Lab 7. Protein Structure data analysis

1. Objective

- Learn how to predict protein structure based on the protein sequence-structure relationship
- Learn what the homology modeling is.
- Learn how to find homologous sequence.
- Learn how to predict protein structure using homology modeling
- Learn how to evaluate the predicted protein model.

2. Background

2.1) Importance of protein structure prediction

The structure of an ordered protein is essential for the understanding of its function. Protein structures are critical for understanding the mechanisms of biological systems and, subsequently, for drug and vaccine design. Even though the number of experimental available proteins is exponentially increasing over the last years in the post-genomic era, where many new complete genomes are available every year and the number of sequences total in the millions, it is impossible to rely on experimental methods alone for structural characterization.

This gap can be partially filled by using computational protein structure prediction. Fold prediction is an important tool for the functional annotation of proteins at the genomic scale. Fold and structure predictions can be used to infer binding interfaces, potential binding partners, and catalytic active sites. *in silico* drug screening can be performed on close homologues of proteins with known structures

2.2) Relationship of protein sequence and structure

Proteins sharing similar sequence identity are likely to share similar structures (Figure 1). Here, the structural similarity is quantitatively evaluated with RMSD (Root mean square deviation) that is the distance difference between the atoms of two aligned protein structures. Protein pairs with a sequence identity higher than 25% are very likely to be structurally similar. The structure of a sequence identity

lesser than 25% (twilight zone) might be similar but can also be different. RMSD in 100% sequence identity can be higher than 0 Å because of experimental errors.

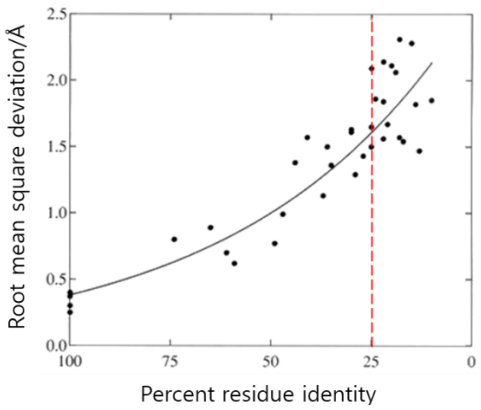


Figure 1. Sequence identity vs structural similarity [1].

2.3) Homology modeling (Template based modeling)

2.3.1) Homology modeling

Homology modeling is a method to predict the tertiary structure of an unknown protein using known 3D structures of template protein with the similar sequence (i.e. homologous sequence). The homology modeling is based on two assumptions: 1) the unknown protein shares some structural features with the homologous protein, and 2) the structures of homologous proteins is already known from an experimental measures such as by X-ray crystallography or NMR structures of homologous proteins.

The procedure of homology modeling is the following:

① Search databases for homologous protein sequences.

② Align homologous protein sequence(s) with the target, sequence of interest.
(tools: BLOSUM60, BLAST, and so on)

③ Build a model of the structure of the protein of interest using the known.
(tools: SWISS-MODEL, MODELLER, and so on)

④ Evaluate and refine model structure.
(tools: SWISS-MODEL, MODELLER, and so on)

The template structure should have a high sequence similarity with the target. If the template structure has low sequence identity with the target sequence, it could lead an inaccurate modeling. Be careful when selecting the template sequence before building a model. Usually, it is acceptable when

the structure has a sequence identity higher than 25% to the target. In addition, template sequence should cover major part of the target sequence. Even if target sequence and template sequence have high sequence identity, result may not be good if they share only small part of the target sequence. Also the quality of the predicted protein structure is influenced by the structural resolution of the template structures. The structures obtained by X-ray crystallography or NMR having a high resolution can provide more accurate structural information to homology modeling.

2.3.2) Principle of homology modeling [2]

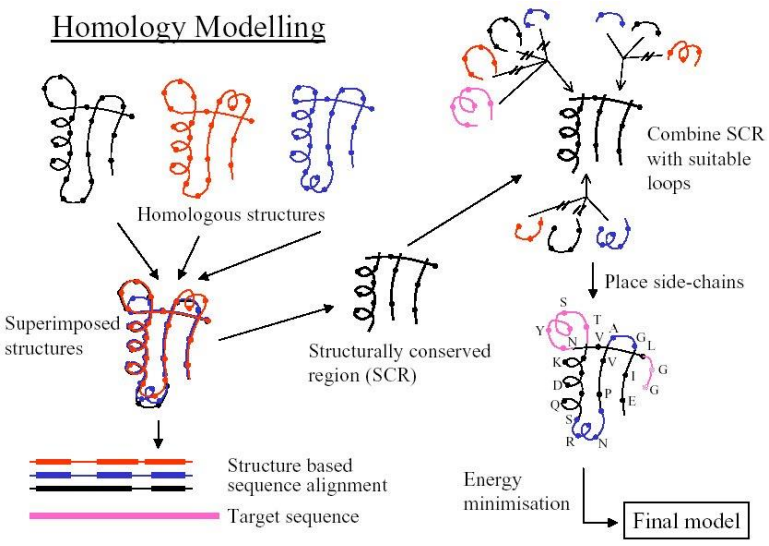


Figure 2. Homology modeling framework

The main idea of the homology modeling is that experimental 3D-structures of known homologous sequences (templates) is utilized to calculate a model for a new sequence (target). The SWISS-MODEL, which will be used in this lab, is based on homology modeling. The procedure for homology modeling is like below:

First, find the template structures that have a high sequence similarity using sequence alignment methods.

Second, identify structurally conserved core regions. Here, the conserved regions are the fragments of conserved sequences between the target and the template sequence. The conserved regions are easily observed by a sequence alignment.

Third, use core structures of the template to build a core structure of the target and leave non-conserved regions (loop regions) for later.

Fourth, combine the core structures with missing loops using an algorithm such as “spare part” to find compatible fragments in a Loop-Database or “ab-initio” rebuilding like Monte Carlo.

Fifth, for side chain modeling, find the most probable side chain conformation, using template structure information, rotamer libraries and energetic and packing criteria. Only a small fraction of all possible side chain conformations is observed in experimental structures. Also, rotamer libraries provide an ensemble of likely conformations and the propensity of rotamers depends on the backbone geometry.

Lastly, fine-tune the built model using an energy minimization method and spatial restraints of neighboring residues. For details of this stage, please refer the below link.

https://swissmodel.expasy.org/static/course/files/PartII_homology_modelling.pdf

3. Prelab activities

3.1) Overall description about the protein structure prediction. Answer the following questions.

- Why do we want to predict the structure of proteins?
- Why sequence alignment can be used in the structure prediction?

3.2) Identification of the most appropriate template.

3.2.1) Protein sequence alignment

- Make protein sequence aligner using smith-waterman algorithm (that was discussed in Lab 6). You already made DNA aligner in Lab 6. Use BLOSUM62.
- Test your protein sequence aligner using “ProteinAligner_testDB.”
- The aligner should take the query file and the template files as input arguments and be able to calculate the alignment score (It is related with the task in mainlab 4.1, see Figure 3).

3.2.2) Browse most similar template

- Based on pair-wise protein aligner, how can you find the most appropriate template? Propose your approach stepwise.

3.2.3) Submit zip file which containing

- Prelab report

- Source code of the protein-pairwise aligner

3.3) Understanding how SWISS-MODEL works.

- Read SWISS-MODEL manual.
  manual site: https://swissmodel.expasy.org/docs/help
  tutorial site: https://swissmodel.expasy.org/docs/tutorial
  information about SWISS-MODEL: https://swissmodel.expasy.org/course

3.4) Installation of Pymol

- Install PyMol (https://www.pymol.org/) to visualize predicted structure. You can download free educational license file (https://pymol.org/edu/?q=educational/) here.

4. Mainlab activities

- Perform homology modeling by utilizing two different ways for sequence alignment and the selection of template protein: 1) Use previously designed protein-pairwise aligner to find proper template for homology modeling from custom database. 2)Use BLAST search to find proper template for homology modeling from PDB database.
- Model protein structures from the template structures with SWISS-MODEL.
- Visualize protein structure results with PyMol.
- Learn criteria evaluating the result of protein modeling and assess modeled structures with the criteria.

4.1) Find template structures for homology modeling.

Suppose we found a new corona virus protein, “protein A”. We have the sequence of protein A but no structures. We want to know the structure of protein A thus we will predict the structure using homology modeling.

4.1.1) Find the template structure using your protein sequence aligner and get its structure.

First, for homology modeling, we need to find the template structure of protein A. The protein sequence of protein A is in the "Query.fa" file. Find appropriate template structures among the "customDB".

① Find the sequence with high similarity using your protein sequence aligner among the customDB. For example, the result should be presented as below. ("Rand", "File name" and "corresponding alignment score.") You should sort the results by the alignment scores and then show the best 10 alignment-scored templates with its score.

```
0 template_124 score: 868
1 template_154 score: 655
2 template_186 score: 578
3 template_055 score: 563
4 template_163 score: 382
5 template_120 score: 324
6 template_084 score: 190
7 template_101 score: 132
8 template_182 score: 124
9 template_181 score: 64
```

Figure 3. The example result of an alignment of a query sequence against the customDB by your protein sequence aligner

Q. Which one is the most desirable sequence for the template structure and why?

② We need the structure information of the template. Usually, the structure information is described in "PDB" file format. To find out the PDB ID of your template, open the template file by text editor. For example, PDB ID of the template_001 is "2gsp".

③ To get the structure information of the template, go to RCSB PDB (https://www.rcsb.org/). RCSB PDB site provides PDB files of proteins having PDB ID. At RCSB PDB, search PDB ID of your template.

④ Download the structure files of your template as "PDB format".

4.1.2) Find the template structure with high sequence similarity using BLAST.

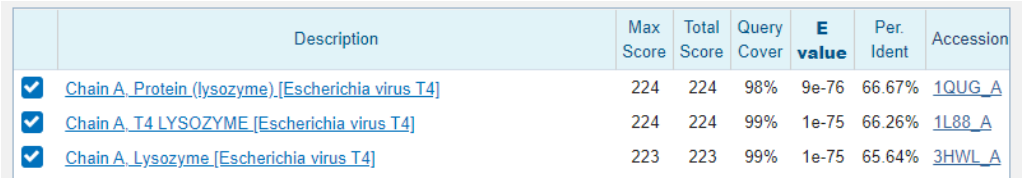
Now, we will use BLAST to find the template structure for the protein A. BLAST (Basic Local Alignment Search Tool) is a tool to find regions of similarity between biological sequences. We will run BLAST against the PDB database that contains known structures to find the template structure.

① Go to BLAST web page (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and click "protein BLAST".

② Insert the sequence of the protein A to the input field “Enter Query Sequence”.

③ **At “Choose Search Set”, click the “Database” field and change it to “Protein Data Bank proteins(PDB)”**. This means that we will run the BLAST against only the protein sequence with a known structure.

④ Click “BLAST” button and wait the result. The example will be like below.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Chain A, Protein (lysozyme) [Escherichia virus T4]	224	224	98%	9e-76	66.67%	1QUG_A
☑	Chain A, T4 LYSOZYME [Escherichia virus T4]	224	224	99%	1e-75	66.26%	1L88_A
☑	Chain A, Lysozyme [Escherichia virus T4]	223	223	99%	1e-75	65.64%	3HWL_A

Figure 4. The example of a result of BLAST

⑤ Find the most desirable template for modeling the structure of protein A among the candidates. Compare the sequences by max score, total score, query cover, e-value, and percent identity. Also, find out what each of criteria means.

Q. Which ones are the most desirable sequences for the templates and why? Explain your own criteria for selecting the templates.

⑥ To get the structure of template you selected, go to RCSB PDB (https://www.rcsb.org/).

⑦ At RCSB PDB, search PDB IDs of your templates. The PDB ID is described at the “Accession” tab in the BLAST result. For example, “1QUG” is the PDB ID (the letter “A” after the PDB ID is a chain id).

⑧ Download the structure files of your template as “PDB format”.

4.2) Predict the protein structure using SWISS-MODEL

Next, we will predict the structure of the protein A using SWISS-MODEL. SWISS-MODEL is a protein structure modeler based on homology modelling. [3]

① Go to SWISS-MODEL (https://swissmodel.expasy.org/) and click the “Start Modelling” button

② In this experiment, we have our own template found from the customDB (4.1.1) or BLAST

(4.1.2). Thus click the “User Template” at the right panel.

③ Insert the sequence of the protein A at “Target sequence(s)”.

④ Click the “Add Template File…” button and upload the template PDB file.

⑤ Set the “Project Title” as your student ID and name (ex. 2020####_JohnWick)

⑥ Click the “Build Model” button to build a model of the protein A. It takes minutes to hours to build the model.

⑦ **Attach the result page URL to the mainlab report.**

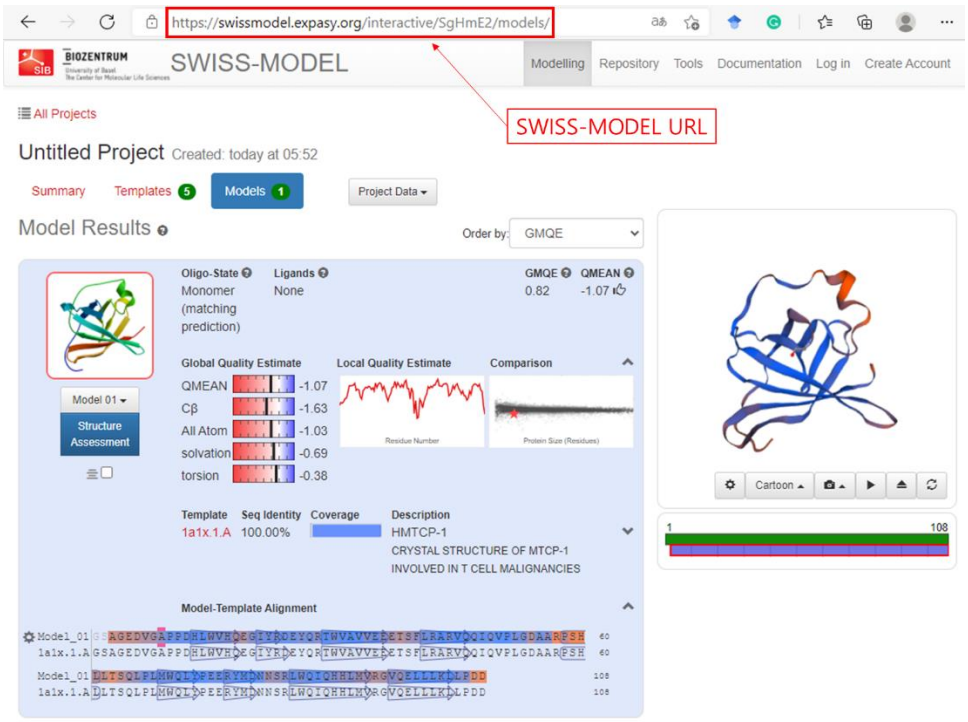


Figure 5. The example of a result page of SWISS-MODEL

4.3) Evaluate the predicted model

Finally, we need to evaluate whether the built model is reliable or not.

① Check the scores of the built model of the protein A at the “Model Results” page such as GMQE, QMEAN, Local Quality Estimate, Comparison, and Model-Template Alignment. You can read the descriptions of each item at “SWISS-MODEL help” page and it is in the top menu “Documentation”.

Q. How we can evaluate the reliability of the predicted model?

Q. Is your model good or not? Which parts are reliable and which parts are not? Describe your reasoning.

Q. Does your model have missing parts of structure? If does, why are there missing parts? How can we reduce the missing parts?

② For further analysis, click the "Structure Assessment" button and see the Ramachandran plot.

Q. What is the Ramachandran plot and how can you evaluate your model using Ramachandran plot? Evaluate your model using Ramachandran plot.

Q. Check the other analysis results and evaluate your model.

③ Now, compare your model with the template you used. Download your model as a PDB format.

④ Open PyMol and load two structures, your model and the template.

⑤ To make comparisons easier, use "align" command to two structures in PyMol.

Q. Compare the structures of your model and the template. Do they look similar or not?

4.4) (Optional) Preprocessing the PDB file

If you see an error message like "Residue numbers must occur in ascending order", you need to edit the PDB file in a right form. It is occurred because some information in the PDB file is not suitable for the SWISS-MODEL. In this case, you need to remove all the rows of "HETATM" and "ANISOU" in the PDB file. HETATM means the information of hetero atoms and ANISOU presents the anisotropic temperature factors. We do not need above two parts in the SWISS-MODEL. Please follow the instruction as below to remove those parts in the PDB file.

4.4.1) Linux or Mac

① Type the below commands at the Terminal. The first command deletes the rows of HETATM and the second command removes the row of ANISOU respectively. Replace the "file_name" with desire name of PDB file.

```
grep -v HETATM file_name.pdb > file_name_1.pdb
grep -v ANISOU file_name_1.pdb > file_name_2.pdb
```

② Use file_name_2.pdb to the template file.

4.4.2) Windows

① Type the below commands at the Windows PowerShell. The first command deletes the rows of HETATM and the second command removes the row of ANISOU respectively. Replace the "file_name" with desire name of PDB file.

```
gc file_name.pdb | ?{$_ -notmatch 'HETATM'} > file_name_1.pdb
gc file_name_1.pdb | ?{$_ -notmatch 'ANISOU'} > file_name_2.pdb
```

② Use file_name_2.pdb to the template file.

5. Reference

[1] Chothia, Cyrus, and Arthur M. Lesk. "The relation between the divergence of sequence and structure in proteins." The EMBO journal 5.4 (1986): 823-826.)

[2] https://swissmodel.expasy.org/static/course/files/PartII_homology_modelling.pdf

[3] Waterhouse, Andrew, et al. "SWISS-MODEL: homology modelling of protein structures and complexes." Nucleic acids research 46.W1 (2018): W296-W303.

[4] Altschul, Stephen F., et al. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.)