International Journal of Leading Research Publication (IJLRP)

E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

Analysis of Plant Argonaute Protein Sequences Motifs

Protip Basu

Assistant Professor
Department of Botany
Siliguri College

Abstract:
Argonaute proteins are found to exist in all types of living organisms, found across all domains and kingdoms of life. Although they fall under the category of uncharacterized proteins, they ubiquitously act as slicers of various non-coding RNAs. This process leads to the silencing of mRNAs, which become function-less. This work deals with the detection of analysis of protein motifs found in the plant argonaute sequences of some plants. Using an identical work plan, that can be undertaken for other domains of life, may help in increasing our knowledge about the Argonaute proteins.

1. INTRODUCTION

Argonautes are proteins found in across all domains of the living world. These proteins are uncharacterized proteins. Argonautes have been found and characterized, in certain organisms, and where they are found to act as slicers of various non-coding RNAs. This process leads to the destroys these mRNAs, leading to 'silencing' viz. rendering them function-less. The Argonautes therefore are part of the RNA induced silencing Complex (RISC), which interferes with the functioning of various mRNAs. Argonautes of different groups may vary in number, but have the same role(s).
The work mentioned in this paper includes detection of various motifs present in these proteins, using the CD search server of the NCBI website.

2. MATERIALS AND METHODS

2.1 Data Mining and Data Curation.

The data-set comprised of Argonautes AGO 1 through 10 from *Arabidopsis thaliana*, the model plant. Protein sequences retrieved from the NCBI-GenPept Protein database, which comprises of sequences from sources including various records and translations from annotated coding regions in different other databases as well.

2.2 BLASTp Analysis

The target database Phytozome v11 and the eleven target species – *Arabidopsis thaliana*, *Brassica rapa*, *Manihot esculenta*, *Glycine max*, *Phaseolus vulgaris*, *Gossypium raimondii*, *Solanum tuberosum*, *Solanum lycopersicum*, *Oryza sativa*, *Sorghum bicolor*, *and Zea mays*, were exposed to subsequent BLASTp and protein Sequence(s) with the best hit were selected, the same as was used in the CPF pathway [2].

2.3 Prediction and Analysis of Domains

Prediction and analysis of motifs was done using the CD search server of the NCBI website. [1]

3. RESULTS AND DISCUSSION

Secondary Structural analysis of the Argonaute proteins: Analysis of Motifs:

Analysis of the Secondary structure of proteins revealed that the Pfam motifs Piwi, Paz, ArgoL1, ArgoL2 and ArgoN motifs are present in each and every plant argonaute protein secondary structure. The motif ArgoMid however was absent in only a few argonautes. Across the different Argonautes analysed, a total of 24 different protein motifs were found (Table 11).

The other peculiar secondary structural features of plant argonaute proteins are (Table 1 to 10):

Argonaute 1: Present in all proteins was the very specific Gly_rich Argo1 motif, being absent only in the Cassava Argonaute 1. DUF3577 motif was present in the *Arabidopsis* Argonaute 1 whereas DUF5026 was present in both Tomato and Potato Argonaute 1.

Argonaute 2: Argonaute 2 of Cassava, Rice and Tomato lacked the ArgoMid motif whereas in *Sorghum*, two motifs, Microtub_bd and DUF4222 are present indicating some diverse functions of the same biomolecule.

Argonaute 3: In Argonaute 3 of *Glycine*, Cassava, Rice, Potato and Tomato there was no ArgoMid motif whereas in *Sorghum*, two motifs, Microtub_bd and DUF4222 are present indicating some diverse functions in the secondary structure of Argonaute 3.

Argonaute 4: In case of the secondary structure of all Argonaute 4 proteins, ArgoMid motif is present. The THF_DHG_CYH motif was present in the secondary structures of *Glycine*, Cotton, Cassava, Rice, *Sorghum* and Maize of Argonaute 4 proteins. As an exceptional case, the Reo_sigmaC and Alpha-2MRAP_N motifs were also found in the *Glycine* Argonaute 4.

Argonaute 5: The ArgoMid motif is present in the secondary structures of all the Argonaute 5 proteins. A Pep_deformylase is present in the *Glycine* Argonaute 5 whereas an AvrD motif is present in the Cotton Argonaute 5. The DUF5026 motif is present in both Tomato and Potato Argonaute 5 protein secondary structure. In the secondary structures of the Argonaute 5 proteins of Maize, Potato and Tomato, an additional motif, the Gly-rich_Ago1 motif was found to be present.

Argonaute 6: In Argonaute 6 proteins of *Brassica*, Cassava, Rice, *Phaseolus*, *Sorghum* and Maize, the THF_DHG_CYH motif was found to be present. Additionally, the DUF1450 motif in case of *Phaseolus* and HHA motif in case of *Sorghum* were found to be present. In all Argonaute 6 proteins, the ArgoMid motif was found to be present.

Argonaute 7: Only the *Brassica* Argonaute 7 lacked the ArgoMid motif which was otherwise ubiquitous in the rest of the species studied. The TBCC_N motif was found in Cotton, Cassava and *Phaseolus* Argonaute 7 proteins. The Gly-rich_Ago1 motif in Maize, the SURF2 motif in Maize, the YmzC motif in Cotton and the DUF3734 motif in Cassava were found as singular occurrences.

Argonaute 8: In Argonaute 8 proteins of *Brassica*, *Glycine*, Cotton, Cassava, Rice, *Sorghum* and Maize, the THF_DHG_CYH motif was found to be present. As an exceptional case, the Reo_sigmaC and Alpha-2MRAP_N motifs were also found in the *Glycine* Argonaute 8. In all Argonaute 8 proteins, the ArgoMid motif was found to be present.

Argonaute 9: In all Argonaute 9 proteins, the ArgoMid motif was found to be present. Singular occurrences include the HHA motif in the Maize Argonaute 9 and the Reo_sigmaC motif in the *Glycine* Argonaute 9. Other than the Argonaute 9 proteins of *Phaseolus*, Tomato and Potato, in all the other species, the THF_DHG_CYH motif was found to be present.

Argonaute 10: E1_UFD motif was present in Argonaute 10 of *Arabidopsis* and *Brassica* and the DUF3577 motif was found in Tomato and Potato Argonaute 10. The Stm1_N motif was found in the *Phaseolus* Argonaute 10. The ArgoMid motif was present in all the Argonaute 10 proteins.

Tables

Analysis of Protein Motifs

Table 1: Protein Motifs found in Argonaute 1

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO01	8
2.	BRARA_AGO01	7
3.	GLYMA_AGO01	7
4.	GOSRA_AGO01	7
5.	MANES_AGO01	6
6.	ORYSA_AGO01	7
7.	PHAVU_AGO01	7
8.	SOLLY_AGO01	8
9.	SOLTU_AGO01	8
10.	SORBI_AGO01	7
11.	ZEAMA_AGO01	7

Table 2: Protein Motifs found in Argonaute 2

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO02	6
2.	BRARA_AGO02	6
3.	GLYMA_AGO02	5
4.	GOSRA_AGO02	6
5.	MANES_AGO02	5
6.	ORYSA_AGO02	5
7.	PHAVU_AGO02	6
8.	SOLLY_AGO02	5
9.	SOLTU_AGO02	6
10.	SORBI_AGO02	8
11.	ZEAMA_AGO02	6

Table 3: Protein Motifs found in Argonaute 3

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO03	6
2.	BRARA_AGO03	6
3.	GLYMA_AGO03	5
4.	GOSRA_AGO03	6
5.	MANES_AGO03	5
6.	ORYSA_AGO03	5
7.	PHAVU_AGO03	6
8.	SOLLY_AGO03	5
9.	SOLTU_AGO03	5
10.	SORBI_AGO03	8
11.	ZEAMA_AGO03	6

Table 4: Protein Motifs found in Argonaute 4

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO04	6
2.	BRARA_AGO04	6
3.	GLYMA_AGO04	9
4.	GOSRA_AGO04	7
5.	MANES_AGO04	7
6.	ORYSA_AGO04	7
7.	PHAVU_AGO04	6
8.	SOLLY_AGO04	6
9.	SOLTU_AGO04	6
10.	SORBI_AGO04	7
11.	ZEAMA_AGO04	7

Table 5: Protein Motifs found in Argonaute 5

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO05	6
2.	BRARA_AGO05	6
3.	GLYMA_AGO05	7
4.	GOSRA_AGO05	7
5.	MANES_AGO05	6
6.	ORYSA_AGO05	6
7.	PHAVU_AGO05	6
8.	SOLLY_AGO05	8
9.	SOLTU_AGO05	8
10.	SORBI_AGO05	6
11.	ZEAMA_AGO05	7

Table 6: Protein Motifs found in Argonaute 6

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO06	6
2.	BRARA_AGO06	7
3.	GLYMA_AGO06	6
4.	GOSRA_AGO06	6
5.	MANES_AGO06	7
6.	ORYSA_AGO06	7
7.	PHAVU_AGO06	8
8.	SOLLY_AGO06	6
9.	SOLTU_AGO06	6
10.	SORBI_AGO06	8
11.	ZEAMA_AGO06	7

Table 7: Protein Motifs found in Argonaute 7

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO07	6
2.	BRARA_AGO07	5
3.	GLYMA_AGO07	6
4.	GOSRA_AGO07	8
5.	MANES_AGO07	8
6.	ORYSA_AGO07	6
7.	PHAVU_AGO07	7
8.	SOLLY_AGO07	6
9.	SOLTU_AGO07	6
10.	SORBI_AGO07	6
11.	ZEAMA_AGO07	8

Table 8: Protein Motifs found in Argonaute 8

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO08	6
2.	BRARA_AGO08	7
3.	GLYMA_AGO08	9
4.	GOSRA_AGO08	7
5.	MANES_AGO08	7
6.	ORYSA_AGO08	7
7.	PHAVU_AGO08	6
8.	SOLLY_AGO08	6
9.	SOLTU_AGO08	6
10.	SORBI_AGO08	7
11.	ZEAMA_AGO08	7

Table 9: Protein Motifs found in Argonaute 9

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO09	7
2.	BRARA_AGO09	7
3.	GLYMA_AGO09	8
4.	GOSRA_AGO09	7
5.	MANES_AGO09	7
6.	ORYSA_AGO09	7
7.	PHAVU_AGO09	6
8.	SOLLY_AGO09	6
9.	SOLTU_AGO09	6
10.	SORBI_AGO09	7
11.	ZEAMA_AGO09	8

Table 10: Protein Motifs found in Argonaute 10

SL. NO.	ARGONAUTE NAME	MOTIFS IN PFAM
1.	ARATH_AGO10	7
2.	BRARA_AGO10	7
3.	GLYMA_AGO10	6
4.	GOSRA_AGO10	6
5.	MANES_AGO10	6
6.	ORYSA_AGO10	6
7.	PHAVU_AGO10	7
8.	SOLLY_AGO10	7
9.	SOLTU_AGO10	7
10.	SORBI_AGO10	6
11.	ZEAMA_AGO10	6

Table 11: List of Motifs found in Argonaute Proteins

SL NO	MOTIF ACRONYM	EXPANDED NAME OF PROTEIN	REMARKS
1.	Alpha2_MRAP	Alpha-2-macroglobulin RAP, N-terminal domain	The alpha-2-macroglobulin receptor-associated protein (RAP) is an intracellular glycoprotein that binds to the 2-macroglobulin receptor and other members of the low density lipoprotein receptor family. RAP is comprised of three domains.
2.	ArgoL1	Argonaute linker 1 domain	ArgoL1 is a region found in argonaute proteins. It is a linker region between the N-terminal and the PAZ domains.
3.	ArgoL2	Argonaute linker 2 domain	ArgoL2 is the second linker domain in eukaryotic argonaute proteins.
4.	ArgoN	N-terminal domain of argonaute	ArgoN is the N-terminal domain of argonaute proteins in eukaryotes.
5.	AvrD	Pseudomonas avirulence D protein (AvrD)	This family consists of several avirulence D (AvrD) proteins primarily found in *Pseudomonas syringae*.
6.	DUF1450	Protein of unknown function (DUF1450)	This family consists of several hypothetical bacterial proteins of around 80 residues in length.
7.	DUF3577	Protein of unknown function (DUF3577)	This family of proteins is functionally uncharacterised.
8.	DUF3734	Patatin phospholipase	This domain family is found in bacteria, and is approximately 110 amino acids in length. The proteins in this family are frequently

			annotated as patatin family phospholipases.
9.	DUF4222	Domain of unknown function (DUF4222)	This short protein is likely to be of phage origin. For example it is found in the B6DZ51 Enterobacteria phage YYZ-2008.
10.	DUF5026	Domain of unknown function (DUF5026)	This family consists of several uncharacterized proteins around 100 residues in length and is mainly found in various Clostridiales species. The function of this family is unknown.
11.	E1_UFD	Ubiquitin fold domain	The ubiquitin fold domain is found at the C-terminus of ubiquitin-activating E1 family enzymes.
12.	Gly-rich_Ago1	Glycine-rich region of argonaute	This domain is found in the N terminus of some argonaute proteins. Argonaute (AGO) proteins are involved in RNA-mediated post-transcriptional gene silencing.
13.	HHA	Haemolysin expression modulating protein	This family consists of haemolysin expression modulating protein (Hha) from Escherichia coli and its enterobacterial homologues, such as YmoA from Yersinia enterocolitica, and RmoA encoded on the R100 plasmid.
14.	Microtub_bd	Microtubule binding	This motor homology domain binds microtubules and lacks an ATP-binding site
15.	Mid	Mid domain of argonaute	The ArgoMid domain is found to be part of the Piwi-lobe of the argonaute proteins.
16.	PAZ	PAZ domain	This domain is named PAZ after the proteins Piwi, Argonaute and Zwille.
17.	Pep_deformylase	Polypeptide deformylase	Peptide deformylase (PDF) is an essential metalloenzyme required for the removal of the formyl group at the N terminus of nascent polypeptide chains in eubacteria.
18.	Piwi	Piwi domain	This domain is found in the protein Piwi and its relatives.
19.	Reo_sigmaC	Reovirus sigma C capsid protein	Protein sigmaC in its native state was shown to be a homotrimer.

20.	Stm1	Stm1	This domain is found at the N-terminal region of the Stm1 protein.
21.	SURF2	Surfeit locus protein 2 (SURF2)	Surfeit locus protein 2 is part of a group of at least six sequence unrelated genes (Surf-1 to Surf-6).
22.	TBCC_N	Tubulin-specific chaperone C N-terminal domain	This N-terminal domain of tubulin-specific chaperone C has a spectrin-like fold and binds to tubulin.
23.	THF_DHG_CYH	Tetrahydrofolate dehydrogenase/cyclohydrolase, catalytic domain	Enzymes that participate in the transfer of one-carbon units require the coenzyme tetrahydrofolate (THF). This entry represents the N-terminal catalytic domain of these enzymes.
24.	YmzC	YmzC-like protein	The YmzC-like protein family includes the *Bacillus. subtilis* YmzC protein O31797 which is functionally uncharacterised.

4. CONCLUSION

In our approach, Argonaute proteins were found out and analyzed. A bulk of sequence data was found in sequence databases that consists of predicted, putative, and redundant sequences. The CPF method used only proper queries.

The same queries were used to search for protein motifs. It can be thus concluded that similar studies, using a similar work plan, can be undertaken for other domains of life, as well, to increase our knowledge about the Argonautes and their various functions.

REFERENCES:

[1] Marchler-Bauer A & Bryant SH: CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 2004 Jul 1; 32 (Web Server issue): W327-31.

[2] Basu et al: “Exploring Computational Protein Fishing (CPF) to identify Argonaute Proteins from Sequenced Crop Genomes” International Letter of Natural Sciences 6 (2015) 27-36.