KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML) Workshop for Life Scientists, Data Scientists, and Bioinformatians

생물정보학&머신러닝 워크샵(오프라인)

Protein structure prediction with AI

백민경 _ 서울대학교

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 오프라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 (2.6 월)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	개회사/공지사항전달			
09:30-10:50 (80)	**Best practice for single-cell data analysis**	**박종은 교수**	**Introduction to ML & DNN (이론)**	**이상근 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**Practice1: Scanpy basic workflow**	**김우석 김성룡 조교**	**CNN (이론)**	**이상근 교수**
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	**Public data, batch correction, cell annotation**	**박종은 교수**	**RNN, GAN, XAI (이론)**	**이상근 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**Practice2: Advanced single-cell analysis**	**김우석 김성룡 조교**	**AI 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)**	**이정현 한성민 조교**

DAY2 (2.7 화)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	**Introduction to protein structure prediction - Homology modeling - Coevolution-guided modeling Early AI-based approaches**	**백민경 교수**	**Pre-trained Models for Transfer Learning (이론)**	**전민지 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**단백질 구조 예측 실습 - MSA generation, template search - homology modeling contact prediction & modeling**	**백민경 교수**	**Pre-trained Models for Transfer Learning (실습)**	**정민수 조교**
12:10-13:40 (90)	점심			
13:40-15:10 (90)	**AI-based protein structure prediction - AlphaFold/RoseTTAFold Applications to PPI prediction & protein design**	**백민경 교수**	**Deep learning in Bioinformatics**	**노미나 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**단백질 구조 예측 실습 II AlphaFold, RoseTTAFold 실습 및 응용**	**백민경 교수**	**Deep learning model을 이용한 실습**	**곽호진 박예솔 조교**

DAY3 (2.8 수)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	**화학정보학 기초(Cheminformatics) 약물특성 및 약물다움(druglikeness) Molecular Notations & Descriptors AI 신약개발을 위한 Databases AI 신약개발을 위한 Programming 기초**	**김동섭 교수**	**마이크로바이옴 기본 이론**	**이선재 교수**
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	**Google Colab에 RDKit 설치 화합물 정보 읽기 실습 Bioactivity database 검색 및 정보 읽기 실습 Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습**	**문채영 나민주 조교**	**16S rRNA amplicon seq. - DADA2**	**서영창 조준우 조교**
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	**AI 신약개발을 위한 기계학습법 기초 QSAR 모델링 기초 AI 신약개발을 위한 딥러닝 모델 Virtual screening (ligand-based, structure-based) 및 de novo design**	**김동섭 교수**	**최신 메타지놈 분석 기법의 현황**	**이선재 교수**
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	**QSAR modeling 전체 과정 실습 화합물의 Bioactivity 예측 모델 개발 Virtual screening 과정을 통한 신약후보물질 발굴 실습**	**문채영 나민주 조교**	**Shotgun metagenome 분석 (Linux)**	**서영창 조준우 조교**

강의개요

Protein structure prediction with AI

단백질은 신호전달, 대사, 면역 등 우리 몸에서 일어나는 거의 모든 생명현상에 관여하고 있는 중요한 생체분자이다. 단백질은 각자의 기능을 수행하기 적합한 3차원 구조를 가지고 있으며, 이러한 구조는 단백질의 서열에 따라 결정되는 것으로 알려져 있다. 즉, 단백질의 기능을 잘 이해하기 위해서는 서열로부터 그 구조를 아는 것이 매우 중요하다. 단백질의 서열을 기반으로 그 3차원 구조를 정확하게 예측할 수 있다면 단백질과 연관된 수많은 생명현상에 대한 답을 찾는데 큰 도움을 주지 않을까?

본 강의에서는 단백질 구조 예측 방법이 어떻게 발전해왔는지 살펴보고, 인공지능이 단백질 구조 예측에 어떤 혁신을 가져왔는지 알아보고자 한다. 또한 인공지능 기반의 단백질 구조 예측이 단백질-단백질 상호작용 예측, 단백질 디자인과 같은 다른 연구분야에 어떤 영향을 주었는지 살펴본다. 강의에서 다루는 방법들을 실제 실습을 통해 사용해보고, 각 방법의 장단점을 알아보고자 한다.

강의는 다음의 내용을 포함한다:

- 단백질 구조 예측의 기본 원리
- 인공지능을 활용한 단백질 구조 예측
- 단백질-단백질 상호작용에의 응용
- 단백질 디자인으로의 응용

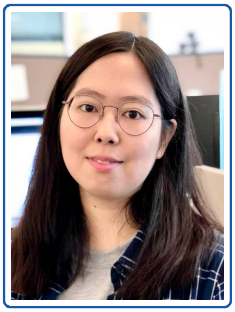
* 교육생준비물: 노트북

* 강의 난이도: 중급

* 강의: 백민경 교수 (서울대학교 생명과학부)

Curriculum Vitae

Speaker Name: Minkyung Baek, Ph.D.



▶ **Personal Info**

Name	Minkyung Baek
Title	Assistant Professor
Affiliation	Seoul National University

▶ **Contact Information**

Address	504-523, 1 Gwanak-ro, Gwanak-gu, Seoul 08826
Email	minkbaek@snu.ac.kr
Phone Number	02-880-6755

Research Interest

Structural bioinformatics, computational biology, protein structure prediction, artificial intelligence

Educational Experience

2013	B.S. in Chemistry, Seoul National University, Korea
2018	Ph.D. in Chemistry, Seoul National University, Korea

Professional Experience

2018-2019	Postdoctoral researcher, Seoul National University
2019-2022	Postdoctoral scholar, University of Washington, USA
2022-	Assistant Professor, Seoul National University

Selected Publications (5 maximum)

1. Minkyung Baek, et al., Accurate prediction of protein structures and interactions using a three-track neural network, Science, 373 (6557), 2021.
2. Ian R. Humphreys†, Jimin Pei†, Minkyung Baek†, Aditya Krishnakumar†, et al., Computed structures of core eukaryotic protein complexes, Science, 374 (6573), 2021. (†co-first authors)
3. Minkyung Baek, Ivan Anishchenko, Hahnbeom Park, Ian R. Humphreys, and David Baker, Protein oligomer modeling guided by predicted inter-chain contacts in CASP14, Proteins: Structure, Function, and Bioinformatics, 89 (12), 2021.
4. Ivan Anishchenko†, Minkyung Baek†, Hahnbeom Park†, et al., Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14, Proteins: Structure, Function, and Bioinformatics, 89 (12), 2021. (†co-first authors)

KSBi-BIML 2023

Protein structure prediction with AI

서울대학교
백민경
(minkbaek@snu.ac.kr)

제 1강.

Intro to Protein Structure Prediction

'인공지능 단백질 모델링' 2021 가장 뜨는 바이오 기술

올해의 과학 성과에 '단백질 구조 예측 인공지능'

단백질 해독은 AI가 맡는다, 양대 과학저널 동시 발표

[사이언스카페] 네이처는 딥마인드, 사이언스는 워싱턴대의 인공지능 발표

[KISTI과학향기]생물학 혁신, 단백질 구조 예측하는 AI 시대

인공지능, 단백질 구조 예측 '열공 중' …신약개발 도우미 역할

'알파폴드2가 인간 전체 단백질 구조 44% 규모 데이터 공개… '인공 단백질' 제작 가능성도

3

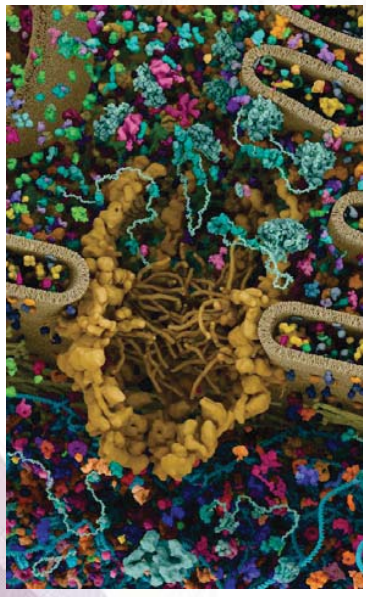
Contents

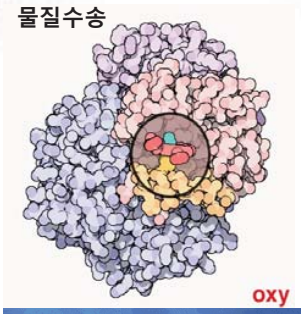
1. 단백질 접힘 (protein folding)의 화학원리 이해
2. 주형기반 단백질 구조 예측 (Homology modeling)
3. Ab initio 단백질 구조 예측 (Free modeling)
4. 유사서열(진화정보)기반 단백질 구조 예측
5. 딥러닝 기반 단백질 구조 예측 – 초기 모델

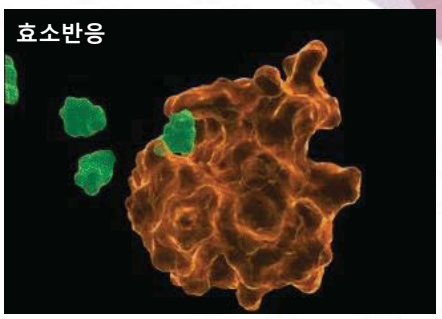
4

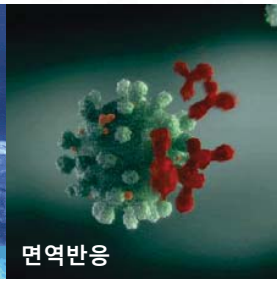
단백질?

생명현상의 핵심 분자



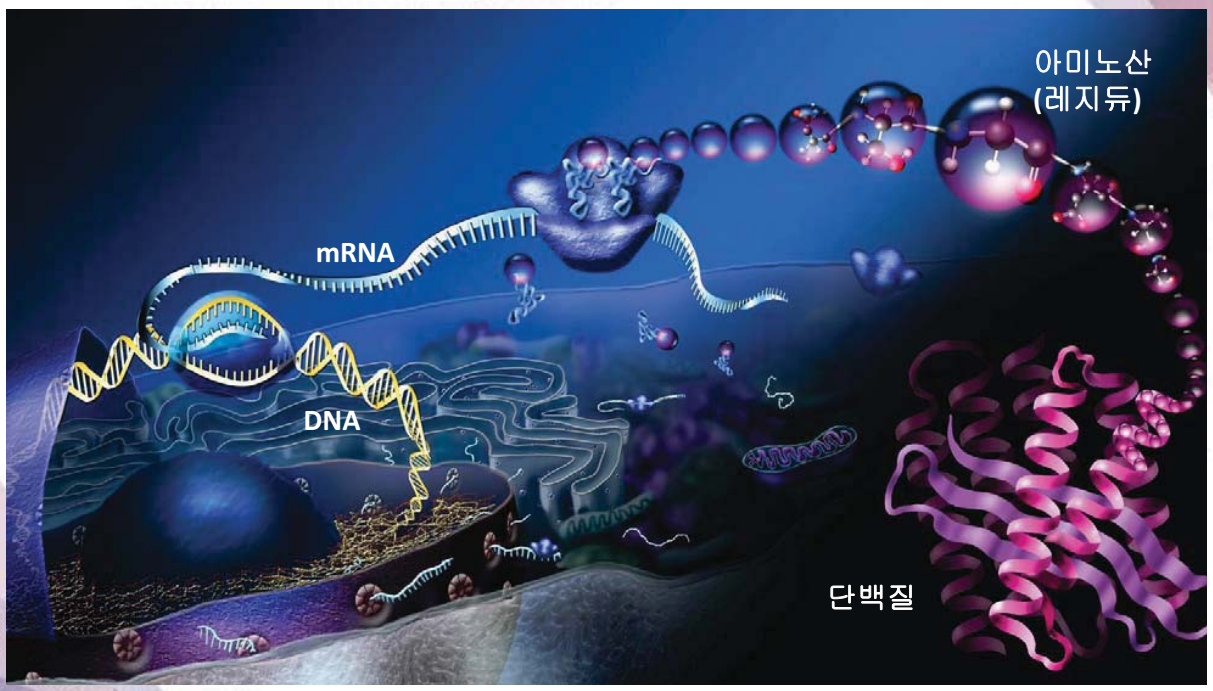






5

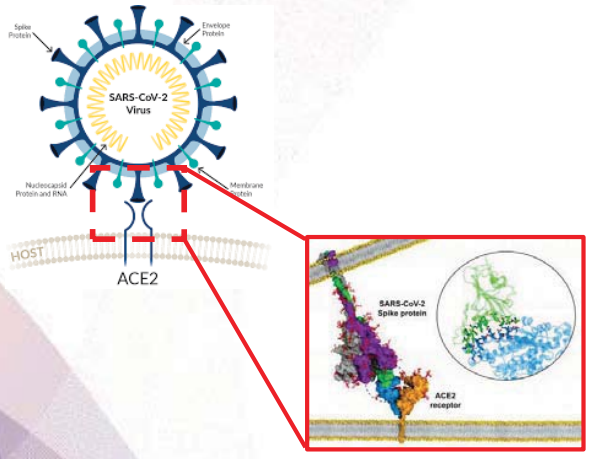
Central Dogma



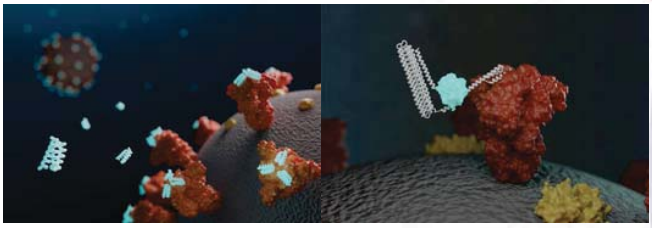
6

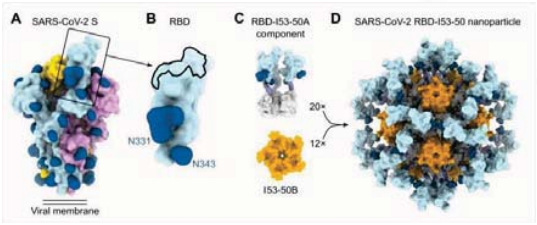
단백질의 구조를 아는 것이 중요한 이유?

생명현상에 대한 더욱 깊은 이해



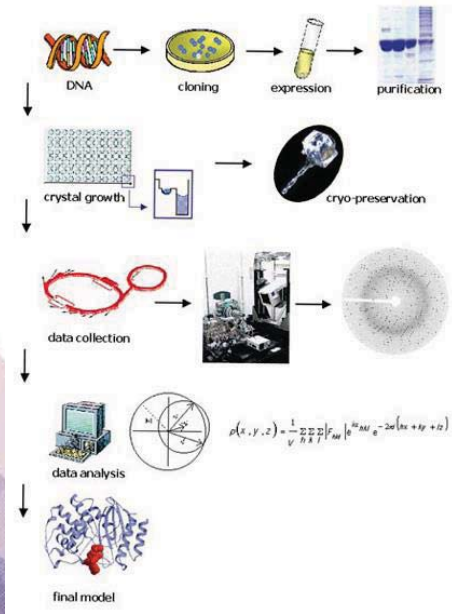
신약/백신/바이오센서 개발로의 응용



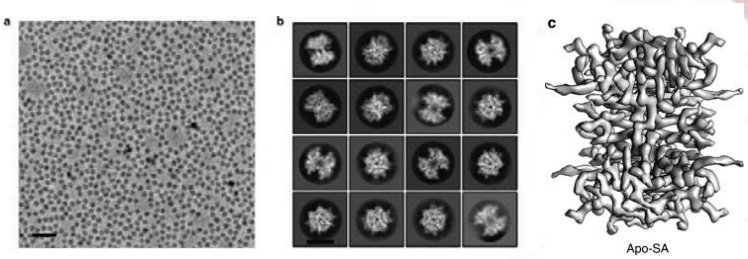


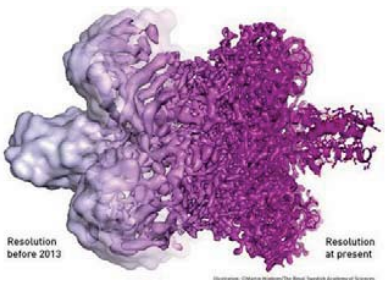
실험을 통한 단백질 구조 결정

X-ray 결정법 (노벨화학상, 1962년)



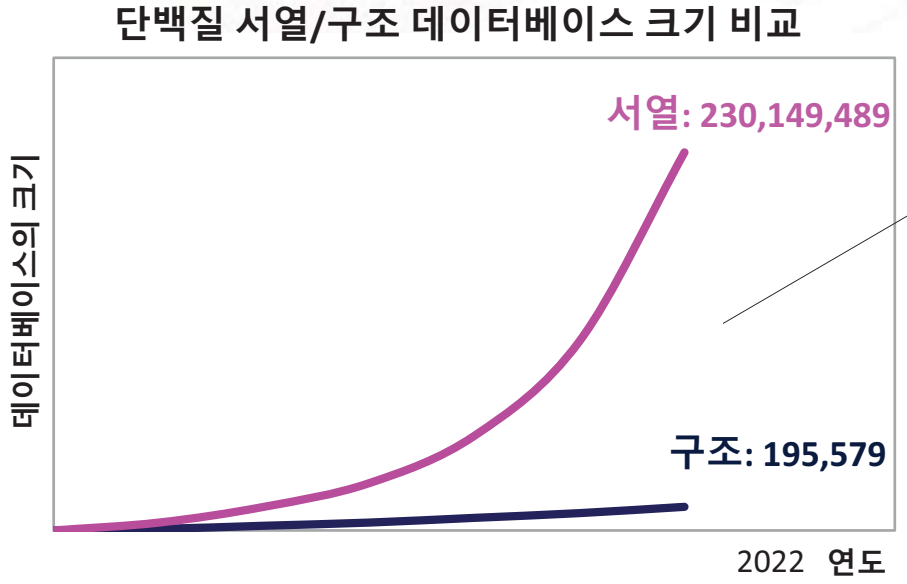
극저온전자현미경 (노벨화학상, 2017년)





비싸다
시간이 오래 걸린다
(수개월 ~ 수년)

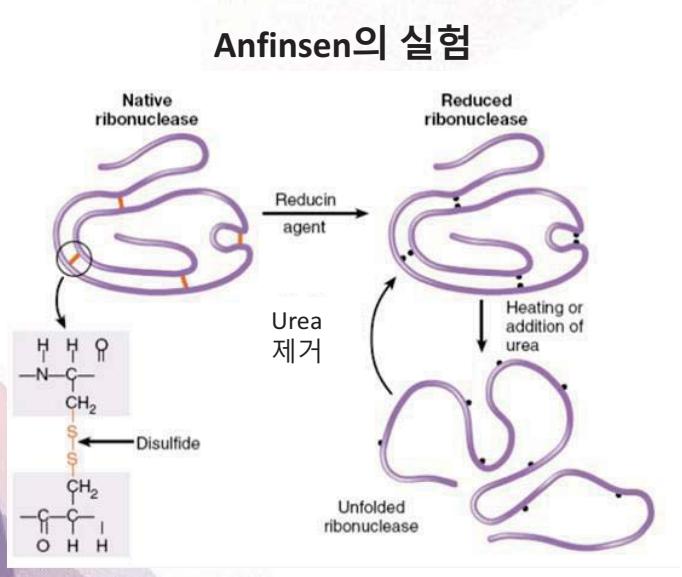
알고있는 단백질의 서열 >> 단백질 구조



컴퓨터 계산을 통한
단백질 구조 예측?

9

컴퓨터 계산을 통해 예측할 수 있을까?



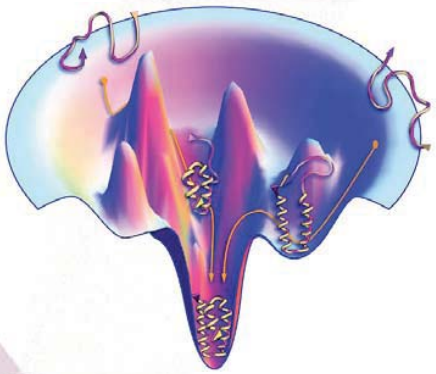
Christian B. Anfinsen
(노벨화학상, 1972)

**"단백질의 3차원 구조는
단백질의 서열에 의해 결정된다"**

10

단백질 구조 접힘의 화학 원리

화학자의 관점: 가장 안정한 상태 (가장 낮은 자유에너지 상태)를 가진다



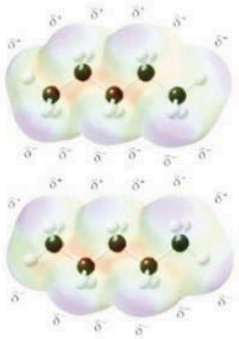
주어진 환경에서 가장 안정한 구조

반데르발스 에너지

수소결합

정전기적 상호작용

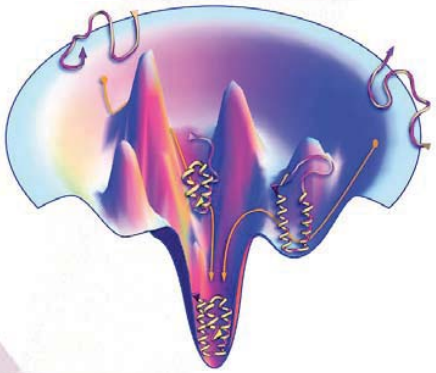
소수성 상호작용



11

단백질 구조 접힘의 화학 원리

화학자의 관점: 가장 안정한 상태 (가장 낮은 자유에너지 상태)를 가진다



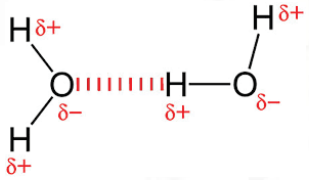
주어진 환경에서 가장 안정한 구조

반데르발스 에너지

수소결합

정전기적 상호작용

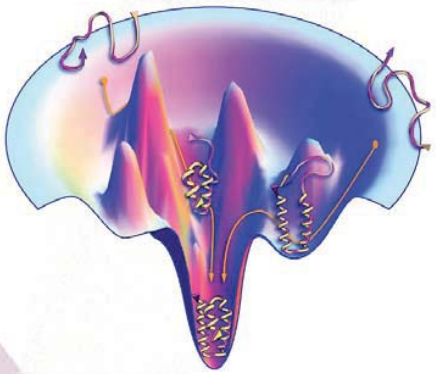
소수성 상호작용



12

단백질 구조 접힘의 화학 원리

화학자의 관점: 가장 안정한 상태 (가장 낮은 자유에너지 상태)를 가진다



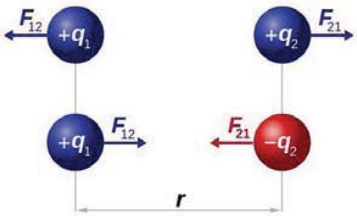
주어진 환경에서 가장 안정한 구조

반데르발스 에너지

수소결합

정전기적 상호작용

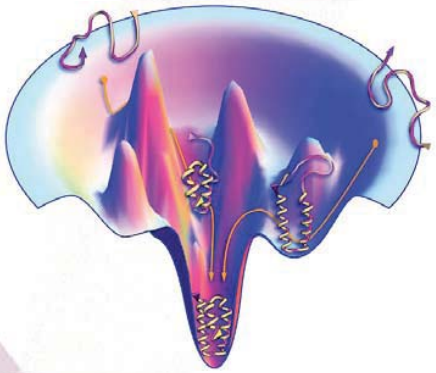
소수성 상호작용



13

단백질 구조 접힘의 화학 원리

화학자의 관점: 가장 안정한 상태 (가장 낮은 자유에너지 상태)를 가진다



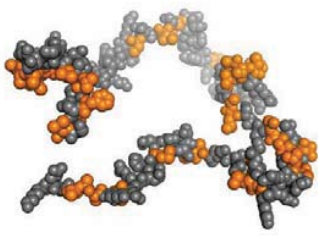
주어진 환경에서 가장 안정한 구조

반데르발스 에너지

수소결합

정전기적 상호작용

소수성 상호작용

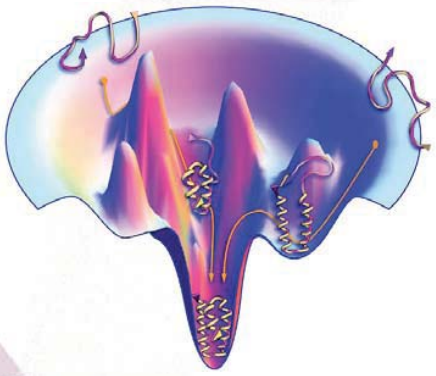


소수성 아미노산
친수성 아미노산

***Entropy*에 의한 효과**

14

MD 시뮬레이션을 통한 단백질 구조 예측?



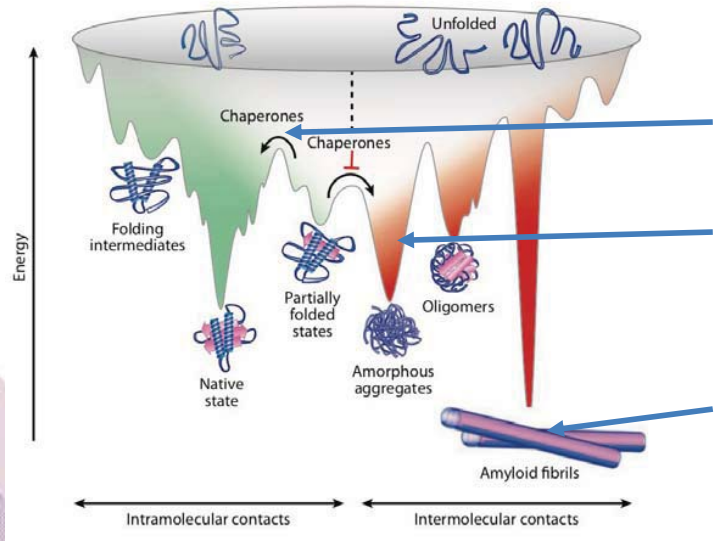
주어진 환경에서 가장 안정한 구조

단백질 접힘 분자동역학 시뮬레이션

매우 작은 단백질에 대해서만 가능 (<50 aa)
시간이 많이 소요 (days to months)
사용하는 Force field (에너지 함수)의 오류

15

샤페론에 의한 단백질 접힘



1. native state로 접히도록 도와줌

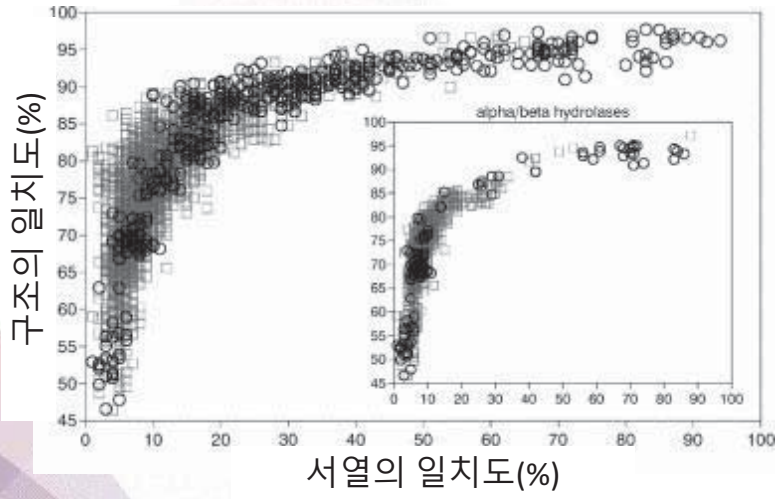
2. 단백질 aggregation 방지

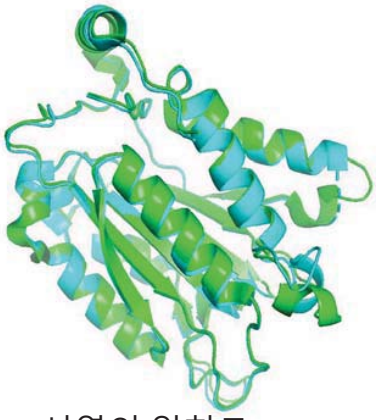
3. 질병을 유발하는 단백질 간 비정상적 상호작용 방지

16

다른 방법은 없을까?

Idea: 단백질의 서열이 비슷하면 그 구조도 비슷하지 않을까?

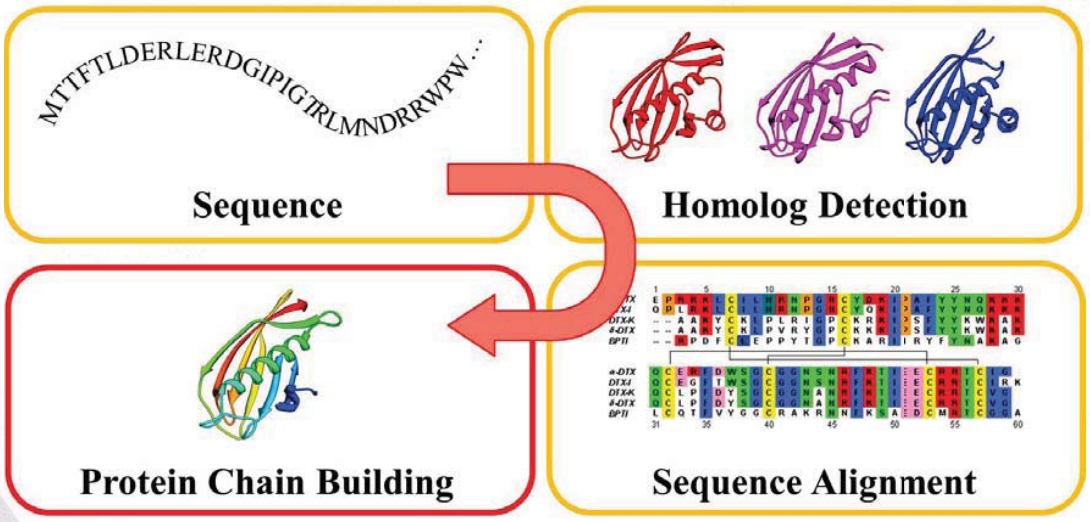




서열의 일치도: 61%

17

호몰로지 모델링 / 주형기반 단백질 구조 예측



18

주형 단백질 찾기 (homolog detection)

- 서열이 유사하면 그 구조도 비슷하다
- 서열이 비슷한 단백질을 구조 데이터베이스(PDB)에서 찾아보자.
  - 어떻게 서열을 비교할 것인가?
  - 서열 유사성이 낮은 경우 (distant homolog) 어떻게 하면 주형 단백질을 더 잘 찾을 수 있을 것인가?

주형 단백질 찾기 (homolog detection)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

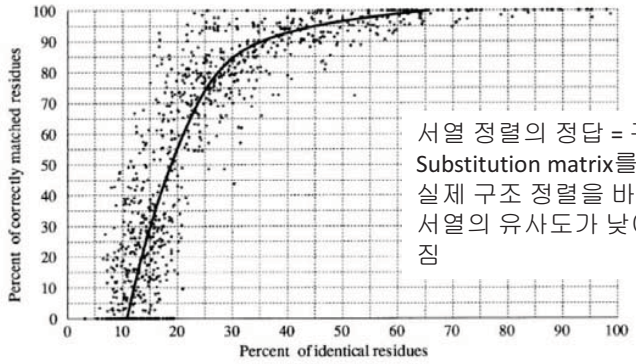
BLOSUM62와 같은 substitution matrix를
평가함수로 사용하여 서열의 유사도 평가

```
NIKE
NICE
```
6+4-3+5
=12점

```
NIKE
LIKE
```
-3+4+5+5
=11점

Distant homolog를 잘 찾아내기 위해서는?

- Substitution matrix: 일반적인 단백질에서 얻어낸 통계 정보
- 모델링하고자 하는 타겟이 속한 family의 정보를 활용하면 더 좋지 않을까?



서열 정렬의 정답 = 구조 정렬
Substitution matrix를 사용하여 정렬된 서열과 실제 구조 정렬을 바탕으로 얻어낸 서열정렬 비교
서열의 유사도가 낮아질수록 서열 정렬의 정확도도 떨어짐

Vogt et. al., JMB, 1995

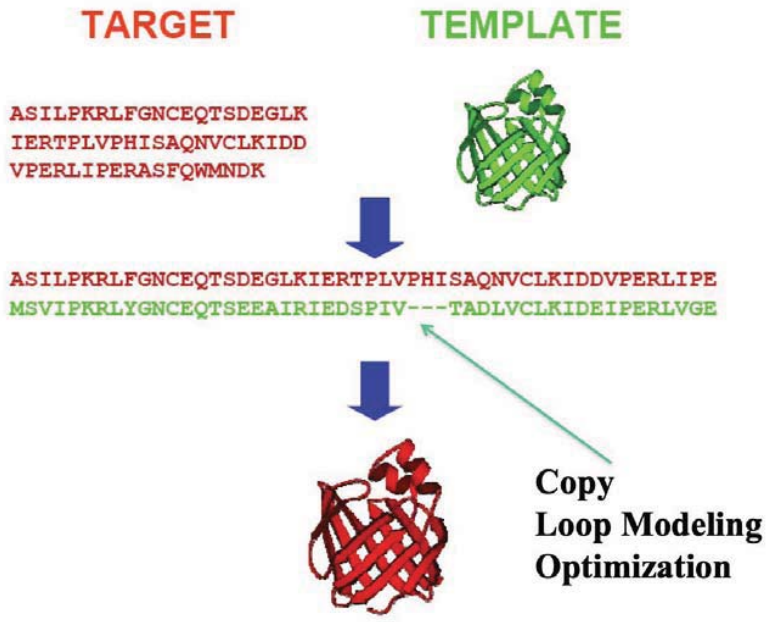
21

Sequence profile 기반 homolog search

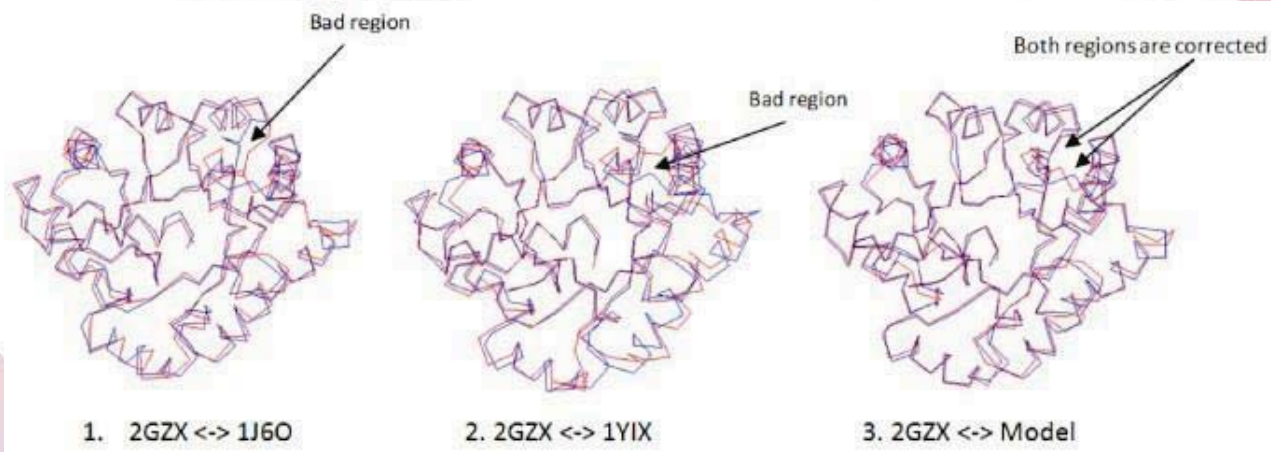
- 하나의 서열 대신 유사한 서열들로부터 얻어진 sequence profile을 사용하자
  - Profile: 유사한 서열 셋이 가지는 특성
  - 타겟 단백질과 유사한 sequence profile을 가지는 단백질: 좋은 주형단백질일 가능성이 높다!
  - Profile-profile alignment (HHsearch, CRFalign, etc)

22

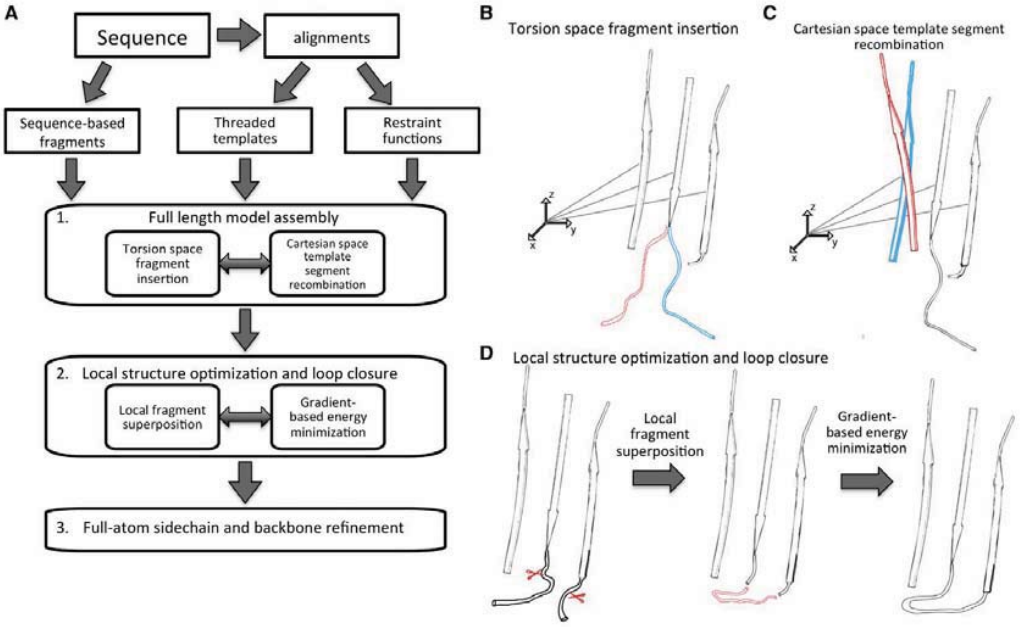
모델 구조 만들기



여러 주형 단백질을 사용하면?

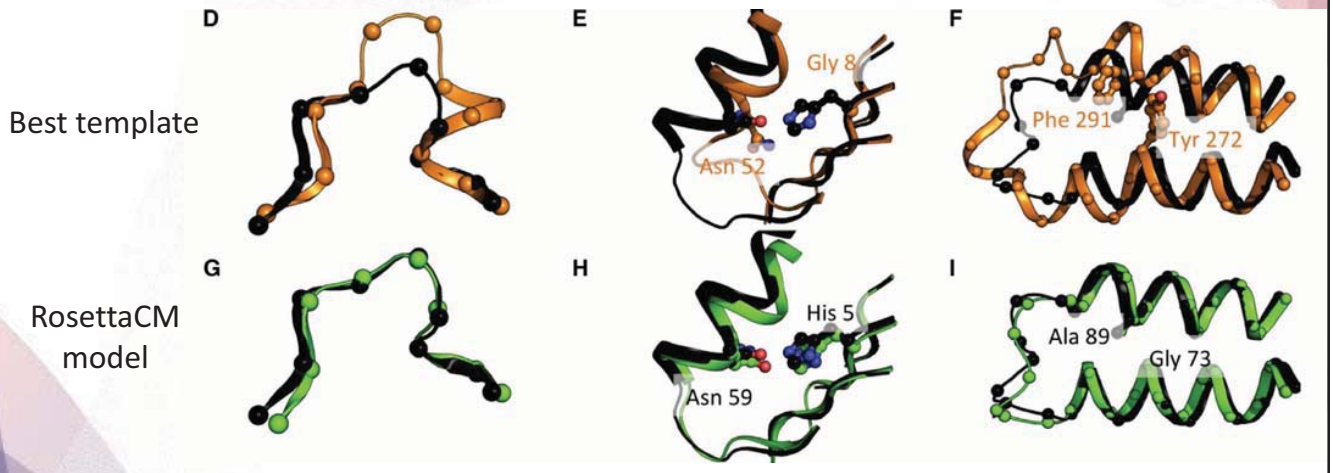


RosettaCM: Multi-template + 추가 구조 정밀화



25

RosettaCM: Multi-template + 추가 구조 정밀화

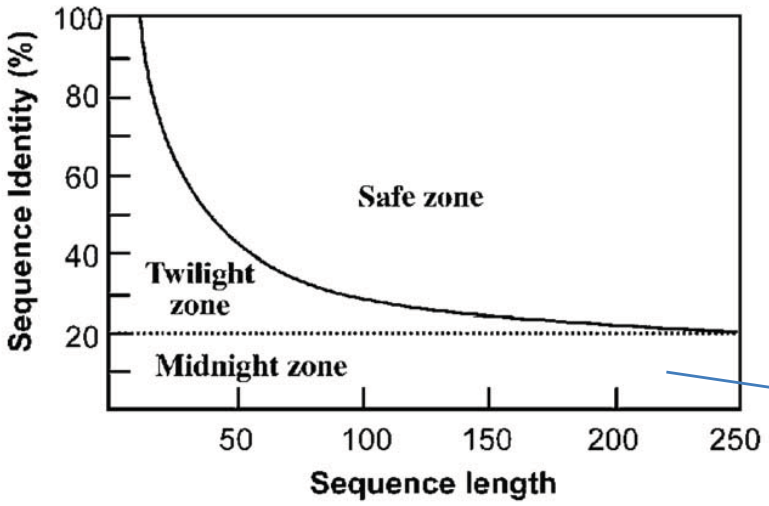


26

Song, Yifan, et al. "High-resolution comparative modeling with RosettaCM." *Structure* 21.10 (2013): 1735-1742.

호몰로지 모델링 / 주형기반 단백질 구조예측

Template 단백질을 찾을 수만 있다면, reasonable한 구조 예측 가능!



여긴 어떻게 모델링?

27

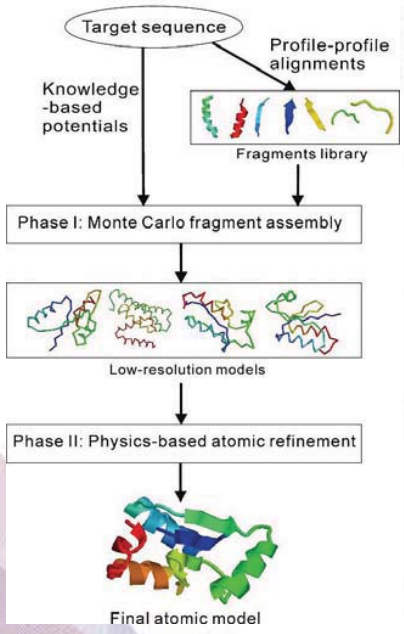
Ab initio 단백질 구조 예측 (Free modeling)

- 단백질 전체에 대한 homology detection은 실패한 경우
  - 짧은 fragment라도 서열이 비슷하면 그 local 구조는 유사하지 않을까?
  - 예측된 2차구조 정보를 활용할 수는 없을까?

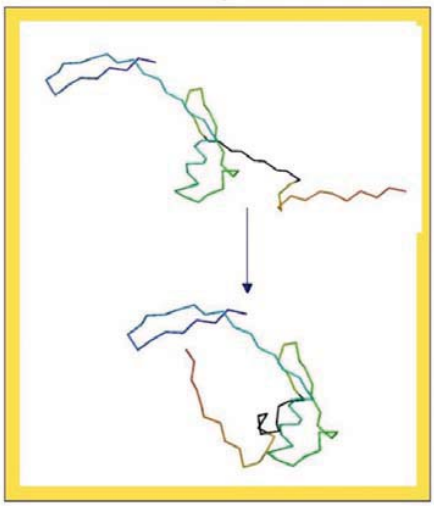
Local 서열 유사성, 2차구조 예측 결과를 기반으로 다양한 fragment 구조를 찾고, 이를 활용해 전체구조를 assembly 해보자!

28

조각 모음 기반 단백질 구조 예측



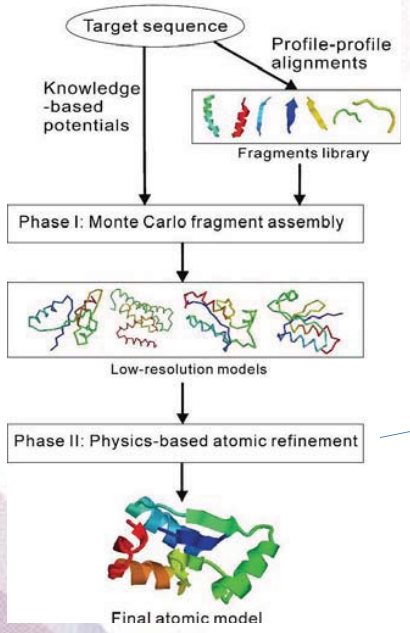
- start with an elongated chain
- make a random fragment insertion
- accept moves which pass the metropolis criterian (random number < exp(-ΔU/RT))
- to converge to low energy solutions decrease the temperature during the simulation (simulated annealing)

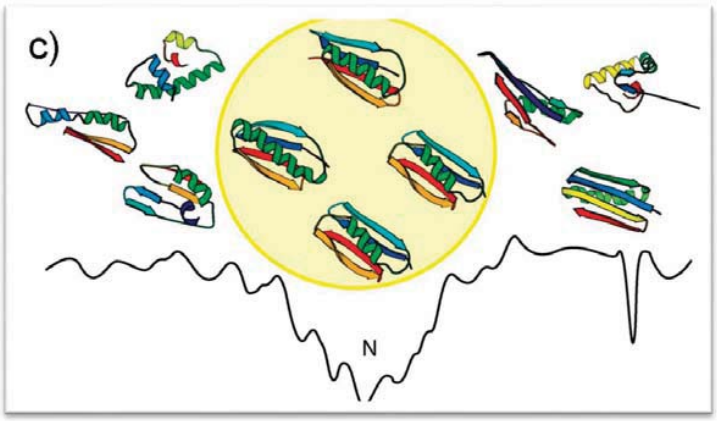


Rohl, Carol A., et al. "Protein structure prediction using Rosetta." *Methods in enzymology*. Vol. 383. Academic Press, 2004. 66-93.

29

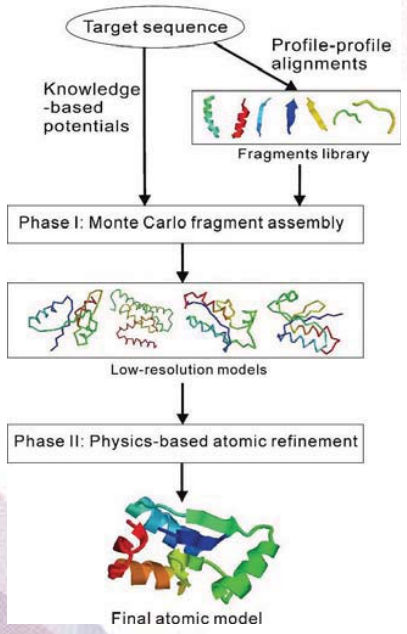
조각 모음 기반 단백질 구조 예측





30

조각 모음 기반 단백질 구조 예측



- 단백질 모델링 문제를 combinatorial problem으로 변환
  - Fragment의 quality
  - 큰 단백질 – too large conformational space
  - Coarse-grained energy의 quality
  - High computational cost

성공확률이 매우 낮음!

31

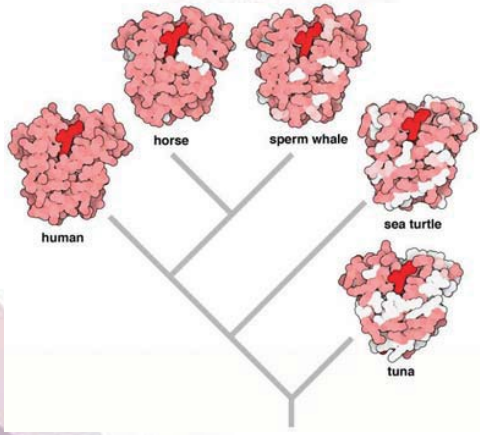
서열로부터 추가적인 구조 정보를 얻을 수는 없을까? (유사서열/진화정보 기반 모델링)

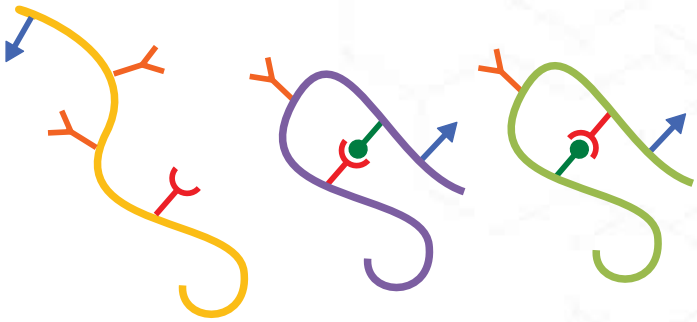
- 서열 데이터베이스 >> 구조 데이터베이스 (100배 이상의 차이!)
- 구조는 없더라도 유사한 서열들을 모아 정렬해본다면, 거기서 구조와 관련된 정보를 얻어낼 수 있지 않을까?
- Coevolution 개념의 등장

32

단백질 진화 역사 속에 숨어있는 구조 정보?

미오글로빈의 구조



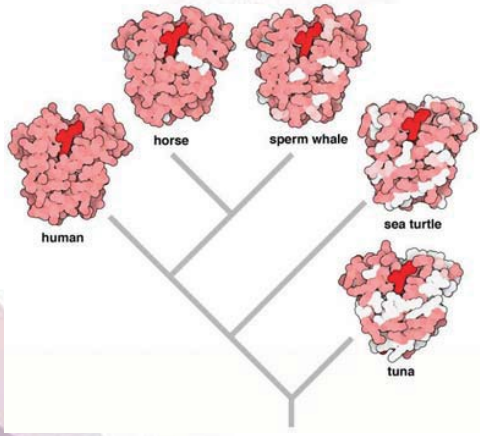


질병유발 정상단백질 구조/기능유지 (살아남음)

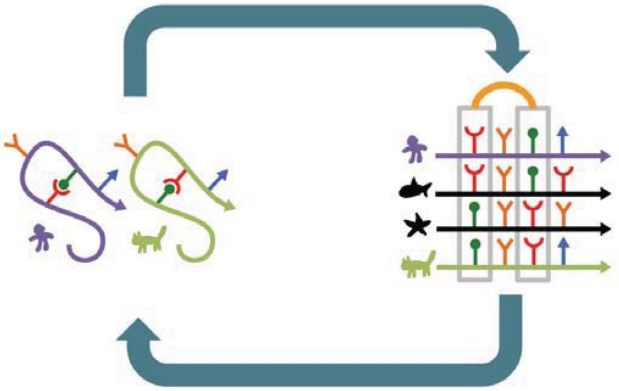
33

단백질 진화 역사 속에 숨어있는 구조 정보?

미오글로빈의 구조



기능 및 구조 유지를 위한
여러 제한요건

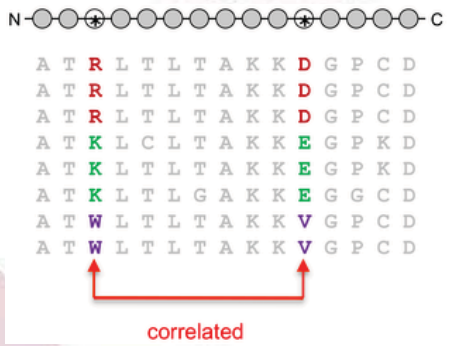


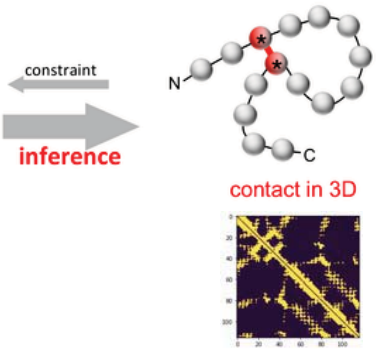
서열모음에 나타난 구조 패턴

34

진화정보 기반의 단백질 구조 예측

Multiple Sequence Alignment (MSA)
진화적으로 연관이 있는 단백질들의 서열모음





***Q: MSA** 서열모음 속에서 어떻게 구조정보를 찾아낼 것인가?*

GREMLIN: 단백질 컨택 예측 방법

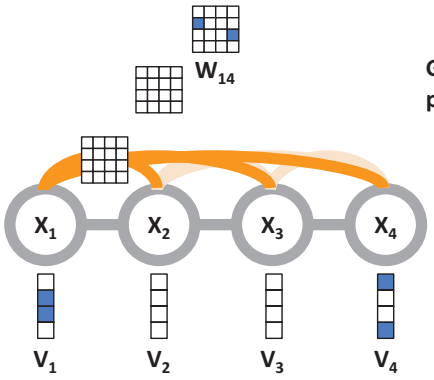
Global statistical model over protein sequences with:

$$P(x_1,x_2,x_3,x_4)=\frac{1}{Z}\exp\left(\sum_i v_i(x_i)+\sum_{i,j} w_{ij}(x_{ij})\right)$$

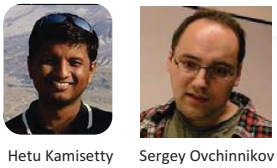
Coupling
w_{ij} = two-body energy

x = position

Conservation
v_i = one-body energy



Given input MSA, learn parameters: v and w


Hetu Kamisetty

Sergey Ovchinnikov

Kamisetty H, **Ovchinnikov S**, Baker D
Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. PNAS (2013)

GREMLIN: 단백질 컨택 예측 방법

GREMLIN uses a learning procedure based on optimizing the pseudolikelihood (5) of **v**, **w**, which, in log space is expressed as the sum of conditional distributions as follows:

$$pll(\mathbf{v},\mathbf{w}|D)=\sum_{n=1}^{N}\sum_{i=1}^{L}\log P\left(x_i^n\left|x_{-i}^n,\mathbf{v},\mathbf{w}\right.\right).$$

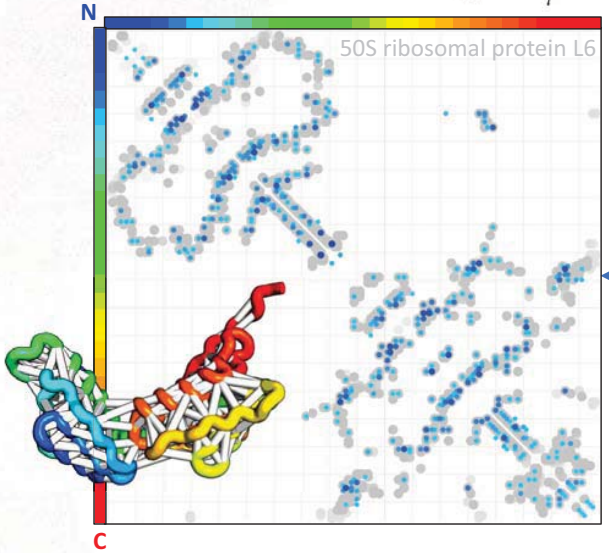
Each conditional distribution models the probability of the observed amino acid at position i in the n^{th} sequence of the alignment, x_i^n, in the context of the amino acids at all other positions in that sequence, x_{-i}^n and depends on the parameters **v**,**w** as:

$$P\left(x_i^n\left|x_{-i}^n,\mathbf{v},\mathbf{w}\right.\right)=\frac{1}{Z_i}\exp\left(\mathbf{v}_i\left(x_i^n\right)+\sum_{j=1,j\neq i}^{L}\mathbf{w}_{i,j}\left(x_i^n,x_j^n\right)\right).$$

37

GREMLIN: 단백질 컨택 예측 방법

Global statistical model $P(x_1,x_2,x_3,x_4)=\frac{1}{Z}\exp(\sum_i v_i(x_i)+\sum_{i,j} w_{ij}(x_{ij}))$



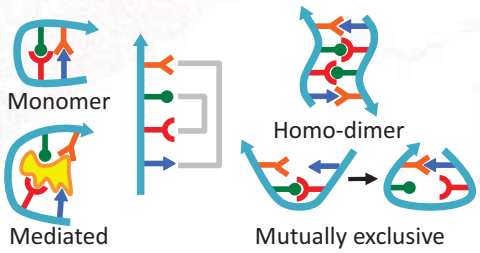
GREMLIN

APC(L2norm(W$_{ij}$))

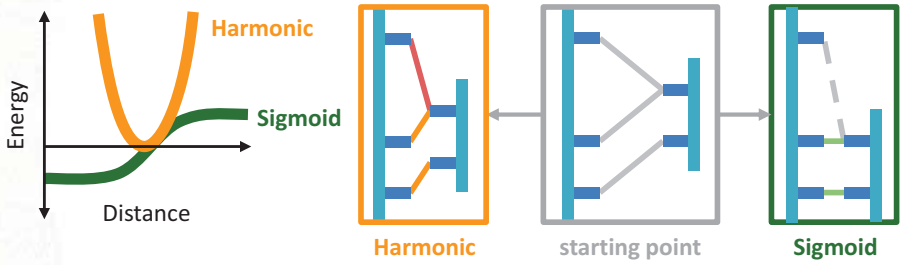
38

예측된 컨택을 기반으로 한 단백질 모델링

모든 예측이
다 맞는건 아님:



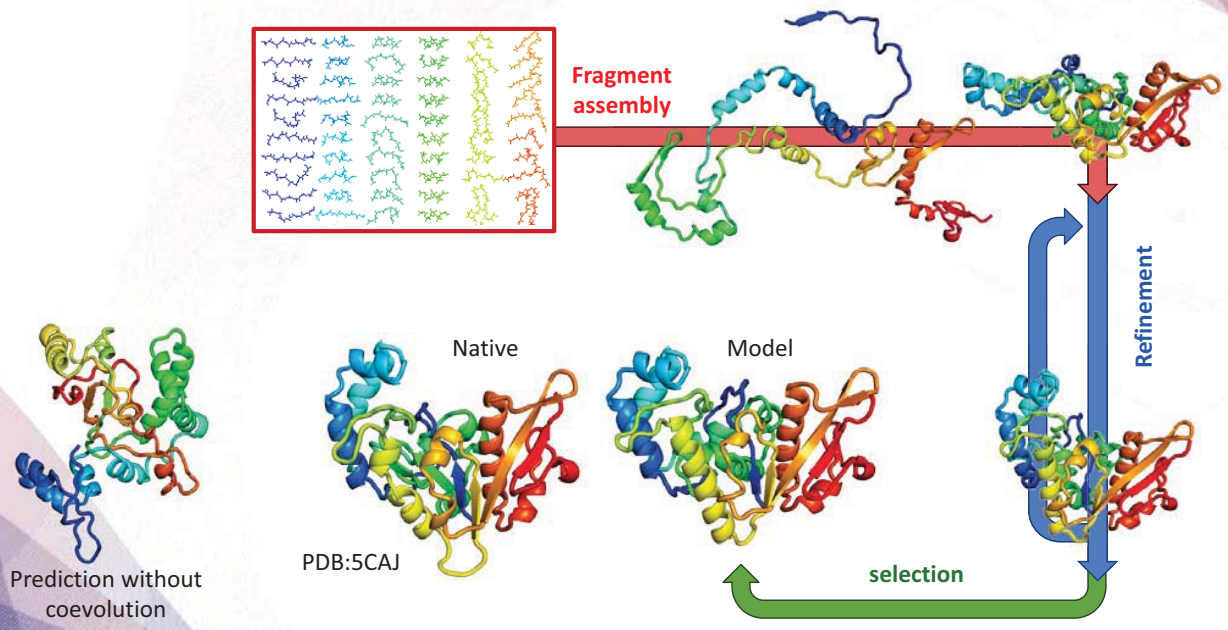
Use sigmoidal restraints to maximize self-consistent contact without distorting the structure:



39

예측된 컨택을 기반으로 한 단백질 모델링

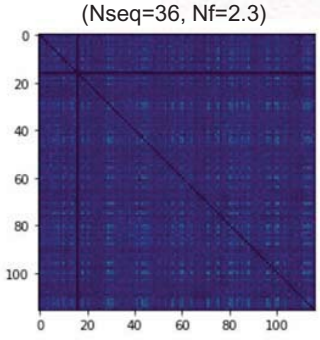
Using Rosetta to sample structures w/ predicted contact as restraints

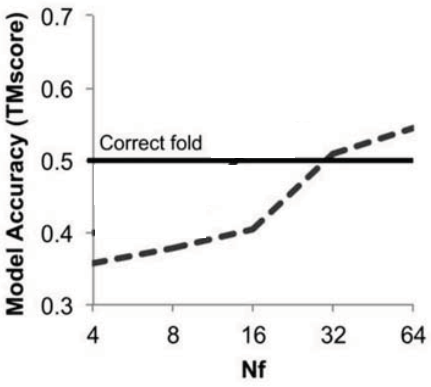


40

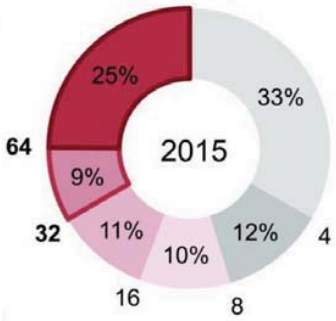
통계기반 방법의 한계

Gremlin predictions on shallow MSAs
(Nseq=36, Nf=2.3)





Nf values for sequences in pfam database



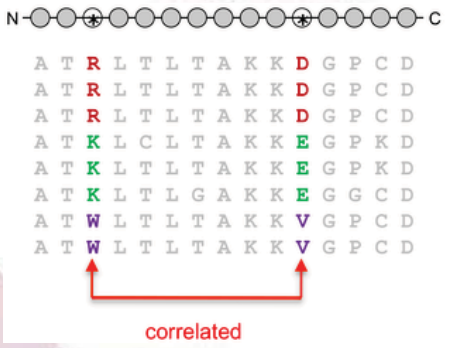
어떻게 해야 이 한계를 극복할 수 있을까?

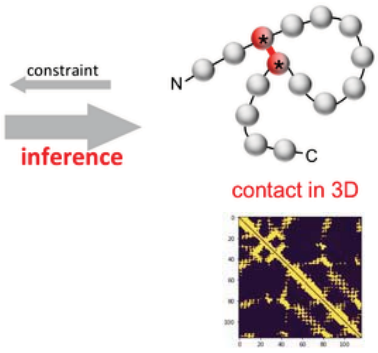
41

Ovchinnikov, Sergey, et al. "Protein structure determination using metagenome sequence data." *Science* 355.6322 (2017): 294-298.

진화정보 기반의 단백질 구조 예측

Multiple Sequence Alignment (MSA)
진화적으로 연관이 있는 단백질들의 서열모음



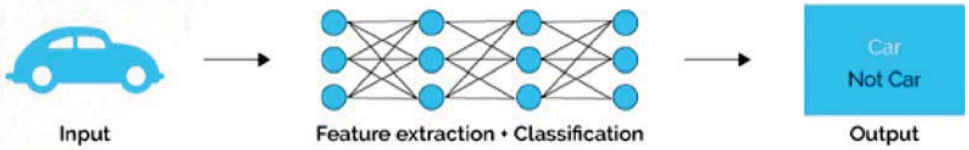


***Q: MSA** 서열모음 속에서 어떻게 구조정보를 찾아낼 것인가?*

42

딥러닝과의 결합

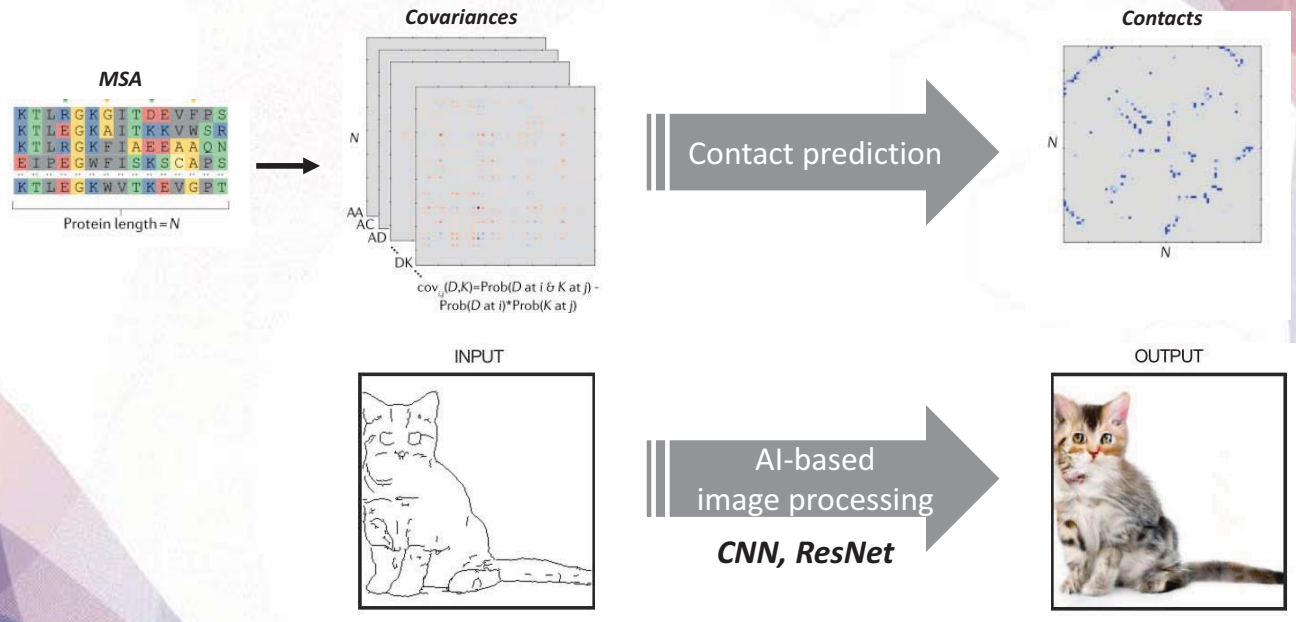
- 목표: 서열모음 **데이터**에 숨어있는 구조 **패턴 찾기**!
  - 패턴찾기에 특화된 기술? → 인공지능!



  - 지금까지 밝혀진 단백질 구조 데이터 ~ 10만여개
    - 이를 학습데이터로 활용하여 단백질 구조 예측을 위한 인공지능 모델을 만들어보자

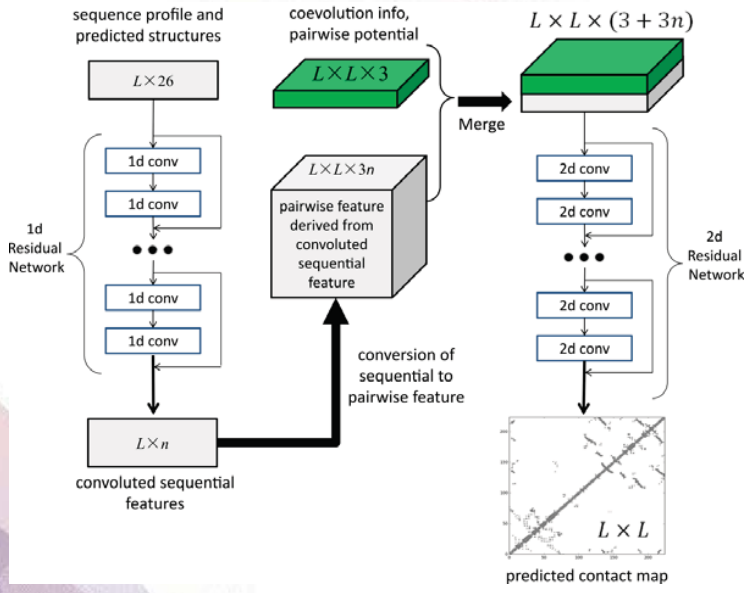
43

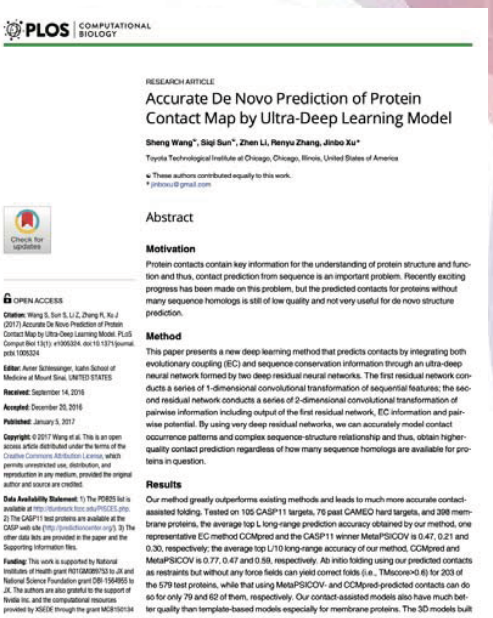
어떤 인공지능 모델을 사용할 것인가? (Late 2010s)



44

RaptorX-Contact (2017)





PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model

Sheng Wang*, Siqi Sun*, Zhen Li, Renyu Zhang, Jinbo Xu*

Toyota Technological Institute at Chicago, Chicago, Illinois, United States of America

* These authors contributed equally to this work.
* jinbocu@gmail.com

Abstract

Motivation

Protein contacts contain key information for the understanding of protein structure and function and thus, contact prediction from sequence is an important problem. Recently exciting progress has been made on this problem, but the predicted contacts for proteins without many sequence homologs is still of low quality and not very useful for de novo structure prediction.

Method

This paper presents a new deep learning method that predicts contacts by integrating both evolutionary coupling (EC) and sequence conservation information through an ultra-deep neural network formed by two deep residual neural networks. The first residual network conducts a series of 1-dimensional convolutional transformation of sequential features; the second residual network conducts a series of 2-dimensional convolutional transformation of pairwise information including output of the first residual network, EC information and pairwise potential. By using very deep residual networks, we can accurately model contact occurrence patterns and complex sequence-structure relationship and thus, obtain higher-quality contact prediction regardless of how many sequence homologs are available for proteins in question.

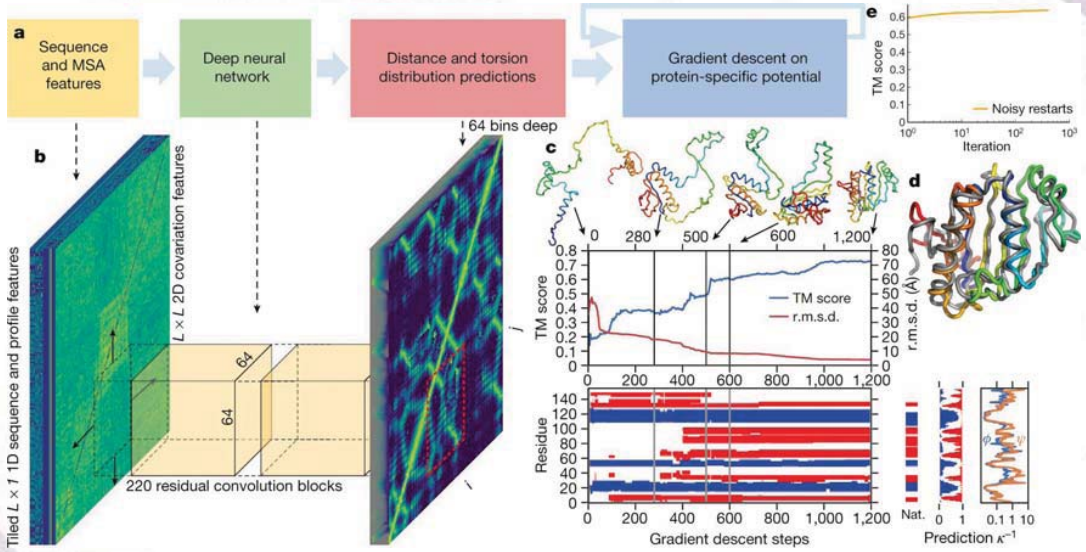
Results

Our method greatly outperforms existing methods and leads to much more accurate contact-assisted folding. Tested on 105 CASP11 targets, 76 past CAMEO hard targets, and 398 membrane proteins, the average top L long-range prediction accuracy obtained by our method, one representative EC method CCMpred and the CASP11 winner MetaPSICOV is 0.47, 0.21 and 0.30, respectively; the average top L/10 long-range accuracy of our method, CCMpred and MetaPSICOV is 0.77, 0.47 and 0.59, respectively. Ab initio folding using our predicted contacts as restraints but without any force fields can yield correct folds (i.e., TMscore>0.6) for 203 of the 579 test proteins, while that using MetaPSICOV- and CCMpred-predicted contacts can do so for only 79 and 62 of them, respectively. Our contact-assisted models also have much better quality than template-based models especially for membrane proteins. The 3D models built

***Contact* 예측 자체의 성능은 향상, but 구조 예측까지 이어지지 못함**

Wang, Sheng, et al. "Accurate de novo prediction of protein contact map by ultra-deep learning model." *PLoS computational biology* 13.1 (2017): e1005324.

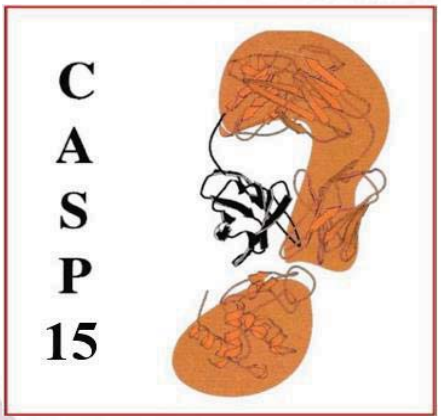
초기 AlphaFold (2018, CASP13)



Distance (distogram) 예측 → restraint E → 에너지 최적화를 통한 구조 예측

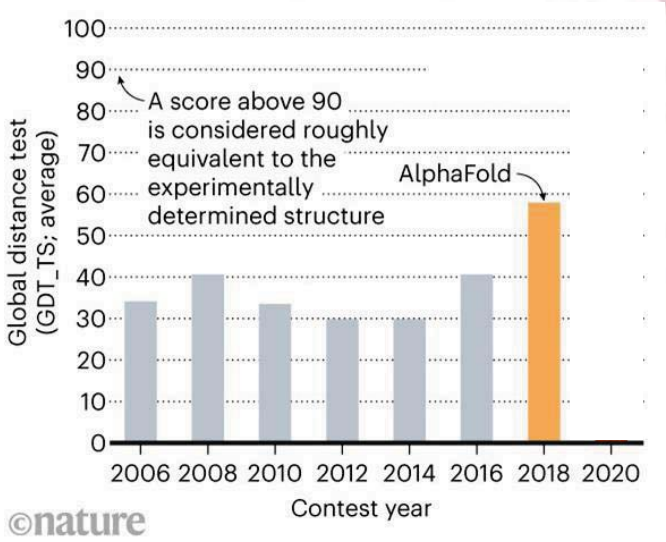
Senior, Andrew W., et al. "Improved protein structure prediction using potentials from deep learning." *Nature* 577.7792 (2020): 706-710.

초기 AlphaFold (2018, CASP13)

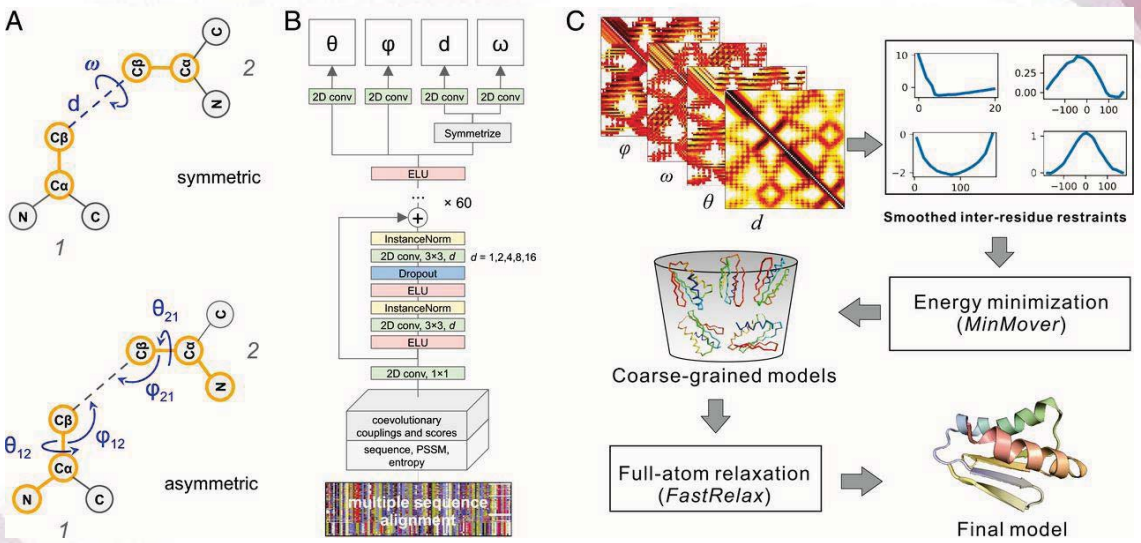


1994년부터 시작된 국제대회

단백질 구조예측 정확도



ResNet 기반의 단백질 구조 예측: trRosetta (2019)

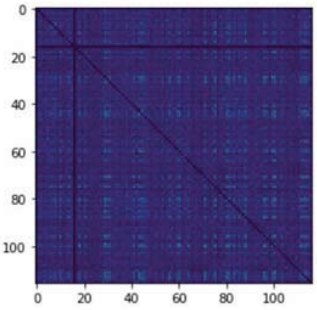


Distance (distogram) + orientation 예측 → restraint E → 에너지 최적화를 통한 구조 예측

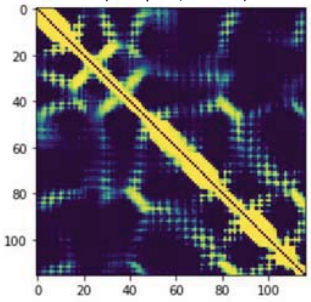
Yang, Jianyi, et al. "Improved protein structure prediction using predicted interresidue orientations." *Proceedings of the National Academy of Sciences* 117.3 (2020): 1496-1503.

ResNet 기반의 단백질 구조 예측: trRosetta (2019)

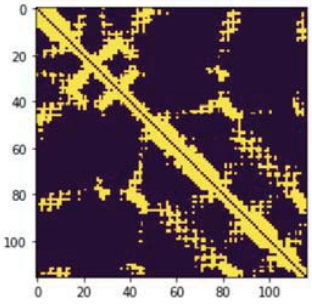
GREMLIN predictions on shallow MSAs
(Nseq=36, Nf=2.3)



trRosetta predictions on shallow MSAs
(Nseq=36, Nf=2.3)

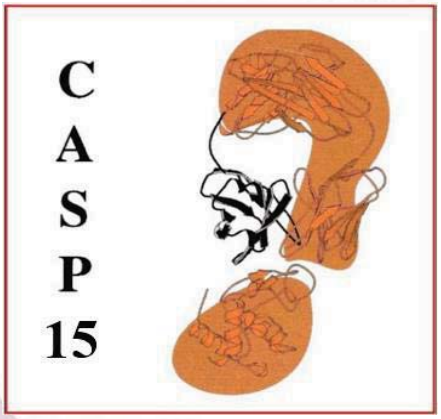


Native contact map



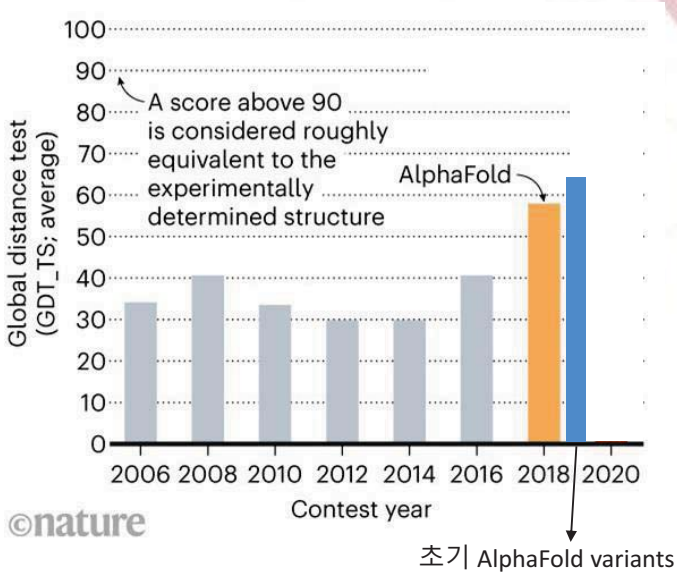
Yang, Jianyi, et al. "Improved protein structure prediction using predicted interresidue orientations." *Proceedings of the National Academy of Sciences* 117.3 (2020): 1496-1503.

알파폴드2의 등장



1994년부터 시작된 국제대회

단백질 구조예측 정확도



제 2강.

Deep Learning-based protein structure prediction & its application

지난 시간…

1. 단백질 접힘 (protein folding)의 화학원리 이해

2. 주형기반 단백질 구조 예측 (Homology modeling)
   - 높은 유사도의 주형 단백질이 있어야만 고정확도 예측 가능

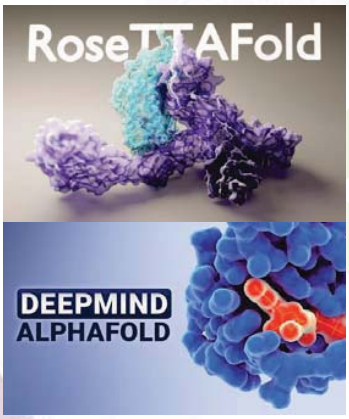
3. Ab initio 단백질 구조 예측 (Free modeling)
   - 단백질 조각 모음 기반. 예측 성공률 매우 낮음

4. 유사서열(진화정보)기반 단백질 구조 예측
   - 진화정보로부터 구조에 대한 패턴 파악/활용

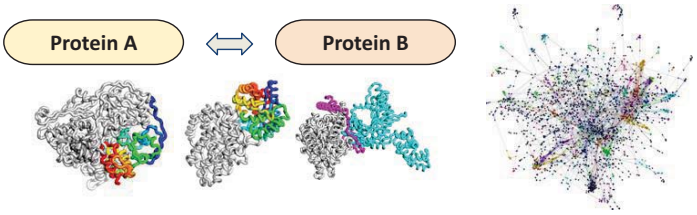
5. 초기 딥러닝 기반 단백질 구조 예측
   - 이미지 프로세싱 딥러닝 모델 (ResNet)을 활용한 단백질 구조 예측

Contents

딥러닝 기반
단백질 구조예측

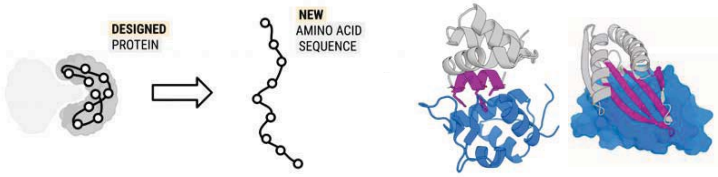


단백질 사이의 상호작용 예측



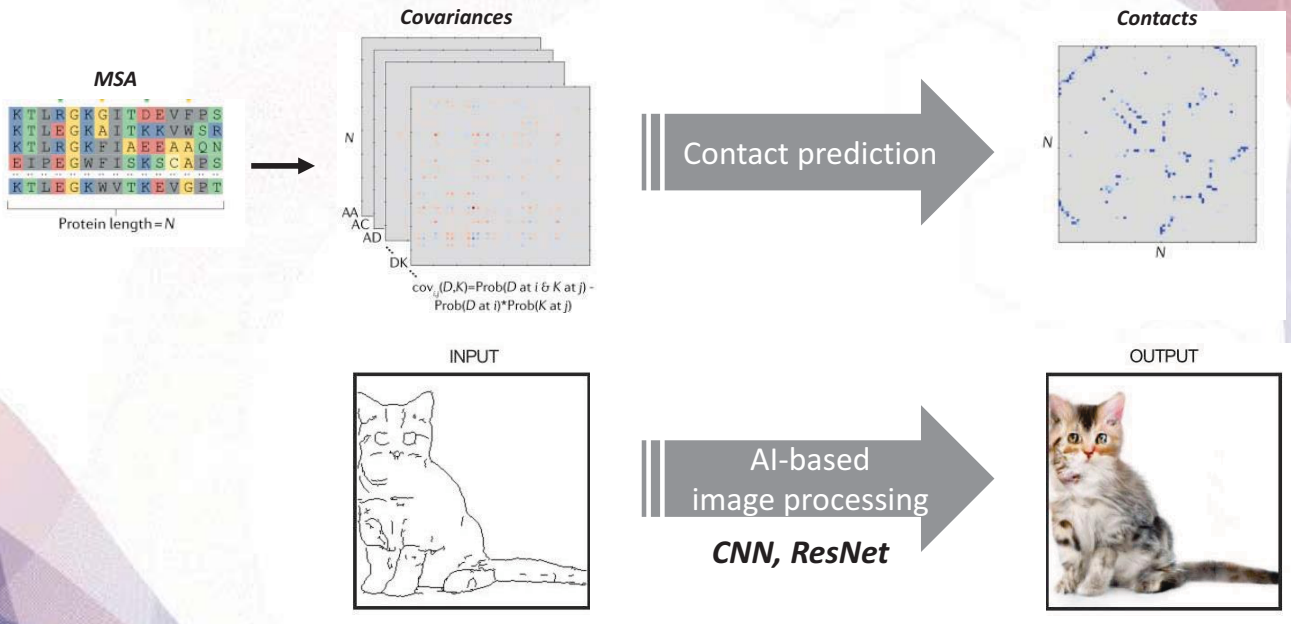
Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

딥러닝 기반 단백질 디자인

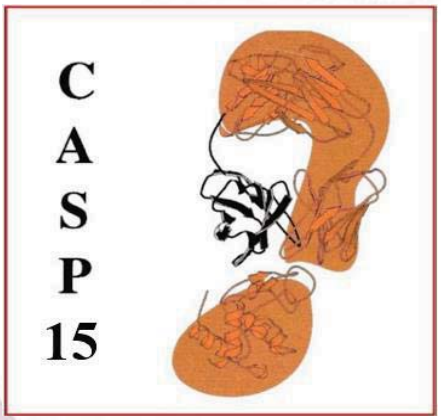


Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., et al, *Science* (2022)

어떤 인공지능 모델을 사용할 것인가? (Late 2010s)

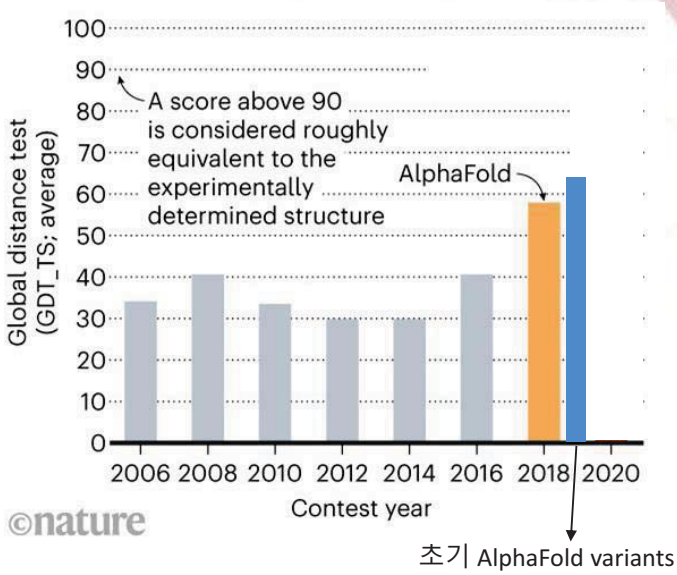


알파폴드2의 등장



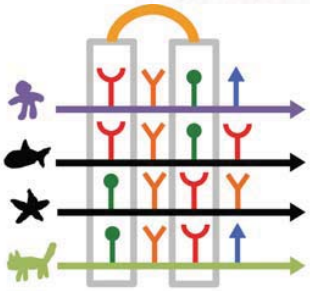
1994년부터 시작된 국제대회

단백질 구조예측 정확도

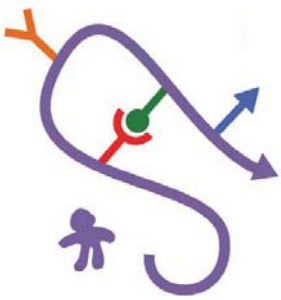


어떤 딥러닝 모델을 사용할 것인가?

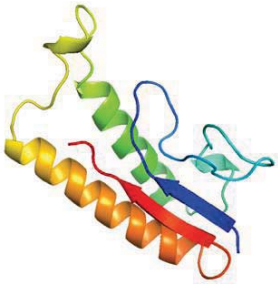
Multiple Sequence Alignment (MSA)
진화적으로 연관이 있는 서열 모음



레지듀 사이의 상호작용 정보

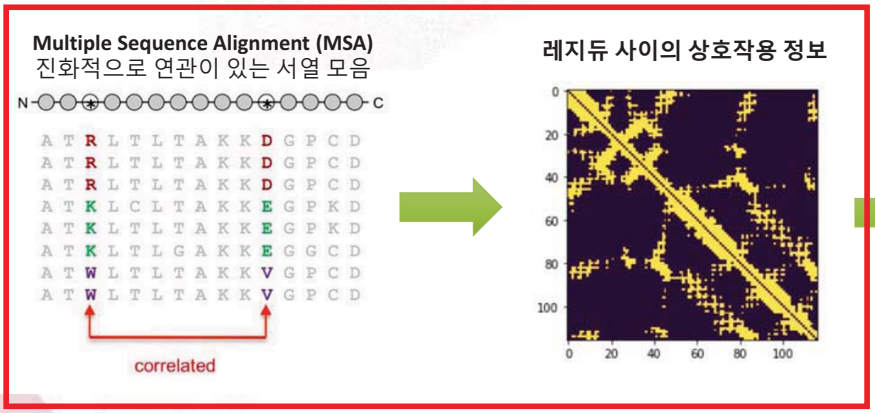


단백질 3차원 구조



어떻게 딥러닝 모델을 설계해야 쉽게 배울 수 있을까?

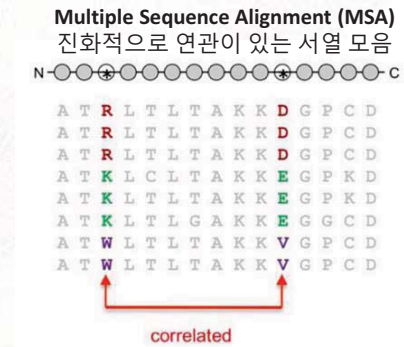
어떤 딥러닝 모델을 사용할 것인가?

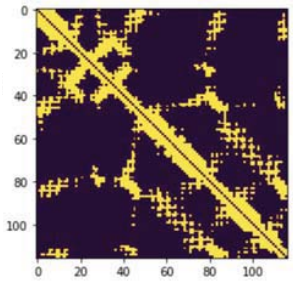


어떻게 딥러닝 모델을 설계해야 쉽게 배울 수 있을까?

57

어떤 딥러닝 모델을 사용할 것인가?





소년은 오리를 보고 놀랐습니다.
오리가 그릴 위에 있었기
때문이죠. 이 광경을 본 소녀는
놀라 소년과 오리를 향해
뛰어왔습니다. 달려오는 소녀
앞에는 노란 테이블이 있었습니다.

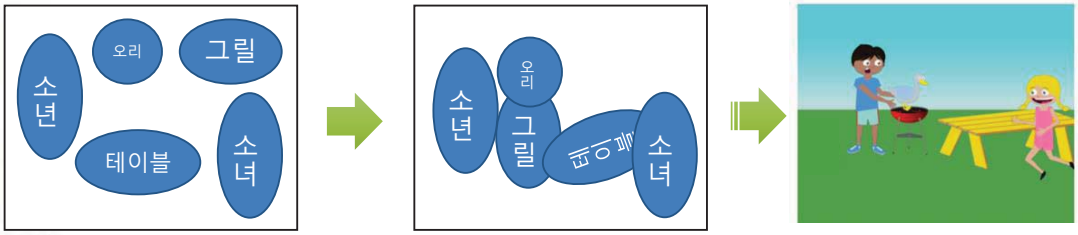
58

묘사를 바탕으로 그림그리기

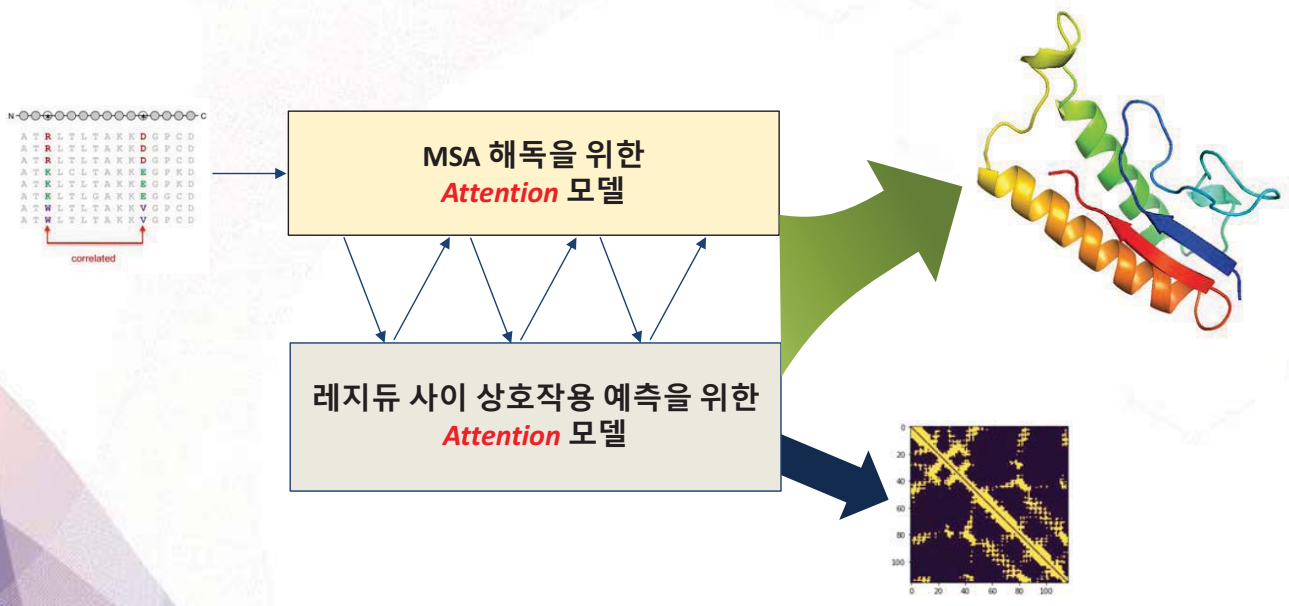
소년은 **오리**를 보고 놀랐습니다. 오리가 **그릴** 위에 있었기 때문이죠. 이 광경을 본 **소녀**는 놀라 소년과 오리를 향해 뛰어왔습니다. 달려오는 소녀 앞에는 노란 **테이블**이 있었습니다.

소년은 오리를 보고 놀랐습니다. **오리가 그릴 위**에 있었기 때문이죠. 이 광경을 본 소녀는 놀라 **소년과 오리를 향해** 뛰어왔습니다. 달려오는 **소녀 앞에**는 노란 **테이블**이 있었습니다.

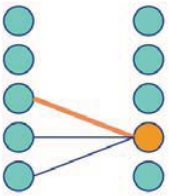
소년은 오리를 보고 **놀랐습니다**. 오리가 그릴 위에 있었기 때문이죠. 이 광경을 본 **소녀는 놀라** 소년과 오리를 향해 **뛰어왔습니다**. 달려오는 소녀 앞에는 **노란 테이블**이 있었습니다.



단백질 구조 예측을 위한 딥러닝 모델 설계



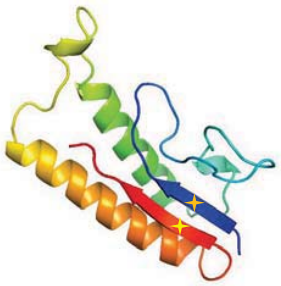
Attention vs Convolutional Network

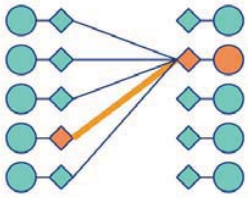


Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours

[Credit: DeepMind, from CASP14 presentation]





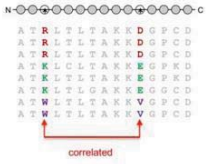
Attention Module (e.g. language)

- data in unordered set
- information flow dynamically controlled by the network (via keys and queries)

[Credit: DeepMind, from CASP14 presentation]

61

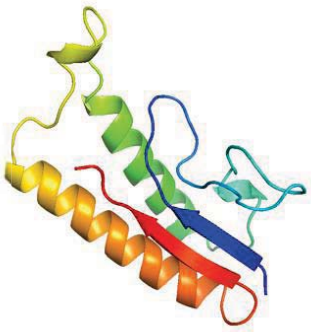
3차원 구조로 어떻게 만들지?

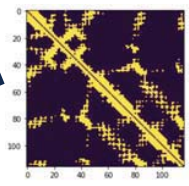


MSA 해독을 위한
Attention 모델

레지듀 사이 상호작용 예측을 위한
Attention 모델

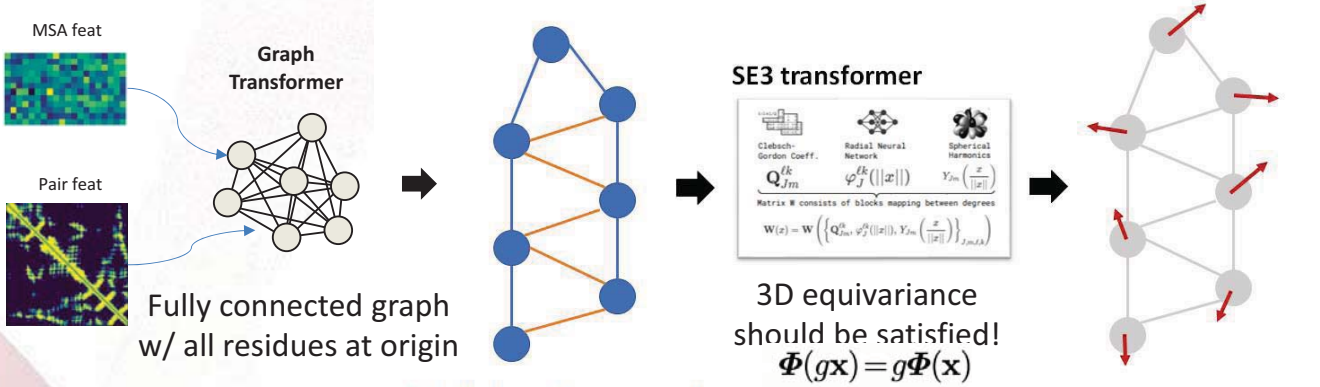
?





62

3차원 구조로 어떻게 만들지?



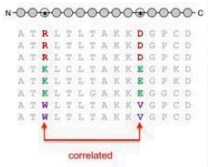
- Initial backbone Cartesian coordinates
- Node + edge information from trunk
- Graph connecting nearby residues

- Predicted offset vectors
- Predicted per-residue errors

63

단백질 구조 예측을 위한 딥러닝 모델 설계

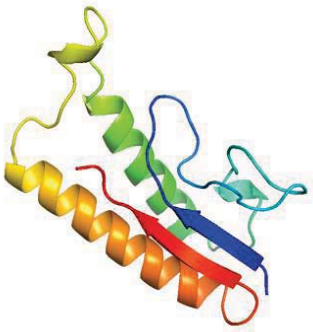
서열모음으로부터
레지듀 사이 상호작용 예측하기

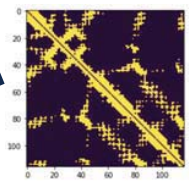


MSA 해독을 위한 Attention 모델

레지듀 사이 상호작용 예측을 위한 Attention 모델

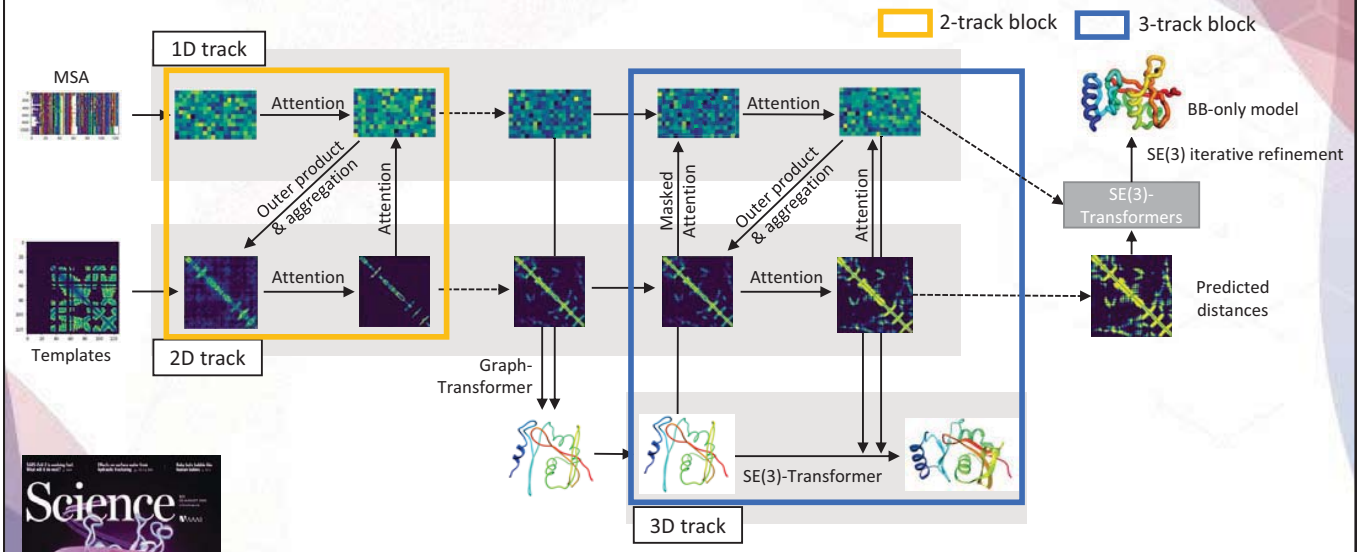
구조 생성모델





64

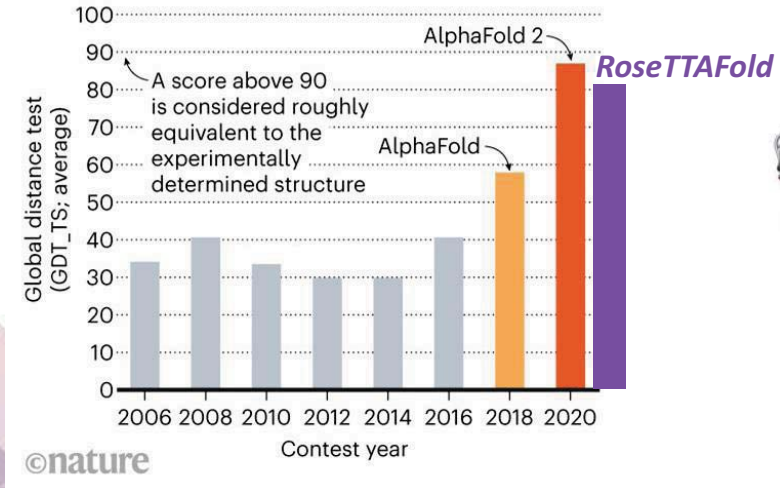
로제타폴드 딥러닝 모델 구조



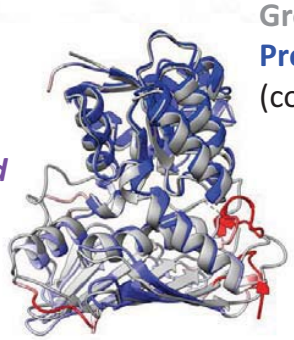
Baek, M., et al, *Science* (2021)

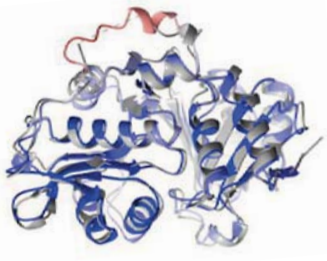
로제타폴드의 성능

단백질 구조예측 정확도

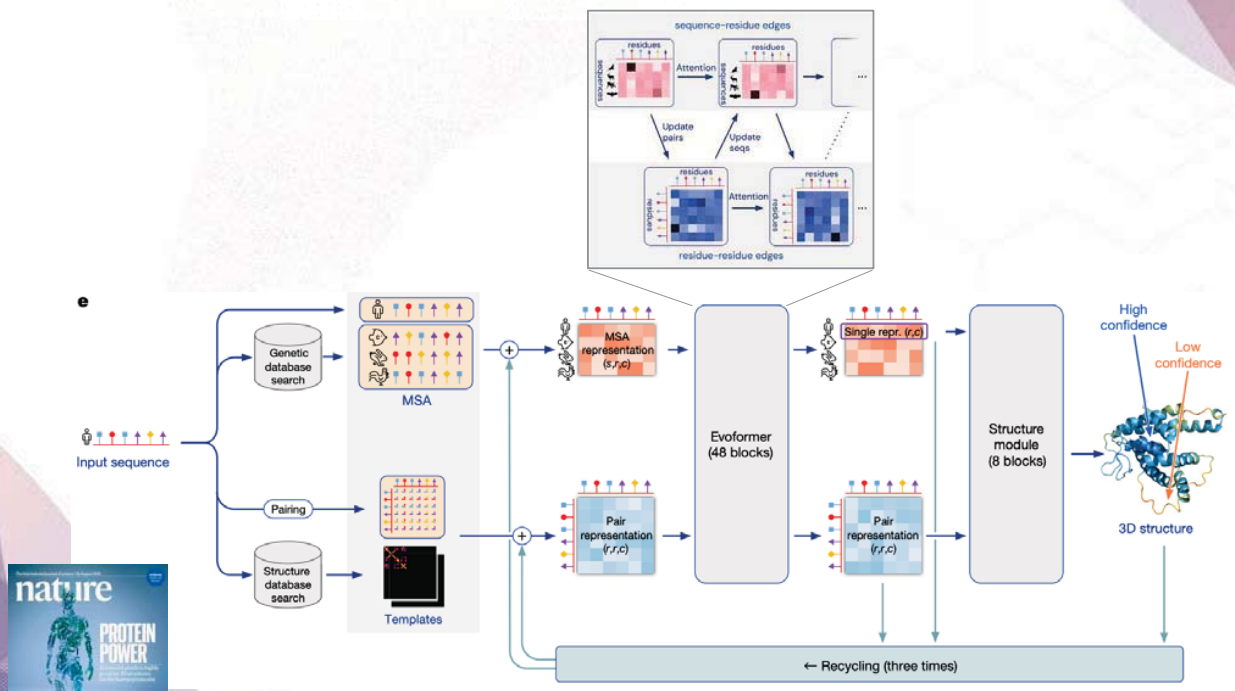


Ground-truth
Predicted
(colored by pLDDT)





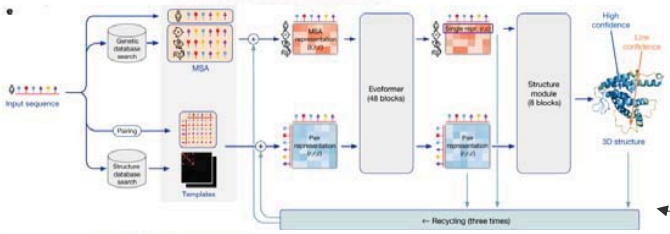
알파폴드2 딥러닝 모델 구조



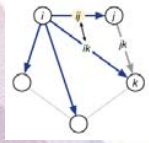
Jumper, J. et al, *Nature* (2021)

알파폴드엔 있지만 로제타폴드엔 없는 것?

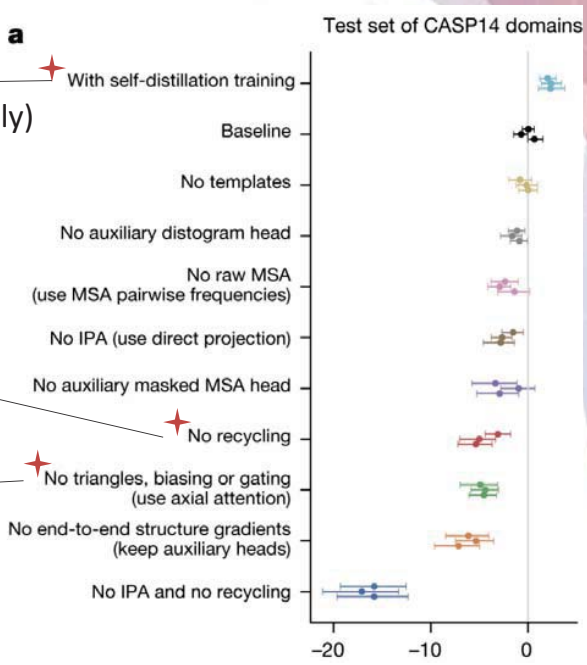
Extended training data
with predicted structures for UniRef50
(millions of proteins, high confidence structures only)



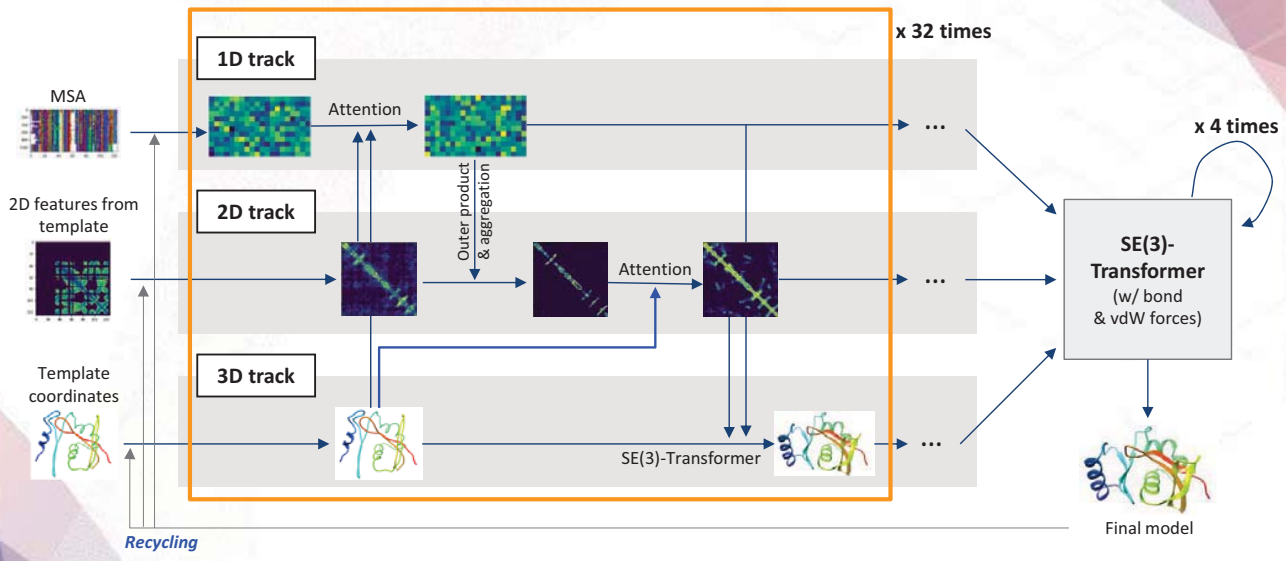
Introduce proper biases for feature updates



Should encode geometric info
(e.g. distances)
Computational cost: $O(L^3)$



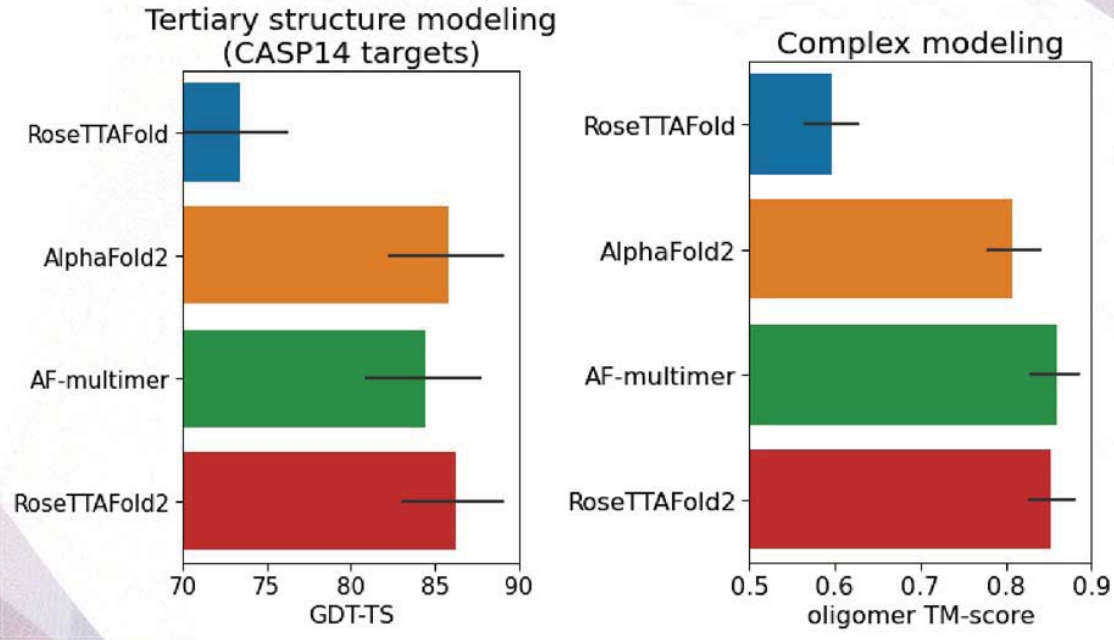
RoseTTAFold2: Improving RF for better protein modeling



Trained on PDB (monomer+**multimer**) + **UniRef50 models**

Baek, M. et al., *In preparation*

RoseTTAFold2: Improving RF for better protein modeling

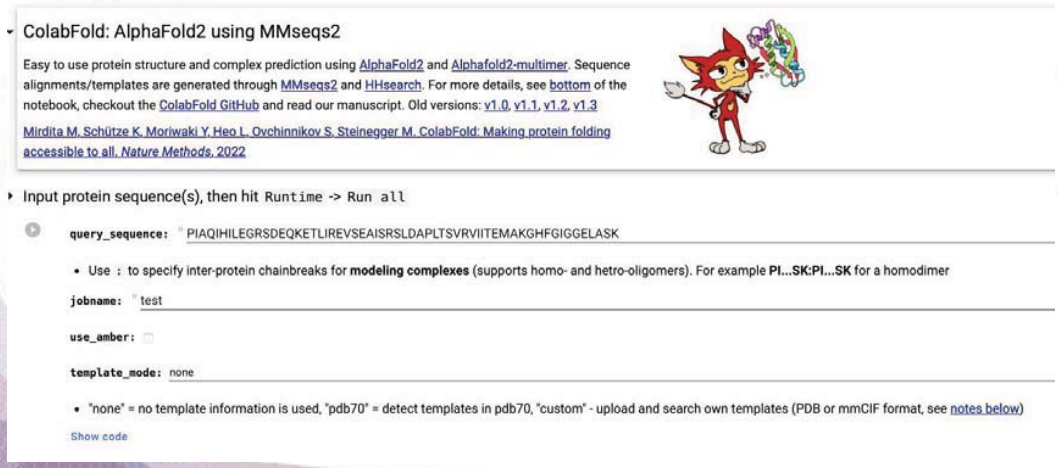


Baek, M. et al., *In preparation*

Anyone can use RoseTTAFold & AlphaFold2

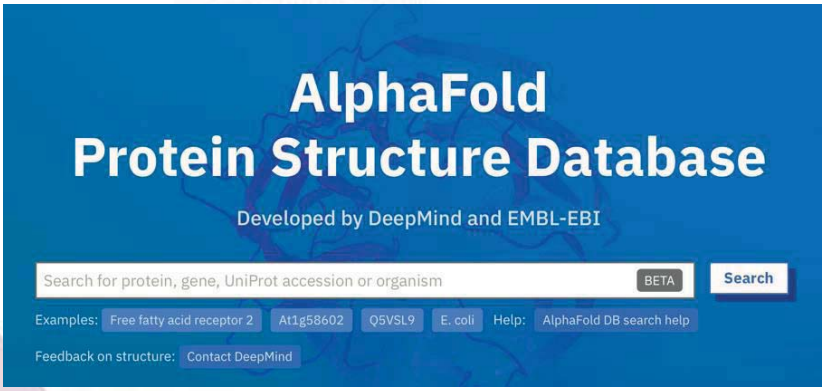
RoseTTAFold
- Source code: https://github.com/RosettaCommons/RoseTTAFold
- Web server: https://robetta.bakerlab.org

AlphaFold2
- Source code: https://github.com/deepmind/alphafold
- ColabFold version: https://github.com/sokrypton/ColabFold (Credit: Sergey Ovchinnikov, Martin Steinegger)

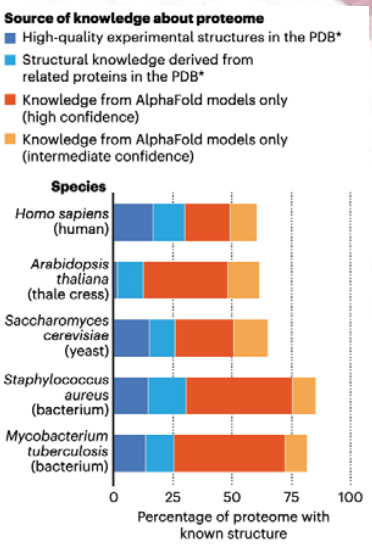


단백질 구조에 대한 수많은 새로운 지식 축적!

https://alphafold.ebi.ac.uk

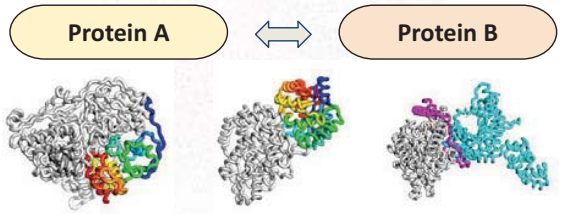


More than 200M protein structures!



단백질 구조 예측의 응용

단백질-단백질 상호작용 및 결합구조 예측



1) Do **A** and **B** interact?

2) What is the structure of **AB**?

Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

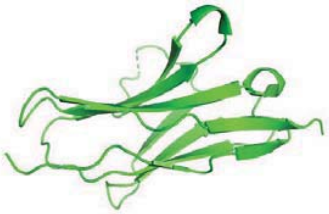
새로운 단백질 디자인

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., Science (2022)

73

단백질의 구조만 알면 되나?

암세포에
유난히 많이 발현되는 단백질



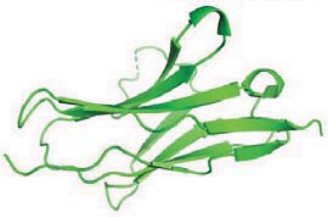
살해T세포에
존재하는 단백질

74

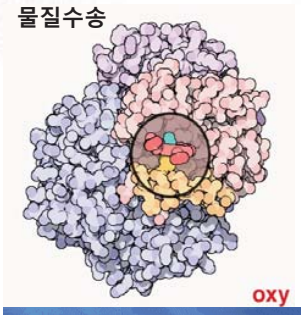
단백질의 구조만 알면 되나?

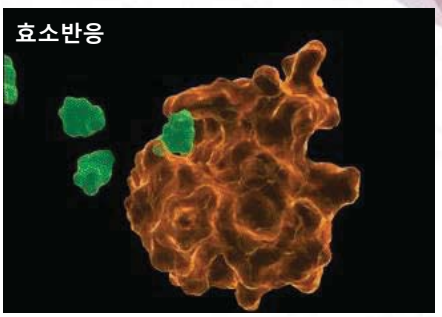
상호작용 예측이 더 중요!

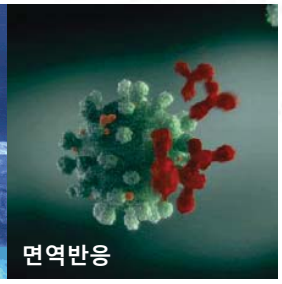
암세포에서
유난히 많이 발현되는 단백질



살해T세포에 존재하는 단백질

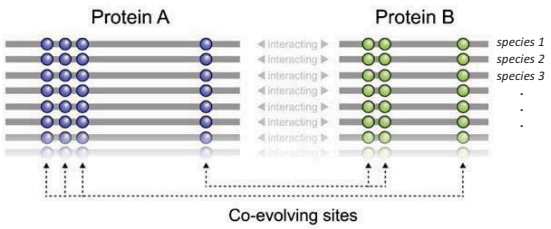


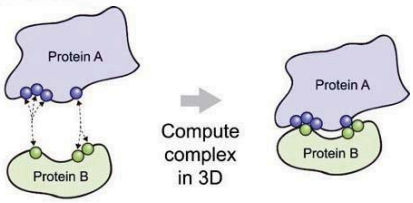


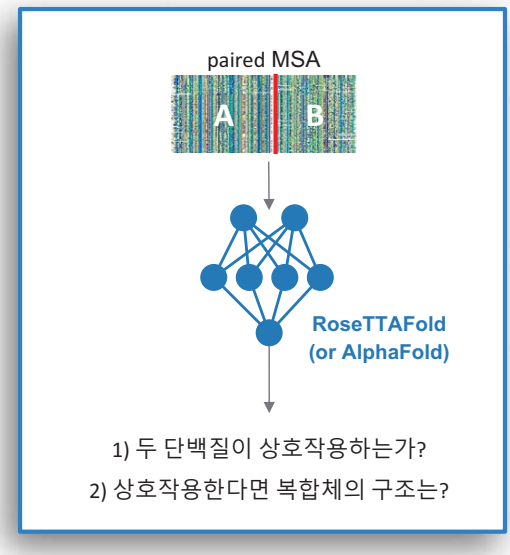


단백질 사이의 상호작용도 예측할 수 있을까?

기능 유지 → 기능에 필수적인 상호작용(구조) 유지 → 서열모음에 구조적 패턴이 나타남



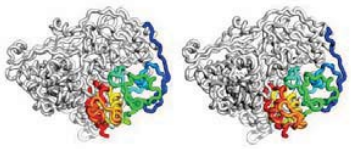




단백질-단백질 상호작용 예측으로의 확장

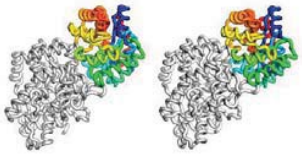
기능 유지 → 기능에 필수적인 상호작용(구조) 유지 → 서열모음에 구조적 패턴이 나타남

Aldehyde oxidoreductase



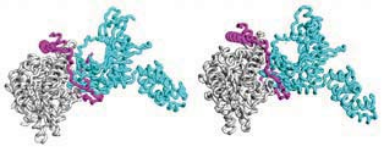
TM-score: 95

Tryptophan synthase

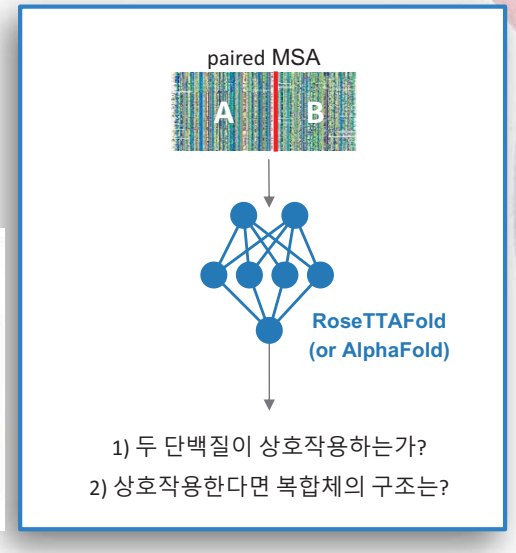


TM-score: 92

tRNA-dependent amidotransferase



TM-score: 89

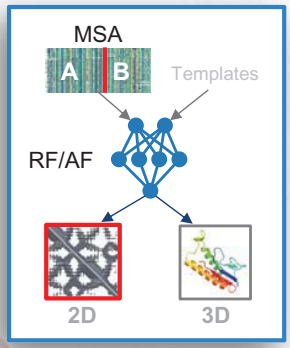


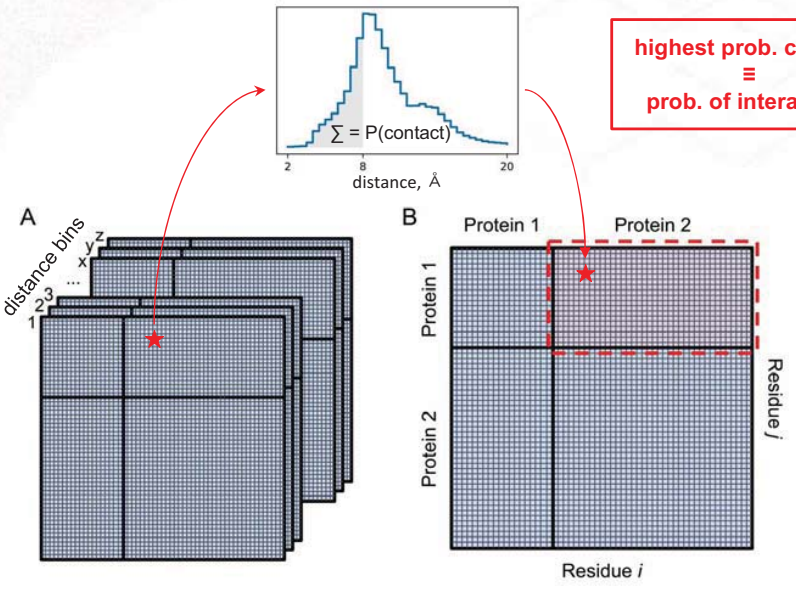
1) 두 단백질이 상호작용하는가?

2) 상호작용한다면 복합체의 구조는?

77

단백질-단백질 상호작용 예측으로의 확장





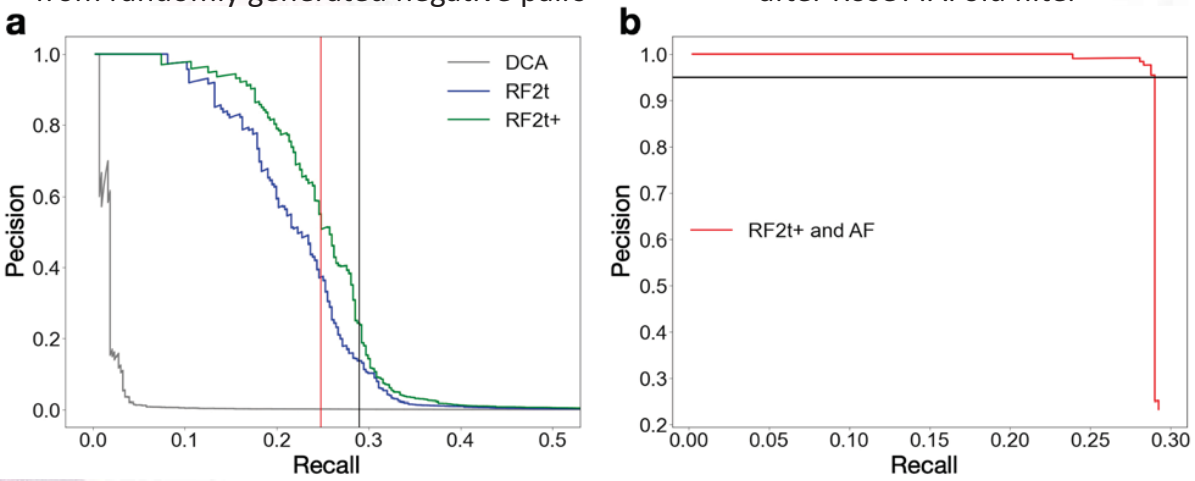
Humphreys, I.R., Pei, J., **Baek, M.**, Krishnakumar, A., et al, (2021), *Science*, (co-first author)

78

단백질-단백질 상호작용 예측으로의 확장

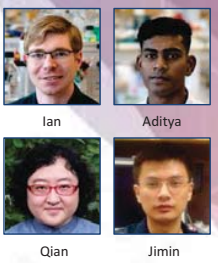
Distinguish 768 gold-standard pairs
from randomly generated negative pairs

Distinguish 717 gold-standard pairs
after RoseTTAFold filter



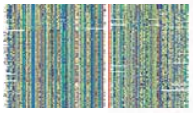
Humphreys, I.R., Pei, J., **Baek, M.**, Krishnakumar, A., et al, (2021), *Science*, (co-first author)

효모(yeast)에 존재하는 단백질들 사이의 상호작용 예측



- 4.3 million protein pairs

pMSA



2-track RoseTTAFold
(10.7M parameters)
~ 10 sec per 1000aa complex

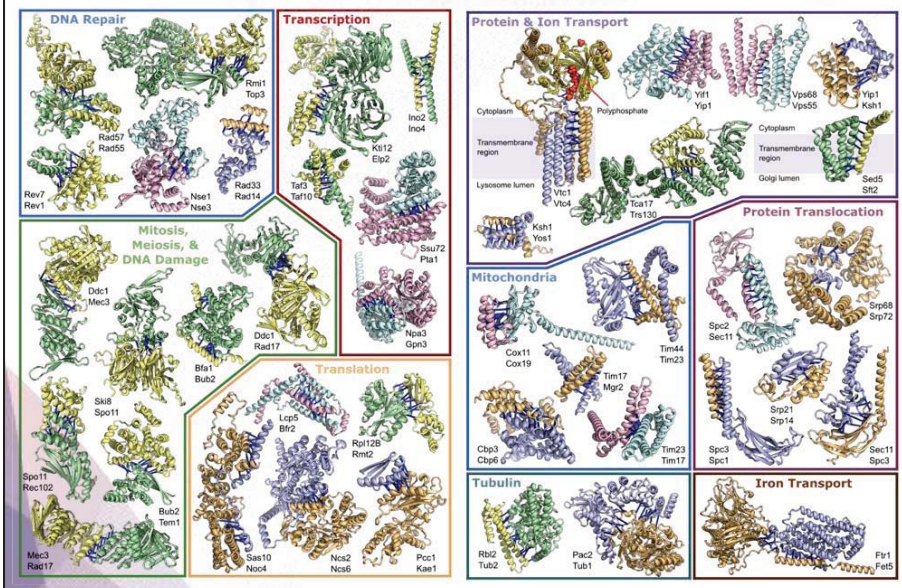
5,495 pairs

AlphaFold
~1800 sec per 1000aa complex

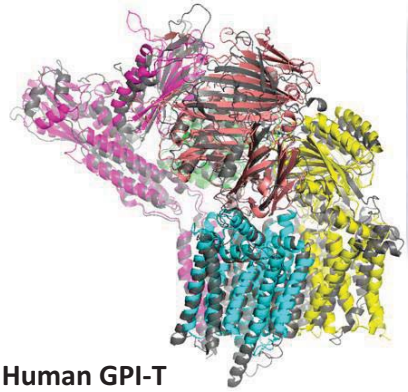
1,505 pairs
likely to interact
+
Manual study

Humphreys, I.R., Pei, J., **Baek, M.**, Krishnakumar, A., et al, (2021), *Science*, (co-first author)

효모(yeast)에 존재하는 단백질들 사이의 상호작용 예측



Yeast GPI-T
(our prediction, Oct 2021)



Human GPI-T
(PDB: 7W72, published Feb 2022)

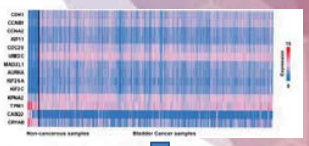
Humphreys, I.R., Pei, J., **Baek, M.**, Krishnakumar, A., et al, (2021), *Science*, (co-first author)

81

신약개발에의 영향?

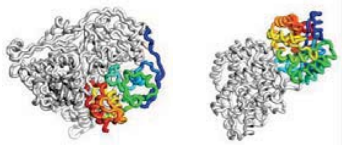
질병 연관 단백질-단백질 상호작용 예측 → 구조 기반 신약 개발

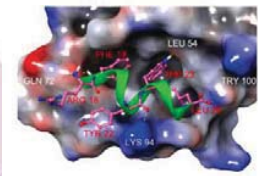
Bioinformatics analysis
of gene expression level



단백질 사이 상호작용 예측

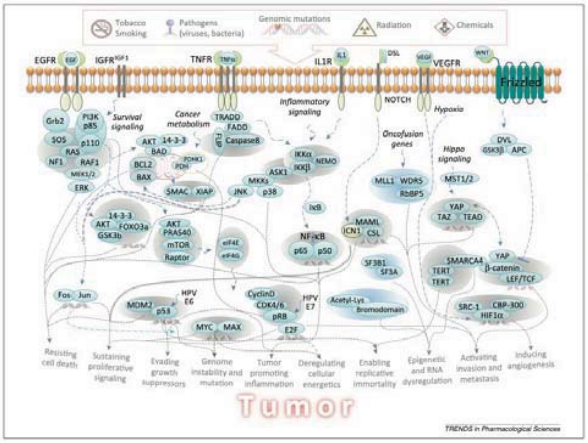
Protein A ⟺ Protein B





구조 기반의
신약 개발
(성공확률 향상)

질병 메커니즘의 자세한 이해 (표적발굴)

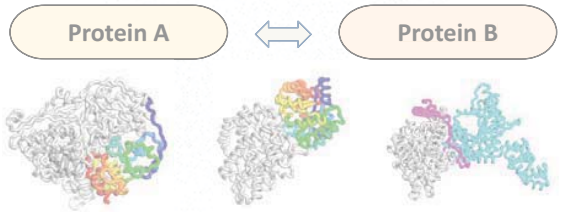


후보물질 발굴 시간 단축, 성공확률 향상

82

단백질 구조 예측의 응용

단백질-단백질 상호작용 및 결합구조 예측

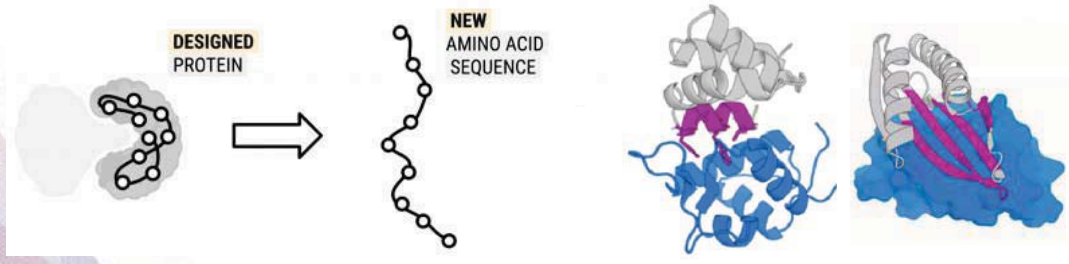


1) Do **A** and **B** interact?

2) What is the structure of **AB**?

Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

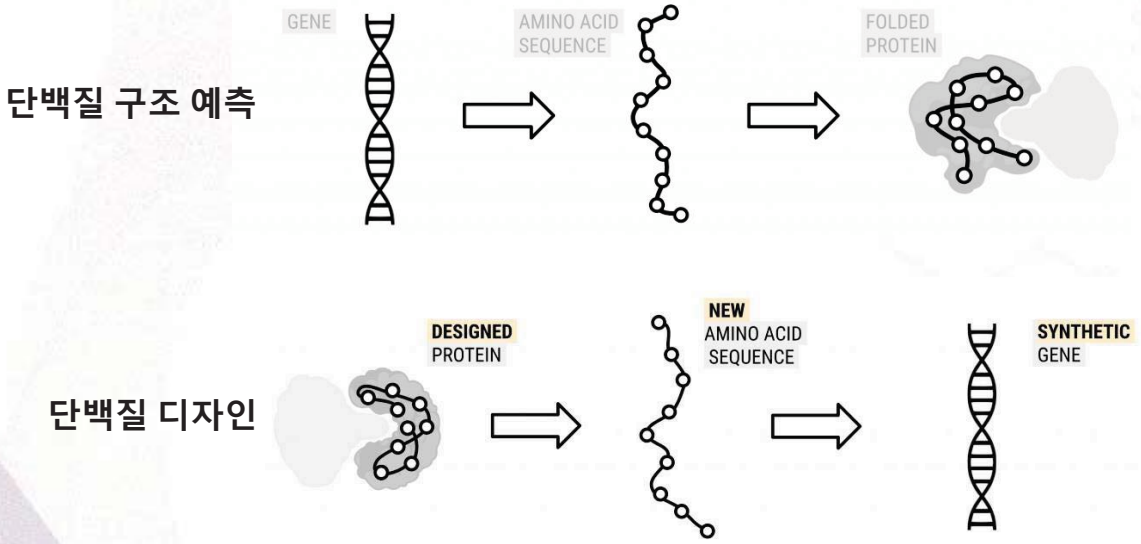
새로운 단백질 디자인



Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., Science (2022)

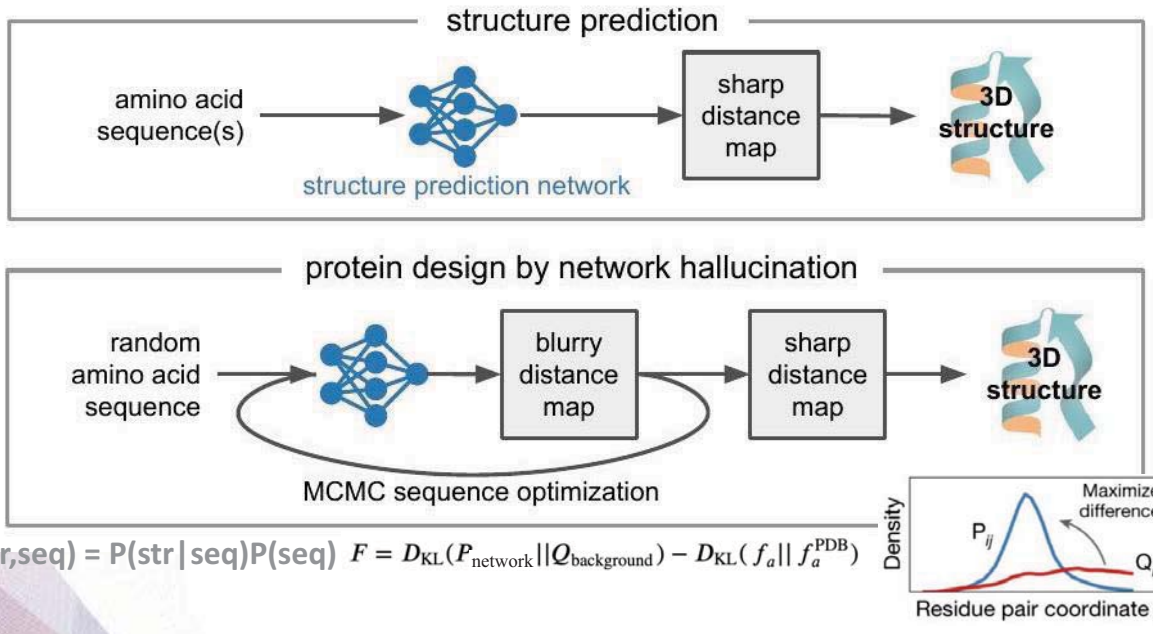
83

단백질 디자인?



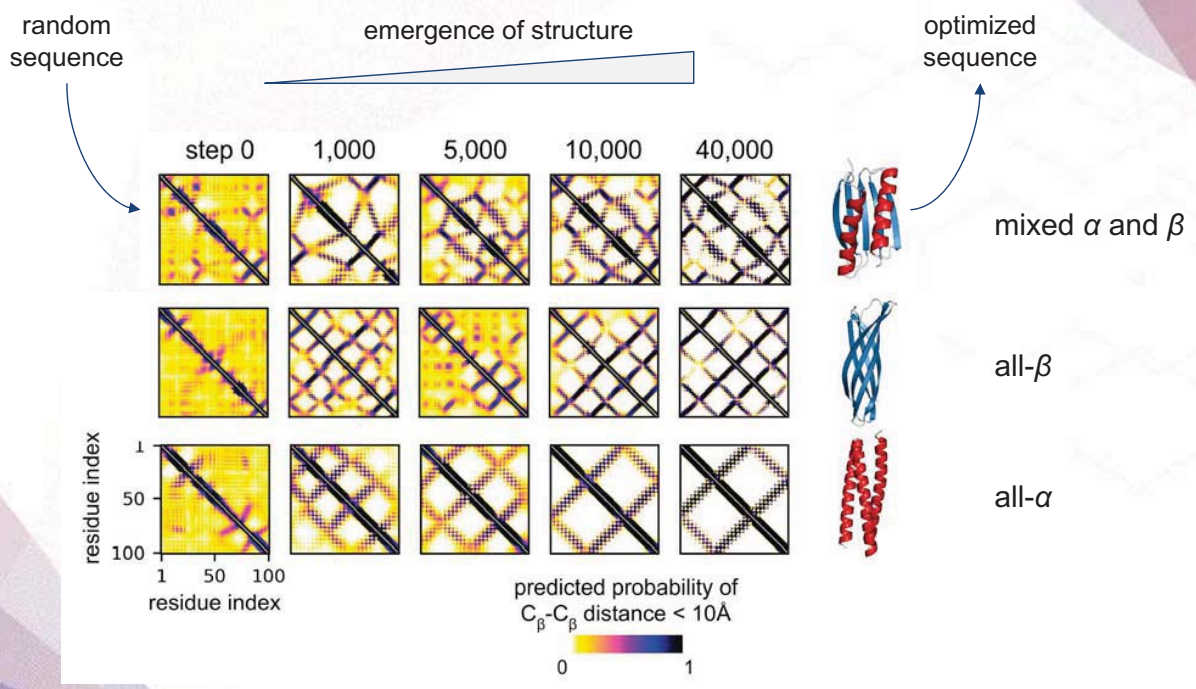
84

Hallucination: 단백질 구조 예측모델을 활용한 디자인



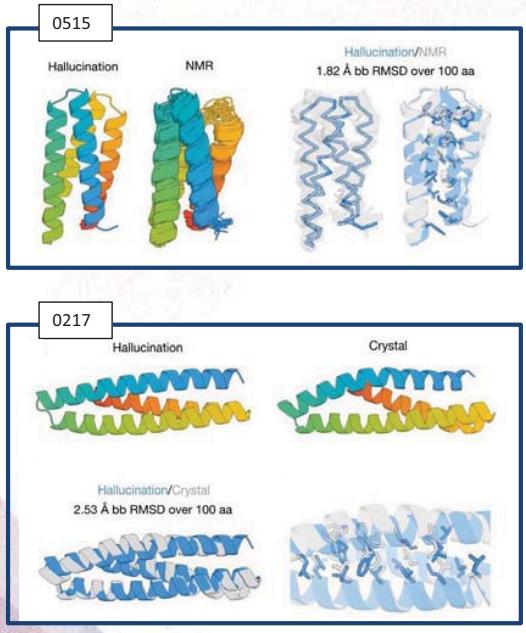
Anishchenko, I., Pellock, S., et al. *Nature* (2021)

Hallucination: 단백질 구조 예측모델을 활용한 디자인

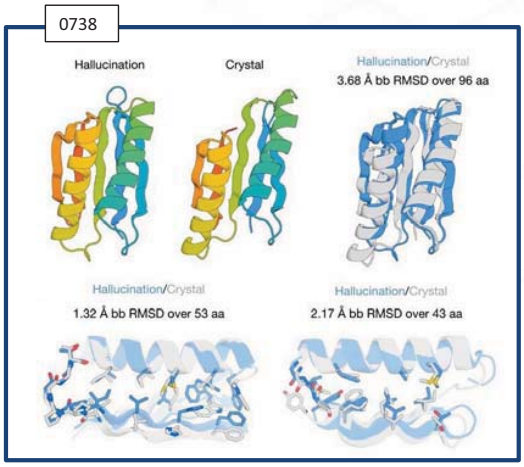


Anishchenko, I., Pellock, S., et al. *Nature* (2021)

Hallucination: 단백질 구조 예측모델을 활용한 디자인



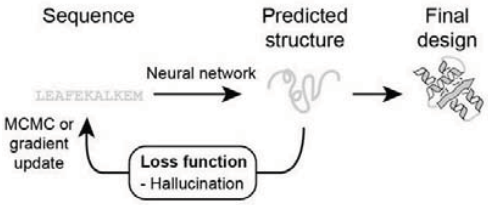
Structures of 3 hallucinations were confirmed experimentally



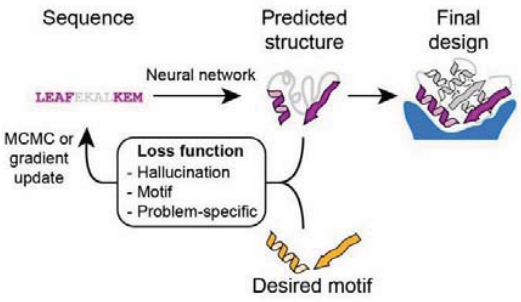
Anishchenko, I., Pellock, S., et al. *Nature* (2021)

87

Constrained Hallucination: 기능 요소의 도입



Free hallucination:
generate novel protein folds



Constrained hallucination:
generate scaffolds harboring pre-specified functional sites

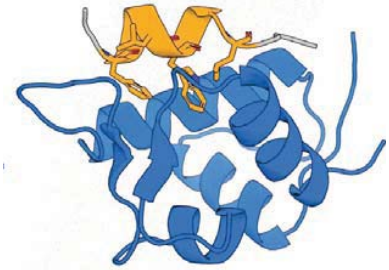
Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., *Science* (2022)

88

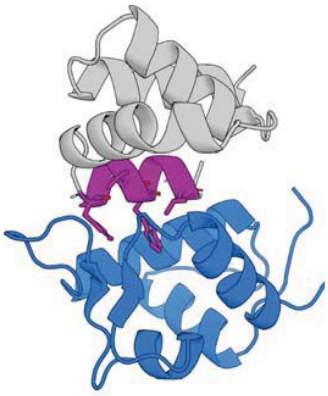
Constrained Hallucination: 결합저해제 설계

Scaffolding p53 helix to bind cancer-signaling protein mdm2

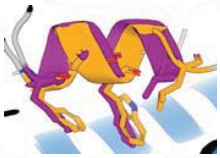
Native motif
Design motif
Binding partner



Native



Hallucination



Native vs designed motif

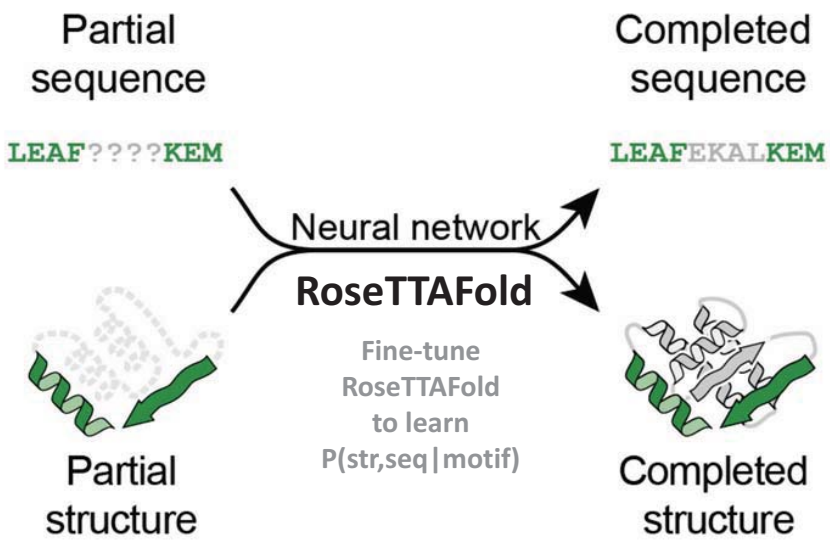
장점: objective function 설계만 잘하면 추가적인 딥러닝 모델 학습 필요 없이 사용 가능
단점: 수천~수만번의 구조예측 필요 → 계산 시간 오래걸림

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., *Science* (2022)

89

Inpainting기술을 활용한 단백질 디자인

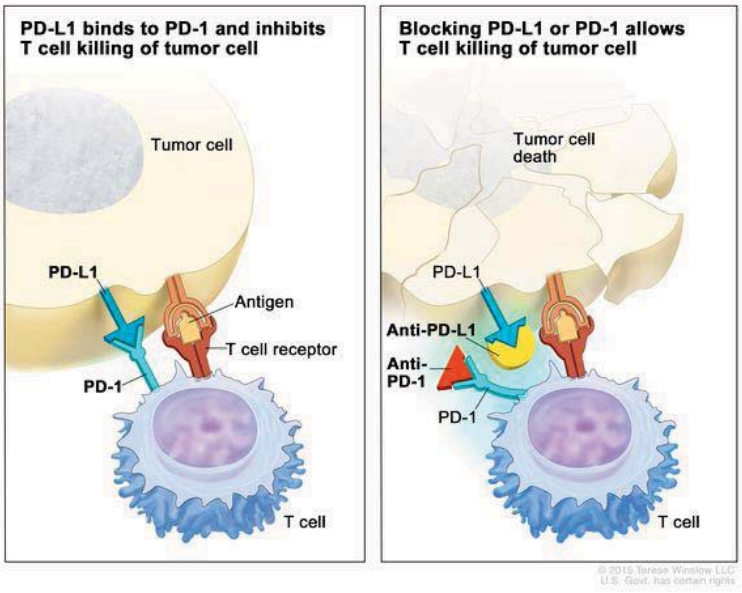
기능 모티프를 가지는 단백질 설계 = 부분 정보를 바탕으로 한 전체 정보 완성 문제 (inpainting)



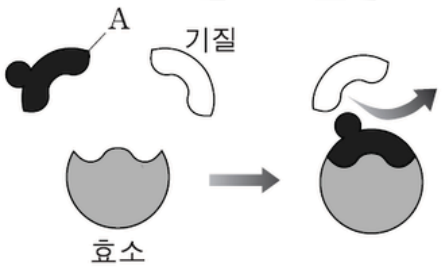
Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., *Science* (2022)

90

Inpainting기술을 활용한 단백질 디자인



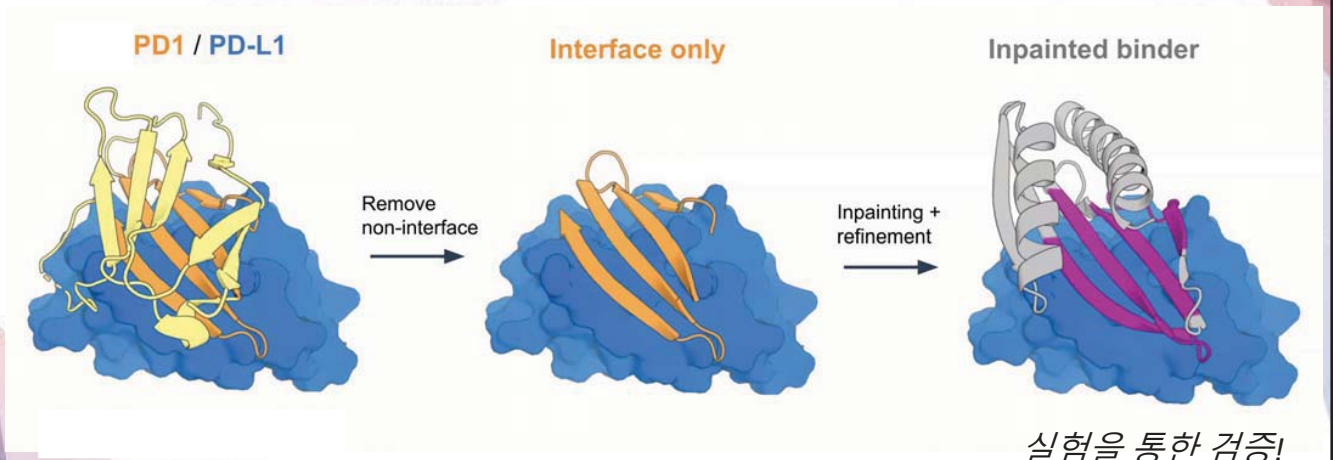
암세포가 가진 PD-L1 단백질이 세포성 면역을 방해
→ PD-1, PD-L1 결합을 방해하는 경쟁적 저해제를 만들자!



Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., *Science* (2022)

91

Inpainting기술을 활용한 단백질 디자인



실험을 통한 검증!

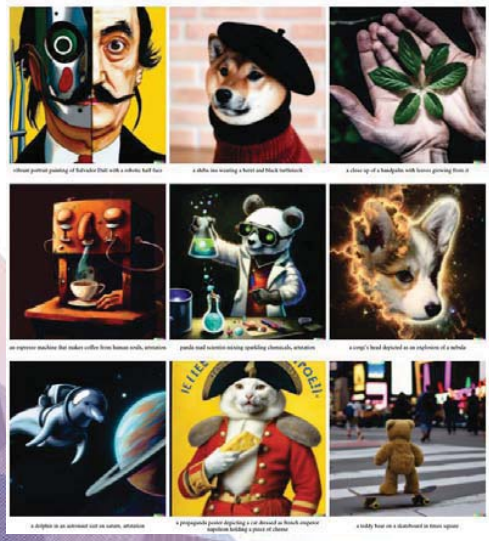
Motif**가 있어야만 디자인 가능* ***& 모티프 종류, 위치가 고정되면 항상 같은 결과!

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., *Science* (2022)

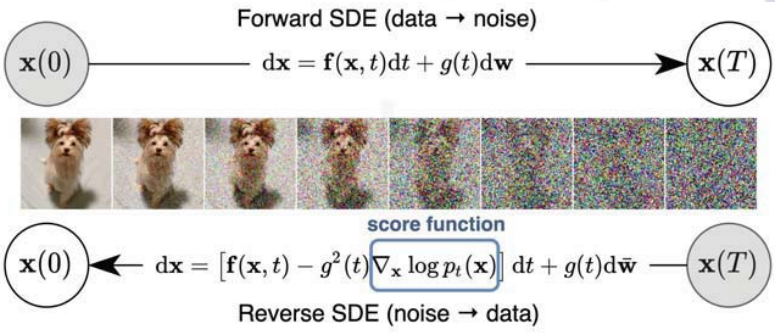
92

폭넓은 응용이 가능한 인공지능 모델은 없을까?

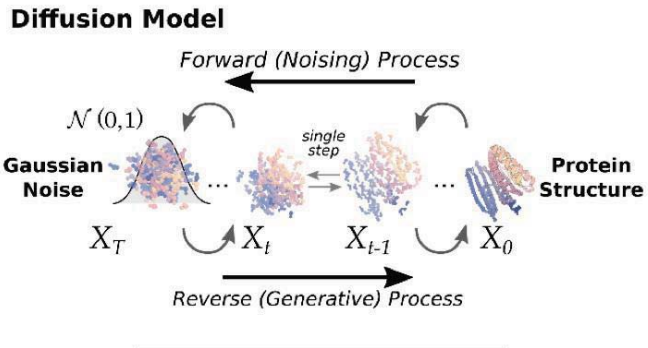
- 원하는 것: 다양한 단백질 구조의 생성
- 인공지능 분야에서 비슷한 연구?: 그림 생성 모델

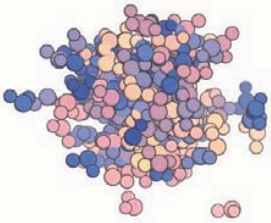


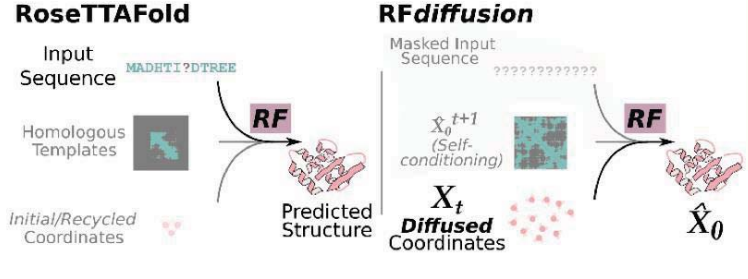
Diffusion Model



RFdiffusion: 생성모델을 활용한 단백질 디자인

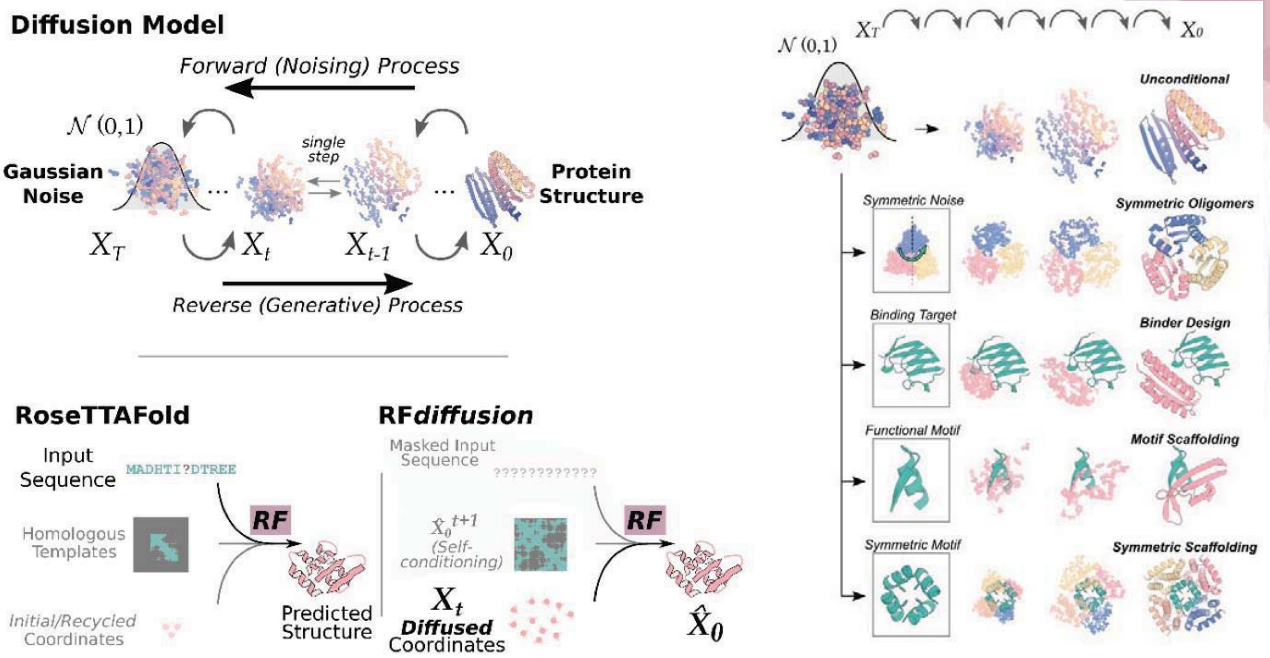






Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

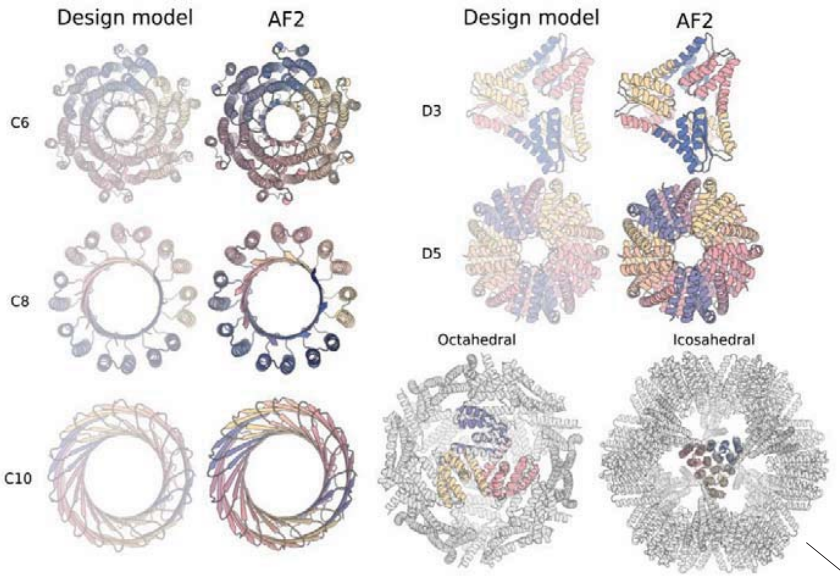
RFdiffusion: 생성모델을 활용한 단백질 디자인

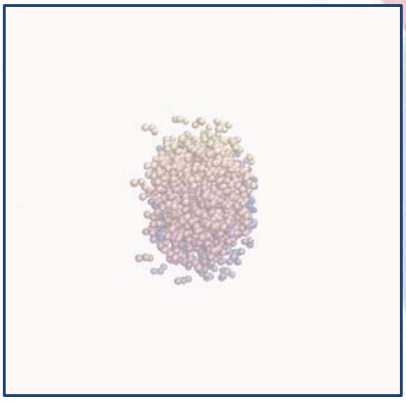


Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

95

RFdiffusion을 활용한 symmetric assembly 디자인



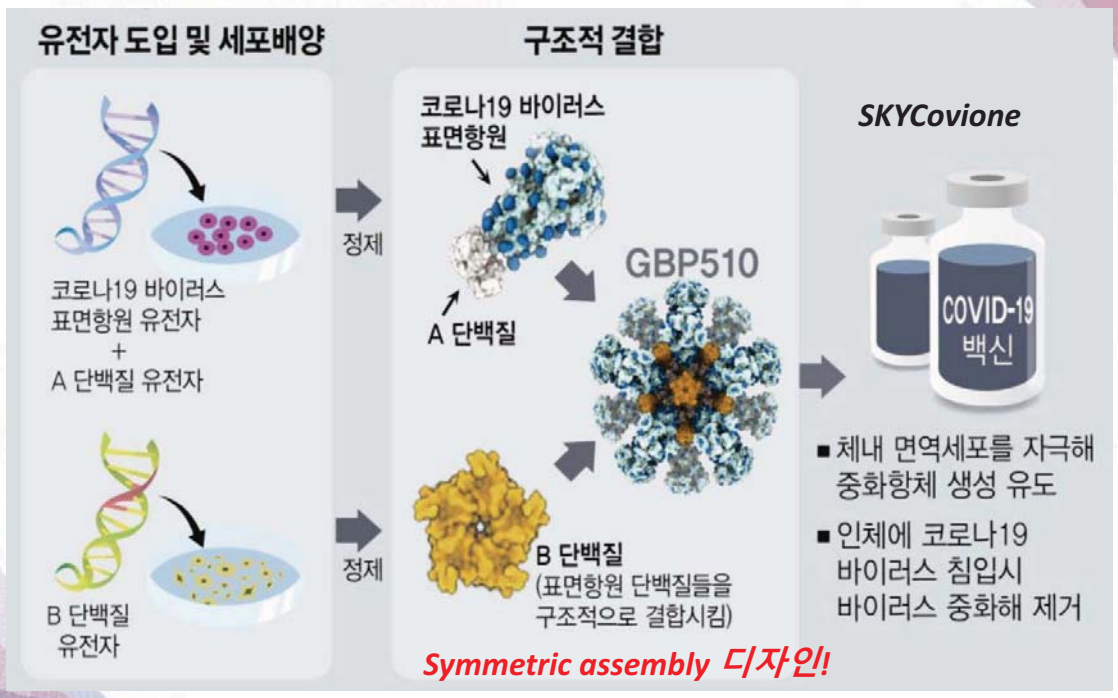


단백질 재조합 백신 디자인에 활용

Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

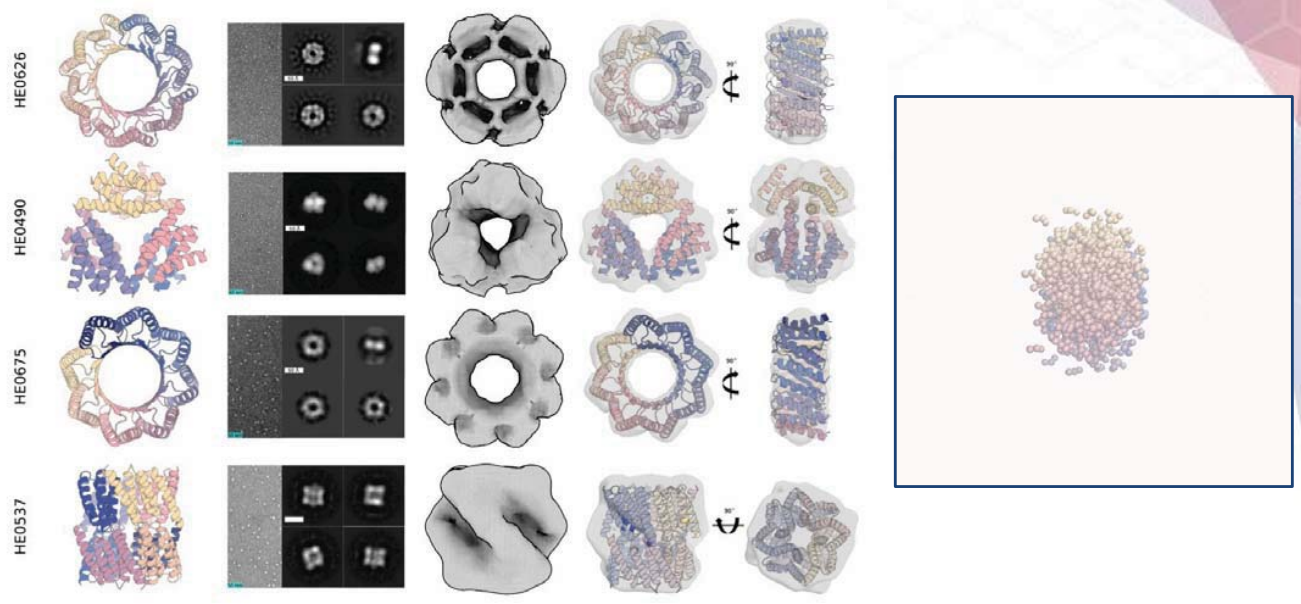
96

RFdiffusion을 활용한 symmetric assembly 디자인



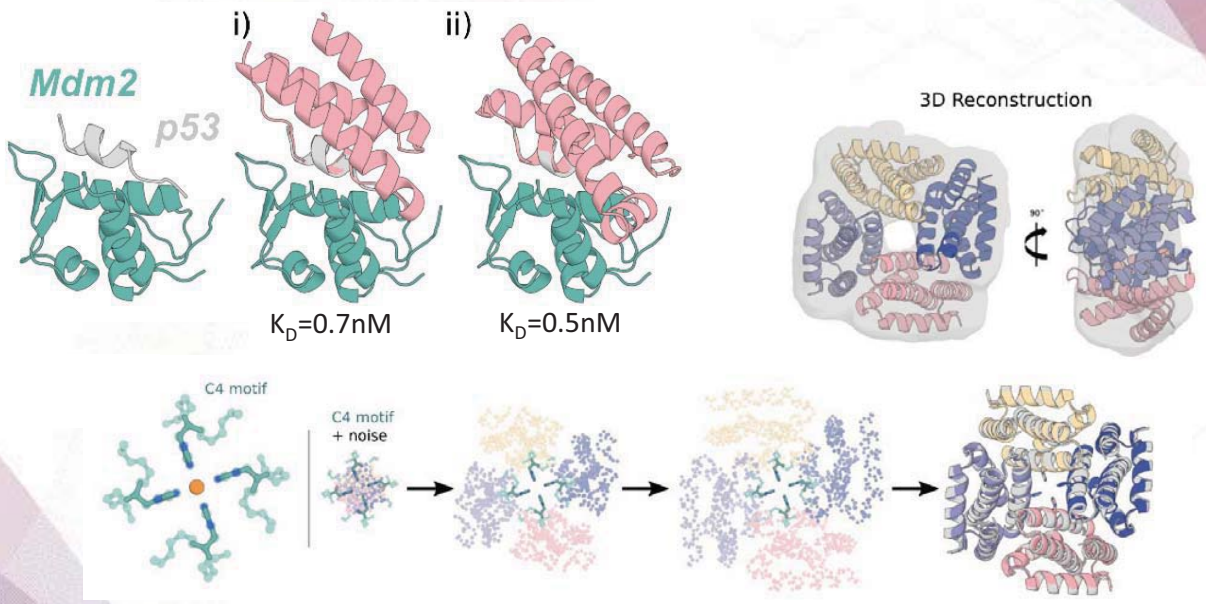
Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

RFdiffusion을 활용한 symmetric assembly 디자인



Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

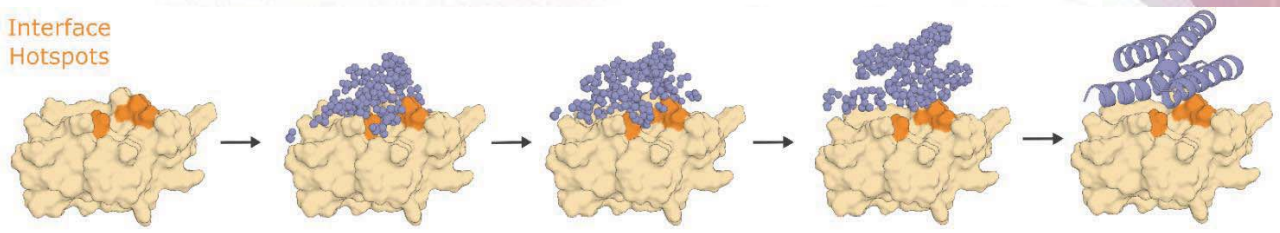
RFdiffusion을 활용한 motif scaffolding

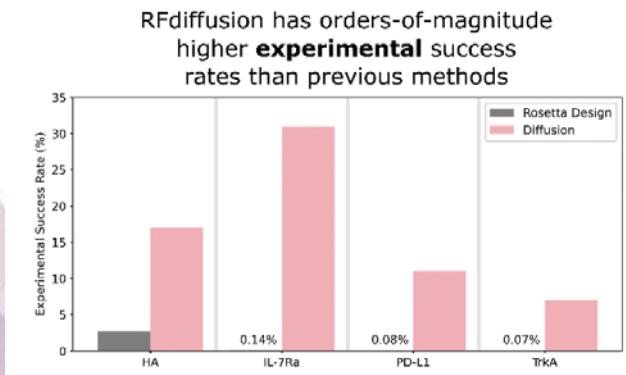


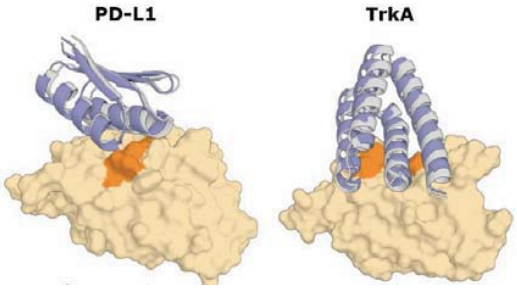
Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

99

RFdiffusion을 활용한 de novo binder 디자인





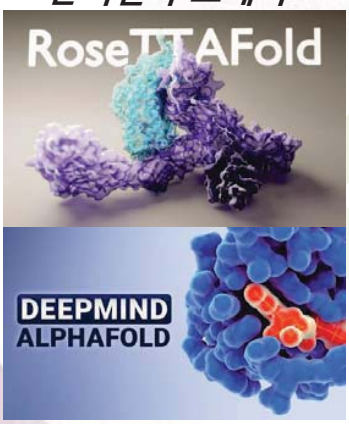


Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

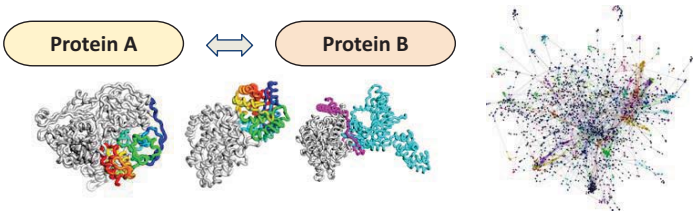
100

단백질 구조 예측의 무궁무진한 응용가능성

인공지능 기반 단백질 구조예측

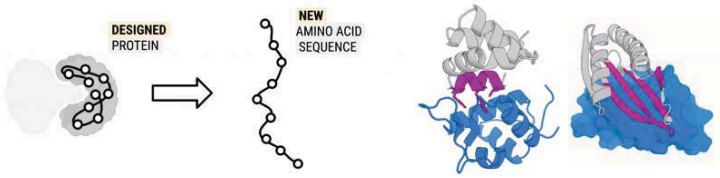


단백질 사이의 상호작용 예측



Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

인공지능 기반 단백질 디자인

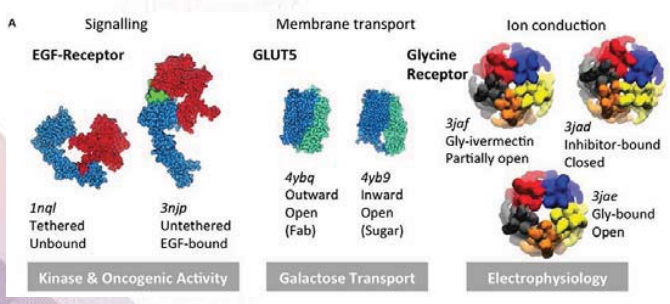


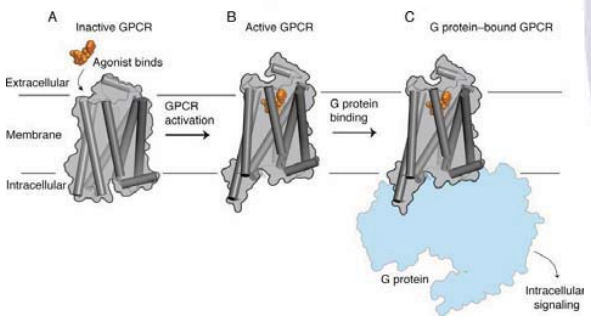
Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., et al, *Science* (2022)

101

남아있는 문제들

- Multi-state 구조 예측
  - 단백질은 굉장히 dynamic한 분자
  - 실제 여러 형태(state)의 구조를 가질 수 있고, 다른 분자와의 상호작용에 따라 그 구조가 변하기도 함

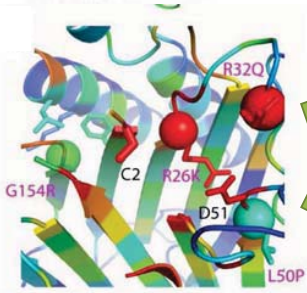




102

남아있는 문제들

- 변이에 따른 단백질의 특성 변화 예측
  - 변이에 따른 구조/안정성/활성도 변화 예측
  - 질병 이해 및 다양한 단백질 engineering을 위해 필요

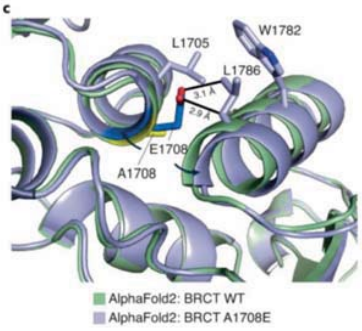


구조변화

안정성 변화

활성도 변화

구조 변화 예측 실패



남아있는 문제들

- 다른 분자와의 상호작용 예측
  - 단백질-단백질 상호작용 예측이 어느정도 가능하긴 하나, 아직 절대적인 성능은 만족스럽지 않음.
  - 특히, 항원-항체 및 숙주-병원체 상호작용 예측이 어려움
  - DNA/RNA/유기분자와의 상호작용 예측도 불가능
  - 신약개발을 위해서는 필수적으로 해결해야할 과제