BioIN 미래를 선도하는 정보 인프라!! Biotech INformation Portal
http://www.bioin.or.kr

BT분야 전문가가 바라본 분야별 동향을 소개합니다.

BioINpro

BioIN + Professional

바이오 신약(Biologics)

2024.7
Vol.140

인공지능과 구조 기반 신약개발: 기회와 도전과제

국가생명공학정책연구센터
National Biotech Policy Research Center

인공지능과 구조 기반 신약개발: 기회와 도전 과제

박한범 / 한국과학기술연구원 선임연구원

1. 배경: 컴퓨터를 이용한 구조 기반 신약개발

가. 구조 기반 신약개발의 의미와 중요성

① 의미와 중요성: 원리를 이용한 신약 개발

약물은 어떻게 우리 몸에서 작용할까? 현대 생물학이 정립되면서 우리가 알게 된 사실에 따르면, 다수의 질병은 **단백질의 기능 저하** 또는 **과다·과소 발현**으로 발생한다. 우리가 알고 있는 약물들 중 대다수는 바로 해당 질병을 일으키는 **표적 단백질과 결합**하여 그 기능을 조절한다. 약물의 작동방식은 크게 두 가지로, 하나는 **표적 단백질의 기능을 억제**하는 것이고, 다른 하나는 **증진**하는 것으로, 단백질의 어느 위치에 약물이 붙는지에 따라 앞서 말한 방식이 대개 달라진다. 그렇지만 공통 조건이 있는데, 바로 어떤 위치가 되었든 **약물이 표적 단백질과 충분히 강하게 결합해야 한다**는 것이다.

여기서 구조 기반 신약개발의 원리가 등장한다. **구조 기반 신약개발이란, 약물과 표적 단백질의 결합을 분자 구조 측면에서 직접적으로 고려하여 이상적인 약물을 찾아내고 설계하는 방법**을 뜻한다. 하지만 어찌 보면 너무나도 당연한 이런 원리가 기존에는 활용되지 못했다. 단백질의 구조를 몰랐기 때문이다. 그래서, 전통적인 신약개발 과정에서는 경험적으로 또는 수많은 시행착오를 통해 약물을 "발견"할 수 있었다. 구조 기반 신약개발은 현대 생화학의 원리에 기반하여, 신약개발에 소요되는 비용과 시간을 획기적으로 절감할 수 있도록 도와준다.

나. 컴퓨터를 이용한 구조 기반 신약개발

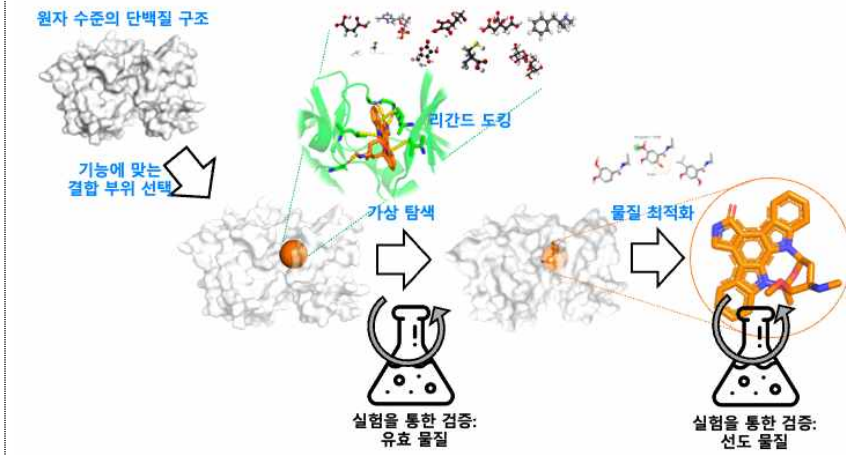
① 전체적인 과정

위에서 간단히 그 원리를 설명했지만, 사실 구조 기반 신약개발 과정이 그렇게 단순한 것은 아니다. 우선 **표적 단백질을 찾아내야 하고, 어디에 어떤 약물이 결합해야 원하는 기능(저해 또는 증진)을 잘 수행할지 알아야 한다**. 무엇보다도, 어렵게 찾아낸 약물이 우리 몸에 해롭지 않을지, 어떤 식으로 합성 및 보존이 가능할지 등등 여러 가지 요소를 모두 검증한 뒤에야 약물로서 쓰일 수 있게 된다. 특히 약물 발굴 단계에서 우리는 수많은 경우의 수를 고려해야 하는 문제에 봉착하게 된다. **인체에는 2만여 개의 단백질**이 있으며, **화합물은** 줄이고 줄여도 **최소 수 억 종류**는 된다. 우리는 궁극적으로는 **2만여 개 단백질 중, 단 한개 표적 단백질의 특정 부위에만 결합하는 화합물을 수 억 개의 화합물 후보군 가운데서 찾아내야 한다.** 모래사장에서 바늘 찾기이다. 여기서 컴퓨터의 역할이 등장한다.

컴퓨터가 구조 기반 신약개발에 활용된지는 40년이 넘게 되었다. 1982년 UCSF Dock[1]이 등장하여, **표적 단백질과 화합물 사이의 결합을 컴퓨터를 통해 시뮬레이션하는 이른바 "리간드 도킹"이 가능해졌다**(리간드라 함은 단백질에 붙는 분자를 통칭하는 단어로, 여기서는 주로 화합물을 의미한다). 이 방법은 기본적으로는 "만약 두 화합물이 결합한다면 **어떤 구조로** 붙을 것인가"에 대한 예측을 제공한다. 이후로 시간이 지나고 발전을 거듭하여, 위에 언급한 신약개발의 여러 요소 가운데 "**어떤 약물이** 붙어야 원하는 기능을 잘 수행할지", 즉 **다수 화합물의 결합 가능성을 비교해 주는 툴로 발전**하였다. 이를 컴퓨터 공간에서 다수의 화합물을 스크리닝한다는 측면에서 **"가상 탐색"**이라고 칭하며, 현재 컴퓨터 기반 신약개발에서 중추적인 역할을 수행하고 있다. 컴퓨터만 충분히 많이 있고 예측력이 정확하다면, 현실에서는 모두 다루는게 거의 불가능한 경우의 수 문제를 컴퓨터에 맡길 수 있게 되었다.

이와 동시에, 신약개발의 다른 단계에 해당하는 **표적 단백질 발굴, 약물의 물성** 및 **합성 가능성을 평가**하는 목적으로도 컴퓨터 방법은 지속적으로 발전해 왔다. 리간드 도킹과는 달리 해당 주제들은 구조 기반으로 표적 단백질과 화합물의 관계를 보는 것은 아니다. 이를테면, 표적 단백질 발굴은 대개 순수한 생물학의 영역이며, 이를 돕는 컴퓨터 방법은 주로 생물 정보학에서 비롯된다. 마찬가지로 약물의 물성

및 합성은 순수한 화학의 영역으로, 화학 정보학 기반의 컴퓨터 방법들이 이를 보조하고 있다. 이 글에서는 구조 기반 신약개발의 핵심 키워드인 **표적 단백질-화합물 상호작용에 초점**을 맞추도록 하겠다.



[그림 1] 컴퓨터를 이용한 구조 기반 신약개발 과정

② 구성 요소와 한계

컴퓨터를 이용한 구조 기반 신약개발은 크게 아래와 같은 단계를 거친다.

1) **표적 발굴**: 표적 단백질 발견, 구조 결정(또는 예측), 표적 위치 탐색
2) **유효 물질* 발굴**: 표적 위치에 대한 가상 탐색을 통한 결합 후보 선별
3) **선도 물질* 최적화**: 유효 물질을 부분적으로 바꿔 물성과 결합력 증진

* 유효물질은 표적 위치에 약하게 결합하는 물질이며 선도물질은 유효물질의 물성을 변경하여 충분한 강도로 표적위치에 결합하여 약물로서 가치가 있는 물질

여기서 중요한 부분은, **컴퓨터가 수행하는 것은 "예측"이라는 점**이다. 우리가 예측에만 의존하여 모든 과정을 진행할 수는 없으므로, **각 단계가 끝날 때마다 실험을 이용한 검증은 필수적인 단계**가 된다. 즉 1번과 2번, 2번과 3번 사이에는 각각 유효, 선도 물질 여부를 검증하는 실험이 반드시 존재해야 한다. 바꿔서 현실적으로 얘기하자면, 실험이 주된 과정이고 컴퓨터는 그 과정에서 수많은 경우의 수를 줄여주는 보조 역할을 한다고도 얘기할 수 있겠다. 즉 수많은 "거짓 양성

(False positives)"과 "거짓 음성(False negatives)"이 나오더라도, 컴퓨터를 이용한 후보군 선택이 동수의 임의 추출보다는 나을 것이라는 기대에 따른 것이다. 따라서 구조 기반 신약개발에 있어서 컴퓨터의 역할이 주인공이 될지, 아니면 조연이 될지는 방법의 예측 정확도에 따라 결정되게 된다. 그리고 최근까지 컴퓨터의 역할은 조연이었다.

2. 연구 동향: 인공지능의 접목을 통한 도약

그렇다면 컴퓨터는 신약개발에서 주인공으로 올라설 수 있을까? 달리 질문하자면, **어떻게 해야 구조 기반 신약개발을 위한 컴퓨터 방법의 예측력을 향상시킬 수 있을까?** 지난 40여 년간 수많은 화학자와 약학자들이 노력했으나 발전 속도는 기대에 미치지 못하였다. 그러나 **2010년대 후반에 이르러 외부 환경의 영향으로 변곡점이 발생하였다. 바로 인공지능이다.**

가. 인공지능과 자연과학 난제 해결: 알파폴드-2

① 단백질 구조예측의 해결

2010년대 중반만 하더라도 인공지능이 과학에 쓰일 것이라는 생각은 소수의 희망적인 사람들의 것이었다. 이른바 딥러닝의 시대가 열리고 알파고가 등장하여 바둑계를 평정했을 때에도, 구글 포토에서 사진의 객체가 자동으로 인식 및 선별될 때에도 인공지능이 과학 연구를 도와줄 것이라는 기대는 하기 쉽지 않았다. 비슷한 시기에 약 물성 예측대회에서 인공지능이 1등을 했을 때에도[2] 소수에 대한 제한적인 활용만 가능할 것이라 생각되었다.

2018년에 등장한 알파폴드-1[3]은 이런 관점을 바꾸었다. 비로소 **과학에 인공지능이 접목되었을 때의 파급력에 대해 자연과학자들이 인지하기 시작**했다. 그러나 이때까지만 해도 기존 방법으로 원래 가능했던 연산을 인공지능이 좀 더 빠르고 잘 한다는 인상을 주는 정도였다. 그리고 **2020년에 알파폴드-2**[4]**가 등장**하였다. 이제는 원래 가능하던 것을 더 잘하는 게 아니었다. 전에는 불가능한 것이 가능해졌다. 이로서 자연과학의 생태계가 바뀌었다. 기존에 쓰던 컴퓨터 알고리즘의 대대적인 수정이 불가피해졌다. 컴퓨터를 잘 쓰지 않던 분야에서도 인공지능 도입을 진지하게 고려하게 되었다. 단백질 구조예측의 거대한 난제가 풀렸으니, 다른 문제가 인공지능으로 안 풀릴 리가 없기 때문이라는 기대에서였다.

② 구조기반 신약개발에의 파급력

구조 기반 신약개발은 단백질 구조예측 문제의 연장선 상에 있다. 즉, 다른 어떤 분야에 비해서도 알파폴드-2에 의해 야기된 생태계 변화의 영향을 가장 직접적으로 받는 분야가 된 것이다. 우선 구조 기반 신약개발 자체가 가능한 표적 단백질이 크게 늘어났다. **구조 기반 신약개발은 원자 수준의 단백질 구조를 반드시 필요로 하는데, 알파폴드-2의 등장과 동시에 구조가 밝혀지지 않은 사람 단백질 수 만개에 대해 높은 신뢰도의 예측 구조를 누구나 활용할 수 있게 되었다.** 결과적으로 구조 기반 신약개발은 표적 단백질 구조가 있어야만 적용할 수 있는 조건부 방법에서 대부분의 표적에 적용 가능한 일반적인 방법으로 바뀌게 된 것이다.

두 번째는 문제의 유사성이다. 단백질 구조예측은 단백질의 구조를 원자 수준으로 예측하는 문제다. **리간드 도킹은** 단백질 대신 **단백질-화합물 결합체의 구조를 역시 원자 수준으로 예측**하는 문제이다. 대상이 조금 확장되었을 뿐 푸는 문제의 종류는 동일하다. 따라서 분야를 이해하고 있는 과학자들은 곧바로 구조 기반 신약개발을 다음 연구목표로 떠올리게 되었다. 이는 알파폴드-2를 개발한 구글 딥마인드의 행적에서도 드러난다. 그들은 알파폴드-2를 공개한지 4년 후, 후술할 **단백질-리간드 모델링이 가능한 알파폴드-3**[5]**를 공개**하였고, isomorphic lab을 설립하여 이를 이용한 신약개발에 나서고 있다.

나. 구조 기반 신약개발 인공지능의 발전

① 화합물 신약: 리간드 도킹 및 가상 탐색 방법의 발전

그러나 예상과 달리, 리간드 도킹 및 가상 탐색을 위한 인공지능의 개발은 처음부터 문제에 봉착하게 된다. 왜냐하면 인공지능은 데이터 과학이기 때문이다. **화합물은** 단백질과는 구성 요소가 너무나도 다르고 또 다양하다. **그 다양성에 비해 단백질과 화합물의 결합구조 데이터베이스는 너무 빈약하다.** 중복을 제거하고 나면 2만여 개 남짓 된다. 알파폴드-2 학습에 활용된 단백질 50만여 개와 대비된다. 후술하겠지만,

이러한 학습 데이터 부족에 의한 한계는 여전히 해결되지 않았다.

과학자들은 기술력으로 이를 돌파하고자 하고 있다. **알파폴드-2의 교훈 중 하나는, 데이터 부족에 의한 한계는 인공지능 기술력으로 상당 부분 극복 가능**하다는 것이었다. 인공지능으로 어떤 문제도 극복 가능하다는 희망적인 비전과 맞물려, 2020년 이후로 수많은 노력과 연구가 진행되었다. 트랜스포머 기법, 그래프 네트워크, 등변위성(equivariance)을 이용한 3차 구조 모델링, Diffusion을 이용한 생성형 모델과 같이, **기존에 컴퓨터 과학에서 발전된 요소들이 빠르게 신약개발 인공지능 속으로 녹아들게 되었다. 그 결과, 2024년 현재 리간드 도킹에서부터 표적 단백질 구조 기반 신약 설계까지 분야 전반에 걸쳐 기존 방법을 대처할 수 있는 인공지능이 경쟁적으로 개발 및 보고되고 있다.**

② 단백질 의약품

앞서 대부분의 내용이 화합물에 맞춰졌지만, 사실 단백질 그 자체도 약물로서 훌륭한 활용성을 지니고 있다. 의아할 수 있겠지만 단백질 의약품은 우리 생활에서 쉽게 찾아볼 수 있다. 바로 항체와 백신이다. 아직 가능성 위주로 탐구되고 있는 펩타이드 신약과 미니 단백질 신약도 20 종류의 아미노산으로 구성될 경우 단백질 신약으로 볼 수 있다. 작동 원리는 화합물의 그것과 다르지 않다. 분해를 통해 면역 체계를 작동시키는 백신을 제외하면, **기본 원리는 표적 단백질의 원하는 위치에 충분한 결합력으로 붙는 것이다. 오히려 표면적이 넓기 때문에 화합물에 비해 더 강하게 선택적으로 결합이 가능하다.**

단백질 의약품이 인공지능과 맞물려 특히 촉망받는 것은 바로 단백질이기 때문이다. 화합물과는 달리 **단백질 구조예측에 쓰였던 인공지능을 그대로 쓰거나 조금만 바꿔서 재활용할 수 있다. 당연히 화합물을 다룰 때 발생하는 데이터 문제도 적다.** 단백질 설계 분야에서 세계 최고의 기술력을 보유한 미국 워싱턴 대학교 단백질 연구소에서는 지난 1년 동안에만 단백질 구조예측 방법을 응용한 인공지능으로 수많은 종류의 단백질 의약품 후보군을 만들어내는 데 성공했다[6,7,8]. **인공지능을 이용한 화합물 신약개발이 아직 발전 단계에 머물러 있는 동안, 인공지능을 이용한 단백질 의약품 설계는 성숙 단계에 접어들고 있다.**

2024년 vol. 140

③ 모든 생체 분자를 모델링할 수 있는 통합 플랫폼의 등장

위에 언급한 인공지능 방법들은 단백질만 볼 수 있거나, 단백질을 고정하고 화합물을 시뮬레이션하거나 하는 제한적인 형태의 모델링 기능만 제공해 주었다. **그러나 생체 내에서 약물이 작용하는 원리는 사실 훨씬 복잡**하다. 단백질은 고정되어 있지도 않으며, 화학적 조성이 신호전달 단계에서 바뀌기도 하고(post-translation modification), 금속 또는 이온 원자에 의해 상호작용이 매개되기도 한다. 또한 우리는 훨씬 복잡한 약물을 설계하기도 한다. PROTAC처럼 단백질-단백질 상호작용을 화합물이 매개하기도 하며, 항체에 특정 화합물을 공유결합으로 연결하여(antibody-drug conjugate) 특정 기능을 유도하기도 하고, 단백질 대신 핵산을 표적으로 삼은 신약을 개발하기도 한다. 즉, **표적 단백질과 약물을 1:1로 붙이는 플랫폼이 아닌, 여러 종류의 생체 분자를 모두 한 번에 구조 변화까지 고려하여 모델링할 수 있는 통합 플랫폼이 필요**한 것이다.

통합 모델링 플랫폼을 표방하는 인공지능 방법들이 올해 한꺼번에 등장한 것은 결코 우연이 아니다. 2024년 봄에 차례대로 등장한 **RF-AA** [9]**와 알파폴드-3**[5]는 현재 신약개발 관련 수요를 반영한 것으로 보인다. 두 방법 모두 **단백질과 기타 분자들(이온, 화합물, 핵산 등) 여러 개로 이뤄진 복합체의 3차 구조를 동시에 모델링**해 준다. 신약개발 측면에서도 유용성이 크다. 리간드의 결합이 표적 단백질의 큰 구조 변화를 수반하는 경우와 같은 어려운 도킹도 가능케 해준다.

다만 과대 해석은 피해야 한다. 진일보한 것은 맞으나 가야 할 길은 멀다. 표적 단백질과 화합물을 동시에 모델링 하면서 생기는 비용을 생각해야 한다. 그리고 구조를 붙여보기만 한다고 해서 결합 여부를 알 수 있는 것은 아니다. 여러모로 수백만 개의 화합물을 가상 탐색하기에 적합한 방법이 아닌 것이다. 구조예측 정확도 역시 추가 검증이 필요하다. 알파폴드-3는 논문에서 리간드 도킹 정확도 70%를 보고하였으나, 그 정확도가 일반적인 약물개발 시나리오에서도 통용될 수 있는 일반적인 정확도인지는 아직 알기 어렵다.

3. 향후 과제

가. 혁신 신약 발굴을 위한 데이터 과적합 문제 해결

① 혁신적인 알고리즘, 부족한 데이터

현재 **화합물 기반 신약개발 방법이 봉착해 있는 가장 큰 문제**를 한 단어로 표현하자면 **"과적합"**이다. 학습한 문제만 잘 풀고 새로운 예제는 잘 못 푼다는 뜻이다. 신약개발 관점에서 얘기하자면 이른바 혁신 신약 발굴에 적용하기 어렵다는 것이다. 신약개발용 인공지능의 알고리즘이 뒤쳐져서가 아니다. 이제 구조 기반 신약개발 문제는 수많은 컴퓨터 과학자들도 관심을 가지는 분야가 되어서 최신 기술력이 빠르게 흡수되고 있다. **과적합의 가장 큰 이유는 데이터 부족 문제가 온전히 해결되지 않았기 때문**이다.

인공지능 구조 기반 신약개발에서 핵심 위치를 차지하는 것은 **리간드 도킹과** 거기서 파생되는 **가상 탐색을 수행해 주는 인공지능**이다. **도킹 인공지능을 학습하기 위해선 단백질-리간드 결합 구조가 많아야 한다.** 가상 탐색 방법을 학습하기 위해선 다수의 표적 단백질에 대해서 각각 결합하는 분자와 결합하지 않는 분자에 대한 데이터를 확보해야 한다. 숫자가 결코 적은 것은 아니다. 결합구조는 8만여 개가 있으며, 결합 데이터는 수백만 개 정도가 있다. 문제는 단순 숫자가 아니다. **첫 번째 문제는** 그 중 태반이 **중복**이거나 의약품의 특징을 학습하는데 **부적합한 분자**라는 것이다. 분자들을 추리고 나면 대략 1/4로 숫자가 줄어든다. 특히나 **특정 단백질군에만** 그 숫자가 크게 **쏠려** 있다. Kinase 단백질 군에 대한 결합 데이터가 전체의 30% 이상이다. 그리고 그에 못지않게 중요한 **두 번째 문제는**, 결합 데이터는 많이 있으나 그 반대에 해당하는 **비결합 데이터는 거의 없다**는 것이다. 결합하지 않는 분자를 모아서 데이터베이스화하는 노력의 가치가 그 동안 과소평가 되었던 것이다.

② 대두되는 과적합 문제, 그리고 해결책

인공지능 학습을 조금이라도 해본 사람은 알 것이다. 우리가 아차 하는 순간에 인공지능은 과적합 된다. 문제의 본질을 이해하기 보다 노력하지 않고 쉽게 그럴듯한 편법을 찾아내는 방향으로 학습이 진행된다. 문제가 복잡해지고 어려울수록 사람조차 인공지능의 과적합 여부를 구별하기 어렵게 된다. 구조 기반 신약개발을 위한 인공지능이 실제로 지난 몇 년간 그랬다.

지난 3년간 수많은 리간드 도킹 방법들이 보고되었다. 그리고 전통적인 비-인공지능 방법의 성능을 뛰어넘었다는 보고가 줄이었다. 그러자 한 그룹에서 정말 그런지 평가를 해보았다[9]. 새로운 단백질-리간드 구조들을 들고 와서 해당 인공지능 방법들로 풀어본 것이다. 결과는 흥미로웠다. 새로운 예제들에서는 그 어떤 인공지능도 전통적인 리간드 도킹 방법을 넘어서지 못했다. 가상 탐색 방법 역시 마찬가지였다. 완벽에 가까운 구별 성능을 보이는 인공지능들이 학계에 보고되었다. 그러나 문제가 곧 드러났다. **비결합 분자에 대한 공용 데이터의 부재**가 발목을 잡았다. 인공지능들은 허술하게 구성된 비결합 분자들 사이에서 결합 분자들을 구별하는 편법을 너무나도 쉽게 터득했다 [10]. 그럴싸한 다른 가짜들을 들고 와서 다시 문제를 풀게 하였더니 금세 인공지능의 분별력은 무작위에 가까워졌다.

재차 강조하지만 **과적합의 이유는 인공지능이** 본질을 이해하기 보다는 **점수를 잘 받는 편법을 터득**했기 때문이다. 인공지능이 터득하기를 기대하는 문제의 **본질이란**, **표적 단백질과 화합물 사이의 물리화학적 상보성**이다. 반대로 **편법**이라 함은 화합물 또는 표적 단백질에 내재되어 있는 **(잘못된) 데이터 편향을 악용**하는 것이다. 그렇다면 해법은 비교적 명확하다. **인공지능에게 데이터 뿐만 아니라 물리화학을 학습**시켜야 한다. **이를 학습 편향**(inductive bias)**을 넣는다**고 표현한다. **데이터가 못하는 부분을 화학자의 직관으로 채우는 것**이다. 그 양상은 여러 형태로 나타난다. 상호작용을 계산하는 부분을 일반적인 행렬 연산이 아닌 물리 수식으로 대체해 주거나[11], 인공지능이 학습하는 목표값 이외의 또 다른 물리적 측정값을 동시에 예측할 수 있도록 유도하는 방식 등이 있다. **앞으로 반드시 탐구되어야 할 중요한 연구 테마**이다.

나. 약물의 물성, 독성 및 면역원성 예측

표적 단백질에 잘 붙기만 한다고 다 약물이 되는 것이 아니다. 실제 신약개발에서 가장 골치를 썩이는 부분은 사실 다른 곳에 있다. 바로 **약물의 물성과 독성**이다. 약물은 물에 잘 녹아야 하고, 세포질을 투과해야 하며, 인체 내에서 예기치 못한 오작동을 하지 않아야 한다. 단백질 의약품은 원치 않는 면역반응을 유발하는 면역원성이 낮아야 한다. 구조 기반 신약개발이 답변해 주기 어려운 질문들이다. 해당 요소들은 신약개발의 전임상·임상 단계에서 수년에 걸쳐 검증된다. 실패율도 높다. 아무리 구조 기반의 **신약개발 인공지능이 극도로 발전하여도 전체 신약개발 과정이 크게 단축되기 어려운 이유**이다.

약물성과 독성 예측 분야에서의 인공지능의 접목은 더딘 편이다. 단백질 구조 예측의 직접적인 파급효과를 누리고 있는 구조 기반 신약개발 분야에 비해 획기적인 변곡점이 없어서일 것이다. 데이터 문제 역시 관련 있다. 미국 보건복지부에서는 수십 년 전부터 곳곳에 흩어져있는 약물성과 독성 관련 데이터를 수집하여 표준 공용 데이터베이스화하는 작업을 진행 중이다. 그러나 아직은 인공지능을 학습하기에는 역부족이다. 인공지능을 접목하고자 하는 많은 시도들이 있었으나 믿고 신뢰할 만한 수준으로 성능이 올라서기엔 요원한 현실이다. 단백질 의약품에 필요한 **면역원성 예측** 역시 마찬가지다.

필자는 구조 기반 신약개발 방법을 만드는 연구자로서, **"독성·면역원성을 구조적인 관점에서 접근할 수 있지 않을까"** 하는 생각을 오래전부터 해왔다. 높은 비율의 화합물 독성은 사이토크롬 P450 또는 핵 수용체(Nuclear receptor)와 같은 단백질들과 약물이 비선택적으로 결합함으로써 일어난다고 알려져 있다. **면역원성의 원인 역시 복잡하지만, 단백질 서열과 주조직 적합성 복합체(MHC)의 상보성**과 같은 **구조적인 관점에서 해석**하는 관점이 있다. 해당 과정들에는(표적 단백질이 아닌) 타 단백질과 의약품의 구조적 상호작용이 중요한 역할을 한다. 구조 기반 신약개발 인공지능이 충분히 기여할 수 있는 분야라고 생각하는 이유이다.

4. 맺음말: 인공지능 신약개발 연구 동향

인공지능 신약개발은 현재 거대 제약회사의 주 관심사 중 하나이다. 애초에 **컴퓨터를 이용한 신약개발은** 이상적으로 봤을 땐 매력적인 분야였으나 그 동안 컴퓨터 방법의 예측 정확도 한계로 제약회사의 많은 이목을 끌지는 못했다. **인공지능에 의한 돌파구가 마련**되면서, 기술력을 가진 테크 업체와 제휴를 맺거나 자체적인 연구팀을 구성하여 구조 기반의 인공지능 기술력을 선점하려는 움직임이 특히 미국 거대 제약회사(BMS, Pfizer, Novartis 등)에서 두드러지고 있다. **학계의 반응은** 제약회사와는 온도 차가 있다. 필자의 생각에는 예전에 비해서 **기술력을 선점하는 것도 중요하지만 기술력을 공유**해야 한다는 시각도 크다. 알파폴드-1와 -2의 논문 게재 과정에서 학계에 이와 관련한 주요 논쟁이 있었다. 그리고 결과적으로 알파폴드-2는 완전한 비영리화 및 코드 공개가 이루어졌다. 이로써 학계에 암묵적인 기준이 세워졌다. 회사에서 개발한 인공지능이라 하더라도 **인공지능을 이용한 과학 연구는 비영리 목적으로 완전한 코드 공개가 이루어져야 한다**는 기준 말이다. 신약개발 인공지능이 단백질 구조 예측 인공지능의 연장선상에 있다는 관점에서처럼, 신약개발 인공지능도 이러한 암묵적인 기준으로부터 자유롭지는 않을 예정이다. 2020년 이전의 연구 과정에서는 비-인공지능 방법들은 논문 투고 과정에 코드 공개가 필수는 아니었다. 그러나 이제는 코드 공개가 필수에 가까워지고 있다.

관련하여 학계에서 또 하나 주시하고 있는 점은 인도적 목적의 활용이다. 그림 및 영상과 같은 매체에 대해 생성형 인공지능이 오용되는 사례가 빈번하게 증가하면서, 신약개발을 위한 생성형 인공지능 역시 비인도적 목적으로 오용될 수 있다는 문제 제기가 있었다. 학계에서는 이에 발 빠르게 대처하였다. **미국 단백질 설계 연구소의 데이비드 베이커 소장의 주도하에 전 세계 과학자들이 생성형 분자 설계 인공지능의 오용 금지에 서명**하였다 [12]. 필자는 서울대 백민경 교수, KAIST 김호민 교수와 더불어 해당 조약에 서명한 174 명의 과학자 중 하나였다.

이제 구조 기반 신약개발과 인공지능은 단순 가능성을 타진하는 수준을 넘어서 큰 주목을 끄는 단계에 접어 들었다. 그리고 점점 많은 비전문가들이 이용할 수 있는 단계로 나아가고 있다. 소비자가 현명해야 좋은 물건을 잘 살 수 있듯이, 이제는 인공지능 개발자 뿐만 아닌 약학계에 종사하는 다수의 활용자가 인공지능의 가치와 한계를 잘 이해해야 하는 시대가 오고 있는 것이다. 이 글이 그 과정에 도움이 되길 바라며 글을 맺는다.

참고문헌

1. Kuntz, ID; Blaney, JM; Oatley, SJ; Langridge, R; Ferrin, TE (1982). "A geometric approach to macromolecule-ligand interactions". Journal of Molecular Biology. 161 (2): 269-88
2. Chen, Hongming; Engkvist, Ola; Wang, Yinhai et al (2018), "The rise of deep learning in drug discovery", Drug Discovery today 23:1241.
3. Senior, Andrew W; Evans, Richard; Jumper, John; (2020). "Improved protein structure prediction using potentials from deep learning". Nature 577: 706-710.
4. Jumper, John; Evans, Richard; Pritzel, Alexander; Green, Tim et al. (2021). "Highly accurate protein structure prediction with AlphaFold". Nature. 596 (7873): 583-589.
5. Abramson, Josh; Adler, Jonas; Dunger, Jack; Evans, Richard et al (2024). "Accurate structure prediction of biomolecular interactions with AlphaFold 3". Nature 630: 493-500.
6. Berger, Stephanie; Seeger,Franziska; Yu, Ta-Yi; Aydin, Merve et al (2024), "Preclinical proof of principle for orally delivered Th17 antagonist miniproteins", Cell 187: 3726-3740.
7. Sahtoe, Danny D.; Andrzejewska, Ewa A.; Han, Hannah L.; Rennella, Enrico et al (2024), "Design of amyloidogenic peptide traps", Nature Chem Biol. 20, 981-990.
8. Susana Vázquez Torres, Leung, Philip J Y; Venkatesh, Preetham; Isaac D Lutz (2023), "De novo design of high-affinity binders of bioactive helical peptides", Nature 626, 435-442.
9. Buttenschoen, Martin; Morris, Garrett M.; Charlotte M. Deane (2024) "PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences", Chem Sci 15, 3130-3139.
10. Chen, Lieyang; Cruz, Anthony; Ramsey, Steven; Dickson, Callum J. et al (2020) Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. Plos ONE 14(8): e0220113.

2024년 vol. 140

11. Moon, Seokhyun; Zhung, Wonho; Yang, Soojung; Lim, Jaechang; Kim, Woo Youn (2021) "PIGNet: a physics-informed deep learning model toward generalized drug-target interaction prediction", Chem. Sci. 13: 3661-3673.
12. https://responsiblebiodesign.ai/

2024년
BioINpro

• 발 행 호 : Vol.140
• 발 행 처 : 한국생명공학연구원 국가생명공학정책연구센터
• 온라인 서비스 : http://www.bioin.or.kr

34141 대전광역시 유성구 과학로 125(어은동) | Tel. 042-879-8368

생명공학정책연구센터
Biotech Policy Research Center