Endogenous Rice (*Oryza Sativa*) miRNAs and their Potential Targets against Rice Tungro Virus using Various String Matching Algorithms

Ramamani Tripathy
Dept. of IT, ITER
Siksha O Anusandhan University,
Bhubaneswar, Odisha, INDIA
ramatripathy1978@gmail.com

Debahuti Mishra
Dept. of CA, ITER
Siksha O Anusandhan University,
Bhubaneswar, Odisha, INDIA
debahuti@iter.ac.in

Rudra Kalyan Nayak
Dept. of IT, ITER
Siksha O Anusandhan University,
Bhubaneswar, Odisha, INDIA
rudrakalyannayak@gmail.com

ABSTRACT

Rice is the major growing crop in South and Southeast Asia and half of the world population takes rice as staple food. But rice production hampered because of rice tungro disease caused by two viruses. MicroRNA(miRNA) act major role in plant resistance against viruses. miRNAs are short around 22 nucleotides RNA molecules found in eukaryotic cells that regulate gene expression by translational inhibition or cleavage of complementary mRNAs. The present work emphasizes on different string matching algorithms such as Boyer-Moore, Knuth-Morris, Rabin-Karp to elucidate the potentiality of rice miRNAs target against rice plant infecting tungro viruses of both Rice Tungro Spherical Virus (RTSV) and Rice Tungro Baciliform Virus (RTBV). 581 number of miRNA sequence from miRBase has been collected and target rice tungro viruses genes like coat proteins CP1,CP2,CP3, poly-protein of RTSV and ORF1, ORF2, ORF3, P12, P24, P46 and P194 of RTBV considered for simulation. The potential endogenous rice miRNAs targeted found by three different algorithms in our approach also compared with the web based sever psRNATarget. The novel target site findings of rice miRNAs and tungro virus will be helpful to manipulate in new biotechnological approaches for enhance rice production.

Categories and Subject Descriptors

F.2.2 [**Numerical Algorithms and Problems**]: Pattern matching

General Terms

Algorithm, Experimentation

Keywords

miRNA:mRNA; Seed Region; Tungro Virus; Boyer-Moore string matching algorithm; Knuth-Morris matching algorithm; Rabin-Karp matching algorithm.

CUBE 2012, September 3–5, 2012, Pune, Maharashtra, India.
Copyright 2012 ACM 978-1-4503-1185-4/12/09…$10.00.

1. INTRODUCTION

In South and Southeast Asia , Rice tungro disease constitute as a serious threat to increase the rice production [1] and the disease is caused by co-infection of two viruses such as: RTBV, a double-stranded DNA-containing virus, belonging to the genus Tungrovirus; and RTSV, a single-stranded RNA virus belonging to the genus Waikavirus [2]. The main characteristics of the disease symptoms consists of severe stunting and discoloration of infected plants, reduced tillering, and small and/or sterile panicles [3] and the symptoms start to appear 1-2 weeks after infection. Yield losses can be as much as 100% when plants are infected in the early seedling stage. Both viruses are transmitted by several green leafhoppers(GLH) (Nephotettix virescens) [4]. RTSV is mainly responsible for transmission by the leafhopper vector where RTBV is responsible for disease symptoms. Because of the constraint of present breeding programs and disease management, rice tungro disease is one of the most harmful diseases of rice production in Asia [5].

To resists viral infections, plants have its' own endogenous miRNAs. MicroRNAs (miRNAs) are approximate 22 nucleotide long non coding RNAs molecules that regulate gene expression at the post transcriptional level and by targeting mRNAs for cleavage in plants [6-7]. More than 581 miRNA sequences have been reported for Rice (Oryza Sativa) and a total of 18226 known miRNAs have been reported for various species in the Mirbase database. In higher eukaryotes about 30% of the total genes are expected to be controlled by miRNAs [8-9]. In addition to regulate the endogenous expression of some genes, miRNAs could have a direct role in viral defense. In this paper, different string matching algorithms such as Boyer-Moore, Knuth-Morris, Rabin-Karp are considered to elucidate the potentiality of rice miRNAs target against rice plant infecting tungro viruses of both Rice Tungro Spherical Virus (RTSV) and Rice Tungro Baciliform Virus (RTBV). 581 number of miRNA sequence from miRBase has been collected and target rice tungro viruses genes like coat proteins CP1,CP2,CP3, poly-protein , of RTSV and ORF1, ORF2, ORF3, P12, P24, P46 and P194 of RTBV considered for simulation. Our three different string matching algorithms compared with the web based sever psRNATarget to found out the potential endogenous rice miRNAs.

The rest of the paper has organized as follows: section 2 gives the idea about related work, section 3 represents preliminary concepts potential miRNA target prediction against complementary sites, finding cleavage region and various string matching algorithms; section 4 represents the proposed model; section 5 give a deeper look into the algorithms; section 6 deals with experimental evaluation and result analysis and section 7 gives the conclusion.

2. RELATED WORK

Perez-Quintero et al.[10] discovered that miRNAs are non-coding short RNAs around 22 nucleotides long that regulate gene expression by translational inhibition in animals or cleavage of complementary mRNAs in case of plants. Here the set of plant miRNA from six plants was targeted against plant viruses and miRanda is a web base tool, which was used for target prediction. Dai X et al. [11] observed that psRNATarget a web base server, which have two activity i.e. reverse complementary matching using a scoring schema and evaluation of unpaired energy. A set of parameters are taken for calculating the complementarity between miRNA and target mRNA. The results indicate that psRNATarget is well capable of performing high-throughput analysis for large-scale datasets. Zhang Y et al. [12] mentioned about the performance of prediction accuracy using several target prediction algorithms like miRanda, TargetScan RNAhybrid and of a selection of integration strategies on these algorithms using multiple data sets. miRNA is to measure the performance of miRNA target prediction algorithms using both the true-positive and false-positive rate and a Bayesian Network classifier is used for better target prediction. Tyagi H et al. [13] explored that tungro is one of the most damaging rice diseases in South East Asia and it is a co-infection involving Rice tungro bacilliform virus and Rice tungro spherical virus. RNA interference has an important role in defending cells against viruses and transposons. By down regulating the gene functions RNAi protect plants against viruses.

3. PRELIMINARIES

3.1 Potential miRNA Target Prediction Against Complementary Sites

Plants do not have an antibody-based immune system, so miRNA-mediated pathway was newly discovered approach to suppress plant viruses. miRNAs target the complementary sites of mRNAs in a specific sequence manner. miRNA act by binding to the complementary sites on the 3' untranslated region (UTR) of the target gene to induce cleavage with near perfect complementary to repress productive translation. RNA silencing is a conserved defense mechanism in plants [1-2][4-5].

3.2 Finding Cleavage Region

miRNAs are derived from double-stranded RNA (dsRNA) and are then processed into 21-22 nt single stranded molecules by Dicer or a Dicer-like enzyme; later, they are incorporated into the RNA-induced silencing complex (RISC) to guide the cleavage or translational repression of the complementary mRNA strand [14]. Potential miRNAs and their target mRNAs base-pairing patterns can be divided into four types: G:C match, A:U match, G:U wobble pairs and mismatch. The position 2-8 nt which is called *seed region* between 5' end of the miRNA and 3' end of the mRNA. Plant miRNA controls two regulatory modes which includes inhibition of translation and repression of mRNA expression [15-16]. The most important features of the miRNA-mRNA pair nucleotide matching information is considered as the seed region. Fig.1 shows the total alignment of miRNA and target mRNA and also defined the detailed overview about bulge ,match and mismatch. Figure 2 shows the exact cleavage position between miRNA 5' and target site of mRNA 3' [4-6].

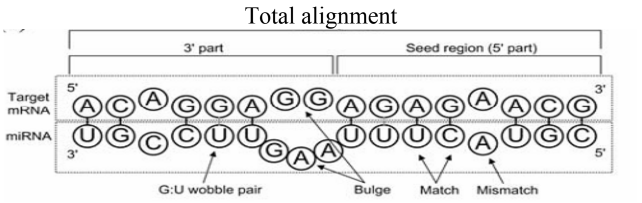


Fig. 1 General scheme of miRNA: mRNA interaction [20]

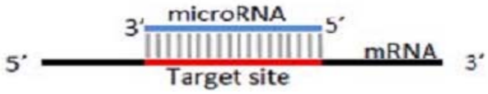


Fig. 2 Perfect matching of seed region [20]

3 String Matching Algorithms

3.3.1 Rabin-Karp String Matching Algorithm

Hashing provides a simple method to avoid a quadratic number of character comparisons in most practical situations, preprocessing phase in O(m) time complexity and constant space, searching phase in O(mn) time complexity, O(n + m) expected running time [17].

3.3.2 Knuth-Morris-Pratt String Matching Algorithm

The design of the Knuth-Morris-Pratt algorithm follows a tight analysis of the Morris and Pratt algorithm. Performs the comparisons from left to right, preprocessing phase in O(m) space and time complexity, searching phase in O(n+m) time complexity [17].

3.3.3 Boyer-Moor String Matching Algorithm

The algorithm scans the characters of the pattern from right to left beginning with the rightmost ,searching phase in O(mn) time complexity, 3n text character comparisons in the worst case when searching for a non periodic pattern, O(n/ m) best performance [17].

4. PROPOSED MODEL

The string matching algorithm model shown is in figure 3. From miRNA database the sequence of rice miRNA (*Oryza Sativa)* and from mRNA gene bank, sequence of tungro virus had been collected [18-19]. Three different string matching algorithms such as Boyer-Moore, Knuth-Morris, Rabin-Karp have been compared with web based tool as psRNATarget finder [8]. The criteria for finding the cleavage region such as maximum expectation, length of sequence and wobble pair were taken to authenticate our result considered the same input with psRNATarget web server (shown in table-1). All three algorithms are described in algorithm 1, algorithm 2 and algorithm 3 in section 5.

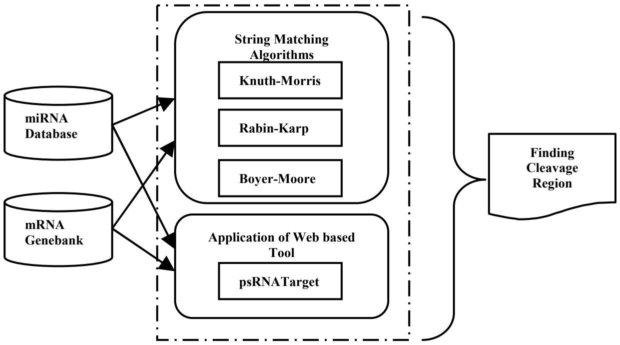


Fig. 3 String matching Algorithm model for miRNA target prediction

5. STRING MATCHING ALGORITHMS

Algorithm 1:

Algorithm: printSequence (Alphabet #, Text *T*[0, ..., *n*-1] #*n*, Pattern *P*[0, ..., *m*-1] #*m*

Return value: Prints output if pattern (with criteria) is a substring of *T*, -1 otherwise

```
1. R = ReverseAndComplementaryPattern(P)
2. for i from 0 to n-1 do
        2.1     Temp = substring of T from i to i+m-1 subject
                to i+m-1 <= n-1
        2.2     Z = MatchingSequence (R, Temp)
        2.3     Count the occurrence 'S', 'W', 'D' from Z
        2.4     if count(S)+count(W) > UPE and count(D) <=
                Expectation then
                2.4.1 Cleavage found at position i
                2.4.2 if z contains 'D' in first and/or last
                      sequence then
                        2.4.2.1 Ignore the first and/or last
                        sequence from R and temp
                2.4.3   Print ComplementaryPattern(R) and
                        temp
3. End
```

Function1: ReverseAndComplementaryPattern(Alphabet #, Pattern *P*[0, ..., *m*-1] #*m*)

Return value: return *P* with reverse and compliment characters i.e replace A->U, G->C, C->G, U->A and return reverse(*P*)

```
1. P = ComplementaryPattern(P)
2. Return reverse sequence of P i.e 3' to 5'
```

Function2: ComplementaryPattern(Alphabet #, Pattern *P*[0, ..., *m*-1] #*m*)

Return value: return *P* with compliment characters i.e replace A->U, G->C, C->G, U->A

```
1. for i from 0 to m-1
        1.1 if p(i) = A then p(i) = U and vice versa
        1.2 if p(i) = G then p(i) = C and vice versa
2. Return P
```

Function 3: MatchingSequence(Alphabet #, Pattern *R*[0, ..., *m*-1] , Pattern Temp[0, ..., *m*-1])

Return value: Return *Z* by putting characters '*S*', '*W*' and '*D*' where '*S*' stand for same/similar, '*W*' stands for wobble pair and '*D*' stand for different.

```
1. Z(0) = 0;
2. for i from 0 to m-1 do
        2.1 if R(i) = Temp(i) then Z(i) = 'S'
        2.2 if R(i) != Temp(i) and isWobblePair then Z(i)= 'W'
        2.3 if R(i) != Temp(i) then Z(i) = 'D'
3. Return Z
```

Algorithm 2:

Algorithm: ReverseAndComplementaryPattern(Alphabet #, Pattern *P*[0, ..., *m*-1] #m)return value: return *P* with reverse and compliment characters i.e replace A->U, G->C, C->G, U->A and return reverse(*P*).

```
1. P = ComplementaryPattern(P)
2. Return reverse sequence of P i.e 3' to 5'
```

Function1: ComplementaryPattern(Alphabet #, Pattern *P*[0, ..., *m*-1] #*m*)

Return value: return *P* with compliment characters i.e replace A->U, G->C, C->G, U->A.

```
1. for i from 0 to m-1
        1.1 if p(i) = A then p(i) = U and viceversa
        1.2 if p(i) = G then p(i) = C and viceversa
2. Return P
```

Function2: computeHash (Alphabet #, Text *T*[0, ..., *n*-1] #*n*, Pattern length #*m*)

Return value: Hash value of the given string.

```
1. h = 0 , R=256
2. for i from 0 to m do
        2.1 h = (R*h + n(i)) mod Q
3. Return h
```

Function3: Check (Alphabet #, Text *T*[0, ..., *n*-1] #*n*, Alphabet #, Pattern *P*[0, ..., *m*-1] #*m* , Index Position *p*)

Return value: True if matches else false

```
1. for i from 0 to m do
        1.1 if n(i+p) != m(i) than return false
        1.2 End if
2. Return true
```

Function4: Search (Alphabet #, Text *T*[0, ..., *n*-1] #*n,* Pattern *P*[0, ..., *m*-1] #*m*)

Return value: Offset position of the matching sequence

```
1. if $n < m$ return $n$ ;
2. $R$ = ReverseAndComplementaryPattern(Text $P$[0, ..., $m$-1])
3. $T1$ = computeHash(Text $T$[0, ..., $n$-1], $m$)
4. $T2$ = computeHash($R$ , $m$)
5. If $T1 == T2$ and check($T$[0, ..., $n$-1], $R$, 0) then return 0
6. Check for hash match; if hash match, check for exact match
   6.1 for $i$ from 0 to $m$ do
       6.1.1 Remove leading digit, add trailing digit, check for match
       6.1.2 $T1 = (T1 + Q – RM * T(i-m) \% Q) \% Q$
       6.1.3 $T1 = (T1* R + T[i]) \% Q$
       6.1.4 Offset = $i – m +1$
       6.1.5 If $T1 == T2$ and check($T$[0, ..., $n$-1], $R$, Offset) then return Offset
7. Return $n$
```

Function5: PrintMatchingSequence (Alphabet #, Text *T*[0, ..., *n*-1] #*n*, Pattern *P*[0, ..., *m*-1] #*m*)

Return value: Matching string with highest value

```
1. Offset = search($T$[0, ..., $n$-1], P[0, ..., $m$-1])
2. for $i$ from offset to offset+ $m$
   2.1 print $T[i]$ and $P[i]$
```

Algorithm 3:

Algorithm: printSequence (Alphabet #, Text *T*[0, ..., *n*-1] #*n*, Pattern *P*[0, ..., *m*-1] #*m*)

Return value: Prints output if pattern (with criteria) is a substring of *T*, -1 otherwise

```
1. $R$ = ReverseAndComplementaryPattern($P$)
2. for $i$ from 0 to $n$-1 do
   2.1 Temp = substring of $T$ from $i$ to $i+m-1$ subject to $i+m-1 <= n-1$
   2.2 $Z$ = MatchingSequence($R$,$Temp$)
   2.3 Count the occurrence '$S$', '$W$', '$D$' from $Z$
   2.4. if count($S$)+count($W$) > $UPE$ and count($D$) <= Expectation then
        2.4.1 Cleavage found at position $i$
        2.4.2 if $z$ contains '$D$' in first and/or last sequence then
              2.4.2.1 Ignore the first and/or last sequence from $R$ and temp
        2.4.3 Print ComplementaryPattern($R$) and temp
   2.5 End if
```

Function1: ReverseAndComplementaryPattern(Alphabet #, Pattern *P*[0, ..., *m*-1] #*m*)

Return value: Return *P* with reverse and compliment characters i.e replace A->U, G->C, C->G, U->A and return reverse(*P*)

```
1. $P$ = ComplementaryPattern($P$)
2. Return reverse sequence of $P$ i.e 3' to 5'
```

Function2: ComplementaryPattern(Alphabet #, Pattern *P*[0, ..., *m*-1] #*m*)

Return value: Return *P* with compliment characters i.e replace A->U, G->C, C->G, U->A

```
1. for $i$ from 0 to $m$-1
       1.1 if $p(i)$ = A then $p(i)$ = U and vice versa
       1.2 if $p(i)$ = G then $p(i)$ = C and vice versa
2. Return $P$
```

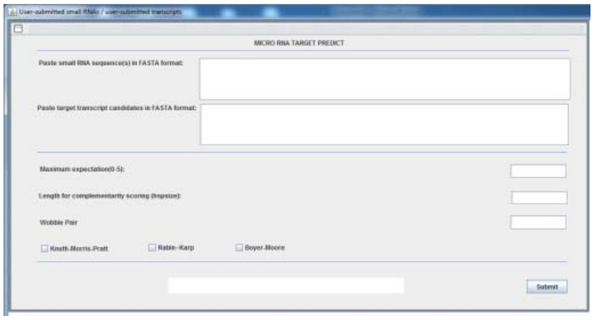
Function3: MatchingSequence(Alphabet #, Pattern *R*[0, ..., *m*-1] , Pattern Temp[0, ..., *m*-1])

Return value: Return *Z* by putting characters '*S*', '*W*' and '*D*' where '*S*' stand for same/similar, '*W*' stands for wobble pair and '*D*' stand for Different

```
1. $Z(0) = 0$;
2. for $i$ from 0 to $m$-1 do
       2.1 if $R(i)$ = Temp($i$) then $Z(i)$ = '$S$'
       2.2 if $R(i)$!= Temp($i$) and isWobblePair then $Z(i)$= ' '
       2.3 if $R(i)$ != Temp($i$) then $Z(i)$ = '$D$'
3. Return
```

6. EXPERIMENTAL EVALUATION AND RESULT ANALYSIS

To elucidate the efficiency of seed matches a script in JAVA was developed. 581 numbers of miRNA sequences of Oryza Sativa have been selected from miRBase[8] and target against tungro virus. The script detects perfect complementarity of the 5`end of a miRNA to a given viral coding sequence. In this paper we have compared the target sites found by different string matching algorithms as discussed in section 4 with web based tool psRNATarget finder [8]. In these algorithms the criteria for matching are *maximum expectation*, *length of nucleotides* and *wobble pair*. Maximum expectation is the threshold of the score. A target site pair will be discarded if its score is greater than the threshold. The default range is in between (0-5), length of miRNA (nt) having range between 19-24 and the maximum number of free wobble pair up to 3. The above criteria are default in all algorithms and applied on the rice endogenous miRNA sequences for targeting the tungro virus genes. A graphical user interface (GUI) for this is has been developed and shown given in figure 4 which takes the input miRNA in FASTA format and target transcript candidate in FASTA format with the parameters, maximum expectation ranges from 3~5 and length complemenatrity score with size ranging from 19~24. The table 1 shows the comparison between proposed string matching algorithms with web based psRNATarget [8]. In all the computational approach out of 581 endogenous miRNA of Rice (*Oryza Sativa*) plant few miRNAs target against different genes of coat proteins CP1,CP2,CP3, poly-protein of RTSV and ORF1, ORF2, ORF3, P12, P24, P46 and P194 of RTBV considered for simulation. Figure 5 shows the snapshot of the matching area found using string matching algorithms.



User-submitted small RNAs / user-submitted transcripts

MICRO RNA TARGET PREDICT

Paste small RNA sequence(s) in FASTA format:

Paste target transcript candidates in FASTA format:

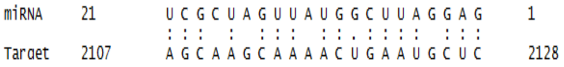
Maximum expectation(0-5):

Length for complementarity scoring (hspsize):

Wobble Pair

□ Knuth-Morris-Pratt □ Rabin-Karp □ Boyer-Moore

Fig. 4 GUI for string matching algorithms



```
miRNA    21       U C G C U A G U U A U G G C U U A G G A G     1
                  : : :   :   : : :   : : . : : : :   : : :
Target   2107     A G C A A G C A A A A C U G A A U G C U C     2128
```

Fig. 5 Finding matching area of string matching algorithms

Table 1: Genes of Rice Tungro Bacilliform virus (RTBV) and Tungro Spherical virus (RTSV) were targeted by Rice (*Oryza Sativa*) Plant miRNAs: A comparison between different string matching algorithms with psRNATarget

Oryza Sativa miRNA	**Knuth-Morris**	**Boyer-Moore**	**Rabin-Karp**	**psRNATarget**
Osa-miRNA2929	X	RTBV ORF4	RTBV ORF4	X
Osa-miRNA5150	RTSV CUTTACK CP3	RTSV POLYPROTEIN	RTSV CP POLYPROTEIN	X
Osa-miRNA5160	X	X	X	RTSV POLYPROTEIN
Osa-miRNA5157	RTBV P24	RTBV ORF4	X	X
Osa-miRNA5158	RTSV CUTTACK CP3	RTSV CP1	RTSV POLYPROTEIN	X
Osa-miRNA5503	RTBV P194	RTBV ORF3	X	RTBV ORF3,P194
Osa-miRNA5523	X	RTSV POLYPROTEIN	X	X
Osa-miRNA169	X	X	X	RTSV POLYPROTEIN
Osa-miRNA5525	RTSV POLYPROTEIN	X	X	RTSV POLYPROTEIN

7. CONCLUSION AND FUTURE WORK

All algorithms including web based tool were able to predict the rice miRNA target sites. By applying different string matching algorithms it was possible to find different new target sites as well as their potency of target will accelerate plant research for manipulation of plant production through developing disease resistance mechanism. The novel rice miRNAs and the targeted gene sequences of tungro viruses can be useful for plant biologist for further clarification of these miRNAs and their potentiality in biotechnological experiments for innovation of new rice varieties with better resistance against tungro viruses.

8. REFERENCES

[1] Hull, R. 2002. *Matthews' Plant Virology* (Academic, London).

[2] Azzam, O., Chancellor, T. 2002. The biology, epidemiology, and management of rice tungro disease in Asia. Plant Dis. ? 86. 88–100.

[3] Mew, T., Leung, H., Savary, S., Cruz, C., Leach, J. 2004. Looking ahead in rice disease research and management*. Crit. Rev. Plant. Sci.* 23. 103–127.

[4] Periasamy, M., Niazi, F. R., Malathi, V. G.2006. Multiplex RT-PCR, a novel technique for the simultaneous detection of the DNA and RNA viruses causing rice tungro disease*.* J. Virol. Methods. 134, 230–236.

[5] Dasgupta, I. 1991. Rice tungro bacilliform virus DNA independently infects rice after Agrobacterium-mediated transfer. *J. Gen. Virol*. 72, 1215–1221.

[6] Cullen., B. R. 2004. Transcription and processing of human microRNA precursors; Mol. Cell 16 861–865.

[7] Kim, S. K., Nam, J. W., Rhee, J. K., Lee, W. J., Zhang, B. T. 2006. miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*. 7. 411.

[8] Griffiths-Jones, S., Grocock, R. J., Van, D. S., Bateman, A. 2006. Enright AJ miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acid Research*. 34, D140-D144.

[9] Yoon, S., Micheli, G. D. 2006. Computational identification of microRNAs and their targets. *Birth Defects Research Part C: Embryo Today: Reviews*. 78, 2. 118 - 128.

[10] Pérez-Quintero, L. A., Neme, R., Zapata, A., López, C., 2010. plant microRNAs and their role in defense against viruses: a bioinformatics approach. BMC.Plant Biology, 10. 138.

[11] Dai, X., Zhao, X. P. 2011. psRNATarget: a plant small RNA targetanalysis server. Nucleic Acids Research. Web Server issue, Vol. 39. 155–159.

[12] Zhang, Y., Verbeek, J. F. 2010. Comparison and Integration of Target Prediction Algorithms for microRNA Studies, Journal of Integrative Bioinformatics, 7(3). 127.

[13] Tyagi, H., Rajasubramaniam, S., Venkat Rajam, M., Dasgupta, I 2008. RNA-interference in rice against Rice tungro bacilliform viruresults in its decreased accumulation in inoculated rice plants. *Transgenic Res*.17. 897–904.

[14] Warthmann, N., Chen, H., Ossowski, S., Weigel, D., Herve, P. 2008. Highly Specific Gene Silencing by Artificial miRNAs in Rice. *PLoS ONE*. 3, 3. e1829.

[15] Zhang, B., Pan, X., Wang, Q., Cobb, G. P., Anderson, T. A. 2006. Computational identification of microRNAs and their targets. Comput. Biol. Chem.30. 395-407.

[16] Tripathy, R., Mishra, D., Nayak, R. K. 2012. A Computational Approach of Rice (Oryza Sativa) Plant miRNA Target Prediction against Tungro Virus*, ICMOC-2012*(Accepted).

[17] Crochemore, M., Hancart, C. 1999. Pattern Matching in Strings, in Algorithms and Theory of Computation Handbook*, M.J. Atallah ed., Chapter 11, CRC Press Inc., Boca Raton, FL*. 11-1--11-28

[18] www.ncbi.nlm.nih.gov.

[19] http://microrna.sanger.ac.uk.

[20] www.wikipedia.org