

Evolutionary and Structural Analysis of Pathogen Proteins.

Final year UG project 2024-25
2024-10-27 (Week 6)

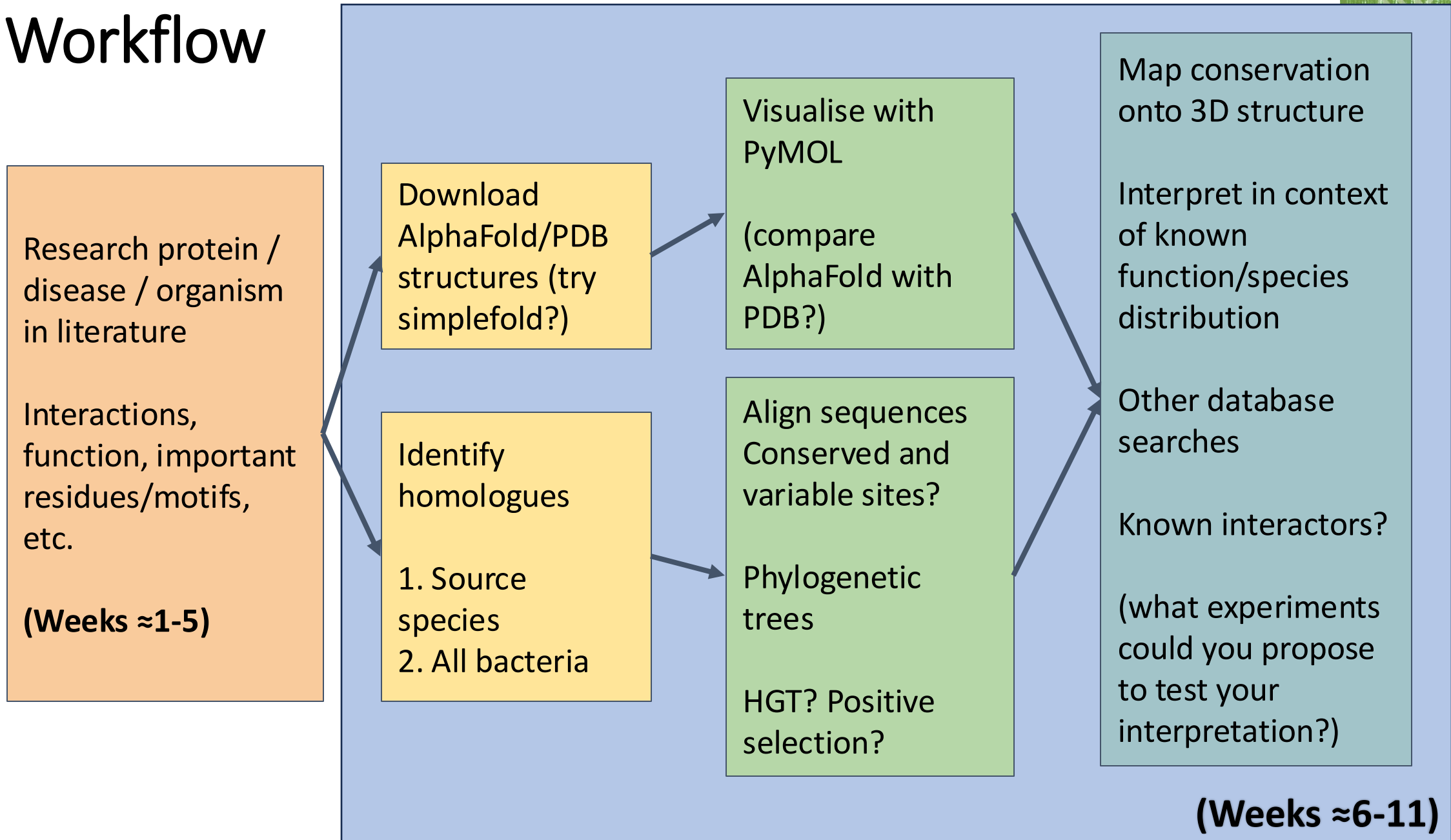
Candidate proteins – start points

Any changes needed?

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

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

Workflow



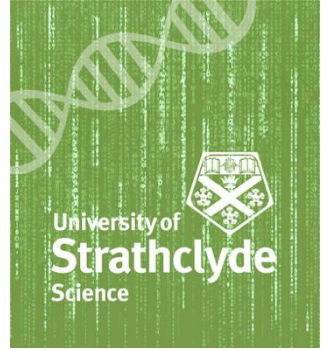
Your questions/comments

(What would you like to talk about?)

Tell everyone about your protein

Identifying and aligning homologues

Useful steps - 1



- Identify UniProt record from PHI-Base
- UniProt has homologues: 100% (15), 90% (102), 50% (107) identity – almost all *Mycobacterium* spp.
 - (what does this tell us?)
- (Similar Proteins -> View all entries -> Download)

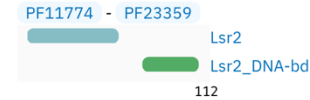
University of
Strathclyde
Science

-
- hypoetical protein [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacteriaceae]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium botniense]
- nucleoid-associated protein Lsr2 [Mycobacterium xenopi]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacteriaceae bacterium]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacteriaceae]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium burgundense]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacillus trivialis]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium sp. MS1601]
- Lsr2 family protein [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium thermoresistibile]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [uncultured Mycolicibacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium terrae]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium koreensis]
- high G+C Gram-positive bacteria | 2 leaves
- histone-like nucleoid-structuring protein Lsr2 [Mycobacteroides abscessus]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacteroides abscessus]
- high G+C Gram-positive bacteria | 3 leaves
- high G+C Gram-positive bacteria | 2 leaves
- histone-like nucleoid-structuring protein Lsr2 [Mycobacteroides abscessus]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacteroides abscessus]
- histone-like nucleoid-structuring protein Lsr2 [Tsukamurella solii]
- histone-like nucleoid-structuring protein Lsr2 [Tsukamurella paurometabola]
- ribosomal protein S10 [Tsukamurella ocularis]
- histone-like nucleoid-structuring protein Lsr2 [uncultured Tsukamurella]
- high G+C Gram-positive bacteria | 2 leaves
- histone-like nucleoid-structuring protein Lsr2 [uncultured Rhodococcus (in: high G+C Gram-positive bacteria)]
- histone-like nucleoid-structuring protein Lsr2 [Rhodococcus sp. NPDC00318]
- histone-like nucleoid-structuring protein Lsr2 [Smargdicoccus niigatensis]
- Lsr2 family protein [Mycobacteriaceae bacterium]
- high G+C Gram-positive bacteria | 2 leaves
- high G+C Gram-positive bacteria | 2 leaves
- high G+C Gram-positive bacteria | 3 leaves
- high G+C Gram-positive bacteria | 3 leaves
- histone-like nucleoid-structuring protein Lsr2 [Antrihabians cavemicola]
- histone-like nucleoid-structuring protein Lsr2 [Antrihabians stalacti]
- histone-like nucleoid-structuring protein Lsr2 [Skermania sp. ID1734]
- histone-like nucleoid-structuring protein Lsr2 [Skermania sp.]
- histone-like nucleoid-structuring protein Lsr2 [Hoyosella]
- high G+C Gram-positive bacteria | 3 leaves
- high G+C Gram-positive bacteria | 3 leaves
- histone-like nucleoid-structuring protein Lsr2 [Williamsia sterculiae]
- histone-like nucleoid-structuring protein Lsr2 [Williamsia sp. 1135]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium fallax]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium avium subsp. hominissus]
- Lsr2 family protein [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Rhodococcus sp. (in: high G+C Gram-positive bacteria)]
- Lsr2 family protein [Mycobacteriaceae bacterium]
- Lsr2 family protein [Mycobacterium sp.]
- Lsr2 family protein [Mycobacterium sp.]
- Lsr2 family protein [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium sarraeanae]
- Lsr2 [Mycobacterium vanbaalenii PYR-1]
- histone-like nucleoid-structuring protein Lsr2 [Actinomyces bacterium]
- hypothetical protein [Mycobacterium sp.]
- Lsr2 protein [Mycobacterium merochaense]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium sp. AZCC_0083]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium sp. MAA66]
- Lsr2 [Mycobacteriaceae bacterium]
- Lsr2 family protein [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium sp. GA-2829]
- Lsr2 family protein [Mycobacterium sp.]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium uberis]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium heidelbergense]
- nucleoid-associated protein Lsr2 [Mycobacterium heidelbergense]
- iron-regulated nucleoid-associated protein Lsr2 [Mycobacterium montefiorensis]
- hypothetical protein MFL01_40290 [Mycobacterium florentinum]
- nucleoid-associated protein Lsr2 [Mycobacterium lacus]
- iron-regulated Lsr2 protein precursor [Mycobacterium tuberculosis]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium tuberculosis]
- sp19WIP1.LSR2.MYCTU.Nucleoid-associated protein Lsr2 OS=Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv) OX=83332 GN=Lsr2 PE=1 SV=1
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium tuberculosis]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium tuberculosis]
- histone-like nucleoid-structuring protein Lsr2 [Mycobacterium tuberculosis]

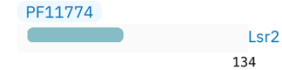
Useful steps - 3

- Protein has a known domain – homologous domains in Pfam (8k)
- Various domain architectures, many the same as my query (7576)
 - (What does this tell us?)
- (Click on link and Download)

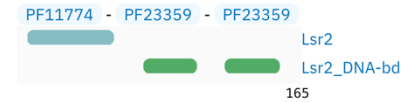
There are 7576 proteins with this architecture (represented by P24094):



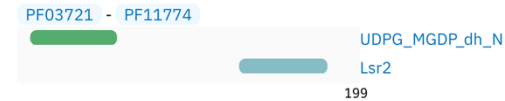
There are 436 proteins with this architecture (represented by A0QRB8):



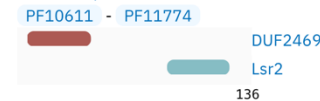
There are 9 proteins with this architecture (represented by A0A7G6WV2):



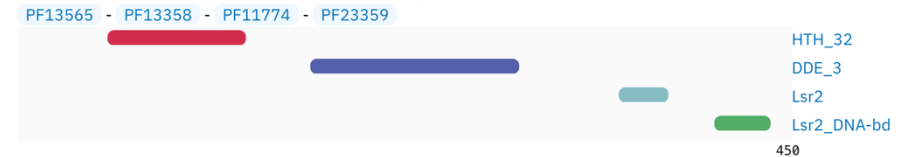
There are 2 proteins with this architecture (represented by A0A166HKT3):



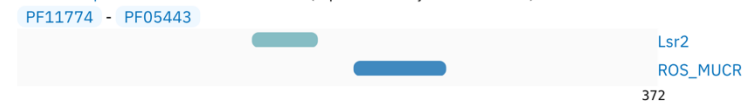
There is 1 protein with this architecture (represented by A0A180FWU6):



There is 1 protein with this architecture (represented by A0ABU1Q480):



There is 1 protein with this architecture (represented by A0A2K9D9C8):



Useful steps - 4

- Homologue databases
 - OrthoDB, InParanoid, EggNOG, etc.
- OrthoDB (5985)
- (Get ortholog groups -> select group -> View protein FASTA)

OrthoDB v12.2

get Gene ▼ P9WIP7 ► Advanced Submit

Gene Info View protein fasta View CDS fasta

Mycobacterium tuberculosis H37Rv (GCF_000195955.2)

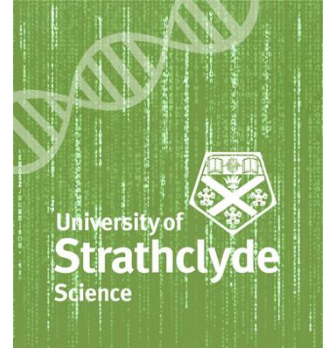
[Rv3597c](#) [4040980:4041319](-) Nucleoid-associated protein Lsr2 » 112 Q [IPR042261](#) [24412](#) [36625](#)

Orthologs in example species Get Ortholog Groups

search species (by name)

Mycobacterium intermedium - group 103947at1763 at Mycobacterium level	
BHQ20_RS12395 [46156:46495](+) Lsr2 family protein »	112 Q IPR024412 42261 36625
Mycobacterium smegmatis MC2 155 - group 164605at1762 at Mycobacteriaceae level	
MSMEG_6092 [6155785:6156130](-) Lsr2 protein »	114 Q IPR024412 42261 36625
Rhodococcus sp. LB1 - group 102564at85007 at Mycobacteriales level	
AZG88_RS35125 [63630:63981](+) Lsr2 family protein »	116 Q IPR042261 24412 36625
Nocardia seriolae - group 102564at85007 at Mycobacteriales level	
1 NS506_RS01565 [309408:309771](-) Lsr2 family protein »	120 Q IPR042261 24412 36625
2 NS506_RS34590 [7449704:7450046](-) Lsr2 family protein »	113 Q IPR042261 24412 36625
Gordonia rhizosphaera NBRC 16068 - group 102564at85007 at Mycobacteriales level	
1 GORHZ_RS17965 [70531:70867](+) Lsr2 family protein »	111 Q IPR042261 24412 36625
2 GORHZ_RS00400 [15175:15544](-) LSR2-like protein »	122 Q IPR024412 42261 36625
Microbacterium sp. No. 7 - group 911508at1760 at Actinomycetes level	
AOA12_RS17560 [3778716:3779058](-) Lsr2 family protein »	113 Q IPR024412 42261 36625

Useful steps - 4



- Concatenate sequence files
 - Galaxy, cat, JalView
- Can also filter identical sequences

 **FASTA Merge Files and Filter Unique Sequences** Concatenate FASTA database files together (Galaxy Version 1.2.0)



 Run Tool

Tool Parameters

Run in batch mode?

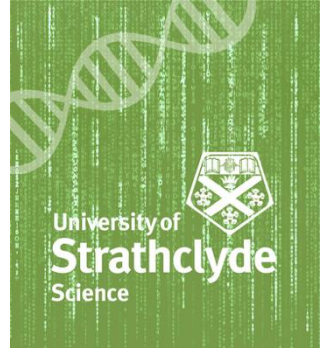
- ☒ Merge individual FASTAs (output collection if input is collection)
- ☐ Merge all FASTAs (always output a single FASTA)

The 'merge all' mode produces one output FASTA for all input FASTA files. The individual mode generates one FASTA file for each set of input FASTAs. For example, if the tool is given 2 collections of 10 FASTAs, it will merge the collections pairwise to create an output collection of 10 FASTAs.

Input FASTA File(s)

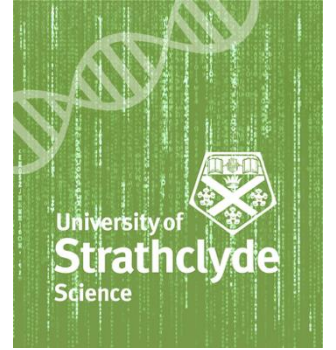
 Insert Input FASTA File(s)

Useful steps - 5



- Align sequences
 - MAFFT, Clustal Omega, MUSCLE, MMSeqs2
- Use standard scoring matrix or incorporate sequence composition?
- Does the alignment look reasonable?
 - If not, what to do?
 - Exclude distant sequences?
 - Align clusters?
 - Can it be “rescued” with trimAl?
 - Would using a structure or other pre-alignment to “anchor” the alignment help?

Topics/Progress to Discuss at Next Meeting



- What would you like to cover?
- **Keep investigating your protein**
- Make a tree for your sequences
 - How does the tree look?
 - Are organisms/species grouped together how you'd expect? Any surprises?
 - Is there evidence supporting gene duplication (two clades, with the same organisms repeated)?
 - Is there evidence supporting horizontal gene transfer (organism in the "wrong" clade)?
- Compare to sequence data
 - Do patterns of sequence variation map onto your tree?
- Galaxy workflow or scripting?
 - Have you created a workflow for your analysis: alignment? making a tree?
 - Have you noted/saved the tool versions/parameter settings for writing your thesis?

Next Week's Group Meeting
Monday 3rd November 13:30 HW324

Useful Links

Useful tools (many others are available)

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

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

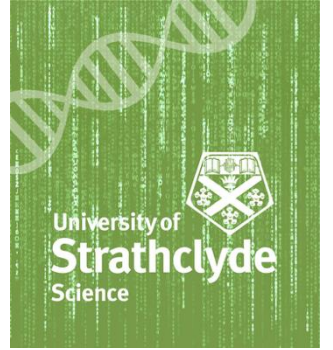
UniProt: <https://www.uniprot.org/>

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

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

- Repository of record for protein structures

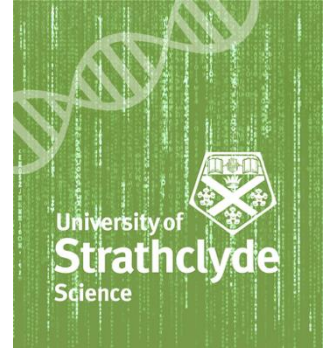
SIPBS CompBiol Sites



- 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](#)