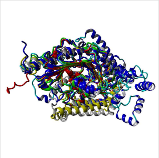


Sep 29, 2021

Homology modeling using SWISS-Model for Biochemistry I

Michael Friedman[1], Chris Berndsen[1]

[1]James Madison University

1 Works for me Share

dx.doi.org/10.17504/protocols.io.byntpven

Chris Berndsen
James Madison University

ABSTRACT

Protocol for homology modeling proteins for use in Biochemistry I at James Madison University. Protocol guides students to use the SWISS-Model web server (citations below).

Citations for servers:

1. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., and Schwede, T. (2018) *SWISS-MODEL: homology modelling of protein structures and complexes.* Nucleic Acids Res. 46, W296–W303.

DOI

dx.doi.org/10.17504/protocols.io.byntpven

PROTOCOL CITATION

Michael Friedman, Chris Berndsen 2021. Homology modeling using SWISS-Model for Biochemistry I. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.byntpven

LICENSE

CREATED

Sep 29, 2021

LAST MODIFIED

Sep 29, 2021

PROTOCOL INTEGER ID

53683

GUIDELINES

This protocol guides students through homology modeling and analysis of the resulting model. This protocol uses the CRX DNA binding domain to generate the results thus the shown images and results will vary.

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven

MATERIALS TEXT

SWISS-MODEL server: https://swissmodel.expasy.org/

Phyre^2 server: http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index

A sequence in FASTA format

Internet connection

Structure viewing program such as YASARA or UCSF Chimera

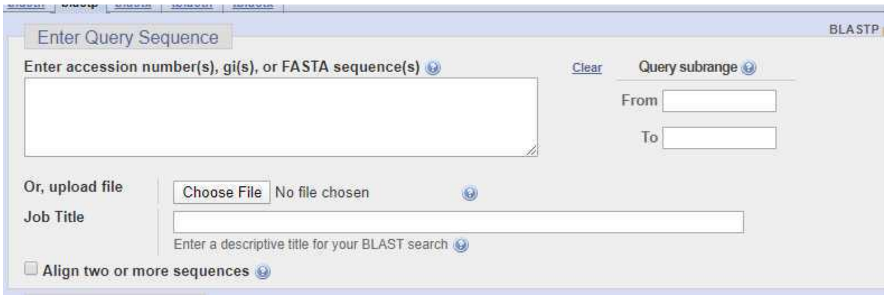
BEFORE STARTING

Gather your sequence in FASTA format (an example is shown below)

```
>seq_name
MASDETEASETEAMDAET
```

NCBI BLAST 10m

1 Navigate to NCBI BLAST (Basic Local Sequence Alignment Tool) and paste your sequence into the "Enter Query Sequence" box.



BLASTP

Enter Query Sequence

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

Clear

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

1.1 The standard settings for the search are shown in the table.

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven

A	B	C
	Default Setting	**What it does**
Enter Query Sequence		
Query Subrange	*(Blank)*	Limits search to a part of the sequence. Can be useful if there are common motifs/domains in the sequence.
Choose Search Set		
Database	*Non-redundant protein sequences (nr)*	Limits search to a sub-set of sequences. For homology modeling searching the Protein Data Bank proteins (pdb) is a good idea if you want to see if your modeling might be successful.
Organism	*(Blank)*	Limit search to a specific organism or other taxonomic group.
Exclude	*(Unchecked)*	Reduce results by removing certain classifications of sequences.
Program Selection		
Algorithm	*blastp*	Setting changes how the databases are searched. blastp is the most straight-forward. PSI-BLAST is useful when the query sequence is not easily aligned to other sequences.

1.2 Record any changes to the settings in Step 2.1 below:

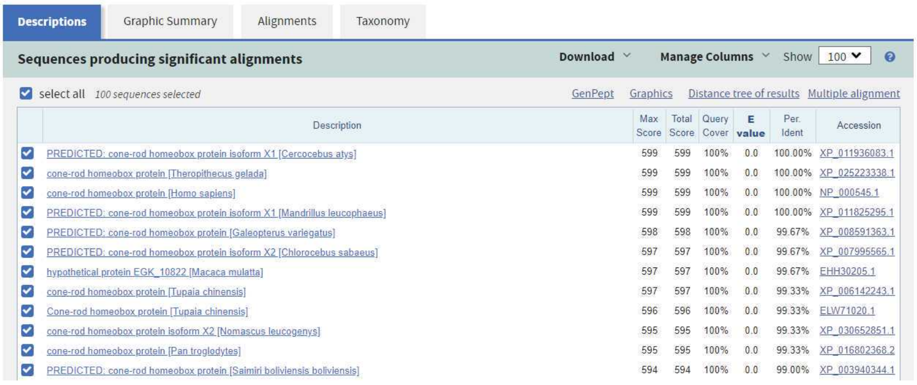
1.3 Press BLAST and wait until the results return.

Thie search can take up to ⏱**01:00:00 hour**

Analysis of BLAST results to ID sequence

2 Results will be returned as shown as below:

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	PREDICTED: cone-rod homeobox protein isoform X1 [Cercocebus atys]	599	599	100%	0.0	100.00%	XP_011936083.1
☑	cone-rod homeobox protein [Theropithecus gelada]	599	599	100%	0.0	100.00%	XP_025223338.1
☑	cone-rod homeobox protein [Homo sapiens]	599	599	100%	0.0	100.00%	NP_000545.1
☑	PREDICTED: cone-rod homeobox protein isoform X1 [Mandrillus leucophaeus]	599	599	100%	0.0	100.00%	XP_011825295.1
☑	PREDICTED: cone-rod homeobox protein [Galeopterus variegatus]	598	598	100%	0.0	99.67%	XP_008591363.1
☑	PREDICTED: cone-rod homeobox protein isoform X2 [Chlorocebus sabaeus]	597	597	100%	0.0	99.67%	XP_007995565.1
☑	hypothetical protein EGK_10822 [Macaca mulatta]	597	597	100%	0.0	99.67%	EHH30205.1
☑	cone-rod homeobox protein [Tupaia chinensis]	597	597	100%	0.0	99.33%	XP_006142243.1
☑	Cone-rod homeobox protein [Tupaia chinensis]	596	596	100%	0.0	99.33%	ELW71020.1
☑	cone-rod homeobox protein isoform X2 [Nomascus leucogenys]	595	595	100%	0.0	99.33%	XP_030652851.1
☑	cone-rod homeobox protein [Pan troglodytes]	595	595	100%	0.0	99.33%	XP_016802368.2
☑	PREDICTED: cone-rod homeobox protein [Saimiri boliviensis boliviensis]	594	594	100%	0.0	99.00%	XP_003940344.1

2.1 Column definitions from the **Descriptions** tab of the results.

A	B
Table column	**What it tells you**
Description	Tells you identify of matching sequence. Predicted or hypothetical in title indicates protein has not been verified.
Max Score	During alignment identities, similarities, and gaps are scored. This indicates the best score if the sequence was aligned multiple times.
Total Score	If many disconnected parts matched, this is the sum of the max scores for those
Query Cover	Indicates the percentage of the query sequence found in the match. 100% means all of the sequence was found.
E value	E(xpect) value tells you how many sequences that would rank higher if this was a random match. 0 or very small numbers are good.
Per. Ident	How much of the sequence was identical in sequence. Need >40% for good homology model.
Accession	The accession number for the sequence. Can be clicked to take you to the info card on that sequence.

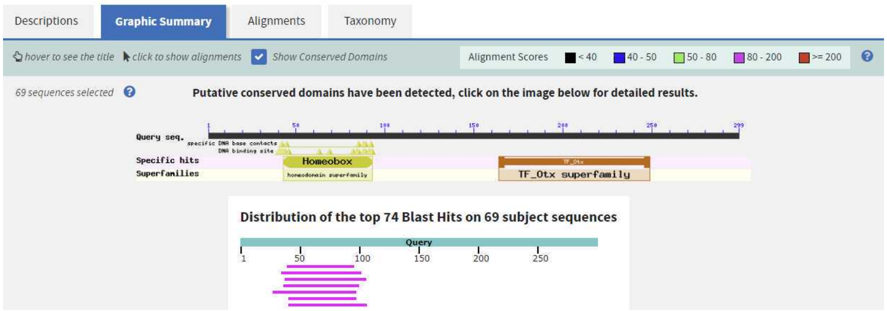
2.2 Record your best 5 sequences and their statistics in the table below.

Sequence Description	Max Score	Total Score	Query Coverage	E value	Per Ident	Accession

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven

3 In the **Graphic Summary** tab, you can view the domains in your sequence.

A **domain** is a part of the sequence with a known fold/shape/structure. A **motif** is a sequence that has a shape or function. Typically domains can fold on their on, while motifs are shorter pieces within domains.



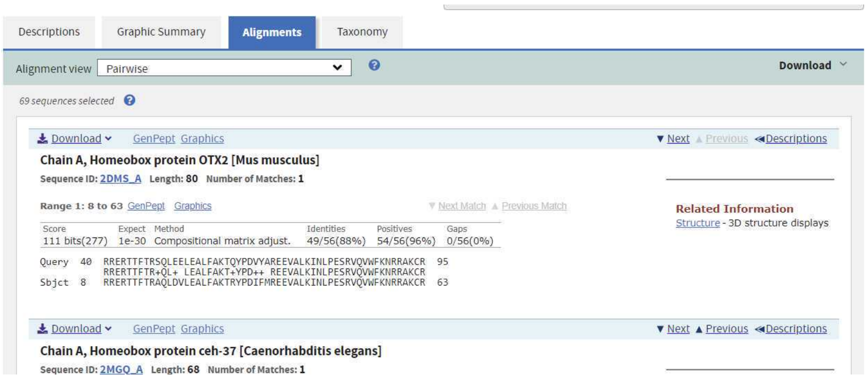
Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 74 Blast Hits on 69 subject sequences

3.1 Record any domains or motifs in the table below along with the approximate position within the sequence. This can help in the modeling and support the accuracy of your model later on.

A	B
Domain/Motif name	**position (this should be a number/set of numbers)**

4 In the **Alignments** tab, the actual sequence alignment (the data) are shown.



Alignment view: Pairwise

Chain A, Homeobox protein OTX2 [Mus musculus]

Sequence ID: 2DMS_A Length: 80 Number of Matches: 1

Range 1: 8 to 63 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
111 bits(277)	1e-30	Compositional matrix adjust.	49/56(88%)	54/56(96%)	0/56(0%)

```
Query  40   RRERTTFTRSQLEELEALFAKTQYPDVYAREEVALKINLPESRVQVWFKNRRAKCR  95
            RRERTTFTR+QL+ LEALFAKT+YPD++ REEVALKINLPESRVQVWFKNRRAKCR
Sbjct  8    RRERTTFTRAQLDVLEALFAKTRYPDIFMREEVALKINLPESRVQVWFKNRRAKCR  63
```

Related Information

Structure - 3D structure displays

Chain A, Homeobox protein ceh-37 [Caenorhabditis elegans]

Sequence ID: 2MGQ_A Length: 68 Number of Matches: 1

4.1 Each alignment shows the following key information:

- **Identities** and their location within the sequence.

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven

- **Positives** and their location within the sequence.
- **Gaps** and their location within the sequence.
- **The alignment**: Your sequence is the top row, the matched sequence in the middle row (+ means similar), and the sequece from the database (called Sbjct).
- **Position number** of the sequence match. These are the numbers at each end of the sequences.

4.2 Press the *Download* link to the top right of the alignment and select *Text* you will get a complete file of your results. Upload this to your folder for this project and name the file:

```
[date]_[sequence_name]_[team_name]_BLAST_alignment.txt
```

Replace **[Group_name]** with your name/group name without the brackets. Replace **[sequence_name]** with the name of the sequence.

4.3 ⚠

Indicate your file location as a link within a note on this step.

THIS IS YOUR DATA FILE FOR THE SEARCH!

Analysis of BLAST results to ID potential modeling templates

5 ↻ and repeat search but limit the Database to Protein Data Bank proteins (pdb). This search will identify proteins of known structure that match your protein and can suggest if your modeling attempt will be successful. Record your sequence matches in the table.

5.1 Accession numbers here lead to the information on the structure which may help when using SWISS-MODEL. These accession numbers are the PDB ID numbers.

Sequence Description	Max Score	Total Score	Query Coverage	E value	Per Ident	Accession

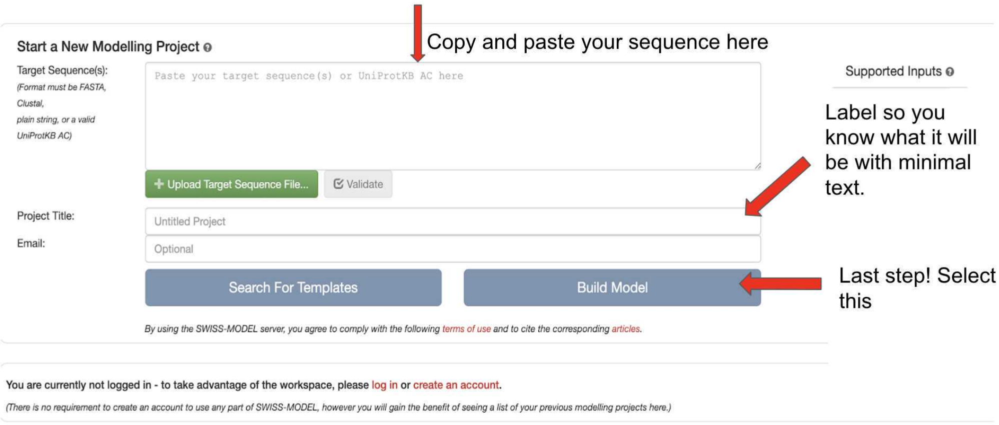
Table for recording results from PDB focused BLAST.

5.2 The top five structures here are potential **templates structures** which you can use to model your sequence. This means these structures are similar at the sequence level to your sequence and *potentially* will result in a similar structure to your sequence.

Homology modeling using SWISS-Model 5m

6 Click on the link for the SWISS-model server to get to a page that looks like 5m

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven



Start a New Modelling Project

Target Sequence(s): (Format must be FASTA, Clustal, plain string, or a valid UniProtKB AC)

Paste your target sequence(s) or UniProtKB AC here

Project Title: Untitled Project

Email: Optional

By using the SWISS-MODEL server, you agree to comply with the following terms of use and to cite the corresponding articles.

You are currently not logged in - to take advantage of the workspace, please log in or create an account.

(There is no requirement to create an account to use any part of SWISS-MODEL, however you will gain the benefit of seeing a list of your previous modelling projects here.)

Copy and paste your sequence here

Label so you know what it will be with minimal text.

Last step! Select this

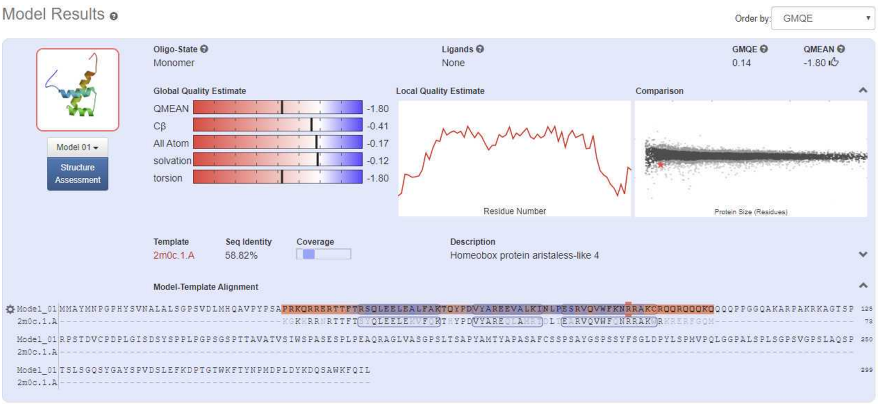
6.1 Follow the instructions on the image above to start the modeling by *Build Model.* Initial steps can take up to **00:20:00**

> Pressing Build Model is auto building using SWISS-Model with the best template according to SWISS-Model. Alternatively, you can Search for Templates and do template selection manually; guided by your templates identified using BLAST.

6.2 If you choose to do manual building via the Search for Templates tab. Record your templates below.

A
Template name (second column in table)

7 Once model building (either manual or automated) is complete, a screen as shown below will appear.



Model Results

Order by: GMQE

Oligo-State: Monomer

Ligands: None

GMQE: 0.14

QMEAN: -1.80

Global Quality Estimate

QMEAN	-1.80
Cβ	-0.41
All Atom	-0.17
solvation	-0.12
torsion	-1.80

Template	Seq Identity	Coverage	Description
2m0c.1.A	58.82%		Homeobox protein aristaless-like 4

Model-Template Alignment

```
Model_01 MMAYMNPGPHYSVNALALSGPSVDLMHQAVPYPSAPRKQRRERTTFTRSQLEELEALFAKTQYPDVYAREEVALKINLPESRVQVWFKNRRAKCRQQRQQQKQQQQPPGGQAKARPAKRKAGTSP 125
2m0c.1.A -----------------------------------KGKKRRNRTTFTSYQLEELEKVFQKTHYPDVYAREQLAMRTDLTEARVQVWFQNRRAKWRKRERFGQM---------------------- 73
Model_01 RPSTDVCPDPLGISDSYSPPLPGPSGSPTTAVATVSIWSPASESPLPEAQRAGLVASGPSLTSAPYAMTYAPASAFCSSPSAYGSPSSYFSGLDPYLSPMVPQLGGPALSPLSGPSVGPSLAQSP 250
2m0c.1.A -----------------------------------------------------------------------------------------------------------------------------
Model_01 TSLSGQSYGAYSPVDSLEFKDPTGTWKFTYNPMDPLDYKDQSAWKFQIL 299
2m0c.1.A -------------------------------------------------
```

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven

7.1 Useful information from this screen

- **GMQE** for Global Model Quality Estimation is scored from zero to 1 and indicates model quality based on the alignment with numbers closer to 1 indicating a more reliable model.
- **QMEAN** indicates the model quality based on structural features and the quality of the chemistry such as torsion angles and solvation. A good model has a number that is more positive, although a good model can have a negative QMEAN score. Less than -4 and model has bad chemistry.
- **Local Quality Estimate** indicates model quality on a per residue basis and can indicate if there are sections of hte model that are problematic (such as the ends of the model in the report above)
- **Model-Template alignment** shows how well the template structure and the sequence align and what parts of the model were used. Blue colors means better alignment while red colors mean worse alignment and modeling. Secondary structure is also indicated with tubes for α-helix and arrows for β-sheet.

8 The grey **Model** button leads to a menu to download information.

Two key options:

1. **PDB format** results in just the homology model, which can be viewed in Mol* or Chimera
2. **Model Report** downloads a .zip with the PDB file model and an HTML based report of the model process including the statistics shown in Step 8.1.

Download both and upload both files to your project.

Name the PDB file:

```
[date]_[sequencename]_[team_name]_SWISS_model.pdb
```

Replace **[Group_name]** with your name/group name without the brackets. Replace **[sequence_name]** with the name of the sequence.

Name the zip file:

```
[date]_[sequencename]_[team_name]_SWISS_data.zip
```

Replace **[Group_name]** with your name/group name without the brackets. Replace **[sequence_name]** with the name of the sequence.

THESE ARE YOUR DATA FILES FOR SWISS MODEL!

8.1 Indicate your file location as a link within a note on this step.

9 The **Structure Assessment** button leads to a new page showing a basic geometric and chemical assessment of the model.

9.1 The Ramachandran plot indicates if the **phi/psi angles** are appropriate for a protein structure and is interactive. The Phi angle is the **dihedral angle** for the rotation of the N-Cα bond while the psi angle is for rotation around the Cα-C bond of the amino acid backbone. The ideal angles for helices, sheets, and coils shown in green areas below are known to minimize steric clashes between atoms.

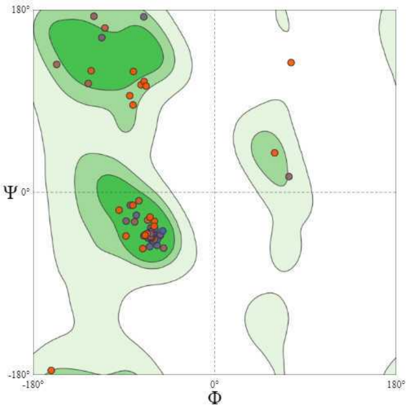
Use the camera tool to record the Ramachandran plot and upload it to your project folder.

Name the image file:

```
[date]_[sequencename]_[team_name]_SWISS_phipsi
```

Replace **[Group_name]** with your name/group name without the brackets. Replace **[sequence_name]** with the name of the sequence. Add a note with the link to your file.

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven



Ramachandran plot

9.2 The **Molprobity Results** are numerical scores based on the model and indicate what percentage of amino acids that fall in the ideal geometry category and have minimal clashes. The check boxes allow for visualization of the bad amino acids and can be useful to see if there are general model problems or localized issues. Localized issues can be fixed, general problems cannot.

9.3 Record your Molprobity numbers.

A	B	C
		Deviant amino acids
Molprobity Score		
Clash Score		
Ramachandran favored		
Ramachandran outliers		
Rotamer outliers		
C-beta deviations		
Bad bonds		
Bad angles		

10 Make sure you have recorded all the required data.

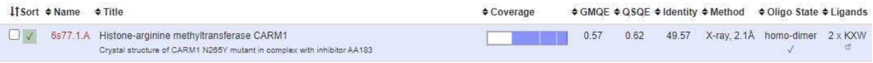
If you have completed the Phyre modeling. Save the record, export to PDF and upload this file to your project in the Notebook files.

Ligand identification using SWISS-Model

11 Clues to functionality can be gleaned from comparing unknown or predicted structures to with previously characterized structures of known function and characteristics. These scans can be biased against novel proteins or proteins with similar structures but distinct functions, but for initial guesses can be powerful. These methods align the structure and/or amino acid sequence to a database of structures with known ligand binding sites and look for structures with the similarity in amino acid composition, position, and over 3-D similarity. The idea being that similar structures lead to similar functions.

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven

12 Return back to your SWISS-Model search and note any models with ligands present. This is noted in the far right column of the templates page.



⇅Sort	⇕Name	⇕Title	⇕Coverage	⇕GMQE	⇕QSQE	⇕Identity	⇕Method	⇕Oligo State	⇕Ligands
□ ✓	6s77.1.A	Histone-arginine methyltransferase CARM1 Crystal structure of CARM1 N265Y mutant in complex with inhibitor AA183		0.57	0.62	49.57	X-ray, 2.1Å	homo-dimer ✓	2 x KXW

13 Identify the ligand by clicking on the ligand name. Record ligands in the top few hits in the table below.

Ligand name

14 Click the Name of the template to see where the ligands bind to the protein.

14.1 Generally ligands are classified into nonfunctional binders, covalent, and non-covalent binders. The latter two categories are the most interesting. Hovering over the ligand name shows the molecule bound to the protein and left-clicking on the name zooms the structure to show the specifics of ligand binding, including the weak interactions between the amino acids and ligand.

15 Download the top two hits with ligands bound from the server and align it to your model in Mol*.

A perfect match in RMSD is 0, while a poor match is one where the RMSD value is >3 Å, however a high RMSD value does not mean there are not regions of local similarity. A visual comparison is always helpful!

16 Observe if there is any match in the ligand/substrate binding sites between your model and the template structures with ligands bound.

Does the ligand “fit” into the aligned sites? It will not be perfect, so look for how bumps could fit into holes or nearby holes! Weak interactions also should be analyzed.

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven

16.1 Record the ligand name and possible interactions between your model and the ligand below.

A	B	C
Ligand name	Source structure PDB ID	Interacting amino acids in the model structure (three letter code and amino acid number)

17 Save this protocol as a PDF and upload it to your project folder.

Citation: Michael Friedman, Chris Berndsen (09/29/2021). Homology modeling using SWISS-Model for Biochemistry I. https://dx.doi.org/10.17504/protocols.io.byntpven