Protein Structure Prediction & Ligand Docking 실습

김유빈, 정우엽

실습에 들어 가기 전에

Download Chimera

- **Go to** https://www.cgl.ucsf.edu/chimera/download.html

Current Production Releases

- See the release notes for a list of new features and other information.
- For more recent changes, use the snapshot and daily builds; they are less tested but usually reliable.

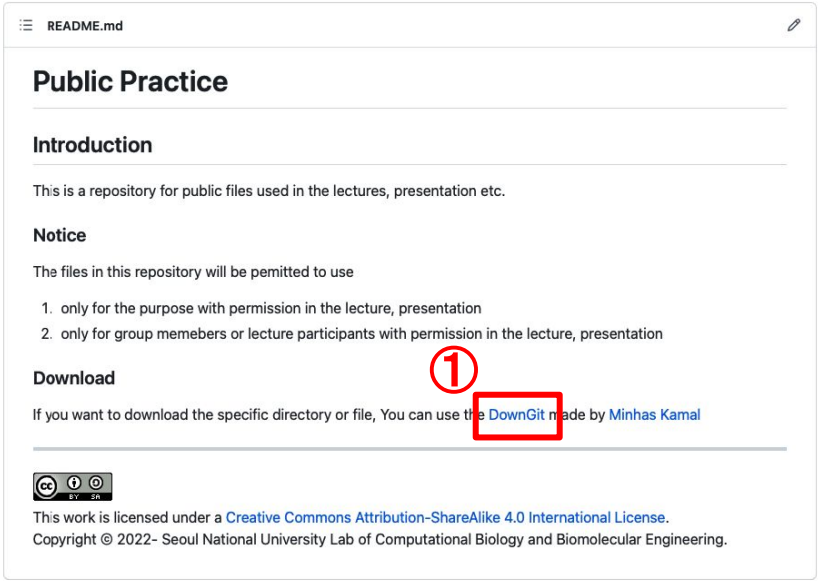
- **64-bit Releases:**

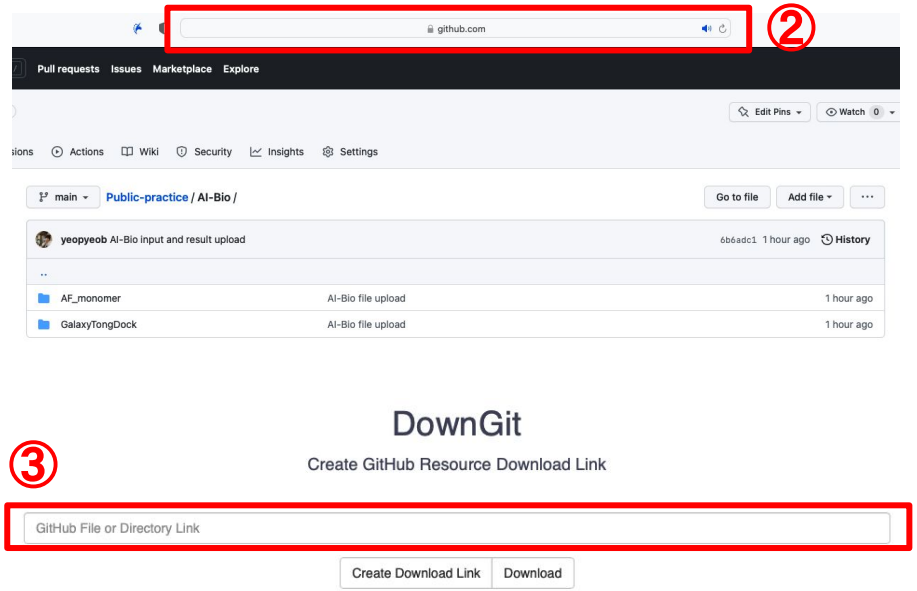
Platform	Installer, Size, and Checksum	Date	Notes
Microsoft Windows 64-bit	chimera-1.16-win64.exe Size: 152332561 bytes MD5: 9672aa27cc7ea1d6cfe9c8680516c741	Dec 17, 2021	Instructions Documentation Runs on Windows 7 or later.
Mac OS X 64-bit	chimera-1.16-mac64.dmg Size: 192170325 bytes MD5: 02cef4e3bf4e2ad5aae44c7104328ade	Dec 17, 2021	Instructions Documentation Runs on Mac OS X 10.12 or later.
Linux 64-bit	chimera-1.16-linux_x86_64.bin Size: 154080130 bytes MD5: 0167c57d7e24c9b69e04fd0dabc5ce87	Dec 17, 2021	Instructions Documentation Compiled on CentOS 5.11.

- **32-bit releases are no longer supported.**

Input & Result File Download

- **Go to** https://github.com/seoklab/Public-practice

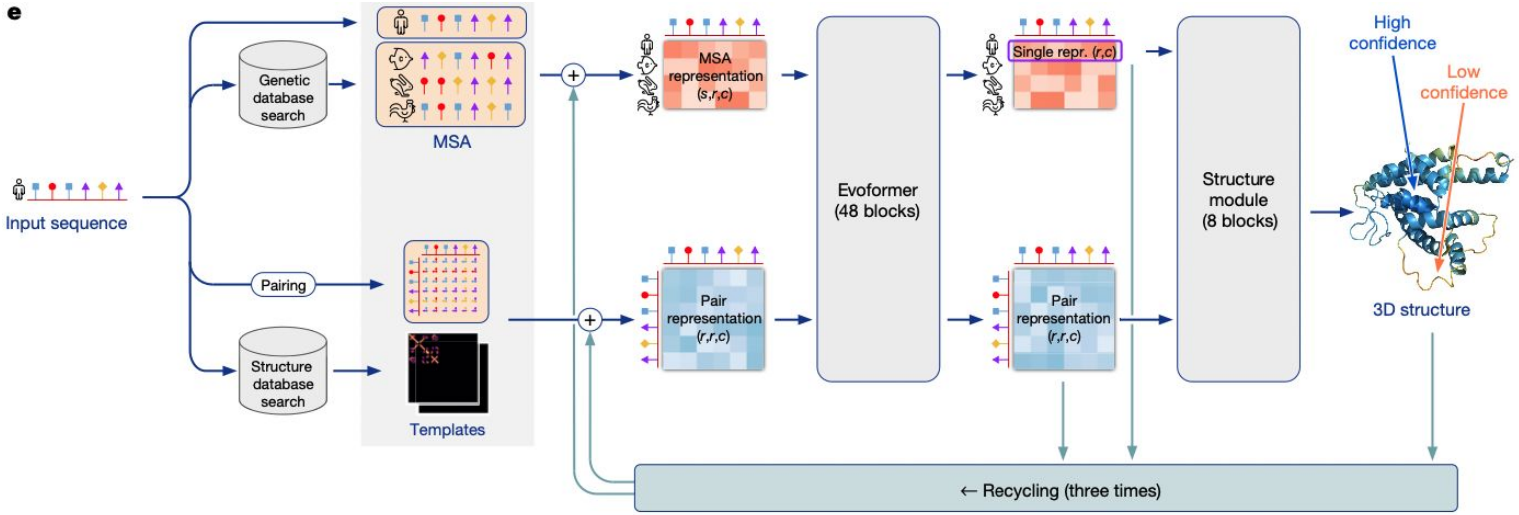




AF2, AF-multimer 실습

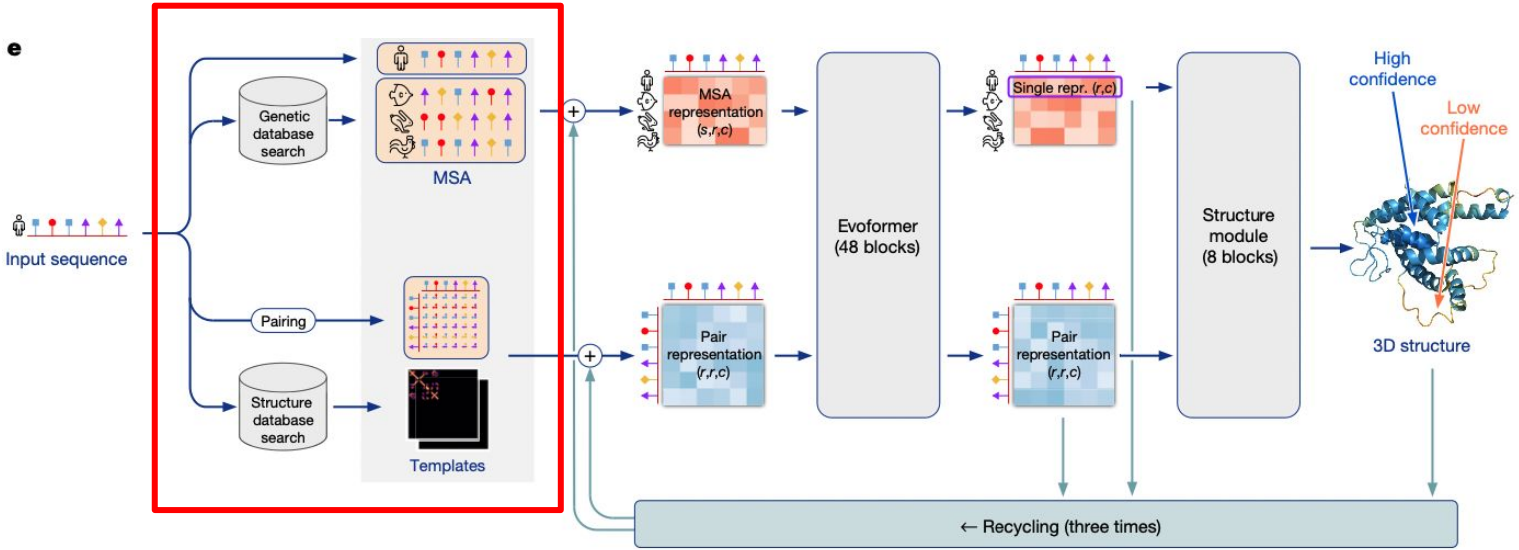
- Antibody-Antigen complex prediction

What is AlphaFold2, AlphaFold-Multimer?



Nature 596, 583–589 (2021)

What is AlphaFold2, AlphaFold-Multimer?



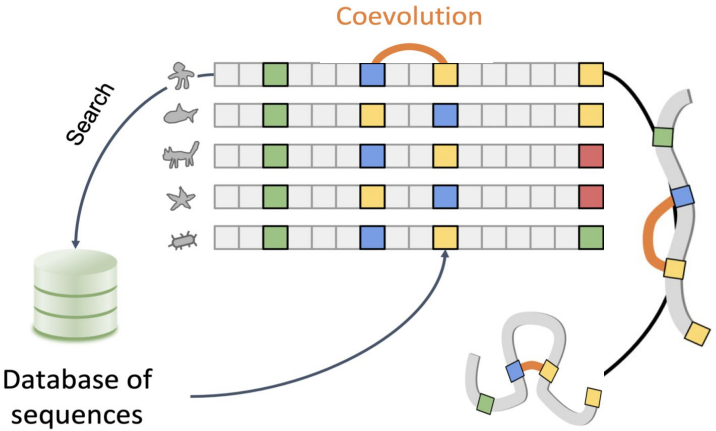
1. Build MSA & Get template information (HHblits, HMMer)

Nature 596, 583–589 (2021)

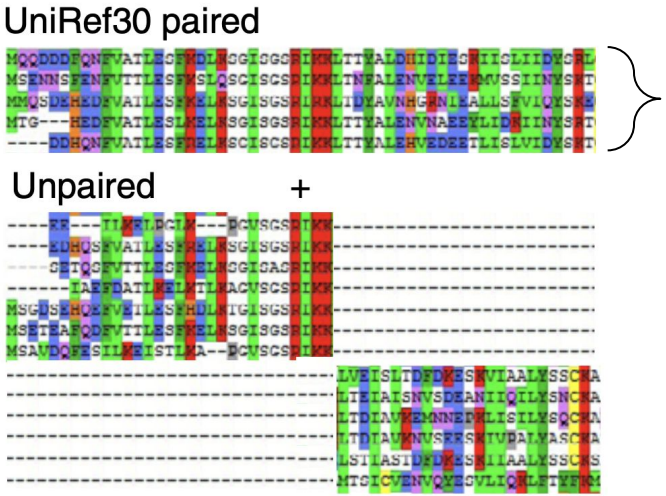
What is AlphaFold2, AlphaFold-Multimer?

1. Build MSA & Get template information (HHblits, HMMer)

Monomer

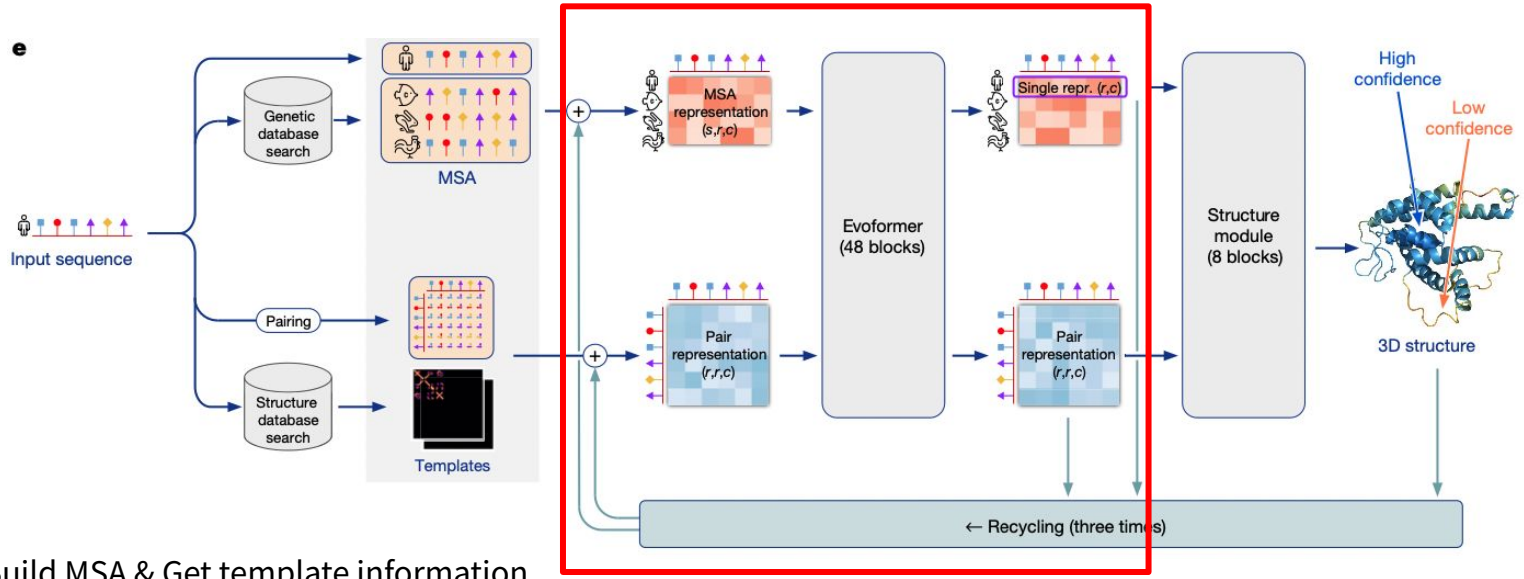


Multimer - Paired MSA



Nat Methods 19, 679–682 (2022).
Nature 596, 583–589 (2021)

What is AlphaFold2, AlphaFold-Multimer?

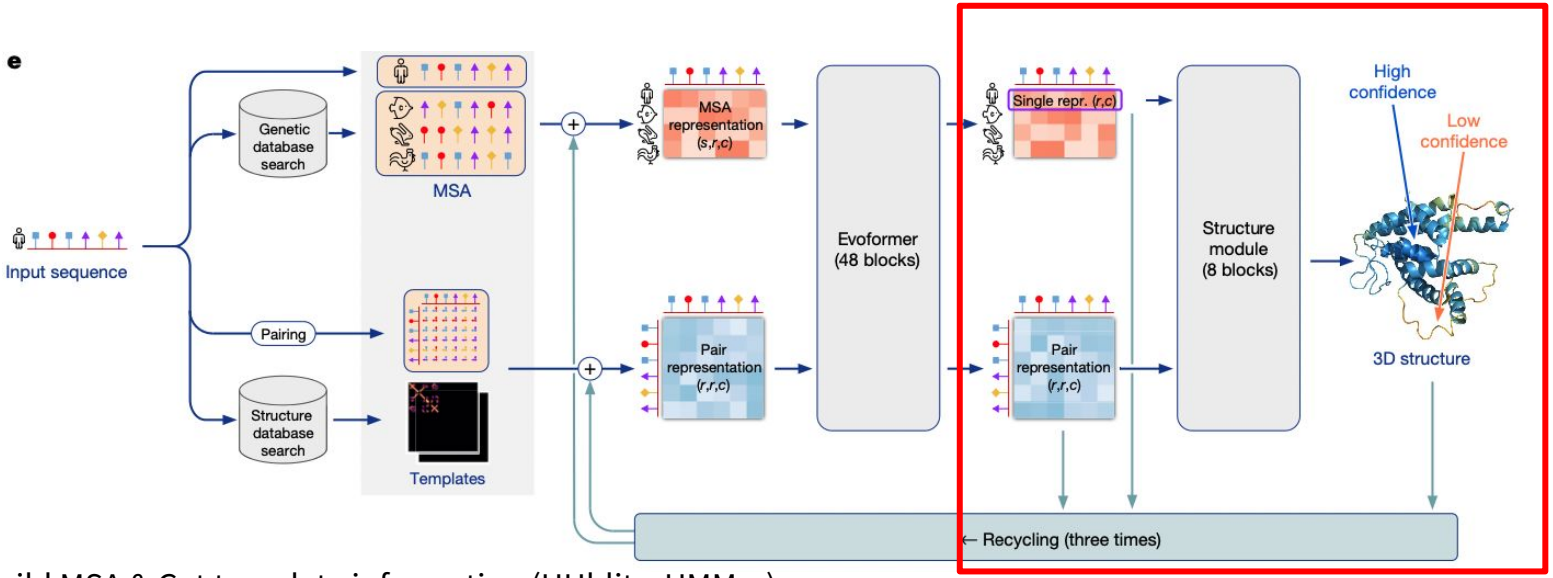


1. Build MSA & Get template information

2. Obtain MSA representation & Pair representation through Evoformer

Nature 596, 583–589 (2021)

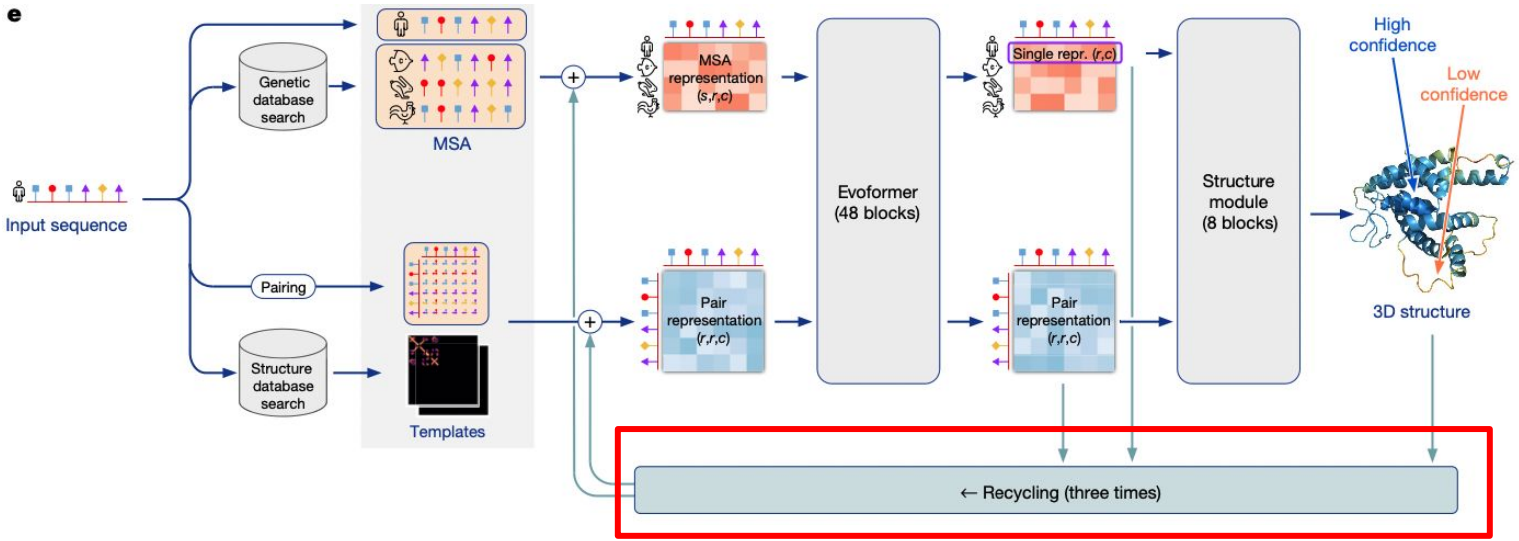
What is AlphaFold2, AlphaFold-Multimer?



1. Build MSA & Get template information (HHblits, HMMer)
2. Obtain MSA representation & Pair representation through Evoformer
3. **Predict structure using Structure module**

Nature 596, 583–589 (2021)

What is AlphaFold2, AlphaFold-Multimer?

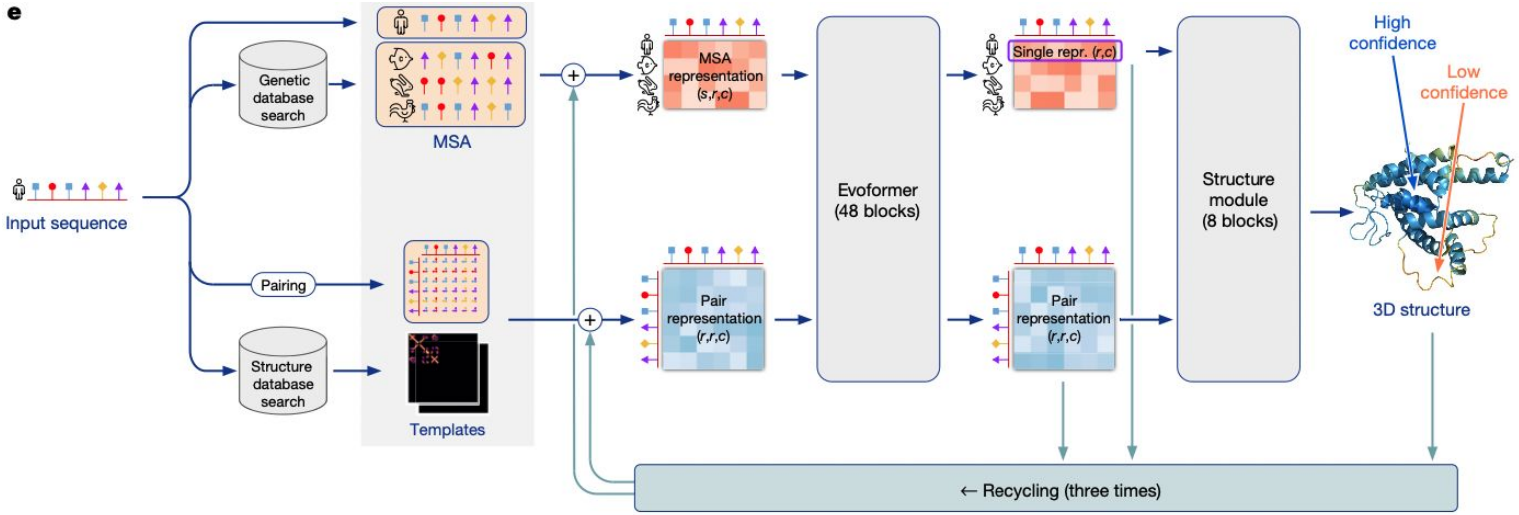


1. Build MSA & Get template information (HHblits, HMMer)
2. Obtain MSA representation & Pair representation through Evoformer
3. Predict structure using Structure module

4. Iterative Refinement through recycling

Nature 596, 583–589 (2021)

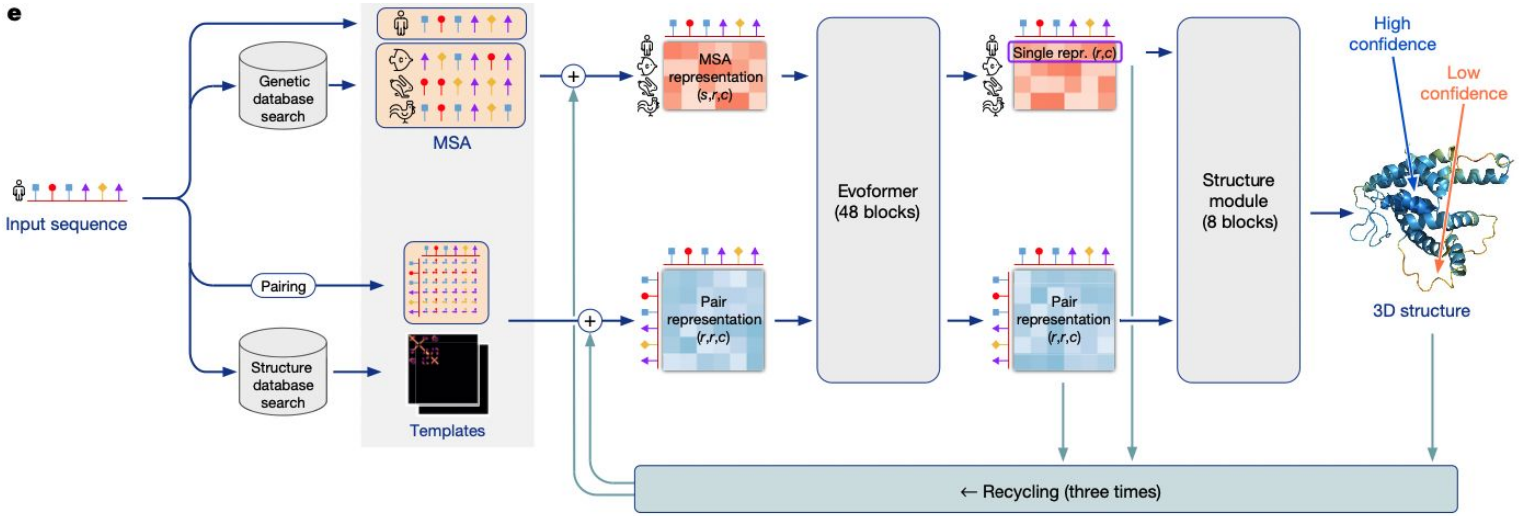
What is AlphaFold2, AlphaFold-Multimer?



1. Build MSA & Get template information (HHblits, HMMer)
2. Obtain MSA representation & Pair representation through Evoformer
3. Predict structure using Structure module
4. Iterative Refinement through recycling

Nature 596, 583–589 (2021)

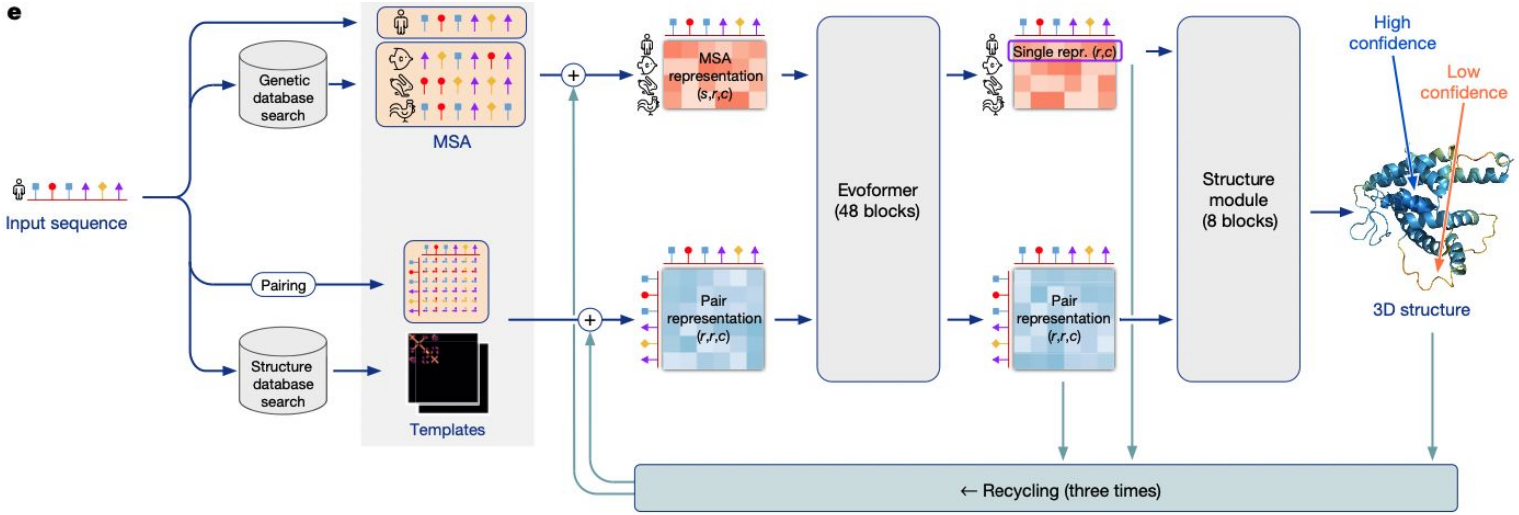
What is AlphaFold2, AlphaFold-Multimer?



1. Build MSA & Get template information (HHblits, HMMer) - **several hours, need ~2TB for database**
2. Obtain MSA representation & Pair representation through Evoformer
3. Predict structure using Structure module
4. Iterative Refinement through recycling

Nature 596, 583–589 (2021)

What is AlphaFold2, AlphaFold-Multimer?



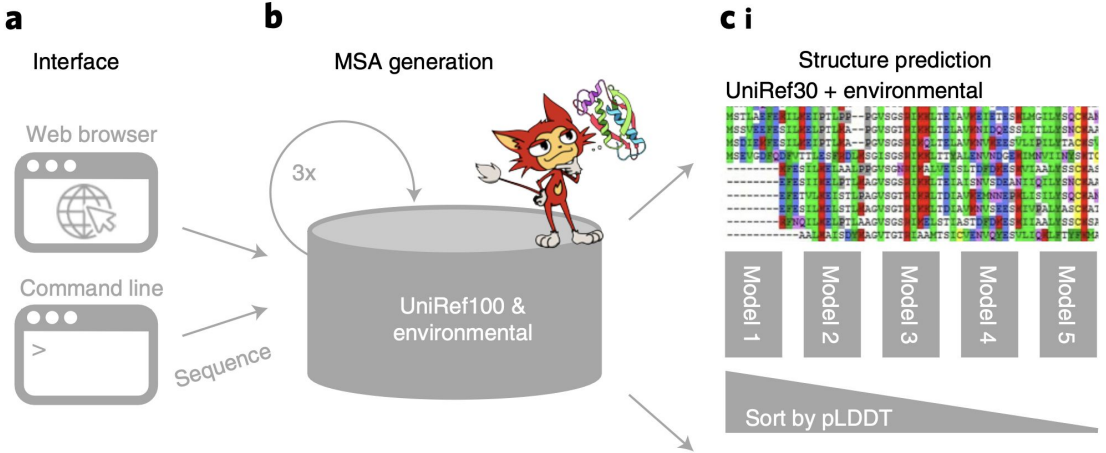
1. Build MSA & Get template information (HHblits, HMMer)
2. Obtain MSA representation & Pair representation through Evoformer
3. Predict structure using Structure module
4. Iterative Refinement through recycling

Need GPUs with a large amount of RAM

Nature 596, 583–589 (2021)

ColabFold

- https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
- Build MSA by MMseqs2 (40-60 folds Faster)
- Speed up structure predictions by avoiding recompilation and adding an early stop criterion (~90 folds faster)



Nat Methods 19, 679–682 (2022).

Model Protein - 7bem

7BEM

Display Files | Download Files

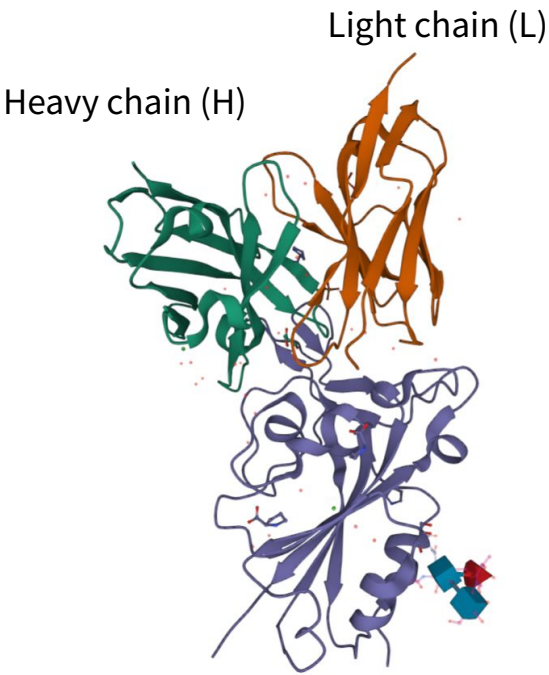
Crystal structure of the receptor binding domain of SARS-CoV-2 Spike glycoprotein in complex with COVOX-269 scFv

PDB DOI: 10.2210/pdb7BEM/pdb

Classification: VIRAL PROTEIN/IMMUNE SYSTEM
Organism(s): Homo sapiens, Severe acute respiratory syndrome coronavirus 2
Expression System: Homo sapiens
Mutation(s): Yes

Deposited: 2020-12-24 **Released:** 2021-03-03
Deposition Author(s): Zhou, D., Zhao, Y., Ren, J., Stuart, D.
Funding Organization(s): Medical Research Council (MRC, United Kingdom), CAMS Innovation Fund for Medical Sciences (CIFMS)

Modeling 7bem with ColabFold & GalaxyTongDock!



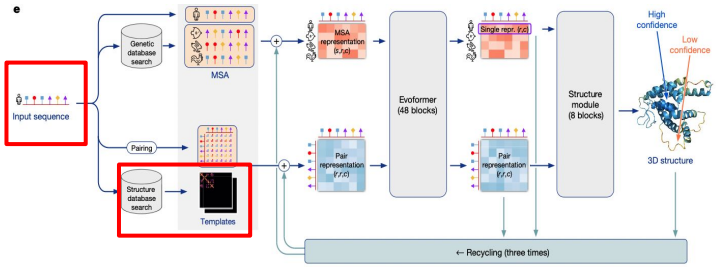
Input Protein sequence

- **Input Files : Public-practice/AI-Bio/AF_multimer/7bem.pdb, 7bem.fa**
  - **Fasta File of 7bem (Protein sequence file)**
  - **RCSB Database for proteins (**https://www.rcsb.org/**) (Protein crystal structure file)**

https://github.com/seoklab/Public-practice/tree/main/AI-Bio/AF_multimer

7bem.fa
>7BEM_1|Chain A[auth H]|COVOX-269 Vh domain|Homo sapiens (9606)
QVQLVESGGGLIQPGGSLRLSCAASGLTVNRNYMSWIRQAPGKGLEWVSVIYSGGSTFYADSVKGRFTI…
>7BEM_2|Chain B[auth L]|COVOX-269 Vl domain|Homo sapiens (9606)
AIQLTQSPSFLSASIGDRVTITCRASQGISSYLAWYQQKPGKAPKLLIYAASTLQSGVPSRFSGSGSGTEF…
>7BEM_3|Chain C[auth E]|Spike glycoprotein|Severe acute respiratory syndrome coronavirus 2 (2697049)
NLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR…

Input Protein sequence



Input protein sequence(s), then hit Runtime -> Run all

query_sequence: " QVQLVESGGGLIQPGGSLRLSCAASGLTVNRNYMSWIRQAPG:AIQLTQSPSFLSASIGDRVTITCRASQGISSYLAWYQQKPGKAPKLLI

- Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: " 7bem

use_amber:

template_mode: pdb70

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (PDB or mmCIF format, see notes below)

코드 표시

← 각 **chain** 사이는 **:** 로 구분

← **jobname (ex. 7bem)**

← **Constrained relaxation of structures using Amber force field (stereochemical violation 제거)**

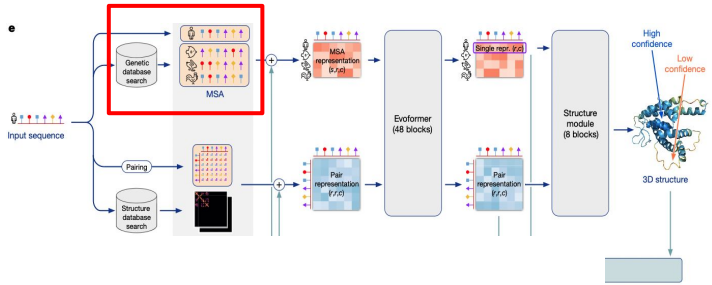
Template_mode:

i) None - no template information is used (Better to use for prediction of Ab-Ag complex)

ii) pdb70 - detect templates in pdb70 (database with maximum pairwise sequence identity of 70%)

iii) custom - upload & search own templates

MSA options



MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: MMseqs2 (UniRef+Environmental)

pair_mode: unpaired+paired

MSA mode:

i) MMseqs2 (UniRef + Environmental) - use UniRef sequences + Environmental sequences

ii) MMseqs2 (UniRef) - only use UniRef sequences

iii) single_sequence - only use single sequence

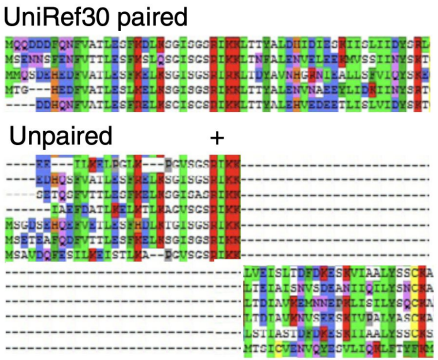
iv) custom - upload own MSA

Pair mode:

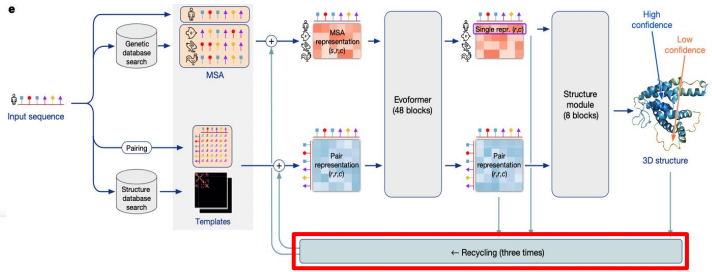
i) unpaired + paired - pair sequences from same species + unpaired MSA

ii) unpaired - separate MSA for each chain

iii) paired - only use paired sequences



Advanced Settings



Advanced settings

model_type: AlphaFold2-multimer-v2

- "auto" = protein structure prediction using "AlphaFold2-ptm" and complex prediction "AlphaFold-multimer-v2". For complexes "AlphaFold-multimer-v[1,2]" and "AlphaFold-ptm" can be used.

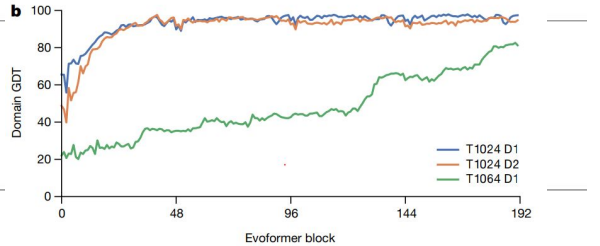
num_recycles: 3

save_to_google_drive: ☑

- if the save_to_google_drive option was selected, the result zip will be uploaded to your Google Drive

dpi: 200

- set dpi for image resolution



Don't forget to hit `Runtime` -> `Run all` after updating the form.

코드 표시

You are logged into Google Drive and are good to go!

Run prediction!

Install dependencies

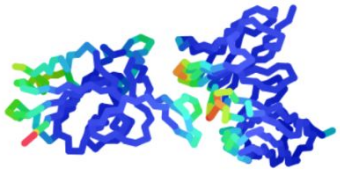
Run Prediction

코드 표시

```
Downloading alphafold2 weights to .: 100%|██████████| 3.82G/3.82G [00:24<00:00, 167MB/s]
2022-10-13 10:12:41,694 Found 8 citations for tools or databases
2022-10-13 10:12:48,648 Query 1/1: 7bem_89674 (length 412)
COMPLETE: 100%|██████████| 450/450 [elapsed: 00:03 remaining: 00:00]
2022-10-13 10:13:22,194 Sequence 0 found templates: ['6wzn_A', '7q9g_E', '7km6_H', '7lx5_A', '1igt_B', '7dv4_H', '1nmb_H', '1nmc_B', '7cu5_A', '5fgb_E', '7n8h_C', '7cho_B', '7n3h_H', '7orb_A', '4jr9_H', '7vsw_B', '7bem_H', '6wzn_A', '4rav_C', '6kyz_B']
2022-10-13 10:13:37,525 Sequence 1 found templates: ['6mg7_L', '6cr1_L', '6b14_L', '7v5n_A', '4rrp_B', '6wzn_A', '7vsw_I', '6tcs_A', '6zdg_C', '3eob_A', '6vun_A', '6wzn_A', '7cwn_F', '6vun_A', '2q20_B', '7shy_D', '4l1h_A', '7bem_L', '4k07_H', '6tcs_A']
2022-10-13 10:14:03,745 Sequence 2 found templates: ['6zdg_D', '7oao_EEE', '7pki_E', '7vmu_B', '7rxd_R', '7sn0_D', '7eam_A', '7d2z_B', '7oaq_EEE', '7vyr_C', '6xe1_E', '7m3i_R', '7n8h_F', '7sbp_A', '7a92_A', '7akj_B', '7kqb_A', '7e7b_B', '7n9t_C', '7e7d_C']
COMPLETE: 100%|██████████| 450/450 [elapsed: 00:01 remaining: 00:00]
2022-10-13 10:14:06,186 Running model_1
2022-10-13 10:17:32,485 model_1 took 206.3s (3 recycles) with pLDDT 89.1, ptmscore 0.581 and iptm 0.442
```

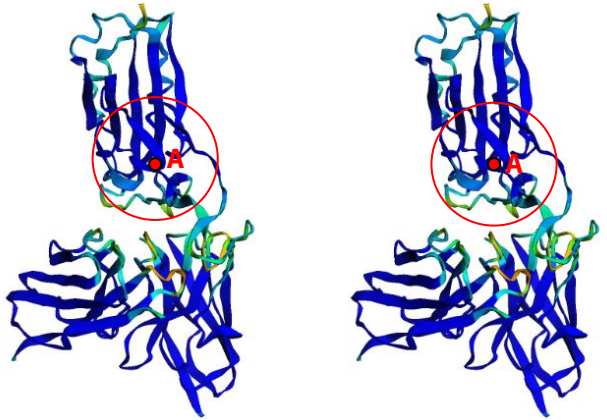
colored by chain

colored by pLDDT



- ~3m for prediction, ~4m for Amber relaxation
- Ranked by pTM score

What is pLDDT, pTM ?



- **pLDDT** (predicted local distance difference test)
  i. predict the per-residue lDDT-Cα score
  ii. local error metric
  iii. value from 0~100 (uncertain ~ certain)

- **pTM** (predicted TM score)
  i. whether the model is confident in overall domain packing
  ii. Value from (0-1] (not matched ~perfectly matched)

$$\text{TM-score} = \text{Max}\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

Bioinformatics 29(21): 2722-2728 (2013)
PROTEINS: Structure, Function, and Bioinformatics 57:702–710 (2004)

Visualization of predicted structures

Display 3D structure

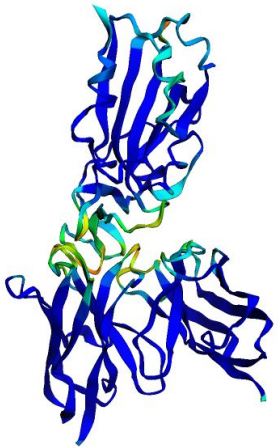
rank_num: 1

color: lDDT

show_sidechains:

show_mainchains:

코드 표시



plDDT: Very low (<50) Low (60) OK (70) Confident (80) Very high (>90)

Display 3D structure

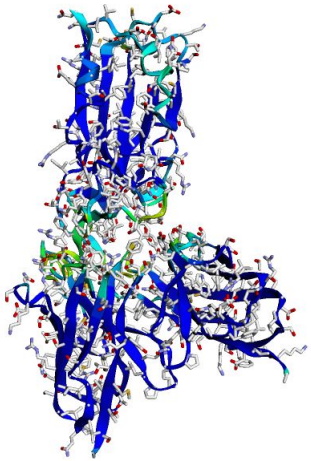
rank_num: 1

color: lDDT

show_sidechains:

show_mainchains:

코드 표시



plDDT: Very low (<50) Low (60) OK (70) Confident (80) Very high (>90)

pLDDT score in Loop

Display 3D structure

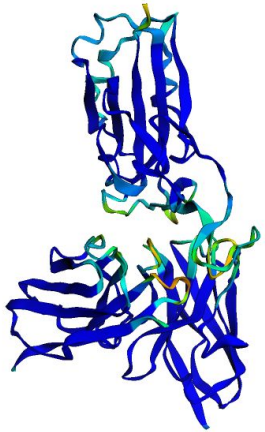
rank_num: 2

color: lDDT

show_sidechains:

show_mainchains:

코드 표시



pIDDT: Very low (<50) Low (60) OK (70) Confident (80) Very high (>90)

Display 3D structure

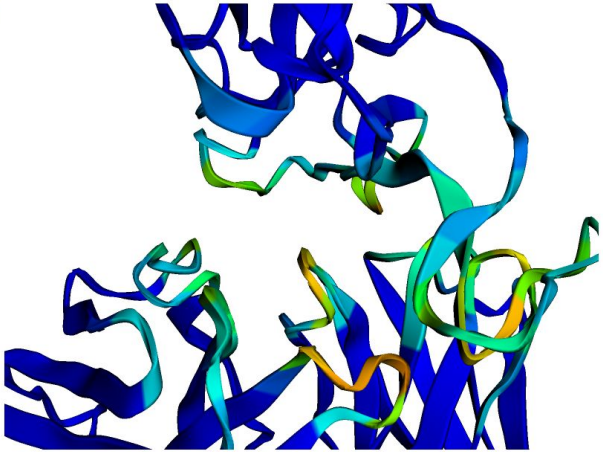
rank_num: 2

color: lDDT

show_sidechains:

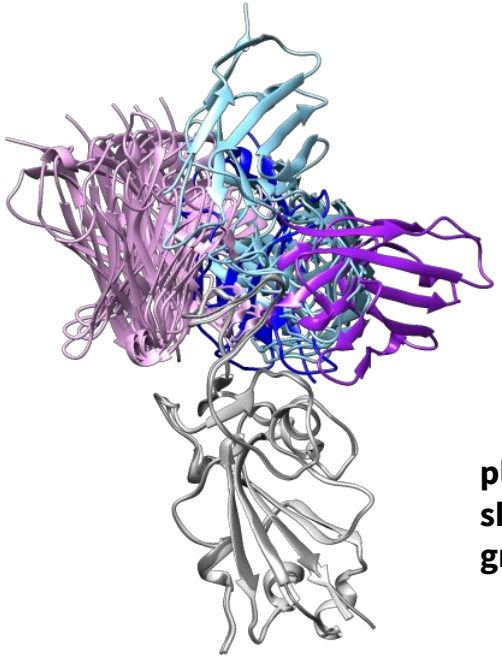
show_mainchains:

코드 표시

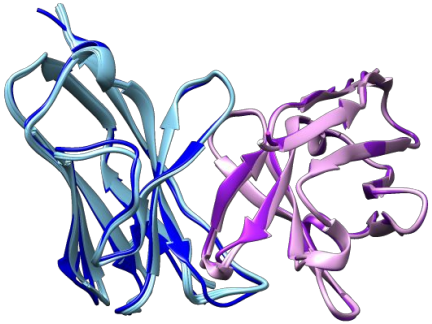


pIDDT: Very low (<50) Low (60) OK (70) Confident (80) Very high (>90)

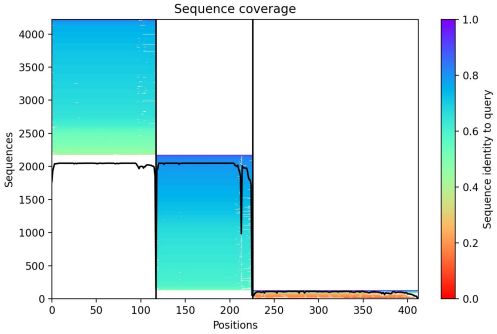
Visualization of predicted structures with Chimera



plum : Heavy chain,
skyblue: Light chain
gray: Antigen

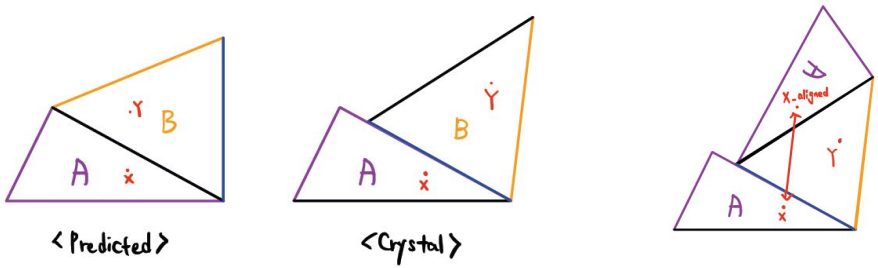


Antibody alignment



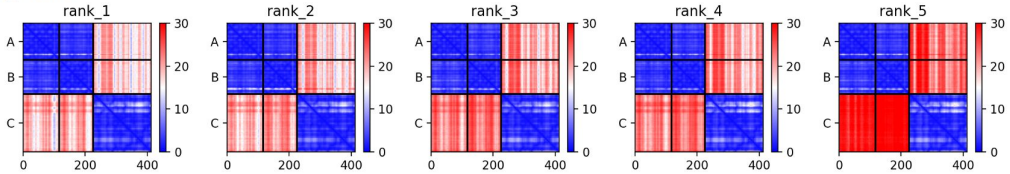
What is PAE ?

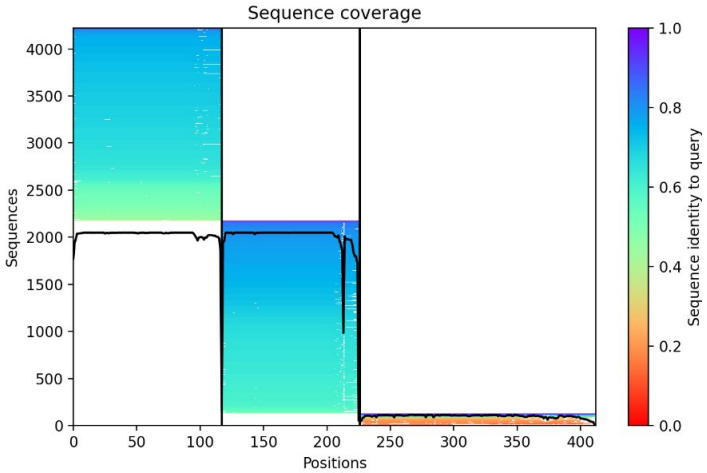
- **PAE** (predicted aligned error)
  - For every pair (x, y) of residues in the structure, PAE is calculated by AlphaFold's estimate of position error at residue x, while the predicted and true structures are aligned on residue y.
  - If the relative position of two domains is confidently predicted, the PAE values will be low (less than 5Å) for pairs of residues from each domain.

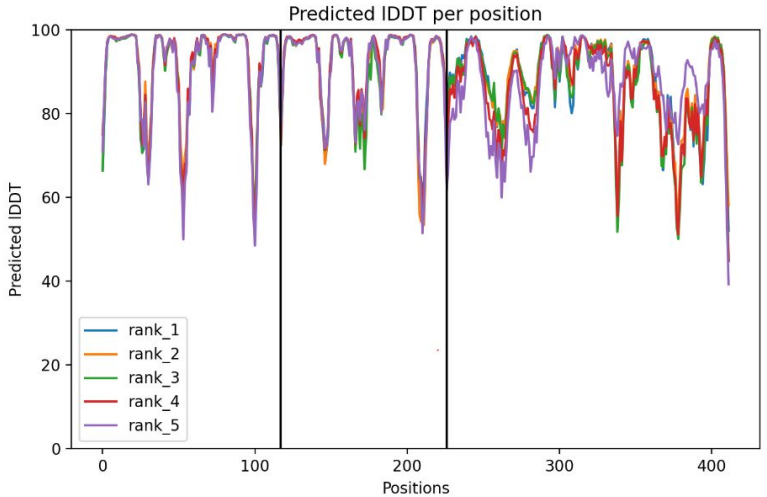


Plots for predicted structures

Plots for 7bem_89674







Successful case - 6yx9

Display Files ▾ Download Files ▾

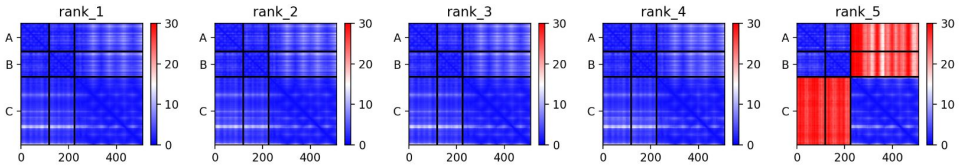
6YX9

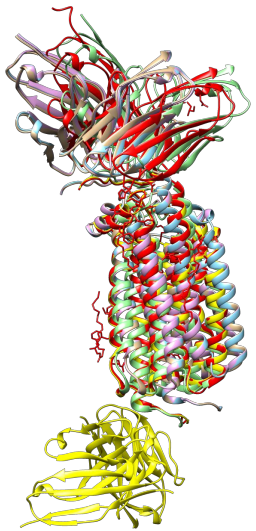
Cryogenic human adiponectin receptor 2 (ADIPOR2) at 2.4 A resolution determined by Serial Crystallography (SSX) using CrystalDirect

PDB DOI: 10.2210/pdb6YX9/pdb

Classification: MEMBRANE PROTEIN
Organism(s): Homo sapiens
Expression System: Drosophila melanogaster
Mutation(s): No
Membrane Protein: Yes PDBTM MemProtMD mpstruc

PAE score for 6yx9 - A: Heavy chain, B: Light chain, C: antigen



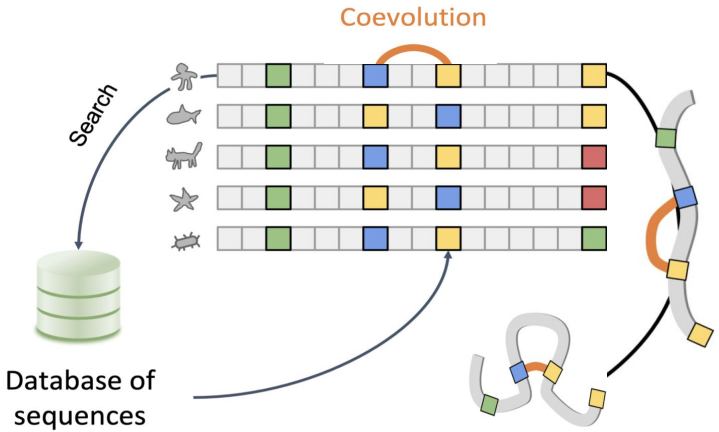


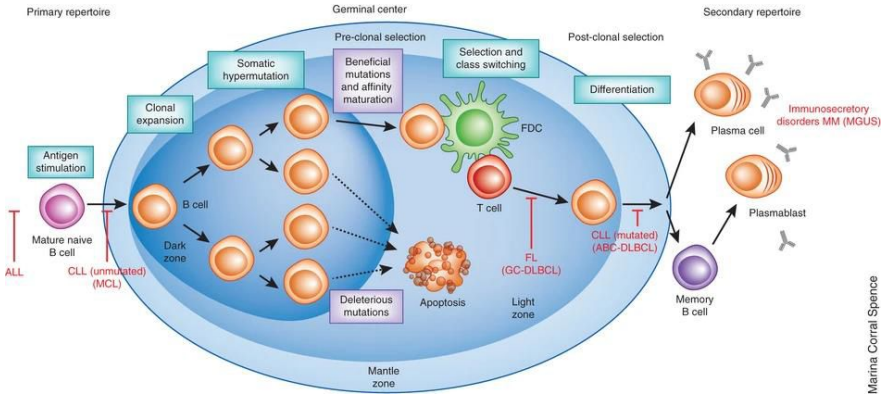
Crystal structure - red
Ranked #5 structure - yellow

- **PAE score can be a criterion of whether the binding geometry is well predicted !**

Why is it difficult to predict Antibody-Antigen complex with AF-multimer?

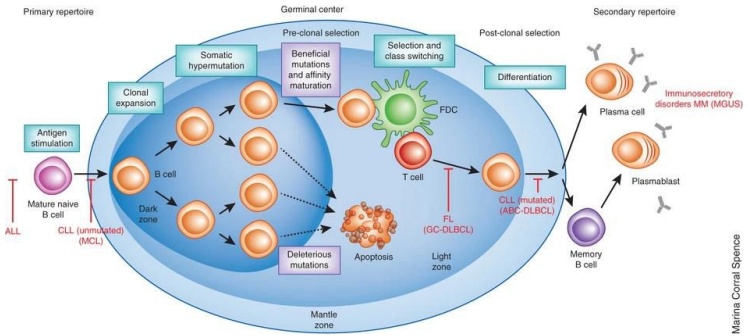
- Main idea of AF2 is to learn coevolution information from MSA
- For antibody-antigen complexes, coevolution information is limited due to the unique generation mechanism of antibodies





Why is it difficult to predict Antibody-Antigen complex with AF-multimer?

- As the antigen enters into our body, the antibody increases its affinity to the antigen rapidly through various mechanisms such as **somatic hypermutation** and **VD(J) recombination**
- This antibody-specific mutation mechanism leads to lack of coevolution information between antibody and antigen



Method	MSA	Template	Top1	Top3	Top5
AF-multimer	Paired	O	4 / 7 / 5 / 65	4 / 7 / 8 / 62	4 / 8 / 8 / 61
AF-multimer	Paired	X	0 / 4 / 5 / 72	0 / 5 / 5 / 71	0 / 6 / 8 / 67
AF-multimer	Diagonal (Unpaired)	O	4 / 7 / 7 / 63	4 / 8 / 9 / 60	4 / 8 / 11 / 58
AF-multimer	Diagonal (Unpaired)	X	0 / 6 / 5 / 70	0 / 6 / 8 / 67	0 / 6 / 9 / 66

*High/Medium/Acceptable/Incorrect according to CAPRI criteria

Summary

- Antibody-antigen complexes is difficult to predict using antibody-antigen due to the **lack of coevolution information**
- The **PAE score** can be a criterion of whether the binding geometry is well predicted
- Other **ab-initio docking method** is necessary for accurate complex structure prediction

GalaxyTongDock 실습

Limitation of AlphaFold-Multimer

- AlphaFold-Multimer
  - Using Co-evolution **Information**

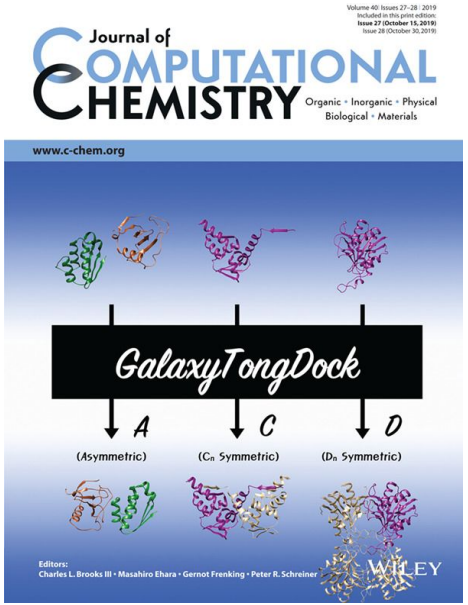
⬇

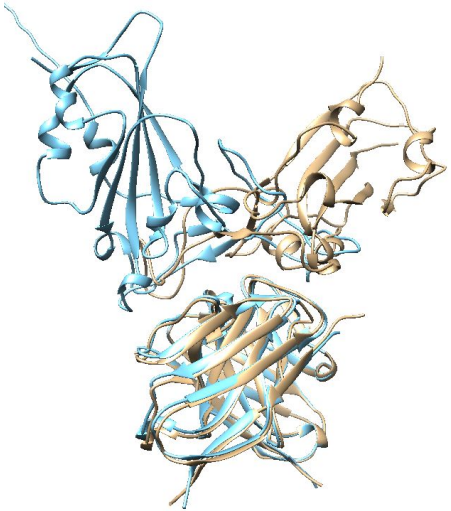
- **Physical Chemistry**

could be breakthrough

⬇

- ***ab-initio* Docking Tool**





GalaxyTongDock : What it is

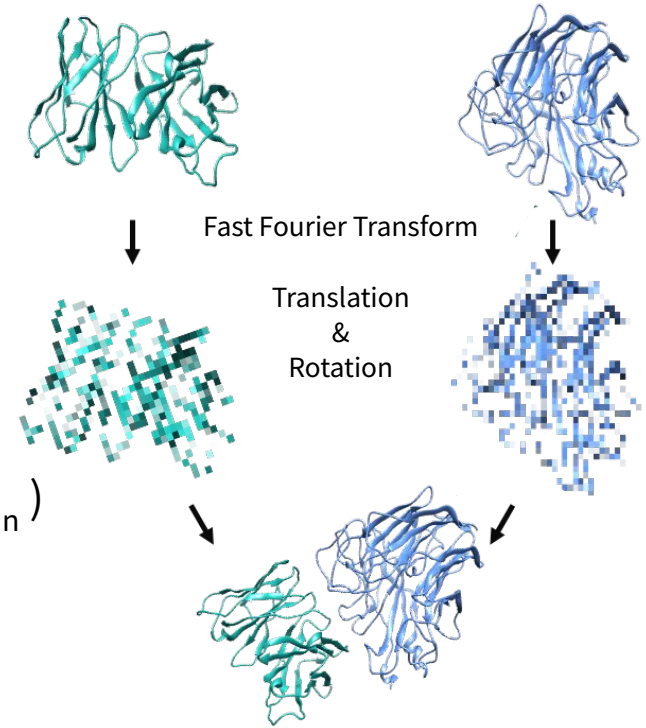
- ***ab-initio*** Docking tool based on 3D FFT

- $E_{GalaxyTongDock}$

  $= E_{SCrep} + w_1 E_{SCattr} + w_2 E_{elec} + w_3 E_{ACE} + w_4 E_{IACE} + w_5 E_{consv}$

- GalaxyTongDock_A(Asymmetric complex)
- GalaxyTongDock_C,D(Symmetric homomer : C_n, D_n)

- Ranking based on Energy and Cluster size



GalaxyTongDock : Web Site

- **Let's Check GalaxyWEB,** https://galaxy.seoklab.org/
  - How to use GalaxyTongDock_A
  - How to use GalaxyTongDock_C, GalaxyTongDock_D
  - Run GalaxyTongDock_A for previous example

- **Input Files : Public-practice/AI-Bio/AF_monomer/7bem_ab.pdb, 7bem_ag.pdb**
  - **AlphaFold2 Modeling of each Monomer**
  - **RCSB Database for proteins (**https://www.rcsb.org/**)**

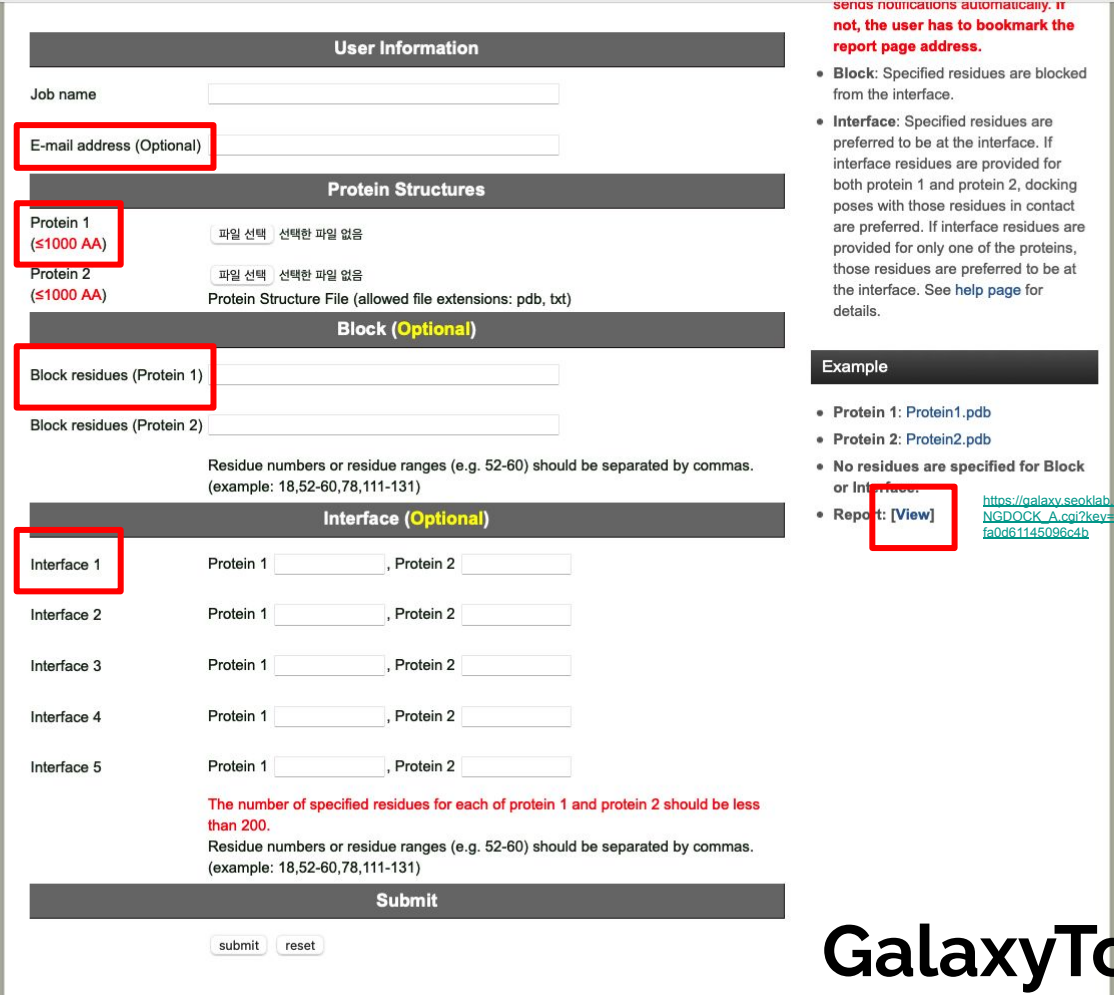
https://github.com/seoklab/Public-practice/tree/main/AI-Bio/AF_monomer

I Think Essential

Due to Memory

Penalty
when at interface

Advantage
when at interface



GalaxyTongDock_A

(subunit 1) and a copy of monomer (subunit 2) is performed, and C_n-symmetric structures are constructed from that. Use GalaxyTongDock_D for D_n symmetry.

User Information

Job name

E-mail address (Optional)

Protein Information

Monomer structure (≤1000 AA) — 파일 선택 선택한 파일 없음 — Protein Structure File (allowed file extensions: pdb, txt)

Oligomeric state (≤12) — Number of monomers. For example, oligomeric state of C_3 symmetry is 3.

Block (Optional)

Block residues

Residue numbers or residue ranges (e.g. 52-60) should be separated by commas.
(example: 18,52-60,78,111-131)

Interface (Optional)

Interface 1 Subunit 1 , Subunit 2

Interface 2 Subunit 1 , Subunit 2

Interface 3 Subunit 1 , Subunit 2

Interface 4 Subunit 1 , Subunit 2

Interface 5 Subunit 1 , Subunit 2

The number of specified residues for each of subunit 1 and subunit 2 should be less than 200.
Residue numbers or residue ranges (e.g. 52-60) should be separated by commas.
(example: 18,52-60,78,111-131)

Submit

submit reset

- E-mail: Average run time is 1h. If e-mail address is given, the server sends notifications automatically. **If not, the user has to bookmark the report page address.**
- **Block**: Specified residues are blocked from the interface.
- **Interface**: Specified residues are preferred to be at the interface. If interface residues are provided for both subunit 1 and subunit 2, docking poses with those residues in contact are preferred. If interface residues are provided for only one subunit, those residues are preferred to be at the interface. See help page for details.

Example

- **Monomer structure**: Monomer.pdb
- **Oligomeric state**: 3
- **No residues are specified for Block or Interface.**
- **Report: [View]** https://galaxy.seoklab.org/cgi-bin/report_TONGDOCK_C.cgi?key=4ad031fba91bf030959cfe762b04f4e0

Homomer

GalaxyTongDock_C

D_n Symmetry

Protein Information

Monomer structure (≤1000 AA): 파일 선택 | 선택한 파일 없음
Protein Structure File (allowed file extensions: pdb, txt)

Oligomeric state (≤12):
Number of monomers. For example, oligomeric state of D_2 symmetry is 4.

Block (Optional)

Block residues:

Residue numbers or residue ranges (e.g. 52-60) should be separated by commas.
(example: 18,52-60,78,111-131)

C-Interface (Optional)

C-Interface 1: subunit 1 , subunit 2
C-Interface 2: subunit 1 , subunit 2
C-Interface 3: subunit 1 , subunit 2
C-Interface 4: subunit 1 , subunit 2
C-Interface 5: subunit 1 , subunit 2

The number of specified residues for each of subunit 1 and subunit 2 should be less than 200.
Residue numbers or residue ranges (e.g. 52-60) should be separated by commas.
(example: 18,52-60,78,111-131)

D-Interface (Optional)

D-Interface 1: half 1 , half 2
D-Interface 2: half 1 , half 2
D-Interface 3: half 1 , half 2
D-Interface 4: half 1 , half 2
D-Interface 5: half 1 , half 2

The number of specified residues for each of half 1 and half 2 should be less than 200.
Residue numbers or residue ranges (e.g. 52-60) should be separated by commas.

are preferred to be at the interface. If interface residues are provided for both subunit 1 and subunit 2, docking poses with those residues in contact are preferred. If interface residues are provided for only one subunit, those residues are preferred to be at the interface.

- **D-Interface**: Specified resiudes in this option are used when structure with D_n symmetry is predicted. The residues are preferred to be at the interface. If interface residues are provided for both half 1 and half 2, docking poses with those residues in contact are preferred. If interface residues are provided for only one 'half', those residues are preferred to be at the interface. See help page for details.

Example

- **Monomer structure**: Monomer.pdb
- **Oligomeric state**: 6
- **No residues are specified for Block or Interface.**
- **Report: [View]** https://galaxy.seoklab.org/cgi-bin/report_TONGDOCK_D.cgi?key=a3e0fa4388cfd9ac8a58c07598d9baed

GalaxyTongDock_D

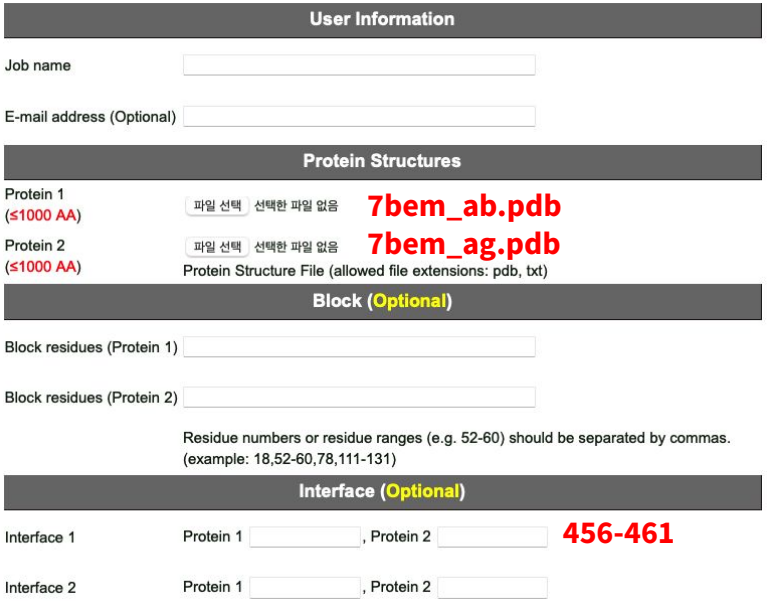
GalaxyTongDock : Run

- **Let's Run!! with GalaxyWeb,** https://galaxy.seoklab.org/
  - Target : 7BEM, fail case of the AlphaFold-Mutimer
  - With interface option & without interface option

- **Without Interface option**

- **With Interface option**

  → How to get Interface residue



GalaxyTongDock : How to Analysis

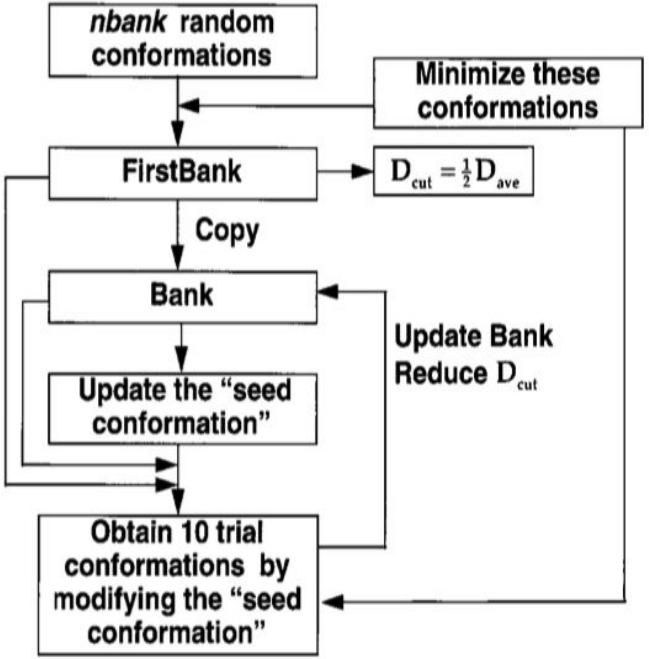
- **Result Page (1. w/o interface option 2. w/ interface option)**
  1. http://galaxy.seoklab.org/cgi-bin/report_TONGDOCK_A.cgi?key=c42ace079232d3e45ac0112e374ace5c
  2. http://galaxy.seoklab.org/cgi-bin/report_TONGDOCK_A.cgi?key=2f846263383b6b8ffe679a50bbc4082b

- **Analysis with Chimera Let's Do it!!**
  - **Comparison with Crystal Structure** or Template, Similar complex
  - **Check the interactions between side chains of each residues**

GalaxyDock 실습

GalaxyDock : What it is

- Protein-Ligand Docking tool based on CSA
  - CSA : Genetic algorithm + Simulated Annealing
- Based on GalaxyDock BP2 Score
  - Hybrid of physics-based, empirical, knowledge-based
- 6+K Degrees of freedom of Ligand
  - Fixed Ring conformation, Rigid Receptor Protein
  - Degrees of freedom
    - Translation + Rotation + Torsion angle



GalaxyDock : Preparing input

- Receptor : Crystal structure or Predicted structure by AF2 etc. → .pdb format
- **Ligand**
  - Prepare SMILES of the target molecule

```
# lig.smi
c1(O)cc(cc2OC(C)(C)[C@H]3[C@H](c12)CC(=CC3)C)C(CCCCCCBr)(C)C
```

  - Convert SMILES to .mol2 using Corina (https://mn-am.com/products/corina/) : 1D→3D

```
corina -i t=smiles lig.smi -o t=mol2 >> lig.mol2
```

  - Preprocess mol2 file : attach hydrogen and assign partial charge with Chimera command

```
del H                  # delete Hydrogen        addh          # attach Hydrogen
addcharge all method gas                                     # assign partial charge
```

GalaxyDock : Run & Analysis

- **Let's Run!! with GalaxyWeb,** https://galaxy.seoklab.org/
  - Input : AF.pdb, lig.mol2, binding site = 75A,94A,386A,406A

- **Result Page**
  1. https://galaxy.seoklab.org/cgi-bin/report_DOCK.cgi?key=bef4b9a3e921020003bfd1742e5a7be8

- **Analysis with Chimera Let's Do it!!**
  - **Comparison with Crystal Structure** or Template, Similar complex
  - **Check the interactions between side chains of each residues and molecule**

Thank you for listening