Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 6 **Hypothesis**

Mining of SSR markers from Expressed Sequence Tags of bamboo species

Oviya Iyappan Ramalakshmi[1]*, Shanmughavel Piramanayagam[2]

[1]Research Scholar, Department of Bioinformatics, Bharathiar University, Coimbatore- 641046, India; [2]Assistant Professor, Department of Bioinformatics, Bharathiar University, Coimbatore- 641046, India; Oviya Iyappan Ramalakshmi - Email: iroviya@yahoo.co.in; Phone: +91-422-2422265; Fax: +91-422-2422387; *Corresponding author.

Received September 08, 2010; Accepted October 11, 2010; Published November 27, 2010

Abstract:
With the ever increasing number of Expressed Sequence Tags (ESTs) from various sequencing projects, ESTs have become valuable and first-hand source of *in-silico* mining of simple sequence repeats (SSR) markers. We examined a total of 3419 EST sequences from three bamboo species, namely, *Phyllostachys edulis, Bambusa oldhamii and Dendrocalamus sinicus* for the presence of di- to hexa- microsatellites. The frequency of SSR containing ESTs varied from 5.36% in *B. oldhamii* to 13.05% in *P. edulis.* No SSRs were found in *D. sinicus.* Tri-nucleotide repeats (49.34%) were most frequent in *P. edulis,* while not much comparable difference in repeats was found in *B. oldhamii.* Flanking primer pairs were also designed *in-silico* for the sequences containing SSRs and their position on the genome hypothesized using similarity searching. SSRs located in open reading frame (ORF) were given functional annotation using Gene Ontology. Polymorphic SSRs were also detected using new pipeline- polySSR. Polymorphism level was very low (2.43%) and the position of the polymorphic SSRs was determined. The development of SSRs and the study of polymorphism will help in the further study of intra- and inter- gene flow, genetic structure, variability, linkage mapping and evolutionary relationships in bamboo.

Keywords: Expressed Sequence Tags (EST), Simple Sequence Repeats (SSR), bamboo, *Phyllostachys edulis, Bambusa oldhamii, and Dendrocalamus sinicus.*

Background:
A large number of plant genomes are under consideration for whole genome sequencing but for most of them, due to their large genome size, the task has been little difficult and time consuming. However, for them, as a prehand, large scale EST sequencing projects has been started as an alternative. Expressed Sequence Tags (EST) are cDNA clones, sequenced randomly in a single pass run. Based on the direction of cloning, it can be 5' EST or 3' EST. Since the ESTs are derived from cDNA, they provide direct evidence for the study of transcriptome and genome.

Microsatellites or Simple Sequence Repeats (SSR) or Short tandem repeats (STR) are 1-6 bp tandemly repeated motifs present in both coding and non-coding regions of prokaryotic and eukaryotic genome. SSR are extensively used as molecular markers because of its multiallelic nature, co-dominant inheritance and relative abundance. Since EST also represent the coding part of the genome, these serve as an important source for mining putative SSR markers and provide first hand insight into the organism's genetic diversity. Here, in this study, we present the mining of EST-SSR markers and detecting polymorphism in microsatellites in 3 bamboo species, namely, *Phyllostachys edulis*, *Bambusa oldhamii*, *Dendrocalamus sinicus.* Two approaches were used in this work. First, by using MISA, SSR markers were predicted and functional annotation of those SSR containing sequences was done. Another approach was to determine polymorphism using the novel pipeline PolySSR.

Methodology:
ESTs for 3 different bamboo species namely, *Phyllostachys edulis*, *Bambusa oldhamii* and *Dendrocalamus sinicus* were downloaded from NCBI's dbEST. A total of 3087 EST sequences of *P. edulis*, 318 EST sequences of *B. oldhamii* and 14 EST sequences of D. sinicus was downloaded from dbEST as on may 29, 2009, dbEST release 052909. dbEST has redundancy in EST sequences. In order to remove the redundancy, CAP3 Assembler **[1]** was used for clustering.

SSR were detected using MIcro SAtellite identification tool (MISA) **[2].** The search criteria were to search for a minimum of 14 bp SSR repeat. Primer pairs for the SSRs were designed using Primer3 (v 0.4.0) **[3]**. The web interface allows the user defined constraints; however, here the default parameters were used.

EST-SSR sequences were subjected to similarity searching against non-redundant (nr) database with constraint of ORGN: *Oryza sativa.* Bamboo being monocotyledon, rice was used as model organism owing to more proximity in phylogeny. It was performed using NCBI's Basic Local Alignment Search Tool (BLAST), variant BLASTX **[4]**. The sequence was considered homologous if the e-value was ≤ 1e-5 and score ≥ 100. Based on the position of SSR in the homology search, they were assigned whether they lie in 5' UTR, 3' UTR or ORF. Only those EST sequences were further analyzed in which SSR were predicted to be in ORF. The functionality was assigned according to Gene Ontology (GO) annotation. Rice was used as model organism for similarity searching. GO annotation of *Oryza sativa* was used to map functions by similarity searching in GRAMENE. Polymorphism exhibited by EST-SSR was mapped using the new pipeline PolySSR **[5].**

Results and Discussion:
For analysis, the EST data was significantly reduced to a nonredundant set of sequences by atleast 30% using CAP3. Mono-SSR repeats were not considered since they do not serve important as molecular markers. The search criteria were kept low to maximize the SSR discovery **[6].** In *D. sinicus*, no SSRs were detected by MISA and excluded for further study.

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 6 **Hypothesis**

Among cereal crops, tri-SSRs were most frequent followed by di- and tetra-SSR **[7]** and in our study also, tri-SSRs were found to be most abundant (49.34%) among all SSRs in *P. edulis* which is in agreement to other poaceae family members like barley, maize, sorghum, rice, wheat **[8,9,10].** In general, AT motif are common in plants **[11, 12].** However in bamboo, AT motif was least occurring. In *P. edulis*, among di-SSRs, AG/CT was most abundant. The result was consistent with the results from poaceae members **[8, 9**, and **10].** AG/CT motif can represent codons GAG, AGA, UCU and CUC in mRNA population and code for R, E, A and L respectively. Since A and L are found in increased amount in proteins, the abundance of AG/CT in the genome can be substantiated. CG repeats are least found in cereal species **[13]** and in our present study also, CG/GC motif was completely absent.

In plants, common motif is AAG. CCG is a specific feature of monocot genome **[14]** but in bamboo species, this trend was not observed. Among tri-SSRs, *P. edulis* showed abundance in AGG/CCT, AGC/CGT and AGT/ATC repeats. These motifs were fairly well represented in rice **[15**], maize **[16],** pearl millet **[13],** and barley **[8].** In our study, AAT/ATT motif was found to be nil. This may be explained because TAA- based variants code for stop codons that have direct effect on protein synthesis **[16].** In this study, tetra- and hexa repeats were found in lowest frequency **(Table 1 see supplementary material).**

In *B. oldhamii,* different SSR motifs very almost evenly distributed. Lack of considerable frequency difference in the occurrence of various SSR motifs in *B. oldhamii* and absence of SSRs in *D. sinicus* can be reasoned with the less number of EST sequences available and analyzed. Flanking primers pairs were designed for SSR containing sequences. For 90.57% of *P. edulis* EST-SSR sequences, primers were designed whereas for all SSR containing sequences of *B. oldhamii* primers were designed *in silico*. Sharma *et al.* (2009) **[17]** designed primers for 81.25% of SSR containing EST from *P. edulis* and *B. oldhamii* and were able to amplify 76.1% of those EST primers in lab. Most SSRs for which primers were designed were found to be < 20 bp length, thus has a smaller chance of mutation or slipped strand mispairing over smaller sequence length. This may lead to more chance of sequence conservation.

The SSR containing sequences for which primers were designed were then analyzed for determining the relative SSR position on genome using sequence similarity search. Most of the SSRs were predicted in 5' UTR followed by 3'UTR. Very few were predicted to be in ORF. Maximum SSRs were found to be in 5'UTR in bamboo species, in accordance to another study **[18].** Almost all AG motifs were found in 5' UTR where as CT motif was found maximally in 5' UTR and very few in 3' UTR. ORF contained mostly trinucleotide repeats **[19**]. High frequency of trinucleotide repeats can be explained as these are less affected by base mutations and hence more conserved. However, disease causing effect of change in SSR sequence in humans has been reported **[20].**
These SSR predicted to be in ORF were mapped against related Gene Ontology (GO) IDs. They were mapped with variable functions which are summarized giving the corresponding GO annotations for sequences in **Table 2 (see supplementary material).**

The polySSR pipeline resulted in 287 contigs and 2127 singlets of *P. edulis.* The contigs were analyzed for polymorphism. In all, only 7 contigs were found to contain 7 polymorphic SSRs. The 7 contigs included 21 *P. edulis* EST sequences, of which 4 EST sequences (189099774, 189100527, 189100246, 189100551) were also included in our final MISA analysis. The SSR motif -GA showed high polymorphism. GA/CT motif was also most polymorphic in rice **[10].** Polymorphism was previously reported in Bamboo species by Sharma *et al.* **[17**]. Primers were designed for the 7 contig sequences and position determination on genome was done. The presence of variation in repeat units of SSR in 5'UTR is said to affect gene transcription and/or translation, if in 3'UTR, they may be responsible for gene silencing or transcription slippage and if in ORF they inactivate or activate genes or truncate protein **[7, 21].** These SSRs can be used as molecular markers if the SSR containing genes are identified and their polymorphism validated. The summary of polySSR result is given in **Table 3 (see supplementary material)**. In *B. oldhamii* and *D. sinicus,* no polymorphic SSRs were found. Thus the divergent biological role of SSRs is important in the study of plant genomics and functionalities.

Acknowledgement:
The authors greatly acknowledge Elodie Salzemann, Wangeningen University, Netherland for using our data and providing us with the polySSR results for the study of polymorphism in EST-SSR in bamboo species. Also, the authors thank Pankaj Bhardwaj, IHBT, Himachal Pradesh for providing valuable suggestions and help.

References:

[1] http://www.genome.clemson.edu/resources/online_tools/cap3
[2] http://pgrc.ipk-gatersleben.de/misa/misa.html
[3] http://fokker.wi.mit.edu/primer3/input-040.htm
[4] http://blast.ncbi.nlm.nih.gov/Blast.cgi
[5] J Tang *et al. BMC Bioinformatics* **9**: 374 (2008) [PMID: 18793407]
[6] JR Ellis & JM Burke. *Heredity* **99**: 125 (2007) [PMID: 17519965]
[7] RK Varshney *et al. TRENDS in Biotechnology* **23**(1):48 (2005) [PMID: 15629858]
[8] T Theil Michalek *et al. Theor. Appl Genet* **106**: 411 (2003) [PMID: 12589540]
[9] N Nicot *et al. Theor Appl Genet* **109**: 800 (2004) [PMID: 15146317]
[10] RV Kantety *et al. Plant Molecular Biology* **48**: 501 (2001) [PMID: 11999831]
[11] A Shanker *et al. Scientia Horticulturae* **113**: 353 (2007)
[12] U Lagercrantz Ellergen *et al. Nucleic Acids Res* **21**: 1111 (1993) [PMID: 8464696]
[13] S Senthilvel *et al. BMC Plant Biology* **8**: 119 (2008) [PMID: 19038016]
[14] Li YC Korol *et al. Mol. Biol. Evol*. **21**(6): 991 (2004) [PMID: 14963101]
[15] S Temnykh *et al. Theor. Appl. Genet* **100**: 697 (2000**)**
[16] ECL Chin Maize *Genome* **39**: 866 (1996) [PMID: 8890517]
[17] V Sharma *et al. Genet* **10**: 721 (2009) [PMID: 18793407]
[18] A Bouck, T Vision *Molecular ecology* **16**: 907 **(**2007) [PMID: 17305850]
[19] G Toth *et al. Genome Res* **10**: 1967 (2000) [PMID: 10899146]
[20] G Bates & H Lehrach. *Bioessays* **16**: 277 (1994)
[21] S Fujimori *et al. FEBS Letters* **554**: 17 (2003) [PMID: 14596907]

Edited by S Datta
Citation: **Oviya & Shanmughavel**. 5(6): 240-243 (2010)

Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 6 **Hypothesis**

Supplementary material:

Table 1: Total SSR and Primer3 summary in Bamboo species.

S.No.		*P. edulis*	*B. oldhamii*
1	Total EST	3087	318
2	Nonredundant (nr) sequences	1899	222
3	Contigs	406	56
4	Singlets	1493	166
5	EST with SSR	229	12
6	% EST with SSR	7.42%	3.77%
7	EST-SSR with primers	194	12
8	nr sequences with SSR	141	10
9	% nr sequences with SSR	7.42%	4.50%
10	Contigs with SSR	53	3
11	% Contigs with SSR	13.05%	5.36%
12	Contig-SSR with primers	48	3
13	Singlets with SSR	88	7
14	% Singlets with SSR	5.89%	4.22%
15	Singlet-SSR with primers	71	7

Statistical summary of occurrence of microsatellites in EST and primer designed for them.

Table 2: Summary of GO IDs associated with bamboo sequences.

	gi / CONTIG ID	CORRESPONDING GO ID
P. edulis	216961680	GO:0003723, GO:0003725, GO:0005622
	216961479, 216961382	GO:0042309, GO:0050825, GO:0050826
	216961132, 216960875, 189100421	GO:0009536, GO:0009579, GO:0016020, GO:0003824, GO:0005198, GO:0005515, GO:0006808, GO:0016564, GO:0044237, GO:0050662
	189100560, contig197	GO:0006412, GO:0006464, GO:0015018, GO:0016020, GO:0016740
	189100399	GO:0000151, GO:0000166, GO:0004672, GO:0004674, GO:0004713, GO:0004842, GO:0005515, GO:0005524, GO:0005618, GO:0005634, GO:0005737, GO:0005739, GO:0005886, GO:0006464, GO:0006468, GO:0006950, GO:0007165, GO:0009719, GO:0016020, GO:0016301, GO:0016567, GO:0030246
	189100364	GO:0003676, GO:0003677, GO:0003700, GO:0005634, GO:0006350, GO:0006950, GO:0008150, GO:0008270, GO:0009628, GO:0009719
	189099984	GO:0003700, GO:0006350
	189099774	GO:0003824, GO:0006118, GO:0016616, GO:0051287
	189099710	GO:0003677, GO:0003700, GO:0005249, GO:0005622, GO:0005634, GO:0006350, GO:0006355, GO:0006412, GO:0006813, GO:0008076, GO:0009725, GO:0040007, GO:0045499
	189007168	GO:0003676, GO:0003723
	189007129	GO:0003676, GO:0003723, GO:0003824, GO:0005634, GO:0006139, GO:0009987, GO:0019538
	189006935	GO:0003824, GO:0005488, GO:0005515, GO:0008270
	189006821	GO:0016068, GO:0003677, GO:0004879, GO:0005634, GO:0006355
	contig36	GO:0042309, GO:0050825, GO:0050826
	contig74	GO:0003824, GO:0005198, GO:0005515, GO:0006808, GO:0009536, GO:0009579, GO:0016020, GO:0016564, GO:0044237, GO:0050662
	contig249	GO:0003723, GO:0003725, GO:0005622
B. oldhamii	113700295	GO:0000003, GO:0005576, GO:0005618, GO:0006629, GO:0006950, GO:0009628, GO:0016740, GO:0016787, GO:0016788
	113700110	GO:0005739, GO:0006457, GO:0009536, GO:0009579, GO:0009987, GO:0016020, GO:0019538, GO:0031072, GO:0051082

Various functions associated with SSR in EST sequences are summarized here.

Bioinformation Volume 5

open access

www.bioinformation.net Issue 6 **Hypothesis**

Table 3: Summary of polySSR results.

Cluster ID	SSR unit	Location	SSR type	SSR start	SSR stop	No of alleles	Allele info [alleleID:repeat times|No.ofseq(seqID)]	Specie info per allele [alleleID:No.of seq|species,No.of seq|another species]
37	GTGC	3'	4	475	491	2	1:4|1(0) 2:3|1(1)	1:1|189100551 2:1|216961050
70	CGC	unknown	3	91	112	2	1:7|1(0) 2:6|1(1)	1:1|189100246 2:1|189100527
105	GA	5'	2	145	159	2	1:5|1(0) 2:7|1(1)	1:1|189100110 2:1|216961217
107	AG	5'	2	98	116	2	1:8|5(0 3 4 5 6) 2:9|2(1 2)	1:1|189100012, 1|189100182, 1|189100315,1|189100419, 1|189100461 2:1|189100094,1|189100130
153	TTG	unknown	3	342	357	2	1:5|1(0) 2:4|3(1 2 3)	1:1|189099774 2:1|189100336, 1|189100416, 1|189100541
211	GT	5'	2	55	67	2	1:5|1(0) 2:6|1(1)	1:1|189006625 2:1|189007317
257	GAA	5'	3	117	129	2	1:3|1(0) 2:4|1(1)	1:1|189006729 2:1|189006920

The SSR motif in contigs exhibiting polymorphism, their position on EST and location on genome is summarized. Allele and specie information tells about the number of alleles, the repeat times of SSR motif and the EST sequences containing each allele is listed.