다른 경우, 가중치 합계를 활용하여 손실의 균형을 조정할 수 있다. 다중 수행 모델은 네트워크가 여러 손실들을 동시에 최적화시켜야 하므로, 훈련이 어려워 트레이드 오프를 만들어야 한다. 위의 문제를 해결하기 위해 다양한 방법들이 제시되었는데, GradNorm이 그 중

하나이다. GradNorm은 훈련과정 중에 가중치를 채택하여 서로 다른 작업에서의 역 전파된 gradient를 같은 크기로 만들어준다. 유전체학에서 다중 수행 모델은 서로 다른 조직에서의 전사인자 결합, 서로 다른 히스톤 표지, DNA 접근성 그리고 유전자 발현정보와 같은 다중 분자적 표현형 예측에 사용되었으며 좋은 성능을 보였다.

다중 수행 모델과 비슷하게, DNN은 다중 유형의 데이터를 입력으로 받아, 이들을 통합하여 상호 보완적인 정보를 이끌어내는데 확장시킬 수 있다. 다중 유형의 데이터들을 통합하는 단순한 방법은 각각의 데이터셋에서의 특징들을 병합시키는 것이다. 이러한 데이터 통합 방식은 각각 데이터 유형의 raw 수준에서는 어려우며, 데이터 특이적으로 전처리 과정을 거친 이후, 통합이 가능한 양식으로 다중 유형의 데이터셋의 입력 포맷을 단일화시킨 뒤 수행이 가능하다. 신경망은 데이터 유형별 전용 층의 출력값을 추가 계층에 사용하여 통합할 수 있다. 이러한 방식을 중간 통합이라고 하는데, 이는 각 데이터에 가장 특이적인 계층을 사용할 수 있으므로 더욱 예측성 있는 특성들을 추출할 수 있다.

2.6. 전이학습을 통합 소규모 데이터 기반 모델 훈련

데이터가 부족한 경우, 처음부터 모델을 훈련시키는 것은 어렵다. 하지만, 비슷한 작업에 대해서 훈련이 완료된 다른 모델 내, 대부분 파라미터를 통하여, 본 데이터 모델링의 초기화 수행이 가능하다. 이러한 방법을 '전이학습'이라고 하는데, 사전 지식을 모델에 통합시킨 것이라고 할 수 있다. 전이학습을 통하여, 무작위 파라미터 초기화를 이용한 모델링보다 빠르게 훈련이 가능하며, 필요로 하는 데이터의 양이 적으며, 일반화가 더 잘 이루어질 수 있다. 생물 분야에서의 대표적인 전이학습 사례로는 ImageNet 경쟁에서 사전에 훈련된 모델을 이용하여 피부 병변을 분류하고 형태학적 특성 규명을 성공적으로 수행하였다. 하지만 중요한 점은 얼마나 많은 옵션 파라미터를 공유시킬 것인지와 작업별로 어떠한 모델을 활용할 것인지에 대한 평가는 여전히 부족하며 추가적인 연구가 필요한 상황이다. 따라서, 다중 수행 학습 및 전이 학습의 잠재성을 파악하기 위해서는 이미 훈련된 기존의 모델들이 최대한 쉽게 공유될 수 있어야 한다. 데이터셋의 크기가 점차 증가하고 예측 모델의 필요성이 점차 증대되고 정확성이 높아짐에 따라, 지난 10년 동안의 데이터와 소프트웨어 공유의 발전처럼 학습된 모델의 공공화가 더욱 강조될 것으로 예상된다.

3. 예측 설명

딥러닝 자체는 데이터의 해석 혹은 mechanistic hypotheses의 수식화를 디자인하는데 특화된 방법은 아니지만, 위의 목적으로도 사후 활용이 가능하다. 위의 과정들을 모델 해석이라 일컬을 수 있다. 모델 해석의 간단한 예시로, 선형 회귀 모델에 사용되는 파라미터들을 사용하여 입력으로 받는 특징들이 예측 결과에 얼마나 영향을 미치는지를 평가할 수 있다.

3.1. 특성 중요도 파악을 통한 입출력 관계 파악

복잡한 모델에서 모델 파라미터를 간접적으로 확인하는 것은 입출력 관계를 분석하는데 필수적이다. 특성의 중요도 점수(기여 점수)는 위의 관계성 파악 문제에 사용되는데, 이러한 특성들 중 모델 예측에 가장 영향력 있는 (기여 점수가 높은) 특성들은 어떠한 근거로 예측이 이루어졌는지 해석하는 과정에 도움을 준다.

특성의 중요도 분석에는 중요도 점수가 입력 변동성에 의한 것인지 backpropagation에 따라 계산되는지에 따라서 두 가지 카테고리로 분류된다. 입력 변동성에 근거한 방법은 전반적으로 입력 특성값에 변화를 주어 결과값의 관측하는 것이다. 예를 들어, DNA 염기서열 기반 모델에서는 단일 염기서열의 치환이나 regulatory motif의 삽입 등이 이에 해당된다. 하지만 이 방법의 치명적인 단점은 computational cost가 매우 높다는 것이다. 반면, backpropagation 기반 방법은 computationally efficient 하며, 신경망에서 backpropagation을 통하여 모든 입력 특성들에 대한 중요도 점수가 계산된다. 가장 단순한 backpropagation 기반 중요도 점수 측정법에는 saliency maps, input-masked gradients 과 같은 방법이 있다.

3.2. 염기서열 Motif 발굴

Motif 발굴은 regulatory DNA 염기서열 분석에 있어 필수적인 생물정보학 분석 파이프라인의 구성요소이다. Motif 예측 모델에서 특성들의 중요도 점수 측정의 경우, motif가 입력 염기서열 중 항상 같은 위치에서만 존재하지 않을 것이므로, 단순히 중요도 점수들을 평균화하는 것만으로는 사용자가 원하는 결과를 얻을 수 없을 것이다. 이러한 한계점 때문에, 많은 사전연구들은 훈련 데이터에서 컨볼루션 층에서 강하게 활성화된 필터들이 혹은 motif들로 직접적으로 해석되는 필터들을 종합적으로 모아서 motif 염기서열을 도출해낸다. 최신 방법 중에서는 중요도 점수들을 모아서 사용하는 TF-MoDISco라는 tool이 개발되었는데, TF-MoDISco는 plain 염기서열에만 의지하는 이전의 motif 발굴 방법과는 다르게, 특성들의 중요도 점수에 기인하여 중요한 위치를 강조하는 예측 모델에 의존한다.

3.3. 해석 가능한 파라미터와 활성함수를 포함한 인공신경망

내부 신경망 활성화의 해석 가능성을 높이기 위해 최근 "DCell"이라는 가시적 신경망이 제안되었다. DCell의 모델 구조는 세포 내 분자적 하위시스템의 계층 구조에 해당된다. 신경망에서 노드들은 신호전달 경로 혹은 큰 단백질 복합체와 같은 분자적 하위 시스템에 해당하며 두 노드 사이의 연결은 상위 시스템이 하위 시스템의 일부분인 경우에만 허용된다. 신경망에서의 뉴런들은 알려진 개념에 해당하기 때문에 활성화와 파라미터들이 해석될 수 있다. 위의 접근법이 향후 질병과 같이 더욱 복잡한 표현형을 예측하고 이해하기 위해 모듈식 모델링 방법과 결합할 수 있는지를 확인하는 것도 흥미로울 것이다.

4. 비지도적 학습

비지도학습의 목적은 사전 정보가 없는 데이터 내에 존재하는 유용한 특성들을 추출하고 학습하여 해당 데이터의 특징을 파악하는 것이다. 전형적인 비지도적 기계학습 방법에는 클러스터링(e.g., K-means clustering) 또는 차원 축소법(PCA, t-SNE, LDA)이 있는데, 딥러닝 알고리즘인 인공신경망은 위의 방법들을 일반화할 수 있다. 예를 들어, 오토인코더는 은닉층(hidden layer)에서 데이터를 저차원으로 임베딩 시킨 다음 다시 원형으로 복구하는 과정을 거치는데, 복원된 데이터는 복원과정에서 불필요한 변이를 자동적으로 거르기 때문에 denoising 된다고 해석할 수 있다.

오토인코더는 microarray와 bulk RNA-seq 데이터에서 missing data의 추론과 유전자 발현 패턴 파악 및 outlier를 확인하기 위해서도 사용되는데, 이와 비슷하게 single cell data에서도 sparse한 데이터를 보강시키기 위하여 imputation을 위해서도 사용되며, 차원 축소 및 클러스터링 개선에 도움을 준다.

앞서 언급된 데이터 노이즈 제거 및 보강의 목적 이외, 인공신경망은 생성모델로서 활용되어 데이터 생산 과정을 학습할 수도 있는데, 대표적인 알고리즘으로 Variational autoencoder (VAE)와 generative adversarial network (GAN)이 있다. VAE는 추가적인 분포가 전제되는 오토인코더로서, 새로운 랜덤 데이터를 생산하는데, 단일세포 및 벌크 RNA-seq 데이터에서 유의미한 확률적 잠재 변수(latent representation)를 발굴하는데 사용될 수 있으며, 이 특성들을 활용해 세포 유형 특이적 및 실험 조건에 따른 단일세포 내의 변화를 예측하는데 활용될 수 있다. 단, VAE를 포함한 다른 여러 가지 인공신경망 모델의 성능은 사용자가 조절하는 hyperparameter들에 의하여 크게 좌우된다.

GAN은 판별기(discriminator)와 생성기(generator), 이렇게 두 가지 신경망으로 구성된 생성모델이다. 이 두 종류의 신경망은 함께 학습되어, 생성기는 현실에 가까운 데이터들을 만들어내고, 판별기는 데이터가 생성기에 의하여 생성된 것인지 실제 존재하는 데이터였는지를 구별해낸다. 다른 알고리즘들에 비하여 GAN은 현재로서는 유전체학 분야에 있어서 사용이 다소 제한적이다. 현재 유전체 분야에서 GAN이 적용되는 사례는 단백질을 만들어내는 DNA 염기서열 생성과 단백질이 결합하는 microarray에 사용되는 DNA probe를 디자인하는데 도움을 줄 수 있다. 단일세포 유전체 분야에서는, GAN을 통하여 단일세포 전사체 데이터 시뮬레이션 및 차원 축소에도 사용되며, 변동(perturbation)을 통한 GAN의 내부 시스템을 해석할 수 있다.

5. 유전체학에서의 딥러닝의 영향

지도적/비지도적 딥러닝 알고리즘들은 유전체 분야에서 다양한 부분에 적용되어왔는데, 본 리뷰에서는 현재와 가까운 미래에 크게 영향을 미칠 것으로 기대되는 세 가지 분야를 소개한다.

1) Non-coding 변이들의 영향력 예측: 복잡한 표현형에 대한 GWAS 연구결과들에 따르면, Non-coding 영역에서 발생하는 변이들이 주를 이루는데, Non-coding 영역에서 발생하는 변이들은 표현형에 미치는 영향력을 예측하기에 어려움이 있다. 따라서 염기서열 기반의 딥러닝 모델은 위와 같은 변이가 복잡한 표현형에서 어떻게 잠재적인 드라이버 역할을 하게 되는지를 예측하는데 유망한 접근법을 제시할 수 있다.

2) 온전히 데이터에 기반한 생물정보학 Tool로의 개선.

3) Richer representations to reveal the structure of high-dimensional data: 예측 모델을 넘어서 비지도적 목적으로의 딥러닝 모델은 여러 중요한 적용사례를 낳았는데, 예를 들어, 오토인코더는 다른 비선형 알고리즘과는 달리, parametric한 성질 때문에 훈련 데이터 셋과 비슷한 분포를 가진 unseen 데이터에도 적용할 수 있다. 또한, 단일세포 유전체 분야에서는 단일 세포 유형을 정의하고 단일세포 전사체 데이터셋의 상태를 분석할 수 있으며, 서로 다른 기관으로부터 생성된 데이터셋을 통합할 때도 사용될 수 있다.

6. 결론 및 향후 방향

유전체학 분야에 딥러닝이 도입되면서 과학적, 경제적 효과가 함께 발생하게 되었는데, 특히 pharmacogenomics 분야에서 유전체에서의 새로운 regulatory 변이를 효율적이면서도 자동화된 방법으로 찾을 수 있으며, 약물 반응과 후성유전체 데이터를 기반으로 한 타겟 물질 발굴에 큰 이점이 있다.

이러한 성과를 얻게 된 딥러닝의 강점은 크게 3가지로 요약될 수 있는데, 첫 번째 요인으로는 end-to-end learning으로서 데이터 처리 과정 및 분석에 소요되는 시간을 단축시킬 수 있다. 두 번째로는 multi-modal 데이터 처리에 용이하다. 과거와 달리, NGS 기술이 발전을 거듭함에 따라 transcriptome, epigenome, proteome 등 다양한 유형의 omics 데이터들이 무수히 쏟아져 나오는 상태에서 multi-modal 데이터 처리는 매우 중요한 과제인데, 딥러닝 기술의 접목을 통해서 데이터 통합과 이에 따른 분석에 이점이 발생한다.

마지막으로 분석 알고리즘에 사용되는 세부적인 수학 내용에 대하여 연구자들의 부담을 덜어줄 수 있다. 유전체학에 종사하는 연구자들의 경우, 알고리즘에 대한 이론적 지식과 통계 및 기계학습에 대한 배경지식을 통하여 데이터 분석을 위한 모델링을 하기에 시간이 충분치 않은데, 계속 발전하고 있는 딥러닝 모델링 프레임은 연구자로 하여금 위의 과정에 소요되는 시간을 단축시켜, 실무분석에 부담을 덜어주는 효과가 있다.

향후 딥러닝의 유전체분야로의 적용 가능성은 매우 유망한데, 떠오르는 문제점 중 하나는 데이터가 개인의 유전체 정보를 담고 있다는 점이다. 개인의 유전체 정보 때문에 개인정보보호의 이슈가 있기에 데이터 전송은 어렵다는 문제가 있어 한 가지 대안으로는 "federated learning"이 있다. Federated learning을 통해서 총 데이터 훈련 시간을 단축시킬 수 있으며, 의료 데이터 보호를 동시에 할 수 있다. 또 다른 대안으로는 데이터 생성모델을 통하여 사람의 유전체 데이터를 시뮬레이션하는 것인데 이는 개인 정보보호에 따른 문제를 걱정하지 않아도 된다는 장점이 있다. 이처럼 지속적인 딥러닝 모델 프레임의 발전이 지속된다면, 딥러닝은 계속적으로 유전체 정보 분석에 있어 "everyday tool"로 자리 잡을 수 있을 것으로 전망된다.

이주성(2021). 딥러닝: 유전체학을 위한 새로운 컴퓨터 모델링 기술. BRIC View 2021-R03
Available from https:// www.ibric.org/myboard/read.php?Board=report&id=3685 (Jan. 21, 2021)

Email: member@ibric.org