인지과학, 제28권 제4호
Korean Journal of Cognitive Science
2017, Vol. 28, No. 4, 299～314.
http://dx.doi.org/10.19066/cogsci.2017.28.4.005

약물-표적 단백질 연관관계 예측모델을 위한 쌍 기반 뉴럴네트워크*

이 문 환　　김 응 희　　김 홍 기†

서울대학교 의생명지식공학연구실

In-silico 기반의 약물-표적 단백질 연관관계 예측은 신약 탐색 단계에서 매우 중요하다. 그러나 기존의 예측모델은 입력 값이 고정적이며 표적 단백질의 특질 값이 가공된 데이터로 한정됨으로써 예측 모델의 확장성과 유연성이 부족하다. 본 논문에서는 약물-표적 단백질 연관관계를 예측하는 확장 가능한 형태의 머신러닝 모델을 소개한다. 확장 가능한 머신러닝 모델의 핵심 아이디어는 쌍기반의 뉴럴 네트워크로써, 약물과 단백질의 미가공 데이터를 사용하여 특질을 추출하고 특질 값을 각각의 뉴럴 네트워크 레이어에 입력한다. 이 방법은 추가적인 지식없이 자동적으로 약물과 단백질의 특질을 추출한다. 또한 쌍기반 레이어는 특질 값을 풍부한 저차원의 벡터로 향상 시킴으로써 입력 값의 차이로 인한 편향 학습을 방지한다. PubChem BioAssay(PCBA) 데이터 셋에 기반한 5-폴드 교차 검증법을 통하여 제안한 모델의 성능을 평가했으며, 이전의 모델보다 우월한 성능을 보였다.

주제어 : 뉴럴네트워크, 약물-표적 단백질 연관관계 예측모델, 단백질 수치화

---

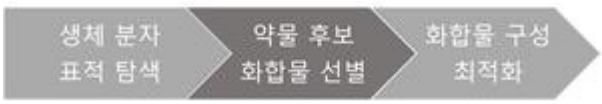
* 이 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2014R1A2A1A11049728).

† 교신저자: 김홍기, 서울대학교 치의학대학원, (08826) 서울특별시 관악구 신림동 150-21
연구 분야: 의료정보학
Tel: 02-880-2319, E-mail: hgkim@snu.ac.kr

서 론

기계 학습 방법은 제약 회사에 의해 축적된 풍부한 약물 측정 데이터를 이용하여 보다 효과적인 약물 개발을 제공 할 수 있다[1][2]. 2015년 데이터 과학 대회 플랫폼인 캐글(Kaggle)에서 주최한 머크 분자 활동량 대회(Molecular Activity Challenge)에서 신약탐색과제가 진행되었다[3]. 우승을 차지한 팀 'gggg'은 컴퓨터 전공자 5명으로 이루어진 팀으로써, 딥러닝을 활용하여 신약탐색을 수행하였다. 이들은 신약탐색에 대한 전문 지식의 개입을 최소화하고 예측 모델의 성능을 최대화하는 방법으로 가장 우수한 성적을 거두었다[4, 5]. 이 모델은 '지난 10년을 통틀어 랜덤 포레스트보다 뛰어난 성능을 보이는 첫 모델' 이라는 평을 듣는다.[6]

신약 탐색은 다음과 같은 3가지 단계의 파이프 라인을 거친다. 첫 번째 단계는 약물이 작용할 수 있는 잠재적인 생체 분자 표적을 확인하는 것이다. 예를 들어, 단백질은 화합물에 의해 유익한 치료 효과를 얻게끔 활성이 변형 된다. 두 번째 단계는 수십만 개의 약물후보 화학물을 가려내는 것이다. 일반적으로 생물학적 고효율 실험기법(high-throughput analysis)을 통하여 목표 단백질과의 상호 작용을 측정한다. 세 번째 단계는 표적과 상호 작용하는 주성분 화합물을 선택한다. 약물후보 화합물의 효능을 최적화하고 부작용을 줄이기 위해서 화학 구조와 골격(scaffold)을 수정한다. 그림 1은 신약 탐색 파이프 라인을 간략하게 표현한 것이다.



(그림 1) The process of the drug discovery

위의 3가지 단계 중에서 2번째 단계인 고효율 실험기법은 약물 선발에서의 시간과 비용이 많이 소요되기 때문에 많은 표적에 대해 시행할 수 없다. 이때, 약물 후보 화합물 선별 단계를 *in-silico* 기반의 가상 선별(virtual screening)로 대체할 수 있다. 가상 선별은 화합물-표적 상호 작용에 대한 모델을 구축하여, 실제 실험에 근접한 결과 값을 도출하는 과정이다. 가상 선별은 예측 정밀도가 높아야만 하며, 표적과 상호 작용하는 새로운 분자 구조를 탐지하는 능력이 요구된다. 가상 선별의 또 다른 장점은 다른 단백질과의 상호작용도 테스트할 수 있는 것이다. 일반적으로 약물은 하나 이상의 단백질과 상호 작용하며 이러한 상호 작용의 대부분은 효능이 떨어지거나 원치 않는 부작용을 일으킨다. 현재, 많은 약물후보 화합물이 임상 시험에서 실패하는데, 그 요인은 미처 발견하지 못한 부작용 때문이다. 이러한 실패는 신약탐색의 효율성을 떨어뜨리는 주요한 원인으로써, 많은 시간과 비용을 필요로 한다. 따라서 가상 선별은 다양한 실험결과에 기초하여 약물 후보 물질의 우선 순위를 정할 수 있으며, 따라서 약물 설계 프로세스의 효율성을

높일 수 있다.

기존의 가상 선별법은 구조 기반 방법과 리간드 기반 방법으로 나눌 수 있다. 리간드 기반 접근법은 생체 분자 표적에서 화합물의 활성을 예측할 때, 표적 단백질의 정보를 제외하고 화합물의 정보에 기반하여 활성값을 예측한다[7]. 구조 기반 방법은 단백질의 3차원 분자 구조를 활용하는 것으로써, 화합물과 표적 단백질 사이의 물리적 상호 작용을 시뮬레이션 한다[8]. 이처럼 약물 발견 및 개발 과정을 촉진하기 위한 화합물 - 표적의 상호작용 예측 모델이 제안 되었고, 다수의 성공적인 예측치도 달성되었다. 그러나 기존의 예측 모델은 주어진 표적 단백질에 한정적이거나 희소한 3차원 데이터에 한정됨으로써 예측 모델의 확장성과 유연성이 부족하다.

본 논문은 신약 탐색의 확장성을 높이는 새로운 딥러닝 모델을 제안한다. 이 접근법은 특질 추출과 예측모델로 이루어져 있다. 첫째, 특질 추출에서는 단백질과 화합물 정보를 벡터화 한다. 특히 단백질의 특질을 아미노산 기반으로 추출함으로써 거의 모든 단백질에 대한 신약 탐색이 가능하다. 둘째, 예측 모델을 쌍(pairwise) 기반으로 구성함으로써 모델의 확장성과 유연성을 높였다. 위에서 주어진 특질 값을 고정된 표에 대한 예측모델로 구성하지 않고 쌍으로 표현하여서, 주어진 단백질에 대한 새로운 화합물을 탐색하는 것뿐만 아니라, 주어진 화합물에 대한 새로운 표적 단백질을 탐색하는 것도 가능하게 되었다.

본 논문은 PCBA 데이터 셋 을 통해서 약 400,000개의 약물과 125개의 단백질로 이루어진 약 4억3천만 쌍에 대한 화합물 - 단백질 상호작용을 평가했다. 그 결과로 본 연구에서 제안하는 방법이 기존 방법 보다 더 정확한 결과를 얻음으로써, 아미노산 기반의 단백질 특질을 활용한 쌍 기반의 예측 모델이 효과적임을 확인하였다. 본 논문의 구성은 다음과 같다. 2절에서는 본 연구와 관련있는 연구들에 대해 소개한다. 3절에서는 본 연구에서 제안하는 특질 추출법과 예측 모델 구성에 대해 상세히 설명한다. 4절에서는 제안하는 방법에 대한 성능을 검증한다. 5절에서는 본 연구의 결론을 맺는다.

연관연구

리간드 기반 예측 모델

리간드 기반 접근법은 표적 단백질의 특질(예: 단백질의 3차원 구조)을 모델링하지 않고 화학물질의 특성만으로 화합물-표적 단백질의 연관성을 예측한다. 화학적 특징은 분자량, 원자 수, 전하 표현, 분자 그래프의 원자 - 원자 관계와 같은 분자의 상태로 표현된다. 지금까지 축적된 화합물-표적에 관련된 많은 데이터는 리간드 기반의 접근법에 적합하다. 이 데이터들은 테이블 데이터로써, 테이블의 열(row)은 화합물의 리스트가 나열되고 행(column)은 표적 단백질의 리스트가

나열된다. 특정 열과 특정 행에 교차되는 값은 해당 화합물과 표적 단백질의 생물작용 값이다. 학습 과정에서는 화합물 값의 유사성을 통하여 화합물-표적 단백질 관계를 학습한다. 새로운 화합물으로 학습된 모델을 평가할 경우에도, 새로운 화합물과 기존에 학습된 화합물의 유사성을 통하여 표적 단백질과의 연관성을 예측한다. 이때, 새로운 화합물이 예측할 수 있는 표적 단백질의 종류는 학습 과정에서 사용한 표적 단백질에 한정한다.

현재까지 머신러닝을 통한 가상 선별법은 리간드 기반 방법이 지배적이며 앞서 소개한 캐글(Kaggle)의 머크 분자 활동량 대회(Molecular Activity Challenge)도 리간드 기반의 방법론으로 제시된 대회이다. 예를 들어 나이브 베이즈 분류기를 이용하여 화합물-표적 단백질 예측 모델을 학습한다[9]. Standard Naive Bayes(SNB) 알고리즘[10]은 한발 더 나아가서 예측 모델에 기여하는 특질에 가중치를 부여하여 예측모델을 구축한다. Robert Lowe는 밀도 추정 함수의 한 종류인 Parzen Window 기법을 이용하여 ChEMBL 데이터 셋의 화합물-표적 단백질 결합 예측과 약물 부작용을 학습하는 모델을 선보였다[11]. Relevance Vector Machine(RVM)[12]은 서포트 벡터 머신에 확률적인 의미를 부여하여 예측모델을 구축하였다. 이는 차후 위험성에 대한 의사결정에 기여를 할 수 있다.

그러나 리간드 기반 접근법은 예측 모델의 학습 과정에서 표적 단백질의 특질을 사용할 수 없는 한계가 있다. 이에 더불어 학습 데이터에 존재하는 표적 단백질에 한정된 예측 모델이라는 한계점도 뚜렷하다. 이러한 한계는 리간드 기반 접근법에 사용되는 데이터 형태가 테이블이기 때문이다. 표적 단백질은 데이터 테이블의 행으로 표현되므로 특질 값을 활용할 수 없으며, 행의 길이가 고정된 상태로 예측 모델의 학습이 진행되므로 학습 이후에는 표적 단백질의 종류를 추가할 수 없다.

3차원 구조 기반 예측 모델

구조 기반(structure based) 접근법은 표적 단백질의 3차원 구조 데이터가 존재할 때 사용한다. 리간드 기반의 접근법은 표적 단백질의 특질 정보를 활용하지 못했던 반면, 구조 기반 접근법은 단백질의 3차원 구조 데이터를 활용하여 예측 모델을 학습한다. 또한, 데이터 형태가 테이블이 아닌 쌍이므로, 학습 이후에도 다양한 표적 단백질에 대한 예측이 가능하다. 구조 기반(structure based)의 가상 선별법은 전통적으로 도킹 프로그램을 사용한다. 도킹 프로그램은 화합물이 표적 단백질의 어떤 부위에 결합하는 것이 가장 잘 맞는지 예측하여 점수를 매긴다[13]. 이 점수는 분자들이 서로 결합할 때 안정적인 화합물을 이룰 수 있도록 유도하는 분자의 선호되는 방법을 표현한다[14]. DeepVS[15]는 도킹 프로그램을 사용하여 화합물-표적 단백질 예측 모델을 학습하는 초창기의 모델이다. 이어서 Atomic CNN[16]은 PDBBind[17] 데이터에서 유래한 3차원 단백질

구조를 모델의 입력으로 사용하여 화합물 - 표적 단백질 예측 모델을 학습한다. AtomNet[18]은 화합물-표적 단백질의 결합 영역 내부에서 여러 개의 접합 후보 영역을 사용한다.

그러나 단백질의 3차원 분자 구조 데이터는 알려진 사례가 많지 않기 때문에 아직까지 광범위한 신약 탐색에는 적절하지 않다. 또한 뉴럴 네트워크를 비롯한 기계학습 모델은 성능이 높은 대신 모델의 복잡도(예. 파라미터 숫자)가 높기 때문에 데이터를 많이 필요하므로, 구조 기반의 접근법은 이와 같은 학습 모델의 성능을 내기에는 적절하지 않다.

제안 모델

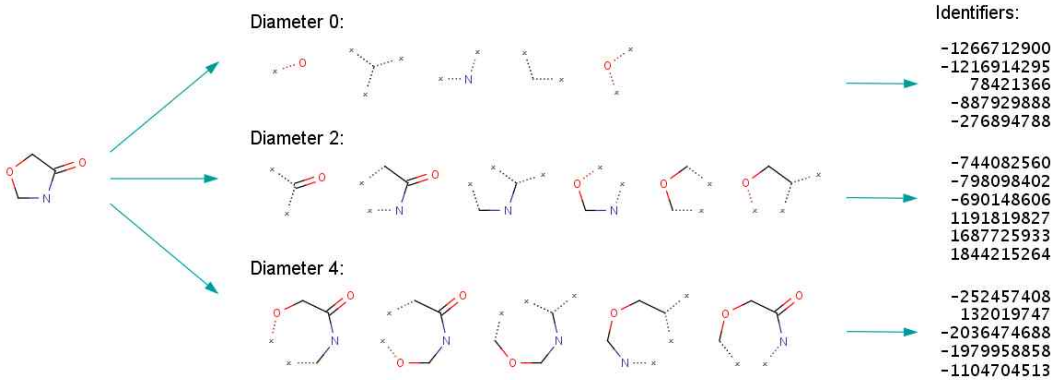
본 논문에서 제시하는 방법은 크게 2가지 단계로 구성된다. 첫 번째 단계는 화합물과 단백질을 위한 특질 추출 모델이다. 화합물은 Extended-Connectivity Fingerprints(ECFPs)[19]기법을 이용하여 특질을 추출하며 단백질은 아미노산 서열에 기반하여 자연 언어 처리(NLP) 방법으로 특질을 추출한다. 두 번째 단계는 추출된 특질에 기반한 화합물 - 단백질의 상호작용 예측 모델이다. 주어진 화합물 - 단백질 쌍은 각각의 특질 따라서 쌍 기반 레이어에 연결된다. 이후 마지막 네트워크 층에서 화합물 - 단백질의 정보를 결합하여 상호작용의 가능성을 예측한다.

약물 특질 추출법

ECFPs는 분자 특성 분석, 유사성 검색 및 구조 활동 모델링을 위해 설계된 화합물 지문으로 분류된다. 이 방법은 약물 발견에서 가장 인기있는 유사성 검색 도구 중 하나이며 다양한 응용 분야에서 효과적으로 사용된다. ECFPs는 크게 3가지 순서로 진행된다. 첫 번째 단계는 원자 식별자의 초기할당(intitial assignment of atom identifiers)으로 주어진 분자구조에 존재하는 비수소 원자에 정수 식별자를 할당한다. 이 식별자는 원자의 다양한 속성들(예. 원자번호, 연결 개수 등)이 해시 함수에 의해서 단일 정수값으로 변환된다.

두 번째 단계는 식별자의 반복적 갱신(iterative update of identifiers)으로 지정된 지름에 도달 할 때까지 초기 원자 식별자를 이웃하는 원자의 식별자와 결합하기 위해 여러 번의 반복이 수행된다. 각 반복은 각 원자 주위의 더 크고 더 큰 원형 인근을 캡처 한 다음 적절한 해싱 방법을 사용하여 단일 정수 값으로 인코딩되며 이러한 식별자는 목록으로 수집된다.

세 번째 중첩 삭제(duplication removal)는 생성 프로세스의 마지막 단계로써, 실제로는 동일한 여러가지 식별자 표현을 제거한다. 만약 동일한 연결성(edge)를 갖거나, 정수 식별자가 동일한 경우 동일한 것으로 간주한다. 이를 그림으로 표현하면 다음과 같다.



(그림 2) ECFP generation step[20]

그림 2는 첫 번째 단계와 두 번째 단계를 표현하는 예이다. 주어진 화합물이 서브 모듈로 분해되고, 각각의 서브모듈에 고유한 단일 정수 값을 식별자(identifiers)로 할당된다. 그림에서 보듯이 1개의 화합물을 무거운 원자를 기준으로 거리가 0인 서브 모듈, 거리가 2인 서브 모듈, 거리가 4인 서브 모듈로 분해한 뒤 각각의 서브모듈에 고유한 식별자를 할당한다. 원자를 기준으로 한 반복적인 연산은 널리 알려진 모간 알고리즘[21]을 사용한다. 표 1은 ECPFs 의 세부 알고리즘이다.

〈표 1〉 ECFPs 알고리즘의 의사코드

Input: molecule, radius R, fingerprint length S

Initalize: fingerprint vector $f \leftarrow 0_S$

for each atom a in molecule

$r_a \leftarrow g(a)$ ▷ lookup atom features

for L=1 to R

for each atom a in molecule

$r_1, \ldots, r_N = neighbors(a)$

$v \leftarrow [r_a, r_1, \ldots, r_N]$ ▷ concatenate

$r_a \leftarrow hash(v)$ ▷ hash function

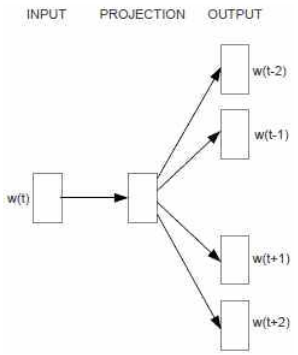
$i \leftarrow \mathrm{mod}(r_a, S)$ ▷ convert to index

$f_i \leftarrow 1$ ▷ write 1 at index

Return: binary vector f

단백질 특질 추출법

본 논문에서는 단백질의 아미노산 배열으로부터 유용한 특징을 배우기 위해 Skip-gram 모델[22]을 사용한다. Skip-gram 모델은 다양한 자연어 처리 작업에서 성공을 거둔바 있는 워드 임베딩 모델이다. 워드 임베딩은 단어의 의미를 수치화하여서 연산에 유용하게 하는 효과가 있다. Skip-gram은 단어가 주어졌을 때 주어진 단어 주위의 맥락 단어를 예측하는 신경망 모델로써, 문장에 존재하는 단어 사이의 문맥 관계를 학습한다.



(그림 3) Skip-gram model[22]

그림 3에 의하면, 단어 $w(t)$가 주어졌을 때, 맥락 단어인 $w(t-2), w(t-1), w(t+1), w(t+2)$를 예측하는 모델이다. 따라서 주어진 단어 $w(t)$ 주변의 맥락 단어가 등장할 확률을 최대화 하는 것이 skip-gram 모델의 목적함수가 된다. 이를 수식으로 표현하면 수식(1)과 같다.

$$Max\ \frac{1}{T}\sum_{t=1}^{T}\sum_{-c \le j \le c, j \neq 0} \log p(w_{t+j}|w_t) \quad (1)\ [22]$$

c는 학습의 맥락 단어의 개수로써, 그림 3에서는 c가 2이다. c가 클수록 더 많은 맥락정보를 포함하며, 정확성 또한 높아진다. skip-gram 모델은 일반적으로 $p(w_{t+j}|w_t)$를 소프트 맥스(softmax) 함수로 표현하며, 이는 수식(2)이다.

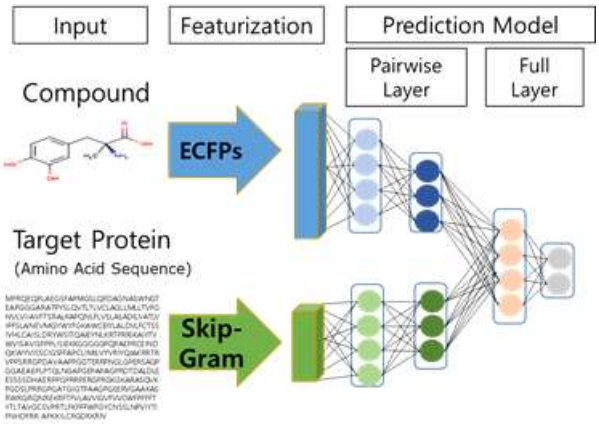
$$p(w_O|w_I) = \frac{\exp({v'}_{w_O}^{T} v_{w_I})}{\sum_{w=1}^{W} \exp({v'}_{w}^{T} v_{w_I})} \quad (2)\ [22]$$

수식(1)에서 w_O와 w_I는 각각 수식(2)의 w_{t+j}와 w_t에 대응되며, v'_{w_O}는 단어 w_O의 출력 레이어에서의 벡터형이고, v_{w_I}는 단어 w_t의 입력 레이어에서의 벡터형이다.

본 연구에서는 Skip-gram을 활용하여 단백질의 특질을 추출하는 방법을 제안한다. 단백질의 아미노산 서열 전체는 문장으로 간주하고, 3개 아미노산 서열을 단어로 간주했다. 각 단백질의 아미노산 서열에 대해서 첫 번째, 두 번째, 세 번째 아미노산 서열에서 시작하여 가능한 모든 단어를 skip-gram을 위한 말뭉치(corpus)로 생성하고, 이에 기반하여 학습을 진행한다. 그 후 단백질 문장은 그 문장에 포함된 단어가 가진 벡터값의 평균에 의하여 벡터값으로 변환된다.

쌍 기반 뉴럴 네트워크 모델

본 논문에서는 짝으로 이루어진 데이터와 레이어를 통해서 학습하는 뉴럴 네트워크 모델을 제시한다. 이 모델은 입력 데이터와 은닉 레이어(hidden layer)를 짝으로 구성함으로써 모델의 확장성과 성능을 높였다. 이 모델의 전체적인 구성은 그림 4와 같다.



(그림 4) Pairwise neural network model

쌍 기반 입력 데이터

본 논문에서는 짝으로 이루어진 입력 데이터를 갖는 뉴럴 네트워크 모델을 제시한다. 이 모델은 입력 데이터의 구조를 개선함으로써 보다 높은 유연성과 확장성을 갖는다. 특히 학습과정에서 사용되지 않은 표적 단백질도 평가과정에서 그대로 사용할 수 있다. 이를 통해서 리간드 기반기법이 갖는 평가과정에서 표적 단백질이 한정적이라는 한계를 극복한다. 또한, 이 모델의 특질 추출 기법은 표적 단백질의 정보를 모르더라도, 아미노산 서열만으로도 활용 가능하므로 3차원 구조기반 기법이 갖는 데이터 부족의 한계점도 극복한다.

모델의 뉴럴 네트워크에 입력되는 데이터 집합이 수식(3)과 같다고 하자.

$$D = \{([c_i, p_i], y_i) \mid i = 1, 2, \ldots, n\} \tag{3}$$

$[c_i, p_i]$는 입력 값으로써, c_i는 d_c차원의 벡터로 표현되는 화합물의 특질 값, p_i는 d_p 차원의 벡터로 표현되는 단백질의 특질 값이다. y_i는 라벨 값으로써, 연관관계를 나타내며 값이 1일 때 관계가 있고, 값이 0일 때 관계가 없음을 나타낸다. 뉴럴 네트워크는 데이터 입력부에서 최종 분류기까지 모두 레이어 L 로 이루어져 있다. 첫 레이어 L_0는 입력 데이터 $[c_i, p_i]$에 대응되며, 다음 레이어 $L_1, \ldots L_h, \ldots L_H$으로 연산되는 과정은 수식(4)와 같다.

$$a_h = f(W_h^T a_{h-1} + b_h) \tag{4}$$

a_h는 레이어L_h를 표현하는 벡터이며, $a_h \in R^{|L_h|}$ 이다. L_h를 위한 가중치 행렬인 W 는 $W_h \in R^{|L_{h-1}| \times |L_h|}$ 으로 구성되며 , 바이어스(bias) 벡터 b는 $b_h \in R^{|L_h|}$이다. $f(\bullet)$는 활성화 함수 $f(x) = \max(0, x)$이며 Rectified linear unit (ReLU)[23]으로도 알려져 있다. ReLU는 최근 딥러닝에서 가장 많이 쓰이는 방법으로써 최적화 수행시 높은 계산 효율성을 보이며 gradient vanishing이 나타나지 않는다. 화합물-표적 단백질 연관관계를 예측하는 것은 수식(5)와 같다.

$$y_{predict} = \sigma(W_H^T a_H + b_H) \tag{5}$$

$y_{predict}$는 각 클래스에 소속될 확률을 나타내며, 현재 모델에서는 주어진 화합물과 표적 단백질의 연관성 유무의 확률을 나타낸다. H는 마지막 히든 레이어이며, σ는 softmax 함수이다.

쌍 기반 레이어 설계

본 논문에서는 두 개의 분리된 쌍 기반 레이어로 시작하여 통합되는 전체 레이어를 지닌 뉴럴 네트워크 모델을 제시한다. 이 모델의 한 레이어는 화합물을 위한 것이며, 다른 하나의 레이어는 단백질을 위한 것이다. 쌍 기반 레이어를 거친 뒤에는 통합 레이어(concatenated full layer)에서 합쳐진 특질 값이 softmax 함수를 통해서 분류를 진행한다. 이와 같은 분리된 설계는 특질 벡터 값의 종류와 차원의 차이를 좁히는 2가지 효과를 거둔다. 첫째로 쌍 기반 설계는 특질 벡터 값의 종류(예. 2진수, 정수, 실수)를 일치시킨다. 특질 벡터 값의 종류가 다르면 표현가능한 값의 범위에도 차이가 생기기 때문이다. 이를테면, 화합물 특질 벡터 c_i는 이진 수(binary)으로 이루어

진 d_c차원의 벡터이기 때문에 표현할 수 있는 범위는 2^{d_c}가 한계이다. 반면에 단백질 벡터 p_i는 실수 값으로 이루어진 d_p 차원의 벡터이다. 따라서 d_p 차원의 값이 아무리 낮다 하더라도, 표현할 수 있는 범위는 ∞^{d_p} 이기 때문에 화합물의 표현범위 2^{d_c}보다 높다. 따라서 효과적인 학습을 위해서 특질 벡터값의 종류를 일치시키는 것은 필요하다. 둘째로, 쌍 기반 설계는 특질 벡터의 차원의 격차를 줄이는 효과가 있다. 만약 화합물 특질 벡터 c_i의 차원 d_c과 단백질 벡터 p_i의 d_p 차원의 크기가 매우 심할 경우, 각 레이어의 노드(node) 개수를 조정함으로써 각 특질 간 벡터 차원의 차이를 줄일 수 있다. 이는 앞서 언급한 특질 값의 표현력의 차이를 좁히는 효과를 거두며, 최종적으로는 모델이 특정 특질 값에 편향되지 않도록 조절하는 효과를 낳는다.

실험 및 평가

본 논문에서 제안한 예측 모델의 성능 평가를 위해서는 대용량의 화합물-표적 단백질 데이터 셋이 필수적이다. 이를 위한 데이터 셋으로 PCBA[24]를 사용했다. PCBA데이터 셋은 Pubchem 데이터 베이스 중에서 BioAssay와 관련된 데이터를 선별하여 직접 실험한 결과를 표현한 데이터이다. PCB 데이터 셋은 고효율 실험기법 (high-throughput analysis)을 통해서 생물학적 활성 여부를 판별했다. 그 데이터 셋 중에서 학습모델의 성능을 평가하기 위해 벤치마크 데이터로 사용되는 약 40,000가지 이상의 화합물과 128가지의 표적 단백질 데이터를 사용했다[25].

화합물의 특질 추출은 화합물의 입력 데이터를 SMILES[26] 형식으로 입력하여 ECFPs 알고리즘을 활용하여 임베딩 벡터의 크기 d_c=1024 값으로 변환했다. ECFPs 알고리즘을 구현한 라이브러리는 RDKit[27]을 사용했다. 단백질의 특질 추출을 위해서 Uniprot[28]의 Swiss Prot 데이터 베이스와 PCBA 데이터 셋에 존재하는 단백질 서열을 이용하여 단백질을 위한 말뭉치(corpus)를 추출하였다. 중복된 단백질은 제외 했으며 총 555,541 개의 단백질 서열이 사용되었다. 임베딩 벡터의 크기 d_p=256, 맥락 창(window size)의 크기 b=35, 최소 등장 횟수(min count) c=2 를 기준으로 단백질 말뭉치를 학습하였다. Skip-gram을 구현한 라이브러리는 Gensim[29]을 사용했다.

예측 모델을 구현하기 위한 라이브러리로 Keras[30] 를 사용했다. 예측 모델 구축을 위한 화합물과 표적 단백질 레이어의 구성은 표 2이며, 예측 모델의 학습을 위한 파라미터는 표 3과 같다.

본 논문에서 제안한 예측 모델의 성능을 검증하기 위하여 2가지 실험을 진행한다. 첫 번째로는 PCBA 데이터 셋에 대한 baseline 모델 4개와 비교하고, 두 번째로 다양한 쌍 기반 레이어의 성능을 비교한다. 본 연구는 분류문제에 대한 예측 모델을 평가하므로, Area Under the Curve(AUC) 점수를 기준으로 모델을 평가했으며, 5-fold cross validation으로 데이터를 나누었다. 본 논문에서 제안하는 baseline과의 비교결과는 표 4와 같다.

〈표 2〉 화합물과 표적 단백질의 레이어 구성

Hidden layers	Number of nodes	
	Compounds	Protein
Input layer	1024	256
Pairwise layer1	512	128
Pairwise layer2	256	64
Concatenated full layer	128	

〈표 3〉 예측 모델의 학습을 위한 파라미터

Category	Value
Learning rate	0.01
Opimizer	Adagrad
Epsilon	1e-08
Dropout rate	0.3
Batch size	64
Epoch	3

〈표 4〉 비교 모델과 ROC-AUC 점수

Model	Mean Train	Mean Validation
Proposed method	**0.983**	**0.926**
Graph Convolution	0.878	0.848
Multitask	0.815	0.797
Bypass	0.813	0.780
Logistic Regression	0.808	0.772

제안된 예측 모델은 모든 baseline 모델[25] 보다 더 높은 성능을 보였다. Logistic regression을 제외한 3개의 모델인 Gragh Convolution, Multitask, By pass 모델은 모두 뉴럴 네트워크를 기반으로 구성되어 있다.

두 번째 모델 평가는 다양한 쌍 기반 레이어의 조합에 대한 성능을 비교했다. 이를테면, 각 특질에 대한 은닉 레이어의 개수를 1~3까지 조절하고, 통합 레이어의 개수를 0~2로 조절하며 다양한 모델의 성능을 비교했다. 그 결과는 표 5와 같다.

〈표 5〉 쌍 기반 모델 차이에 따른 ROC-AUC 점수

Pairwise Layers	Concatenated Full layers	Train Validation	Mean Validation
2	1	0.983	0.926
3	1	0.992	0.909
1	1	0.981	0.902
2	0	0.988	0.902
3	0	0.989	0.891
0	2	0.943	0.858

본 논문에서 제안한 쌍 기반 레이어 모델은 다양한 방법으로 구성가능하다. 쌍 기반 레이어와 통합 레이어를 모두 갖춘 모델(2-1, 3-1, 1-1)이 높은 순위의 성능을 보였으며, 둘 중 하나의 레이어만으로 이루어진 모델은 모두 낮은 순위의 성능을 보였다. 그 중에서도 각 특질간의 쌍 기반 레이어가 없이 통합 레이어 만으로 이루어진 모델(0-2)은 각 특질간의 쌍 기반 레이어만으로 이루어진 모델(2-0, 3-0) 보다도 더욱 낮은 성능을 보였다.

결 론

본 논문은 화합물-표적 단백질 상호작용 예측모델의 확장성을 높이기 위해 데이터 기반의 새로운 딥러닝 모델을 제안한다. 이 접근법은 화합물과 표적 단백질에 대한 전문가의 지식이나 추가적인 정보없이 원본 데이터에서부터 유의미한 특질을 추출한다. 특히 표적 단백질의 경우, 기존 모델에서는 단백질의 특질이 사용되지 않거나 제한적인 단백질에 대해서만 사용이 되었으나, 본 논문은 단백질의 아미노산 서열을 기반으로 특질을 추출함으로써 표적의 대상을 거의 모든 단백질로 확장시켰다. 이에 더해서 예측모델을 쌍 기반으로 구성함으로써 모델의 확장성과 유연성을 높였다. 모델이 주어진 행과 열에 고정되지 않기 때문에, 새로운 화합물을 탐색하는 것뿐만 아니라, 주어진 화합물에 대한 새로운 표적 단백질을 탐색하는 것도 가능하게 되었다. 이와 같은 방법은 대규모 데이터 베이스와 같은 방대한 양의 화합물-표적 단백질 데이터에 적합하다. 본 연구에서는 약 400,000개의 약물과 125개의 단백질로 이루어진 PCBA 데이터 셋을 통해서 화합물-단백질 상호작용을 평가했다. 그 결과로 기존의 방법들보다 더 높은 성능을 보임으로써 아미노산 서열 기반의 단백질 특질을 활용하여 특질을 추가하는 것과 쌍기반 레이어 기반의 예측 모델이 효과적임을 확인하였다. 본연구의 향후 방향은 기계학습에 따라 추출된 특질 간의 생물학적 의미를 살피는 것이 추후 연구로 필요하다. 또한 화합물과 표적 단백질의 상호작용은 각 분자 간의 특정 영역에서 연결되므로 그 연결영역에 특화된 특질을 집중적으로 학습하는 예측

모델에 대한 연구가 필요하다.

참고문헌

[1] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal. (2014). "Clinical development success rates for investigational drugs," NatureMichael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos, Thuy B Tran, et al. Predicting new molecular targets for known drugs. Nature, 462(7270):175-181, 2009. *Biotechnology, 32*(1), pp. 40-51.

[2] Eugen Lounkine, Michael J Keiser, Steven Whitebread, Dmitri Mikhailov, Jacques Hamon, Jeremy L Jenkins, Paul Lavan, Eckhard Weber, Allison K Doak, Serge Cˆot´e, et al. (2012). "Large-scale prediction and testing of drug activity on side-effect targets." *Nature, 486*(7403): 361-367.

[3] Kaggle Merck Molecular Activity Challenge, *https://www.kaggle.com/c/MerckActivity*

[4] No Free Hunch, Deep Learning How I Did It: Merck 1st place interview, http://blog.kaggle.com/*2012/11/01/deep-learning-how-i-did-it-merck-1st-place-interview/*

[5] No Free Hunch, Merck Competition Results - Deep NN and GPUs come out to play, *http://blog.kaggle.com/2012/10/31/merck-competition-results-deep-nn-and-gpus-come-out-to-play/*

[6] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. (2015). "Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships." *Journal of Chemical Information and Modeling 55*, 263-274.

[7] J. L. Jenkins, A. Bender, and J. W. Davies. (2007). "In silico target fishing: Predicting biological targets from chemical structure," Drug Discovery Today: Technologies, vol. 3, no. 4, pp. 413-421.

[8] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. (2004). "Docking and scoring in virtual screening for drug discovery: methods and applications," *Nature Reviews Drug discovery, 3*(11), pp. 935-949.

[9] F. Nigsch, A. Bender, J. L. Jenkins, and J. B. O. Mitchell. (2008). "Ligand-target prediction using winnow and naive bayesian algorithms and the implications of overall performance statistics," *Journal of Chemical Information and Modeling, 48*(12), pp. 2313-2325.

[10] H. Y. Mussa, J. B. O. Mitchell, and R. C. Glen. (2013). "Full "Laplacianised" posterior naive Bayesian algorithm," *Journal of Cheminformatics, vol. 5*, pp. 37+, Aug.

[11] R. Lowe, H. Y. Mussa, F. Nigsch, R. C. Glen, and J. B. Mitchell (2012). "Predicting the mechanism of phospholipidosis," *Journal of Cheminformatics, 4*(1), p. 2.

[12] R. Lowe, H. Y. Mussa, J. B. O. Mitchell, and R. C. Glen. (2011). "Classifying molecules using a

sparse probabilistic kernel binary classifier," *Journal of Chemical Information and Modeling, 51*(7), pp. 1539-1544.

[13] Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. (2012). "Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review." **The AAPS Journal** 14, 133-141.

[14] Lengauer T, Rarey M. (Jun 1996). "Computational methods for biomolecular docking". *Current Opinion in Structural Biology, 6*(3), 402-6.

[15] Pereira JC, Caffarena ER, dos Santos CN. Boosting. (2016). "Docking-Based Virtual Screening with Deep Learning." *Journal of Chemical Information and Modeling 56*, 2495-2506.

[16] Gomes, J., Ramsundar, B., Feinberg, E. N., & Pande, V. S. (2017). "Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity." arXiv preprint arXiv:1703.10603.

[17] Wang R, Fang X, Lu Y, Yang C-Y, Wang S. (2005). "The PDBbind Database: Methodologies and Updates." *Journal of Medicinal Chemistry 48*, 4111-4119.

[18] Wallach, Izhar, Michael Dzamba, and Abraham Heifets. (2015). "AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery." arXiv preprint arXiv:1510.02855

[19] D. Rogers and M. Hahn. (May 2010). "Extended-connectivity fingerprints.," *Journal of Chemical Information and Modeling, 50*, pp. 742-754.

[20] ChemAxon documents, https://www.chemaxon.com/jchem/doc/user/*ECFP_files/ecfp_generation.png*

[21] Morgan, H. L. (1965). "The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service." J. Chem. Doc. 5: 107-112.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. (2013). "Distributed representations of words and phrases and their compositionality." In Advances in neural information processing systems, pages 3111-3119.

[23] Vinod Nair and Geoffrey E Hinton. (2010). "Rectified linear units improve restricted boltzmann machines." In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 807-814.

[24] Yanli Wang, Evan Bolton, Svetlana Dracheva, Karen Karapetyan, Benjamin A. Shoemaker, Tugba O. Suzek, Jiyao Wang, Jewen Xiao, Jian Zhang, Stephen H. Bryant. (January 2010). "An overview of the PubChem BioAssay resource", Nucleic Acids Research, Volume 38, Issue suppl_1, 1 Pages D255-D266.

[25] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2017). "MoleculeNet: A Benchmark for Molecular Machine Learning." arXiv preprint arXiv:1703.00564.

[26] Anderson E, Veith GD, Weininger D. (1987). "SMILES: A line notation and computerized interpreter

for chemical structures." Duluth, MN: U.S. EPA, Environmental Research Laboratory-Duluth. Report No. EPA/600/M-87/021.

[27] RDKit, *http://www.rdkit.org/*

[28] UniProt Consortium. (2014). "Uniprot: a hub for protein information." Nucleic acids research, page gku989,

[29] Gensim, *https://radimrehurek.com/gensim/models/ word2vec.html*

[30] Keras, http://keras.io

1차 원고 접수: 2017. 09. 22
1차 심사 완료: 2017. 10. 17
최종 게재 확정: 2017. 10. 24

(*Abstract*)

Pairwise Neural Networks for Predicting Compound-Protein Interaction

Munhwan Lee Eunghee Kim Hong-Gee Kim

Seoul National University

Predicting compound-protein interactions *in-silico* is significant for the drug discovery. In this paper, we propose an scalable machine learning model to predict compound-protein interaction. The key idea of this scalable machine learning model is the architecture of pairwise neural network model and feature embedding method from the raw data, especially for protein. This method automatically extracts the features without additional knowledge of compound and protein. Also, the pairwise architecture elevate the expressiveness and compact dimension of feature by preventing biased learning from occurring due to the dimension and type of features. Through the 5-fold cross validation results on large scale database show that pairwise neural network improves the performance of predicting compound-protein interaction compared to previous prediction models.

Key words : Neural network, Compound-target protein interaction, Protein embedding