



May 20, 2020

Homology modeling for Biochemistry I

Michael Friedman¹, Chris Berndsen¹¹James Madison University

1 Works for me This protocol may be deleted by the owner

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 SWISS-Model and PHYRE2 web servers (citations below).

The protocol directs users to save data in OSF or the [Open Science Framework](#). This is the preferred project management tool for the class and is required for JMU students using this for the course. Other users can use whichever system is preferred.

Citations for servers:

1. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., and Schwede, T. (2017) *Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology*. Sci. Rep. 7, 10480.
2. Benkert, P., Biasini, M., and Schwede, T. (2011) *Toward the estimation of the absolute quality of individual protein structure models*. Bioinformatics 27, 343–350.
3. 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.
4. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015) *The Phyre2 web portal for protein modeling, prediction and analysis*. Nat. Protoc. 10, 845–858.

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.

The protocol directs users to save data in OSF or the [Open Science Framework](#). This is the preferred project management tool for the class and is required for JMU students using this for the course. Other users can use whichever system is preferred.

MATERIALS TEXT

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

Phyre² 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

Open Science Framework account (JMU students only)

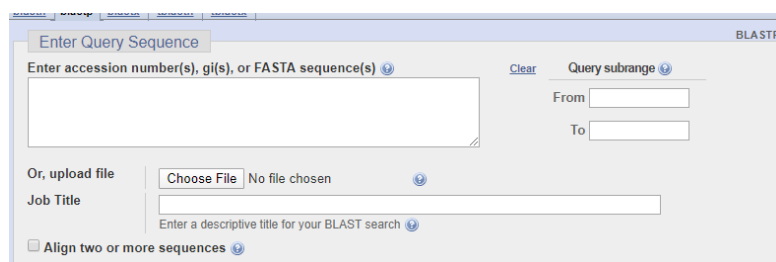
BEFORE STARTING

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.



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

	Default Setting	What it does
Enter Query Sequence		
<i>Query Subrange</i>	<i>(Blank)</i>	Limits search to a part of the sequence. Can be useful if there are common motifs/domains in the sequence.
Choose Search Set		
<i>Database</i>	<i>Non-redundant protein sequences (nr)</i>	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.
<i>Organism</i>	<i>(Blank)</i>	Limit search to a specific organism or other taxonomic group.
<i>Exclude</i>	<i>(Unchecked)</i>	Reduce results by removing certain classifications of sequences.
Program Selection		

Algorithm	blastp	Setting changes how the database s are searched. blastp is the most straight-forward. PSI-BLAST is useful when the query sequence is not easily aligned to other sequence s.

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

1.3 Press BLAST and wait until the results return.

This 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:

Descriptions	Graphic Summary	Alignments	Taxonomy				
Sequences producing significant alignments							
Download Manage Columns Show 100							
select all 100 sequences selected							
GenPept Graphics Distance tree of results Multiple alignment							
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	PREDICTED_cone-rod homeobox protein isoform X1 [Cercopithecus aethiops]	599	599	100%	0.0	100.00%	XP_011936083.1
<input checked="" type="checkbox"/>	cone-rod homeobox protein [Theropithecus gelada]	599	599	100%	0.0	100.00%	XP_025223338.1
<input checked="" type="checkbox"/>	cone-rod homeobox protein [Homo sapiens]	599	599	100%	0.0	100.00%	NP_000545.1
<input checked="" type="checkbox"/>	PREDICTED_cone-rod homeobox protein isoform X1 [Mandrillus leucophaeus]	599	599	100%	0.0	100.00%	XP_011825295.1
<input checked="" type="checkbox"/>	PREDICTED_cone-rod homeobox protein [Galeoscoptes variegatus]	598	598	100%	0.0	99.67%	XP_008591363.1
<input checked="" type="checkbox"/>	PREDICTED_cone-rod homeobox protein isoform X2 [Chlorocebus sabaeus]	597	597	100%	0.0	99.67%	XP_007995565.1
<input checked="" type="checkbox"/>	hypothetical protein EGK_10822 [Macaca mulatta]	597	597	100%	0.0	99.67%	EHH130205.1
<input checked="" type="checkbox"/>	cone-rod homeobox protein [Tupaia chinensis]	597	597	100%	0.0	99.33%	XP_006142343.1
<input checked="" type="checkbox"/>	Cone-rod homeobox protein [Tupaia chinensis]	596	596	100%	0.0	99.33%	ELW71620.1
<input checked="" type="checkbox"/>	cone-rod homeobox protein isoform X2 [Nomascus leucogenys]	595	595	100%	0.0	99.33%	XP_030652851.1
<input checked="" type="checkbox"/>	cone-rod homeobox protein [Pan troglodytes]	595	595	100%	0.0	99.33%	XP_016802368.2
<input checked="" type="checkbox"/>	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.

Table column	What it tells you
--------------	-------------------

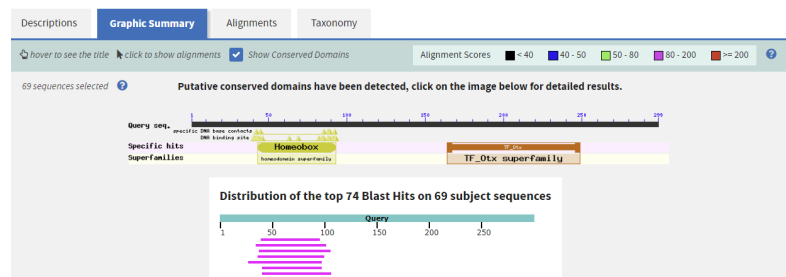
<i>Description</i>	Tells you identify of matching sequence. Predicted or hypothetical in title indicates protein has not been verified.
<i>Max Score</i>	During alignment identities, similarities, and gaps are scored. This indicates the best score if the sequence was aligned multiple times.
<i>Total Score</i>	If many disconnected parts matched, this is the sum of the max scores for those
<i>Query Cover</i>	Indicates the percentate of the query sequence found in the match. 100% means all of the sequence was found.
<i>E value</i>	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.
<i>Per. Ident</i>	How much of the sequence was identical in sequence. Need >40% for good homology model.
<i>Accession</i>	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

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

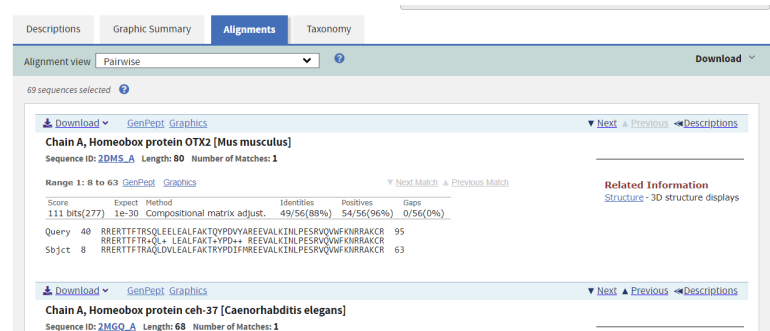
A **domain** is a part of the sequence with a known fold/shape/structure. A **motif** is a sequence that has a shape or function. Typically domains can fold on their on, while motifs are shorter pieces within domains.



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.

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.



4.1 Each alignment shows the following key information:

- **Identities** and their location within the sequence.
- **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 sequence 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 OSF folder for this project and name the file:

BLAST_alignment_[Group_name]_[sequence_name].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 OSF 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  **go to step #1** 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

6 Having identified the sequence and potential templates, now it is possible to start modeling the sequence to generate a potential sequence.

6.1 Follow the steps for the preferred server.



For the biochemistry course modeling project, both servers should be used.

Step 6.1 includes a Step case.

Phyre
SWISS-Model

step case

Phyre

This will outline the steps for modeling the structures using Phyre²

7 Go to the [Phyre2 server](#). This should take you to a page that looks like this.

Cambridge 2019 Workshop | Older Workshops | Phyre2 paper

E-mail Address:

Optional Job description:

Amino Acid Sequence:

Or try the sequence finder

Modelling Mode: ☒ Normal ☐ Intensive

Please tick as appropriate: ☒ NOT for Profit ☐ FOR Profit (Commercial) ☐ Other

Phyre Search Reset

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

Red arrows indicate the necessary things to change and select.

7.1 Paste your sequence into the *Amino Acid Sequence* box as shown above.

7.2 Provide your:

- **email address** so the results and model can be sent to you
- A **job description** so you can keep track of your data
- Which **mode** you want to use. Intensive takes longer but can give better results for models with few templates. Choose normal unless you identified less than 3 templates from BLAST.
- Select **NOT for profit** if you are a JMU student

Record your job description in this step as a note.

8 Something like this will appear. Your results will be sent to you via email. Time to retrieve the result varies depending on the server but usually is more than ⌚ 02:00:00

Phyre²

Job Status

Email	fried2ma@dukes.jmu.edu
Job Description	CRX_DBD_Phyre
Unique Job ID	2b3f9fcf3c81cbf1
Date	Mon Jan 27 15:51:50 GMT 2020

Estimated total processing time: 2.4 hours ± 2.1 hours ⓘ

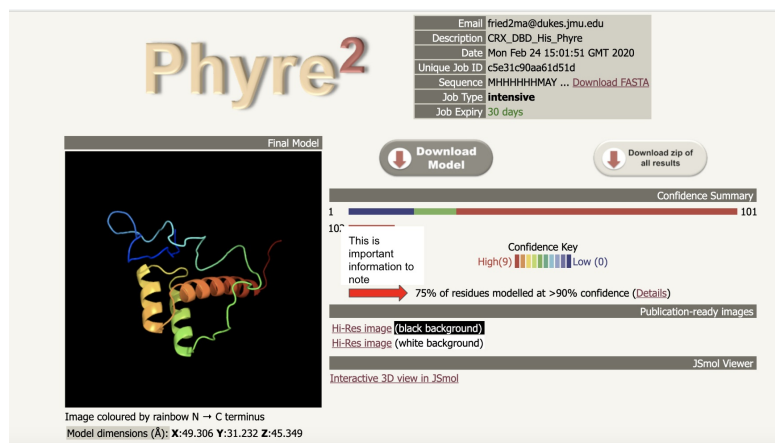
2. Building hidden Markov model of sequence

A link to results will be mailed to you when the job is finished

Or bookmark this page to return to it at any time

Analysis of Phyre2 results

9 Heres a sample of the results linked in the emailed results. Make sure to download this model for compairson. If you find that there are other models that this was built from that you prefer feel free to use their links for compairson too!



- 9.1 Note the percentage of residues modelled and the location of low confidence regions from the scheme in the **Confidence Summary** box.

Percent of residues modeled:	
Low confidence region locations:	

- 9.2 Download the model and the zip of all results and upload these files into OSF.

Name the .pdb file as:

PHYRE_model_[Group_name]_[sequence_name].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 as:

PHYRE_results[Group_name]_[sequence_name].zip

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

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

THIS IS YOUR DATA FILE FOR THE PHYRE modeling!

- 10 In the **Sequence analysis** section, you can download the sequence alignment file used in the modeling.

- 11 The **Secondary structure and disorder prediction** section, you can see what the predicted secondary structure is along with the confidence in that prediction (9 is high, 0 is low). Also, the disorder prediction is shown with ? suggesting disorder and the confidence in that prediction (9 is high, 0 is low).

A PDF of this figure can be download using the symbol on the left.

- 11.1 Upload your PDF to OSF.

Name the .pdf file as:

PHYRE_SecStrPred_[Group_name]_[sequence_name].pdf

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

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

12 In the **Domain Analysis** section, you can move the cursor over each red part of the aligned region and see the predicted domains.



This should match the domains identified in step 3!

12.1 Record the code and the domain name for the top 5 hits in the table.

Code	Domain/motif

13 In the **Detailed Template information** table, there is important information about the templates.

Use this link to download if you prefer this over the built model

This is important information to note

#	Template	Alignment Coverage	3D Model	Confidence	% I.D.	Template Information
1	d1pufA	Alignment		99.9	32	Fold: DNA/RNA-binding 3-helical bundle Superfamily: homeodomain-like Family: Homeodomain PDB header: dna binding protein Chain: A; PDB Molecule: homeobox protein barh-like 1; PDBTitle: solution structure of the homeobox domain of homeobox2 protein barh-like 1 Run Investigator
2	c2dmrA	Alignment		99.9	31	PDB header: dna binding protein Chain: A; PDB Molecule: homeobox protein barh-like 1; PDBTitle: solution structure of the homeobox domain of homeobox2 protein barh-like 1 Run Investigator
3	c2m34A	Alignment		99.9	39	PDB header: transcription Chain: A; PDB Molecule: homeobox protein gbx-1; PDBTitle: rnm structure of the homeodomain transcription factor gbx1 from homo2 sapiens Run Investigator
4	c2mgaA	Alignment		99.9	69	PDB header: dna binding protein Chain: A; PDB Molecule: homeobox protein ceh-37; PDBTitle: structure of ceh37 homeodomain Run Investigator

13.1 Take a screen shot of the table showing the top 5 hits and upload the photo to OSF.

Name the file as:

PHYRE_templateinfo_[Group_name]_[sequence_name]

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

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

14 Save your record, export it as a PDF, and place it in the OSF folder for your notebook files. *If this is part of the modeling project, make sure that you also modeled using SWISS-model*