University of
Strathclyde
Science

Evolutionary and Structural Analysis of Pathogen Proteins.

Final year UG project 2024-25
2024-10-13 (Week 5)

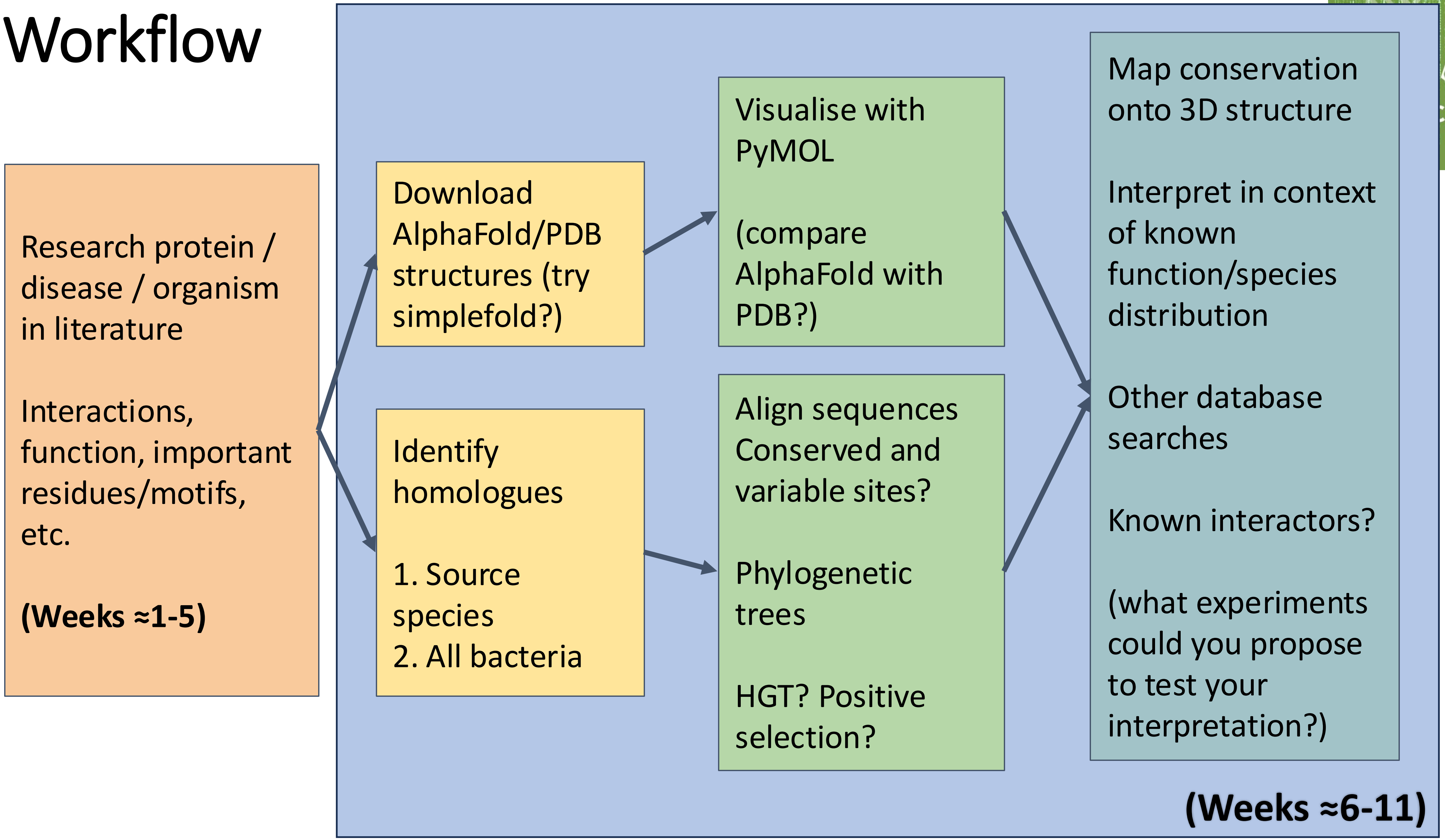
Candidate proteins – start points

Any changes needed?

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/

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)

Your questions/comments

(What would you like to talk about?)

Thesis Introductions

Examples

- Lsr2 (Ailsa)
- SpA (Yann)

- Note:
  - Use of figures
  - Density of references
  - Overall depth of description

Sequence Alignment

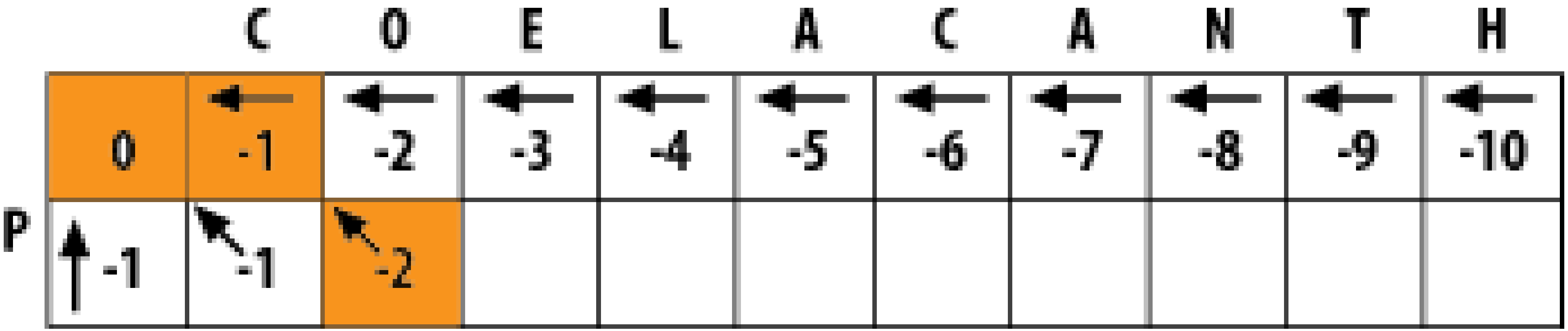
Pairwise sequence alignment

- We have two words we want to align:
  - **COELACANTH**
  - **PELICAN**
- We have a scoring scheme:
  - Matching identical letters scores **+1**
  - If we have different letters aligned, that scores **-1**
  - Inserting a gap in the alignment scores **-1**
- All sequence alignment is a mathematical operation: *the maximization of an alignment score*

Initialise a matrix

		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1										
E	↑ -2										
L	↑ -3										
I	↑ -4										
C	↑ -5										
A	↑ -6										
N	↑ -7										

Start filling the matrix

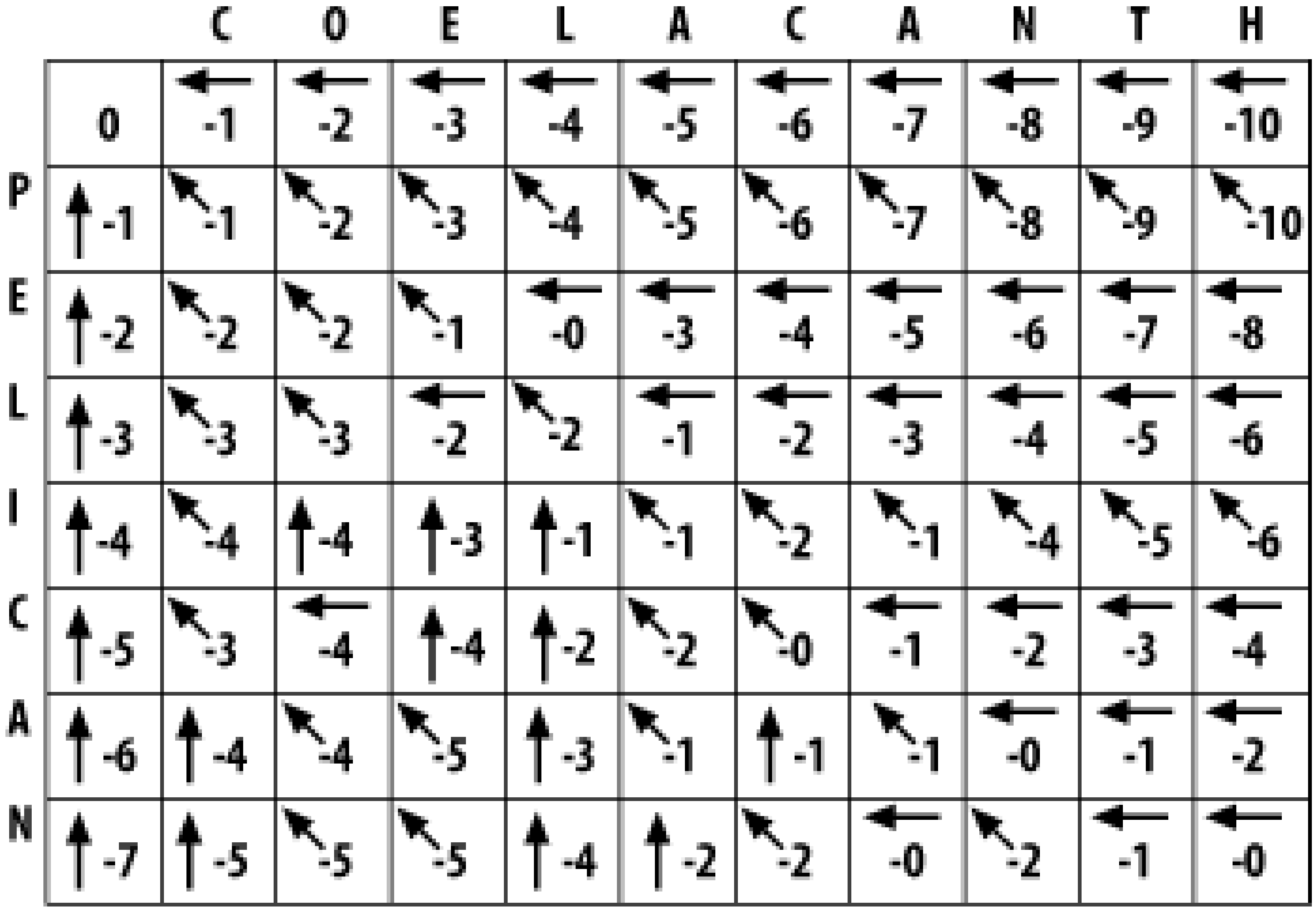


		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2								

```
CO
-P
```

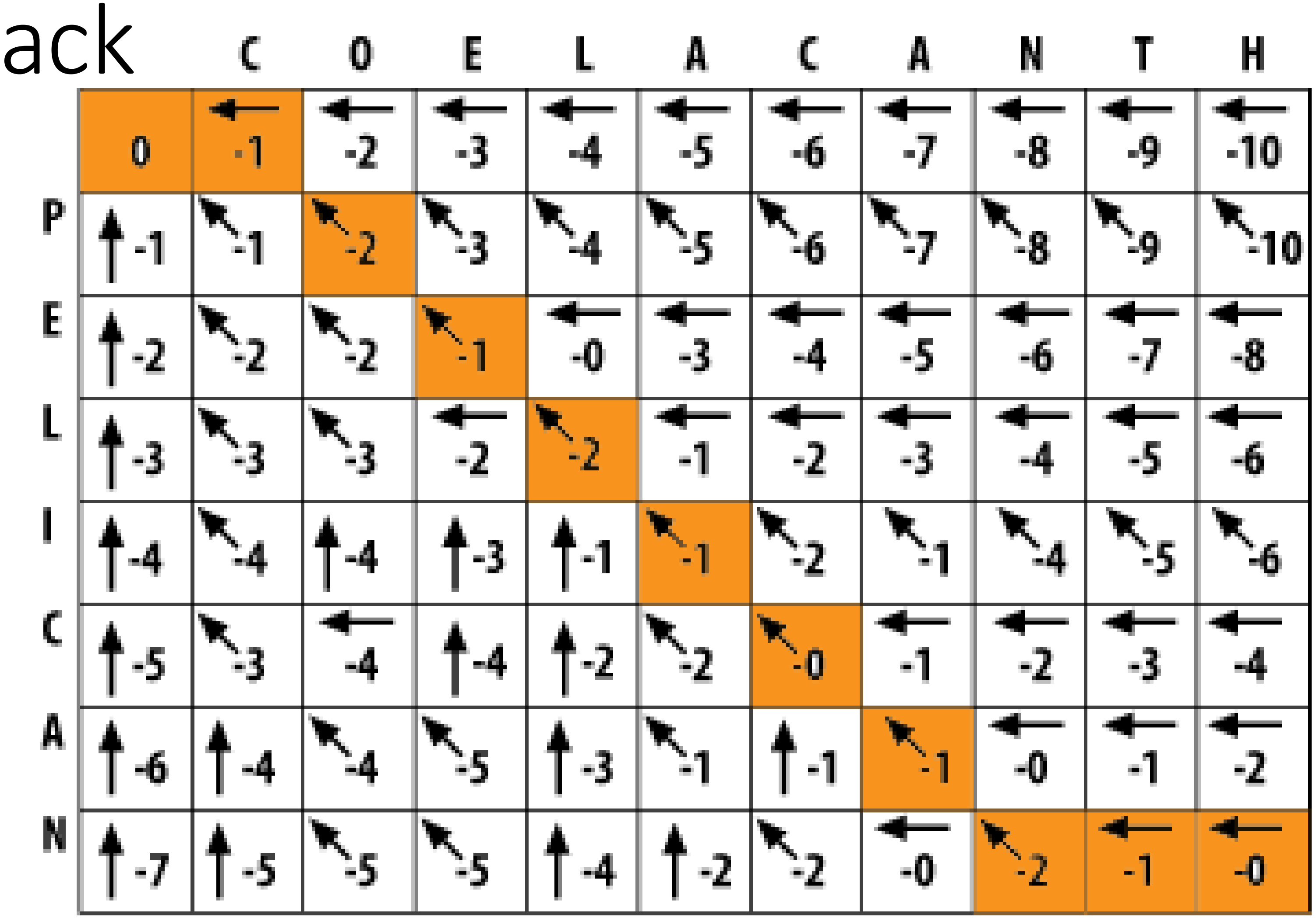
```
CO
P-
```

A full matrix



		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6	↖ -7	↖ -8	↖ -9	↖ -10
E	↑ -2	↖ -2	↖ -2	↖ -1	← -0	← -3	← -4	← -5	← -6	← -7	← -8
L	↑ -3	↖ -3	↖ -3	← -2	↖ -2	← -1	← -2	← -3	← -4	← -5	← -6
I	↑ -4	↖ -4	↑ -4	↑ -3	↑ -1	↖ -1	↖ -2	↖ -1	↖ -4	↖ -5	↖ -6
C	↑ -5	↖ -3	← -4	↑ -4	↑ -2	↖ -2	↖ -0	← -1	← -2	← -3	← -4
A	↑ -6	↑ -4	↖ -4	↖ -5	↑ -3	↖ -1	↑ -1	↖ -1	← -0	← -1	← -2
N	↑ -7	↑ -5	↖ -5	↖ -5	↑ -4	↑ -2	↖ -2	← -0	↖ -2	← -1	← -0

Traceback



		C	O	E	L	A	C	A	N	T	H
	0	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
P	↑ -1	↖ -1	↖ -2	↖ -3	↖ -4	↖ -5	↖ -6	↖ -7	↖ -8	↖ -9	↖ -10
E	↑ -2	↖ -2	↖ -2	↖ -1	← -0	← -3	← -4	← -5	← -6	← -7	← -8
L	↑ -3	↖ -3	↖ -3	← -2	↖ -2	← -1	← -2	← -3	← -4	← -5	← -6
I	↑ -4	↖ -4	↑ -4	↑ -3	↑ -1	↖ -1	↖ -2	↖ -1	↖ -4	↖ -5	↖ -6
C	↑ -5	↖ -3	← -4	↑ -4	↑ -2	↖ -2	↖ -0	← -1	← -2	← -3	← -4
A	↑ -6	↑ -4	↖ -4	↖ -5	↑ -3	↖ -1	↑ -1	↖ -1	← -0	← -1	← -2
N	↑ -7	↑ -5	↖ -5	↖ -5	↑ -4	↑ -2	↖ -2	← -0	↖ -2	← -1	← -0

```
COELACANTH
-PELICAN--
```

Where was the biology in that?

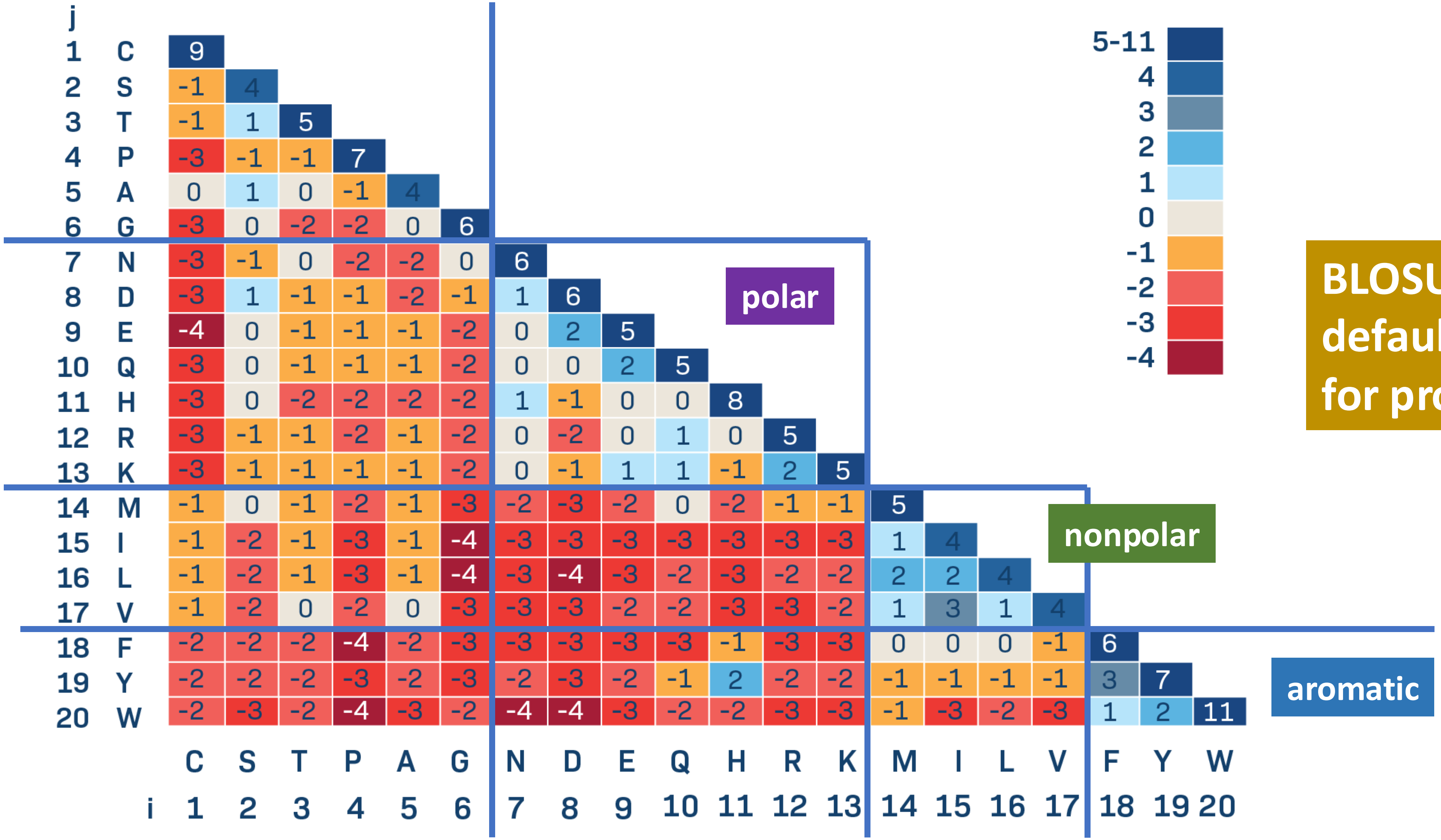
Honest question. Where do ***you*** think it was?

Pairwise sequence alignment

- We have two words we want to align:
  - **COELACANTH**
  - **PELICAN**
- **We have a scoring scheme:**
  - **Matching identical letters scores +1**
  - **If we have different letters aligned, that scores -1**
  - **Inserting a gap in the alignment scores -1**
- All sequence alignment is a mathematical operation: *the maximization of an alignment score*

This was all the biology:
What score do we give the same/ similar symbols in the alignment?
Represents evolutionary pressure on the sequence

BLOSUM62 Substitution Matrix



j																					
1	C	9																			
2	S	-1	4																		
3	T	-1	1	5																	
4	P	-3	-1	-1	7																
5	A	0	1	0	-1	4															
6	G	-3	0	-2	-2	0	6														
7	N	-3	-1	0	-2	-2	0	6													
8	D	-3	1	-1	-1	-2	-1	1	6												
9	E	-4	0	-1	-1	-1	-2	0	2	5											
10	Q	-3	0	-1	-1	-1	-2	0	0	2	5										
11	H	-3	0	-2	-2	-2	-2	1	-1	0	0	8									
12	R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
13	K	-3	-1	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
14	M	-1	0	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
15	I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
16	L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
17	V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
18	F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
19	Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
20	W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

BLOSUM62: a common default scoring scheme for protein sequences

University of
Strathclyde
Science

Next Week’s Group Meetings

Tuesday 20th October 13:30 HW324

Thursday 23rd October 10:30 HW324

Topics to Discuss at Next Meeting

- Tell me about your protein
    - What scientific questions do you think you could address in your research? (This could be inspired by something you’ve read in a paper describing your protein – or something more general)

- Phylogenetic Trees?
- What would you like to cover?

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