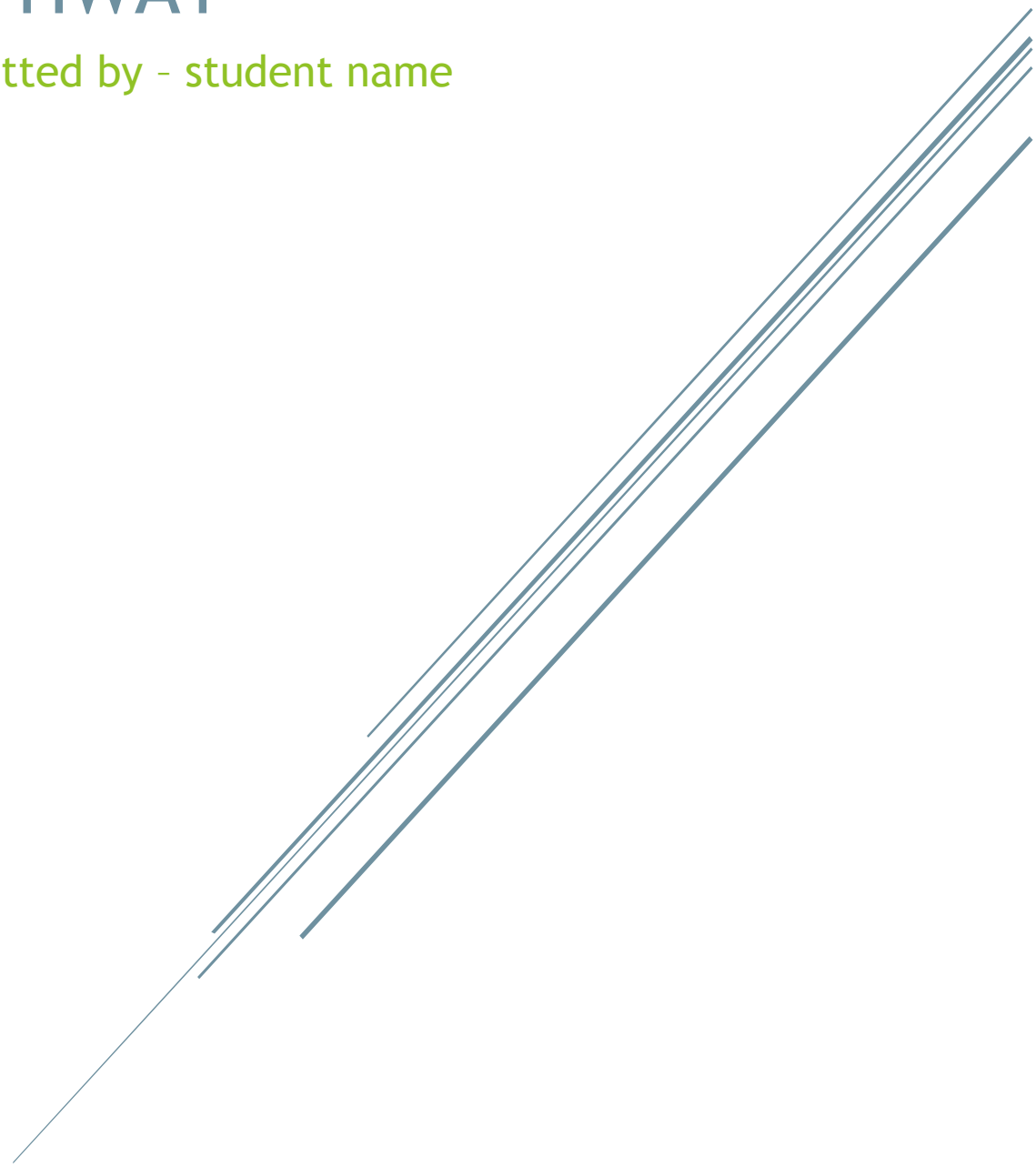
FORMAMIDASE – UNDERSTANDING THE PROTEIN INTERACTION OF THE HIDDEN ENZYME IN THE KYNURENINE PATHWAY

Submitted by – student name



REGISTRATION NUMBER - xxxxxx

INTRODUCTION

The enzyme formamidase catalyzed the chemical reaction between formamide and water. This enzyme belongs to the hydrolase family. Despite the fact that the KYNFA enzymatic activity (also called as aryl formamidase) linked with the second stage of the kyn pathway was identified in early research. Also, the particular gene responsible for it was unknown for a very long time. Recently, studies have shown that this is an enzyme that could be used in different species. (Sorci et al., 2010) The main function of formamidase is to catalyze the hydrolysis of N-formyl-L-kynurenine to L-kynurenin, the second step is tryptophan breakdown. Kynurenine can be oxidized further to form nicotinic acid, NADH and NADPH. The main pathways involved are the shikimate pathway and the kynurenine pathway. The enzyme commission number is EC 3.5.1.49 – formamidase. (Wu et al., 2005) The disease that formamidase causes is tuberculosis and other neurodegenerative diseases but in the recent studies there has not been much research going on about what type of diseases formamidase causes. (Dykhno et al., 1964) The pathways are poorly recognized in fungi and have more evidence shown in bacteria and mammals.

Akanthomyces lecanii belongs to the *Cordycipitaceae, Hypocreales, Sordariomycetes,* and *Ascomycota* families of fungi. *Cordyceps confragosa, Torrubiella confragosa, Lecanicillium lecanii*, and other names have been used in the past. This species is interesting because it may infect a variety of insects (especially scale insects), has a wide geographic distribution. This fungus also has anti-parasitic nematode and anti-fungal disease properties (Zhang et al., 2020). This strain was first isolated in China.

Protein localization and protein-protein interaction in a cell are directly linked to the function and knowing the subcellular location of proteins is very essential for obtaining a complete picture of the cells (Shekari et al., 2014). So, understanding the protein subcellular localization is very important for both the function of the protein and the overall organization of the cell. The localization can easily be seen in large number of proteins (Scott et al., 2005). There are three main reasons why protein-protein interaction is important. a) The function and behavior of the protein can be understood, b) Prediction of the biological processes that are not known can be retrieved, c) The characteristics of protein complexes and the pathways can be retrieved or also can be used as a draft map. Understanding how gene functions and regulatory are integrated at the organismal level requires studying the interactome, which is the entire network of molecular physical interactions between biological entities in cells and organisms (Sevimoglu & Arga, 2014). Hence, protein-protein interaction and protein localization give more insight on the pathway that formamidase enzyme is involved in. Therefore, the research question that the report aims to solve: **Studying the protein-protein interaction of formamidase in Akanthomyces lecanii and understanding the reasoning behind them.**

METHODOLOGY

ASSEMBLY OF SEQUENCE - The two reads, forward and reverse transcript were downloaded from the directory and then uploaded in the Galaxy server and were assembled to contigs using genome assembler (Version 1.0) (Bankevich et al., 2012) with default parameters.

DNA ANNOTATION - The reference transcript was used to search against the protein database using BLASTX (Altschul et al., 1990) to give the protein of interest, and while doing the process in blastx, the default parameters were used. The protein was selected on the criteria of having the highest identity percentage and E value. BLOSUM62 matrix was used as a scoring matrix for building query from BLAST search.

PROTEIN DOMAINS – The protein sequence was uploaded to protein domain database search engines like PFAM (Version 34.0) (Mistry et al., 2021) and PROSITE (Release 2021_03 of 02-Jun-2021) (Sigrist et al., 2010) separately with default search parameters.

TOPOLOGICAL SIGNALS AND PROTEIN LOCALISATION - Signal IP (version 5.0) (Almagro Armenteros et al., 2019) was used to predict the presence of signal peptides and the location of their cleavage sites in proteins. Fungi group was chosen and other parameters were set to default. Protein localization was predicted using WoLF PSORT (Horton et al., 2007) and Deeploc (v1.0) (Almagro Armenteros et al., 2017). The organism for these tools was set to fungi and the rest were default parameters. TMHMM (version 2.0) was used for the prediction of transmembrane helices in protein. String-db (version 11.5) (Szklarczyk et al., 2021) was used to find the protein-protein interactions.

MULTIPLE SEQUENCE ALIGNMENT AND PHYLOGENETIC TREE - The multiple sequence alignment and Phylogenetic tree were assessed using ClustalX (2.1)(Larkin et al., 2007) and FigTree (version 1.4.4) (FigTree,07/10/2021). Two trees were built, one for closely related species and one for distantly related species. Firstly, a BLAST search was done to find the homologous sequences. For distantly related species Uniprot/Swissprot was used and for closely related species protein nr database was used. For distantly related sequences, hits > 50% similar was used and for closely related sequences hits > 80% similarity were chosen. For the phylogenetic analysis a bootstrap NJ tree was built with the output format of bootstrap labels that were changed to node. The tree was then visualized using FigTree with node labels and bootstrap values.

HOMOLOGY MODELLING AND QUALITY CHECK - Due to the low E value (0) and high GA341 score (>0.7), models based on templates 2wkn.1. A was chosen from the Swissmodel (Waterhouse et al., 2018) website. There was only one model based on the template. The models were then downloaded and their appropriate PDB templates files were used. Using Pymol (V2.5.2 super) function alignment query, the template and model were visually and structurally aligned. ProSA (Wiederstein & Sippl, 2007) web tools and PROCHECK (Laskowski et al., 1993) web tools were used to do quality checks, and Ramachandran plots were drawn using default settings using ProSA web tools and PROCHECK web tools.

RESULT AND DISCUSSSION

DNA ASSEMBLY:

The forward and reverse reads were assembled using SPADES genome assembler. These reads were then assembled into a single contig of length 1234 nucleotides and the coverage of the assembly was 12.5454. Manually the calculations are shown in appendix (1)

DNA ANNOTATION USING BLASTX: The assembled sequence was then uploaded to BLASTX tool and the search was done with default parameters. The best hit matched the formamidase enzyme that belonged to the organism *Aknathomyces lecanii* which had a bit score value of 850 and an E value of 0 and has a similarity of 100%. The protein was 412 amino acids long.

DOMAIN IDENTIFICATION AND SIGNAL PEPTIDE:

DOMAIN IDENTIFICATION:

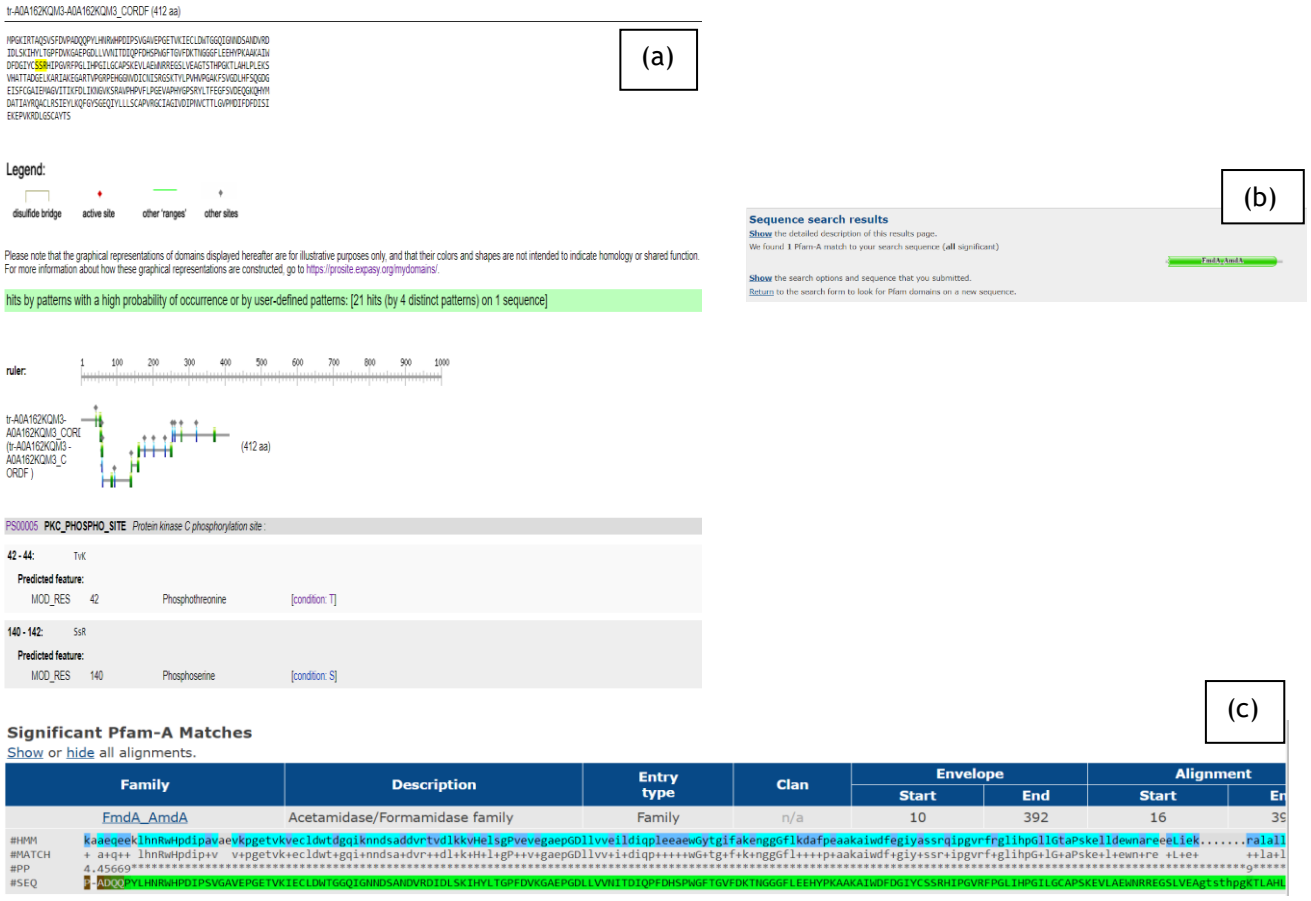


Fig1: a – Result page of prosite where 21 hits by 4 pattern hits were found. b & c – Results of significant domains obtained from Pfam.

There were no hits found on SMART. The PROSITE found 21 hits of the protein domain kinase C-phosphorylation site (location 42-44) and also many different domains like the N-myristylation site, N-glycosylation site were found. The Pfam predicted the Acetamidase/formamidase family which had a Bit score value of 584.0 and the alignment location is from 16-392.

PROTEIN LOCALISATION AND SIGNAL PEPTIDE

165073481328764 WoLFPSORT prediction cyto: 21.5, cyto_nucl: 12.333, cyto_pero: 12.166

27 Nearest Neighbors

id	site	distance	identity	comments
FKBP_YEAST	cyto	288.545	11.4078%	[Uniprot] SWISS-PROT45:Cytoplasmic. GO:0005634; C:nucleus; Evidence:IDA
AMPL_YEAST	vacu	303.933	14.5914%	[Uniprot] SWISS-PROT45:Vacuolar. GO:0000324; C:vacuole (sensu Fungi); Evidence:IDA
SODC_CANAL	cyto	307.042	10.9223%	[Uniprot] SWISS-PROT45:Cytoplasmic.
OAT_YEAST	cyto	323.024	13.8889%	[Uniprot] SWISS-PROT45:Cytoplasmic. GO:0005634; C:nucleus; Evidence:IDA
PGM2_YEAST	cyto	336.195	15.993%	[Uniprot] SWISS-PROT45:Cytoplasmic.
IDHC_YEAST	cyto	337.184	11.6667%	[Uniprot] SWISS-PROT45:Cytoplasmic.
PGM1_YEAST	cyto	349.323	15.9649%	[Uniprot] SWISS-PROT45:Cytoplasmic.
SODC_NEUCR	cyto	354.201	12.3786%	[Uniprot] SWISS-PROT45:Cytoplasmic.
IDH2_CANTR	pero	355.888	10.9005%	[Uniprot] SWISS-PROT45:Peroxisomal.
PSB6_YEAST	cyto_nucl	360.254	14.0777%	[Uniprot] SWISS-PROT45:Cytoplasmic and nuclear.
EF1A_AJECA	cyto	365.234	14.8387%	[Uniprot] SWISS-PROT45:Cytoplasmic.
ADH2_KLUMA	cyto	366.192	15.0485%	[Uniprot] SWISS-PROT45:Cytoplasmic.
HOSM_YARLI	mito	368.236	11.6331%	[Uniprot] SWISS-PROT45:Mitochondrial.
DYL1_YEAST	cyto	368.301	6.79612%	[Uniprot] SWISS-PROT45:Cytoplasmic. GO:0005623; C:cell; Evidence:NAS
GLNA_AGABI	cyto	369.271	14.3204%	[Uniprot] SWISS-PROT45:Cytoplasmic.
ADH1_KLUMA	cyto	378.697	16.9903%	[Uniprot] SWISS-PROT45:Cytoplasmic.
EF1A_ASPOR	cyto	382.74	13.6659%	[Uniprot] SWISS-PROT45:Cytoplasmic.
PSB1_YEAST	cyto_nucl	385.794	15.0485%	[Uniprot] SWISS-PROT45:Cytoplasmic and nuclear.
VAOX_PENSI	cyto_pero	386.136	13.3929%	[Uniprot] SWISS-PROT45:Peroxisomal and cytoplasmic.

Predicted proteins

tr_A0A162KQM3_A0A162KQM3_CORDF

Prediction: Cytoplasm, Soluble

Localization	Cytoplasm	Peroxisome	Extracellular	Plastid	Mitochondrion	Nucleus	Endoplasmic reticulum	Lysosome/Vacuole	Cell membrane
Likelihood	0.6992	0.2444	0.0186	0.01	0.0093	0.0093	0.0035	0.003	0.0023

Type	Soluble	Membrane
Likelihood	0.9929	0.0071

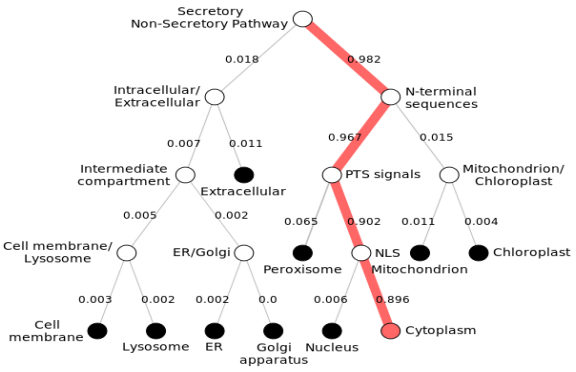


Fig2: The protein localization was determined by WoLF PSORT and DeepLoc tools. The fig. to the left shows that WoLF PSORT finds the nearest neighbors that are preset in the cytoplasm. The figure to the right presents the DeepLoc results which shows that the protein has a likelihood to localize in the cytoplasm.

Signal IP found no signal peptides and TMHMM also did not show presence of any transmembrane domains. By using the protein localization tools like WoLF PSORT and DeepLoc was found that the protein has the highest probability of localizing in the cytoplasm with a likelihood of 0.6992

To show protein localization of the enzymes and to show the protein-protein interaction between the tool String-db was used. The enzymes selected for this study were based on the location which was found to be in the cytoplasm. (table1)

Protein Name	Accession number/Identifier
1) Glutamine Synthetase	A0A168HKS7
2) Malic enzyme	A0A168DVR3
3) Fumarate hydratase	A0A168FIL1
4) Nitrite reductase	A0A162LR72
5) Phosphoenolpyruvate carboxykinase	A0A162KSX6

Table1- Enzymes from the String-db

(1) **WoLF PSORT Prediction**

k used for kNN is: 27
165072550628764 details mito: 17.5, cyto_mito: 11.5, pero: 5, cyto: 4.5

(2 **WoLF PSORT Prediction**

k used for kNN is: 27
165072566928763 details mito: 11, cyto: 7.5, cyto_nucl: 7, nucl: 5.5

(5) **WoLF PSORT Prediction**

(4)

k used for kNN is: 27
165072578628763 details cyto: 12, cyto_nucl: 8.833, cyto_pero: 8.833, mito: 6, nucl: 4.5, pero: 4.5

ito: 2

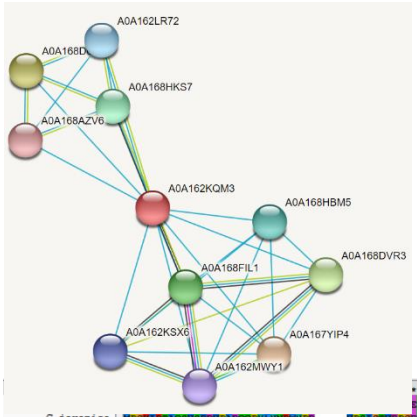


Fig3: 1,2,3,4 – They show the results of the location different enzymes mentioned in the table 1. (cytoplasm). Bottom – Shows the structure of the protein-protein interaction with the different enzymes that have interaction from the curated database.

MULTIPLE SEQUENCE ALIGNMENT AND PHYLOGENETIC TREE

Closely related species:

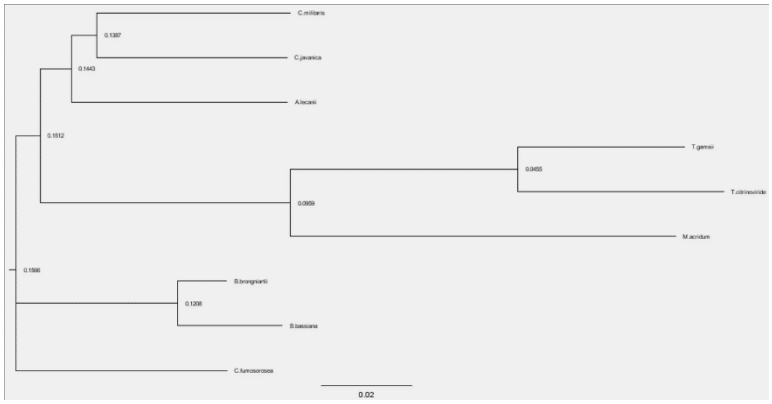


Fig4: Top- Multiple sequence alignment was done for closely related species. Bottom – Shows the phylogenetic tree with C. fumosorosea as the outgroup

Distantly related species:

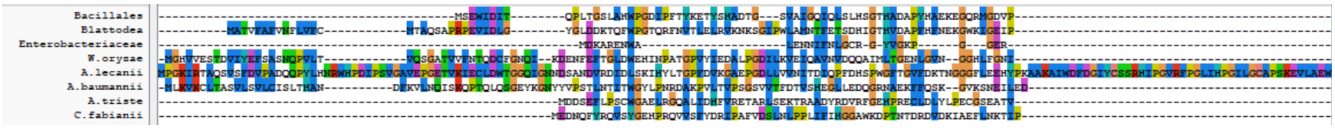
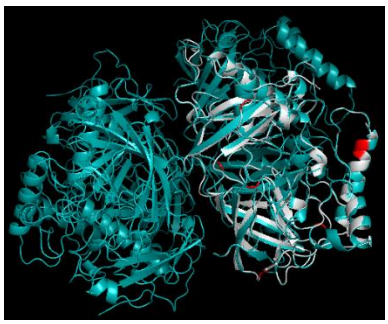
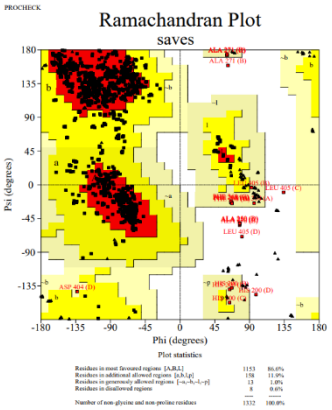


Fig5: Top – A multiple sequence alignment was done for distantly related species of A. lecanii. Bottom – Phylogenetic tree was built for the same sequences.

A Multiple sequence alignment with *A. lecanii* showed strong relation between different organisms of the same fungus family which were considered as the closely related organisms and the distantly related organisms were retrieved from Uniprot which were from different species. The outlier in both the closely related and distantly related species was shown to be different. The species table is shown in appendix (2)

COMPARITIVE 3D MODELLING AND PROTEIN STRUCTURE VALIDATION





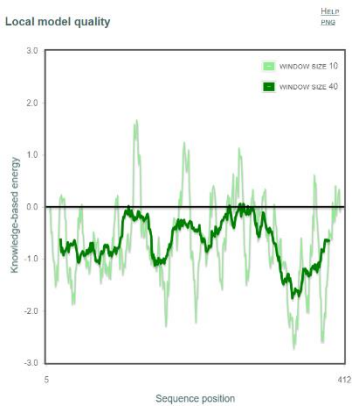
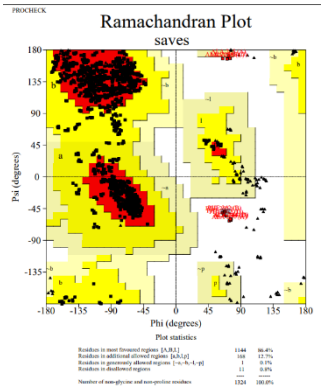
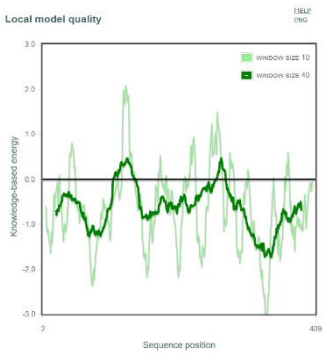


Fig 6: Left top- Shows the aligned 3D structure of the model (aqua blue) and template (orange) and the active site (red) with an RMSD value of 0.174 (range 1-2). Right – Shows the results of the quality checks by Ramachandran plot and PROSA's local-model quality plot of each of the template (bottom) and the model (top). The colored regions show the allowed regions.





The template and model were built in Swiss Model. The template and model were aligned in Pymol. The active site residues were found through literature. The PROSA and Ramachandran plots were done for the model where more than 90% of amino acids were in the allowed regions of the Ramachandran plot. The PROSA score for the model was -7.46 and 2wkn.1 template was -8.02 while the energy model showed peaks almost similar to the template but mostly in the negative region.

DISCUSSION

DNA ASSEMBLY:

The estimated coverage for the gene length was found to be 43.176. This is very high when compared to the actual coverage of SPADES (12.5454). The estimated number of contigs for the coverage of 12.5454 was found to be 4.071775763*10^-27 contigs which is close to zero and not possible score. Hence, we get a single contig for this coverage. The calculation is shown in appendix (1)

DNA ANNOTATION:

BlastX was used to annotate and the best hit was matched with formamidase enzyme of the Akanthomyces lecanii with a 100% identity. The bit score and E value were 850 and 0 respectively. BlastX was the basis to annotate the respective raw reads that were given initially.

PROTEIN LOCALISATION AND SIGNAL PEPTIDE

The SMART tool for domain did not give any results maybe because it had only limited set of curated domains. PROSITE and Pfam found the presence of protein kinase phosphorylation site with 20 other hits against 4 patterns with a region between 42-44 which means that they have smaller database of patterns and profiles. PROSITE also found that protein kinase C exhibits a preference for the phosphorylation of serine or threonine residues that are found close to C terminal. The results from Pfam had a bit score value of 584 with a region between 16-392. These results showed that they have a larger database patterns and profiles. Also, this particular domain predicted that it hydrolyzes formamide with the production of ammonia which can be used as a nitrogen source.
The protein localization results of WoLF PSORT showed the most of the sequences was soluble and found to localize mostly in the cytoplasm. The results of DeepLoc also gave similar results as WoLF PSORT.
This related back to my research question where it has shown that formamidase has protein interactions with the cell wall membrane proteins and cytosolic proteins because it is present in the fungus cytoplasm and cell wall and these interactions might be due to potential roles in the nitrogen pathway which was confirmed by the tool String.

MULTIPLE SEQUENCE ALIGNMENT AND PHYLOGENY

The *Akanthomyces lecanii* was aligned with other organisms of the same family. Hence, the results showed a close relation with species of the fungus family but showed distant relation with other organisms that were taken from Uniprot. These organisms were from the family of bacteria.
The node length of closely related species showed better results than distantly related species. The outlier that was obtained in closely related species was *C. fumosorosea* and there were two outliers in the distantly related species that were *A. triste* and *C. fabianii*. The tree of both closely related species and distantly related species showed distinct orthologs and paralogs with the other organisms. The table of organisms is shown in appendix (2)

3D MODELLING AND QUALITY CHECK

Through the quality check the model was found to be the optimum with a GMQE value of 0.85 which is greater than 0.7. The template and model were aligned with RMSD value of 0.174 >0.7 which means that the distance between the template and model was 0.174 Units away from the alignment. The Ramachandran plot indicated that a greater number of residues were present in the allowed regions. The PROSA showed a Z-score which was within the optimum region of z-scores. The local model quality of both the model and template was almost similar but the model graph showed a few peaks in the negative region.

RESEARCH QUESTION

The research question revolves around "Studying the protein-protein interaction of formamidase in Akanthomyces lecanii and understanding the reason behind it."
To tackle the question, the tool string results showed that formamidase interacts with Malic enzyme, Glutamate synthetase, Phosphoenolpyruvate carboxykinase (ATP), Nitrite reductase and fumarate hydratase all of which had a homology score of >0.90 which is optimal. Most of these proteins had interactions from the curated databases (shown in blue). This network has significantly more interactions than expected which means that they are biologically connected. Therefore, the protein has mainly interactions with cell wall membrane proteins and cytosolic proteins as it is present in the fungus cytoplasm and cell wall and these interactions might be due to potential roles in the nitrogen metabolism pathway. In future, there could be research done on other interactions related to formamidase. It was experimentally proven that by isolating a cell wall protein interacted with formamidase at the fungal surface.

ABSTRACT

Formamidase is an enzyme of the hydrolase family which is associated with the Kynurenine pathway. Members of the fungi family such as the Akanthomyces lecanii are involved in this pathway as either an antifungal or as a pest for insects. As the location of the species is very important to find out the function and the pathway involved. Experiments have shown that this fungus is located mostly in the cytoplasm of the cell or in the mitochondria. A protein-protein interaction studies/experiment revealed that *Akanthomyces lecanii* has interaction with a few species which are also located in the cytoplasm such as the Glutamine synthetase, Malic enzyme, Fumarate hydratase, Nitrite reductase, Phosphoenolpyruvate carboxykinase. The interaction between these organisms had been derived from curated databases. The likelihood of the location of the protein was also found to be withing the range. Therefore, in this research it was found that it mainly has interactions with cell wall membrane proteins and cytosolic proteins as its present in the fungus cytoplasm and cell wall and these interactions might be due to potential roles in the nitrogen metabolism pathway.

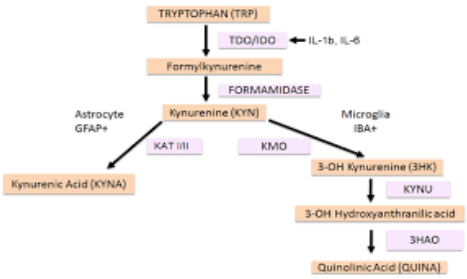


Fig7: Left – Kynurenine pathway. Right – Microscopic view of Aknathomyces lecanii

APPENDIX

1) COVERAGE CALCULATION

Gene Length – 1234 ; No. of reads – 444 ; Length of each read – 60

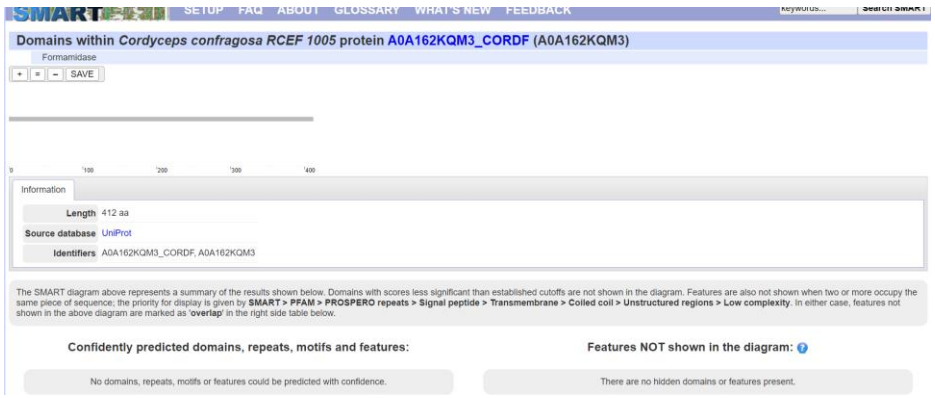
Statistical coverage = (444 x 2 x 60)/1234 = 43.176

2) SPECIES Table

Distantly related species	Closely related species
1) *Aknathomyces lecanii*	1) *Akanthomyces lecanii*
2) *Acinetobacter baumanii*	2) *Cordyceps militaris*
3) *Amblyomma triste*	3) *Beauveria brongniartii*
4) *Cyberlindnera fabianii*	4) *Cordyceps javanica*
5) *Bacillales*	5) *Cordyceps fumosorose*
6) *Blattodea*	6) *Beauveria bassiana*
7) *Enterobacteriaceae*	7) *Trichoderma gamsii*
8) *Weisella Oryzae*	8) *Metarhizium acridum*
	9) *Trichoderma citrinoviride*

3) GALAXY RESULT

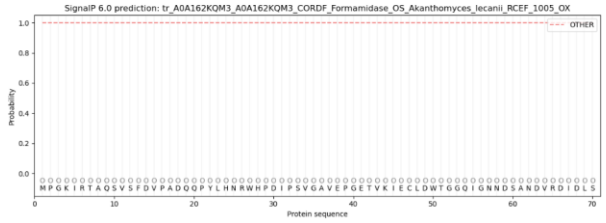
4) SMART RESULT



5) SIGNAL IP RESULT

Protein type	Other	Signal Peptide (Sec/SPI)
Likelihood	1	0.0001

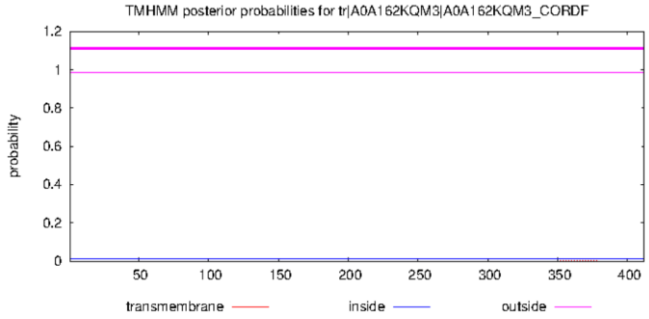
Download: PNG / EPS / Tabular



6) TMHMM RESULT

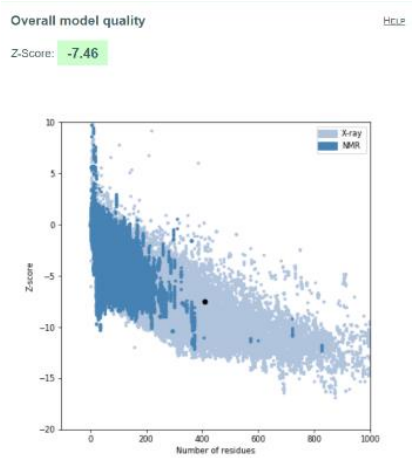
TMHMM result

```
# tr|A0A162KQM3|A0A162KQM3_CORDF Length: 412
# tr|A0A162KQM3|A0A162KQM3_CORDF Number of predicted TMHs:  0
# tr|A0A162KQM3|A0A162KQM3_CORDF Exp number of AAs in TMHs: 0.03539
# tr|A0A162KQM3|A0A162KQM3_CORDF Exp number, first 60 AAs:  0
# tr|A0A162KQM3|A0A162KQM3_CORDF Total prob of N-in:        0.01387
tr|A0A162KQM3|A0A162KQM3_CORDF  TMHMM2.0       outside     1   412
```

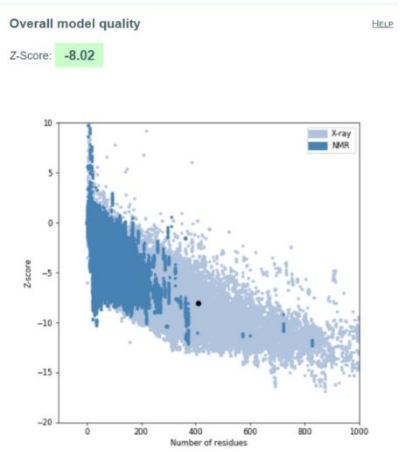


7) Overall Model Quality

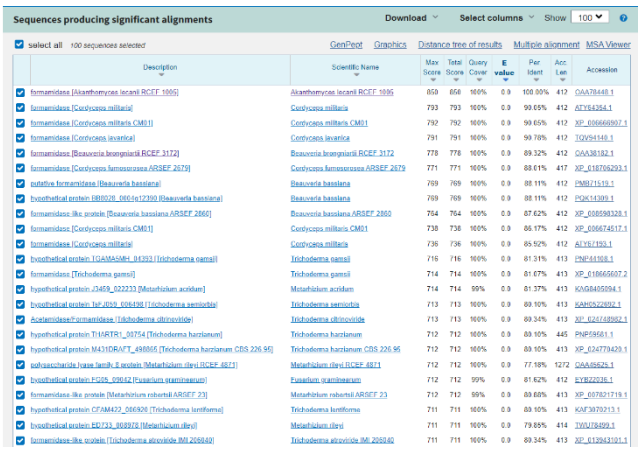
Model

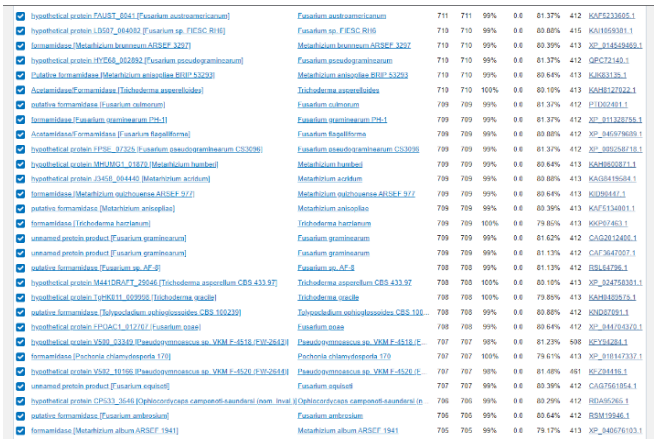


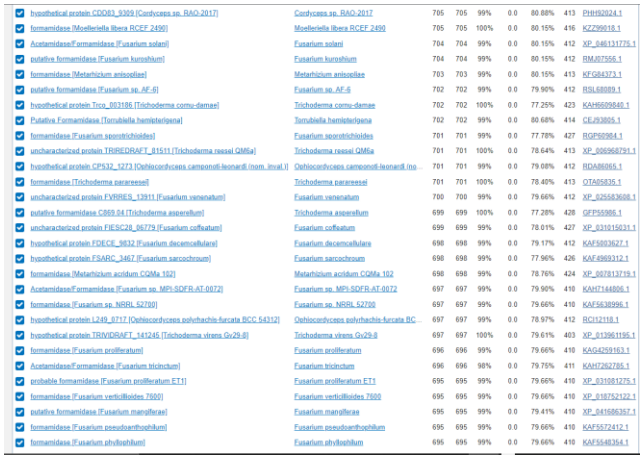
Template

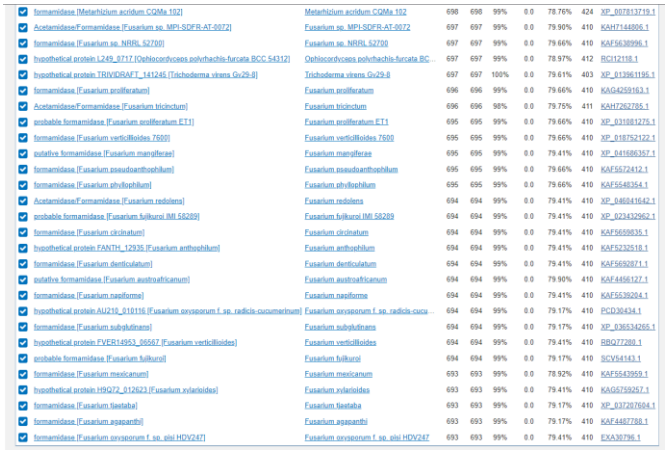


8) Blast Search sequences

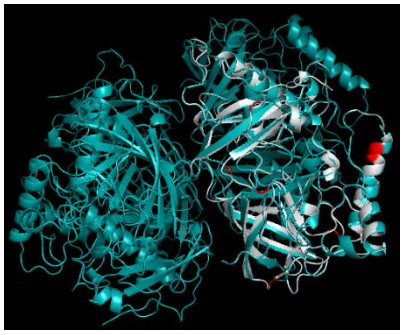


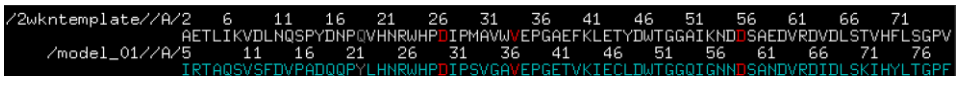






9) 3D HOMOLOGY MODELLING – Template (grey), Model (teal), active site residues (red)





REFERENCES

Dykhno, M. M., Zemtsova, O. M., & Zaides, V. M. (1964). [DIFFERENTIATION BETWEEN ACID-RESISTANT SAPROPHYTES, MYCOBACTERIUM TUBERCULOSIS AND ATYPICAL STRAINS BY THE DETERMINATION OF FORMAMIDASE ACTIVITY]. *Problemy Tuberkuleza*, *42*, 63–66.

Scott, M. S., Calafell, S. J., Thomas, D. Y., & Hallett, M. T. (2005). Refining protein subcellular localization. *PLoS Computational Biology*, *1*(6), e66. https://doi.org/10.1371/journal.pcbi.0010066

Sevimoglu, T., & Arga, K. Y. (2014). The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, *11*(18), 22–27. https://doi.org/10.1016/j.csbj.2014.08.008

Shekari, F., Baharvand, H., & Salekdeh, G. H. (2014). Organellar Proteomics of Embryonic Stem Cells. In *Advances in Protein Chemistry and Structural Biology* (Vol. 95, pp. 215–230). Elsevier. https://doi.org/10.1016/B978-0-12-800453-1.00007-5

Sorci, L., Kurnasov, O., Rodionov, D. A., & Osterman, A. L. (2010). Genomics and Enzymology of NAD Biosynthesis. In *Comprehensive Natural Products II* (pp. 213–257). Elsevier. https://doi.org/10.1016/B978-008045382-8.00138-6

Wu, G., Huang, Q., Tang, Y., Unno, H., & Kusunoki, M. (2005). Crystallization and preliminary crystallographic study of a recombinant predicted acetamidase/formamidase from the thermophile *Thermoanaerobacter tengcongensis*. *Acta Crystallographica Section F Structural Biology and Crystallization Communications*, *61*(1), 106–108. https://doi.org/10.1107/S1744309104030519

Zhang, Y.-J., Yang, X.-B., & Zhang, S. (2020). Complete mitogenome of the entomopathogenic fungus *Akanthomyces lecanii*. *Mitochondrial DNA Part B*, *5*(1), 1021–1022. https://doi.org/10.1080/23802359.2020.1721349

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. (2017). DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics*, *33*(21), 3387–3395. https://doi.org/10.1093/bioinformatics/btx431

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, *37*(4), 420–423. https://doi.org/10.1038/s41587-019-0036-z

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). WoLF PSORT: Protein localization predictor. *Nucleic Acids Research*, *35*(Web Server), W585–W587. https://doi.org/10.1093/nar/gkm259

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*(21), 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283–291. https://doi.org/10.1107/S0021889892009944

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, *38*(suppl_1), D161–D166. https://doi.org/10.1093/nar/gkp885

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, *49*(D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, *35*(Web Server), W407–W410. https://doi.org/10.1093/nar/gkm290

FigTree (no date). Available at: http://tree.bio.ed.ac.uk/software/figtree/ (Accessed: 6 October 2021)

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., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427