Gene Cell Tissue. 2023 January; 10(1):e122297. https://doi.org/10.5812/gct-122297.

Published online 2022 July 23. **Research Article**

Identification of the Properties and Function of the Unknown Protein with Accession Number AT2G15110.1 on the TAIR Website

Shahnam Azizi-Dargahlou[1] and Bahman Fazeli-Nasab[2, 3, *]

[1]Department of Biotechnology, Azarbaijan Shahid Madani University, Tabriz, Iran
[2]Department of Agronomy and Plant Breeding, Agriculture Institute, Research Institute of Zabol, Zabol, Iran
[3]Department of Biotechnology and Plant Breeding, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran

[*] *Corresponding author*: Department of Agronomy and Plant Breeding, Agriculture Institute, Research Institute of Zabol, Zabol, Iran. Email: bfazeli@uoz.ac.ir

Received 2022 January 01; **Revised** 2022 March 05; **Accepted** 2022 March 08.

Abstract

Background: Nowadays, with the help of genomics and proteomics, numerous genes and proteins have been discovered, however, the function and role of most of them are still unknown. Using bioinformatics tools can be a major step in the identification of the function of these genes and proteins.
Methods: In this study, we applied various bioinformatics software to identify the unknown protein properties with the AT2G15110.1 accession number on the Arabidopsis Information Resource website. Operations, such as identification of general protein properties, blasting amino acid sequences, detection of motifs and domains in the sequence, examination of the second and third protein structures, exploration of ligands, assessment of proteins involved with the target protein as well as recognition of the target protein location in the cell, were carried out.
Results: The results showed that the query protein had no significant homology in terms of sequence, three-dimensional structure, and any interaction with known proteins. Additionally, it was observed that the presence probability of this protein in the nucleus organelle was more than in other organelles, and it only has one domain of unknown function.
Conclusions: The results of this study can be basic information for other researchers who seek to identify this unknown protein and determine the right pathway for the identification of mentioned protein function through bioinformatics tools or laboratory methods. Based on the results of this study, laboratory methods are recommended for subsequent studies.

Keywords: *Arabidopsis thaliana*, Bioinformatics, Database, Protein Function, Unknown Protein

1. Background

Currently, more than six million unique protein sequences are stored in public databases, and this number is growing rapidly (ncbi.nlm.nih.gov/RefSeq). In addition, despite numerous advances in the determination of structures, only the structure of 50,000 proteins has been laboratory and experimentally validated. This apparent difference between the identification of protein sequences and the determination of their structure has led researchers to turn to computer-aided data analysis (1-4). The determination of the function of all these proteins by laboratory methods can be extremely time-consuming, costly, and even impossible in some cases. Two basic strategies are widely used to predict the role of proteins, one is based on the similarity of sequences and the other one is based on the structural similarity of proteins (5-7).

Unknown-function proteins are listed in different databases under different headings, including hypothetical, putative, and unknown proteins, which have a more well-known sequence regarding the order of the mentioned names. Hypothetical proteins are proteins that have been predicted by nucleic acid sequencing methods, and no chemical experimental evidence has shown the existence of these proteins; nevertheless, the existence of unknown proteins has been proven in vitro, although their role and function are still unknown. In this study, an unknown protein was randomly selected from the Arabidopsis Information Resource (TAIR) website (arabidopsis.org), belonging to the plant *Arabidopsis thaliana*.

Arabidopsis thaliana is a flowering plant in the Brassica family that includes economically important genera of Brassica and mustard. *Arabidopsis thaliana* was the first plant whose genome was sequenced. Although this plant is not important in terms of economic value, it is an ideal plant for research due to its small size, short growth period, and small genome. This plant has five chromosomes in the haploid stage, which has 135 Mb with 32,000 genes

Table 1. General Characteristics of the Unknown Protein Identified by AT2G15110.1 on Arabidopsis Information Resource Website[a]

Target Protein Sequence with Access Number AT2G15110.1 on the TAIR Website	Properties			External IDs
	Calculation of Molecular Weight	Calculation of Isoelectric Ph	Length In Terms of Amino Acids	
MPPRKVVREVFLIDGKFEKYKTSLSTSSRLLLLRGAHQ IPHQIPLISPEPTVCPENPPPGHPSEEDRFSLSLNDLL QLYAVKKGRTKGTFFLSPRKGFRVFDDFPDKDEQWR KSYFFFPVNDLTYGNKTGLFVSEWAARTDLGWESLT IDRIRASGRRIRSRTDLAVSSPPFCPRIIDKADMSLPSS RQTTNKASVSAGKKPETPTSGSGKTIKDPAGKDSEKR AADKKRKQPEETNPSPPRSSRPRHEEKGAKLKGIVKE APQNLVVLSSRESETCESERRNVPLPAPPMTFADTMR TLVPPGSAIAPFDEMKEVNKENYLRFARKLGKLILEF NSVFCSHEDQLFDKDVEIESFKRSEDENAKAVEKAN KVMNRMKAAELQVQKLEVNNIDLTAKLKAGKNAY LDAIEKETQARADLRTCKEKMKKMEEEQAEMIVAA RTDERRKVRAQFHDFSSKYGNFFKESEEVETLKVRV AEAKANRELLEEIEKGEIPDLSKELESVRADEEKFAR HAAEPKTPRPDPTELTSLLADTPSEVAAESIPPAEVAII DEGGSNKGSTSEAGIAAMFPVDVEKDSGKTE	64421.5	8.19	583 aa	GenPept: 238479252; UniProtKB: F4IHH0-1

[a] TAIR, The Arabidopsis Information Resource (https://www.arabidopsis.org)

encoding proteins (8-11). In this study, an unknown protein with an access number or ID or AT2G15110.1 ID was selected from the TAIR site. Then, attempts were made using different databases to identify the properties and functions of the mentioned protein.

2. Methods

Firstly, the unknown protein was selected from arabidopsis.org, and then its access number was extracted from various databases. Operations, such as identification of the general properties of the desired protein, blasting the amino acid sequence, exploration of the motifs and domains in its amino acid sequence, examining the second, including alpha-helix and beta-sheets, and third structure of the target protein, recognition of the ligands, assessment of proteins involved with the target protein, and the determination of the protein location in the cell, were carried out.

3. Results and Discussion

In the initial investigation in order to understand the general characteristics of this unknown protein, the data shown in Table 1 were obtained.

Protein blasting was then performed at the National Center for Biotechnology Information website (blast.ncbi.nlm.nih.gov). Among the records that had the most similarity with the target sequence, records 1 - 9 were unknown proteins. 15 of the records that were most similar to the target protein were selected and the phylogenetic tree of these records was drawn in MEGA 7 software (version 7) (12) after the alignment operation (Figure 1). It was observed that the unknown protein of interest did not has any common ancestor with any of the known proteins of these 15 records.

The unknown protein of interest blast was performed with dedicated access number of NP_179115.2 ** on the National Center for Biotechnology Information (NCBI) website. After the operations, the phylogenetic tree of these 15 records was drawn using MEGA software (version 7). It was observed that the protein in question had no common ancestor with any of the known proteins of these 15 records.

* The asterisk indicates the searched protein.

** This is the unknown protein of interest access number on the NCBI website that is different from that of the TAIR website.

In the next step, the aforementioned protein was searched for motifs and domains in different databases. In all domain searched databases, it was observed that the searched protein had only a domain of unknown function (DUF) called DUF601 (Figure 2).

The DUF domains are protein domains that have no known function. This family of domains is collected in the Pfam database (pfam.xfam.org) and is prefixed with a DUF followed by a registration number, such as DUF2992 or DUF1220. The protein we are looking for has only one DUF and is named DUF601 in the Pfam database. This domain is located at position 186-469 of the amino acid sequence. Additionally, this unknown protein sequence had three low complexity region (LCR) (Figure 2). In proteins, LCRs are places where one or more amino acids are found in abundance. Due to their high abundance and potential for the ability to propagate in a short time through replication slippage, they can significantly contribute to increasing protein sequence length and producing new protein functions. However, little information is available on the overall impact of LCRs on protein evolution (13).

The identification of the function of the unknown protein was continued by examining its three-dimensional (3D) structure at the Phyre2 website (14), and the results are shown in Figure 3. At this stage, no significant similarity was observed between the desired protein and the proteins

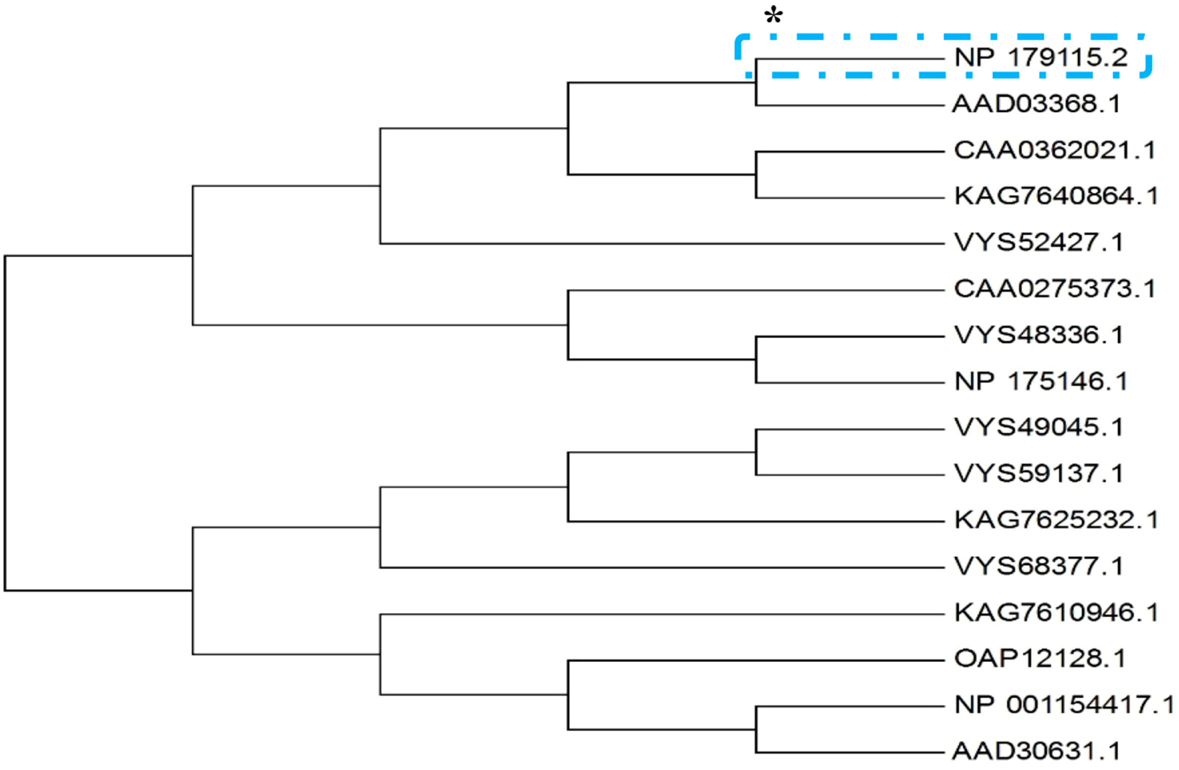


Figure 1. Phylogenetic tree related to the initial 15 records obtained from the unknown protein of interest blast using MEGA software (version 7)

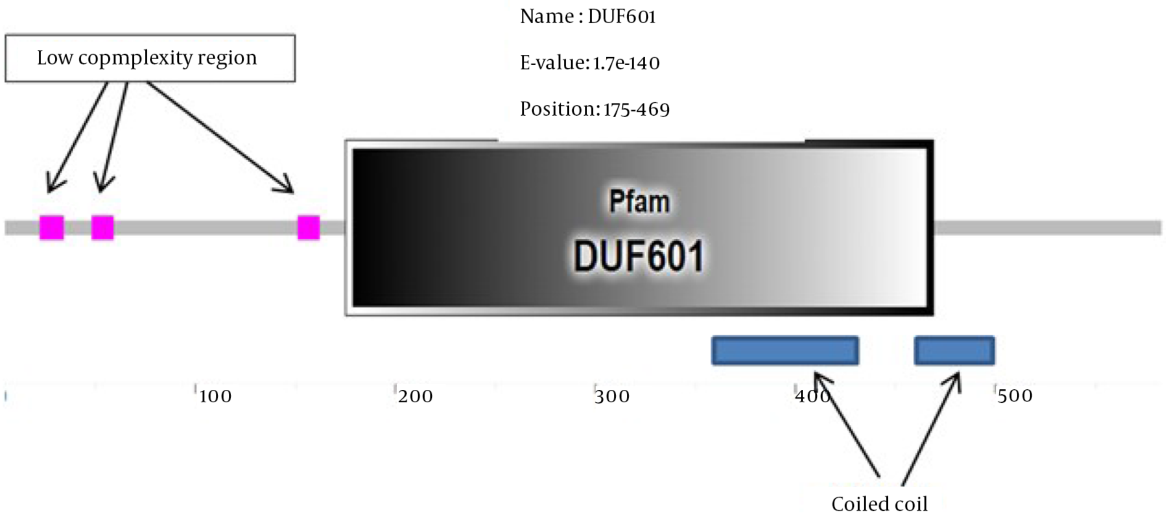
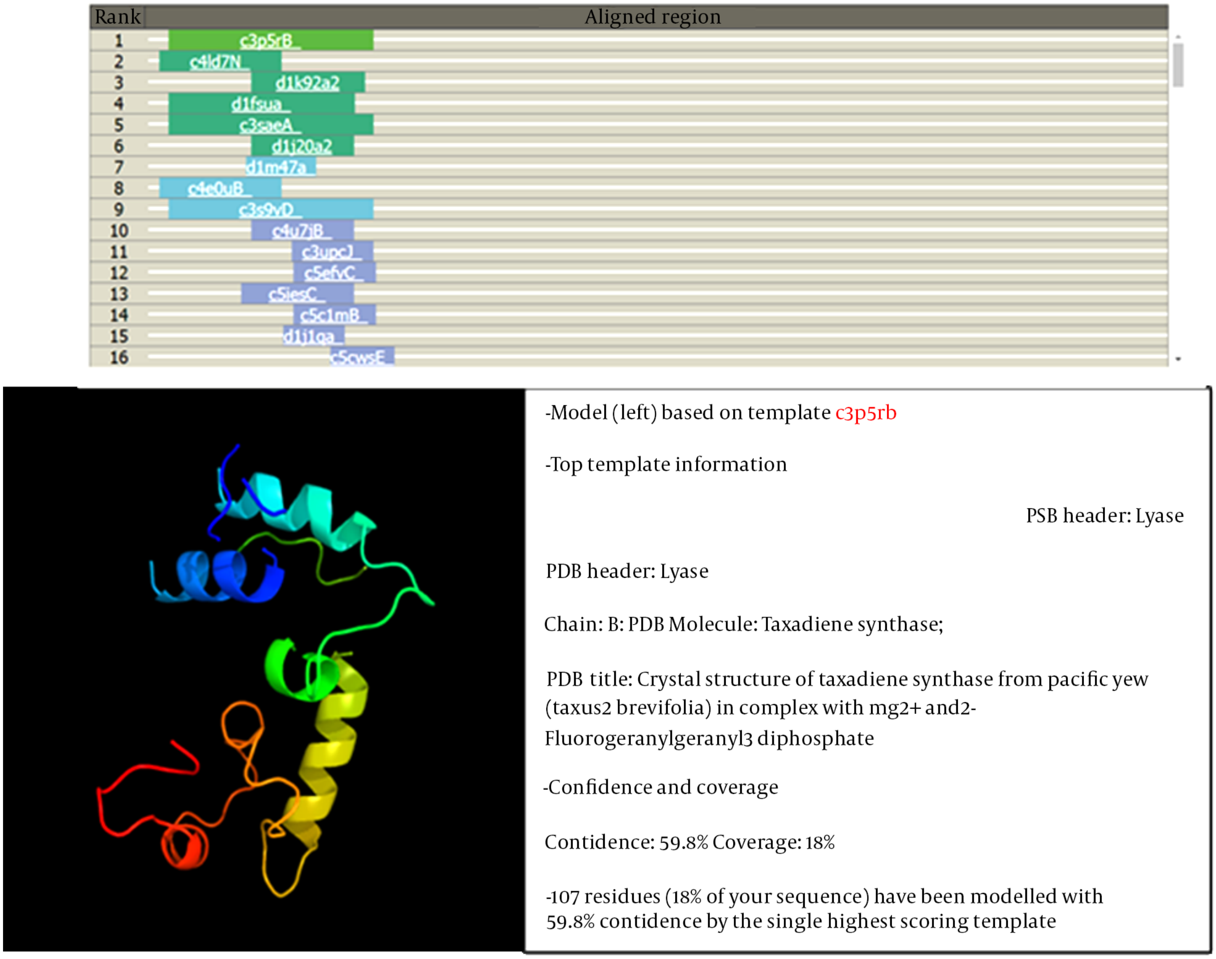


Figure 2. Results from the Pfam database to identify the domain in the requested protein, including a domain of unknown function (DUF) called DUF601 and two sequences containing a coiled-coil structure and three structures of low complexity region.



Rank	Aligned region
1	c3p5rB_
2	c4ld7N_
3	d1k92a2
4	d1fsua_
5	c3saeA_
6	d1j20a2
7	d1m47a_
8	c4e0uB_
9	c3s9vD_
10	c4u7jB_
11	c3upcJ_
12	c5efvC_
13	c5iesC_
14	c5c1mB_
15	d1j1qa_
16	c5cwsE_

-Model (left) based on template c3p5rb

-Top template information

PSB header: Lyase

PDB header: Lyase

Chain: B: PDB Molecule: Taxadiene synthase;

PDB title: Crystal structure of taxadiene synthase from pacific yew (taxus2 brevifolia) in complex with mg2+ and2-Fluorogeranylgeranyl3 diphosphate

-Confidence and coverage

Contidence: 59.8% Coverage: 18%

-107 residues (18% of your sequence) have been modelled with 59.8% contidence by the single highest scoring template

Figure 3. Results of studying the three-dimensional (3D) structure of the searched protein with the 3D structure of known proteins.

in the databases regarding 3D structure.

Among the identified models, the only model that was somewhat similar to the 3D structure based on the searched protein was c3p5Rb. This sample covered 18% of the searched protein, so its function could not be generalized to the searched protein.

In the next step, the investigation was continued by following the relationship of the sought protein with other proteins that were carried out in UniProtKB database data (uniprot.org/uniprot) and from the STRING section. Figure 4 depicts the obtained results.

None of the proteins that interacted with the studied protein were known proteins, and only two of the mentioned proteins had the domain of unknown function (DUF601) in common with the target proteins, which are shown in red colour (Figure 4).

Finally, the location of the searched protein was examined using the data from different databases (15-17) in the cell, and various data were obtained. Most of the results confirmed the presence of the searched protein in the nucleus organ (Figures 5 and 6).

These data show that the presence of protein in chloroplast and mitochondrial organs and the secretion of that protein are very low, and the probability that it is in other organs is higher. This probability is 0.696; however, the validity of the results is low due to the value of RC, which is equal to 4.

Len, sequence length; cTP, chloroplast transit peptide; mTP, mitochondrial targeting peptide; SP, secretory pathway; a signal peptide; RC, reliability class, from 1 to 5, where 1 indicates the strongest prediction; Sign, meaning any other location

As it can be observed, most databases almost confirm the presence of the searched protein in the nuclear organ.

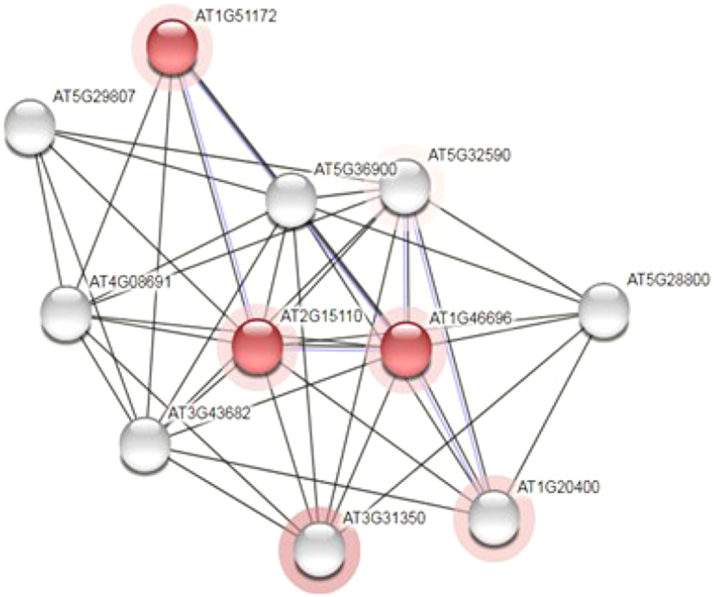
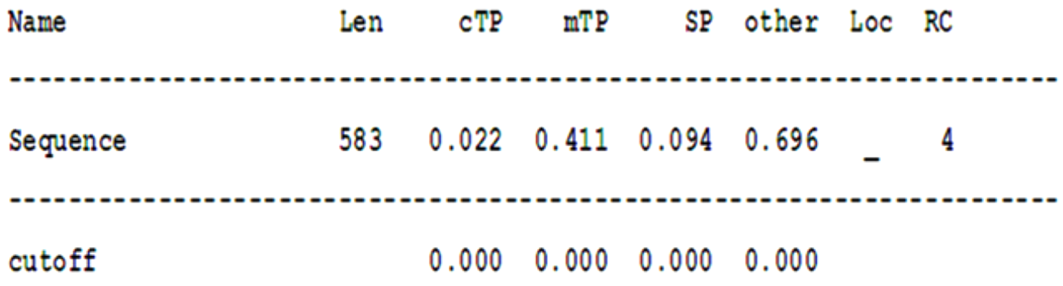


Figure 4. Results of Protein-Protein Interaction Test on UniProt Website with the searched protein (AT2G15110.1)



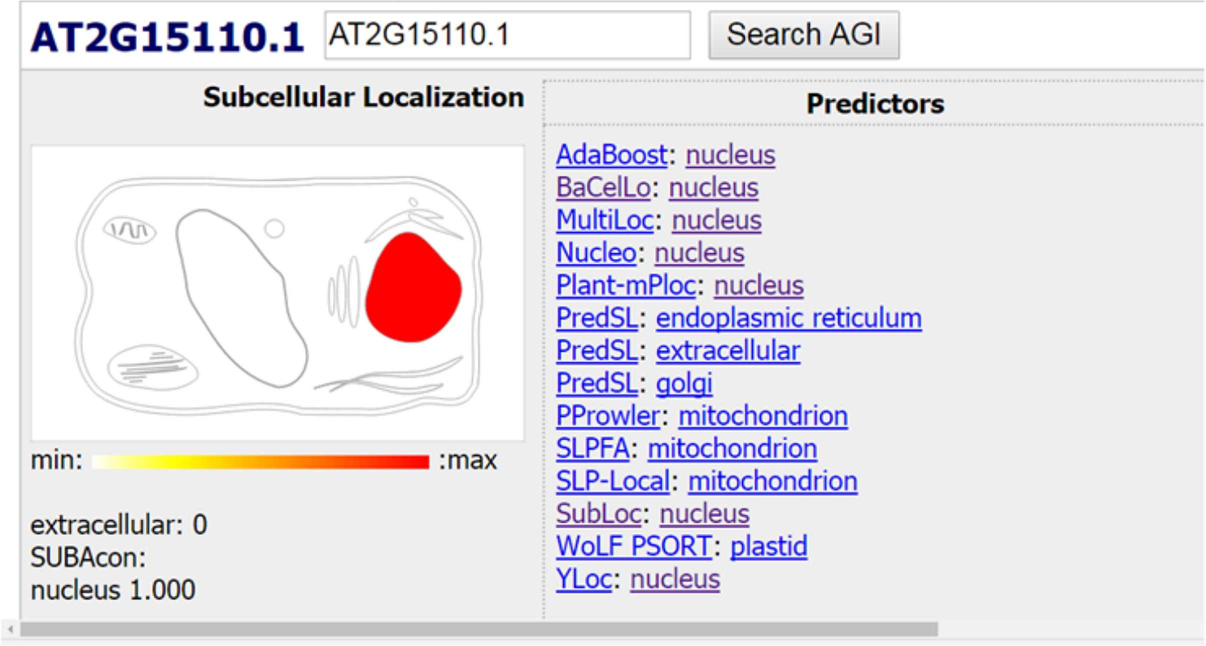
```
Name                  Len    cTP    mTP     SP  other  Loc  RC
----------------------------------------------------------------
Sequence              583  0.022  0.411  0.094  0.696   _    4
----------------------------------------------------------------
cutoff                      0.000  0.000  0.000  0.000
```

Figure 5. Determination of the location of the searched protein using TargetP Server Software (version 1.1)

3.1. Conclusions

Based on the data obtained from various databases and the interpretation of these results, it was shown that the protein searched under access number AT2G15110.1 on the TAIR website bears no resemblance in sequence to known proteins and the 3D structure of proteins identified to date. Nevertheless, it turned out that this protein has only one DUF called DUF601, and there was the possibility of the presence of this protein in the nuclear organ due to its protein sequence in most databases of protein location identification.

However, to detect the function of this unknown protein, the researchers can repeat the same operations performed on the protein over time because countless new data are added to the databases every day, and this new data might help accurately identify the function of this protein. Since the requested protein has only one domain, clues can be provided to the function of this protein by knocking it out or knocking it down. DUF domains often lack basic functions and might not be recognizable by mentioned methods. Nevertheless, we can find its location and function by attaching reporter genes to it and transiently expressing it, or by using other methods to analyze protein function in a laboratory environment.



Subcellular Localization

extracellular: 0
SUBAcon:
nucleus 1.000

Predictors

AdaBoost: nucleus
BaCelLo: nucleus
MultiLoc: nucleus
Nucleo: nucleus
Plant-mPloc: nucleus
PredSL: endoplasmic reticulum
PredSL: extracellular
PredSL: golgi
PProwler: mitochondrion
SLPFA: mitochondrion
SLP-Local: mitochondrion
SubLoc: nucleus
WoLF PSORT: plastid
YLoc: nucleus

Figure 6. Search results for the location of the searched protein in the data of different databases

Acknowledgments

This paper was prepared based on the material and equipment of the Agricultural Biotechnology Research Institute, University of Zabol, Iran, for which the authors would like to express their gratitude.

Footnotes

Authors' Contribution: All authors had an equal role in study design, work, statistical analysis, and manuscript writing.

Conflict of Interests: All authors declare no conflict of interest.

Data Reproducibility: All the data are embedded in the manuscript.

Funding/Support: This study was supported by the Research Institute of Zabol.

References

1. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc*. 2009;**4**(3):363–71. [PubMed ID: 19247286]. https://doi.org/10.1038/nprot.2009.2.
2. Reynolds CR, Islam SA, Sternberg MJE. EzMol: A Web Server Wizard for the Rapid Visualization and Image Production of Protein and Nucleic Acid Structures. *J Mol Biol*. 2018;**430**(15):2244–8. [PubMed ID: 29391170]. [PubMed Central ID: PMC5961936]. https://doi.org/10.1016/j.jmb.2018.01.013.
3. Du Z, Su H, Wang W, Ye L, Wei H, Peng Z, et al. The trRosetta server for fast and accurate protein structure prediction. *Nat Protoc*. 2021;**16**(12):5634–51. [PubMed ID: 34759384]. https://doi.org/10.1038/s41596-021-00628-9.
4. Zheng W, Zhang C, Bell EW, Zhang Y. I-TASSER gateway: A protein structure and function prediction server powered by XSEDE. *Future Gener Comput Syst*. 2019;**99**:73–85. [PubMed ID: 31427836]. [PubMed Central ID: PMC6699767]. https://doi.org/10.1016/j.future.2019.04.011.
5. Sael L, Chitale M, Kihara D. Structure- and sequence-based function prediction for non-homologous proteins. *J Struct Funct Genomics*. 2012;**13**(2):111–23. [PubMed ID: 22270458]. [PubMed Central ID: PMC3375349]. https://doi.org/10.1007/s10969-012-9126-6.
6. Bilal I, Xie S, Elburki MS, Aziziaram Z, Ahmed SM, Jalal Balaky ST. Cytotoxic effect of diferuloylmethane, a derivative of turmeric on different human glioblastoma cell lines. *Cellular, Molecular and Biomedical Reports*. 2021;**1**(1):14–22. https://doi.org/10.55705/CMBR.2021.138815.1004.
7. Ercisli MF, Lechun G, Azeez SH, Hamasalih RM, Song S, Aziziaram Z. Relevance of genetic polymorphisms of the human cytochrome P450 3A4 in rivaroxaban-treated patients. *Cellular, Molecular and Biomedical Reports*. 2021;**1**(1):33–41. https://doi.org/10.55705/CMBR.2021.138880.1003.
8. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*. 2000;**408**(6814):796–815. [PubMed ID: 11130711]. https://doi.org/10.1038/35048692.
9. Sun XD, Yuan XZ, Jia Y, Feng LJ, Zhu FP, Dong SS, et al. Differentially charged nanoplastics demonstrate distinct accumulation in Arabidopsis thaliana. *Nat Nanotechnol*. 2020;**15**(9):755–60. [PubMed ID: 32572228]. https://doi.org/10.1038/s41565-020-0707-4.
10. Kobayashi F, Tanaka T, Kanamori H, Wu J, Handa H. Reference Genome Sequencing and Advances in Genomic Resources in Common Wheat-Chromosome 6B Project in Japan. *Jpn Agric Res Q*. 2021;**55**(4):285–94. https://doi.org/10.6090/jarq.55.285.

11. Schulman AH. Genome Size and the Role of Transposable Elements. *Genetics and genomics of brachypodium*. Springer; 2015. p. 81–106.
12. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016;**33**(7):1870–4. [PubMed ID: 27004904]. [PubMed Central ID: PMC8210823]. https://doi.org/10.1093/molbev/msw054.
13. Rado-Trilla N, Alba M. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evol Biol*. 2012;**12**:1–10. [PubMed ID: 22920595]. [PubMed Central ID: PMC3523016]. https://doi.org/10.1186/1471-2148-12-155.
14. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. 2015;**10**(6):845–58. [PubMed ID: 25950237]. [PubMed Central ID: PMC5298202]. https://doi.org/10.1038/nprot.2015.053.
15. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*. 2000;**300**(4):1005–16. [PubMed ID: 10891285]. https://doi.org/10.1006/jmbi.2000.3903.
16. Tanz SK, Castleden I, Hooper CM, Vacher M, Small I, Millar HA. SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in Arabidopsis. *Nucleic Acids Res*. 2013;**41**(Database issue):D1185–91. [PubMed ID: 23180787]. [PubMed Central ID: PMC3531127]. https://doi.org/10.1093/nar/gks1151.
17. Hooper CM, Tanz SK, Castleden IR, Vacher MA, Small ID, Millar AH. SUBAcon: a consensus algorithm for unifying the subcellular localization data of the Arabidopsis proteome. *Bioinformatics*. 2014;**30**(23):3356–64. [PubMed ID: 25150248]. https://doi.org/10.1093/bioinformatics/btu550.