Final ‘Find a Gene’ Project

Renee Zuhars (PID A17329856)

Question 1

Beginning my search, I knew I wanted to limit my organism to some kind of fungus, because they are very understudied and I think their diversity and versatility are fascinating!

I decided to narrow my search to those proteins that help the fungus *Ophiocordyceps Unilateralis* “zombify” a host insect by taking over the hosts’ neurological systems, eventually killing the host.

protein name: serine/threonine-protein kinase MAK, partial

species: *Ophiocordyceps Unilateralis*

accession number: ADI72911.1

function: The role of MAK-like kinases in this species is to induce behavioral changes in the host by interfering with Mitogen- Activated Protein Kinase signaling pathways. (ChatGPT)

Figure 1: Ophiocordyceps Unilaterialis

Question 2

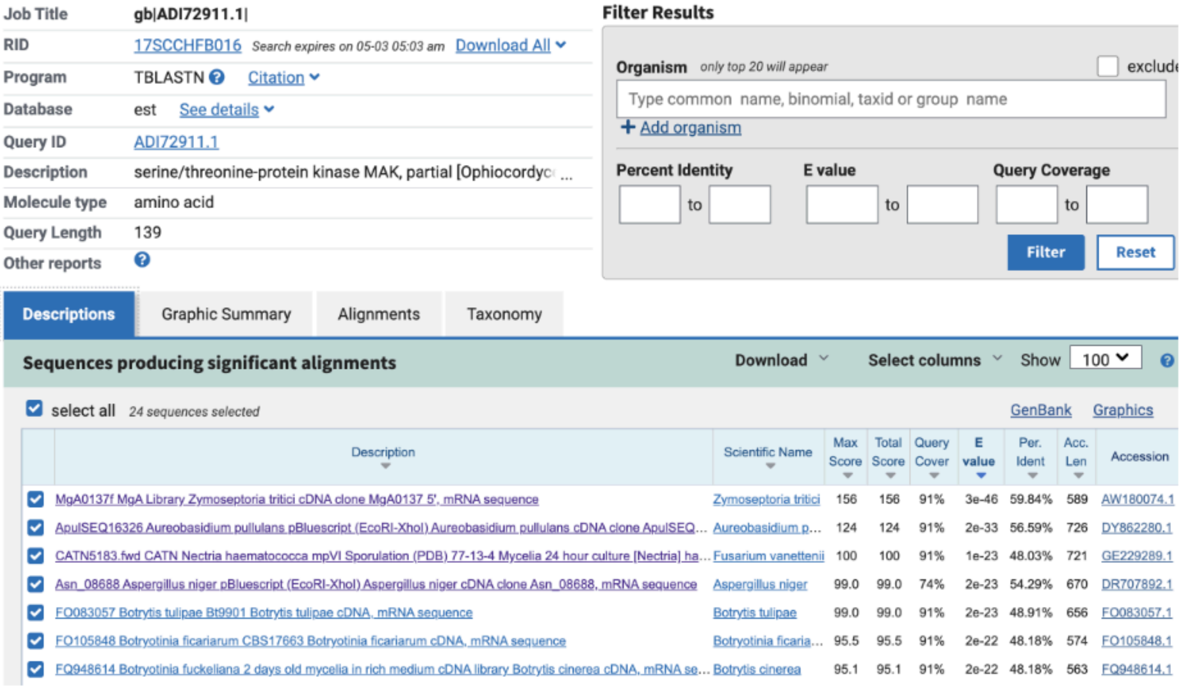
Attempting to find a homologous protein:

blast method: NCBI tblastn

database: est

limits/restrictions: none

My BLAST results were as followed:



Job Title	gb|ADI72911.1|
RID	17SCCHFB016 *Search expires on 05-03 05:03 am* Download All
Program	TBLASTN Citation
Database	est See details
Query ID	ADI72911.1
Description	serine/threonine-protein kinase MAK, partial [Ophiocordyc ...
Molecule type	amino acid
Query Length	139
Other reports	

Sequences producing significant alignments

☑ select all *24 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	MgA0137f MgA Library Zymoseptoria tritici cDNA clone MgA0137 5', mRNA sequence	Zymoseptoria tritici	156	156	91%	3e-46	59.84%	589	AW180074.1
☑	ApulSEQ16326 Aureobasidium pullulans pBluescript (EcoRI-XhoI) Aureobasidium pullulans cDNA clone ApulSEQ...	Aureobasidium p...	124	124	91%	2e-33	56.59%	726	DY862280.1
☑	CATN5183.fwd CATN Nectria haematococca mpVI Sporulation (PDB) 77-13-4 Mycelia 24 hour culture [Nectria] ha...	Fusarium vanettenii	100	100	91%	1e-23	48.03%	721	GE229289.1
☑	Asn_08688 Aspergillus niger pBluescript (EcoRI-XhoI) Aspergillus niger cDNA clone Asn_08688, mRNA sequence	Aspergillus niger	99.0	99.0	74%	2e-23	54.29%	670	DR707892.1
☑	FO083057 Botrytis tulipae Bt9901 Botrytis tulipae cDNA, mRNA sequence	Botrytis tulipae	99.0	99.0	91%	2e-23	48.91%	656	FO083057.1
☑	FO105848 Botryotinia ficariarum CBS17663 Botryotinia ficariarum cDNA, mRNA sequence	Botryotinia ficaria...	95.5	95.5	91%	2e-22	48.18%	574	FO105848.1
☑	FQ948614 Botryotinia fuckeliana 2 days old mycelia in rich medium cDNA library Botrytis cinerea cDNA, mRNA se...	Botrytis cinerea	95.1	95.1	91%	2e-22	48.18%	563	FQ948614.1

Figure 2: BLAST results, original query

Alignment of choice: MgA0137f MgA Library Zymoseptoria tritici cDNA clone MgA0137 5', mRNA sequence

E Value: 3e-46

Percent Identity: 59.84%

Percent Coverage: 91%

Question 3

Here is some information about the homolog I am looking into:

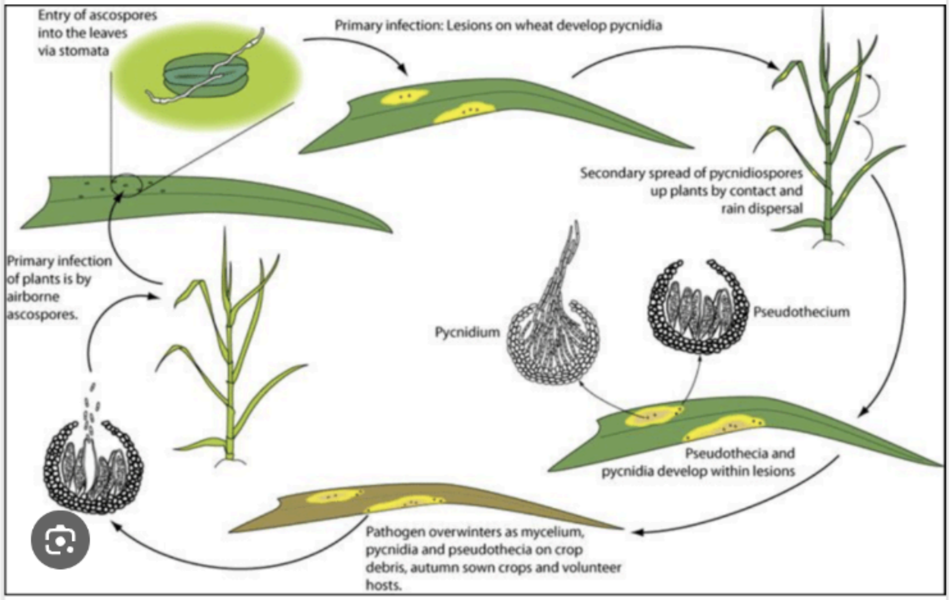

Entry of ascospores into the leaves via stomata
Primary infection: Lesions on wheat develop pycnidia
Secondary spread of pycnidiospores up plants by contact and rain dispersal
Primary infection of plants is by airborne ascospores.
Pycnidium
Pseudothecium
Pseudothecia and pycnidia develop within lesions
Pathogen overwinters as mycelium, pycnidia and pseudothecia on crop debris, autumn sown crops and volunteer hosts.

Figure 3: Zymoseptoria Triciti

Name: MgA0137f MgA Library Zymoseptoria tritici cDNA clone MgA0137 5', mRNA sequence

Species Derived from: Zymoseptoria tritici : this is a pathogenic fungus that attacks wheat plants. It is resistant to multiple fungicides, and causes septoria leaf blotch.

FASTA format sequence, translated using EMBOSS Transeq:

AW180074.1_1 MgA0137f MgA Library Zymoseptoria tritici cDNA clone MgA0137 5', mRNA sequence

```
RQLSVNSQGNHYAEIHRQEAERALVGASALKSPTGSQRESFFSHLRKRARRLSGRNSGVI
TPSMDAMETSAGCVPWAANKQTTFDTHSIASAAADPSSDPNFAELDRALQSVRYSLDAAA
NATQQARKPTNRVVEQPSLKRHHSLPHGVRHKTNPTTVYHDEH*STPRAADTRPPTKKKN
SRRSHELSASRRTAFSX
```

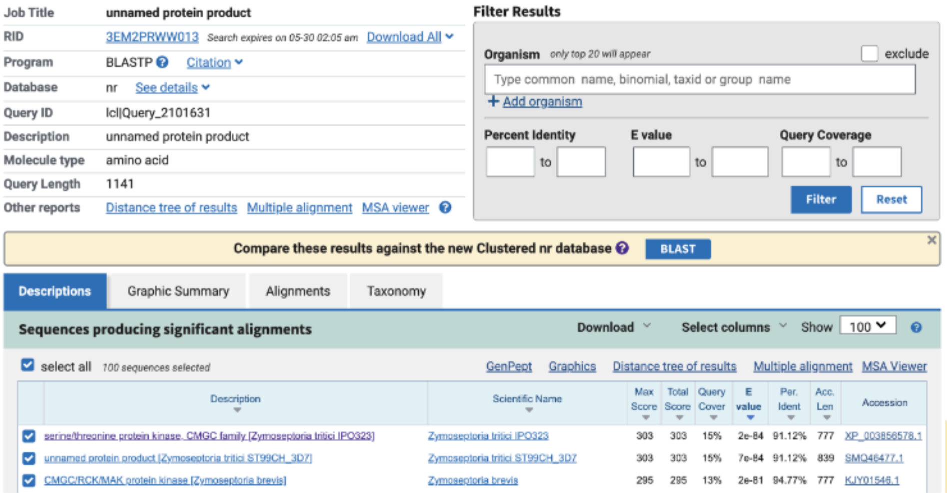
```
DNYQSTHRAITTPKFTGRKLSVLWLAQALSSHRLAAKEKASSLICARGREDFPAATQVSS
HLQWMLWKPALGAFLGLLTNKPPSTPTRSRLPQPIRHQTPISLSWIVHCKVYDTAWMPPR
TRLNKLGSLRTALSNHHSVTTRFLTALDTRPTQPPYTTTSTEARHEQPIQDPRRRRRI
LDEVMNSAHLAARRSR
```

```
TIISQLTGQSLRRNSPAGSACSGWRKRSQVTDWQPKRKLLLSSAQEGEKTFRPQLRCHH
TFNGCYGNQRWVRSLGCQTNHLRHPLDRVCRSRSVIRPQFRAGSCTAKCTIQPGCRRE
RDSTSEAYEPRSATIIEASPLASSRRTQDQPNHRIPRRALKHATSSRYKTPDEEEEF
STKSTQRISPHGVLX

RERRAARCAEFMTSSRILLLRRGSCIGCSWRASVLVVVYGGWVGLVSNAVRKRVVTLQ*W
LLNYAVRRLPSLLSRVRGGIQAVSYTLQCTIQLSEIGV**RIGCGRRDRVGVEGGLFVSS
PRNAPSAGFHSIHRCDDTVAAGKSSRPLAQMREEAFSLAASRLESACANQSTLSFLP
VNFGVVIALVDLS

SRTPCGEMRVHDFVENSSSSSGVLYRLLVACFSARRGIRWLGWSCVRREEASGDASMM
VAQLRGSASLVESRSRRHPGCIVHFAVHDPAQRNWGLMTDRLRQTRSSGCRRWFVCQ
PKERTQRWFPHPLKVHLSCGRKVFSPSCADERRSFLFGCQSVTERLRQPEHAQLPA
GEFRRSDCPVSLIIVX

ENAVRRDALSSLRREFFFFVGGLVSAARGVLQCSSWYTVVGLVLCLTPGSEWRFNDG
CSTTRFVGFLACVAFAAASRLYRTLCSARSSSAKLGSDDGSAAADAIEWVSKVVCLLAA
QGTHPALVSIASIEGVMTPELRPESLLALLRR*EKKLSLWLPVGDLRALAPTRARSASCR
ISALPCELTDNCR
```

Question 4

To determine if this protein is novel:

I used NCBI blastp, in the nr database.



Job Title	unnamed protein product
RID	3EM2PRWW013 *Search expires on 05-30 02:05 am* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_2101631
Description	unnamed protein product
Molecule type	amino acid
Query Length	1141
Other reports	Distance tree of results Multiple alignment MSA viewer

Sequences producing significant alignments

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	serine/threonine protein kinase, CMGC family [Zymoseptoria tritici IPO323]	Zymoseptoria tritici IPO323	303	303	15%	2e-84	91.12%	777	XP_003856578.1
☑	unnamed protein product [Zymoseptoria tritici ST99CH_3D7]	Zymoseptoria tritici ST99CH_3D7	303	303	15%	7e-84	91.12%	839	SMQ46477.1
☑	CMGC/RCK/MAK protein kinase [Zymoseptoria brevis]	Zymoseptoria brevis	295	295	13%	2e-81	94.77%	777	KJY01546.1

Figure 4: BLAST results, novel query

There is no match with 100% identity!

Question 5

I will use MUSCLE at EBI to produce a multiple sequence alignment of the following proteins:

My novel protein

>Zymoseptoria unknown protein (novel protein from BLAST results)

My original sequence for ophiocordyceps:

>ADI72911.1 serine/threonine-protein kinase MAK, partial [Ophiocordyceps unilateralis]

Other proteins of interest, based on BLAST results of original and novel proteins:

1. >KAK4508075.1 hypothetical protein PRZ48_001812 [Zasmidium cellare]
2. >XP_047755397.1 Sporulation protein kinase pit1 [Fulvia fulva]
3. >KJY01546.1 CMGC/RCK/MAK protein kinase [Zymoseptoria brevis]
4. >KAK4981489.1 hypothetical protein LTR28_003093, partial [Elasticomyces elasticus]
5. >KAI5369935.1 putative serine/threonine-protein kinase, active [Septoria linicola]

Alignment

Here is the alignment I obtained after running the above 7 proteins through MUSCLE at EBI (labeled by species), displayed in a code chunk because PDF formatting was giving me issues

```
"
Zymoseptoria        LLRRGSCIGCSWRASVLVVVYGGWVGLVSNAVRKRVVTLQWLLNYAVRRLPSLLSRVRGG
Elasticomyces       -----------------VASHGNHYADAHRHEAEQALN----------------------
Zymoseptoria_br     -----------------VNSQGNHYAEIHRQEAERALV----------------------
Ophiocordyceps      ------------------------------QEAERALS----------------------
Septoria            -----------------VNSQGNHYAELHRQEAERALN----------------------
Zasmidium           -----------------VNSQGNHYADIHRQEAERALT----------------------
Fulvia              -----------------VNSQGNHYADMHRQEAERALT----------------------

Zymoseptoria        IQAVSYTLQCTIQLSEIGVRIGCGRRDRVGVEGGLFVSSPRNAPSAGFHSIHRCDDTVAA
Elasticomyces       -----------------------------GRNG---LASPTSSQRGSFFAHLRKRARRLS
Zymoseptoria_br     -----------------------------GASA---LKSPTGSQRESFFSHLRKRARRLS
Ophiocordyceps      -----------------------------GANG---RKSPTGTLLESFFSHLRKRARRLS
Septoria.           -----------------------------GASG---LKSPTGSQRESFFSHLRKRARRFS
Zasmidium           -----------------------------GANG---LKSPTGSQRESFFSHLRKRARRLS
```

```
Fulvia              -----------------------------GATG---LQSPTGSQRESFFSHLRKRARRLS

Zymoseptoria        GKSSRPLAQMREEAFSLAASRLESACANQSTLSFLPVNFGVVIALVDLSSRTPCGEMRVH
Elasticomyces       GRNQAPVSPSVDD---------------------IEASAG-------------CAP----
Zymoseptoria_br     GRNSGVITPSMDA---------------------METNAG-------------CVP----
Ophiocordyceps      GRNQGPMSPGAED---------------------LEANAG-------------CAP----
Septoria            GKPSGLASPTAED---------------------MEANVG-------------CAP----
Zasmidium           GRNQGPMSPGAED---------------------IEASVG-------------CAP----
Fulvia              GRNSGPMSPSAED---------------------AEANVG-------------CAP----

Zymoseptoria        DFVENSSSSSGVLYRLLVACFSARRGIRWLGWSCVRREEASGDASMMVAQLRGSASLVES
Elasticomyces       -------------------------------WAS-NRQSMAIE-----------------
Zymoseptoria_br     -------------------------------WAA-NKQPVF-D-----------------
Ophiocordyceps      -------------------------------WSS-NRGSIQ-E-----------------
Septoria            -------------------------------WTT-NRQSIP-D-----------------
Zasmidium           -------------------------------WSTNNRGSIQ-E-----------------
Fulvia              -------------------------------WASNNRQSVQ-E-----------------

Zymoseptoria        RSRRHPGCIVHFAVHDPAQRNWGLMTDRLRQTRSSGCRRWFVCQPKERTQRWFPHPLKVH
Elasticomyces       -----SLAITTHATDPSSDPNFAELDRALQNVRYSLDAGSYSNNNVQK------PVQKV-
Zymoseptoria_br     -----THSVASAAADPSSDPNFAELDRALQSVRYSLDAAANATQQARK------PTNRV-
Ophiocordyceps      -----PQPIEAVASDPSSDPNFAELDRALQNVRYSLDATANTSNNQPK------HPTKM-
Septoria            -----AQAIAPTAADPSVDPNFAELDRALQSVRYSLDATAGAMPTQPK------PPVKM-
Zasmidium           -----PQPIAPAAADPSTDPNFAELDRALQNVRYSLDAAANPANMQPK------HPTKM-
Fulvia              -----PQSIASVAVDPSSDPNFAELDRALQNVRYSLDAAAGAANPQPK------QPTKM-

Zymoseptoria        LSCGRKVFSPSCADERRSFLFGCQSVTERLRQPEHAQLPAGEFRRSDCPVSLIIVXENAV
Elasticomyces       ---------PSNPMLKR-----------------HHSLPFGQDERISPVPAVNGPISSRT
Zymoseptoria_br     ---------VSNPSLKR-----------------HHSLPHGVDDKTQ-------ANHRIP
Ophiocordyceps      ---------ASNPSLKR-----------------HQSSHSG-------------------
Septoria            ---------ASNPALKR-----------------HHSLPYGKEEL---------SVVNRT
Zasmidium           ---------PSNQSLKR-----------------HHSLPYGKEEIMSQT---GGSTSNRT
Fulvia              ---------ASNPTLKR-----------------HHSVPCSKEEVTSNP-----SMANRT

Zymoseptoria        RRDALSSLRREFFFFVGGLVSAARGVLQCSSWYTVVGLVLCLTPGSEWRFNDGCSTTRFV
Elasticomyces       RRS------------------------------------VRQAPHPGHRYETPDEEEELL
Zymoseptoria_br     RRA------------------------------------LKHATSS--RYETPDEEEELL
Ophiocordyceps      ------------------------------------------------------------
Septoria            RRS------------------------------------VKQAPSN-IRYETPCEEDELL
Zasmidium           RRS------------------------------------LRHAPSS--RYETPCEEDELL
Fulvia              RRS------------------------------------LRHAPSS--RYETPCEEDELL
```

```
Zymoseptoria        GFLACVAFAAASRLYRTLCSARSSSAKLGSDDGSAAADAIEWVSKV--VCLLAAQGTHPA
Elasticomyces       DEVLASAHRAARRLDRYIQQDNSPLPSVTSQQERARPPVQQVTSDP--GCFVPYLTPSPS
Zymoseptoria_br     DEVMTSAHLAARRL---------------DNEQLSRPPLPHVISEPVTVYTAPYLTPSHS
Ophiocordyceps      --------------------------------------------------------PAPS
Septoria            DEAIASAHQAVTRL-------------DNGITQPARPHLPHVTSEP--TYNVPYLTPSPS
Zasmidium           DEALASVHAAATRL------DKGTS-NVAGTYQPSRPPIGHVTSEP--VYPAPYLTPSHS
Fulvia              DEALSNAHQAAQNL------DNAPSMNTTATYQPARPLLPQAISEP--AYAAPYLTPSPS

Zymoseptoria        LVSIASI
Elasticomyces       KDRNG--
Zymoseptoria_br     KDQMSLD
Ophiocordyceps      RKP----
Septoria            KDHMAVD
Zasmidium           KDQMNVT
Fulvia              KDQMNIS
"
```

```
[1] "\nZymoseptoria        LLRRGSCIGCSWRASVLVVVYGGWVGLVSNAVRKRVVTLQWLLNYAVRRLPSLLSRVRGG\nEla
```

Question 6

Given the alignment above, I can now create a phylogenetic tree using the EBI's Simple Phylogeny tool.

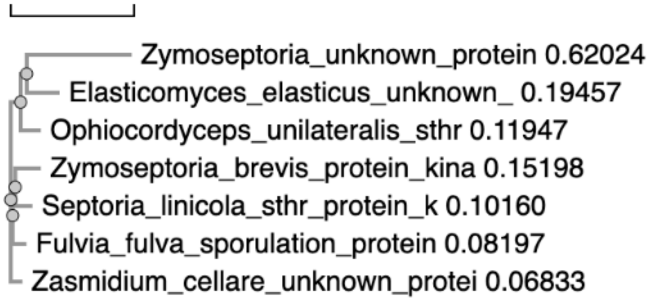

Zymoseptoria_unknown_protein 0.62024
Elasticomyces_elasticus_unknown_ 0.19457
Ophiocordyceps_unilateralis_sthr 0.11947
Zymoseptoria_brevis_protein_kina 0.15198
Septoria_linicola_sthr_protein_k 0.10160
Fulvia_fulva_sporulation_protein 0.08197
Zasmidium_cellare_unknown_protei 0.06833

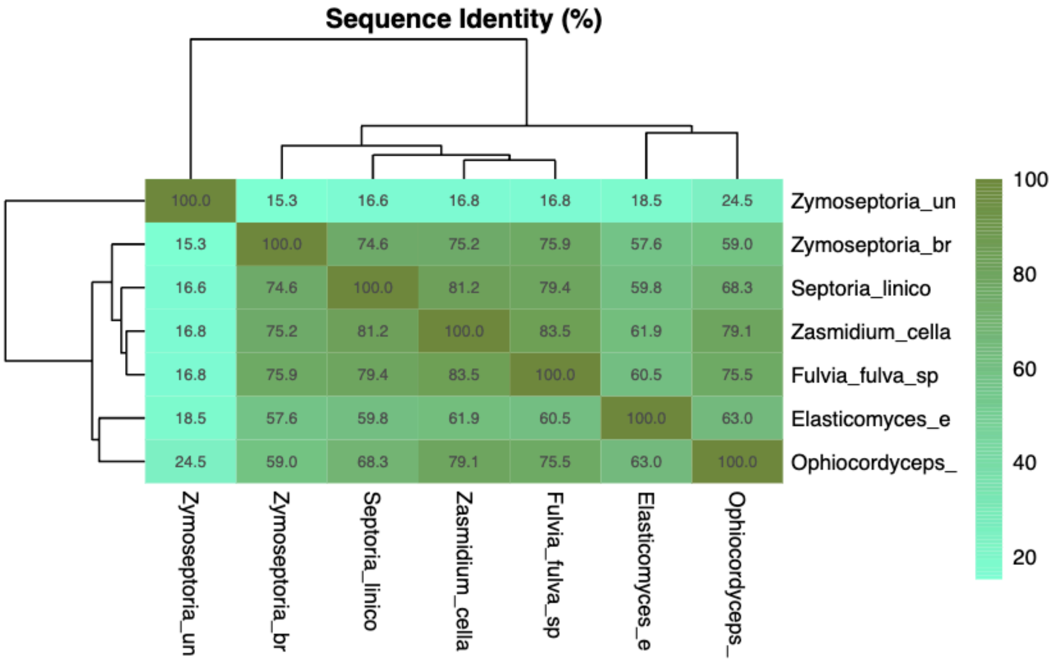
Figure 5: Phylogenetic tree for various fungal proteins

Question 7

To generate a heatmap, the msa must be in fasta format. One of the MUSCLE outputs is ‘the alignment in FASTA format converted by Seqret’, so I used that.

```
library(bio3d)

# Read multiple sequence alignment into R

msa <- read.fasta("finalseq.fasta")

# Calculate sequence identity matrix

seq_id_m <- seqidentity(msa)
```

```
library(pheatmap)

#Generate a heatmap of the msa (nicer enhancements courtesy of ChatGPT)

idm_percent <- seq_id_m * 100
rownames(idm_percent) <- substr(msa$id, 1, 15)
colnames(idm_percent) <- substr(msa$id, 1, 15) #shortening the labels

pheatmap(idm_percent,
         display_numbers = TRUE,       # show exact % in each cell
         number_format = "%.1f",       # format: 1 decimal place
         color = colorRampPalette(c("aquamarine", "darkolivegreen4"))(100),  # color gradien
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean",
         main = "Sequence Identity (%)",
         fontsize = 8,
         border_color = NA)            # cleaner look without gridlines
```



	Zymoseptoria_un	Zymoseptoria_br	Septoria_linico	Zasmidium_cella	Fulvia_fulva_sp	Elasticomyces_e	Ophiocordyceps_
Zymoseptoria_un	100.0	15.3	16.6	16.8	16.8	18.5	24.5
Zymoseptoria_br	15.3	100.0	74.6	75.2	75.9	57.6	59.0
Septoria_linico	16.6	74.6	100.0	81.2	79.4	59.8	68.3
Zasmidium_cella	16.8	75.2	81.2	100.0	83.5	61.9	79.1
Fulvia_fulva_sp	16.8	75.9	79.4	83.5	100.0	60.5	75.5
Elasticomyces_e	18.5	57.6	59.8	61.9	60.5	100.0	63.0
Ophiocordyceps_	24.5	59.0	68.3	79.1	75.5	63.0	100.0

Quick note: because the sequence of my novel protein is quite long (over 2000 characters), some shortening is needed before it can be processed by many of the tools that I'll be using going forward.

I was unsure of the best way to shorten my sequence, so I asked ChatGPT to do it for me. It identified a central 1000-amino acid segment that avoids the N-terminal signal peptide and C-terminal tail, which are less structured and may not be essential for the core function.

The resulting novel sequence, printed below, will be saved as "zymo_novel".

```
zymo_novel <- "VITPSMDAMETSAGCVPWAANKQTTFDTHSIASAAADPSSDPNFAELDRALQSVRYSLDAAANATQQARKPTNRVVE
```

Question 8

I am going to try to create a consensus for all my aligned sequences (msa), using the Bio3d package `consensus()`.

```
cons <- consensus(msa)
```

Because this sequence appears to have a lot of gaps in it, I will use the protein sequences that I aligned instead. I have saved them as text in some of the code chunks above.

I will use the **blast.pdb()** function to search the seven fungal protein sequences against the pdb database.

```
# blast.pdb(zymo_novel) yielded no results

# blast.pdb(ophiocordyceps) yielded no results

# blast.pdb(zasmidium)
```

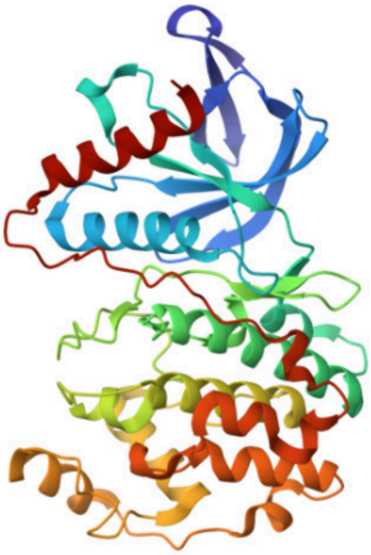


Figure 6: pdb match 1, 7W5C

Sequence details

PDB ID: 7W5C

E Value: 6.985e-33

Experimental technique: X-RAY DIFFRACTION

Resolution: 2.20 Å

Source organism: Arabidopsis thaliana

% identity: 34%

```
# blast.pdv(fulvia)
```

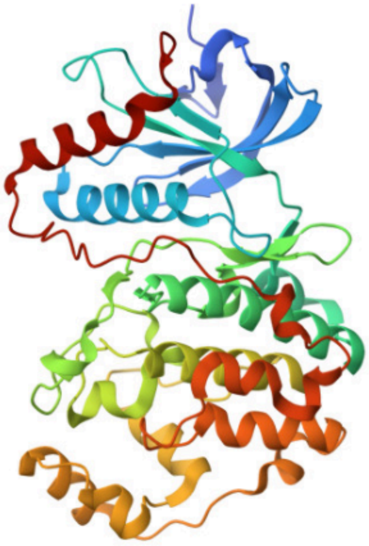


Figure 7: pdb match 2, 7E75

Sequence details

PDB ID: 7E75

E Value: 2.141e-28

Experimental technique: X-RAY DIFFRACTION

Resolution: 2.48 Å

Source organism: Homo Sapiens

% identity: 34%

```
# blast.pdb(z_brevis)
```

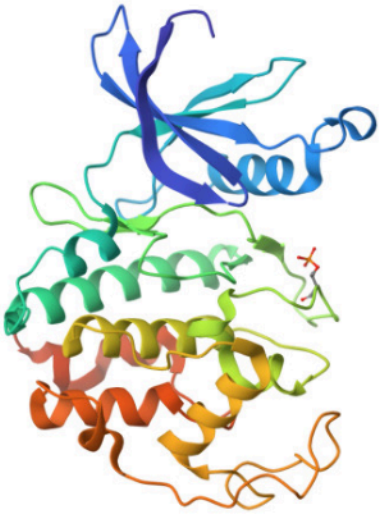


Figure 8: pdb match 3, 7NJ0

Sequence details

PDB ID: 7NJ0

E Value: 1.682e-40

Experimental technique: ELECTRON MICROSCOPY

Resolution: 3.60 Å

Source organism: Homo Sapiens

% identity: 34%

```
# blast.pdb(elasticomyces) yielded no results

# blast.pdb(septoria) yielded no results with a higher identity score than those above
```

Question 9

I will run "zymo_novel" through AlphaFold.

It took a couple hours, but my output included several files that were saved to my project directory.

I then ran "/test_62382/test_62382_unrelaxed_rank_001_alphafold2_ptm_model_2_seed_000.pdb" through the online Mol* viewer (the conserved regions are rendered in spacefill):

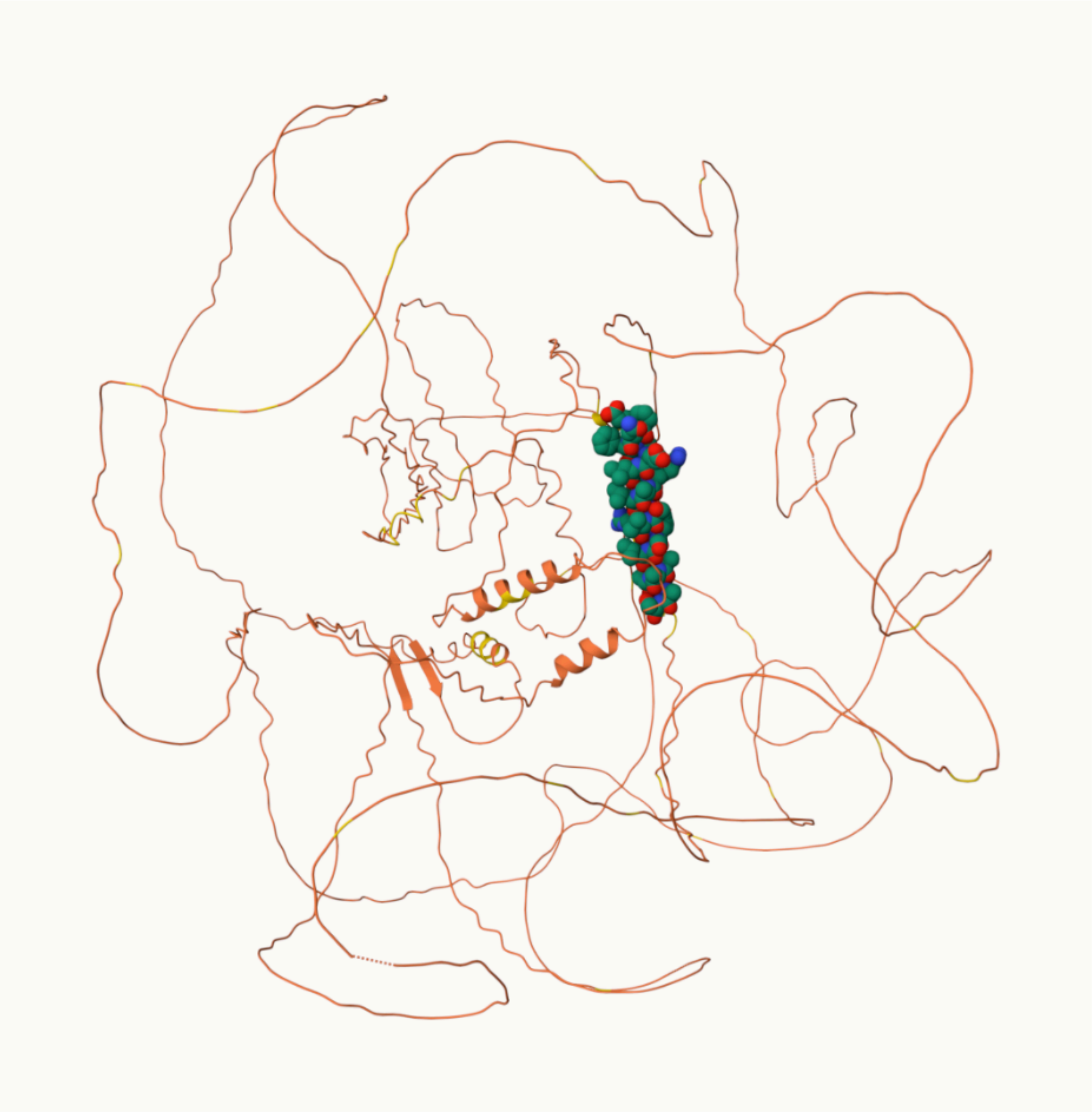


Figure 9: Final Unknown Protein Structure- colored by PLDDT confidence

Here it is without the spacefill, so the PLDDT coloring can be seen more clearly:

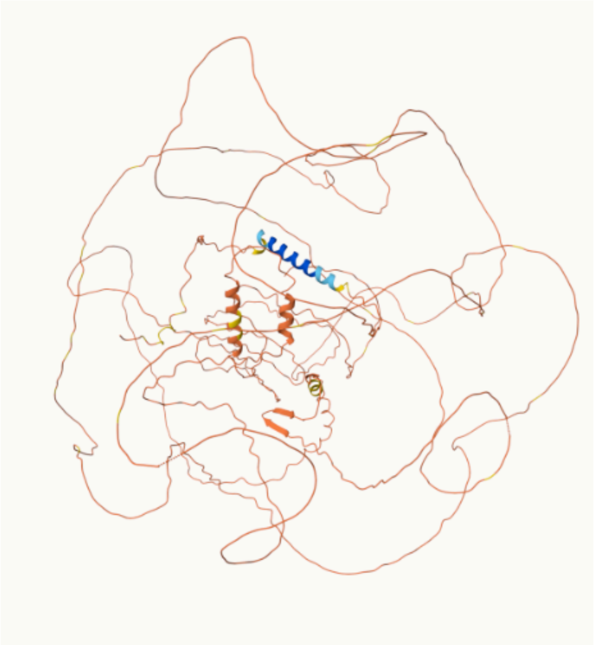
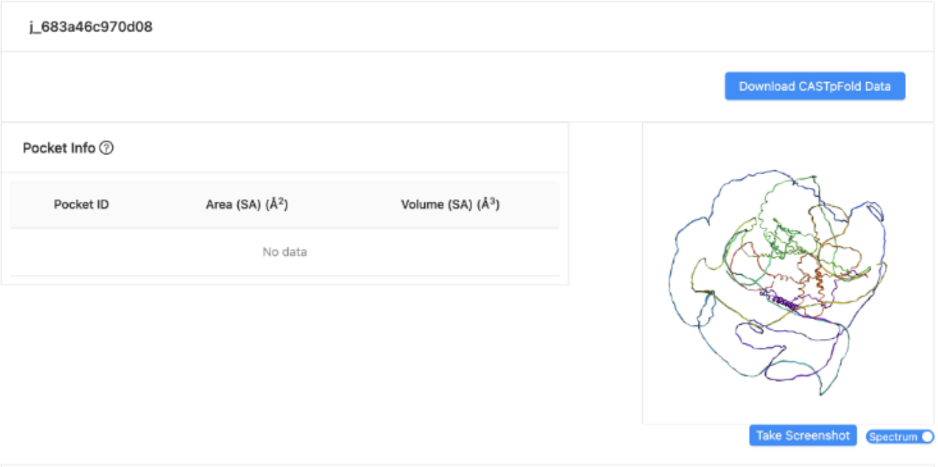


Figure 10: Final Protein Structure (no spacefill)

Question 10

(i) I attempted to use the CASTPfold server to predict binding sites in my protein, but got the following result:



j_683a46c970d08

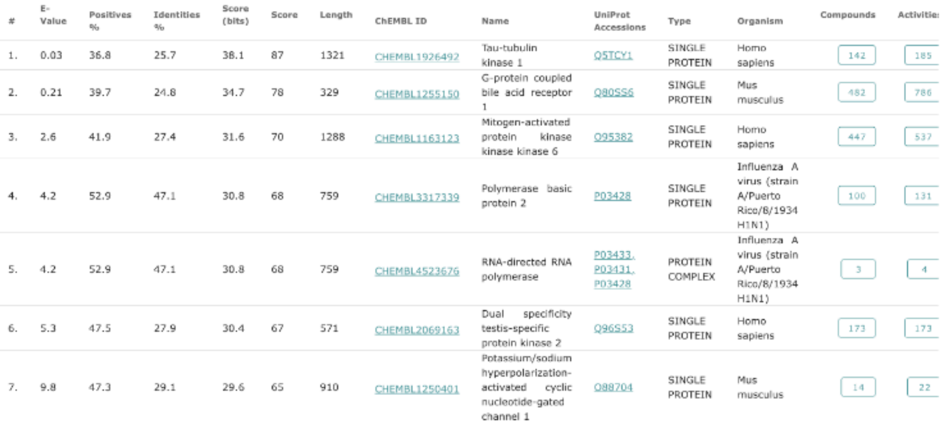
Pocket Info ⓘ

Pocket ID	Area (SA) (Å^2)	Volume (SA) (Å^3)
	No data	

Figure 11: CASTPfold for my novel protein

So, unfortunately my protein appears to have no small molecule binding sites.

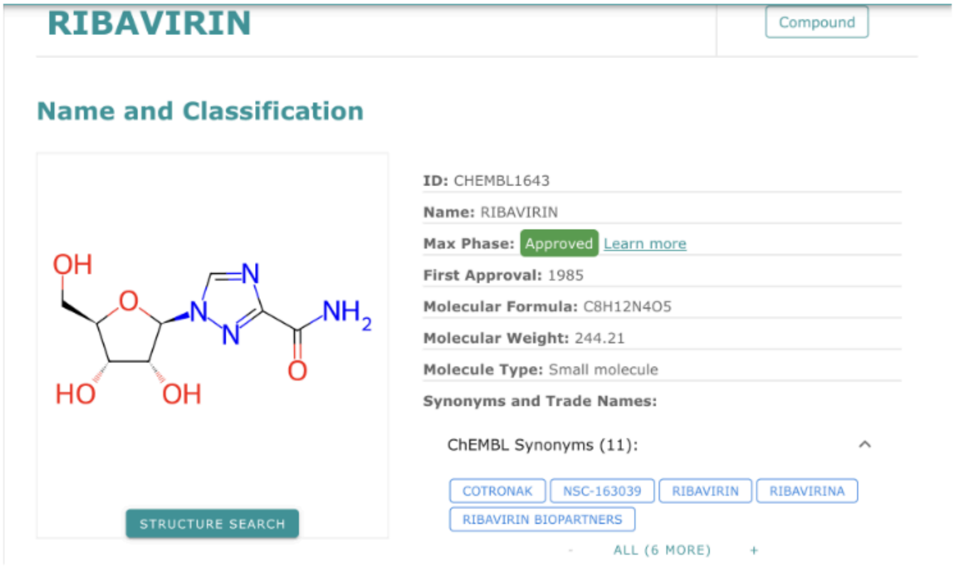
(ii) I then performed a CHEMBEL "target" search of my novel sequence. The resulting list consisted of 7 targets. I decided to focus on targets 4 and 5 (shown below) because they have the highest percent identity.



#	E-Value	Positives %	Identities %	Score (bits)	Score	Length	ChEMBL ID	Name	UniProt Accessions	Type	Organism	Compounds	Activitie
1.	0.03	36.8	25.7	38.1	87	1321	CHEMBL1926492	Tau-tubulin kinase 1	Q5TCY1	SINGLE PROTEIN	Homo sapiens	142	185
2.	0.21	39.7	24.8	34.7	78	329	CHEMBL1255150	G-protein coupled bile acid receptor 1	Q80SS6	SINGLE PROTEIN	Mus musculus	482	786
3.	2.6	41.9	27.4	31.6	70	1288	CHEMBL1163123	Mitogen-activated protein kinase kinase kinase 6	O95382	SINGLE PROTEIN	Homo sapiens	447	537
4.	4.2	52.9	47.1	30.8	68	759	CHEMBL3317339	Polymerase basic protein 2	P03428	SINGLE PROTEIN	Influenza A virus (strain A/Puerto Rico/8/1934 H1N1)	100	131
5.	4.2	52.9	47.1	30.8	68	759	CHEMBL4523676	RNA-directed RNA polymerase	P03433, P03431, P03428	PROTEIN COMPLEX	Influenza A virus (strain A/Puerto Rico/8/1934 H1N1)	3	4
6.	5.3	47.5	27.9	30.4	67	571	CHEMBL2069163	Dual specificity testis-specific protein kinase 2	Q96S53	SINGLE PROTEIN	Homo sapiens	173	173
7.	9.8	47.3	29.1	29.6	65	910	CHEMBL1250401	Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 1	O88704	SINGLE PROTEIN	Mus musculus	14	22

Figure 12: My CHEMBL search

Target 4 appears to have the most useful data, as it has a defined compound that is effective against it, called Ribavirin.



RIBAVIRIN

Name and Classification

ID: CHEMBL1643

Name: RIBAVIRIN

Max Phase: Approved Learn more

First Approval: 1985

Molecular Formula: C8H12N4O5

Molecular Weight: 244.21

Molecule Type: Small molecule

Synonyms and Trade Names:

ChEMBL Synonyms (11):

COTRONAK | NSC-163039 | RIBAVIRIN | RIBAVIRINA | RIBAVIRIN BIOPARTNERS

Figure 13: Details about Ribavirin

This compound functions as an RNA polymerase inhibitor.

(iii) Overall, I do not believe my protein is druggable. My sequence is very long, and it was cut down so alphafold could process it. The resulting predicted structure did not have many conserved regions that I could use as starting points for further research. My CASTPfold search also did not yield any pockets. My CHEMBL searches yielded interesting results, but the highest percent identity listed was 47.1%, which does not place much confidence in the results. Perhaps if the whole sequence was run though a server with the capability to process it, a more accurate structure could be predicted.