In-silico prediction of structural and functional aspects of a hypothetical protein of *Arabidopsis thaliana* (L) Heynh.

A .Bhattacharjee, H .Choudhury, U. Maheswari, S. R. Joshi

Abstract

Arabidopsis thaliana is a model plant for scientific research due to the presence of many desirable characteristics like rapid development, small plant size, mutable seeds, mutations quickly becoming homozygous etc. An *in-silico* technique was initiated to characterize a hypothetical protein to deduce its structural and functional information. The hypothetical protein analysed in the present study showed domain characteristics of ankyrin repeats family with beta-hairpin-alpha-hairpin repeat {multiple repeats of beta (2) - alpha (2) motif}. The protein showed five domains of 32 amino acid repeat units. The modelled protein revealed the presence of maximum number of random coils as its secondary structural elements. The existence of ankyrin repeats indicates its role in protein-protein interaction essential for various metabolic processes in organism.

Keywords: *Arabidopsis thaliana;* hypothetical protein, ankyrin repeats, protein-protein interaction.

Introduction

Arabidopsis thaliana (L) Heynh. is a small flowering plant that is widely used as a model organism [1] in plant biology. It has a small genome of about 125 Mb organized into five chromosomes and contains an estimated 25,500 genes. More than 30 megabases of annotated genomic sequence has already been deposited in GenBank by a consortium of laboratories in Europe, Japan, and the United States. The entire genome has been sequenced which has enhanced the importance of *Arabidopsis* as a model for plant biology. Its genome was sequenced in the year 2000. Most of the DNA encodes 25,498 genes and very little junk DNA is present. Many characters of this unique plant such as prolific seed production, rapid development, small plant size, mutable seeds, normally self pollinated thereby making mutations quickly becoming homozygous and its expression, make it an ideal model organism for scientific research.

The large scale genome sequencing project has generated a plethora of information both in terms of genes and proteins. There are however, a vast amount of proteins whose function and structure has not been unearthed yet. There is, therefore, an urgent need to characterize these hypothetical proteins whose only primary information in the form of sequence is available. The results generated will be helpful in gaining insight into the various metabolic, gene regulatory mechanism and the functional aspects of this unique model organism. Therefore, the present work involves the extensive use of tools and graphical software for a complete annotation of the hypothetical protein (Acc. No. gi|4206201|gb|AAD11589.1|).

Environmental stresses such as heat, cold, drought and high salinity influence plant growth and productivity. Plants respond and adapt to these stresses to survive under adverse conditions at physiological and biochemical levels. These stresses have been shown to induce the expression of genes with various functions in a variety of plants (Thomashow, 1999; Zhu, 2002; Shinozaki and Yamaguchi-Shinozaki 2003; Bartels and Sunkar, 2005). Some research with *Arabidopsis* has provided unexpected insights into cellular mechanisms shared with other organisms. For example, a protein complex initially identified through genetic analysis of the constitutive photomorphogenic class of *Arabidopsis* mutants has been found throughout eukaryotes and may provide clues to complex signal transduction networks active in humans (Wei *et al.,* 1998). A retinal photoreceptor that may serve to entrain the circadian clock in mammals was recently identified on the basis of, in part, similarity to the *CRY2* photoreceptor of *Arabidopsis* (Miyamoto and Sancar, 1998). Plant biologists have long realized that cellular mechanisms common to eukaryotes are often characterized first in yeast or animal systems and then later extended to plants. The advent of *Arabidopsis* functional genomics and the availability of large numbers of *Arabidopsis* mutants defective in known gene products provide a unique opportunity for plant biologists to contribute to research efforts in a variety of related disciplines. As a result, it will become increasingly important for those studying other groups of organisms to keep abreast of continuing advances in plant biology. Many biotechnology companies are counting on *Arabidopsis* research to help solve practical problems related to agriculture, energy and the environment. Significant advances have already been reported in applied research efforts including molecular cloning of disease resistance genes (Meinke *et al.,* 1998) engineering of plants

resistant to cold temperatures (Thomashow *et al.,* 2001) production of specialized hydrocarbons (Nawrath *et al.,* 1994) and stimulation of premature flowering in trees and other plants with extended life cycles (Weigel and Nilsson 1995). If patent applications are any indication of the practical benefits of *Arabidopsis* research, then the economic value of this simple weed has already been demonstrated (Meinke et al., 1998). One of the original ideas behind using *Arabidopsis* as a model system was to facilitate the identification of related genes of importance in crop plants.

Materials and Methods

To analyze the hypothetical protein and assign its functional and structural role, various tools and software were used. The primary sequence of the hypothetical protein (Acc. No. gi|4206201|gb|AAD11589.1|) was obtained from the GenBank at National Centre for Biotechnology Information (NCBI). The sequence was compared for detecting homologous sequences found in data bases using Basic Local Alignment Search Tool (BLAST). For calculation of the physico-chemical properties of the protein tools like Protparam were used. The secondary structure of the protein was analyzed using tools such as GOR. BLAST from NCBI was used to compare the query sequence with the database sequence to find its homologues. Conserved domains were detected from the BLAST analysis. Protein fold was recognized using SCOP (Structural Classification of Proteins) tool. Since only primary sequence information was available from NCBI, and no structure in the form of X-Ray crystallographic data was available from the Protein Data Bank (PDB), hence modelling of the protein had to be done to deduce the three dimensional structure of the protein. Here homology modelling was done using SWISS PDB Viewer and by selection of seven suitable templates obtained from Protein Data Bank. The structure of the protein developed was refined further to improve the model. The 3-D coordinate file was visualized in RASMOL.

Result

The similarity search for the sequence was carried out with the help of BLAST tool The results indicated the similarity to ankyrin repeat family protein of A. thaliana and showed moderate sequence similarity to ACD (accelerated cell death) (table 1). The domain analysis on the BLAST site indicated the dominance of ankyrin repeats (detected at five different regions in the sequence) at different location on the protein. The first domain identified in the protein sequence was positioned at 96 to 221 residues, the second domain at 115 to 255, the third domain at 150 to 290, the fourth at 233 to 357 and the fifth domain was located at 298 to 427th residues (figure 1). The hits on similar architecture of domain revealed multi-domain characteristics. For classification of protein, Structural Classification of Protein (SCOP) algorithm was used. It was found that the fold is formed of beta-hairpin-alpha-hairpin repeat [multiple repeats of beta (2)-alpha (2) motif]. The class of the protein was evaluated to be alpha and beta protein [(a+b) (mainly antiparallel beta -sheets, i.e. segregated alpha and beta regions)]. The super families for the domain were ankyrin repeat [(repeats organized in elongated structures and Plakin repeat) (repeats associate forming globular sub domains)] (table 2). The physicochemical properties of the hypothetical protein revealed the number of amino acid to be 564,

Table 1. BLAST result of the hypothetical protein

Accession Number	Similar hits	Score	E- value
gb|AAD11589.1|	hypothetical protein (Arabidopsis thaliana)	1147	0.0
gb|AAD11589.1	ankyrin repeat family protein (A. thaliana)	1070	0.0
gb|AAD48974.1|AF162444_6	contains similarity to Pfam family P…	579	2e-163
ref|NP_567285.1|	ankyrin repeat family protein (A. thaliana)	578	4e-163
ref|NP_192253.1|	ankyrin repeat family protein (A. thaliana)	516	2e-144
ref|NP_193175.2|	ankyrin repeat family protein (A. thaliana)	501	7e-140
ref|NP_192257.1|	ankyrin repeat family protein (A. Thaliana)	498	5e-139
ref|NP_192254.1|	ankyrin repeat family protein (A. thaliana)	464	7e-129
Ref|NP_567430.1|	ACD6 (ACCELERATED CELL DEATH 6)	457	1e-126

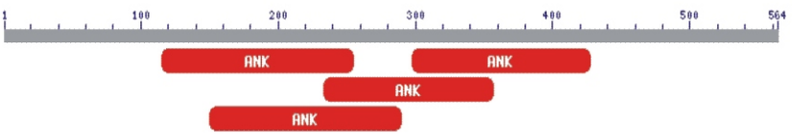

Figure 1: Putative conserved domains detected

Table 2: Structural Classification of the hypothetical protein using SCOP Database

Fold	Lineage	Super families
beta-hairpin-alpha-hairpin repeat multiple repeats of beta(2)-alpha(2) motif	1. Root: scop 2. Class: Alpha and beta proteins (a+b) Mainly antiparallel beta sheets (segregated alpha and beta regions)	1. Ankyrin repeat repeats organized in elongated structures 2. Plakin repeat repeats associate forming globular subdomains

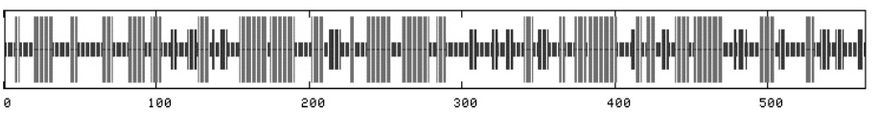

Figure 2: Graphical representation of secondary elements in the protein

¦= helix; ¦= Sheath; ¦= Coil
Sequence length: 564 residues

Table 3. Physicochemical properties of the protein (amino acid composition)

Amino acid	Number of residues	Percentage of residue
Ala (A)	39	6.9%
Arg (R)	38	6.7%
Asn (N)	34	6.0%
Asp (D)	31	5.5%
Cys (C)	13	2.3%
Gln (Q)	19	3.4%
Glu (E)	35	6.2%
Gly (G)	33	5.9%
His (H)	20	3.5%
Ile (I)	27	4.8%
Leu (L)	79	14.0%
Lys (K)	34	6.0%
Met (M)	11	2.0%
Phe (F)	17	3.0%
Pro (P)	20	3.5%
Ser (S)	40	7.1%
Thr (T)	24	4.3%
Trp (W)	3	0.5%
Tyr (Y)	14	2.5%
Val (V)	33	5.9%
Asx (B)	0	0.0%
Glx (Z)	0	0.0%
Xaa (X)	0	0.0%

Mol. Wt. of 63356.7 and the theoretical isoelectric points as 8.55. The maximum number of amino acids present in the sequence was found to be that of serine (7%) and the least was that of tryptophan (0.5%). Total number of positively charged residues (A + L) was 72 and the negatively charged residues (A + G) were 66. The instability index of the protein was computed to be 44.49. This classified the protein to be unstable. The grand hydropathicity was calculated to be 0.297 (table 3). The secondary structural analysis of the protein was done and random coil was found to be most frequent (45.21%), followed by alpha helix (40.60%). Extended strand (Ee) was found to be least frequent (14. 18%) (figure 2). The structure for the hypothetical protein was deduced by homology modelling. The structural information was obtained from the templates from PDB. The identities of these templates were 38%, which was fairly a good score to begin modelling. The modelled protein structure showed 92 H-bonds but no S-S bonds, 162 groups and 1252 atoms. The secondary structure is composed of 10 helices, 18 turns but no strands. In helices, there were 504 atoms and in turns, 347 atoms (figure 3). Seven templates from PDB were used to model the hypothetical protein. The templates selected exhibited various functions like cell cycle inhibition, cyclin dependent kinase inhibition, hormone growth factor and ankyrin repeats.

Figure 3: Three dimensional structure of the modeled hypothetical protein

Discussion

The analysis of the hypothetical protein showed sequence similarity mostly to the ankyrin repeat protein belonging to *A. thaliana.* The hits score with 100% alignment identity was observed between the two sequences, which indicates that protein with similar function exist in *A. thaliana*. The domain identified in the protein was characteristics of the ankyrin repeat (ANK) found in various diverse groups of proteins. Their presence in functionally diverse proteins such as enzymes, toxins and transcription factors strongly suggests domain shuffling. ANK repeat occurred in five consecutive copies along the entire length of the protein in an overlapping fashion (figure 1). One feature of the internal repeats is a predicted central hydrophobic alpha-helix, which is likely to interact with other repeats. The function of the ankyrin-like repeats is comparable with a role in protein-protein interactions required for various cellular processes, an observation in parity with Leila *et al.,* (2004) which shows a promise for further research. The repeats identified in the protein are tandemly repeated modules of 33 amino acids. The fold recognized was beta-hairpin-alpha-hairpin repeat with ankyrin repeat super family (table 2). The dominance of coiled regions indicates the high level of conservation and stability of the protein structure (figure 2 and 3). Ankyrin repeat is one of the most widely existing protein motifs in nature and exclusively functions to mediate protein-protein interactions, some of which are directly involved in the development of human cancer and other diseases (Li and Tsai, 2006). Further research involving development of appropriate strategies for studying these repeats using *A. thaliana* as model organism could be of significance in relation to the genes encoding the domains and its transfer.

Acknowledgement

The authors are thankful to the department of Biotechnology and Bioinformatics, and the Bioinformatics Centre, North-Eastern Hill University, Shillong, for providing research facilities.

References

Bartels, D. and R. Sunkar, 2005. Drought and Salt Tolerance in Plants. Crit. Rev. *Plant Sci.* **24**:2358.

Li. J., A. Mahajan and M.D. Tsai, 2006. Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry*. **45**(51):15168-78.