단백질 구조 및 디자인 연구에서의 인공지능 활용

남궁석

알파폴드, 그 이후

2020 년 딥마인드의 인공지능 알파폴드 (Alphafold)가 오랫동안 난제로 여겨지던 단백질 구조 예측 문제, 즉 아미노산 서열로부터 단백질의 3 차원 구조를 알아내는 문제를 풀었다는 것은 인공지능 기술의 큰 성취로 받아들여졌고 많은 관심을 불러일으켰다.

그렇다면 알파폴드가 등장한지 어언 3 년에 가까워지고 있는 2023 년 현재, 과연 알파폴드가 불러일으킨 '단백질 구조 예측 혁명' 은 구조생물학, 더 나아가 생명과학/공학에 어떤 영향을 미쳤을까? 오랫동안 단백질 구조를 풀고, 이를 이용한 연구를 한 입장에서 과연 알파폴드라는 기술이 어떤 파급효과를 가져왔는지를 알아보기로 하자.

알파폴드는 구조생물학 연구를 어떻게 바꾸었는가?

먼저 알파폴드의 유용성을 제일 먼저 실감한 사람들은 실험적인 방법을 통하여 단백질 구조를 풀던 구조생물학자였다.

그전까지 단백질 구조를 푸는 표준적인 방법이던 X 선 결정학 (X-ray Crystallography)은 단백질을 결정으로 만들고, 결정에 강한 X 선을 쪼아, 결정 안의 단백질 원자에 충돌하여 발생하는 회절 (Diffraction) 정보를 분석하여 이를 단백질 내의 전자 분포를 추측하는 방법이었다. 단백질 결정에서 형성된 X 선 회절 정보는 단백질의 구조에 대한 정보를 가지고 있으나, 이 정보는 불완전한 정보이다. 어떤 물체가 빛을 받아 형성된 그림자는 분명히 어떤 물체의 구조에 대한 정보를 가지고 있지만, 완벽한 정보를 가지고 있지 않은 것처럼, X 선 회절의 정보는 파장의 진폭 (Amplitude) 의 정보는 있지만 위상 (Phase) 에 대한 정보는 결여되어 있다. 그러나 단백질의 구조를 재구성하려면 진폭과 위상에 대한 정보가 다 필요하며, 이러한 위상에 대한 정보를 구하는 것은 전통적으로 단백질 결정학에서 가장 어려운 문제로 알려져 있었다.

현재까지 X 선 회절의 위상 정보를 구하기 위해서는 단백질에 X 선을 좀 더 효율적으로 튕겨내는 중금속을 넣어 나오는 회절 패턴과 일반적인 단백질의 회절 패턴을 비교하여 그 차이를 분석하고 이를 이용하여 위상 정보를 유추하는 방법을 이용했다. 그러나 이 방법은 그리 쉬운 방법이 아닌 관계로, 단백질 결정을 만들고 회절 데이터를 얻은 후에도 위상을 얻지 못하여 구조를 풀지 못하는 경우가 종종 있었다.

만약 구조를 풀고 싶은 단백질과 구조가 유사한 단백질이 있는 경우라면, 이를 이용하여 위상을 계산하여 간단히 전자 밀도를 얻을 수 있다. (이를 분자치환법 Molecular Replacement 이라고 부른다). 만약 우리가 풀고 싶은 단백질의 구조 모델 (단백질의 전체가 아닌 일부라도

된다)이 있으면 이를 '열쇠' 로 이용하여 X 선 회절 정보의 위상을 계산할 수 있는 것이다. 만약 규명하려는 단백질과 구조가 매우 유사한 단백질의 구조가 풀렸다면, 이 구조를 이용하여 비교적 간단하게 위상을 계산할 수 있었다. (호랑이를 한번도 본 적이 없지만 고양이가 어떻게 생겼는지 알고 있다면, 호랑이 그림자를 보고 고양이의 모양을 모델로 이용하여 호랑이의 생김새를 추측하는 것과 비슷하다). 그러나 이미 규명된 단백질과 유사한 구조의 단백질이 아닌 경우 이러한 방법을 사용할 수는 없었다.

이 상황에서 알파폴드가 등장하였고, 알파폴드로 예측된 모델은 충분히 정확도가 높기 때문에, 이를 이용하여 분자 치환법으로 위상을 얻을 수 있었다. 결정의 회절 데이터까지 얻었지만 위상 결정에 실패하여 구조를 풀지 못했던 회절 데이터를 이용하여 알파폴드로 만든 모델로 바로 구조를 풀었다는 사례들이 속속 등장하였다. 알파폴드의 등장은 X 선 결정학에서 가장 어려운 부분이었던 위상 계산 문제를 아주 손쉽게 극복하는 수단을 제공해 준 셈이었다.

최근 구조 생물학의 대세가 된 초저온전자현미경 (Cryo-Electron Microscopy, Cryo-EM) 등의 방법으로 단백질의 구조를 푸는 경우에도 알파폴드는 매우 유용하게 사용되기 시작하였다. 초저온 전자현미경으로 단백질 구조를 푸는 작업은 전자현미경으로 관찰된 단백질 입자의 이미지로부터 단백질의 윤곽을 구축하고, 여기에 원자 모델을 구성하는 일이다. 만약 단백질 입자의 윤곽이 고해상도로 형성되는 경우에는 정확한 모델을 그리 어렵지 않게 끼워 맞출 수 있지만 그렇지 않은 경우 정확한 모델을 만드는데 매우 많은 시간과 노력이 필요하다. 그러나 알파폴드를 이용하여 단백질 구조에 대한 모델을 만들 수 있다면, 이를 초저온 전자현미경으로 만든 윤곽에 매우 손쉽게 끼워맞출 수 있다.

결론적으로 알파폴드는 좋은 '초기 가설' 이 되는 단백질 구조 모델을 제공하는 셈이다. 이전에는 실험 데이터를 기반으로 단백질 구조를 구축했다면, 알파폴드 시대에는 알파폴드가 미리 예측한 단백질 구조 모델로부터 시작하여, 이를 실험적 증거로 검증하는 형식으로 구조 생물학의 패러다임이 바뀐 셈이다.

여러가지 실험 방법이 총동원되어 기존에는 엄두도 내지 못했던 세포 내의 거대 구조 복합체의 구조가 밝혀진 사례도 있다. 그 대표적인 사례는 세포 내에서 가장 큰 단백질 복합체인 핵공 복합체 (Nuclear Pore Complex)다. 핵공 복합체는 세포 내의 핵과 세포질 간의 물질을 교환하는 일종의 톨게이트와 같은 역할을 하는 구조물이다. 세포 내에 있는 단백질 구조물 중에서 가장 크며, 약 30 종류의 단백질 약 1,000 개에 의해 구성되어 있으며, 분자량은 무려 1 억 2 천만 (포도당의 분자량은 180, 일반적인 단백질은 1 만에서 10 만 정도의 분자량이다) 에 달한다. 독일의 막스 플랑크 생물물리학 연구소의 연구진들은 세포 내 핵공 복합체의 전체적인 윤곽을 저온 전자 토모그래피 (Cryo-Electron Tomography, Cryo-ET) 라는 실험 기술로 파악한 다음, 30 종류의 핵공 단백질 구성 요소들을 알파폴드로 예측한 다음, Cryo-ET 로 알아낸 윤곽에 끼워맞추어 핵공 복합체 모델을 구성하였다. 거대한 퍼즐의 윤곽에 알파폴드로 예측한 부품들을 끼워 넣는 것과 비슷한 일이다. 이렇게 구축된 핵공 복합체의 모델은 현재까지 구축된 핵공 복합체의 모델 중 가장 정교한 모델이었다. 이렇게 알파폴드와 실험 구조생물학의 기술들은 함께 단백질 구조를 푸는데 같이 어우러져 사용되고 있다.

단백질의 상호작용의 분석

세포 내에 존재하는 단백질은 대개 홀로 행동하지 않는다. 같은 종류의 단백질 가닥끼리 복합체를 이루거나, 혹은 다른 종류의 단백질 가닥이 만나서 복합체를 이루어 세포 내에서 존재하고 작동한다. 자동차와 같은 복잡한 기계는 기계를 구성하는 수많은 부품들이 모여서 작동하고, 이러한 부품들 사이의 상호작용이 바로 자동차를 움직이게 하는 메커니즘이 되는 것처럼 세포 내에 존재하는 단백질 간의 상호작용을 아는 것은 바로 생명현상의 본질을 파악하는 중요한 단계이다.

그렇다면 알파폴드는 단백질 상호 작용을 예측할 수 있는가? 애초에 알파폴드가 개발될 때는 단백질 상호작용의 예측을 염두에 두고 제작되지 않았다. 그러나 알파폴드가 공개된 후 알파폴드를 테스트하던 연구자들은 서로 결합하는 것으로 알려진 두 개의 단백질의 아미노산 서열을 마치 하나의 단백질인 것처럼 이어붙여 구조를 예측하면, 단백질의 결합 구조를 비교적 정확하게 예측할 수 있음을 발견했다. 이후 딥마인드는 알파폴드 초기 버전이 나온 이후 이를 업데이트하여 단백질의 복합체 구조를 보다 정확히 예측할 수 있는 보완 버전을 만들었다.

그렇다면 어떻게 알파폴드는 애초에 의도하지 않았던 단백질 상호작용을 예측할 수 있을까? 이것은 알파폴드가 작동하는 기본 원리와 관련되어 있다. 알파폴드는 단백질의 구조를 결정하는 아미노산 사이의 상호작용을 단백질의 진화 정보에 의해서 유추한다. 서로 다른 진화 경로를 걸어온 비슷한 단백질은 아미노산 서열이 달라도 거의 유사한 구조를 유지해야 한다. 단백질을 구성하는 아미노산 서열이 달라져도 같은 구조를 유지하기 위해서는 단백질 구조 내에서 서로 접하는 아미노산들은 진화 과정에서도 특수한 관계를 가진다.

가령 어떤 단백질의 20 번째 아미노산인 (음성 전하를 가진) 글루탐산과 81 번째 아미노산인 (양성 전하를 가진) 라이신이 서로 이온 결합으로 상호작용을 하고 있다면, 진화 과정 중에서 글루탐산이 아스파르트산과 같은 다른 아미노산으로 변화한다면, 이와 상호작용하는 아미노산인 81 번째 라이신 역시 상호작용이 그대로 유지될 수 있도록 다른 아미노산으로 바뀌어야 한다는 것이다. 이렇게 진화 과정 사이에서 서로 같이 변하는 (공변화 Covariation 라고 칭한다) 아미노산들은 단백질의 3 차원 구조를 암시하는 정보이며, 알파폴드를 포함한 대부분의 단백질 구조 예측 알고리즘은 이러한 단백질 진화 정보 속에 숨겨진 단백질의 3 차원 구조를 이용하여 구조를 예측한다.

이러한 아미노산 간의 공변화 관계는 하나의 단백질 내부에서만 존재하는 것이 아닌, 세포 내에서 서로 결합하는 다른 단백질 사이에서도 존재하게 된다. 이러한 공변화 정보를 이용하여 단백질의 구조를 예측하는 알파폴드는 서로 같이 진화해 온 단백질 간의 상호작용 역시 비교적 정확히 예측할 수 있는 것이다.

알파폴드의 '단백질 상호 작용 예측' 기능을 이용하여 많은 연구자들은 특정한 단백질끼리 상호작용하는지를 미리 테스트하기 시작했고, 이러한 테스트는 실험으로 검증되기 시작하였다.

기존에 어떤 단백질이 상호작용하는지를 알아보기 위해서는 여러가지 번거로운 실험이 필요했던 것에 비해서, 단백질 간의 상호작용을 컴퓨터에서 예측할 수 있다는 것은 매우 강력한 파급효과를 가져왔다. 단백질 여러 개로 구성된 단백질 복합체 역시, 하나 하나씩 상호작용을 예측하여 구조를 예측할 수 있다는 것은 매우 큰 의미를 가진다.

현존하는 인공지능 기반의 단백질 구조 예측 방법의 한계는 무엇인가?

인공지능 기반의 단백질 구조 예측 알고리즘은 이전의 물리 기반 / 상동 구조 기반의 알고리즘에 비해서 매우 정확한 단백질 구조를 예측한다. 그러나 이들이 아직 극복하지 못한 한계점은 존재한다. 이 중 하나는 단백질이 가질 수 있는 여러 개의 구조 상태를 예측하는 것이다. 많은 단백질들은 상황에 따라서 구조가 변화하며, 이러한 것은 해당 단백질의 기능에 매우 큰 영향을 미친다. 가령 혈액 내에서 산소를 결합하여 운반하는 단백질인 헤모글로빈 (Hemoglobin) 은 산소가 결합한 상태와 그렇지 않은 상태가 구조가 바뀌며, 이러한 구조의 변화가 혈액 중에서 산소를 운반하는 성질과 직접적인 관계가 있다. 세포 밖의 신호를 받아들여서 세포 안으로 전달하는 주된 단백질인 G 단백질 연계 수용체 (GPCR) 는 우리 몸 속에 800 종이 있으며, 이들은 수많은 약물의 표적이 되는데, 이들 역시 인지하는 물질에 결합하여 활성화되었을 때와 비활성화된 상태가 서로 구조가 다르다.

그러나 현존하는 단백질 구조 예측 방법은 단백질 구조 데이터베이스에 올라와 있는 여러가지 상태의 단백질 구조를 물리적인 방법이 아닌 뉴럴 네트워크에 의해서 참고하여 구조를 예측하므로, 이로 인해 예측되는 구조는 세포 내에서 가질 수 있는 상태 중의 한 종류 (좀 더 많은 구조의 예가 데이터베이스에 올라와 있는 상태), 혹은 그 중간 상태의 하나로 예측되어 버린다. 이러한 상황은 단백질이 여러가지 상황에서 어떻게 구조를 변화시키면서 생물학적인 기능을 수행하는지를 구조 예측으로 아는 데는 한계가 있다는 것이다.

또 다른 한계라면 현재의 구조 예측은 단백질 가닥의 구조만을 예측할 수 있으며, 단백질과 상호작용할 수 있는 수많은 생체 물질 (당, DNA, RNA, 여러가지 대사물질 등등) 과의 상호작용을 예측할 수 없다는 것이다. 많은 단백질들은 이러한 단백질 이외의 물질과 상호작용하면서 생명 현상에 참여하는데, 이러한 것의 예측은 현재의 단백질 구조 예측 방법에서 제공되고 있지 않다. 물론 단백질과 이외의 물질 간의 상호작용을 예측하는 방법들도 개발되고는 있지만, 아직 그 정확도가 현재의 단백질 구조 예측 방법론처럼 정확하지는 않다.

결국 이러한 문제는 인공지능 기반의 구조 예측 알고리즘은 현재까지 실험적으로 규명된 구조 정보에 기반한 예측이며, 단백질과 여러 생체 물질의 물리적인 특성은 감안하지 않기에 생기는 한계인 셈이다.

결론적으로 현재 존재하는 단백질 구조 예측 방법은 분명히 구조생물학자 및 단백질 구조의 이용자인 생물학자들에게 큰 혁신을 가져오긴 했지만, 아직 실험적인 구조 규명이 전혀 필요 없을 정도는 아니다. 구조생물학자가 아닌 많은 생물학자들은 알파폴드가 등장한 이후 단백질 구조를 실험적으로 규명하는 구조생물학자는 이제 할 일이 없어지는 것이 아닐까 하는 생각을

하곤 하지만, 아직 실험적인 구조 규명이 필요한 분야는 많이 남아 있다. 설령 이러한 예측이 발전하여 지금 한계로 지적되는 부분에서도 비교적 정확한 예측이 가능하게 되는 시점에서도 실험적으로 이러한 예측을 확인해야 할 필요성은 존재한다.

알파폴드는 과연 신약 개발에 큰 보탬이 되는가?

알파폴드가 등장한 이후 알파폴드가 신약 개발의 게임 체인저가 될 것이라는 류의 기사가 많이 등장했다. 오랫동안 세기의 난제로 남아있던 단백질 구조 예측 문제를 AI 의 힘으로 하루아침에 해결했으므로 이제 신약개발 역시 이에 비견할 속도로 빨라질 것이라는 류의 이야기이다. 그러나 이러한 희망적인 실제 신약 개발의 현장에서 보면 상당한 과장이 들어 있다.

신약개발은 여러가지 조건을 최적화하는 문제로써, 약물의 표적이 되는 단백질의 구조는 신약개발에 필요한 중요한 정보이긴 하지만, 이것이 오늘날 신약개발의 병목 지점이라고 보기는 힘들다. 즉, 특정한 표적 단백질에 결합하여 이를 저해하는 물질을 찾는 단계는 신약 개발의 첫 단계에 불과하고, 신약 개발에서 정말 많은 시간과 비용이 소모되는 단계는 가장 후반부 단계인 개발된 신약 후보물질이 과연 인체에서 효과를 가지는지, 독성이나 부작용은 없는지 등을 인간을 대상으로 시험하는 임상 시험 단계이다.

그리고 현행의 단백질 구조 예측 방법으로 예측된 단백질 구조는 이전의 예측 방법에 비해서 정확도가 대폭 향상되기는 했지만, 신약 개발에 그대로 사용하기에는 많은 한계가 있다. 단백질 구조는 약물과 결합하는 상태에서는 미세한 구조의 변화를 일으키고, 이러한 미세한 변화는 약물이 단백질을 결합하는지에 큰 영향을 미치는데, 이러한 약물에 의한 단백질 구조의 변화는 현재의 인공지능 기반의 구조 예측 방법만으로 쉽게 예측하기 힘들다. 실제로 알파폴드로 예측된 단백질 구조가 화합물의 결합을 예측하는데 그다지 최적이 아니라는 연구 결과도 나왔다. 알파폴드에 의해서 예측된 단백질 구조나 실험으로 결정된 단백질 구조를 이용하여 화합물이 어떻게 단백질에 결합하는지를 테스트한 결과, 알파폴드로 예측한 단백질 구조는 화합물이 붙어 있지 않은 상태로 실험으로 결정된 단백질 구조와 비슷한 정확도를 기록했다. 문제는 화합물이 붙어 있는 상태에서 실험으로 결정된 단백질 구조는 다른 두 경우보다 훨씬 높은 정확도를 기록했다는 것이다. 약물이 붙게 되면 단백질 구조가 미세하게 달라지는 것 때문에 알파폴드로 예측한 구조는 생각만큼 약물 탐색에 유용하지는 않다는 것을 의미한다.

그러나 이러한 단백질 구조 예측 기술을 기반으로 신약 개발에 도전하는 시도는 계속 이어지고 있다. 딥마인드의 창립자인 데미스 허사비스는 2021 년 인공지능 기술을 기반으로 신약을 개발하려는 구글 산하의 스타트업인 '이소모픽 랩' (Isomorphic Lab) 을 설립하였다. 이들이 구체적으로 어떤 방법론으로 신약을 개발할지는 아직 확실하지 않다. 분명한 것은 알파폴드를 중심으로 한 인공지능 기술이 중요한 역할을 할 것이라는 정도이다.

단백질 구조 예측은 우리의 미래를 어떻게 바꿀 것인가?

분명한 것은 정확한 단백질 구조 예측은 당장의 생명과학 연구의 속도를 증가시키는 혁신적인 발전임에는 분명하지만, 이러한 발전이 어떻게 신약 개발과 같이 대중이 체감할 수 있는 혁신으로 이어질지는 조금 기다려 보아야 한다는 것이다. 흥미롭게도 단백질 구조 예측의 발전은 소분자 신약 개발과 같이 이전에 단백질 구조 예측이 가능해지면 용이하게 도리 것이라고 생각되던 분야보다는 기존에는 기대하지 않았던 새로운 응용분야에서 먼저 가시적인 성과를 내고 있다.

그 중의 하나가 '단백질 디자인' (Protein Design)이다. 단백질 디자인은 간단하게 말하면 아미노산 서열 로부터 단백질 3 차원 구조를 예측하는 구조 예측 과정의 역반응으로써, 특정한 단백질 구조를 만들 수 있는 아미노산 서열을 예측하는 방법이다. 이전에는 단백질 구조의 예측 자체가 어려웠으므로, 구조 예측의 역함수인 단백질 디자인 역시 정확하게 되지 못했다. 그러나 단백질 구조 예측이 정확해진 지금에는 특정한 구조를 가진 단백질을 만들 수 있는 아미노산의 서열을 찾아내는 일 자체가 매우 간단해지게 된다.

게다가 최근 발전한 이미지 생성 인공지능의 발전과 더불어, 이 원리를 단백질에 적용하여 자연계에는 존재할 법하지만, 실제로 발견되지는 않은 인공 단백질의 구조를 생성하는 방법들도 등장하였다. 이러한 기술들이 접목되어 자연계에 존재하지 않았던 구조를 가진 전혀 새로운 단백질을 디자인하는 기술들 역시 등장하였다. 단백질은 생명의 기본 단위인 세포 속에서 '부품'으로 작용한다. 그런데 인간이 진화 과정을 뛰어넘어 자연계에 존재하지 않는 단백질을 디자인하여 만들고, 여기에 기능을 불어넣을 수 있게 되었다는 것은 매우 중요한 의미를 가진다. 그동안의 '유전공학' 기술은 결국 자연계의 생물이 가진 유전자를 약간 뜯어고치거나 오타를 수정하는 정도의 수준이었다면, 단백질 디자인은 생명체라는 텍스트에서 '문장' 의 역할을 하는 단백질을 인공적으로 만들 수 있는 기술인 셈이다.

이러한 기술이 단백질 구조 예측의 발전으로 매우 용이하게 수행될 정도로 발전하였다는 것은 어떤 의미일까? 인류는 드디어 생명의 언어를 문장 단위로 쓸 수 있게 되었다는 것을 의미한다. 자연계에 존재하는 단백질 정도를 겨우 다룰 수 있던 인류가 원하는 특성을 가진 단백질을 스스로 만들어 낼 수 있게 된 순간은 마치 1950 년대 트랜지스터가 발명된 순간과 비슷한 상황인 셈이다. 트랜지스터 발명 이후 전자공학이 급속도로 발전하여 21 세기의 IT 문명의 근간이 된 것과 마찬가지로 생명의 기본 부품인 단백질을 자유자재로 만들어 낼 수 있게 된 이후 세상의 변화 속도는 엄청나게 빨라질 것이다.

참고문헌

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.
2. Callaway, E. (2022). What's next for the AI protein-folding revolution. *Nature*, *604*, 234-238.

3. Mosalaganti, S., Obarska-Kosinska, A., Siggel, M., Taniguchi, R., Turoňová, B., Zimmerli, C. E., ... & Beck, M. (2022). AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science*, *376*(6598), eabm9506.
4. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., ... & Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. *biorxiv*, 2021-10.
5. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., ... & Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, 1-3