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

Shahnam Azizi-Dargahlou¹ and Bahman Fazeli-Nasab², *

¹Department of Biotechnology, Azarbaijan Shahid Madani University, Tabriz, Iran
²Department of Agronomy and Plant Breeding, Agriculture Institute, Research Institute of Zabol, Zabol, 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, despite the existence 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 these proteins.

Methods: This study used various bioinformatics software to identify the unknown protein properties with the AT2G15110.1 accession number on the Arabidopsis Information Resource website. Operations, such as the identification of general protein properties, blasting amino acid sequences, the identification of motifs and domains in the sequence, the examination of the second and third protein structures, the identification of ligands, the identification of proteins involved with the protein, and the identification of the protein site 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 would 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 based on the similarity of sequences and the other 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 species of Brassica and mustard. A. 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 encoding proteins (8-11). In this study, an unknown protein with an
access number or ID or AT2G5110.1 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 extracted from arabidopsis.org, and then its access number was extracted from various databases. Operations, such as the identification of the unknown protein properties on the sequence, the identification of the general properties of the protein, amino acid sequence blast, the identification of the motifs and domains in the sequence, the examination of the second protein structure, including alpha helix and beta plates, the identification of the third protein structure, the identification of the ligand, the identification of proteins involved with the target protein, and the identification of the protein site in the cell, were carried out. Finally, putting together the data obtained from different databases, an attempt was made to describe all the characteristics and functions of the target protein or part of it.

3. Results and Discussion

In the initial study and with general knowledge 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 bore the most similarity to the target sequence, nine records were among the unknown proteins. Among the above-mentioned records, 15 records were selected. In MEGA software (version 7) (12), after performing the phylogenetic tree subchromatic operation, these records were drawn (Figure 1). It was observed that the required protein did not have any common ancestor with any of the known proteins of these 15 records.

The required protein blast was performed with access number 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 an unknown protein 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 numbered DUF601 in the Pfam database. The latter is in the amino acid sequence at positions 469-186. Additionally, the required protein sequence had three regions called the 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...
**Figure 1.** Phylogenetic tree related to the initial 15 records obtained from the required protein 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.
was observed between the desired protein and the 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. It covered the required protein, and the function of this structure could not be generalized to the required protein due to the small amount of protein coverage.

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 second domain of unknown function (DUF)601 in common with the required proteins, which are shown in red.

Finally, the position 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. There is a probability of 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.

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 because DUF domains often lack basic functions and might not be recognizable by these methods. Nevertheless, some of its conditions and functions can be detected by attaching the reporter genes to it, or other methods of protein function analysis can be used in laboratories.
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 University of Zabol with a grant number IR-UOZ-GR-2735.

References


