Class Today

- Lecture 12: protein modeling

- Announcement
  1. Final exam: Open-book exam, **Tuesday (June 4th) at 2 pm (Beijing Time)**
     - 3 hours
     - in a campus computer room, but personal electronic devices are NOT allowed.
     - You are allowed to search the internet for information but not to copy text directly. Write the answers ONLY on the booklet provided in your own words and always acknowledge the source(s) from which your answers are derived.

- Assignment:
  1. Coursework 2 is due on Tuesday (May 7th) at 5 pm

LECTURE 12: PROTEIN MODELING

Can we predict protein structure from sequence?

- If there is a high degree of sequence identity between two proteins, their overall folds will be similar.
- There are many cases of two proteins having virtually identical overall folds and closely related functions despite having no statistically significant degree of sequence identity/similarity.

- The known protein structures and canonical protein folds are used to derive structure from sequence by different approaches.
  - Homology modeling
  - Threading
  - Fragment libraries based
  - *Deep learning based*
  - …

Homology Modeling of Proteins

Definition:

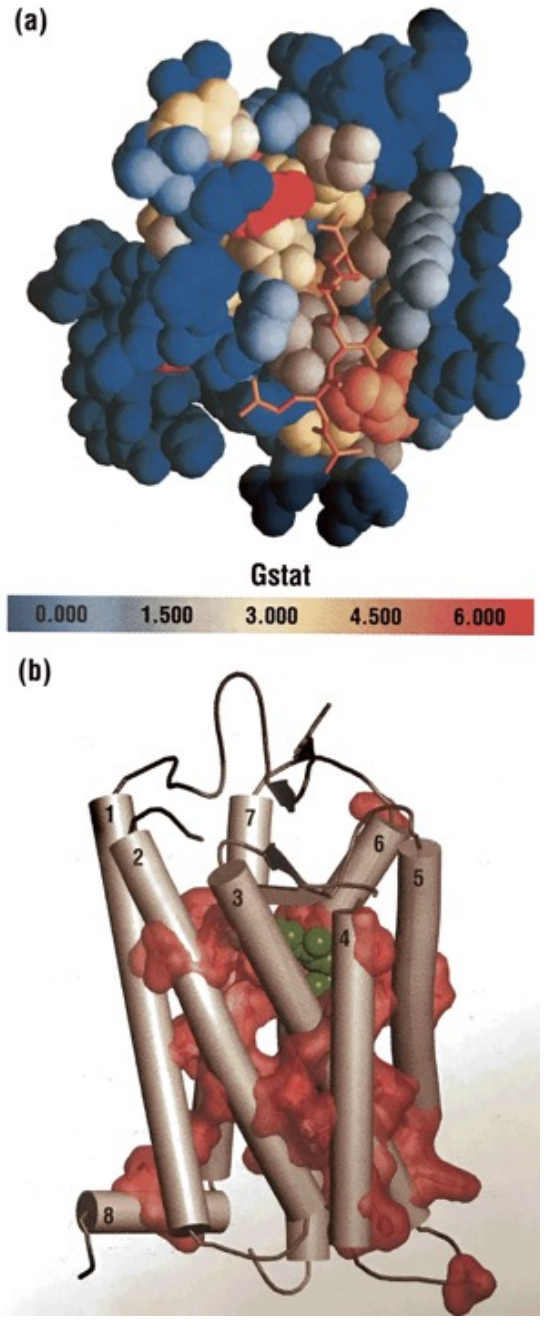
- A computational method for modeling the structure of a protein based on its sequence similarity to one or more other proteins of known structure.

Why a Model:

- A Model is desirable when X-ray crystallography, NMR spectroscopy or Cryo-EM cannot determine the structure of a protein in time or at all. The built model provides a wealth of information of how the protein functions with information at residue property level. This information can then be used for mutational studies or for drug design.

- ***Note***: homology models can not (*not reliable*) be used to study conformational changes induced by ligand (proteins or small molecules) binding, pH changes, or post-translational modification, or the structural consequences of sequence insertions and deletions.

- Evolutionary data for a protein family can be used to measure statistical interactions between amino acid positions. It is based on two empirical observations:
  1. The functional coupling of two positions, even if distantly located in the structure, should mutually constrain evolution at the two positions, and this should be represented in the statistical coupling of the underlying amino-acid distributions in the multiple sequence alignment, which can then be mapped onto the protein.
  2. A lack of evolutionary constraints at one position should cause the distribution of observed amino acids at that position in the multiple sequence alignment to approach their mean abundance in all proteins, and deviances from the mean values should quantitatively represent conservation.



Gregory A Petsko, Dagmar Ringe. Protein structure and function. New Science Press Ltd 2004.

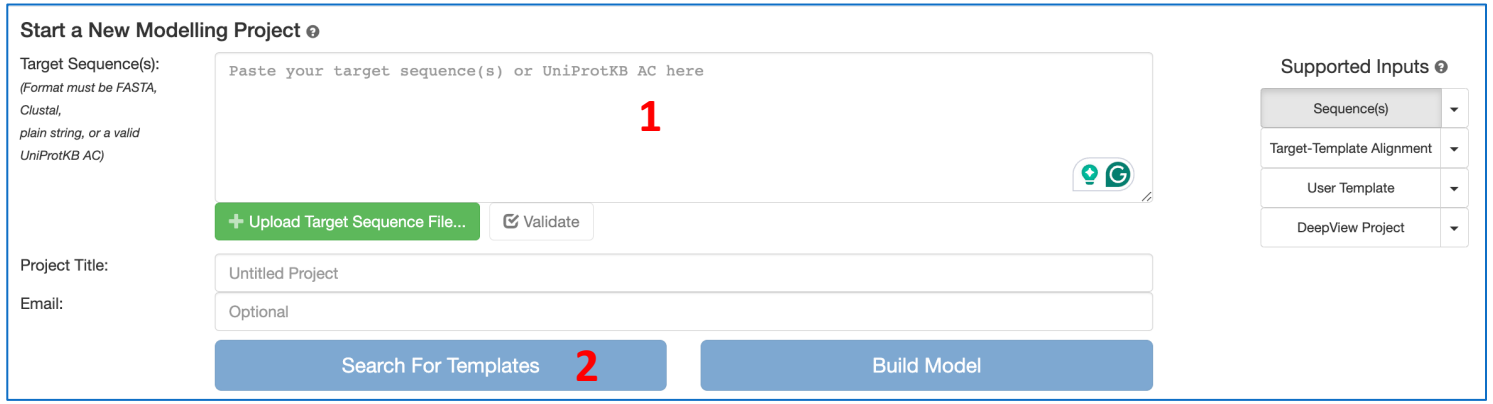
Steps in Homology Modeling

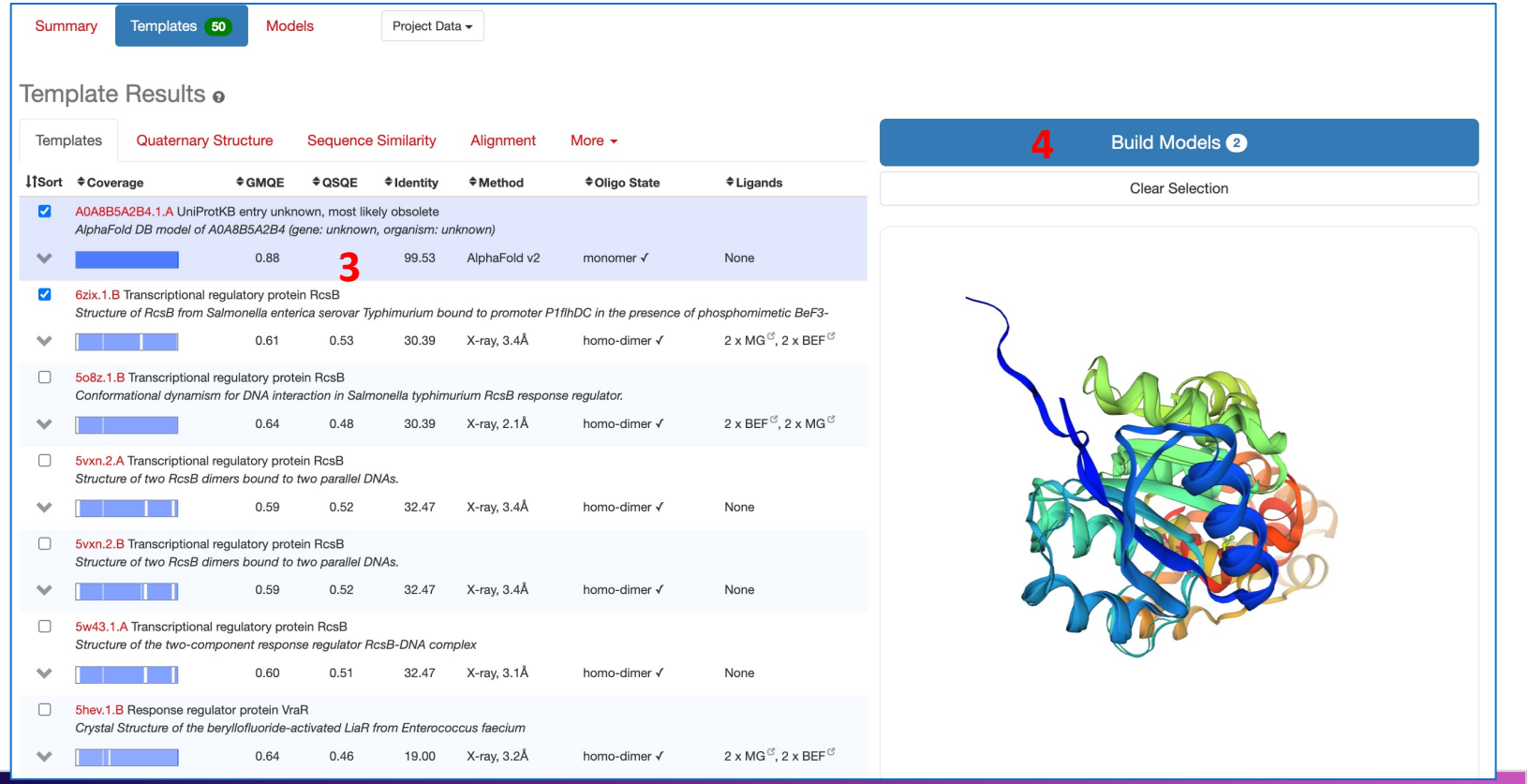
- Template selection
- Sequence alignment
- Backbone model building
- Loop modeling and side chain refinement
- Model refinement using energy function (Energy minimization)
- Model Evaluation

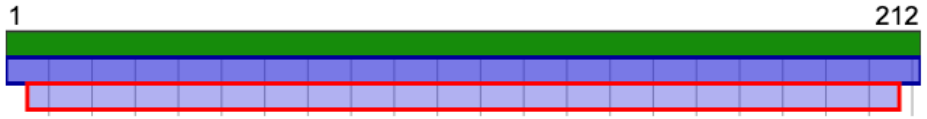
SWISS-MODEL

- https://swissmodel.expasy.org/interactive

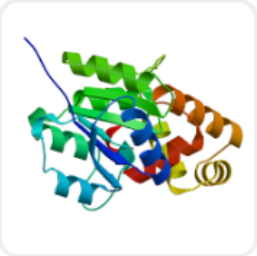
- Steps:
  1. Input primary sequence
  2. Template search
  3. Template selection
  4. Model Building
  5. Model quality estimation







Model 01



Structure Assessment

Compare ☐

Download files ▾

Oligo-State
Monomer

5

GMQE
0.88

Template

A0A8B5A2B4.1.A UniProtKB entry unknown, most likely obsolete

AlphaFold DB model of A0A8B5A2B4 (gene: unknown, organism: unknown)

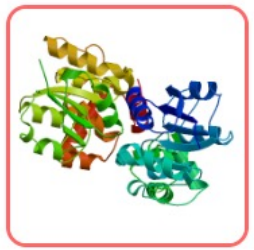
Seq Identity
99.53%

Coverage

Biounit Oligo State	Monomer
QSQE	-
Method	AlphaFold v2
Seq Similarity	0.60
Coverage	1.00
Range	1-212

Model-Template Alignment

Model 02



Structure Assessment

Compare ☐

Download files ▾

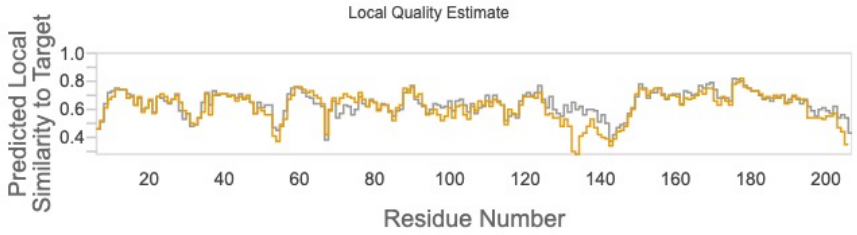
Oligo-State
Homo-dimer
(matching prediction)

GMQE
0.61

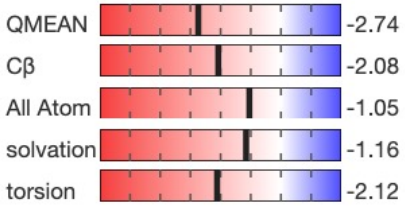
5

QMEANDisCo Global:
0.63 ± 0.05

QMEANDisCo Local



QMEAN Z-Scores



Template

6zix.1.A Transcriptional regulatory protein RcsB

Structure of RcsB from Salmonella enterica serovar Typhimurium bound to promoter P1flhDC in the presence of phosphomimetic BeF3-

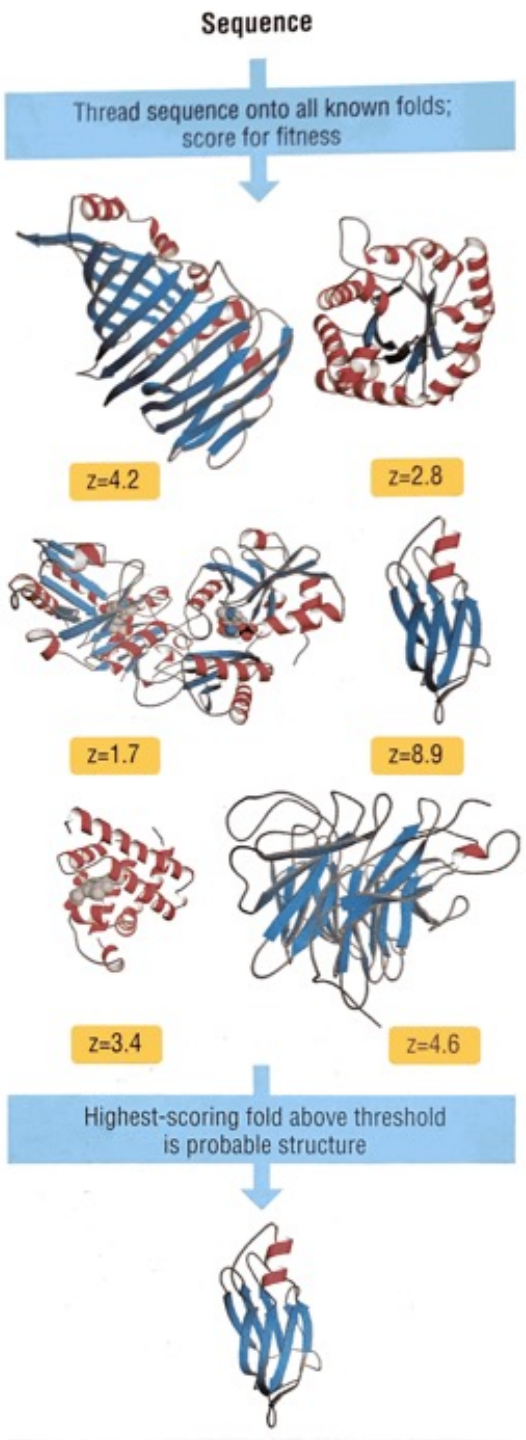
Seq Identity
30.39%

Coverage

Model-Template Alignment

Profile-Based Threading

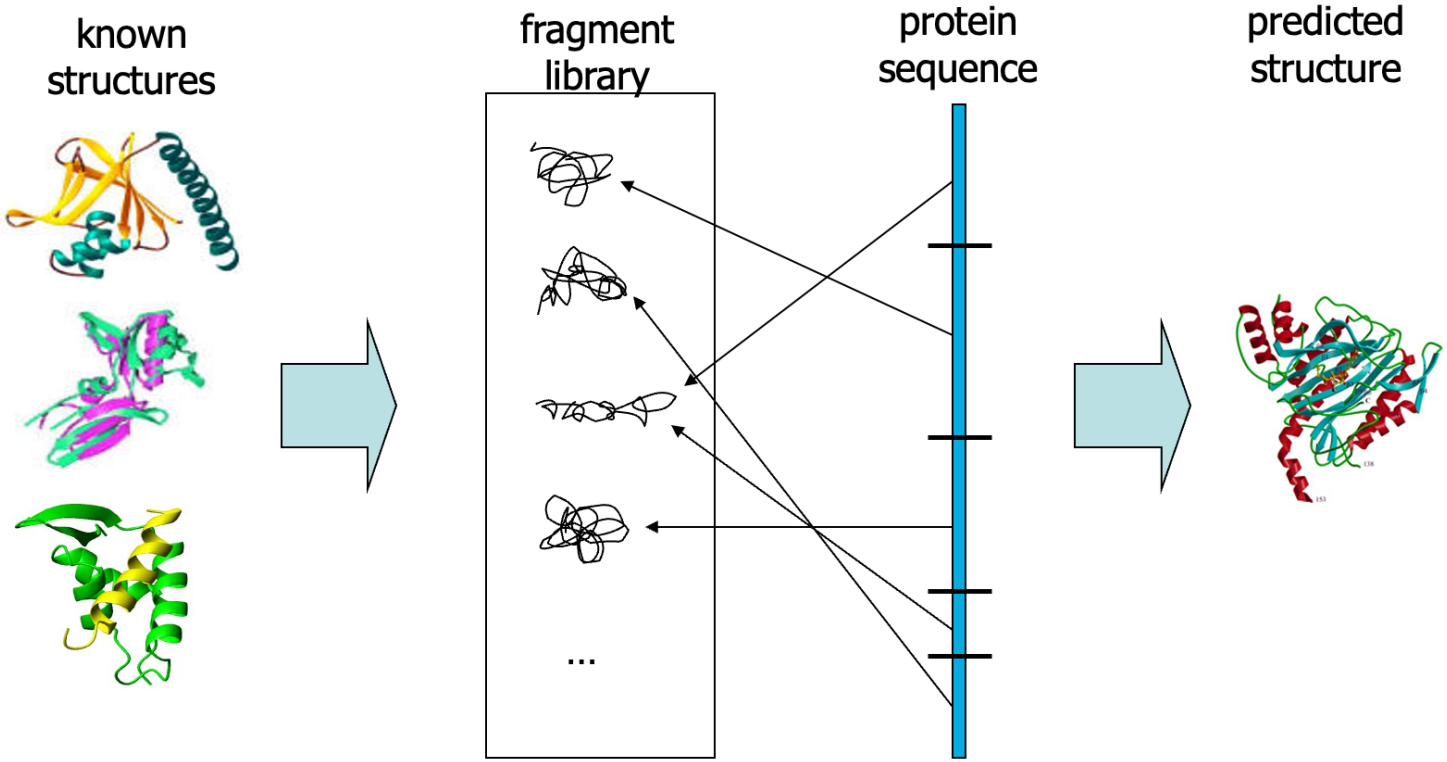
- Threading uses the structure to compute energy function during alignment
- predicts the structure of a sequence even if no sequence homologs are known
- In this method, a computer program forces the sequence to adopt every known protein fold in turn, and in each case a scoring function is calculated that measures the suitability of the sequence for that particular fold.
  - A high score of z value indicates high possibility that the sequence adopts this fold.



Fragment Libraries for Modeling

- Fragment libraries for short segments are extracted from the protein structure database.

- The conformational space defined by these fragments is then searched with an energy function that favors compact structures with paired strands and buried hydrophobic residues.
  - ~1000 independent simulations are carried out for each query sequence, and the resulting structures are clustered.



Some Modeling Programs

Modeller
(https://salilab.org/modeller/)

Swiss-Model
(https://swissmodel.expasy.org/)

RoseTTAFold Server (registration needed)

(https://robetta.bakerlab.org/submit.php)

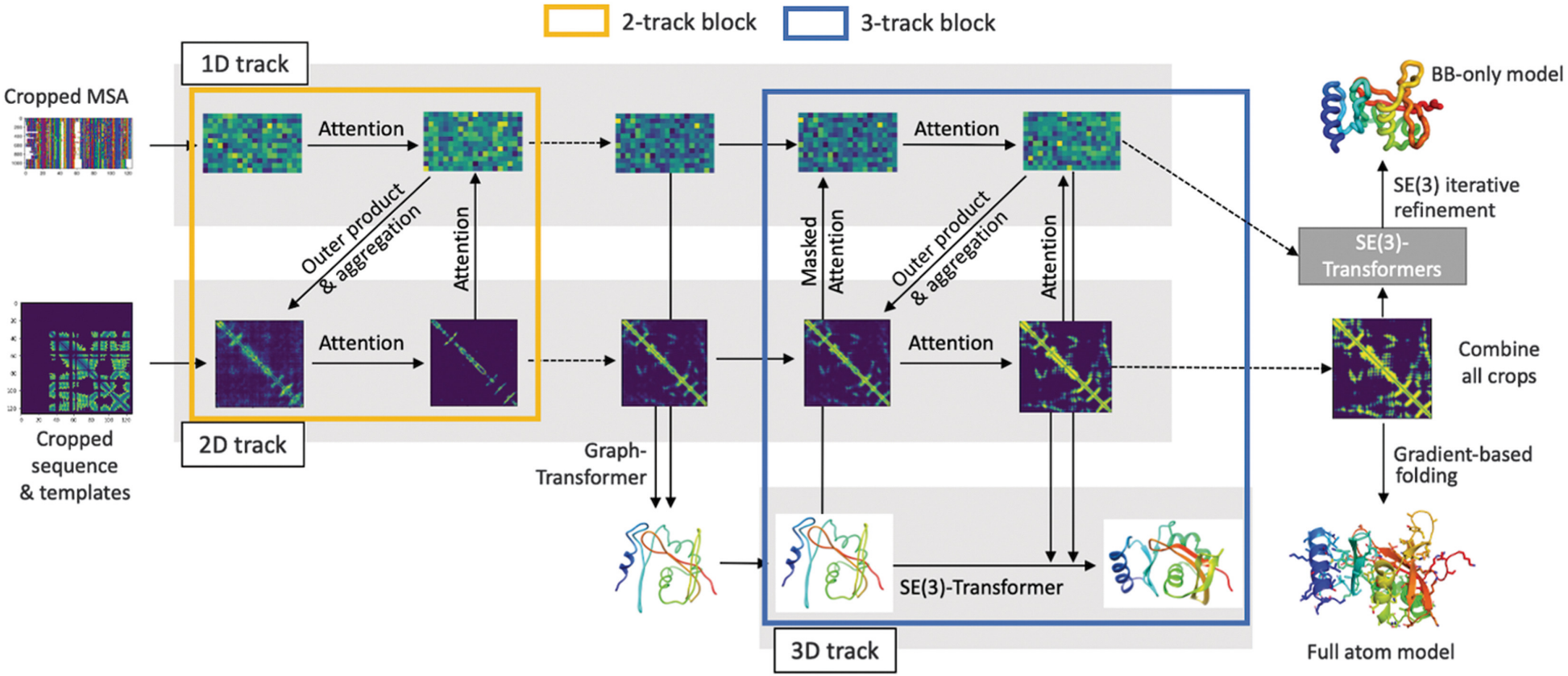
AlphaFold2
(http://github.com/sokrypton/colabFold)

database(https://alphafold.ebi.ac.uk/)

RoseTTA Fold

- RoseTTAFold is a “three-track” neural network.

- 1D: sequences
- 2D: distances
- 3D: coordinates



Google

rosettafold server

All Images Videos Shopping News More Tools

Calculator Free Github

About 21,100 results (0.29 seconds)

Robetta
https://robetta.bakerlab.org

Robetta - Baker Lab

Features include relatively fast and accurate deep learning based methods, **RoseTTAFold** a
TrRosetta, and an interactive submission interface that allows ...

Login · Frequently Asked Questions · Structure Prediction Queue · Register

Robetta Project Structure Prediction

Submit a job for structure prediction

Please do not submit jobs under different user accounts. Such jobs will be removed.

Required

Target Name

Protein sequence

or upload FASTA Choose File No file chosen

Optional

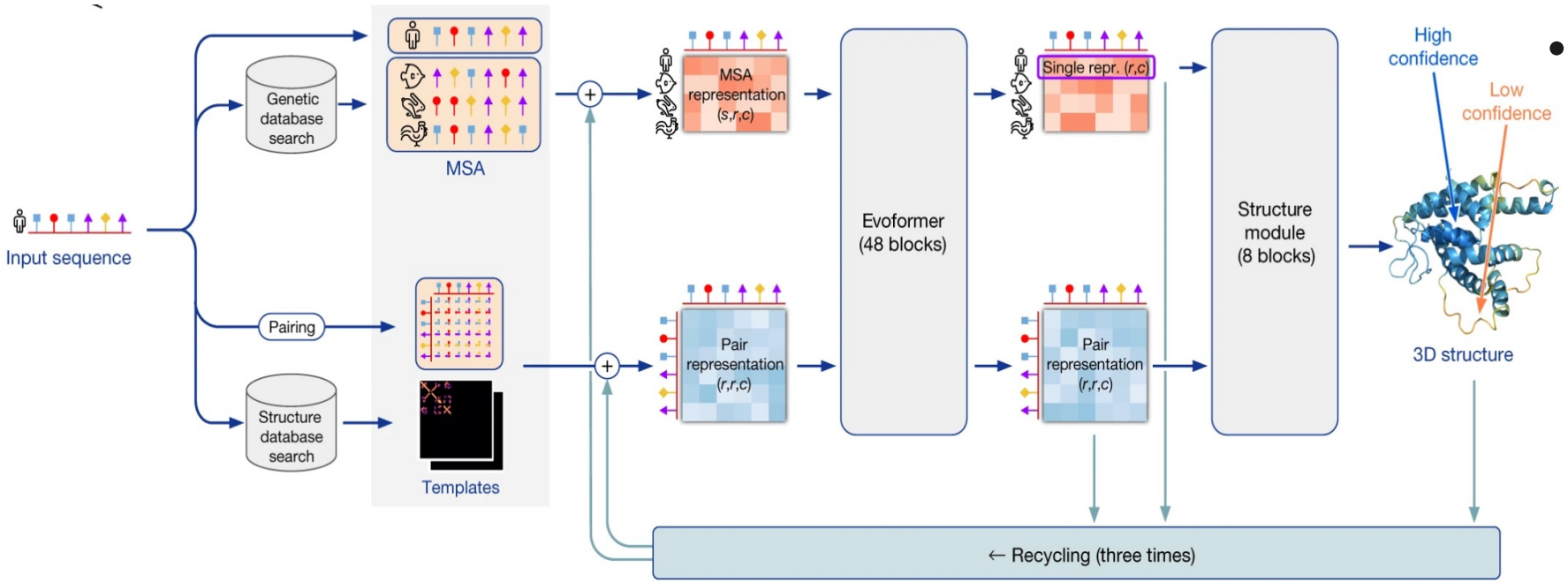
RoseTTAFold ☑ CM ☐ AB ☐ Predict domains ☐

Upload MSA Choose File No file chosen

Submit 3 + 2 = Keep private ☐

AlphaFold2

- AlphaFold2 is a multicomponent artificial intelligence (AI) system that uses machine learning to predict a protein's 3D structure based on its primary amino acid sequence.



- AlphaFold is **NOT** a homology modelling tool: it can successfully operate without using any template structures and even predict previously unknown protein folds.

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Nikolov, S., Jain, R., Adler, J., Back, T., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

2. Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, *13*(1), 1-11. https://doi.org/10.1038/s41467-022-28865-w

Four Ways of Using AlphaFold 2

1. use the AlphaFold database (https://alphafold.ebi.ac.uk)
2. use homology search.
   - If your protein of interest is not listed in UniProt, you can use the homology search functions provided by EMBL-EBI at https://www.ebi.ac.uk/Tools/sss/fasta/.
3. use AlphaFold Colab or ColabFold (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=G4yBrceuFbf3)
   - Accuracy is not as high as that of the original AlphaFold 2
4. install and run it yourself. You can choose either the Docker version or the non-Docker setup.
   - Adjust parameters
   - But it is costly and requires knowledge

AlphaFold2_advanced.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive

Share

RAM Disk

AlphaFold2_advanced

This notebook modifies deepmind's original notebook (**before AlphaFold-Multimer existed**) to add experimental support for modeling complexes (both homo and hetero-oligomers), option to run MMseqs2 instead of Jackhmmer for MSA generation and advanced functionality.

See ColabFold for other related notebooks

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

Limitations

- This notebook does **NOT** use Templates.
- This notebook does **NOT** use AlphaFold-Multimer for complex (protein-protein) modeling.
- For a typical Google-Colab session, with a `16G-GPU`, the max total length is **1400 residues**. Sometimes a `12G-GPU` is assigned, in which the max length is ~1000 residues.
- Can I use the models for **Molecular Replacement**? Yes, but be CAREFUL, the bfactor column is populated with pLDDT confidence values (higher = better). Phenix.phaser expects a "real" bfactor, where (lower = better). See post from Claudia Millán on how to process models.

Install software

Please execute this cell by pressing the *Play* button on the left.

Show code

Enter the amino acid sequence to fold

sequence: PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

jobname: test

homooligomer: 1

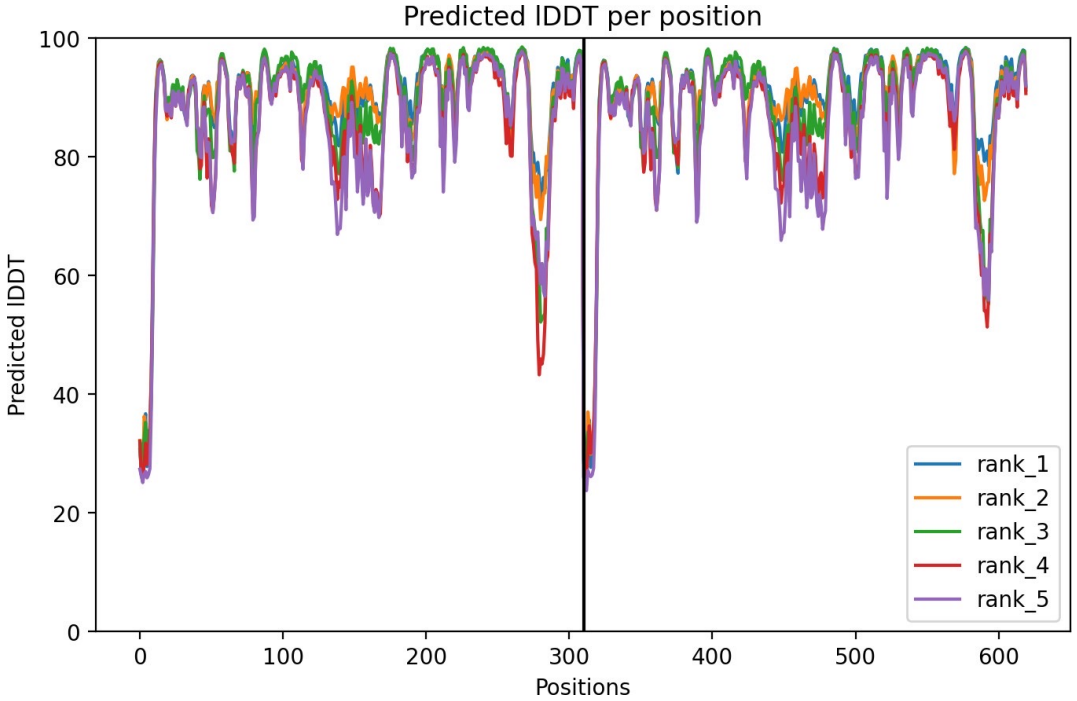
- `sequence` Specify protein sequence to be modelled.
  - Use `/` to specify intra-protein chainbreaks (for trimming regions within protein).
  - Use `:` to specify inter-protein chainbreaks (for modeling protein-protein hetero-complexes).
  - For example, sequence `AC/DE:FGH` will be modelled as polypeptides: `AC`, `DE` and `FGH`. A seperate MSA will be generates for `ACDE` and `FGH`. If `pair_msa` is enabled, `ACDE`'s MSA will be paired with `FGH`'s MSA.
- `homooligomer` Define number of copies in a homo-oligomeric assembly.
  - Use `:` to specify different homooligomeric state (copy numer) for each component of the complex.
  - For example, **sequence:** `ABC:DEF`, **homooligomer:** `2:1`, the first protein `ABC` will be modeled as a homodimer (2 copies) and second `DEF` a monomer (1 copy).

Show code

AlphaFold2

- Ranked by either average pLDDT (predicted local distance difference test), pTM (multimer)
  - Considers local environment < 4 Å around Cα
  - Higher pLDDT/pTM is better

- pLDDT scoring
  - < 50 Disordered/bad quality
  - 50-70 low quality
  - 70-90 Backbone probably correct
  - >90 High quality



- AlphaFold2 can be used for:

1. Structure prediction of novel proteins
2. Models for CryoEM and Crystallography
3. Guidance for protein construct design
4. Investigating intrinsically disordered proteins
5. Predict oligomer state/complex structure
6. Predict alternative protein conformation
7. Predict effects of mutation

etc…

Akdel, M., Pires, D. E., Pardo, E. P., Jänes, J., Zalevsky, A. O., Mészáros, B., Bryant, P., Good, L. L., Laskowski, R. A., Pozzati, G., Shenoy, A., Zhu, W., Kundrotas, P., Serra, V. R., Rodrigues, C. H., Dunham, A. S., Burke, D., Borkakoti, N., Velankar, S., . . . Beltrao, P. (2022). A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, *29*(11), 1056-1067. https://doi.org/10.1038/s41594-022-00849-w

Confirming Catalytic Residues/Binding Sites

- Active site residues in a structure can sometimes be recognized computationally by their geometry
- Site-directed mutagenesis can identify residues involved in binding or catalysis

- Docking programs(e.g., MOE) model the binding of ligands
  - Each ligand is divided into a small set of rigid fragments that are docked separately into the binding site, allowing a degree of flexibility at the positions that join them.
  - Can be used to find possible new compounds for drug development

  - ***Note***: This method cannot take unknown conformational changes of the protein into account. So, it could only be used to find candidates.

AlphaFold2 predicts	AlphaFold2 struggles to predict	AlphaFold2 doesn't predict
• Single protein chains • Protein multimers • Multisubunit protein-protein complexes	• Multiple conformations for the same sequence • Effects of point mutations • Antigen-antibody interactions	• Protein-DNA and protein-RNA complexes • Nucleic acid structure • Ligand and ion binding • Post-translational modifications • Membrane plane for transmembrane domains

https://www.ebi.ac.uk/training/online/courses/alphafold/an-introductory-guide-to-its-strengths-and-limitations/strengths-and-limitations-of-alphafold/

Optional Assignment but recommended

- Try to predict the protein structure with RossettaFold and AlphaFold. The AA sequence is shown below:

HHHHHHGSLQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRR
LMEAFAKRQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQIGGMA
SEARGGLGAPPLQSARSLPGPAPCLKHFPLDLRTSMDGKCKEIAEELFTRSLAESE
LRSAPYEFPEESPIEQLEERRQRLERQISQDVKLEPDILLRAKQDFLKTDSDSDLQL
YKEQGEGQGDRSLRERDVLEREFQRVTISGEEKCGVPFTDLLDAAKSVVRALFIR
EKYMALSLQSFCPTTRRYLQQLAEKPLETRTYEQGPDTPVSADAPVHPPALEQH
PYEHCEPSTMPGDLGLGLRMVRGVVHVYTRREPDEHCSEVELPYPDLQEFVAD
VNVLMALIINGPIKSFCYRRLQY