University of
Strathclyde
Science

Evolutionary and Structural Analysis of Pathogen Proteins.

Final year UG project 2025-26

Introduce Yourself

Project Expectations

Learning Agreement

Outlines responsibilities of students and staff

Please read and sign the learning agreement (MyPlace), and send a signed copy to me (.jpg/.png signature is fine)

[download link]

I’ll return a signed copy to you.

Then **you** upload the double-signed copy on MyPlace **as a PDF file**

[upload link]

Other Expectations - 1

- **Maintain a lab notebook – electronic or handwritten**
    - Benchling is popular, used in industry, and has a free tier: https://www.benchling.com/academic
    - Plain text files are perfectly fine (e.g. Notepad++ if on Windows)
    - Jupyter notebooks/Quarto are flexible and welcomed
    - Word documents are tolerable, but note they are proprietary format and not easily archiveable
- **Back up your work!**
    - Read/complete the Data Management Plan template (project webpages)
    - University shared drives (OneDrive)
    - External hard drives are good
    - I expect to receive your raw data files, and project output files, at the end of the project

Other Expectations - 2

- **I expect you to work together**
  - You are all working on different proteins/systems
  - Sharing information about techniques, approaches, software, etc. is not plagiarism – it's peer learning
- **Be kind to yourself and others**
- **Communicate clearly, openly, and honestly**
  - If I don't know there's a problem, I can't help – so tell me
  - Time moves very quickly – if you have a question, ask it; **don't wait for the next group meeting**
- See the learning agreement for more…

The Project

Overview

Pathogens are in arms races with their hosts

The weaponry is often proteins

Understanding how the weapons work helps understand disease, and identify candidate drug targets

Protein function is a consequence of sequence and structure

Looking at **sequence evolution** helps identify conserved and variable residues; conserved sites are presumably under selective pressure

Having a **3D protein structure** helps locate residues (e.g. internal vs external) and interpret potential selective pressures, which may imply druggable importance, and/or suggest future experiments

The Importance of AlphaFold

Protein structures are often difficult and expensive to obtain

AlphaFold does an excellent job of predicting structures in many cases, so shortcuts this process for thousands of proteins

We can now use AlphaFold predictions to help interpret sequence-based evolutionary analyses (e.g. positive selection)

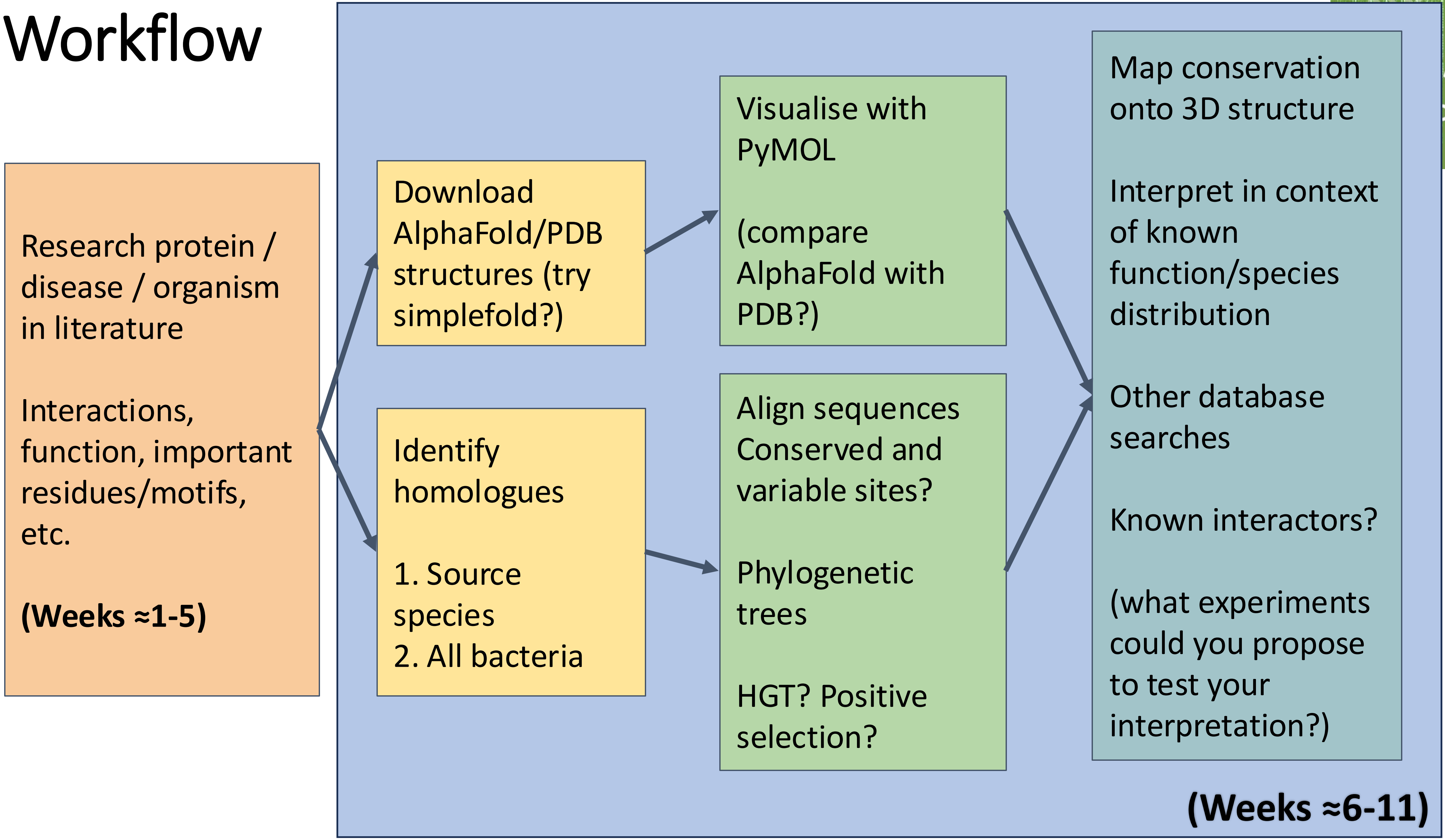
We couldn't do this project in this way prior to 2021!

- https://www.youtube.com/watch?v=j9UHcxucKZE Protein Structure Prediction in a Post-AlphaFold2 World (54min)

- https://www.ebi.ac.uk/training/events/how-interpret-alphafold-structures/ (use the Watch Video link) How to interpret AlphaFold structures (100min)

Maybe take a look at simplefold: https://github.com/apple/ml-simplefold

Workflow



Research protein / disease / organism in literature

Interactions, function, important residues/motifs, etc.

(Weeks ≈1-5)

Download AlphaFold/PDB structures (try simplefold?)

Identify homologues

1. Source species
2. All bacteria

Visualise with PyMOL

(compare AlphaFold with PDB?)

Align sequences Conserved and variable sites?

Phylogenetic trees

HGT? Positive selection?

Map conservation onto 3D structure

Interpret in context of known function/species distribution

Other database searches

Known interactors?

(what experiments could you propose to test your interpretation?)

(Weeks ≈6-11)

Candidate proteins

University of Strathclyde Science

These proteins all:

- Have an entry in PHI-base with evidential support for a role in virulence (you can find references in the PHI-base records)
- Have an AlphaFold prediction at the EMBL AlphaFold DB or on UniProt
- Have homologues in UniProt (you can find references, other info here, also)

These proteins might:

- Also have a solved PDB structure
- Not have many homologues in UniProt

You can look for your own protein of interest, if you prefer, but please contact Leighton to check that it's suitable.

Candidate proteins – start points

Organism	Host	Gene/Protein	PHI accession	Student
Escherichia coli	*Homo sapiens*	*espY*	PHI:8647	LB
Shigella flexneri	*Homo sapiens*	*ipaJ*	PHI:9253	LT
Candida albicans	*Mus musculus*	*sap6*	PHI:10193	IM
Pseudomonas aeruginosa	*Homo sapiens*	*tplE*	PHI:6646	AE
Vibrio vulnificus	*Mus musculus*	*vvhA*	PHI:6877	JT

http://www.phi-base.org/

Useful tools (many others are available)

University of Strathclyde Science

GalaxyEU: https://usegalaxy.eu/

- Sequence alignment (e.g. MAFFT), phylogenetics (e.g. RaxML), positive selection (e.g. codeML)

iTOL: https://itol.embl.de/

- Visualisation/annotation of phylogenetic trees

PyMOL: https://pymol.org/2/ and/or ChimeraX: https://www.cgl.ucsf.edu/chimerax/

- Protein structure visualisation/annotation

Jalview: http://www.jalview.org/

- Visualisation of multiple sequence alignments

Windows vs Mac vs Linux… GUI vs terminal…

Useful sites/databases

PHI-base: http://www.phi-base.org/

- Proteins involved in host-pathogen interactions, with linked evidence

EMBL AlphaFold: https://www.alphafold.ebi.ac.uk/

- AlphaFold predictions for proteins from model organisms

UniProt: https://www.uniprot.org/

- Protein sequence (including homologous sequences) and functional information with evidence

RCSB/PDB: https://www.rcsb.org/

- Repository of record for protein structures

SIPBS CompBiol Sites

- BM432 Project Pages
  - https://sipbs-compbiol.github.io/bm432-project/

- An incomplete little book of bioinformatics
  - https://sipbs-compbiol.github.io/little-bioinformatics-book/

Project Management Tools

You may want tools to…

- Manage your time
  - E.g. Pomodoro technique (e.g. BeFocused, Pomofocus, Forest)
- Schedule work
  - Reminders (macOS, MS Office)
  - Calendar (macOS, MS Office), with email alerts
  - Trello, Asana, etc.
- Manage your project data and information effectively
  - How to name files
  - Project management guidelines (BM432, 2022-23 session; me and Dr Feeney)
  - How to keep a lab notebook
  - Keeping a computational biology lab notebook: https://doi.org/10.1371/journal.pcbi.1004385
  - Organising a lab book

Next Week’s Group Meetings

Monday 6th October, 1:30pm, HW324

Thursday 9th October, 10:30am, HW324

Topics to Discuss at Next Meeting

- How the literature search is going
- How are you managing your time?
- Share advice
  - How to find useful papers
  - What databases are helpful
  - What software tools might be useful