Prediction of Functional Features of Plant Argonaute Sequences

Protip Basu

Assistant Professor
Department of Botany
Siliguri College

Abstract:
Argonautes are proteins found in animals, protists, archaea, eubacteria, fungi and plants alike. These proteins are mostly categorized as uncharacterized proteins in most organisms. In those organisms, where Argonautes have been found and characterized, they act as slicers of various non-coding RNAs, in conjunction with the latter's target, the cognate mRNAs. This process leads to the destruction of these mRNAs, and thus they are 'silenced', rendering them function-less. The work mentioned in this paper includes detection of nuclear export signals (NESs), acetylation sites and above all, phosphorylation sites. 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.

1. INTRODUCTION

Argonautes are proteins found in animals, protists, archaea, eubacteria, fungi and plants alike. These proteins are mostly categorized as uncharacterized proteins in most organisms. In those organisms, where Argonautes have been found and characterized, they act as slicers of various non-coding RNAs, in conjunction with the latter's target, the cognate mRNAs. This process leads to the destruction of these mRNAs, and thus they are 'silenced', rendering them function-less.

Thus, Argonautes are part of the RNA induced silencing Complex or RIS Complex (RISC), which interferes with the functioning of various mRNAs. Intriguingly, argonautes of different taxonomic groups vary in number, but have the same function(s).

The work flow that had been undertaken has resulted in sequence identification and characterization of the same, pertaining to ten crop species of our country, along with the model plant as well.

The work mentioned in this paper includes detection of nuclear export signals (NESs), acetylation sites and above all, phosphorylation sites.

A nuclear export signal (NES) is a short target peptide containing 4 hydrophobic residues in a protein that targets it for export from the cell nucleus to the cytoplasm, and these can be detected in protein sequences using the NetNES server.

Acetylation is an important modification of proteins and occurs as a co-translational and post-translational modification of proteins and such sites of acetylation can be detected using the NetAcet server.

When an amino acid residue is phosphorylated by a protein kinase by the addition of a covalently bound phosphate group, the process being a reversible process, the reaction is called phosphorylation. Phosphorylation sites on fasta sequences of proteins can be detected using the NetPhos server.

2. MATERIALS AND METHODS

2.1 Data Mining and Curation.

The initial data-set comprised of Argonaute (AGO 1 – AGO 10) from *Arabidopsis thaliana*, the model plant. Protein sequences retrieved from the NCBI-GenPept Protein database, which is comprised of sequences from sources including various records and translations from annotated coding regions in different other non-protein databases.

2.2 BLASTp Analysis.

The target database Phytozome v12 and the eleven target species – *Arabidopsis thaliana* and ten crop plant species *Brassica rapa, Manihot esculenta, Glycine max, Phaseolus vulgaris, Gossypium raimondii, Solanum tuberosum, Solanum lycopersicum, Oryza sativa, Sorghum bicolor, and Zea mays*), all important Indian agricultural species were exposed to subsequent BLASTp and protein Sequence(s) with the best hit were selected (One protein sequence hit/Argonaute type/Species), the same as was used in the CPF pathway [4].

2.3 Prediction of Nuclear Export Signals (NES), Acetylation sites and Phosphorylation sites.

Nuclear Export Signals (NES), Acetylation sites and Phosphorylation sites were predicted with different web-servers [1, 2, 3].

A type of protein localization signal, NES or Nuclear Export Signals, are often involved in important processes such as signal transduction and cell cycle regulation as are the Phosphorylation and Acetylation sites.

3. RESULTS AND DISCUSSION

3.1. Prediction of NES in Plant Argonaute Proteins (Tables 1 to 10).

Argonaute 1: It was observed that wherever Isoleucine (I) and Leucine (L) are occurring together as NESs, they are separated by a single amino acid – in *Arabidopsis*, *Glycine* and *Phaseolus*. In other cases viz. Maize, Cotton, Rice, Potato and Tomato, single NES is Leucine (L). The *Brassica* Argonaute protein had no NES. *Sorghum* had three NESs, one Isoleucine and two Leucines whereas Cassava had five NESs, one Isoleucine and four Leucines.

Argonaute 2: Two Leucine (L) NESs were found only in *Phaseolus*. In other cases viz. *Brassica*, Cotton, Rice and Tomato, single NES is mostly Leucine (L). Maize, *Glycine*, *Sorghum*, Potato and Cassava Argonaute proteins had no NES. *Arabidopsis* also had a single NES, but it is Isoleucine (I).

Argonaute 3: Two Leucine (L) NESs were found only in *Phaseolus*. In other cases viz. *Brassica*, Cotton, Rice and Tomato, single NES is also Leucine (L). *Arabidopsis*, Maize, *Glycine*, *Sorghum*, Potato and Cassava Argonaute proteins had no NES.

Argonaute 4: Two NESs were found in Cotton, one Isoleucine (I) and one Leucine (L) separated by one amino acid and three NESs were found in Cotton, one Isoleucine (I) and two Leucines (L), where one Leucine (L) was located one amino acid upstream of the Isoleucine (I). There were no NESs in the other nine species.

Argonaute 5: Two NESs were found only in *Arabidopsis*, one Lysine (K) and one Leucine (L) and in *Phaseolus*, both Leucines (L). In other cases viz. Maize, Cassava, Rice, Potato and Tomato, single NES is Leucine (L) and in Cotton, single NES is Isoleucine (I). *Brassica*, *Glycine* and *Sorghum* Argonaute proteins have no NES.

Argonaute 6: All NESs in Argonaute 6 proteins were Leucine (L). Two NESs were found only in Cotton and One NES was found in *Arabidopsis*, Potato and *Sorghum*. *Brassica*, Maize, *Glycine*, Rice, *Phaseolus*, Tomato and Cassava Argonaute proteins have no NES.

Argonaute 7: All NESs in Argonaute 6 proteins were Leucine (L). Two NESs were found only in *Brassica* and One NES was found in Maize, *Phaseolus* and Tomato. *Arabidopsis*, Cotton, *Glycine*, Rice, Potato, *Sorghum* and Cassava Argonaute proteins have no NES.

Argonaute 8: Two NESs were found only in Cotton, one Isoleucine (I) and one Leucine (L) and single NES, Leucine (L) in *Brassica*. *Arabidopsis*, Maize, *Glycine*, Rice, Potato, *Phaseolus*, Tomato, Cassava and *Sorghum* Argonaute proteins have no NES.

Argonaute 9: Two NESs were found only in Cotton, one Isoleucine (I) and one Leucine (L) and Maize, both Leucines (L), in both cases the NESs were separated by a single amino acid. In other cases viz. *Brassica* and Potato, single NES is Leucine (L) and in *Arabidopsis*, single NES is Isoleucine (I). Rice, *Glycine*, *Phaseolus*, Tomato, Cassava and *Sorghum* Argonaute proteins have no NES.

Argonaute 10: The highest number of NESs was found in case of Argonaute 10. This is significant because the function of Argonaute 10 is functionally different from the other argonautes, in that it is a microRNA locker. Cassava had three NESs with two Leucines (L) and one Isoleucine (I). Four NESs, three Leucines (L) and one Isoleucine (I), were found in argonautes of the species *Arabidopsis*, *Brassica*, Cotton, Potato and *Phaseolus*. Five NESs were found in Maize, *Glycine*, Rice, *Sorghum* and Tomato, of which all except *Glycine* [three Leucines (L) & one Isoleucine (I) & Alanine (A) each] had four Leucines (L). Tomato's fifth NES being Isoleucine (I) and Glutamic acid (E) being the fifth NES in all the three monocot species.

3.2. Prediction of Acetylation Sites in Plant Argonaute Proteins.

Argonaute 1: Only the Cotton Argonaute 1 has a predicted acetylation site in the 2nd amino acid i.e. Serine. No acetylation sites were found in the argonaute 1 proteins of the other species.
Argonaute 2: None predicted.
Argonaute 3: None predicted.
Argonaute 4: *Arabidopsis*, *Glycine*, Cotton and Cassava Argonaute 4 proteins had acetylation sites in the 3rd position, which was a Serine in each case.
Argonaute 5: *Brassica* Argonaute 5 proteins has a single acetylation site in the 3rd position and similar sites were present at the 2nd position in *Glycine*, Cotton, *Phaseolus* and Cassava.
Argonaute 6: None predicted.
Argonaute 7: None predicted.
Argonaute 8: *Brassica*, *Glycine*, Cotton, *Phaseolus* and Cassava Argonaute 8 proteins have a single acetylation site at the 3rd Serine residue.
Argonaute 9: *Arabidopsis*, *Glycine*, Cotton, *Brassica* and Cassava Argonaute 9 proteins have a single acetylation site at the 3rdSerine residue.
Argonaute 10: None predicted.
Thus, we can state that acetylation sites are not found in the plant Argonaute proteins, Argonaute 2, Argonaute 3, Argonaute 6, Argonaute 7 and Argonaute 10.

3.3. Prediction of Phosphorylation Sites in Plant Argonaute Proteins (Tables 11 to 20).

Argonaute 1: *Arabidopsis* and Maize had the highest number of phosphorylation sites – 65 each, with Cassava having least, 53. The number of serine residues was the highest ranging from 28 to 43 across the species whereas the lowest in number were tyrosine residues, which ranged between 6 in *Sorghum* upto13 in *Phaseolus*. Threonine residues were also found to be potential phosphorylation sites, ranging from 13 to 20 in number across the plant species.
Argonaute 2: *Phaseolus* had the highest number of phosphorylation sites at 67, and Tomato with the least, 43. The number of serine residues was the highest ranging from 25 to 47 across the species whereas the lowest in number were tyrosine residues, which ranged between 4 in the two cruciferous genera and going up to 12 in Cotton. Threonine residues were also found to be potential phosphorylation sites, ranging from 8 in Cassava to as high as 21 in Rice.
Argonaute 3: *Arabidopsis* had the highest number of phosphorylation sites, 74, with Tomato having least, 43. The number of serine residues was the highest ranging from 25 to 49 across the species whereas the lowest in number were tyrosine residues, which ranged between 4 in *Brassica* then going up to 12 in Cotton. Threonine residues ranged from 8 in Cassava to 21 in Rice.
Argonaute 4: Cassava has the highest number of phosphorylation sites, having 58 of them, with *Brassica* having least, 39. The number of serine residues was the highest ranging from 27 to 42 across the species whereas the lowest in number were tyrosine residues, which ranged between 2 in Rice to 8 in Cassava. Threonine residues were as low as 5 in Cotton and as high as 12 in Cassava.
Argonaute 5: 77 phosphorylation sites were present in Cassava and Tomato having least, 55. The number of serine residues was the highest ranging from 28 to 52 across the species whereas the lowest in number were tyrosine residues, which ranged between 6 in *Arabidopsis* and *Phaseolus* up to 10 in Cassava.

Threonine residues were also found to be possible phosphorylation sites, ranging from 13 to 23 in number across the species.

Argonaute 6: *Arabidopsis* and Maize had the highest number of phosphorylation sites – 55 each, with Potato having least, 42. The number of serine residues was the highest ranging from 23 to 39 across the species whereas the lowest in number were tyrosine residues, which ranged between 3 in *Sorghum* up to 8 in *Phaseolus*. Threonine residues were also found ranging from 9 to 21 in number across the species.

Argonaute 7: *Glycine* had the highest number of phosphorylation sites – 65, with *Brassica* having least, 45. The number of serine residues was the highest ranging from 28 to 34 across the species whereas the lowest in number were tyrosine residues, which ranged between 4 in *Brassica* up to 19 in *Glycine*. Threonine residues were also found to be possible phosphorylation sites, ranging from 6 in *Arabidopsis* to 18 in *Glycine*.

Argonaute 8: Cassava had the highest number of phosphorylation sites – 58, with *Phaseolus* having least, 34. The number of serine residues was the highest ranging from 22 to 39 across the species whereas the lowest in number were tyrosine residues, which ranged between 3 in *Phaseolus* and 8 in both *Arabidopsis* and Cassava. Threonine residues were least (5) in Cotton and ranged up to 12 in both *Brassica* and Cassava.

Argonaute 9: *Arabidopsis* and Maize had the lowest number of phosphorylation sites – 39 each, with Cassava having most, 58. The number of serine residues was the highest ranging from 24 to 38 across the species whereas the lowest in number were tyrosine residues, which ranged between 3 in Maize up to 8 in Cassava. Most Threonine residues were found in both *Brassica* and Cassava, 12 whereas the lowest were in Cotton amounting to 5.

Argonaute 10: *Arabidopsis* and Tomato had the highest number of phosphorylation sites – 55 each, with Maize having least, 45. The number of serine residues was the highest ranging from 22 to 34 across the species whereas the lowest in number were tyrosine residues, which ranged from 5 to 8. Threonine residues were also found to be possible phosphorylation sites, ranging from 12 to 16 in number across the species.

Tables

Prediction of NES.

Table 1 – Distribution of NES in across the different plant species in case of Argonaute 1.

Name of Species	Number of NESs	Positions
Arabidopsis	2	323 – I, 325 – L
Maize	1	708 – L
Glycine	2	680 – L, 682 – I
Gossypium	1	704 – L
Oryza	1	709 – L
Potato	1	680 – L
Phaseolus	2	689 – L, 691 – I
Sorghum	3	736 – L, 737 – L, 738 – I
Tomato	1	680 – L
Cassava	5	278 – L, 628 – L, 629 – L, 630 – L, 632 – I

Table 2 – Distribution of NES in across the different plant species in case of Argonaute 2.

Name of Species	**Number of NESs**	**Positions**
Arabidopsis	1	995 – I
Brassica	1	863 – L
Gossypium	1	815 – L
Oryza	1	720 – L
Phaseolus	2	772 – L, 810 – L
Tomato	1	881 – L

Table 3 – Distribution of NES in across the different plant species in case of Argonaute 3.

Name of Species	**Number of NESs**	**Positions**
Brassica	1	863 – L
Gossypium	1	815 – L
Oryza	1	720 – L
Phaseolus	2	772 – L, 810 – L
Tomato	1	881 – L

Table 4 – Distribution of NES in across the different plant species in case of Argonaute 4.

Name of Species	**Number of NESs**	**Positions**
Gossypium	2	214 – I, 216 – L
Oryza	3	217 – L, 612 – L, 614 – I

Table 5 – Distribution of NES in across the different plant species in case of Argonaute 5.

Name of Species	**Number of NESs**	**Positions**
Arabidopsis	2	631 – K, 638 – L
Maize	1	708 – L
Gossypium	1	624 – I
Oryza	1	696 – L
Potato	1	680 – L
Phaseolus	2	112 – L, 122 – L
Tomato	1	680 – L
Cassava	1	636 – L

Table 6 – Distribution of NES in across the different plant species in case of Argonaute 6.

Name of Species	Number of NESs	Positions
Arabidopsis	1	620 – L
Gossypium	2	19 – L, 631 – L
Potato	1	214 – L
Sorghum	1	810 – L

Table 7 - Distribution of NES in across the different plant species in case of Argonaute 7.

Name of Species	Number of NESs	Positions
Brassica	2	625 – L, 627 – L
Maize	1	728 – L
Phaseolus	1	279 – L
Tomato	1	830 – L

Table 8 - Distribution of NES in across the different plant species in case of Argonaute 8.

Name of Species	Number of NESs	Positions
Brassica	1	871 – L
Gossypium	2	214 – I, 216 - L

Table 9 - Distribution of NES in across the different plant species in case of Argonaute 9.

Name of Species	Number of NESs	Positions
Arabidopsis	1	201 – I
Brassica	1	871 – L
Maize	2	651 – L, 653 – L
Gossypium	2	214 – I, 216 – L
Potato	1	214 – L

Table 10 - Distribution of NES in across the different plant species in case of Argonaute 10.

Name of Species	Number of NESs	Positions
Arabidopsis	4	625 – L, 626 – L, 627 – L, 629 – I
Brassica	4	607 – L, 608 – L, 609 – L, 611 – I
Maize	5	606 – L, 607 – E, 608 – L, 609 – L, 610 – L

Glycine	5	608 – L, 609 – L, 610 – L, 611 – A, 612 – I
Gossypium	4	629 – L, 630 – L, 631 – L, 633 – I
Oryza	5	612 – L, 613 – E, 614 – L, 615 – L, 616 – L
Potato	4	252 – I, 597 – L, 598 – L, 599 – L
Phaseolus	4	608 – L, 609 – L, 610 – L, 612 – I
Sorghum	5	621 – L, 622 – E, 623 – L, 624 – L, 625 – L
Tomato	5	270 – L, 275 – I, 620 – L, 621 – L, 622 – L
Cassava	3	625 – L, 627 – L, 629 – I

Prediction of Phosphorylation sites.

Table 11 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 1.

Species	**S**	**T**	**Y**	**Total**
Arabidopsis	42	15	8	65
Brassica	39	14	9	62
Maize	43	13	9	65
Glycine	33	18	11	62
Gossypium	41	13	9	63
Oryza	37	13	8	58
Potato	32	20	8	60
Phaseolus	30	17	13	60
Sorghum	34	16	6	56
Tomato	28	19	8	55
Cassava	32	14	7	53

Table 12 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 2.

Species	**S**	**T**	**Y**	**Total**
Arabidopsis	33	20	4	57
Brassica	43	15	4	62
Maize	41	17	7	65
Glycine	44	10	8	62
Gossypium	36	10	12	58
Oryza	38	21	10	69
Potato	30	16	11	57
Phaseolus	42	15	10	67
Sorghum	41	14	8	63
Tomato	25	13	5	43
Cassava	47	8	10	65

Table 13 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 3.

Species	S	T	Y	Total
Arabidopsis	49	17	8	74
Brassica	43	15	4	62
Maize	41	17	7	65
Glycine	44	10	8	62
Gossypium	36	10	12	58
Oryza	38	21	10	69
Potato	28	15	5	48
Phaseolus	42	15	10	67
Sorghum	41	14	8	63
Tomato	25	13	5	43
Cassava	47	8	10	65

Table 14 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 4.

Species	S	T	Y	Total
Arabidopsis	27	11	5	43
Brassica	28	6	5	39
Maize	39	11	5	55
Glycine	28	11	5	44
Gossypium	31	5	6	42
Oryza	42	9	2	53
Potato	29	9	5	43
Phaseolus	30	10	6	46
Sorghum	35	11	5	51
Tomato	30	8	5	43
Cassava	38	12	8	58

Table 15 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 5.

Species	S	T	Y	Total
Arabidopsis	39	19	6	64
Brassica	44	14	9	67
Maize	43	13	9	65
Glycine	36	13	7	56
Gossypium	37	17	9	63
Oryza	34	18	9	61
Potato	32	20	8	60
Phaseolus	40	14	6	60
Sorghum	34	23	8	65
Tomato	28	19	8	55
Cassava	52	15	10	77

Table 16 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 6.

Species	S	T	Y	Total
Arabidopsis	28	21	6	55
Brassica	28	12	7	47
Maize	39	11	5	55

Glycine	35	12	5	52
Gossypium	23	14	7	44
Oryza	37	9	7	53
Potato	28	9	5	42
Phaseolus	27	12	8	47
Sorghum	35	15	3	53
Tomato	34	9	7	50
Cassava	35	12	5	52

Table 17 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 7.

Species	S	T	Y	Total
Arabidopsis	34	6	6	46
Brassica	30	11	4	45
Maize	32	12	6	50
Glycine	28	18	19	65
Gossypium	31	16	7	54
Oryza	29	11	8	48
Potato	30	16	11	57
Phaseolus	29	16	10	55
Sorghum	31	14	9	54
Tomato	26	16	12	54
Cassava	34	13	8	55

Table 18 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 8.

Species	S	T	Y	Total
Arabidopsis	22	9	8	39
Brassica	30	12	5	47
Maize	39	11	5	55
Glycine	28	11	5	44
Gossypium	31	5	6	42
Oryza	37	9	7	53
Potato	29	9	5	43
Phaseolus	22	9	3	34
Sorghum	35	11	5	51
Tomato	30	8	5	43
Cassava	38	12	8	58

Table 19 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 9.

Species	S	T	Y	Total
Arabidopsis	24	11	4	39
Brassica	30	12	5	47
Maize	28	8	3	39
Glycine	26	10	5	41
Gossypium	31	5	6	42
Oryza	37	9	7	53
Potato	28	9	5	42
Phaseolus	30	10	6	46

Sorghum	35	11	5	51
Tomato	28	10	4	42
Cassava	38	12	8	58

Table 10 - Distribution of Phosphorylation sites in across the different plant species in Argonaute 10.

Species	**S**	**T**	**Y**	**Total**
Arabidopsis	34	16	5	55
Brassica	33	14	5	52
Maize	23	14	8	45
Glycine	28	12	8	48
Gossypium	30	13	6	49
Oryza	23	16	7	46
Potato	32	13	7	52
Phaseolus	29	12	7	48
Sorghum	22	16	8	46
Tomato	34	14	7	55
Cassava	30	12	6	48

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, partial and redundant sequences. The CPF method used only proper queries.

The same queries were used to search for NESs, acetylation sites as well as phosphorylation sites. When the consequent results will be used for doing structural studies of Argonaute proteins, the said sites can be mutated and thus will lend a different dimension to the subsequent molecular dynamic simulation studies. 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] Blom N et al: Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of Molecular Biology 1999, 294(5): 1351-1362.

[2] Kiemer L et al: NetAcet: Prediction of N-terminal acetylation sites. Bioinformatics. 2005, 21(7):1269-70.

[3] La Cour T et al: Analysis and prediction of leucine-rich nuclear export signals, Protein Eng. Des. Sel. 2004, 17(6):527-36.

[4] 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.