# 보도자료

미래를 개척하는 지식 공동체
서 울 대 학 교
SEOUL NATIONAL UNIVERSITY

| | |
|---|---|
| 보도일시 | 즉시보도 |
| | 2023. 9. 14.(목) |
| 문의 | 연구단장/연구책임자 생명과학부 마틴 스타이네거 교수(02-880-4438)/교신저자 |
| | 연구단/연구진 여진기 연구원(02-880-4439 / 공동 제1저자 |

## ■ 제목/부제

| | | |
|---|---|---|
| 제목 | *국문* | **2억개 단백질 예측 구조에서 새로운 진화 연결고리 제시하다** |
| | *영문* | Proposing New Evolutionary Connections from 214 Million Predicted Protein Structure Database |
| 부제 | *국문* | Foldseek cluster를 통한 대규모 단백질 데이터 연구 |
| | *영문* | Large-Scale Protein Data Research with Foldseek Cluster |

## ■ 요약

| | |
|---|---|
| 연구<br>필요성 | 단백질 구조를 통해 알 수 있는 단백질의 기능과 진화에 대한 정보는 모든 생명 현상에 밀접한 연관이 있다. 기존에 많은 시간과 노력을 들여 밝히던 단백질 구조를 **최근 알파폴드와 같은 인공지능 소프트웨어를 통해 빠르게 예측**할 수 있게 되었다. 그렇게 예측된 단백질 구조 데이터가 2억 개에 달한다. 이 데이터를 종합적으로 분석한다면 단백질과 생명 현상에 대한 새로운 이해를 도출할 수 있다. |
| 연구성과/<br>기대효과 | **기존 기술로는 2억 개의 단백질 구조 데이터를 클러스터링 하는 데에 10년 이상이 걸렸다. 우리 연구실에서 같은 연산이 5일이 걸릴 수 있도록 폴드시크 클러스터(Foldseek Cluster)라는 알고리즘을 개발**하였고, 대형 데이터베이스에 대한 전체적인 시각에서 연구를 할 수 있었다.<br>기존에 연구되지 않은 단백질 군을 밝히고, 새롭게 단백질에 기능 주석을 달았다. 특히, 인간 면역 단백질과 구조적으로 유사한 박테리아 단백질을 찾아 인간 면역 기능의 새로운 진화 가능성을 보여주었다. 이러한 발견은 진화생물학, 의학 등 다양한 분야에서 활용될 수 있을 것으로 기대된다. |

| | |
|---|---|
| Abstract<br>(영문) | Proteins are vital for cellular processes, and understanding their structure aids in studying function and evolution. AlphaFold's database offers 214 million predicted protein structures. We developed Foldseek cluster, clustering millions of structures. It revealed 2.27M structural clusters, 31% of which lack annotations, potentially indicating novel structures. Evolutionary analysis suggests ancient origins, with 4% being species-specific. This resource also helps predict domain families and their relationships, uncovering remote homology examples. Notably, human immune-related proteins share structural similarities with prokaryotic species, showcasing the resource's value in studying protein function and evolution across the tree of life. |
| Journal Link | https://doi.org/10.1101/2023.03.09.531927 |

## ■ 본문

소위 “탄단지”, 3대 영양소로 더 친숙한 단백질은 세포를 구성하고 작동시키는 일종의 “분자 기계”로 에너지 생성, DNA 복제, 질병에 대한 방어작용 및 여러 생물학적 과정에서 핵심적인 역할을 수행한다. 분자 기계로서 단백질은 각 기능에 맞는 다양한 종류의 크기, 구조를 갖는다. 서울대학교와 취리히 연방 공과대학교(ETH Zurich) 및 유럽 생물정보학 연구소(EMBL-EBI) 팀은 **수억 개의 단백질 구조를 비교할 수 있는 새로운 접근 방식을 Nature 저널에 발표**했으며, 이를 활용하여 자연에 존재하는 다양한 유형의 단백질 구조를 구조적 유사성에 따라 분류하였다. 본 연구에서 연구진은 새로운 분석 방식을 통해 기존에 알려지지 않았던 면역 체계에 관여하는 인간 단백질과 박테리아 종에서 발견되는 단백질 사이에 유사성을 발견했다.

단백질은 형형색색 구슬로 이어진 팔찌처럼 아미노산 사슬로 만들어진 다음 3차원 구조로 접혀 기능을 수행하게 된다. 전체 생물종에서는 2억 개가 넘는 단백질이 알려져 있고, 사람의 경우만 해도 약 20,000가지의 단백질 유전정보를 갖고 있다. 유전체에 존재하는 단백질 서열을 해독하는 것은 상대적으로 쉬운 일이었지만, 서열로부터 각 단백질의 3차원 구조를 결정하는 것은 훨씬 더 어려운 작업이었다. 하지만 인공지능 기반의 AlphaFold2는 50여년 간 난제였던 단백질 구조 예측의 정확도를 획기적으로 높였고, 이 예측 기술을 사용하여 딥마인드(DeepMind)와 EMBL-EBI의 연구 그룹은 2억 개가 넘는 단백질 구조의 데이터베이스를 구축할 수 있었다. 딥마인드의 연구는 플라스틱 분해 효소와 항생제 내성 극복 효과를 지닌 단백질 등 개별 구조 분석을 통해 많은 발견을 이루어냈지만, 더 큰 잠재력을 가진 종합적인 분석에는 미흡했다.

서울대학교 연구팀은 수억 개 이상의 단백질 구조에 대해 효율적으로 유사성을 계산할 수 있는 방법을 개발했고, 이를 활용하여 총 2억 개가 넘는 구조들이 200만 개 가량의 서로 다른 클러스

터로 그룹화될 수 있음을 발견했다. “기존의 방법을 사용하면 2억 개 단백질의 상호 유사성을 계산하는데 10년 가량 걸릴 것으로 추정되지만, 우리의 새로운 방법을 사용하면 5일 만에 이를 끝낼 수 있었습니다.“라고 서울대학교 마틴 스타이네거(Martin Steinegger) 교수는 밝혔다.

일반적으로 진화 과정에서 단백질의 서열이 구조보다 변화가 많고 변화의 속도 또한 더 빠른 것으로 알려져 있다. 이로 인해 기존 서열 비교를 통해서는 단백질 간 관계가 명확하지 않은 경우에도, 구조를 비교하면 유연관계(조상-후손관계)를 밝혀 먼 친척에 해당되는 단백질을 식별할 수 있게 된다. 즉, 이와 같은 구조 유사성 연구를 통해 단백질의 고대 진화에 관한 간접적 증거를 찾을 수 있다. “먼 옛날 단백질들이 언제, 어떻게 분화가 시작되었는지를 되돌아볼 수 있는 일종의 진화 역사에 대한 타임머신을 발견한 것과 같습니다.“ 라고 취리히 대학 페드로 벨트라오(Pedro Beltrao) 교수는 말했다.

연구팀의 분석 결과, 인간의 면역 관련 단백질 중 많은 수가 박테리아에서 발견되는 단백질과 비슷한 3차원 형태를 갖고 있음을 발견했다. 이러한 점은 우리의 면역체계가 기존의 가설보다 훨씬 오래된 단백질에서 분화되었을 수 있으며 병원체에 맞서 싸우는 메커니즘이 종 간에 더 광범위하게 공유될 수 있다는 것을 시사한다.

본 연구는 면역 관련 단백질을 포함한 전체 단백질의 진화에 대한 새로운 가설을 제시했을 뿐 아니라 전체 데이터를 공개함으로써, 진화생물학, 의학 등 다양한 분야에서 더욱 흥미로운 발견의 기대를 모으고 있다. 연구결과에 대한 상세설명은 [Nature, 2023, 10.1038/s41586-023-06510-w]의 출판물을 참조.

## □ 연구결과

**Clustering-predicted structures at the scale of the known protein universe**

Inigo Barrio-Hernandez*, Jingi Yeo*, Jürgen Jänes, Milot Mirdita, Cameron L. M. Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao and Martin Steinegger

(Nature, in press)

단백질 구조를 통해 알 수 있는 단백질의 기능과 진화에 대한 정보는 모든 생명현상에 밀접한 연관이 있다. 인공지능으로 예측된 2억 개의 거대한 단백질 구조 데이터는 생물학적 발견의 잠재력을 가지고 있다. 본 연구에서는 2억 개의 데이터베이스를 모두 클러스터링 하여 전체적이고 큰 규모의 시각으로 분석하였다. 단백질 진화의 새로운 종 간 연결과 탐색되지 않은 미지의 단백질과 도메인에 대한 연구 결과를 제시하여 의학, 약학 등 다양한 분야에서 활용될 것으로 기대된다.