A Database and Artificial Intelligence Analysis of an Unknown Protein in *Landoltia punctata*

Alex Zatuchney

Received June 23, 2024
Accepted September 10, 2024
Electronic access September 30, 2024

Landoltia punctata, or Spotted Duckweed, is a small, fast-growing aquatic plant native to freshwater bodies. Due to its fast-growing properties and high starch accumulation, the plant is of interest to researchers, with potential applications in bioremediation, biofuel production, and antibiotic manufacturing. Because of its potential significance, this paper seeks to determine an unknown protein's structure, function, and role in *Landoltia punctata* via database searches and artificial intelligence structure and protein disorder prediction models. Since proteins perform various functions within a plant cell, understanding their roles will lead to a better understanding of how *Landoltia punctata* can be utilized in these fields of interest. Database searches yielded several matches of medium significance, but many matches had varying functions. The structure and disorder predictors suggested that the unknown protein was likely intrinsically disordered. Furthermore, they predicted that the beginning of the protein binds to nucleic acids and the end of the protein binds to other proteins, indicating its potential role in either the transcription or translation processes in *Landoltia punctata*. These findings may mean that researchers can modify the protein to increase output of potentially desired products. However, due to the lack of distinct significant database matches and because these artificial intelligence models are not perfectly accurate, these results remain only partially conclusive. This study recommends that future researchers experimentally determine the structure and function, and role in *Landoltia punctata*.

Introduction

Duckweeds are an aquatic plant group that are of interest to researchers due to their potential applications in biofuel production and bioremediation. Specifically, *Landoltia punctata* (LP) is relevant due to its unique genetic composition. The plant is the only organism in the Landoltia genus and thus is genetically distant from other duckweed genera[1]. This genetic profile results in unique morphological characteristics that may be useful to researchers. LP has multiple potential applications: bioremediation, being able to absorb both phosphorus and nitrogen rapidly, especially in conjunction with other plants[1]; biofuel production, being able to produce ethanol for combustion[2]; and potential antibiotic and aquaculture application, containing flavonoids that have antibiotic and antioxidant properties[3,4]. Understanding what causes the organism's behaviors and functions—that is, the cellular processes behind them—may allow researchers to be able to better utilize the plant in these potential applications. As proteins are largely responsible for these cellular processes, better understanding their role within LP may thus benefit future researchers who aim to use the organism.

Proteins

Because each protein's function is dictated by its structure, understanding the structure of a protein may yield insights into the molecular processes within a cell and may allow researchers to modify the protein to better suit an organism like LP to their needs. Outside of function, determining the structure of a protein is also vital for understanding the structural evolution of proteins[5]. For example, because alpha helices typically interact with DNA or RNA, tracing the history of alpha helices in a protein family may yield insight into how the function of a specific protein category changed over time[6].

Despite their importance, many sequenced proteins have not had their structures determined[5]. Although methods like Nuclear Magnetic Resonance Imaging exist, they are expensive, require a high-quality sample, and are subject to human error[7,8]. As such, researchers have been investigating alternative methods for predicting protein structure. Recent developments in artificial intelligence have created a more accurate prediction model called AlphaFold[9]. Evans et al. demonstrated that AlphaFold is significantly more accurate at predicting protein structure than previous prediction methods and is comparable in accuracy to experimental methods[10].

Understanding when and where a protein is translated may also help determine the function of a protein. One way to experimentally determine where a protein is expressed is through a western blot analysis, as Wang et al. did with Lemna aequinoctialis, another duckweed species[11]. Western blot analyses aim to separate and identify specific proteins in an organism via gel electrophoresis and antibody binding. By doing so, researchers

can monitor the presence and thus expression of a protein in a region of interest of an organism[11].

Database Searches to Predict Protein Function

Often, knowing protein structure is not enough to determine function alone, and a researcher must perform additional analysis to do so. One potential method to determine protein function is database searches, which aim to find proteins with similar sequences to a query sequence. Theoretically, similar proteins (homologs) have similar structures and functions[12]. Thus, if one were to find a statistically significant protein match with a known function, they would be able to predict the query sequence's function.

One method of determining the match significance is to utilize its E-value. As defined by NCBI, this is the number of matches similarly significant to the one displayed that one can expect to find when searching through a database of a given size due to chance[13]. Thus, if the E-value was 1, one could expect at least one match of equal or greater significance in a database of similar size due to random chance. E-values vary due to the size of the query sequence and the number of entries within the database searched, so it is difficult to quantify the threshold for significant values. However, values of 1E-50 and above are generally deemed very significant for both DNA and protein sequences, and values from 1E-10 to 1E-50 suggest a relationship between the two sequences[14]. Additionally, the appearance of multiple related matches with similar E-values signals that the inputted sequence may belong to a family of proteins. Though methods to experimentally determine protein function exist, database searches are a fast, easy, and inexpensive method to conclude whether such analyses are necessary.

Protein-Protein Interactions

Proteins rarely exist in a vacuum, usually interacting with other proteins to perform their functions. Protein-protein interactions (PPIs) are vital for understanding a protein's function and relevance within an organism[15]. Understanding them within the context of LP would provide researchers with insight into how each protein contributes to the overall function of a cell, potentially opening avenues for proteins to be modified to better suit the plant to one's needs, like biofuel production or bioremediation.

There are multiple methods one can utilize to determine PPIs. To identify potential interactions between uninvestigated proteins, one may conduct a pull-down assay, where a researcher would isolate and fluorescently tag a protein of interest, place it in an *in vitro* solution containing other proteins from the organism of origin, and re-extracted via centrifuge if it binds to other proteins, as Louche et al. did in their study[16]. Alternatively, one may employ a structural approach: if two unknown proteins are structurally similar to two proteins that are known to interact, then it is likely that those unknown proteins also interact with each other[17]. Artificial intelligence models can also predict PPIs. Multimer versions of AlphaFold or other models like RoseTTAFold are used to predict protein interactions and complexes[18]. These models could allow researchers to predict these interactions within LP with high accuracy and without utilizing expensive equipment.

Intrinsically Disordered Proteins

It is important to note that not all proteins have a defined structure. Intrinsically disordered proteins (IDPs) are proteins where some or all regions do not have a fixed or ordered three-dimensional structure, typically in the absence of other interacting molecules. Despite this, IDPs are functional, usually conforming to a defined shape when paired with another molecule, like DNA or other proteins[19]. Disordered regions have been proven to interact with nucleic acids as well[20]. IDPs are often heavily involved in PPIs and are important in the regulation of DNA transcription, the formation of links between two proteins, and cell signaling[21]. Thus, understanding the role IDPs play in LP could allow researchers to modify them to make those cellular processes more efficient.

Researchers have created artificial intelligence programs to predict IDPs. The most accurate disorder predictor currently available is flDPnn, a deep-learning neural network program. The model also predicts what function the disordered region carries out, like binding to specific compounds or forming linker regions between two proteins[22]. Hu et al. found that flDPnn has an average area under the receiver operating characteristic curve (AUC) of 0.814. AUC acts as a measure of accuracy and ranges from 0.5 for a random prediction and 1.0 for a perfect prediction[22]. Notably, flDPnn's AUC values for DNA and RNA binding were significantly higher (0.87 and 0.86, respectively) than its protein binding AUC values (0.79).

A eukaryotic protein has, on average, 32% of its residues disordered, meaning that simply using AlphaFold may not yield accurate results regarding the structure of some segments of proteins[23]. Thus, it is important to identify potential disordered regions within proteins with programs like flDPnn when predicting structure to maximize accuracy.

Gap

Despite these recent advancements in sequencing and prediction, many proteins are unsequenced, and an estimated 43% of eukaryotic proteins have not had their structures observed or simulated[24]. One such unknown protein is coded for by DNA sequence JZ987503.1 (the protein will be referred to as Unknown Protein or UKP). JZ987503.1 was sequenced and published by the author as part of the Waksman Student Scholars Program

(WSSP) by isolating bacterial plasmid copies (cDNA) of the sequence[25]. No other identifying characteristics of JZ987503.1 are known. UKP has not had its structure or function determined. This is demonstrated by the lack of significant named matches for JZ987503.1 in BLAST, a tool developed by NCBI that searches through all published DNA sequences to find matches. When searching for matches to JZ987503.1, only one database yielded statistically significant DNA matches with function-denoting names. However, all of these matches came from WSSP, where high school students "determine if the sequences are similar to genes from other organisms using bioinformatic programs and accessing databases"[26].

As such, JZ987503.1 and its matches must have obtained their names from another match somewhere in the NCBI databases. If one converts JZ987503.1 into a protein sequence, they can find a named but statistically insignificant match in an NCBI database. This match is not very significant, with an E-value of 7E-10, and no other significant named matches exist in the top 100 matches in the protein version of BLAST for UKP. Because of this, one cannot accurately state the structure (if the protein is ordered) and function of the protein coded by JZ987503.1 and what interactions it participates in.

As such, there is an apparent knowledge gap in the prior research concerning UKP because no "true" significant named matching sequences exist on any NCBI database. The author believes UKP is relevant due to its organism of origin. Although this protein and the genetic sequence coding for it have no immediate distinguishing characteristics that make them of special interest, understanding any protein's structure, function, and role within LP may allow researchers to uncover the structural evolution of the protein as well as modify or target it to allow LP to better suit its potential biofuel, bioremediation, and antibacterial applications[1,2,4,5].

Because the researcher does not have access to laboratory tools, this paper seeks to determine the structure of the ordered regions in UKP, if present, and how the protein contributes to the overall function and survival of a cell in LP that expresses it by using a combination of large database searches and artificial analysis tools. It will first convert JZ987503.1 from a DNA sequence into a protein sequence and use AlphaFold and flDPnn to predict its structure before searching through publicly available protein databases and performing additional analysis, if necessary.

Only the length of sequence JZ987503.1 is currently known. Given this information, the author hypothesizes that because JZ987503.1, and by extension, UKP are relatively short sequences, UKP could be part of the lysosome, binding to and degrading other cellular material. Many lysosomal proteins, like SNAPIN or assembly subunits, are around 120-150 residues long, which is approximately the length of UKP[27,28].

Results

Due to the explorative nature of this study, results build on each other and thus are presented in chronological order. Additionally, because the control sequence used does not contribute to the conclusion and only validates the accuracy of the methodology, all results regarding it are in the methods section.

NCBI's ORF finder found that ORF1 was most likely the reading frame that coded for a protein. Because JZ987503.1 was derived from a reverse transcription of an mRNA sequence, the sequence is guaranteed to run left to right, meaning that ORFs four through six are not possible outputs[25]. Out of the ORFs one through three, ORF1 is the most likely to code for a protein; ORF1 is 129 residues long, ORF2 is 43 residues, and ORF3 is 31 residues long. As protein domains are typically longer than 40 residues, ORFs 2 and 3 are less likely to be coding ORFs. The BLASTp results for each of these ORFs confirm this, as ORF1 had several significant but unnamed protein matches while ORFs 2 and 3 had no matches above 1E-05. The researcher then used ORF1 as the sequence for UKP for the rest of the methods. Refer to Fig. 1 for the exact sequence.

AlphaFold predicted that UKP had around a 50-residue-long, high-confidence alpha helix structure in the front, followed by a long, low-confidence tail. There appear to be partial formations of alpha helices throughout this tail, though the low-confidence level suggests that this region has no ordered structure. Refer to Fig. 2 for the full protein sequence structure and Fig. 3 for the confidence intervals for the protein and predicted aligned error.

The most significant matches in my searches came from the TrEMBL database in UniProt. The four most significant matches were CRT10 with an E-value of 1E-25; "DNA dependent protein kinase catalytic subunit" with an E-value of 4E-21; "TXP2 C-terminal domain-containing protein" with an E-value of 4E-19; "Putative histone acetyltransferase HAC-like 1" with an E-value of 1E-18; and "Genome assembly, chromosome: A01" with an E-value of 4E-15. All E-values were generated by NCBI's BLAST "Align Two Sequences" tool. Refer to Figs. 4-7 for the AlphaFold predicted structures of each of these proteins.

FlDPnn predicted that UKP was entirely disordered. Around the first 30 residues are predicted to bind to a nucleic acid, with a high DNA and RNA binding propensity, and around the last 52 residues are predicted to bind to proteins or form linker regions, with high protein binding propensities for that area. The middle section of the protein is most likely to form a linker region between two proteins. Refer to Fig. 8 for the flDPnn predictions for UKP and Fig. 9 for a graphical representation of the propensities generated in Google spreadsheets.

Database searches containing around the last 52 residues yielded no significant matches. However, there were several matches for the database search containing the first 41 residues. All match E-values were standardized with NCBI's "Align Two Sequences" tool and presented in order of significance. The

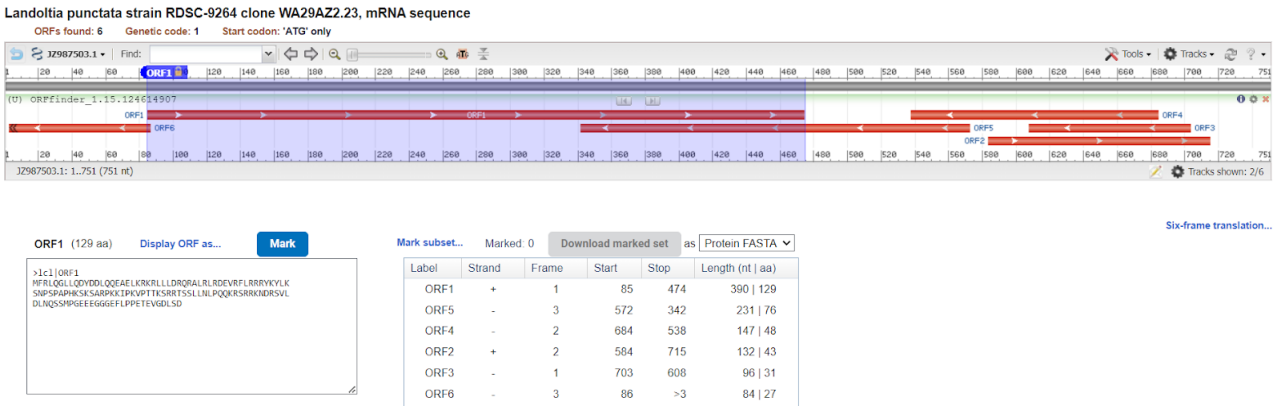

Fig. 1 NCBI's ORF Finder Results

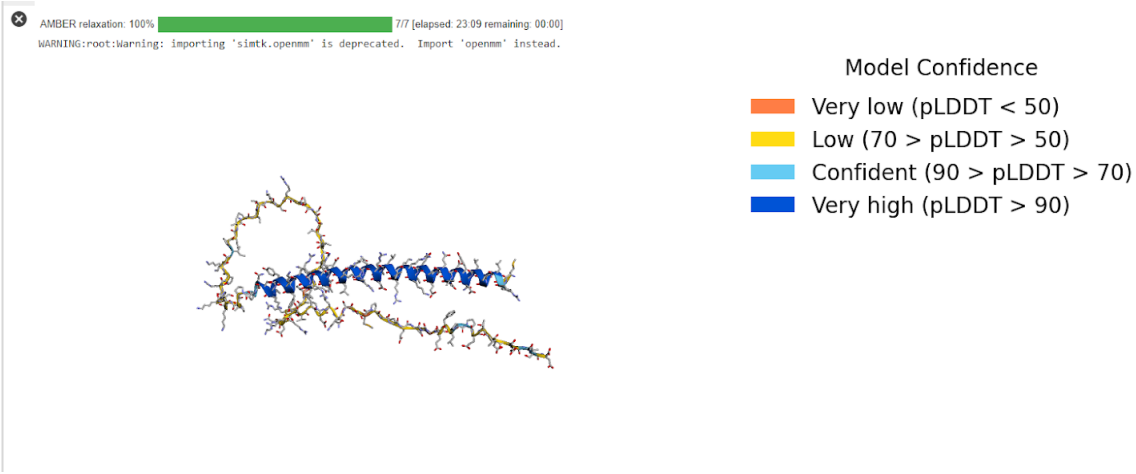

Fig. 2 Predicted Structure For UKP, Generated by AlphaFold.

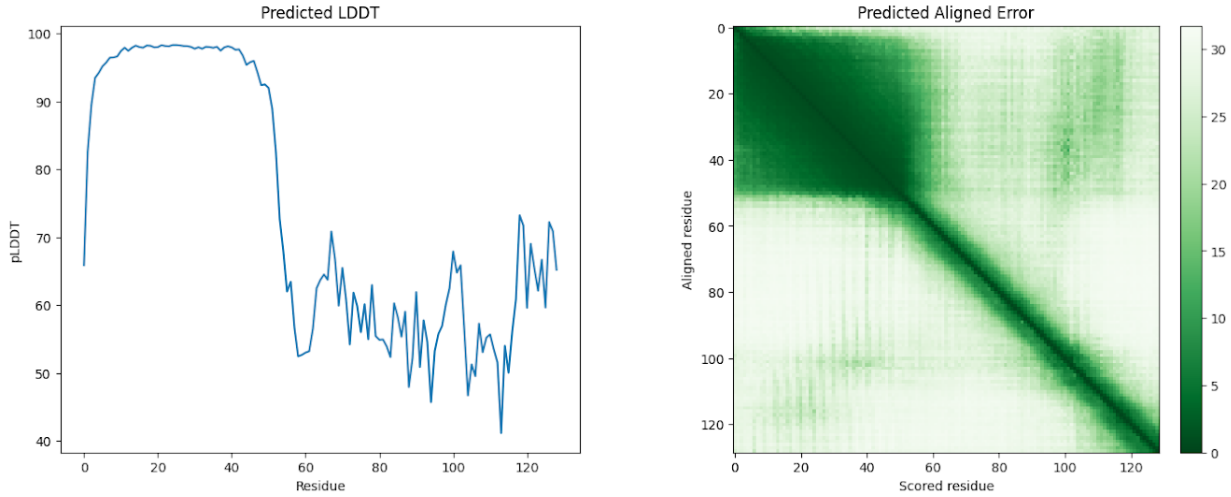

Fig. 3 Confidence Interval and Predicted Aligned Error for UKP, Generated by AlphaFold

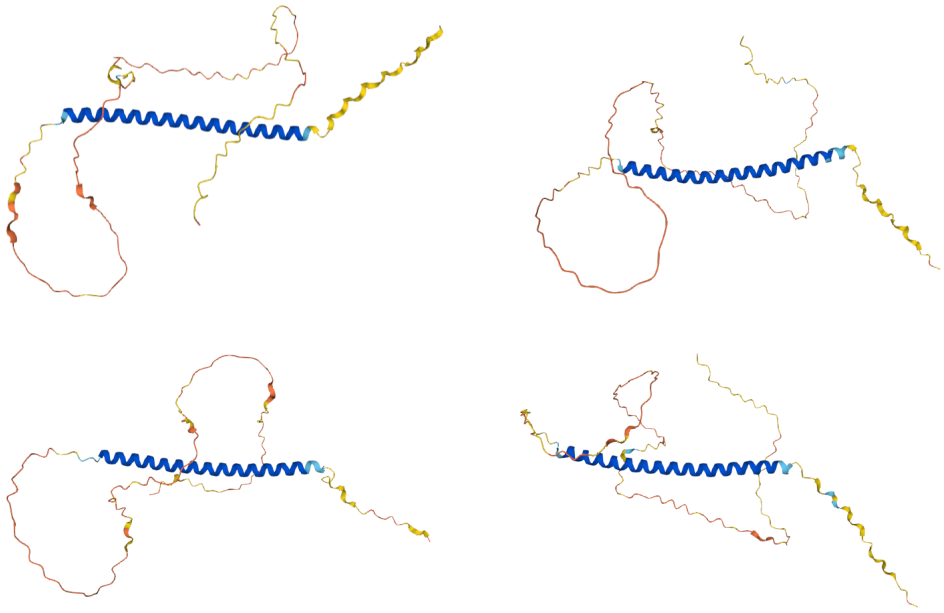

Fig. 4 - 7: From Left to Right, Top to Bottom, the Structures for the Four Proteins, from Highest E-value to Lowest: CRT10, "DNA dependent protein kinase catalytic subunit," "TXP2 C-terminal domain-containing protein," "Putative histone acetyltransferase," and "Genome assembly, chromosome: A01"

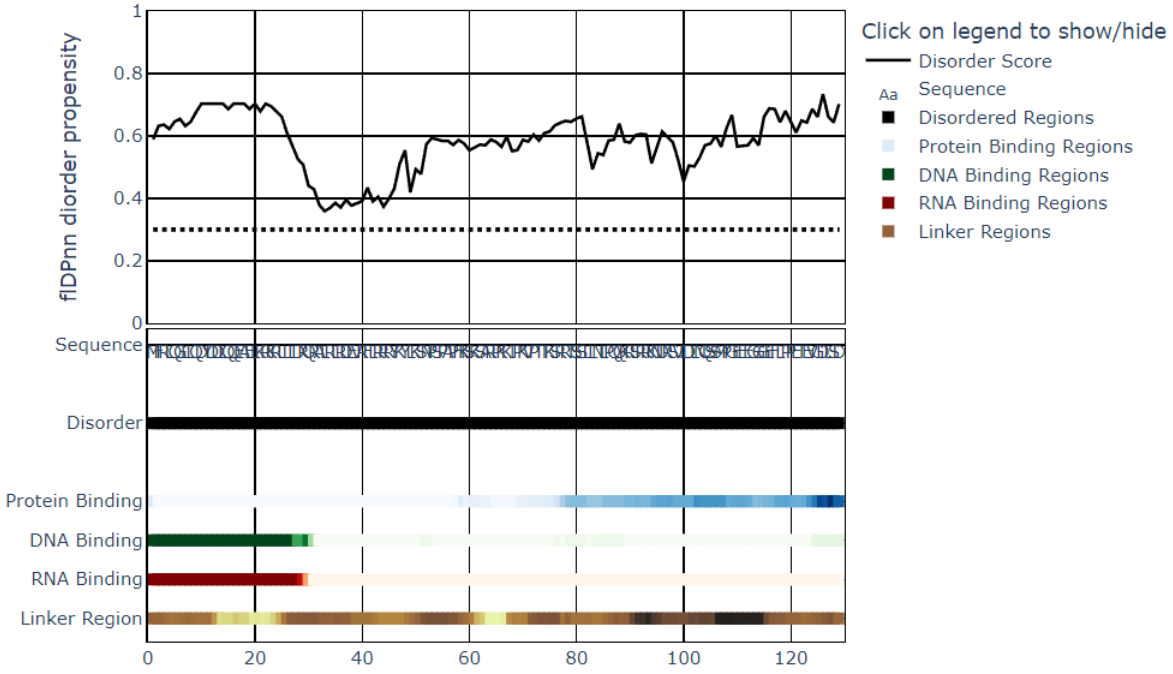

Fig. 8 flDPnn's Results for UKP

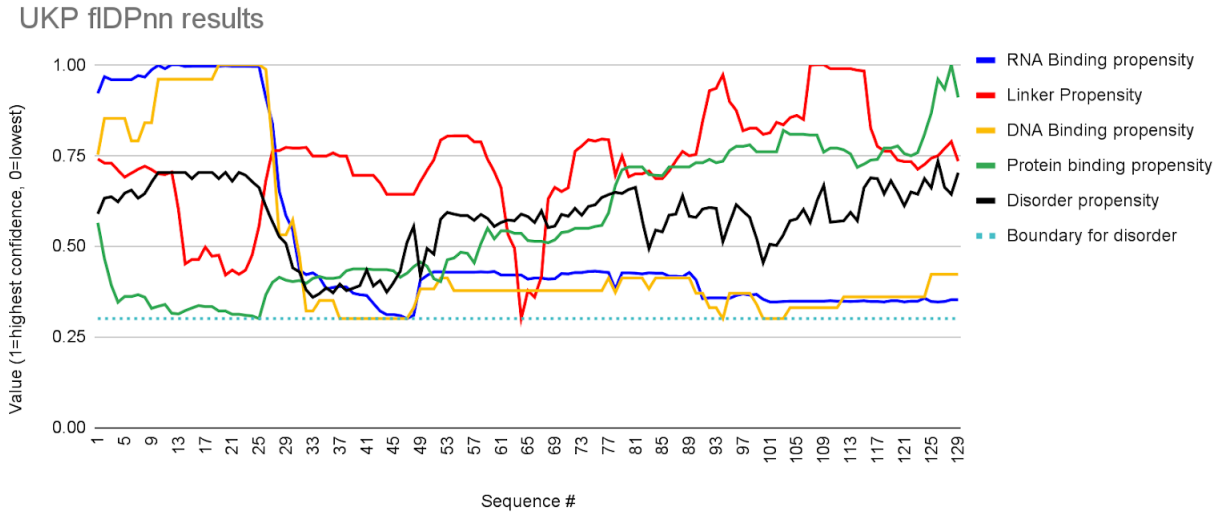

Fig. 9 Graph Generated in Google Spreadsheet for the Varying Propensities of UKP. Data Points were Extracted from the CSV file Generated by the Program

first match is CRT10, with an E-value of 9E-13, "Shugoshin C-terminal domain-containing protein" at 4E-11, "Non-specific serine/threonine protein kinase" at 4E-11, "TXP2 C-terminal domain-containing protein" at 4E-11, "DNA-dependent protein kinase catalytic subunit" at 2E-11, and "Putative histone acetyltransferase HAC-like 1" at 3E-10. The flDPnn results for those sequences are below. Refer to Figs. 10-15 for the results.

The functions of these protein matches vary significantly; CRT10 binds to ribosomal RNA and degrades mutant variants[29]. TPX2 binds to other proteins and spindle fibers during mitosis, though not all interactions with this protein are known[30]. Shugoshin proteins help maintain nuclear stability and protect chromatid cohesion, ensuring that chromosomes remain intact[31]. Putative histone acetyltransferase HAC acetifies DNA, binding itself to DNA to denature and thus deactivate the nucleic acid[32]. The non-specific serine/threonine protein kinase binds to amino acids. Although the fact that many of the proteins are known to bind with nucleic acids (or do not have their functions known) may suggest a weak correlation, because these proteins are not related regarding specific function, no conclusions can be directly generated from these results.

Discussion

Conclusion

This research suggests that UKP is likely intrinsically disordered and forms complexes with other proteins. Additionally, UKP may bind to nucleic acids and proteins.

The most apparent conflict of results lies within the conclusions generated by AlphaFold and flDPnn. AlphaFold predicted that UKP contains an ordered alpha helix region spanning around the first 50 residues while flDPnn suggested that the protein is entirely disordered. This region is especially relevant when considering the confidence intervals of both programs. AlphaFold's only confident region was this first domain, and the nucleic acid binding propensities on flDPnn for it were nearly 1.0 while its disorder propensity was relatively high. This suggests that this region shares properties with both training datasets, which, if interpreted literally, would mean that the domain is both ordered and disordered. Despite this conflict, this region may still interact with nucleic acids, as alpha helices typically interact with DNA or RNA and disordered regions have been proven to interact with nucleic acids as well[6,20]. Because of this possibility, UKP could potentially be involved in transcription or translation processes that regulate the expression of other proteins.

However, flDPnn is more likely to be correct in its prediction of protein structure due to the large, low-confidence region in AlphaFold's output. This domain suggests that the majority of UKP does not share properties with AlphaFold's nearly entirely ordered training dataset. No large, low-confidence region exists for flDPnn, meaning that UKP has a higher likelihood of sharing properties with its entirely disordered training dataset. As such, UKP may be more likely to not contain an alpha helix.

However, the results generated from this study are only partially conclusive due to the nature of the data collected. No artificial intelligence model used had 100% accuracy, and the database results were not significant enough to generate inferences or conclusions about the functions of UKP or its inter-

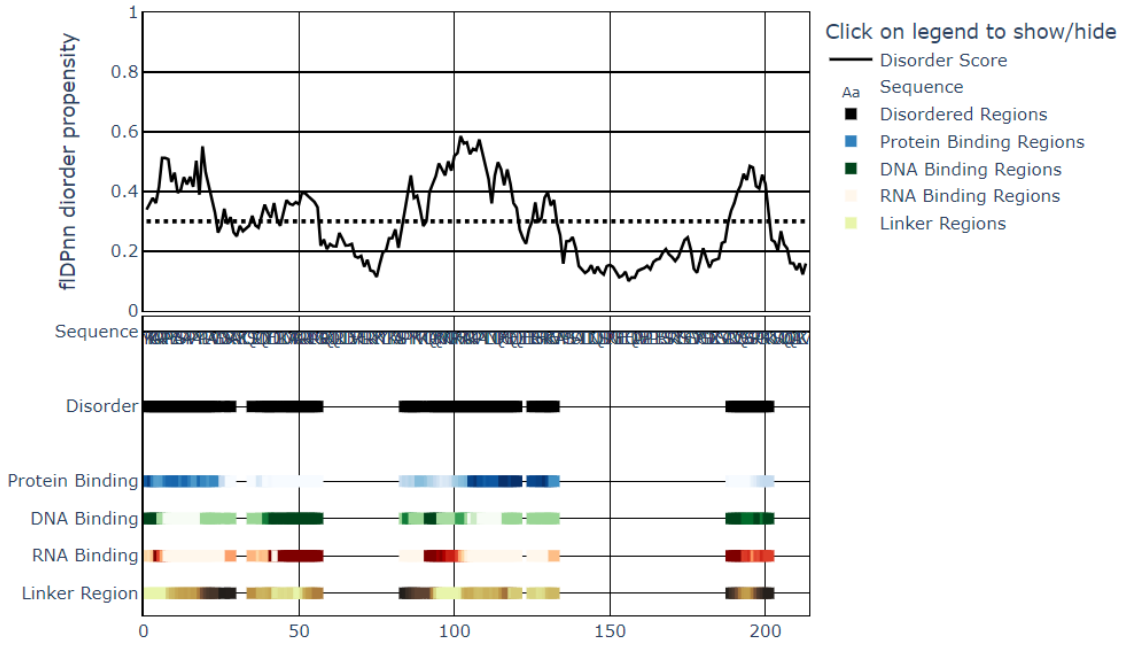

Fig. 10 FlDPnn Results for Protein CRT10. Match Begins at Residue 31 and Ends at Residue 71

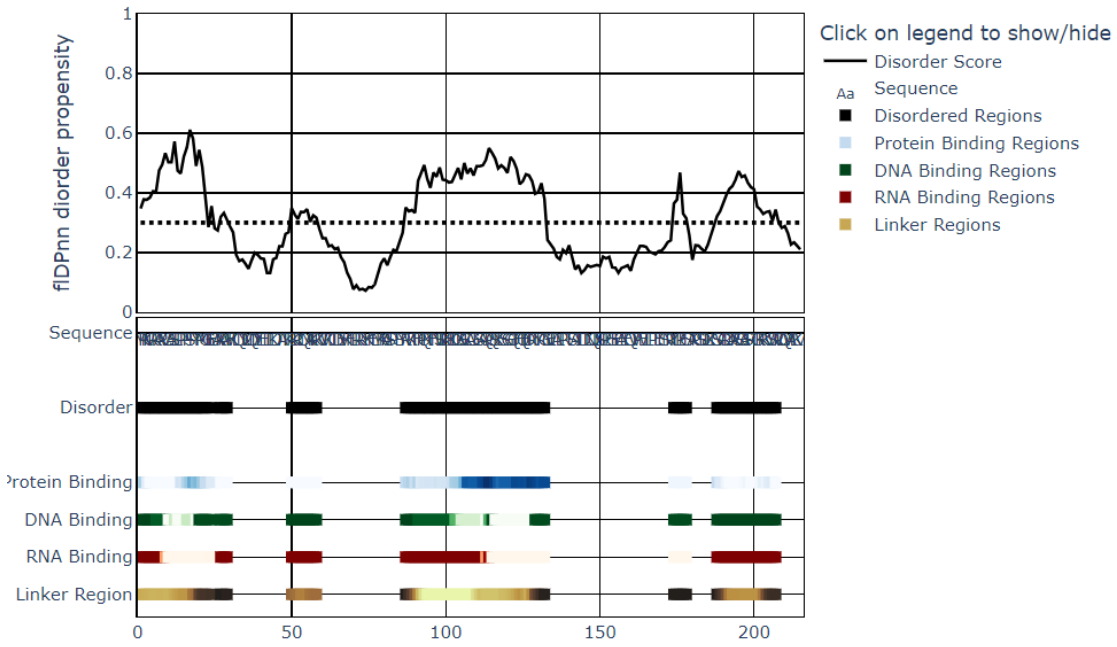

Fig. 11 FlDPnn Results for "DNA-dependent protein kinase catalytic subunit." Match Begins at Residue 29 and Ends at Residue 69. Notably, some Data is not Generated due to a Lack of Predicted Disorder

>Shugoshin_C_terminal_dom

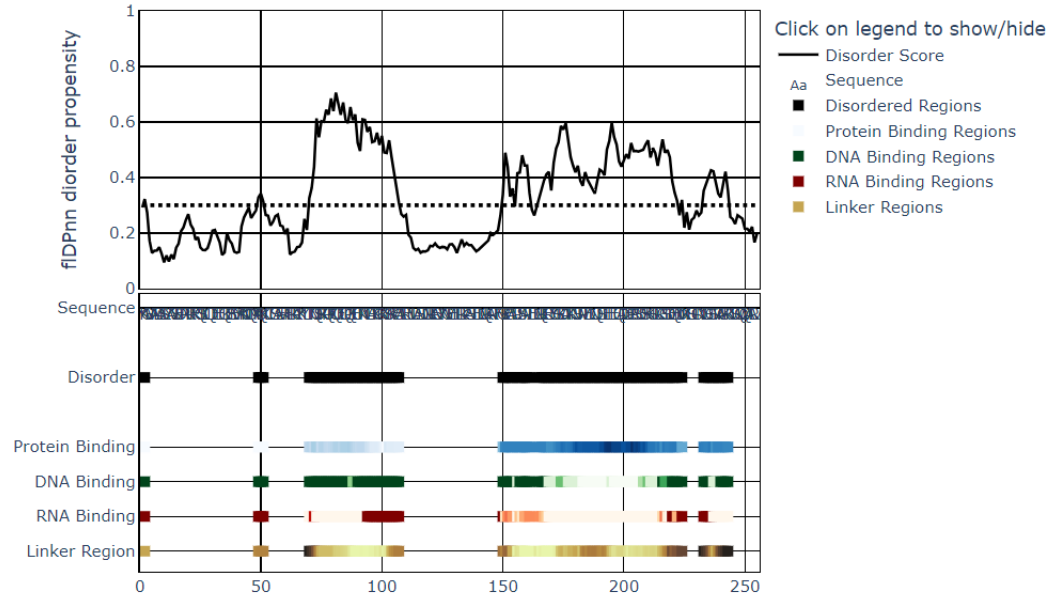

Fig. 12 FlDPnn Results for "Shugoshin C-terminal domain-containing protein." Match Begins at Residue 22 and Ends at Residue 62. Notably, Data is not Generated due to a Lack of Predicted Disorder

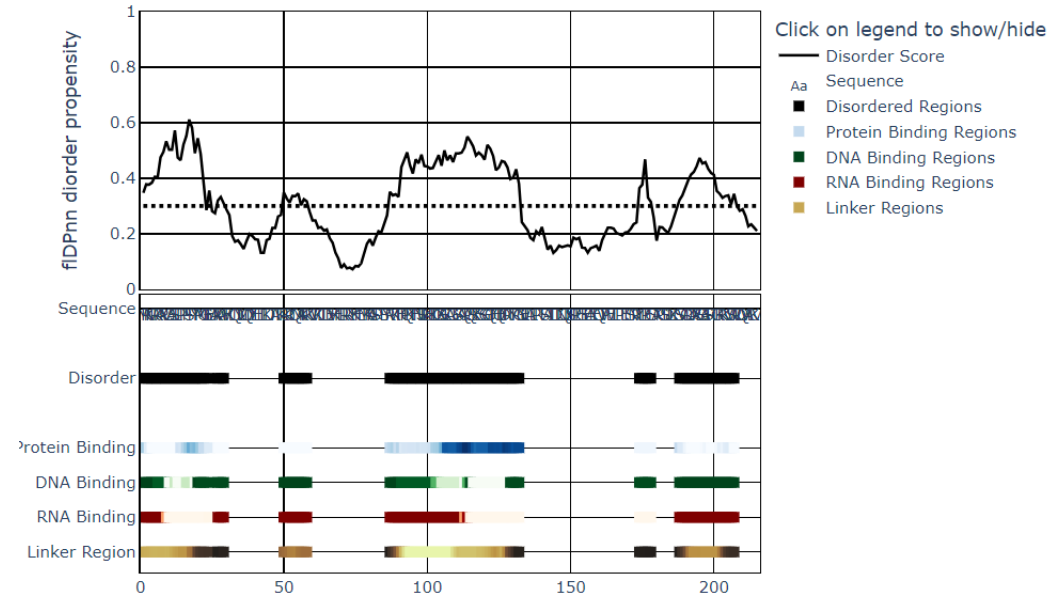

Fig. 13 FlDPnn Results for "Non-specific serine/threonine protein kinase." Match Begins at Residue 28 and Ends at Residue 68.

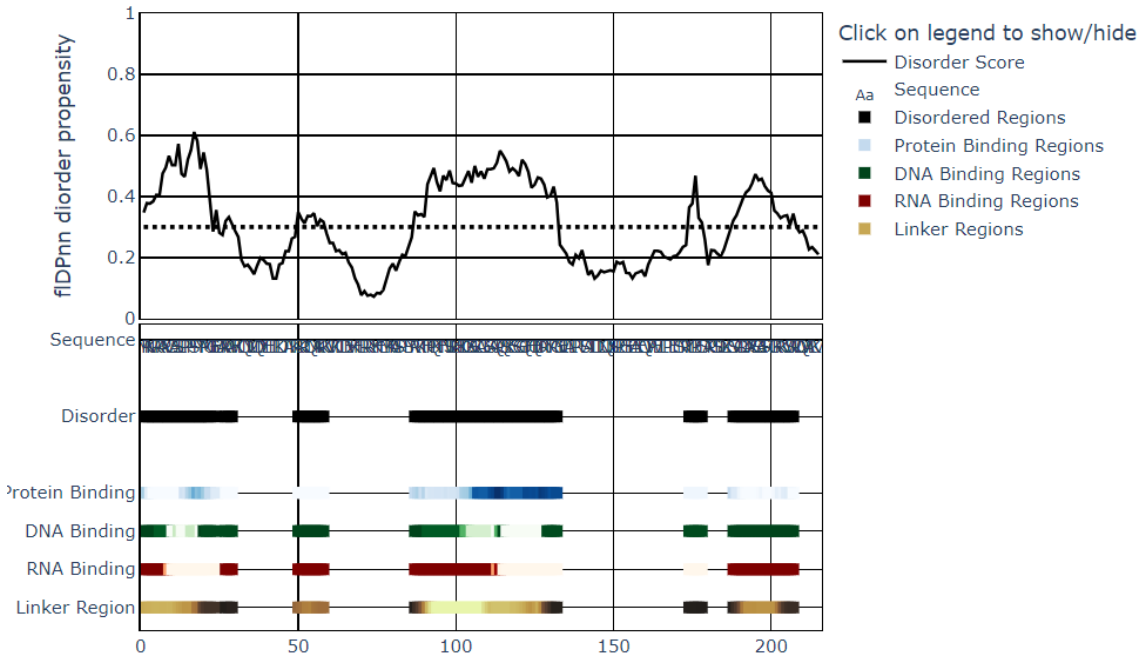

Fig. 14 FlDPnn Results for "TXP2 C-terminal domain-containing protein." Match Begins at Residue 24 and Ends at Residue 63

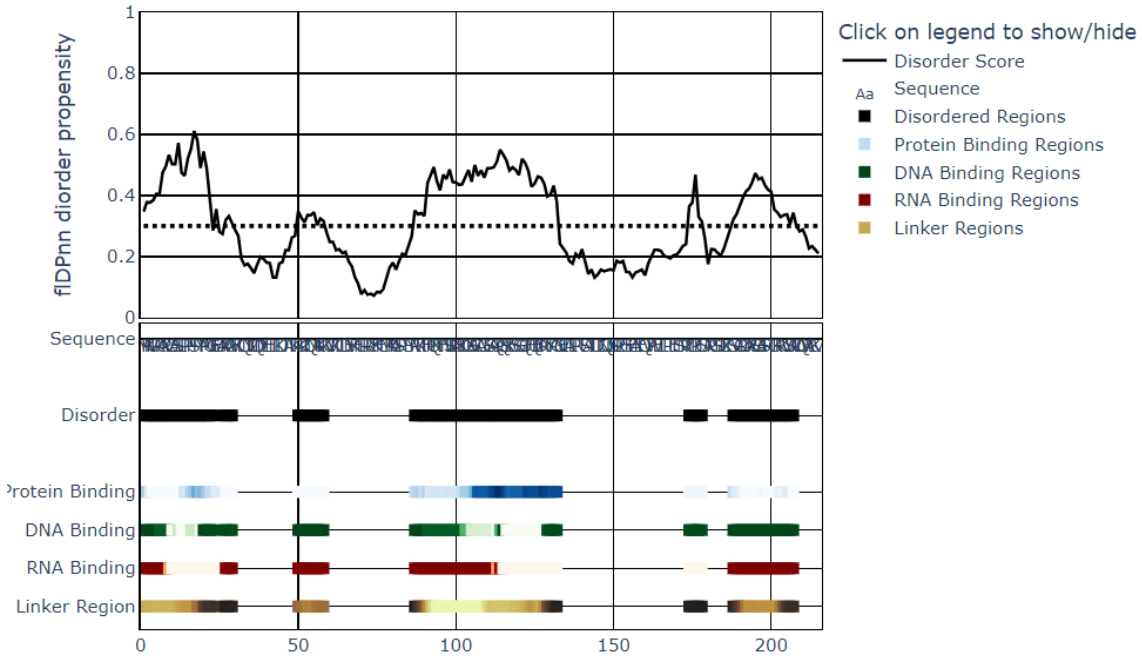

Fig. 15 FlDPnn Results for "Putative histone acetyltransferase HAC-like 1." Match Begins at Residue 34 and Ends at Residue 73

actions in *Landoltia punctata*. Thus, though the structure of UKP is very likely to be disordered, this study can only give suggestions, not confirmations, for the protein's function and how it contributes to the overall survival of the cell. Further research is appropriate to experimentally confirm UKP's role in LP.

Future Research

Experimental determination is the most appropriate way to conclusively determine the function of UKP and what other proteins it interacts with within *Landoltia punctata*. Because the results of this study suggest that UKP could potentially be involved in common cellular processes, researchers could justify utilizing expensive machinery to perform further testing.

First, a researcher could synthesize many copies of UKP. This can be done by inserting JZ987503.1 into a bacterial plasmid, where it can be transcribed and translated from a DNA sequence to an amino acid sequence. This methodology is somewhat common, having been used since the 1980s to synthesize proteins coded in the plasmid genome33. Because JZ987503.1 was derived from a bacterial plasmid, the same plasmid could be used when generating copies of the protein[25]. After collection, researchers could attempt to determine the structure of the protein through NMR (nuclear magnetic resonance imaging). Because NMR is uniquely *in vivo* and thus can be performed in solution to avoid protein damage, it could also be used to determine whether the protein is intrinsically disordered7. Afterward, researchers could use a pull-down assay, following the methodology used by Louche et al., where UKP would be tagged fluorescently, placed in an *in vitro* solution containing other proteins from LP, and re-extracted via centrifuge if it binds to other proteins[16]. By doing this, researchers could "pull down" and identify interacting proteins, helping determine the relevance of UKP within LP. Similar experiments can be conducted with nucleic acids to determine whether UKP binds to DNA, RNA, or both. Finally, researchers can conduct a western blot analysis on UKP to determine where and when the protein is expressed in LP.

Aside from UKP, future researchers could use the combination of flDPnn and AlphaFold as a preliminary step when conducting proteome-wide analyses to identify potentially relevant proteins. As the programs were demonstrated to be generally accurate and can be run natively, quickly, and in parallel, they could be used to predict binding propensities and structures before conducting experiments, allowing for stronger candidates to be identified prior to experimental evaluation.

Limitations

The conclusions of this study are limited by the size of existing databases and the accuracy of protein structure and disorder predictors used.

Though the database matches showed relatively significant results, due to their varying functions and the lack of highly significant matches, conclusions cannot be confidently drawn from the database searches alone. If this study were to be performed in the future, the results for this step in the methods would likely change as more proteins are investigated.

AlphaFold's largely low confidence prediction suggests that UKP does not share many properties with the proteins in its training dataset outside of its alpha helix region. AlphaFold trained on experimentally determined structures from the Protein Data Bank (PDB)[9]; 96.3% of the structures of the average protein on PDB are ordered34. Thus, flDPnn, which exclusively trained off of disordered proteins and predicted that UKP is intrinsically disordered, is more likely to be correct across the entirety of the protein. Although AlphaFold's output warrants consideration because of the high-confidence alpha helix, the program always assumes the protein is ordered. Thus, the validity of its results is questionable.

FlDPnn's predictions are more likely to be correct, but the program is not perfectly accurate. FlDPnn is the AUC curve is 0.814, which is less than 1.0[22]. Because flDPnn predicted high DNA/RNA binding region propensities and also because flDPnn's AUC values for DNA and RNA binding are significantly higher than its protein binding AUC values, the suggestion that UKP binds to nucleic acids is the most likely to be true. However, one must still account for the potential error of flDPnn when considering the validity of the conclusions of this study. Thus, the conclusion generated from the program remains partial.

Furthermore, there is a possibility that JZ987503.1 is an incomplete sequence. The sequence was derived from the reverse transcription of messenger RNA (mRNA), which is directly synthesized from a DNA sequence[25]. mRNA is an intermediate step when converting a DNA sequence into a protein sequence. Due to RNA's instability, JZ987503.1's mRNA complement may have degraded and lost a segment of itself before UKP's start codon appears, meaning that UKP may actually start beyond the beginning of the published DNA sequence and UKP could be significantly longer than JZ987503.1 suggests. This is unlikely, as the preliminary NCBI BLAST search for the study yielded similar, unnamed sequences of similar lengths in the est databases. If this mRNA degraded, every other match in the est database would have also had to degrade at almost the same spot.

There is also a possibility that the protein coded by JZ987503.1 is actually on ORF2 or ORF3. There exist other proteins that have around 40 residues. Signal peptides, for example, are around 20-40 residues long[33]. Though unlikely, it is important to consider this possibility when evaluating the conclusion of this study.

Implications

Though the conclusions are at most partial, since UKP likely interacts with nucleic acids, the protein could be involved in the transcription or translation processes that regulate the expression of other proteins. In this case, understanding UKP's role in *Landoltia punctata* could help researchers better understand how the plant can carry out cellular functions. For example, UKP could work with other proteins to increase the production of growth-controlling proteins during cell growth phases. If this is true, knowing how UKP interacts in LP would potentially allow researchers to modify the protein to increase protein output, thus increasing the growth rate of LP and making the plant more suitable for bioremediation and biofuel production[2]. Furthermore, if UKP helps regulate starch-producing or accumulating proteins, researchers could optimize the starch storage process in LP to make the plant more suitable for biofuel production[2]. Similar lines of reasoning can be applied to other potential applications of LP. Though the results generated by this study are not immediately valuable, this study suggests further research into UKP due to its potential to be involved in these applications.

Methods

Since only the sequence of UKP is available, traditional experimental methods were not feasible without a laboratory and the tools necessary to synthesize JZ987503.1. Instead, this study used secondary analysis in the form of database searches and structure prediction models such as AlphaFold to determine UKP's structure, function, and relevance within LP. Because the study solely relies on publicly available artificial intelligence programs and secondary data, no further ethical considerations were taken into account when gathering data.

First, the researcher translated JZ987503.1 into a protein sequence by inputting the GenBank code into NCBI's open reading frame (ORF) finder. This step was also used to determine the gap of this paper. ORF finder is a program that considers all six possible reading frames of a DNA sequence and visualizes all proteins that could be produced. Out of the three possible sequences, ORF1 was the longest, with ORF2 and ORF3 being around 40 residues long, shorter than the typical protein domain[34]. This means that ORF1 is the most likely to code for a functional protein sequence. To ensure this was true, the author conducted a BLASTp search on each ORF output. ORF1 outputted several, albeit uncharacterized, matches, while neither ORF2 or ORF3 outputted any matches above 1E-05. So, the researcher considered UKP to have the sequence of ORF1. Refer to Fig. 16 to see all possible reading frames.

After determining the sequence of UKP, the researcher ran it through the AlphaFold 2.3 Colab notebook to predict its three-dimensional structure. Pentony et al. utilized structure prediction models in a study to predict the structure of proteins, so it can be inferred that using artificial intelligence is justifiable, provided the limitations are considered[5]. Though the collab notebook version of AlphaFold is slightly simplified, its accuracy on monomer strands is not affected[9]. The researcher used the monomer folding option and set the iteration cycle count to its maximum of 20 to ensure the highest accuracy. With the generated structure, the researcher then searched several protein databases to find similar proteins and compared their AlphaFold predicted structures with the one of UKP. Using UKP's FASTA protein sequence, he conducted a BLAST search through all publicly available protein databases that contain plant sequences: the EBI AlphaFold database, all UniProt databases (including UniProtKB, SwissProt, and tREMBL), the Protein Data Bank (including computed structures), NCBI's protein databases (nr, tsa_nr, landmark, refseq_protein, etc.), the Database of Interacting Proteins, DisProt and MobiDB (for IDPs), NextProt, Prosite, Protein Information Resource (PIR), ModBase, SuperFamily, and SCOP. The researcher then used the 3-dimensional .pdb file produced with AlphaFold to search the FoldSeek and CATH structure databases. If he found any matches with names that denoted function, he ordered them in order of significance, determined by the match's E-value. Since he searched several databases, he standardized the E-values for each match by using NCBI's "Align Two Sequences" tool.

The author also ran a control sequence on this simplified Colab notebook to ensure the methodology would not produce errors. He selected protein 1TKG, as it is similarly monomeric and relatively short (224 residues long), to compare the accuracy of AlphaFold[35]. When repeating the steps used with UKP, the program outputted a result with an RMSD of 0.434 when compared to the ground truth. This signals that this version of AlphaFold is accurate with protein structure prediction and the methodology used with UKP should not produce any errors.

However, the database searches yielded many named proteins with different functions and similar significance and structures. Furthermore, AlphaFold had a large, low-confidence region in its prediction for UKP. See Fig. 17 for AlphaFold's prediction of UKP. AlphaFold's low confidence region suggested that UKP does not share many similarities with proteins in its training dataset. Since AlphaFold trained off of PDB entries, 96.3% of which are ordered, this result, along with the lack of significant database matches, signaled that UKP may be intrinsically disordered and that further analysis was necessary[9,36]. So, the researcher ran UKP through flDPnn to help confirm or deny AlphaFold's prediction of UKP's structure.

To once again ensure accuracy, 1TKG was run through flDPnn with the same parameters as UKP. The program predicted that 1TKG was largely ordered and that the small disorder region (which may be explained by the induced-fit nature of enzymes) binds solely to RNA. As 1TKG is known to bind to modified adenosines, this output supports the program's accuracy[35]. See Fig. 18 for the 1TKG flDPnn results.

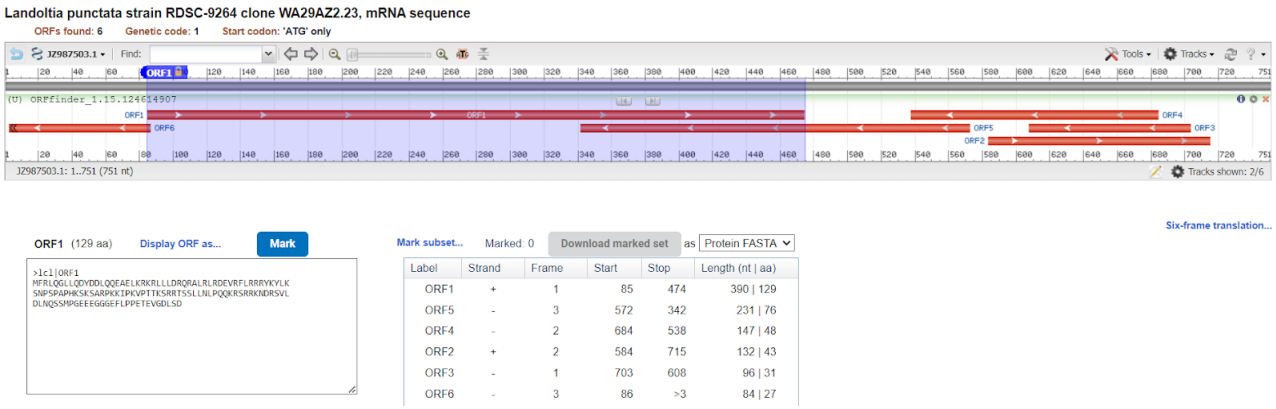

Fig. 16 NCBI's ORF Finder Results Show that ORF1 is 129 Amino Acids Long

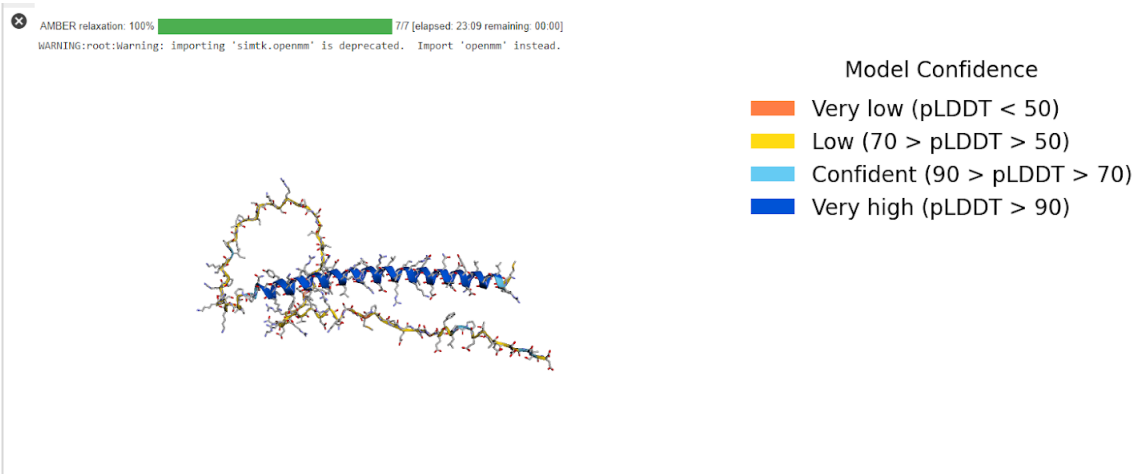

Fig. 17 AlphaFold Three-dimensional Prediction of UKP

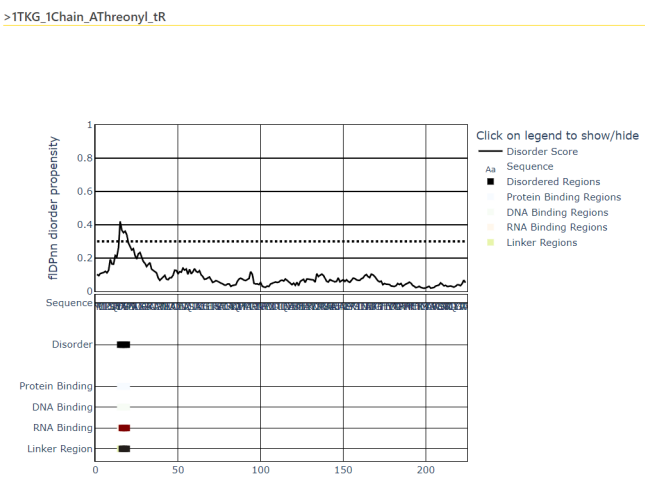

Fig. 18 FlDPnn Results for 1TKG

FlDPnn predicted that all of UKP was disordered and that approximately the first 30 residues likely bound with either DNA or RNA. Similarly, AlphaFold predicted that around the first 50 residues were an alpha helix. These predictions signaled that the beginning of the protein was an area of interest, potentially being the functional region in UKP. Furthermore, the last 52 residues were predicted to likely bind to proteins by flDPnn, suggesting that this portion of the protein may also be functional. To eliminate distracting matches, the researcher searched the previous databases with just the first 41 residues, the midpoint between the length of interest predicted by AlphaFold and flDPnn, and just the last 52 residues, looking for named matches. Like the full sequence search, he standardized the E-value by using NCBI's "Align Two Sequences" tool.

Once again, the researcher was given several results with similar E-values but different functions. He had initially intended to use BioGRID, a database of known protein interactions, to determine what proteins UKP was likely to interact with, but BioGRID only takes the name or code of a protein as a valid

input. Unfortunately, because he could not find a distinct match in my database searches, he could not confidently determine the precise function or name of UKP and was unable to continue with this step. This concluded the data collection process.

Acknowledgments

Thank you to Dr. Peter Kahn, a former professor of Rutgers University, who guided my research methodology and gave me direction from across the country, and to Dr. Andrew Vershon, who taught the WSSP program where I learned about *Landoltia punctata* and proteins and sequenced JZ987503.1. I would also like to thank the creators of AlphaFold and flDPnn. This paper would not exist had they not laid the foundation for my analysis. Finally, thank you to my parents, who supported me throughout the process, and to my cat, who was of no help at all, but was very cute.

References

1 A. F. Miranda, N. R. Kumar, G. Spangenberg, S. Subudhi, B. Lal and A. Mouradov, *Plants*, 2020, **9**, 437.

2 A. Faizal, A. A. Sembada and N. Priharto, *Saudi Journal of Biological Sciences*, 2021, **28**, 294–301.

3 N. Wang, G. Xu, Y. Fang, T. Yang, H. Zhao and G. Li, *Molecules*, 2014, **19**, 6623–6634.

4 N. M. Al-Abd, Z. M. Nor, M. Mansor, F. Azhar, M. S. Hasan and M. Kassim, *BMC Complementary and Alternative Medicine*, 2015, **15**, year.

5 M. M. Pentony, P. Winters, D. Penfold-Brown, K. Drew, A. Narechania, R. DeSalle, R. Bonneau and M. D. Purugganan, *Genome Biology and Evolution*, 2012, **4**, 360–371.

6 C. M. Alberini and E. Klann, *Alpha helix - an overview — ScienceDirect topics*, https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/alpha-helix#:~:text=The%20NH2%20terminus%20of, 2014, Accessed: 2023-09-21.

7 D. H. Lysak, K. Downey, L. S. Cahill, W. Bermel and A. J. Simpson, *Nature Reviews Methods Primers*, 2023, **3**, year.

8 M. K. Singh and A. Singh, *Nuclear magnetic resonance spectroscopy - an overview — ScienceDirect topics*, https://www.sciencedirect.com/topics/materials-science/nuclear-magnetic-resonance-spectroscopy, 2016, Accessed: 2023-09-21.

9 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

10 R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis, *Protein complex prediction with AlphaFold-multimer*, https://doi.org/10.1101/2021.10.04.463034, 2021, Accessed: 2023-09-21.

11 K.-T. Wang, M.-C. Hong, Y.-S. Wu and T.-M. Wu, *Plants*, 2021, **10**, 1576.

12 W. R. Pearson, *Current Protocols in Bioinformatics*, 2013, **42**, 3.1.1–3.1.8.

13 NCBI, *Frequently asked questions — BLASTHelp documentation*, https://blast.ncbi.nlm.nih.gov/doc/blast-help/FAQ.html, Accessed: 2023-09-21.

14 Qiagenbioinformatics.com, *E-value*, https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/_E_value.html, 2021, Accessed: 2023-09-21.

15 M. Shatnawi, *Chapter 6 - review of recent protein-protein interaction techniques*, https://www.sciencedirect.com/science/article/abs/pii/B9780128025086000065, 2015, Accessed: 2023-09-21.

16 A. Louche, S. P. Salcedo and S. Bigot, *Methods in Molecular Biology*, 2017, **1615**, 247–255.

17 L. Lu, H. Lu and J. Skolnick, *Proteins: Structure, Function, and Genetics*, 2002, **49**, 350–364.

18 I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde, V. G. Stancheva, X. H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernández, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, A. M. E, J. C. Fromme, T. L. Hendrickson, Q. Cong and D. Baker, *Science*, 2021, **374**, year.

19 H. J. Dyson and P. E. Wright, *Nature Reviews Molecular Cell Biology*, 2005, **6**, 197–208.

20 H. J. Dyson, *Molecular BioSystems*, 2012, **8**, 97–104.

21 P. E. Wright and H. J. Dyson, *Nature Reviews Molecular Cell Biology*, 2014, **16**, 18–29.

22 G. Hu, A. Katuwawala, K. Wang, Z. Wu, S. Ghadermarzi, J. Gao and L. Kurgan, *Nature Communications*, 2021, **12**, 4438.

23 W. Basile, M. Salvatore, C. Bassot and A. Elofsson, *PLOS Computational Biology*, 2019, **15**, e1007186.

24 N. Perdigão and A. Rosa, *High-Throughput*, 2019, **8**, 8.

25 A. Zatuchney, A. Vershon and J. Mead, *Landoltia punctata clone WA29AZ2.23 ribosomal RNA Small methyltransferase-like mRNA, partial sequence, mRNA sequence*, https://www.ncbi.nlm.nih.gov/nuccore/JZ987503.1?report=GenBank, 2023, Accessed: 2023-09-21.

26 WISE, *WISE — waksman student scholars program*, https://wssp.rutgers.edu/wise#:~:text=The%20WISE%20programs%20engage%20high, 2024, Accessed: 2023-09-21.

27 *Q9VQF9 · SNAPN_DROME*, https://www.uniprot.org/uniprotkb/Q9VQF9/entry, Accessed: 2023-09-21.

28 *Q6QNY1 · BL1S2_HUMAN*, https://www.uniprot.org/uniprotkb/Q6QNY1/entry, Accessed: 2023-09-21.

29 *CRT10 — SGD*, https://www.yeastgenome.org/locus/S000005424, Accessed: 2023-09-21.

30 P. Wadsworth, *Current Biology*, 2015, **25**, R1156–R1158.

31 Y. Yao and W. Dai, *Cell Cycle*, 2012, **11**, 2631–2642.

32 W. E. Hinckley, K. Keymanesh, J. A. Cordova and J. A. Brusslan, *Plant Direct*, 2019, **3**, year.

33 C. Dolan, C. S. Burke, A. Byrne and T. E. Keyes, *Plant Cell Biology*, 2017, **2**, year.

34 D. Xu and R. Nussinov, *Folding and Design*, 1998, **3**, 11–17.

35 *RCSB PDB - 1TKG: crystal structure of the editing domain of threonyl-tRNA synthetase complexed with an analog of seryladenylate*, https://www.rcsb.org/structure/1tkg, 2014, Accessed: 2023-09-21.

36 Y. Zhang, B. Stec and A. Godzik, *Structure*, 2007, **15**, 1141–1147.