International Letters of Natural Sciences
ISSN: 2300-9675, Vol. 33, pp 27-36
doi:10.18052/www.scipress.com/ILNS.33.27

Online: 2015-01-27

Exploring Computational Protein Fishing (CPF) to identify Argonaute Proteins from Sequenced Crop Genomes

Protip Basu[1]*, Sayak Ganguli[1], Sohini Gupta[2], Abhijit Datta[3]
[1] DBT-Centre for Bioinformatics, Presidency University, Kolkata, India
[2] PG Department of Botany, Barasat Government College, Barasat, India
[3]Department of Botany, Jhargram Raj College, Jhargram, India
*E-mail address: protipbasu@gmail.com

ABSTRACT

Plant RNA interference has been a very well studied phenomenon since its discovery. We are well versed with the types of small noncoding RNAs that are prevalent in the plant systems and their pathways of biogenesis and subsequent actions. However, apart from model plant systems such as *Arabidopsis* and *Oryza*, very little information is available regarding the other members of the RNA interference machinery; specially Argonaute proteins which acts as the major stabilizing factor for execution of the interference. This work focuses on the exploration of the sequenced crop genomes available on the web using a hybrid approach of computational protein fishing and genome mining. The results indicate that this hybrid approach was successful in the identification of argonaute proteins in the crop genomes under study.

Keywords: Computational Protein Fishing (CPF); Argonautes; RNA Interference; Crop Plant Genomes

1. INTRODUCTION

In the past decade or two, there has been a huge leap in the generation of sequence data because of the advent of advanced sequencing pipelines like Next-Generation Sequencing, deep-sequencing, RNA-Seq, etc. (Korpelainen et al, 2014). But, the growth of properly annotated sequence databases and availability of crystallographic or predicted structural data of the resultant proteins has not grown concurrently with the availability of completely sequenced genomes. Keeping up with these trends and also because of their ubiquitous presence across all the domains of life, we selected the Argonaute proteins as the target for our analysis (Mallory and Vaucheret, 2010).

Common wisdom suggests that genes that can replicate (make their own copies) themselves also form their complementary RNAs by the process of transcription, thus losing the introns (non-functional elements), leading to mRNA transcripts containing the coding sequence or cds (coding functional elements) bordered on both sides by the untranslated regions (UTRs, non-coding functional elements). These coding RNAs i.e. the cds are translated to form peptides culminating into generation of proteins. The non-coding RNAs

that are produced have various lengths, being segregated into long and short non-coding RNAs, the latter having plenty of regulatory roles. It is here that the Dicer proteins pop in resulting in mass-scale trimming ('dicing') and shortening of these 'precursor' non-coding RNAs (ncRNAs) into there shorter, 'mature' forms (Lee et al, 2004).

Hence, the fact that these RNAs do not code for any proteins, but are formed nevertheless vouches for their significance in the cellular physiology. From this point forward, the Argonaute proteins take over the operational control of the mature ncRNAs leading to self-regulatory measures of the cell, induced by these RNAs and causing necessary interference in metabolic processes (Baumberger and Baulcombe, 2005). These measures are aptly called RNA interference pathways or RNAi pathways (Ganguli and Datta, 2012b) and the assemblage of the Dicers and its associated proteins (varying in different organisms), the corresponding ncRNAs (being of various types, Bartel, 2004; Chen et al, 2010), alongwith the corresponding Argonaute (AGO) protein, is called the RNA induced silencing complex (RISC) (Ganguli and Datta, 2012a).

All sorts of RNA are very reactive and hence 'sticky' but a RISC is never complete without the target mRNA (which is to be 'silenced', thus causing the 'interference'), and a suitable ncRNA which is complementary to the target mRNA sequence, i.e. 'anti-sense' in nature (Meister and Tuschl, 2004; Baulcombe, 2004; Saleh et al, 2006). Hence, proper understanding of Argonautes is of paramount importance given their role in the RNAi machinery (Okamura et al, 2004).

Thus, to locate the Argonaute proteins, we went about our task of fishing out the proteins. Gene fishing in bioinformatics, in case of browsing and locating genes across genomes (Jakt & Nishikawa, 2008) and target fishing in cheminformatics, in case of trying to find out unknown biological targets for known chemical compounds (mechanism of action unknown, Jenkins et al, 2006) being used as effective drugs in certain diseases are approaches that have been used earlier. But, in our approach of Computational Protein Fishing or CPF, we have fished out Argonaute proteins along with their genomic and transcriptomic information.

In case of plant genomes, it has been observed that about four dozen species have been completely sequenced but what lies embedded within these sequenced genomes (Church and Gilbert, 1984) is still not elucidated. Being citizens of a country, which boasts itself to be an agricultural nation; we narrowed down our focus to ten crops which are grown in this vast geography also adding to the analysis the first plant genome (*Arabidopsis thaliana*) to be completely sequenced.

2. MATERIALS AND METHODS

2.1. Data Mining and Data Curation

The initial data-set was composed of *Arabidopsis thaliana* Argonaute (AGO 1 – AGO 10) Protein sequences downloaded from the GenPept database of NCBI and these served as the query sequences.

Table 1. Names and lengths of query sequences.

SL. NO.	NAME	DESCRIPTION	LENGTH
1.	AGO01	gi|15221177|ref|NP_175274.1| protein argonaute 1 [Arabidopsis thaliana]	1048
2.	AGO02	gi|145336300|ref|NP_174413.2| argonaute 2 [Arabidopsis thaliana]	1014
3.	AGO03	gi|15221662|ref|NP_174414.1| argonaute 3 [Arabidopsis thaliana]	1194
4.	AGO04	gi|18401305|ref|NP_565633.1| argonaute 4 [Arabidopsis thaliana]	924
5.	AGO05	gi|30683679|ref|NP_850110.1| argonaute 5 [Arabidopsis thaliana]	997
6.	AGO06	gi|42569579|ref|NP_180853.2| argonaute 6 [Arabidopsis thaliana]	878
7.	AGO07	gi|15222321|ref|NP_177103.1| protein argonaute 7 (protein ZIPPY) [Arabidopsis thaliana]	990
8.	AGO08	gi|42568003|ref|NP_197602.2| protein argonaute 8 [Arabidopsis thaliana]	850
9.	AGO09	gi|28396616|emb|CAD66636.1| ARGONAUTE9 protein [Arabidopsis thaliana]	896
10.	AGO10	gi|12643935|sp|Q9XGW1.1|AGO10_ARATH RecName: Full=Protein argonaute 10; AltName: Full=Protein PINHEAD; AltName: Full=Protein ZWILLE	988

2.2. BLASTp Analysis

Phytozome v10 was used as the target database and the eleven relevant species – one model organism (*Arabidopsis thaliana*) and ten crop plant species (both food and cash - *Brassica rapa, Manihot esculenta, Glycine max, Phaseolus vulgaris, Gossypium raimondii, Solanum tuberosum, Solanum lycopersicum, Oryza sativa, Sorghum bicolor,* and *Zea mays*), all relevant in the Indian agricultural perspective were the target species in the subsequent BLASTp that was performed using the above query sequences. Protein Sequence(s) with the best hit were selected (One protein sequence hit/Argonaute type/Species).

Table 2. List of Plant species and supplementary information.

Sl. no.	Organism	Common Name	Broad group	Trivial sub-group
1	*Arabidopsis thaliana*	Thale cress	Dicot	Crucifer
2	*Brassica rapa FPsc*	Turnip mustard	Dicot	Crucifer
3	*Glycine max*	Soybean	Dicot	Legume
4	*Gossypium raimondii*	Cotton	Dicot	-
5	*Manihot esculenta*	Cassava	Dicot	-
6	*Oryza sativa*	Rice	Monocot	-
7	*Phaseolus vulgaris*	Common bean	Dicot	Legume
8	*Solanum lycopersicum*	Tomato	Dicot	Solanaceous plant
9	*Solanum tuberosum*	Potato	Dicot	Solanaceous plant
10	*Sorghum bicolor*	Cereal grass	Monocot	-
11	*Zea mays*	Maize	Monocot	-

2.3. Characterization of functional and non-functional elements

The total number of introns was counted and the total length of genomic, transcript and coding sequences of the protein sequence hits as well as the peptide lengths were noted, so as to find quantitative variations between genomic elements and resultant protein lengths.

3. RESULTS AND DISCUSSION

3.1. The measure of introns

The number of introns present in the genes of Argonaute 1 (Range: 20 – 22 in number) in monocots was 22 and in case of crucifers as well as in legumes was found to be 21 and Argonaute 4 (Range: 20 – 23 in number) of monocots and crucifers was found to be constant at 22.
In crucifers and legumes, the introns count in case of Argonaute 6 (Range: 20 – 23 in number) was constant at 22 while in case of Argonaute 5 (Range: 19 – 22 in number), both crucifers had 19 introns and both legumes had 21 introns.

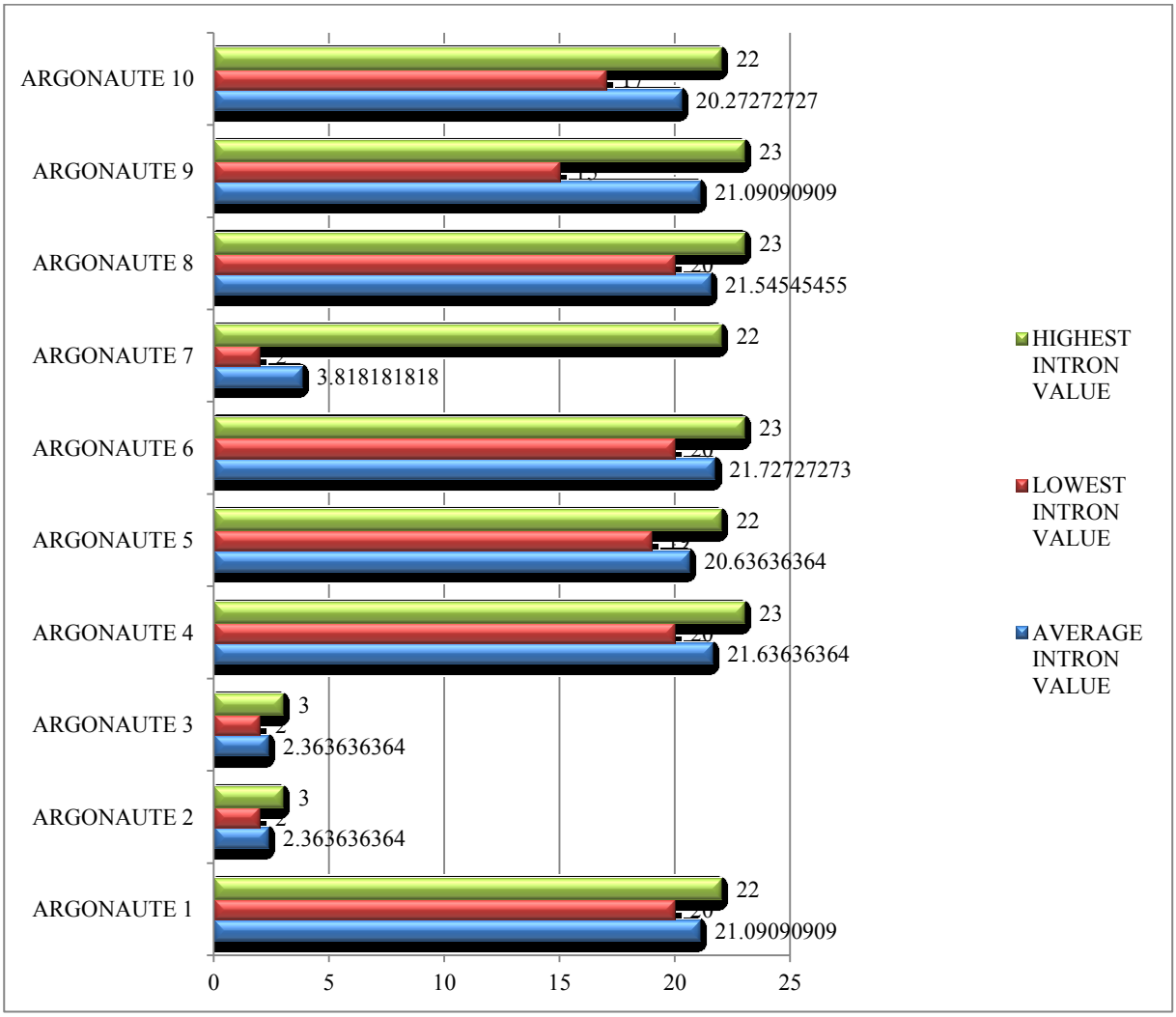

Figure 1. Argonaute-wise Comparison of Intron values showing Lowest, Highest and Average values.

In solanaceous plants the intron count was found to be same in case of Argonaute 4, Argonaute 5 and Argonaute 8 (Range: 20 – 23 in number) being 21, 20 and 21 respectively (corresponding peptide lengths being same too) and in case of Argonaute 6, Argonaute 7

(Range: 2 – 22 in number) and Argonaute 10 (Range: 17 – 22 in number) being 21, 2 and 20 respectively.
Argonaute 7 had the most rigid intron count with maize having 22 interruptions in the argonaute gene to the others' 2 and Argonaute 9 having a highly flexible intron count range of 15 – 23, 22 being the modal value.
All plants had the same number of interruptions in their Argonaute 2 genes as they had in their respective Argonaute 3 genes.

3.2. UTR lengths

One of the most consistent pattern observed was, that other than the mRNA transcript of the Argonaute 5 gene, all the other mRNA transcripts in case of cassava, lacked either one or both of the UTRs.

3.3. Correlating the peptide lengths

Plants can be classified into two broad categories – monocotyledons (monocots) and dicotyledons (dicots); thus the initial observations from the calculated data focused on identifying the differences in properties of all the argonaute protein sequences under study in the selected taxa at this level. It was observed that no specific global trends were identified; however, specific argonaute sequences displayed interesting characteristics as documented below.

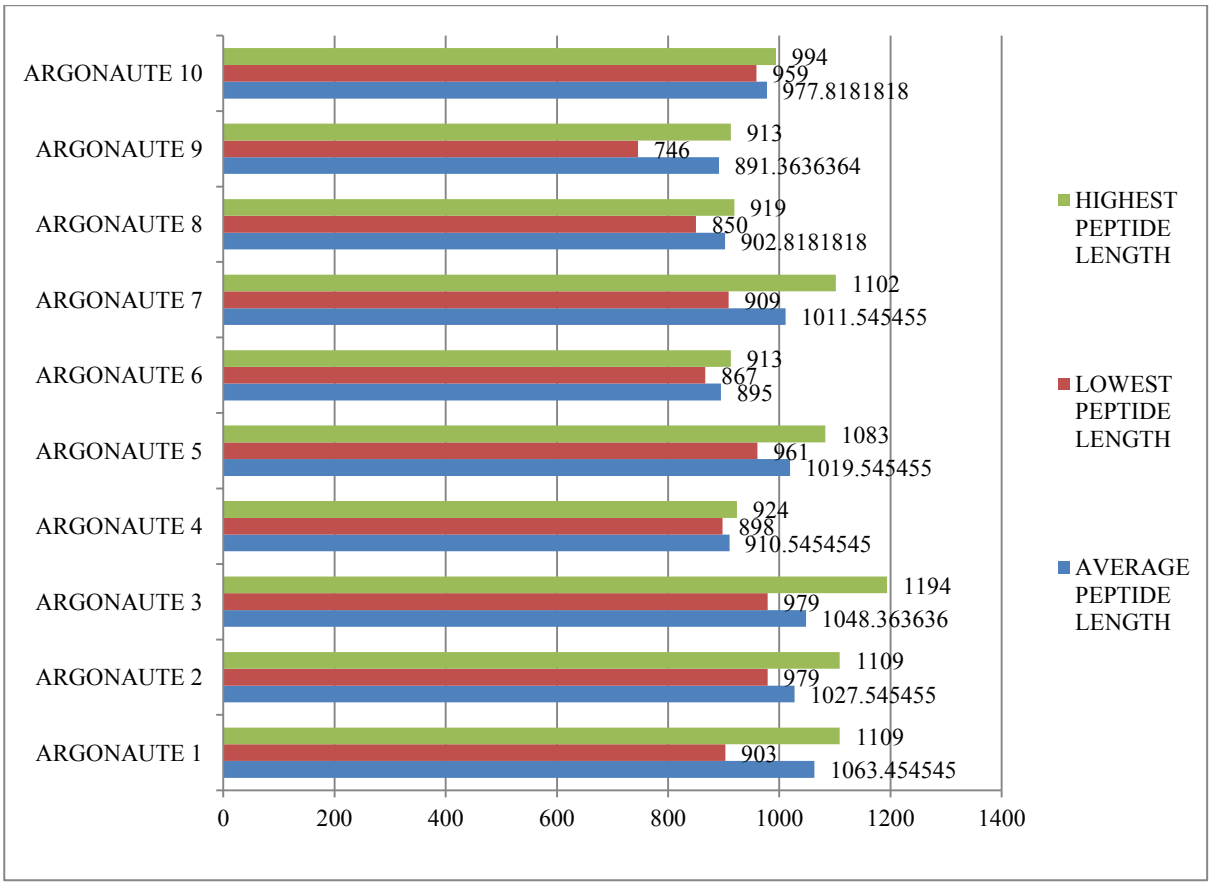

Figure 2. Argonaute-wise Comparison of Peptide lengths showing Lowest, Highest and Average lengths.

Argonaute 1: All the peptides varied in length, ranging from 903 – 1109, but the solanaceous plant argonaute proteins had the same length (1054), also indicating that monocots had the longer peptides.

Argonaute 2: The same trend of longer monocot argonautes continued but the shortest dicot argonaute was 979 amino acids long.

Argonaute 3: The longest Argonaute 3 (length 1194 amino acids) belonged to Thale cress, the model plant, but the range of lengths in the crop plants remained the same as Argonaute 2 (979 – 1109).

Argonaute 4: Cotton and the solanaceous plants had the same peptide length of 913 and the entire range here was smaller at 898 – 924.

Argonaute 5: Like argonautes 1 and 2, all monocots had longer peptides with the range being 961 – 1083, and again tomato and potato argonaute proteins had the same length (1054), being the only dicot argonautes having lengths more than a 1000 amino acids.

Argonaute 6: Argonaute 6 proteins of maize and cotton surprisingly had the same length of 898 amino acids and the overall range was 867 – 913.

Argonaute 7: The range of peptide length (909 – 1048) in case of all plants except maize may have some relation to the fact that all these plants had 2 introns in the corresponding gene, whereas maize argonaute 7 has a length of 1102 amino acids and its gene had 22 introns.

Argonaute 8: Cotton, potato and tomato had the same peptide length of 913 as had turnip mustard and soybean at 906. The range here was 850 – 919.

Argonaute 9: The peptide length of Argonaute 9 in case of maize 746 amino acids whilst the others had a range of length 896 – 913. The solanaceous plants and cassava have a 909 amino acid long Argonaute 9 while soybean and turnip mustard had a similar 906 amino acid long peptide.

Argonaute 10: The range of length of peptides was 959 – 994 whereas both the legumes had a similar length of 974 amino acids.

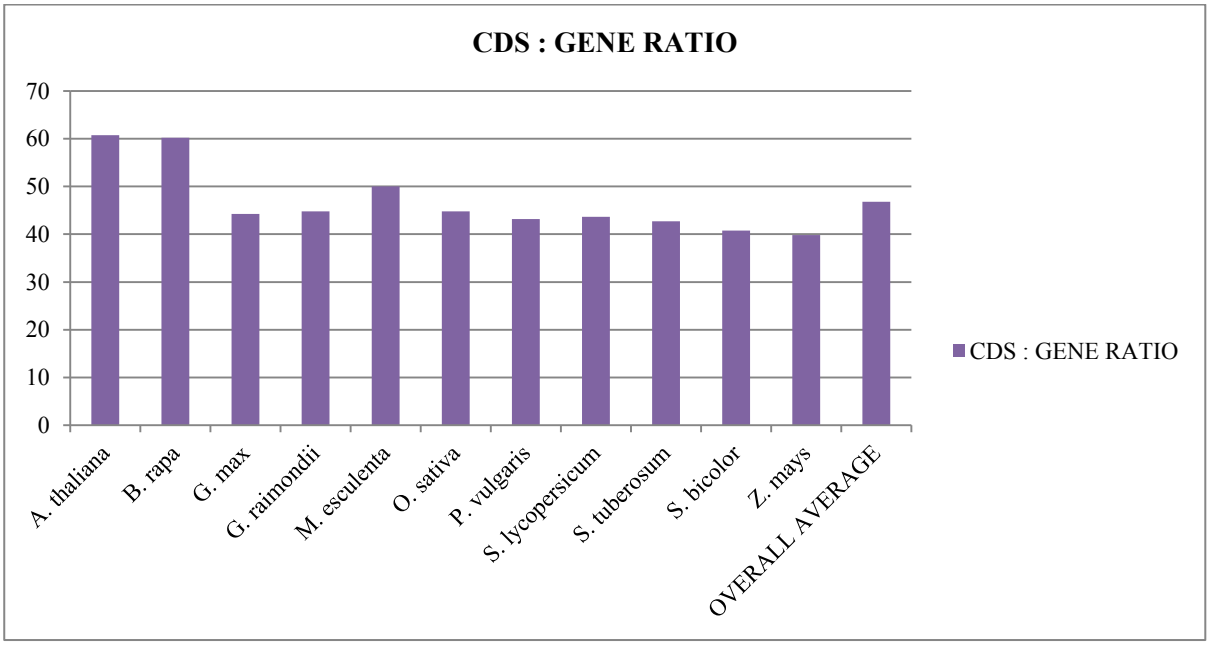

Figure 3. Species – wise comparison of CDS: Gene Ratio.

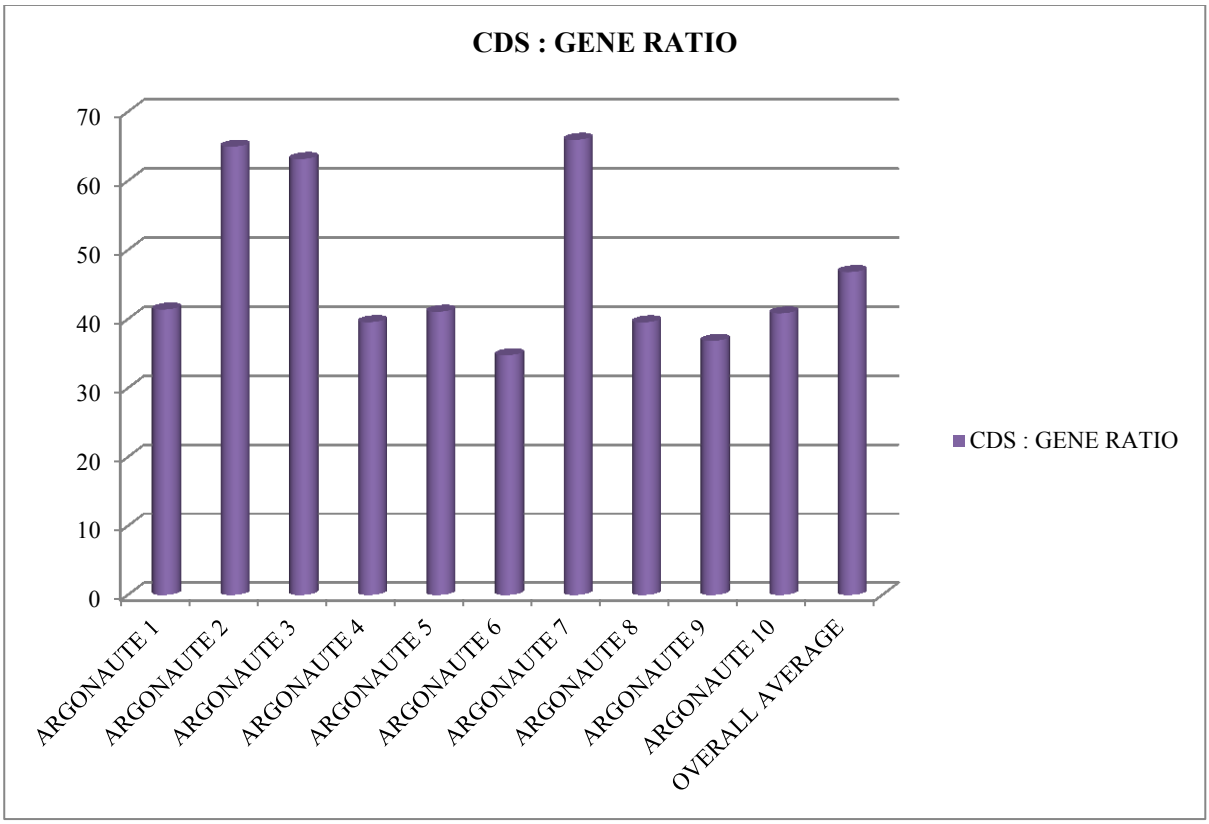

Figure 4. Argonaute – wise comparison of CDS: Gene Ratio.

From an evolutionary point of view the variations in the sequence length and intron number can be attributed to the phylogenetic similarities that the plant taxa under study possess. The results also show that a considerable amount of genomic length is expendable and consists of non-functional elements, which means a large fraction of Argonaute genes consist of non-coding portion as is evident from the Gene: CDS ratio.

Table 3. Argonaute-wise Average number of introns, average length of genomic elements & peptides and Average CDS: Gene Ratio.

NAME OF THE ARGONAUTE	INTRONS	GENE	TRANSCRIPT	5'UTR	3'UTR	CDS	PEPTIDE	CDS : GENE RATIO
ARGONAUTE 1	21.09090909	8012.727	4013.545455	671.1818182	258.545455	3193.36364	1063.454545	41.38540358
ARGONAUTE 2	2.363636364	4884	3371.727273	60	226.090909	3085.63636	1027.545455	64.96567661
ARGONAUTE 3	2.363636364	5140.545	3425.272727	56	221	3148.09091	1048.363636	63.17700533
ARGONAUTE 4	21.63636364	7065.545	3427.636364	640	290.272727	2734.63636	910.5454545	39.56045048
ARGONAUTE 5	20.63636364	8036.091	3858.818182	550.3636364	248.272727	3061.63636	1019.545455	41.07275896
ARGONAUTE 6	21.72727273	8055.727	3196.545455	454.8181818	278.818182	2688	895	34.78081159
ARGONAUTE 7	3.818181818	4972.273	3566.818182	317	212.181818	3037.63636	1011.545455	65.95373246
ARGONAUTE 8	21.54545455	7149.455	3376.363636	649.4545455	294	2711.45455	902.8181818	39.53737769
ARGONAUTE 9	21.09090909	7467.273	3474.545455	754.3636364	268.181818	2677.09091	891.3636364	36.85843402
ARGONAUTE 10	20.27272727	7645	3577.363636	290.0909091	350.818182	2936.45455	977.8181818	40.83656248
OVERALL AVERAGES	**15.65454546**	**6842.8636**	**3528.864**	**444.3272727**	**264.8181818**	**2927.4**	**974.8**	**46.81282132**

Table 4. Species-wise Average number of introns, average length of genomic elements & peptides and Average CDS: Gene Ratio.

NAME OF THE PLANT	INTRONS	GENE	TRANSCRIPT	5'UTR	3'UTR	CDS	PEPTIDE	CDS : GENE RATIO
A. thaliana	15.1	5079.2	3185.3	91.9	156.7	2936.7	977.9	60.75918171
B. rapa	15.1	5027.3	3274.3	133.4	233.9	2907	968	60.19197156
G. max	15.7	6915.7	3488	289.3	356	2842.5	946.5	44.22374337
G. raimondii	16.2	7201.7	3482.2	212.2	339.1	2930.7	975.9	44.76315169
M. esculenta	14.8	6121.8	3071.7	47.1	136.8	2887.8	961.6	50.03604363
O. sativa	16.1	7252.9	3427.8	137.7	241.5	3003.6	1000.2	44.76752408
P. vulgaris	15.7	7074.1	3271.9	187.5	190.6	2893.8	963.6	43.19317657
S.lycopersicum	15.2	7245.1	3613	43.7	276.9	2946.6	981.2	43.63635306
S. tuberosum	15	7584.7	3361.7	1606.3	279.7	2932.8	976.6	42.73113383
S. bicolor	15.9	8090.5	4168.5	788.2	504.6	2994.3	997.1	40.79907711
Z. mays	17.4	7678.5	4473.1	1350.3	197.2	2925.6	974.2	39.83967792
OVERALL AVERAGES	**15.65454545**	**6842.863636**	**3528.86**	**444.3272727**	**264.8181818**	**2927.4**	**974.8**	**46.81282132**

Table 5. Summary of Averages.

PARAMETER	OVERALL AVERAGE VALUE
AVERAGE INTRON VALUE	**15.65454545**
AVERAGE GENE LENGTH (NUCLEOTIDE)	**6842.863636**
AVERAGE TRANSCRIPT LENGTH (NUCLEOTIDE)	**3528.86**
AVERAGE 5'UTR LENGTH (NUCLEOTIDE)	**444.3272727**
AVERAGE 3'UTR LENGTH (NUCLEOTIDE)	**264.8181818**
AVERAGE CDS LENGTH (NUCLEOTIDE)	**2927.4**
AVERAGE PEPTIDE LENGTH (AMINO ACID)	**974.8**
AVERAGE CDS : GENE RATIO	**46.81282132**

4. CONCLUSIONS

During our query selection and optimization phase, we found that there is a body of sequence data in sequence databases that consists of predicted, putative, partial and above all redundant sequences. The CPF method hence required a proper group of sequences as queries and the results obtained thus provided us with some hitherto unknown information.

There is an inherent need for availability of properly annotated sequence data which can be correlated to the data related to the biomolecules which are the phenotypic expressions of their corresponding genes. The latter consists of structural models and their source protein sequences. The proteins that have been discovered using the CPF method shall serve as properly characterized and reliable target sequences to be used for predicting structural models of the same.

The role of Argonautes in Stress-related pathways (Jeong et al, 2010), Developmental Pathways (Borges et al, 2011), DNA methylation pathways (Havecker et al, 2010) and anti-viral pathways of plants underscores their importance in plant immunity as well as marks

them out as potential targets of viral silencing suppressor proteins (Voinnet, 2005, Gupta et al, 2014).

The method can also be applied to other groups of uncharacterized and less understood proteins.

Acknowledgement

The authors acknowledge the Department of Biotechnology, Government of India for the infrastructure facility at the DBT-Centre for Bioinformatics, Presidency University, Kolkata, which was utilized in execution of the above work.

References

[1] Korpelainen E, Tuimala J, Somervuo P, Huss M, Wong G, 2014.

[2] Jenkins JL, Bender A, and Davies JW, , 3(4) 2006 : 413 -421.

[3] Jakt LM & Nishikawa S, 2008, Cancer Sci; 99(5) 2008 : 829-835.

[4] Baumberger, N. and Baulcombe, D.C. Proc. Natl. Acad. Sci. 102 (2005): 11928-11933.

[5] Lee, Y.S., Nakahara, K., Pham, J.W., Kim, K., He, Z., Sontheimer, E.J., and Carthew, R.W.. Cell 117 (2004) : 69-81.

[6] Okamura, K., Ishizuka, A., Siomi, H., and Siomi, M.C. Genes & Dev. 18 (2004): 1655-1666.

[7] Voinnet, O. Nat. Rev. Genet. 6 (2005): 206-220.

[8] Meister, G. and Tuschl, T. Nature 431 (2004): 343-349.

[9] Baulcombe, D. Nature 431 (2004) : 356-363.

[10] Saleh, M.C., van Rij, R.P., Hekele, A., Gillis, A., Foley, E., O'Farrell, P.H., and Andino, R. Nat. Cell Biol. 8 (2006): 793-802.

[11] Havecker ER, Wallbridge LM, Hardcastle TJ, Bush MS, Kelly KA, Dunn RM, Schwach, F., Doonan, J.H., and Baulcombe, D.C, Plant Cell 22 (2010), 321-334.

[12] Mallory, A, and Vaucheret, H, Plant Cell, 22, (2010) 3879-3889.

[13] Bartel, DP, Cell 116 (2004): 281-297.

[14] Borges F, Pereira PA, Slotkin RK, Martienssen RA, Becker JD, J Exp Bot 62 (2011) : 1611-1620.

[15] Chen HM, Chen LT, Patel K, Li YH, Baulcombe DC, Wu SH, Proc Natl Acad Sci USA 107 (2010): 15269-15274.

[16] Church GM, Gilbert W, Genomic sequencing. Proc Natl Acad Sci, USA 81(1984): 1991-1995.

[17] Jeong DH, German MA, Rymarquis LA, Thatcher SR, Green PJ, Methods Mol Biol 592 (2010): 203-230.

[18] Ganguli S and Datta A,"Advances in Life Sciences: Principles and Applications" Eds: Tayung K, Barik BP and Mohapatra UB. 2012a, 1 - 12.

[19] Ganguli S and Datta, 2012b, "RNAi Technology" Ed: Gupta et.al. 347 - 356 (Chapter 19).

[20] Gupta S, Ganguli S and Datta A, 2014,“Agricultural Bioinformatics” Eds: Kavi Kishore et.al, 21 - 32 (Chapter 4).

(Received 26 December 2014; accepted 15 January 2015)