

Homology modelling using Modeller in Chimera

Exercises adopted from Chimera User's Guide, documentation & tutorials and Modeller Tutorials.

Exercise 1: Modeling of *Lottia pelta* malate dehydrogenase Modeller using Chimera web service

This tutorial includes running Blast sequence search and Modeller comparative modeling calculations from Chimera. Internet connectivity is required to fetch data and to access Blast, Modeller, and other web services. Although no software installation (other than Chimera itself) is needed to follow the tutorial, Modeller use requires a license key. Academic users can obtain a license key free of charge by registering at the Modeller website.

Modeling of protein malate dehydrogenase from *Lottia pelta*

Start Chimera.

Choose **File... Fetch by ID** from the menu and use the resulting dialog to fetch the sequence of the target, the Malate dehydrogenase from lottia peta: **UniProt ID B8R5G9**.

If you want to verify the ID before fetching, click the **Web Page** button on the fetch dialog to see the corresponding page at UniProt. (One way to determine the ID in the first place is by searching at the UniProt site.)

The sequence is displayed in **Multalign Viewer**, and its UniProt feature annotations listed in the **Region Browser**. The **S** column checkboxes in the **Region Browser** can be used to show feature annotations as colored boxes in the sequence window. Close the **Region Browser**; it can be accessed any time from the sequence window **Info** menu.

The next step is to find a known protein structure suitable for use as a modeling template. We will use Chimera's **Blast Protein** tool to search the Protein Data Bank (PDB; a database of known structures) for sequences similar to the target.

From the sequence window menu, choose **Info... Blast Protein**, click **OK** to use **B8R5G9** as the query, and **OK** again to perform the search using default settings, including **pdb** as the database to search. Searching the **pdb** sequences should take only a few seconds. Searching the **nr** database, which also contains a huge number of sequences without known structures, would take much longer.


In the Blast results dialog, the hits are listed from best to worst. Click the **Columns** button to reveal several checkboxes for controlling which columns of information are shown. Hide (uncheck) **Description**, then show **Resolution** and **Chain names**. The two best hits, PDB entries 5MDH and 4MDH, contain structures malate dehydrogenase. It is possible to use multiple templates, but we will use just 5MDH as it is a slightly higher resolution structure compared to 4MDH (PDB entry 5MDH, chain A). In the Blast results dialog, click to highlight the corresponding row, then at the bottom of the dialog:

1. click **Show in MAV** to display the query-hit sequence alignment from Blast in another **Multalign Viewer (MAV)** window
2. click **Load Structure** to fetch 5MDH from the PDB and open it in Chimera
3. click **Quit** to dismiss the Blast results dialog

Use the sequence window **Headers** menu to hide the **Consensus** and **Conservation** lines, then scroll or resize the window to show the whole alignment. Check the alignment for the query B8R5G9 and 5MDH sequence.

From the sequence window menu, choose **Structure... Modeller (homology)** to open the Chimera interface to comparative modeling with Modeller. The target should be set to query **B8R5G9** and click **5MDH** in the dialog to choose it as the template.

Click the **Advanced Options** button to reveal additional settings. **Run Modeller via web service** indicates using a web service hosted by the UCSF RBVI. No local installation is required to run the web service, but it is necessary to enter a **Modeller license key**, available free of charge to academic users upon registration at the Modeller website. After entering the license key, click **OK** to launch the calculation with default settings. Five comparative models will be generated.

The Modeller run may take several minutes and is handled as a background task. Clicking the information icon  near the bottom of the Chimera window will bring up the **Task Panel**, in which the job can be canceled if desired.

When the five models have been generated, they will be opened in Chimera and their evaluation scores shown in a **Model List** dialog. The models can be viewed individually or collectively by choosing rows in the dialog with the mouse. The different scores from Modeller use different criteria and will not necessarily agree on which models are best:

- **GA341** - model score derived from statistical potentials; a value > 0.7 generally indicates a reliable model, >95% probability of having the correct fold
- **zDOPE** - normalized Discrete Optimized Protein Energy (DOPE), an atomic distance-dependent statistical score; negative values indicate better models

To calculate RMSD between model and template and to rescore the models, choose **Fetch Scores... zDOPE and Estimated RMSD/Overlap** from the **Model List** menu. Rescoring uses a

web service provided by the Sali lab at UCSF. After a minute or few, more favorable zDOPE values are obtained, along with the additional scores:

- **Estimated RMSD** - TSVMMod-predicted C α root-mean-square deviation (RMSD) of the model from the native structure
- **Estimated Overlap (3.5 Å)** - TSVMMod-predicted native overlap (3.5 Å), fraction of C α atoms in the model within 3.5 Å of the corresponding atoms in the native structure after rigid-body superposition

The comparative models are atomically detailed and can be subjected to various analyses in Chimera.

Now we are going to color the best model (lowest estimated RMSD score) as spheres colored by amino acid hydrophobicity, from **dodger blue** for the most hydrophilic to **white** to **orange red** for the most hydrophobic. The **Model List** dialog was used to show only this model of the five, then the following commands were used to hide the template and adjust the model's appearance:

Command: **~modeldisp #0**

Command: **disp**

Command: **~ribbon**

Command: **rangecol kdHydrophobicity min dodger blue mid white max orange red**

Command: **preset apply pub 1**

Command: **repr sphere**

Exercise 2: Homology modeling of tyrosine phosphatase from *Staphylococcus aureus* using standalone Modeller

Create a directory and save all the files from this exercise within that directory.

1. Retrieve the query sequence:

Download the protein sequence of *protein-tyrosine-phosphatase PtpB* from *Staphylococcus aureus*. Open <http://www.uniprot.org/> in your browser, and search for the *PtpB* sequence. (Uniprot Id :P0C5D3)

The screenshot shows the UniProtKB entry for P0C5D3 (PTPB_STAAU). The page includes a search bar at the top, navigation links (BLAST, Align, Retrieve/ID mapping, Peptide search), and a 'Format' button. The main content area displays the protein name 'Low molecular weight protein-tyrosine-phosphatase PtpB', gene name 'ptpB', and organism 'Staphylococcus aureus'. It also shows the protein's status as 'Reviewed' with an annotation score of 4.0 and experimental evidence at the protein level. The 'Function' section is expanded, showing the catalytic activity: 'Protein tyrosine phosphate + H₂O = protein tyrosine + phosphate.' with a link to 1 publication. The 'Enzyme regulation' section is also expanded, showing 'Inhibited by N-ethylmaleimide and sodium orthovanadate.' with a link to 1 publication.

Click Format and save the sequence as a file in fasta format.

2. Template selection

Open Ncbi Blast page in your browser.
<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Standard Protein BLAST

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear **Query subrange**

From

To

Or, upload file No file selected.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Protein Data Bank proteins(pdb)

Organism

Optional Enter organism name or id—completions will be suggested Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Models (XM/XP) Uncultured/environmental sample sequences

Optional

Entrez Query [YouTube](#) [Create custom database](#)

Optional Enter an Entrez query to limit search

Copy the contents of the target fasta file ptpb.fasta into the Enter Query Sequence box. Or you can upload the fasta file directly. Change the “Choose Search Set” Database to Protein Data Bank proteins (pdb). Click the BLAST button. We are going to use a simple protein-protein BLAST for finding similar sequences. Run Blast. Once the got the results page, Select the top entry and Download the complete fasta file for the hit sequence. Note down the PDB accession code.

Now you should have the target and template sequences in fasta format files.

3. Download the template PDB File

Open protein data bank (PDB) in your browser and search for the template PDB file.







<http://www.rcsb.org/pdb/home/home.do>

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

RCSB PDB An Information Portal to 127184 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands

Advanced Search | Browse by Annotations

Download the protein in *pdb* file format. Open the protein file in a text editor and remove everything other than the template chain coordinates and save it again.

4. Sequence alignment

Open *Clustal Omega* page in your browser. Clustal Omega is a multiple sequence alignment program for proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences.

The screenshot shows the Clustal Omega web interface. At the top, there is a teal header with the text "Clustal Omega" and navigation links for "Input form", "Web services", "Help & Documentation", "Feedback", and "Share". Below the header, the breadcrumb "Tools > Multiple Sequence Alignment > Clustal Omega" is visible. The main heading is "Multiple Sequence Alignment". A brief description states: "Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools." An important note follows: "Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB." The interface is divided into three steps:
1. **STEP 1 - Enter your input sequences**: Includes a dropdown menu for "Enter or paste a set of" (currently set to "PROTEIN") and a large text area for "sequences in any supported format:". Below this is a file upload section: "Or, upload a file:" with a "Browse..." button and "No file selected."
2. **STEP 2 - Set your parameters**: Features an "OUTPUT FORMAT" dropdown menu set to "Clustal w/o numbers". A note states: "The default settings will fulfill the needs of most users and, for that reason, are not visible." There is a "More options..." link with the text "(Click here, if you want to view or change the default settings.)"
3. **STEP 3 - Submit your job**: Contains a checkbox for "Be notified by email (Tick this box if you want to be notified by email when the results are available)" and a "Submit" button.

The sequences to be aligned can be entered directly into this box or the file containing the sequences in fasta format can be uploaded. Running the tool is usually an interactive process, the results are delivered directly to the browser when they become available.

Copy and paste the sequences in fasta format directly in the box and submit your job. Choose output format as Pearson/Fasta.

Download the alignment file and save the file as 'alignment.fasta'.

5. PIR File Format

Convert the alignment.fasta file into a PIR file format. Modeller programs supports the PIR file format. Save the file as alignment.pir

Here is an example PIR file format.

```
>P1;5fd1
structureX:5fd1:1      :A:106  :A:ferredoxin:Azotobacter vinelandii: 1.90: 0.19
AFVVTDNCIKCKYTDCEVEVCPVDCFYEGPNFLVIHPDECIDCALCEPECPAQAI FSEDEVPEDMQEFIQLNAELA
EVWPNITEKKDPLPDAEDWDGVKGLQHLER*

>P1;1fdx
sequence:1fdx:1      : :54   : :ferredoxin:Peptococcus aerogenes: 2.00:-1.00
AYVINDSC--IACGACKPECPVNI IQGS--IYAIDADSCIDCGSCASVCPVGAPNPED-----
-----*
```

The first line of each sequence entry specifies the protein code after the >P1; line identifier. The line identifier must occur at the beginning of the line. The second line of each entry contains information necessary to extract atomic coordinates of the segment from the original PDB coordinate set. The fields in this line are separated by colon characters, `:`.

The fields are as follows:

Field 1:

A specification of whether or not 3D structure is available and of the type of the method used to obtain the structure (structureX, X-ray; structureN, NMR; structureM, model; sequence, sequence). Only structure is also a valid value.

Field 2:

The PDB filename or code.

Fields 3-6:

The residue and chain identifiers (see below) for the first (fields 3-4) and last residue (fields 5-6) of the sequence in the subsequent lines.

Field 7:

Protein name. Optional.

Field 8:

Source of the protein. Optional.

Field 9:

Resolution of the crystallographic analysis. Optional.

Field 10:

R-factor of the crystallographic analysis. Optional.

6. Run Modeller

Use a text editor to create the modeller script file called "model-single.py". Copy the following lines into the file.

```
from modeller import *
from modeller.automodel import *
env = environ()
a = automodel(env, alnfile='alignment.ali', knowns='template', sequence='target', assess_methods=(assess.DOPE, assess.GA341))
a.starting_model = 1
a.ending_model = 10
a.make()
```

Use text editor to edit *model-single.py* file. Change 'alnfile', 'knowns', 'sequence' to your alignment file, template and target names. Open a terminal and then type:

```
mod9.18 model-single.py
```

After a short wait your model should be ready. You will get the results in the running folder.

Models were name as **B9999000*.pdb*.

Read the log file. How many models you made? Which one has best score? Open your model in chimera to visualize and compare it with the template structure.