Original Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

COMPUTATIONAL PREDICTION OF MI-RNAS RESPONSIBLE FOR GENE SILENCING IN *HELIANTHUS TUBEROSUS* BY USING CROSS PHYLOGENY APPROACH WITH *ARABIDOPSIS THALIANA*

SUMIT GOVIL[1], SHAILESH KUMAR[2*]

[1]School of Life Sciences,Jaipur National University, Jaipur, Rajasthan, India
[2]Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur, Rajasthan India

ABSTRACT

MicroRNAs (mi-RNAs) are conserved biomolecules that regulate gene expression at level of translation. miRNA play a critical role at various levels in an organism. This is to deduce mi-RNAs by computational methods where they are compared for homology in EST sequences of an organism. The *Helianthus tuberosus* called as father of plant migration, having number of silent genes during its acclimatization to new world. A genomic approach is used to classify and target various identifiers at the nucleic acid level which govern the entire acclimatization of this plant. Raw Express Sequence Tags (ESTs) of *Helianthus tuberosus* were downloaded from dbEST Database, after refinement, clustering, and assembly by TGICL program, contigs and singletons were generated which are stored in a local database. Mature miRNA and Stem Loop of *Arabidopsis thaliana* were compared to generated local database by BLAST, were 50 mi-RNAs precursors were identified which aligned to multiple contigs where mature mi-RNA also showed their presence. The region of contigs alignment with the mi-RNA precursors were given in Mirfinder program for calculation of free energy values. Total 10 contig region identified where presence of mi-RNA precursor was feasible taking into consideration both the BLAST e-values and Mirfinder energy score.

KEYWORDS: ***Contigs, Singletons, Clustering, Alignment, Assembly***

SHAILESH KUMAR*

Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur, Rajasthan, India

Received on: 06-06-2017
Revised and Accepted on: 01-09-2017
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.4.b166-172

INTRODUCTION

Helianthus tuberosus L. member of compositae family is popularly accepted as father of plant migration. It is phenotypically rich variety of *Helianthus annus* L. It is expert in acclimatization due to enrichment of highly expressed heat shock protein (HSP) genes and alternative splicing is also reported up to 70 percent. Advanced approaches suggest the relation of mi-RNAs with various important cell phenomena (source: http://www.funpecrp.com). Because these are related with regulation of various genes, so proportionally affects the various pathways[1-2]. This information will be helpful to develop the Genetic and Protein networks[3], which may lead to understand at phenotypic level changes and complete system biology[4-5]. It has been postulated that in near future they will be used extensively for therapeutic as well as control measures. Micro RNAs are endogenous non-protein coding RNAs which have a high complementarity to 3`UTR [6] regions of miRNAs with the aid of numerous co-factors. This results in regulation of gene expression at post transcriptional level which makes them so valuable as a product. Though many experimental and computational processes have been developed to identify miRNAs none of them can claim a good accuracy. It has helped in gathering immense knowledge in genetics, transcriptomics [7] and physiology of these species. The mi-RNAs of these species are now publicly available through the MIRBASE database [8] of mi-RNA precursors and mature miRNAs (http://www.mirbase.org/). So, the complete MIRBASE data was downloaded and from it experimentally and computationally validated precursors and mature mi-RNA of *A. thaliana* were selected to predict miRNAs of *H. tuberosus*. The miRNAs of *A. thaliana* were compared [9] with the assembled ESTs of *H. tuberosus* to identify regions in the *Helianthus* genome where experimentally validated mi-RNA of *Arabidopsis* shows its presence. The *Helianthus* EST entries were preprocessed to remove repeats and vector sequences and sent for clustering and assembly. Both Mature and precursor miRNA matches [10] were checked out in the resultant clustered contigs and singletons. The results were compared to the other miRNA through BLAST executable [11] to generate a consensus for the predictions[12].

MATERIAL & METHODS

Dataset and Program Used in the study

Here we have downloaded the EST sequences [13] of *Helianthus tuberosus* present at dbEST on NCBI database (http:// www.ncbi.nlm.nih.gov/) which includes 40,373 sequences. Sequences obtained were processed through SEQCLEAN tool for validation and trimming of DNA sequences from a flat file database (FASTA format). SEQCLEAN was designed primarily for "cleaning" of EST databases, when specific vector and splice site data are not available, or when screening for various contaminating sequences is desired. The resultant sequences cleaned sequences were submitted to RepeatMasker a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked. The options -norna, -species Helianthus, -gc 39 were used specifically so that small RNA sequences are not screened, only repeats of *Helianthus* species are checked against and the total GC content of all the ESTs are taken into consideration. The output sequences from RepeatMasker[14] were subjected to clustering and assembly using the TGICL pipeline[15] package for efficient clustering and assembly of EST sequences. The options -p 98 -l 40 -v 20 were used for stringent clustering and assembly. This package automates clustering and assembly of a large EST/mRNA dataset. The clustering is performed by a slightly modified version of NCBI's megablast [16], and the resulting clusters are then assembled using CAP3 assembly program. TGICL starts with a large multiFASTA file (with an option of peer quality file) and gives outputs asassembly files produced by CAP3 program. The clustering and assembly phases of TGICL can execute parallelly by distributing the searches and assembly jobs into multiple CPUs, as TGICL program can take advantage of either SMP (Symmetric multiprocessing) machines or PVM (Parallel Virtual Machine) clusters. All the contigs and the singletons were put together in a single file where all the contigs start with the letter C and all singletons start with the letter S. These sequences were formatted into BLAST search able database files using the utility XDFORMAT from WU-BLAST v2.0[17]. The mi-RNA precursors and mature mi-RNA of A. thaliana were BLAST searched against this database. The mature mi-RNA was also BLAST searched against mi-RNA precursors where all precursors were formatted into database files by XDFORMAT.

Parameters and command used for comparison

MiRNA precursor BLAST search was performed with the WU-blast BLASTn program with the given command parameters.

blastndatabase_filequery_file W=7 qrecmax=93 -o output file

here in the command W represents word size 7 and maximum query at one go is 93, -o is representing output followed by file name. For mature mi-RNA BLAST search was performed with the wu-blast BLASTn program with the given command and parameters.

blastn database file query file W=7 E=1000 qrecmax=93 -o output file

here also the wordsize was kept at 7 for greater sensitivity in nucleotide searches and for mature miRNA e-value was kept at thousand to get optimum number of hits generated as the query sequence size is very small, qrecmax was used to process multiple queries at one go which included 93 mi-RNA precursors and 88 mature mi-RNA of *A. thaliana*. BLAST result of all sequences of the mi-RNA precursors were compared to that of the

mature mi-RNA BLAST result and all the contigs or singletons. Out of 93 miRNA precursors 27 had more than one contig or singleton common with the result of all mature miRNA and 25 had only one contig or singleton common. The best hit results for the precursors in BLAST search taking into consideration the short length of the precursors of *Arabidopsis* were chosen. The corresponding mature mi-RNA within these precursors were

blastn database file query file W=7 qrecmax=93 -o output file blastn database file query file W=7 E=1000 qrecmax=93 -o output file

identified and stored in tabulated form. The regions in the contigs or singletons matching with the corresponding mi-RNA precursor were given for RNA secondary structure prediction using MirFinder algorithm.

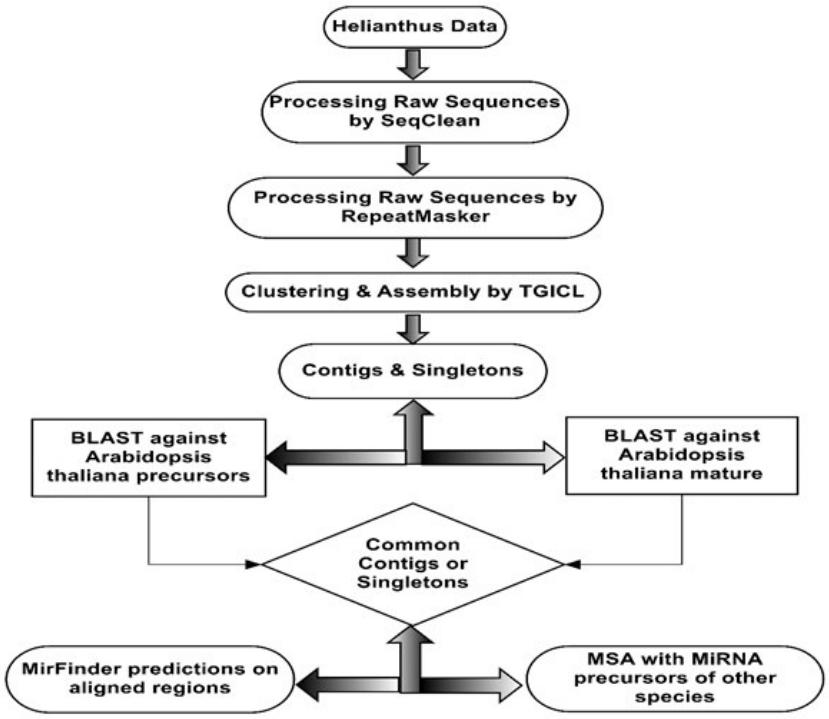

Figure1
Work Flow to predict MiRNA for Helianthus tuberosus.

RESULTS

After passing the 40,373 raw EST[18] sequences through SeqClean software the 461 trimmed sequences and 9 trashed sequences were obtained. In this way 40,364 ESTs are filtered. All the resulting data followed the principle of "clear range theory" of DUST algorithm [19] (Figure 2).

gi|125484477|gb|EL513260.1|EL513260	0.00	1	698	698
gi|125484476|gb|EL513259.1|EL513259	0.00	1	962	962
gi|125484475|gb|EL513258.1|EL513258	0.18	1	557	557
gi|125484474|gb|EL513257.1|EL513257	0.47	101	638	739
gi|125484473|gb|EL513256.1|EL513256	0.00	1	926	926
gi|125484472|gb|EL513255.1|EL513255	0.00	1	756	756
gi|125484471|gb|EL513254.1|EL513254	0.00	1	611	611
gi|125484470|gb|EL513253.1|EL513253	0.13	1	787	787
gi|125484469|gb|EL513252.1|EL513252	0.00	1	482	482
gi|125484468|gb|EL513251.1|EL513251	0.00	1	715	731
gi|125484467|gb|EL513250.1|EL513250	0.00	1	960	960
gi|125484466|gb|EL513249.1|EL513249	0.00	1	282	282
gi|125484465|gb|EL513248.1|EL513248	0.22	1	908	908
gi|125484464|gb|EL513247.1|EL513247	0.00	1	833	833

Figure 2
seqcl_sequences_fasta.cln file

The seqcl_sequences_fasta.cln file provides detailed sequence trimming, trashing (coordinates, reasons for trashing, contaminants names, etc.) as shown in Figure 2.

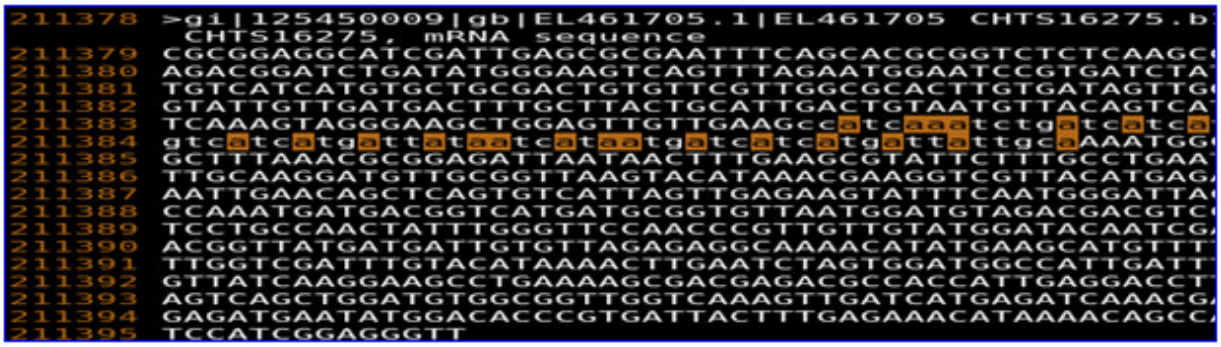

Figure 3
Masked file generated from Repeatmasker.

After seqclean we used repeatmasker program to remove IS elements, LINES/SINES and other repetitive elements. Resulting Numerical data contains only 4039 gi IDs which got masked out of 62598. That represents total 6.25% sequences have been masked, one example masked file is represented in Figure 3. The Interspersed Repeats or mutative elements are also checked against the available databases. The retrovirus elements arealso checked against the available databases.

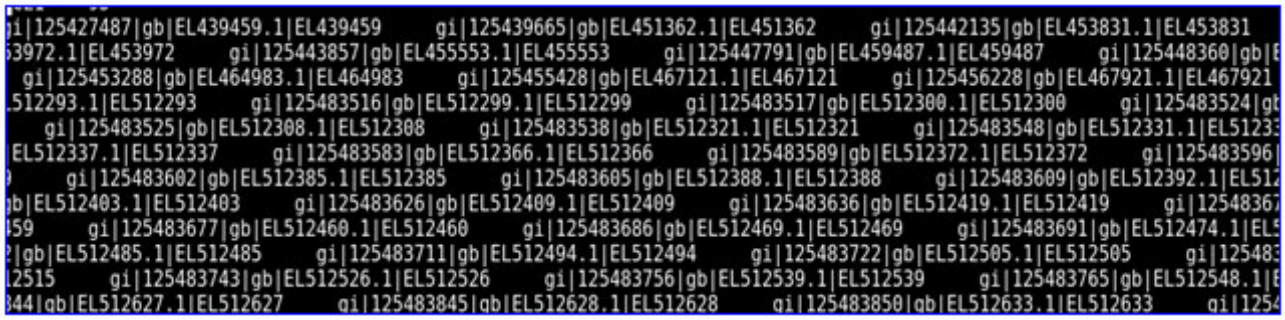

Figure 4
Entire_cluster file generated from TGICL.

The software tool used for EST clustering are the TIGR Assembler[15], Phrap and CAP3[20] (merged with TGICL), all originally designed for fragment assembly approach. In Figure 4 the Cluster 95 is showing all its ESTs in the form of their respective gi IDs

Figure 5
ASM file generated from TGICL

This Ace-file contains the cluster of 5 gi-IDs that are aligned to generate one contig. This file also contain one transcript file (cap_info file) which contains the information about overall overlaps and total number of chimeric overlaps removed. It also shows the information about the location of overlapping clips and their total length represented in Figure 5.

Figure 6

Blast(e) results of A. thaliana miRNA with locally generated contigs database of Helianthus tuberosus

These precursor and mature miRNAs were aligned through BLAST against Contigs in a local contig database. The figure 5 is showing cluster 2411, contig 1 which is of the length of 1218 nucleotide 94% identities is matched in this blast. We used Mirfinder program to calculate the free energy (dG) value. This dG value was considered as main parameter in Mirfinder. We found that all the dG values are under the range of -12.4 to -23.3. According to energy threshold of Mirfinder (dG< -25.0 kcal/mole) none of the structure could be skipped.

Table 1

Total number of predicted miRNA from Helianthus tuberosus by using miRFinder and the concerned clusters of Contigs with their dG values .

Concerned cluster and contig in ESTs	Predicted MicroRNA from *Helianthus* ESTs with Position	dGvalue
CL1628Contig1 760 - 3'PTR bp	>Position: 632-730 bp miRNA1: UGAUGUUCAAUUUUUCUAGGCUUUUAUUCCAAACAUUGGUCGUGG miRNA2:GAUACAAGGUGCAGCUACCCAUUGUUUGGGACAAAAGUUUGCUGAGAAGUUUGA	-22.1
CL4481Contig1 844 - 3'PTR bp	>Position:1528-1630bp miRNA3:UGCGUCGCAAAUGGCUGCUCUUGACUAUAUUGUAUCUGUU miRNA4:GAAAGUGAUGUAUUUGUACAUUCUUACCCUGGCAACAUGGCGCGAGCGGUUGAAGGCCAUCGU	-18.7
CL530Contig1 5'PTR - 546 bp	>Position:4167-4270bp miRNA5: CUCUGCUGAUCUGGGGGGAUUCCCUCCUUGUCCUGGAUUU miRNA6:UUGCCUUGACAUUAUCGAUAGUGUCUGAACUUUCCACCUCUAAGGUGAUGGUCUUUCCGGUCAA	-23.3
CL1474Contig1 5'PTR - 48 bp	>Position:2041-2150bp miRNA7:UUGCCAUUUUGCCAGCUUAUUUGAAUGCAUUAGUUGGGAAGGGAGUUCACGUUGUUACUGUAAAUGAUUAUUUGGCUCGG miRNA8:CGUGAUUGUGAAUGGGUUGGCCAAGUUCCA	-12.0
CL976Contig1 1045 - 3'PTR bp	>Position:711-818bp miRNA9: GUAGUCGGCGAGAGUGCGGCCGUCUUCCAGCUGUUUUCCG miRNA10:GCGAAGAUGAGACGUUGUUGGUCCGGUGGGAUUCCUUCUUUGUCUUGGAUCUUGGCUUUCACAUUGUC	-13.4
CL295Contig1 928 - 3'PTR bp	>Position:911-1018bp miRNA11: GUAGUCGGCGAGAGUGCGGCCGUCUUCCAGCUGUUUUCCG miRNA12:GCGAAGAUGAGACGUUGUUGGUCCGGUGGGAUUCCUUCUUUGUCUUGGAUCUUGGCUUUCACAUUGUC	-20.2
CL2896Contig1 518 - 3'PTR bp	>Position:956-1062bp miRNA13:UAAUUACCUUAGUGGUCCAGCAAUGUCCUUGACCAUUGUGGAUAGCAUCAACUA miRNA14:AGCCAGUGUAAUUCCAGUUCUUGAUCACAUUGUAUGGUCCAUCAUGGAUUUCA	-21.0

The overall seven clusters are found to contain common hits which are generated by local blast of precursor and mature miRNAs of A. thaliana in local contig database. Cluster 1474 and cluster 976 are less energetically favorable while the cluster 1628 and cluster 530 are more energetically favorable. Cluster 976 is least stable structure according to energy criteria.

Table 2

Predicted miRNA from Helianthus tuberosus by using MiRFinder and the concerned clusters of Contigs.

CLUSTER No	TRACED microRNAs	
>CL1628Contig1	miRNA1	UAUUCCAAACAUUGGUCGUGG
	miRNA2	CUACCCAUUGUUUGGGACAAA
>CL4481Contig1	miRNA3	GGCUGCUCUUGACUAUAUUGU
	miRNA4	GCAACAUGGCGCGAGCGGUUG
>CL530Contig1	miRNA5	CUGAUCUGGGGGGAUUCCCUC
	miRNA6	GUGAUGGUCUUUCCGGUCAA
>CL1474Contig1	miRNA7	GCCAUUUUGCCAGCUUAUUUG
	miRNA8	UGAAUGGGUUGGCCAAGUUCC
>CL976Contig1	miRNA9	GCGAGAGUGCGGCCGUCUUCC
	miRNA10	GGAUCUUGGCUUUCACAUUGUC
>CL295Contig1	miRNA11	GCGAGAGUGCGGCCGUCUUCC
	miRNA12	GGAUCUUGGCUUUCACAUUGUC
>CL2896Contig1	miRNA13	CUUGACCAUUGUGGAUAGCAU
	miRNA14	GUGUAAUUCCAGUUCUUGAUC

The stem loop structures of various precursor miRNAs generated within these clusters are analyzed by using Mirfinder.

Figure 7
miRNA from cluster 1628 (dG = -22.1)

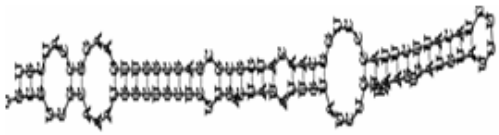

Figure 8
miRNA from cluster 4481 (dG = -18)

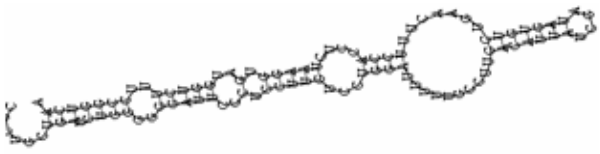

Figure 9
miRNA from cluster 530 (dG = -23.3)

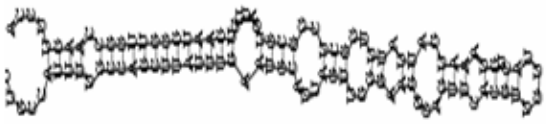

Figure 10
miRNA from cluster 1474 (dG = -2.4)

Figure 11
miRNA from cluster 976 (dG = -13.4)

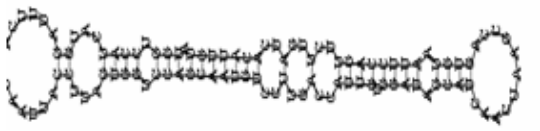

Figure 12
miRNA from cluster 295 (dG = -20.2)

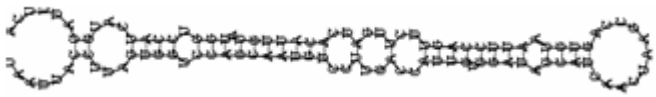

Figure 13
miRNA from cluster 2896 (dG = -21.0)

DISCUSSION

Forpredicting miRNA from EST data we have started with 40,373 EST sequences of *Helianthus tuberosus.* These raw sequences were processed through software for removal of error in the sequences, by taking reference of NCBI based UniVec (vector sequence) database. 9 sequences were trashed by the program and 461 sequences were trimmed by SEQCLEAN program. After cleaning all sequences by SEQCLEAN shown in Figure 2, repeated sequences are removed by comparing sequences against Repeat database libraries for version 3.1.2 by using Repeatmasker tool of alignment. Here, 765 simple repeats and 6.25% regions of low complexity were identified by the program which are masked and will not affect upcoming processes. The regions were masked in lowercase letters shown in Figure 3. The TIGR Assembler Phrap and CAP3 as used to assemble the sequence. This only puts together sequences linked to other sequences by overlaps of at least 98% identity, at least 40bp, and at most 20bp overlap distance of sequence. 6619 singletons along with 2292 cluster (Figure 4) where every cluster contains minimum 1 and maximum 8 contigs was used. This assembled data is further stored as local dataset of contigs. Now, this local database of EST based contigs are aligned with mature and precursor mRNA of *A. thaliana* a reference plant. Total 7 unique miRNAs were obtained that are available in the contigs of *H. tuberosus*. The miRNAs interacts primarily at 3' UTRs and 5' UTRs playing role in controlling gene expression. UTRs length plays key role in number of miRNAs binding of [21] . The structure of these was predicted by MIRFINDER program and they were found energetically very stable in Figure 7-13. These methods can be used for predicting miRNA for any species and gene using clustering and prediction methods. Glutathione peroxidase (GPx) activity plays a vital role in detoxification of human body[22] and their activity may also be influenced by miRNA binding. The prediction of putative miRNA targeting GPx family genes will be our focus in future studies and will be compared with existing miRNA in humans.

CONCLUSION

The entire methodology for predicting mi-RNAs from EST data had resulted to seven novel mi-RNAs in *Helianthus tuberosus*. Mir Finder based analysis as used to represents the energy scores of predicted miRNAs. These predicted mi-RNAs from EST data obtained in this study can be useful for study of UTR based target hybridization and detection of silent genes. These miRNAs are can be used for gene silencing and help in better understanding the system biology of

Helianthus tuberosus plant which are evolutionary very important.

ACKNOWLEDGEMENT

The authors are thankful to Jaipur National University, Jaipur, India and Amity University Rajasthan, Jaipur, India, for providing full support to carry out this research.

CONFLICTS OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Barozai MY, Baloch IA, Din M. Identification of MicroRNAs and their targets in Helianthus. Mol Biol Rep. 2012 Mar 1;39(3):2523-32.
2. Sarangzai AM. Profiling the carrot (Daucuscarota L.) microRNAs and their targets. Pak. J. Bot. 2013 Jan;45(S1):353-8.
3. Botstein D, Cherry JM, Ashburner M, Ball CA, Blake JA, Butler H, Davis AP, Dolinski K, Dwight SS, Eppig JT. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25-9.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA. Gene Ontology: tool for the unification of biology. Nat genet. 2000 May;25(1):25.
5. Hammond SM, Caudy AA, Hannon GJ. Post-transcriptional gene silencing by double-stranded RNA. Nat Rev. Genet. 2001 Feb 1;2(2):110.
6. Mazumder B, Seshadri V, Fox PL. Translational control by the 3′-UTR: the ends specify the means. Trends Biochem Sci. 2003 Feb 28;28(2):91-8.
7. Sugnet CW, Kent WJ, Ares M, Haussler D. Transcriptome and genome conservation of alternative splicing events in humans and mice. InPacific Symposium on Biocomputing 2004 Jan (Vol. 9, pp. 66-77).
8. Griffiths-Jones S, Grocock RJ, Van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006 Jan 1;34(suppl 1):D140-4.
9. Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 2002 Jul 1;14(7):1457-67
10. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ. Processing of primary microRNAs by the Microprocessor complex. Nat. 2004 Nov 11;432(7014):231-5.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.
12. Lu DP, Read RL, Humphreys DT, Battah FM, Martin DI, Rasko JE. PCR-based expression analysis and identification of microRNAs. J RNAi Gene Silencing. 2005 Jul 28;1(1):44-9.
13. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR. Complementary DNA sequencing: expressed sequence tags and human genome project. Sci. 1991 Jun 21;252(5013):1651-6.
14. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinf. 2009 Mar 14:4-10.
15. Sutton GG, White O, Adams MD, Kerlavage AR. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. Genome Sci Technol.1995;1(1):9-19.
16. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000 Feb 1;7(1-2):203-14.
17. Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W. WU-Blast2 server at the European bioinformatics institute. Nucleic Acids Res. 2003 Jul 1;31(13):3795-8.
18. Neto ED, Correa RG, Verjovski-Almeida S, Briones MR, Nagai MA, da Silva W, et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Natl Acad Sci. 2000 Mar 28;97(7):3491-6.
19. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol. 2006 Jun 1;13(5):1028-40.
20. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999 Sep 1;9(9):868-77.
21. Kumar S, Kachhwaha S, Kothari SL. Quantitative Estimation for Impact of Genomic Features Responsible for 5′ and 3′ UTR Formation in Human Genome. InProceedings of Fifth International Conference on Soft Computing for Problem Solving 2016 (pp. 299-309). Springer, Singapore. (10.1007/978-981-10-0448-3_24)
22. Kumar S, Bhadhadhara K, Govil S, Mathur N., Pathak A N, In-silico studies for comparative analysis of glutathione peroxidaes isozymes in homo sapiens. Int J Pharm Bio Sci 2014 Oct; 5(4): (B) 352-63

Reviewers of this article

Dr. Sanjay Mishra, Ph.D.

Professor, Department of Biotechnology, IFTM University, Delhi Road (NH- 24), Moradabad- 244102, U.P., India

Asst. Prof. Dr. Deepansh Sharma, M.Sc, M.Phil, Ph.D.

Assistant Professor, School of Biotechnology and Bioscience, Lovely Professional University, Phagwara, Punjab, India

Prof.Dr.K.Suriaprabha

Asst. Editor , International Journal of Pharma and Bio sciences.

Prof.P.Muthuprasanna

Managing Editor , International Journal of Pharma and Bio sciences.

We sincerely thank the above reviewers for peer reviewing the manuscript