보도자료

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

보도일시	즉시 보도
	2023. 5. 10.(수)
문의	담당자: 마틴 스타이네거 교수 (02-880-4438)
	연구단장/연구책임자 마틴 스타이네거 교수(02-880-4438) / 교신저자
	연구단/연구진 마틴 스타이네거 교수 (02-880-4438) / 교신저자

■ 제목

Foldseek 을 이용한 빠르고 정확한 단백질 구조 검색

■ 요약

연구 필요성	AlphaFold2와 ESMFold 같은 단백질 구조 예측 방법의 발전은 공개적으로 이용 가능한 단백질 구조 데이터베이스 크기의 전례 없는 성장을 이끌었다. 단백질의 구조는 단백질의 아미노산 서열보다 진화 과정 속에서 더 잘 유지되기 때문에, 구조를 비교하면 서열을 비교할 때보다 유연 관계가 먼 단백질 사이의 유사성을 더 잘 감지할 수 있다. 단백질 구조의 이런 특성을 활용한다면, 방대한 단백질 구조 데이터베이스에서 연구자들은 더 많은 양의 유연 관계가 먼 단백질들을 찾을 수 있으며, 이를 통해 새로운 단백질의 기능을 기존에 알고 있는 단백질로의 기능으로부터 유추하는 등 단백질 기능 연구 분야에 큰 도움을 줄 것으로 예상된다. 하지만 기존의 방법으로는 2억 개 이상의 단백질 구조를 포함하는 방대한 데이터베이스 전체를 한 번 탐색하는데 몇 달이 걸려 이런 계산 속도 문제를 해결하기 위한 방안이 필요하였다.
연구성과/ 기대효과	서울대학교 마틴 스타이네거 연구팀과 괴팅겐 막스 플랑크 연구소가 개발한 Foldseek (foldseek.com)은 매우 빠른 검색 속도를 최대의 장점으로 하는 혁신적인 도구이다. 검색 민감도는 기존의 가장 민감한 도구보다 미세하게 낮지만, Foldseek은 검색 시간을 몇 달에서 몇 초로 단축시켰다. Foldseek 의 빠른 속도와 우수한 민감도는 메타 유전체학, 분자 의학 및 환경 유전체학과 같은 분야에 결정적인 영향을 미칠 것으로 기대되며, 더 나아가 다양한 기능과 편리한 웹 서버 (search.foldseek.com) 를 갖추었으며, 연구 목적에 맞게 조정 가능한 사용자 정의 워크플로우를 만들 수 있어 향후 구조 기반 단백질 분석에 있어 중요한 도구로 자리매김할 것으로 기대된다.

■ 본문

□ **본문**

서울대 마틴 스타이네거 교수팀, 초고속 단백질 구조 검색엔진 "Foldseek" 개발,
기존 대비 10만배 속도 향상으로 초격차 벌려

AlphaFold2와 ESMFold 같은 단백질 구조 예측 방법의 발전은 공개적으로 이용 가능한 단백질 구조 데이터베이스 크기의 전례 없는 성장을 이끌었다. 단백질의 구조는 단백질의 아미노산 서열보다 진화 과정 속에서 더 잘 유지되기 때문에, 구조를 비교하면 서열을 비교할 때보다 유연 관계가 먼 단백질 사이의 유사성을 더 잘 감지할 수 있다. 단백질 구조의 이런 특성을 활용한다면, 방대한 단백질 구조 데이터베이스에서 연구자들은 더 많은 양의 유연 관계가 먼 단백질들을 찾을 수 있으며, 이를 통해 새로운 단백질의 기능을 기존에 알고 있는 단백질로의 기능으로부터 유추하는 등 단백질 기능 연구 분야에 큰 도움을 줄 것으로 예상된다.

하지만 기존의 방법으로는 2억 개 이상의 단백질 구조를 포함하는 방대한 데이터베이스 전체를 한 번 탐색하는 데 몇 달이 걸린다. 이런 계산 속도)문제를 해결하기 위해 서울대학교 마틴 스타이네거 연구팀과 괴팅겐 막스 플랑크 연구소는 Foldseek (foldseek.com) 이라는 혁신적인 단백질 구조 검색 도구를 개발했다. 이 혁신적인 도구의 최대 장점은 매우 빠른 검색 속도이다. 검색 민감도는 기존의 가장 민감한 도구보다 미세하게 낮지만, Foldseek은 검색 시간을 몇 달에서 몇 초로 단축시켰다.

빠른 속도를 달성하기 위해 Foldseek는 간단한 트릭을 활용했다. 단백질은 아미노산 사슬이 접혀서 안정화된 3차원 구조를 이루게 되는데, 연구진은 그 3차원 구조를 하나의 서열로 표현해내는 방법을 개발했다. 연구진은 단백질의 3차원 구조 속에서 각 아미노산이 주변 아미노산과 가지는 위치 관계들을 스무개의 특정한 패턴을 기준으로 나눈 뒤 각 집합에 구조 알파벳을 부여했다. 그런 다음, 단백질의 아미노산 서열을 구조 알파벳의 서열로 변환하여 단백질의 구조를 담은 서열을 얻어내었다. Foldseek은 이렇게 얻은 구조를 담은 서열을 초고속 서열 검색 도구로 비교하여 3D 구조를 직접 비교하는 것보다 훨씬 빠른 속도를 달성했다.

Foldseek의 빠른 속도와 우수한 민감도는 메타 유전체학, 분자 의학 및 환경 유전체학과 같은 분야에 결정적인 영향을 미칠 것으로 기대된다. 더 나아가, 다양한 기능과 편리한 웹 서버 (search.foldseek.com) 를 갖추었으며, 연구 목적에 맞게 조정 가능한 사용자 정의 워크플로우를 만들 수 있도록 하였다. 이런 사용자 편의성과 10만 배 빠른 구조 검색 속도를 바탕으로 Foldseek은 향후 구조 기반 단백질 분석에 있어 매우 중요한 도구로 자리매김 할 것으로 기대된다. 자세한 정보는 [Nature Biotechnology, 2023, 10.1038/s41587-023-01773-0]의 최근 발표를 참조.

□ **연구결과**

Foldseek: fast and accurate protein structure search

Michel van Kempen, Stephanie Kim, Charlotte Tumescheit, Milot Mirdita, Johannes Söding, Martin Steinegger

(*Nature Biotechnology,* 2023, 10.1038/s41587-023-01773-1)

As structure prediction methods are generating millions of publicly available protein structures, searching these databases is becoming a bottleneck. Foldseek aligns the structure of a query protein against a database by describing the amino acid backbone of proteins as sequences over a structural alphabet. Foldseek decreases computation times by four to five orders of magnitude with 86%, 88% and 133% of the sensitivities of DALI, TM-align and CE, respectively.

□ 용어설명

단백질 상동성(Protein homology) : 진화 과정 속에서 단백질에 돌연변이가 쌓이면 비슷하지만 조금씩 다른 단백질들이 생겨나게 되는데, 이를 상동 단백질 (homologous protein) 이라고 부른다. 진화의 기간이 길거나 돌연변이가 축적되는 속도가 빠를 수록 변화가 많이 축적되어 상동 단백질 사이의 유사성은 점점 희미해지는데, 이를 원거리 상동성 (remote homology) 라고 부른다. 두 단백질의 유사성이 희미해질수록 아미노산 서열 비교를 통한 유사성 확인은 어려워진다. 반면, 단백질의 구조는 진화 과정 속에서 더 잘 유지되기 때문에, 단백질의 구조를 비교하면 원거리 상동성을 더 잘 찾아낼 수 있다.

□ 연구자

○ 성 명 : 마틴 스타이네거
○ 소 속 : 서울대학교 생명과학부 교수
○ 연락처 : 02-880-4438, martin.steinegger@snu.ac.kr

○ 성 명 : 김서우
○ 소 속 : 서울대학교 기초과학연구원
○ 연락처 : 02-880-4439, stepkim92@gmail.com

○ 성 명 : 요하네스 소딩
○ 소 속 : 막스플랑크 연구소, 괴팅겐 대학 교수
○ 연락처 : +49-551-201-2803 soeding@mpinat.mpg.de

○ 성 명 : 미헬 반 켐펜
○ 소 속 : 막스플랑크 연구소, 괴팅겐
○ 연락처 : +49-551-201-2894
michel.van-kempen@mpinat.mpg.de