Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

RetroPred: A tool for prediction, classification and extraction of non-LTR retrotransposons (LINEs & SINEs) from the genome by integrating PALS, PILER, MEME and ANN

Pradeep Kumar Naik[1, *], Vinay Kumar Mittal[1] and Sumit Gupta[1]

[1]Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Distt.-Solan, 173 215, Himachal Pradesh, India; Pradeep Kumar Naik* - E-mail- pknaik73@rediffmail.com; * Corresponding author

received January 04, 2008; revised January 14, 2008; accepted January 19, 2008; published January 27, 2008

Abstract:
The problem of predicting non-long terminal repeats (LTR) like long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) from the DNA sequence is still an open problem in bioinformatics. To elevate the quality of annotations of LINES and SINEs an automated tool "RetroPred" was developed. The pipeline allowed rapid and thorough annotation of non-LTR retrotransposons. The non-LTR retrotransposable elements were initially predicted by Pairwise Aligner for Long Sequences (PALS) and Parsimonious Inference of a Library of Elementary Repeats (PILER). Predicted non-LTR elements were automatically classified into LINEs and SINEs using ANN based on the position specific probability matrix (PSPM) generated by Multiple EM for Motif Elicitation (MEME). The ANN model revealed a superior model (accuracy = 78.79 ± 6.86 %, Q_{pred} = 74.734 ± 17.08 %, sensitivity = 84.48 ± 6.73 %, specificity = 77.13 ± 13.39 %) using four-fold cross validation. As proof of principle, we have thoroughly annotated the location of LINEs and SINEs in rice and Arabidopsis genome using the tool and is proved to be very useful with good accuracy. Our tool is accessible at http://www.juit.ac.in/RepeatPred/home.html.

Keywords: prediction; non-LTR retrotransposons; classification; LINEs; SINEs; artificial neural network

Background:
Long interspersed elements (LINEs) and short interspersed elements (SINEs) are non-LTR retrotransposons that reside within cells of a host organism, copying and inserting themselves into the host genome. Studies have revealed their ubiquity in many eukaryotic organisms, both plants and animals. However, the identification of repetitive elements still remains the *cinderella* of genome annotation. This can be due to both its technical (algorithmic) inherent complexity and to the prominent interest in determining coding portions of the genome. However, the situation is surprising for different reasons. Repetitive sequences are an important feature of eukaryotic genomes accounting for a large proportion of the genome; at least 50% of the human **[1]** and about 80% in some plants **[2]** genome seems to be composed by repetitive elements. They played an important role in the evolutionary game **[3]**. Moreover, some repetitive sequences are also an important tool in genomic analysis and discovery **[4]**. Finally, under a "technical" perspective, repetitive sequences in most cases represent a serious problem in the genome assembly steps. Understanding retrotransposable elements (RE) and their biological role has now become imperative in furthering research in functional and molecular genomics. One way of furthering our knowledge of RE biology is through the computational analysis of REs in the complete genomic sequences. By detailed comparison of the abundance and distribution of REs we can infer the fundamental biological properties that are shared or that differ among species.

The annotation of genomic repeats, typically relies on the results of a single computational program, RepeatMasker (http://www.repeatmasker.org/). Recently it has been reported that RepeatMasker may be ''neither the most efficient nor the most sensitive approach'' for annotation of genomic repeats **[5]**. However, with the development of several new methods for transposable elements and repeat detection **[6-9]**, it is now possible to apply a ''combined evidence'' approach to elevate the quality of RE annotations to a level comparable to that of gene models. Stratigically we have developed a RE annotation pipeline. This integrates the combined computational evidence derived from PALS **[10]**, Piller **[9]**, MEME **[11]** and ANN for detection of non-LTR elements and their classification into LINEs / SINEs.

Methodology:
Implementation of the tool
The stand-alone tool "RetroPred" is implemented in three separate phases (Figure 1). The first phase is meant for

identification of repeats in the genome using (PALS) **[10]** and (PILER) **[9]**. The input genomic sequence is aligned locally to itself using PALS which detect the position of transpose repeat signature within the genome. The output file is parsed by PILER to extract all the dispersed transposable elements from the genome and cluster together similar repeats. The repeats are processed using MEME **[11]** with energy value 0.01 for discovery of conserved pattern in a window size of 50. In the third phase the predicted genomic repeats are classified into LINEs and SINEs based on their conserved signature using ANN.

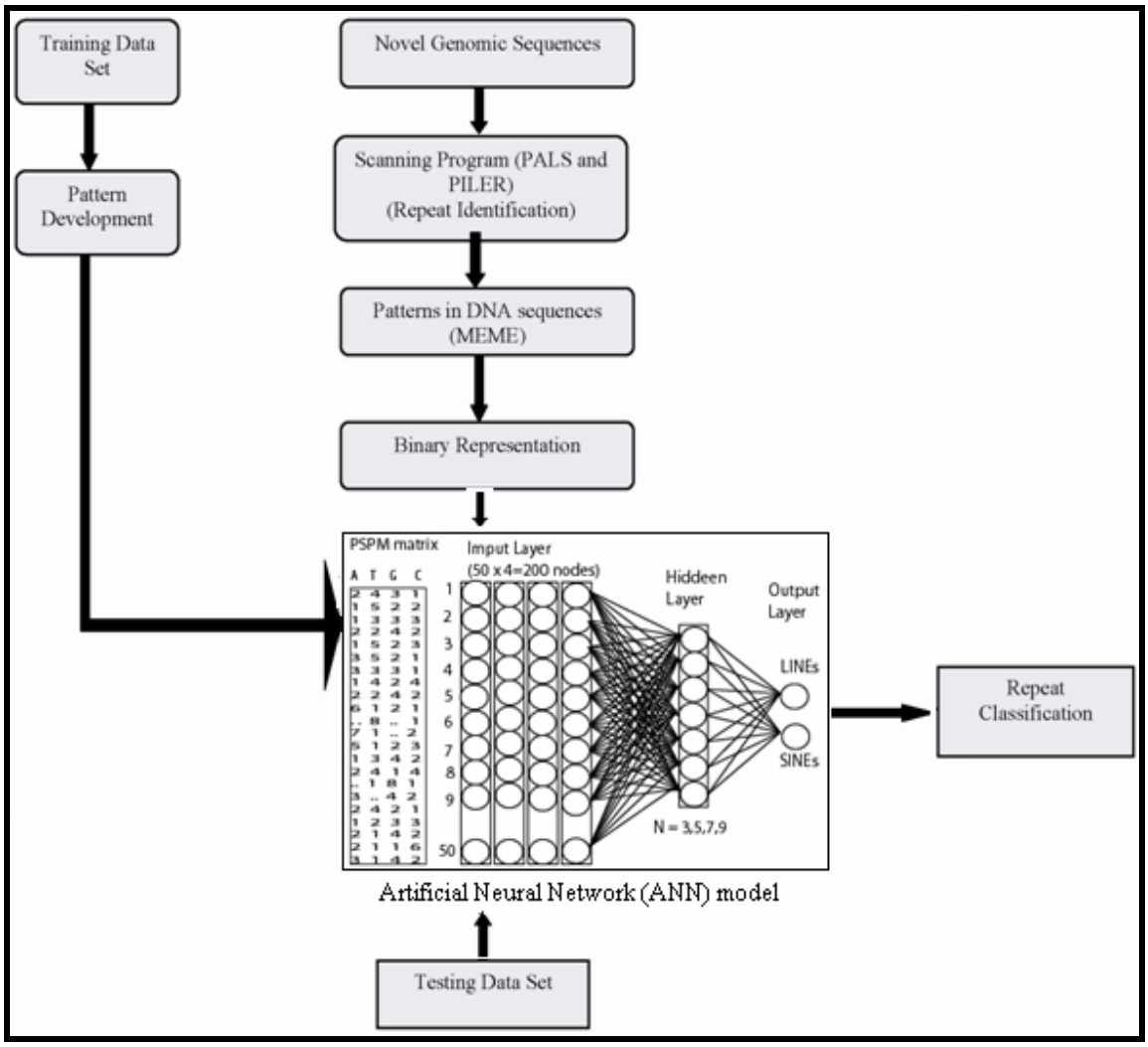

Figure 1: The flow diagram used for identification and configuration of artificial neural network (ANN) for classification of predicted non-LTR retrotransposons into LINEs and SINEs.

Dataset for identification of genomic repeats (LINEs and SINEs)

The genomic repeats belonging to LINEs and SINEs were obtained from several sources: Repbase (update database release 8.12) (downloaded from http://www.girinst.org); TIGR; and from MIPS (MIPS Repeat Element Database) containing a total of 253 LINEs and 350 SINEs sequences (Table 1 in supplementary material). We have taken 70 sequences of terminal repeats (non LINEs and non SINEs) from the Repbase as negative sequences.

Position specific probability matrix (PSPM)

The position specific probability matrices were built separately each for LINEs and SINEs using MEME. The matrix has 4 x M real-number elements, where M is the length of the sliding window (M = 50). Each element represents the probability of each nucleotide base appearing at each possible position for an occurrence of motifs using 0^{th} order Markov model. The steps followed for generation of PSPM matrix are described in Figure 2.

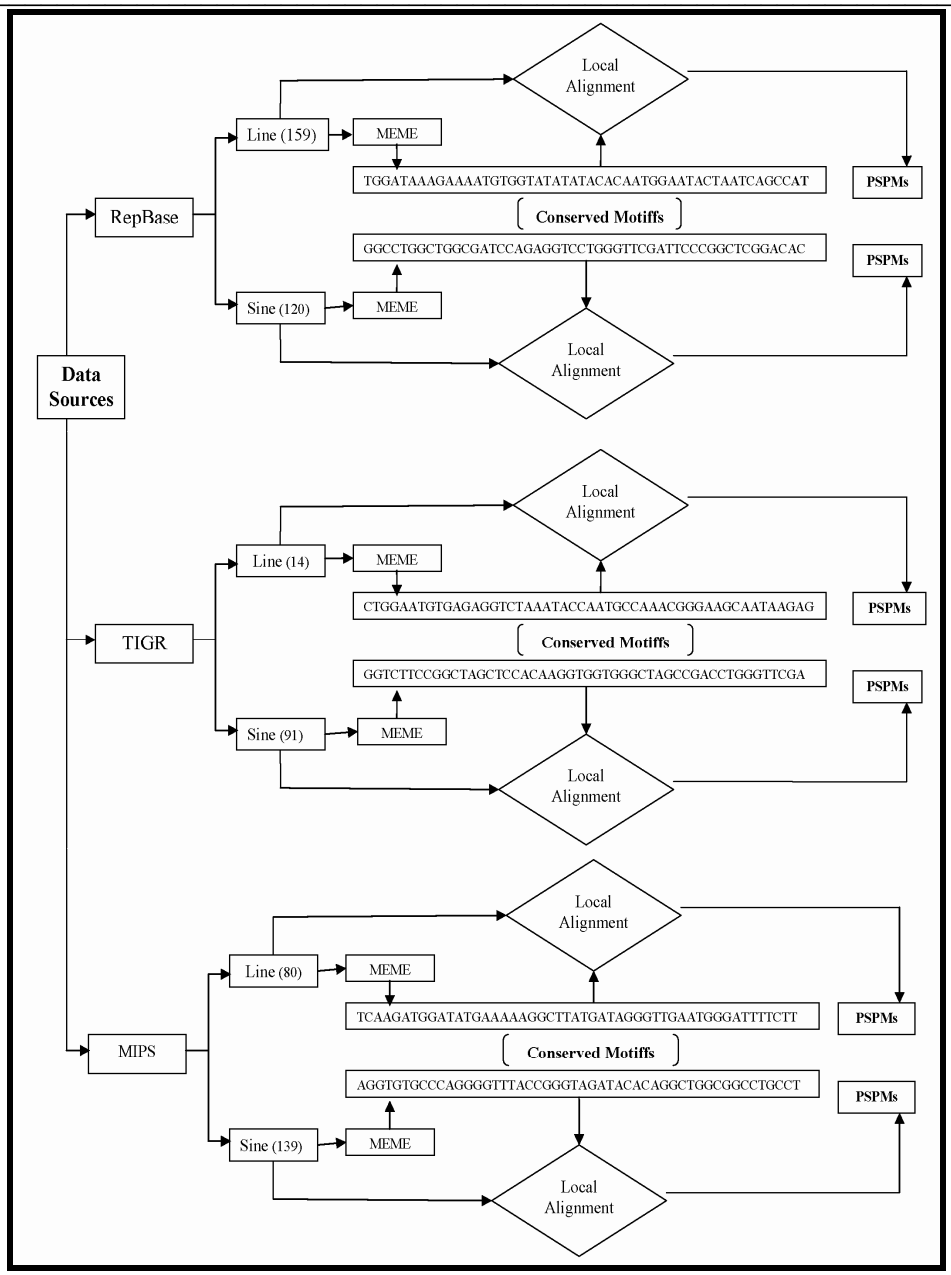

Figure 2: The steps followed for generation of position specific probability matrix (PSPM) of the datasets from three different sources using Multiple EM for Motif Elicitation (MEME).

Neural network architecture

The implementation of ANN was realized using the software package SNNS version 4.2 from Stuttgart University. The PSPM matrix generated by MEME was used as input to the neural network. The ANN configuration consists of 200 inputs and 2 output nodes to discriminate between LINEs and SINEs from the training sets (Figure 1). The number of nodes in the hidden layer was varied from 3 to 9 in order to find the optimal network that allows most accurate assignment of LINEs and SINEs (Table 2 in supplementary material). During the learning phase, a value of 1 was assigned for the LINEs and SINEs sequence whereas, 0 for the non-LINEs and non-SINEs. For each configuration of the ANN 110 independent training runs were performed to evaluate the average predictive power of the network. The corresponding counts of the false/true positive and negative predictions were estimated using 0.1 and 0.9 cut-off values for non-repeats and repeats respectively.

Fourfold cross-validation

A four-fold cross-validation technique has been used to validate the developed ANN model. The dataset is randomly divided into four subsets (C1, C2, C3 and C4). Each set is an unbalanced set that consist of about 60 percent of LINES/SINEs and 40 percent of non-LINEs/non-SINEs. The ANN was trained with three subsets and was validated (based on performance measure) for minimum error on testing set. This has been done four times to test for each subset. The final prediction result was averaged over four testing sets.

Performance measures

The prediction results of ANN model developed in the study were evaluated using the equations given in the supplementary material.

Results and discussion:

The ANN model develop in this study (200-7-2) is trained with the PSPM matrix calculated using MEME. When applying a fourfold cross-validation test, the network reached an overall accuracy of 95.31 ± 0.78 % for prediction of LINEs and 94.53 ± 1.44 % for SINEs prediction. The prediction results are presented in Table 3 (see supplementary material). The net has achieved an MCC of 0.9351 ± 0.0355 for LINEs and 0.8835 ± 0.0306 for SINEs prediction. The other performance measures were: Qpred = 97.99 ± 1.53 %, sensitivity = 94.17 ± 1.44 % and specificity = 97.04 ± 2.28 % for prediction of LINEs. However, performance measures of the network for prediction of SINEs were: Qpred = 95.03 ± 1.87 %, sensitivity = 96.37 ± 2.29 % and specificity = 93.02 ± 3.29 %. The value of the learning parameter was set to 0.1. The vast majority of the predictions of LINEs and SINEs have been contained within 0.9 to 1.0. However, the predicted output range for non-LINEs/non-SINEs is 0.0 to 0.1 (Table 3 in supplementary material). This illustrates that 0.1 and 0.9 cut-offs values provide very adequate separation of two bioactive classes using ANN. Performance of networks for prediction of LINEs and SINEs has been evaluated by calculating the area under the receiver output characteristic (ROC) curve. The areas under the curve is 0.97 for prediction of LINEs and 0.84 for prediction of SINEs; revealing a better discrimination of network system.

The reliability of developed tool for prediction, classification and extraction of genomic repeats (LINEs and SINEs) was performed by running the program on the complete genomic sequences of Rice and Arabidopsis downloaded form GenBank. The predicted results are shown in Table 4 under supplementary material (see the website http://www.juit.ac.in/RepeatPred/results.htm for more details). Our tool has predicted a total of 255 LINEs (0.114 % of entire genome) and 671 SINEs (0.292% of entire genome) out of 12 chromosomes from rice genome. Form the complete genome of Arabidopsis (5 chromosomes) the tool also predicted a total of 46 LINEs 0.04 % of genome) and 65 SINEs (0.082 % of the genome). The tool produces a graphical representation of the entire chromosome indicating the location of LINEs and SINEs in the chromosome (Figure 3) (see the website http://www.juit.ac.in/RepeatPred/results.htm for more details). By clicking the corresponding element one should extract the repeat sequence. Although the tool has been tested for two genomes, it can be used for prediction of LINEs and SINEs among other genomes too.

Figure 3: Graphical output of the program detecting the location of LINEs and SINEs on the chromosome. The red regions represent the location of SINEs and green region represent the LINEs in the chromosomal DNA. The position of the SINEs and LINEs are in the unit of mega basepair (Mbp).

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net

Robust *de novo* computational identification and classification of genomic repeats is an important unsolved problem. The most obvious difficulties are caused by multiple interacting evolutionary processes. For example, most repeats due to mobile elements were presumably intact at the time they were inserted into the genome, but today are often found as fragmented, degraded copies that may be adjacent to repeats belonging to other families and/or embedded in segmental or tandem duplications. Functional regions within segmental duplications may be conserved, producing a repeat signature that can mimic a mobile element. Raw genomic data could be searched for the presence/absence of these conserved features trying a *de novo* identification of putative non- LTR retroelements **[12]**. The developed tool "RetroPred" introduced a new approach to genomic repeat identification, classification and extraction of their sequences. In contrast to methods that involve self-alignment of a single genome, our comparative method searches for the conserved signature of LINEs and SINEs and are rely on the sequence similarity between different occurrences of retroelements in the genome. The results demonstrate that the developed ANN-based model is adequate and can be considered an effective tool for 'in silico' annotation of LINEs and SINEs from the complete genome.

Availability:

The program (standalone) is implemented on the Web server RetroPred, available at http://www.juit.ac.in/RepeatPred/home.html by using CGI/Perl script. Users can download the entire program and used for detection, classification and extraction of corresponding LINEs and SINEs sequence from the entire genome.

Conclusion:

Currently, there is no reliable systematic way for detection and classifying retroelements into LINEs and SINEs. Strategically, we have developed a neural network, fully automated computational method capable of classifying predicted genomic repeats into their subfamilies (LINEs and SINEs) based on their conserved sequence patterns. A user-friendly program RetroPred has been developed on the basis of this study. We have designed our system to be manually curated in an efficient manner for detection, classification and extraction of LINEs and SINEs, a goal that has important implications for experimental studies of genome and chromosome biology.

References:

[01] International Human Geneome Sequencing Consortium, *Nature*, 409: 860 (2001)
[02] R. Flavell, *Annu. Rev. Plant Physiol.*, 31: 569 (1980)
[03] J. C. Avise, *Science*, 294: 86 (2001)
[04] A. Kumar & H. Hirochika, *Trends in plant science*, 6:127 (2001)
[05] N. Juretic, *et al., Bioinformatics*, 20: 155 (2004)
[06] Z. Bao & S. R. Eddy, *Genome Res.*, 12: 1269 (2002)
[07] E. M. McCarthy & J. F. McDonald, *Bioinformatics*, 19: 362 (2003)
[08] O. Andrieu, *et al., BMC Bioinformatics*, 5: 94 (2004)
[09] R.C. Edgar & E.W. Myers, *Bioinformatics*, 21: I152 (2005)
[10] K. Rasmussen, *et al., Proc. RECOMB.*, (2005)
[11] L. Timothy, *et al., Proceedings of the Second International Conference on Intelligent Systems for*
[12] *Molecular Biology*, pp-28, AAAI Press, Menlo Park, California (1994)
[13] P. K. Mandal, *et al., Nucleic Acids Research*, 34: 5752 (2006)

Edited by P. Kangueane

Citation: Naik *et al.*, Bioinformation 2(6): 263-270 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Supplementary material

Equations

Performance measures:

The prediction results of ANN model developed in the study were evaluated using the following statistical measures.

(1) Accuracy of the methods: The accuracy of prediction for neural network models was calculated as follows:

$$Q_{ACC} = \frac{P+N}{T}$$

where T = (P+N+O+U), Where *P* and *N* refer to correctly predicted LINEs/SINEs and non-LINEs/non-SINEs, and *O* and *U* refer to over and under predictions, respectively.

(2) The Matthews correlation coefficient (MCC) is defined as:

$$MCC = \frac{(P \times N) - (O \times U)}{\sqrt{(P+U) \times (P+O) \times (N+U) \times (N+O)}}$$

(3) Sensitivity (Q_{sens}) and specificity (Q_{spec}) of the prediction methods are defined as:

$$Q_{sens} = \frac{P}{P+U}$$

$$Q_{spec} = \frac{N}{N+O}$$

(4) Q_{Pred} (Probability of correct prediction) and Q_{obs} (Percentage over coverage) are defined as:

$$Q_{pred} = \frac{P}{P+O} \times 100$$

$$Q_{obs} = \frac{P}{P+U} \times 100$$

Tables

Organism	LINEs	SINEs
(a) RepBase		
Arabdopsis Thaliana	12	20
Canis familiaris	2	7
Danio rerio	11	4
Homo sapiens	81	39
Monodelphis domestica	11	8
Mus musculus	11	4
Oryza sativa (Rice)	11	11
Rattus norvegicus	3	11
Sus scrofa	2	14
Triticum monococcum	10	2
Zea mays	5	0
Total	**159**	**120**
(b) MIPS		
Oryza Sativa	30	16
Brassica	26	63
Arabdopsis Thaliana	20	60
Solanum ochranthum	4	0
Total	**80**	**139**
(c) TIGR		
Oryza Sativa	14	91
Total	**14**	**91**
Grand Total	**253**	**350**

Table 1: The sources of dataset used for identification of genomic repeats and classification into LINEs and SINEs.

Hidden Nodes	Specificity	Sensitivity	Accuracy Q(Total)	Q(Pred)
(a) LINES				
3	0.8445	0.9496	0.9107	91.0219
5	0.9302	0.9611	0.9486	95.4180
7	0.9704	0.9417	0.9531	97.9967
9	0.6628	0.9340	0.8251	81.6108
(b) SINES				
3	3	3	3	3
5	5	5	5	5
7	7	7	7	7
9	9	9	9	9

Table 2: The variation in performance of the network with increasing hidden nodes for both LINES and SINES.

	Accuracy	Specificity	Sensitivity	Qpred	MCC	Prediction range (LINEs & SINEs)	Prediction range (non-LINES & non-SINES)
LINEs							
C1	0.9537	1.0000	0.9231	100.0000	0.9094	0.9785 - 0.9986	0.00 - 0.1111
C2	0.9626	0.9767	0.9531	98.3871	0.9767	0.9755 - 0.9857	0.00 - 0.04
C3	0.9528	0.9524	0.9531	96.8254	0.9020	0.9691 - 0.9874	0.00 - 0.02
C4	0.9434	0.9524	0.9375	96.7742	0.9524	0.9510 - 0.9914	0.00 - 0.09
Mean ± SD	0.9531± 0.0078	0.9704 ± 0.0228	0.9417 ± 0.0144	97.9967 ± 1.531	0.9351 ± 0.0355		
SINEs							
C1	0.9655	0.9767	0.9589	98.5915	0.9274	0.9975 - 0.9998	0.0029 – 0.003
C2	0.9483	0.9070	0.9726	94.6667	0.8887	0.9821 - 0.9949	0.0050 - 0.0855
C3	0.9561	0.9070	0.9859	94.5946	0.9068	0.9756 - 0.9831	0.0070 - 0.02
C4	0.9316	0.9302	0.9324	95.8333	0.8549	0.9173 - 0.9855	0.0054 - 0.01
Mean ± SD	0.9453 ± 0.0144	0.9302 ± 0.0329	0.9637 ± 0.0229	95.0315 ± 1.8684	0.8835 ± 0.0306		

Table 3: Performance measures of ANN model for classification of LINEs and SINEs using four fold cross validation.

Chromosome	Length of Chromosome (bp)	Number Of LINEs predicted	Length of LINEs (bp)	Percentage of LINEs	Number of SINEs predicted	Length of SINEs (bp)	Percentage of SINEs
Oryza sativa							
Chromosome1	43596840	34	60707	0.139246	80	208288	0.477759
Chromosome 2	35925420	25	21435	0.059665	92	98729	0.274817
Chromosome 3	36345540	25	27903	0.076771	62	75515	0.20777
Chromosome 4	47244300	22	137198	0.290401	66	81534	0.17258
Chromosome 5	29874180	22	12826	0.042933	33	21814	0.07302
Chromosome 6	31246800	11	31063	0.099412	39	59325	0.189859
Chromosome 7	29688660	19	17114	0.057645	53	126160	0.424943
Chromosome 8	28309260	24	60101	0.212302	50	83368	0.29449
Chromosome 9	23011260	13	9726	0.042266	43	92465	0.401825
Chromosome 10	22876560	22	14689	0.06421	29	74270	0.324655
Chromosome 11	28462200	29	25601	0.089947	71	85320	0.299766
Chromosome 12	27497280	9	19763	0.071873	53	113162	0.411539
Total	**3.84E+08**	**255**	**438126**	**0.114072**	**671**	**1119950**	**0.291594**
Arabidopsis thaliana							
Chromosome 1	30432570	3	1472	0.004837	21	23271	0.076467
Chromosome 2	19547030	9	13701	0.070092	17	17275	0.088377
Chromosome 3	23470880	6	4712	0.020076	12	8789	0.037446
Chromosome 4	18585110	14	13845	0.074495	6	4658	0.025063
Chromosome 5	26992730	14	14348	0.053155	9	43444	0.160947
Total	**1.19E+08**	**46**	**48078**	**0.040392**	**65**	**97437**	**0.08186**

Table 4: Prediction result of LINEs and SINEs in rice and Arabidopsis genome.