BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 8(6)

# Mining for SSRs and FDMs from expressed sequence tags of *Camellia sinensis*

**Jagajjit Sahu, Ranjan Sarmah, Budheswar Dehury, Kishore Sarma, Smita Sahoo, Mousumi Sahu, Madhumita Barooah, Mahendra Kumar Modi & Priyabrata Sen***

Agri-Bioinformatics Promotion Programme, Department of Agricultural Biotechnology, Assam Agricultural University, Jorhat-785013, Assam, India; Priyabrata Sen - Email: pbsen14@yahoo.co.in; Phone; +91-(376)-2340001; FAX: (376)-2340001, 2340101; *Corresponding author

Received March 16, 2012; Accepted March 21, 2012; Published March 31, 2012

**Abstract:**
Simple Sequence Repeats (SSRs) developed from Expressed Sequence Tags (ESTs), known as EST-SSRs are most widely used and potentially valuable source of gene based markers for their high levels of crosstaxon portability, rapid and less expensive development. The EST sequence information in the publicly available databases is increasing in a faster rate. The emerging *computational* approach provides a better alternative process of development of SSR markers from the ESTs than the conventional methods. In the present study, 12,851 EST sequences of *Camellia sinensis*, downloaded from National Center for Biotechnology Information (NCBI) were mined for the development of Microsatellites. 6148 (4779 singletons and 1369 contigs) non redundant EST sequences were found after preprocessing and assembly of these sequences using various computational tools. Out of total 3822.68 kb sequence examined, 1636 (26.61%) EST sequences containing 2371 SSRs were detected with a density of 1 SSR/1.61 kb leading to development of 245 primer pairs. These mined EST-SSR markers will help further in the study of variability, mapping, evolutionary relationship in *Camellia sinensis*. In addition, these developed SSRs can also be applied for various studies across species.

**Keywords:** Expressed Sequence Tags (ESTs), Simple Sequence Repeats (SSRs), Functional Domain Markers (FDM), *Camellia sinensis.*

## Background:

SSRs (Simple Sequence Repeats) otherwise known as Microsatellites are short repeat motifs that are present in both protein coding and non-coding regions of DNA sequences. SSRs show a high level of length polymorphism due to mutations of one or more repeats. The use of SSRs as molecular marker is being favored due to their multiallelic nature, reproducibility, high abundance and extensive genome coverage **[1].** Expressed Sequence Tags (ESTs) are single pass sequence of cDNA classes that provide direct information of gene expression and serve as the main source for microsatellites **[2].** The traditional methods of developing simple sequence repeat (SSR) markers are usually time consuming and labor-intensive. Generally these processes involve genomic library construction, hybridization with the repeated units of nucleotides and sequencing of the clones. The *computational* approach for developing SSR markers from ESTs provides a better platform than the conventional approach. The EST databases stores EST sequences which are redundant, so they contain repetitive units. The EST sequences can be obtained from the databases and undergo preprocessing and assembly to develop potential SSR markers. Numerous tools (both standalone and web-based) are available for the mining of EST data to design EST-SSR markers at a large scale. The free computational programs and large set of EST data available on the web helps the researchers to perform data mining very easily from their local system rapidly in a very low cost. The tools like cross_match and trimest provided non redundant high quality EST sequence which do not contain the vector contamination and poly-A, -T talils. CAP3 tried to assemble the

EST sequneces having overlapping region and produces Contigs by joining the sequeneces. The sequences that do not have common portions are remained as Singletons. MISA helped in detection of the Simple Sequence Repeats from both contigs and singletons.

EST-SSR markers are potential candidates for gene tagging and comparative studies in various species as they are gene specific. Studies have been reported regarding the use of EST-SSR markers in a large number of commercially important plants. Best known for its commercial value, *Camellia sinensis* (Tea), a perennial ever green shrub is the perfect companion in the daily life. It is perhaps the second most consumed beverage in the whole world with some important health benefits. Tea promotes a healthier immune system by lowering blood pressure and cholesterol. It is mainly produced in a measure amount in India and Sri Lanka in many varieties. Huge amount of EST data in tea have been generated and stored in the databases. Present study deals with the detection of SSRs from the EST sequences of *Camellia sinensis* available in the public domains and functional annotation of the SSR containing ESTs. The annotation helps to know the putative functions of the ESTs and to find the important functional domain markers (FDM) related to the SSR-ESTs leading to gene ontology study. The gene ontology covers three domains: biological process: operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms, cellular component: the parts of a cell or its extracellular environment and molecular function: the elemental activities of a gene product at the molecular level, such as binding or catalysis.

## Methodology:

### *Data source*

Sequence data were collected from the public domain GeneBank at NCBI (National Center for Biotechnology Information) **[3].** A search for EST sequences of *Camellia sinensis* resulted in 12851 sequences. All the sequences were downloaded to the local machine for further analysis.

### *Processing EST sequence*

6148 non-redundant EST sequences were searched for vector sequence and removal of the vector sequence was done using cross_match software **[4].** The vector sequences were obtained from the UniVec database **[5].** Then the trimest program from EMBOSS was used for removing the poly-A and poly-T ends of the EST sequences **[6].** This is freely available on the web which can be downloaded and installed in the PC. It allows to trim the poly-A and poly-T ends from the given sequence according to the parameters given.

### *EST assembly*:

After the EST sequences were fully processed up to trimming, they were subjected to CAP3 for assembly **[7].** CAP3 is a DNA sequence assembly program, freely available for academic use. CAP3 results Contig files, Singlets files, Qual files, Info files and Out files.

### *SSR detection*:

The 6148 unique EST sequences *i.e.* 4779 singlets and 1369 contigs were searched for microsatellite sequences using MISA (MIcroSAtellite identification tool) **[8].** MISA is a freely downloadable perl script available on internet. Alongwith misa.pl another file misa.ini was also downloaded which contains the search parameters.

### *Primer designing:*

Primer pairs were designed from the obtained SSR sequences using Primer3 tool **[9].** The parameters were changed according to own interest. The primer size parameter was changed to min-17, opt-21 and max-27. The GC% was changed to min-45%, max-65%. Then the SSRs were searched for both forward and reverse primers.

### *Functional annotation:*

The Functional Domain Markers were found from the SSR containing sequence using InterProScan at EBI **[10].** InterProScan provides the platform to analyze the functional domains with the help of the member databases such as BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, HAMAP, PatternScan, SuperFamily, SignalPHMM, TMHMM, HMMPanther and Gene3D. The SSR-ESTs sequences were searched for the significant matches using a special type of BLAST program BLASTx at National Center for Bioechnology Information against non-redundant protein database entries **[11].** BLASTx searches protein database using a translated nucleotide query. The BLASTx was performed keeping the value of identity parameter > 70%. The SSR-FDM contig sequences found from Interproscan were annotated for Biological process, Cellular component and Molecular function using QuickGO (http://www.ebi.ac.uk/QuickGO) at the EBI server **[12].** QuickGO is a fast web-based browser for Gene Ontology terms and annotations, which is provided by the UniProtKB-GOA group.

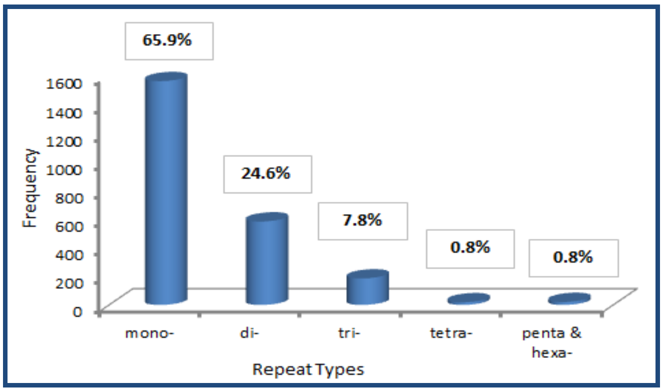

**Figure 1:** Frequency distribution of different repeat types identified in EST sequences of *Camellia sinensis.*

## Discussion:

Powerful computational tools were used to mine the publicly available *Camellia sinensis* EST data. In the present study, 2371 SSRs were found from the total screened EST data **Table 1 (see supplementary material)**. The EST sequences containing SSRs are generally referred to as SSR-ESTs, whereas the markers developed from SSR-ESTs are called EST-SSRs. Total number of 6148 sequences of size 3822.68 kb was examined for microsatellites, out of which 1636 (26.61%) numbers of sequences were found to contain SSRs with a density of 1 SSR/1.61 kb. **(Figure 1)** shows the frequency distribution of different repeat types identified and (**Figure 2)** shows the frequency distribution of mono, di, tri, tetra, penta and hexa-nucleotide repeat motifs in EST sequences of *Camellia sinensis* The most frequent repeat type found within the EST sequences were mono- nucleotide repeats (65.9%) followed by di-nucleotide repeats (24.6%), tri-nucleotide repeats (7.8%), tetra-nucleotide repeats (0.8%) and penta- and hexa-nucleotide repeats (0.8%). The frequency of identified SSR motifs are provided in the **Table 2 (see supplementary material)**. The detected SSR motifs put insight into the frequency distribution of different types of nucleotide repeats in *Camellia sinensis*. The analysis showed that the frequency is not so high but better than many other species. Ignoring the mono-nucleotide repeats, the di-nucleotide repeats were the most abudant repeat types. Among the di-nucleotide repeat types, CT and AG have the highest frequency. This adds emphasis to the fact that in plant EST-SSRs, AG/CT repeats have been found to be most frequent **[13, 14].** According to the previous reports, in most of eukaryotes, the GC repeats are abudant but in this study, however, the GC repeats were completely absent **[15, 16].** Compared to di-meric repeats, tri-nucleotide repeats are less in *Camellia sinensis*. Studies revealed that in plant the most common tri-SSRs are AAG and CCG **[17].** In Tea AAG repeat was not present and though CCG repeats were present but with less frequency. The most frequent tri-repeats in tea were GAA, TCT and CAC.

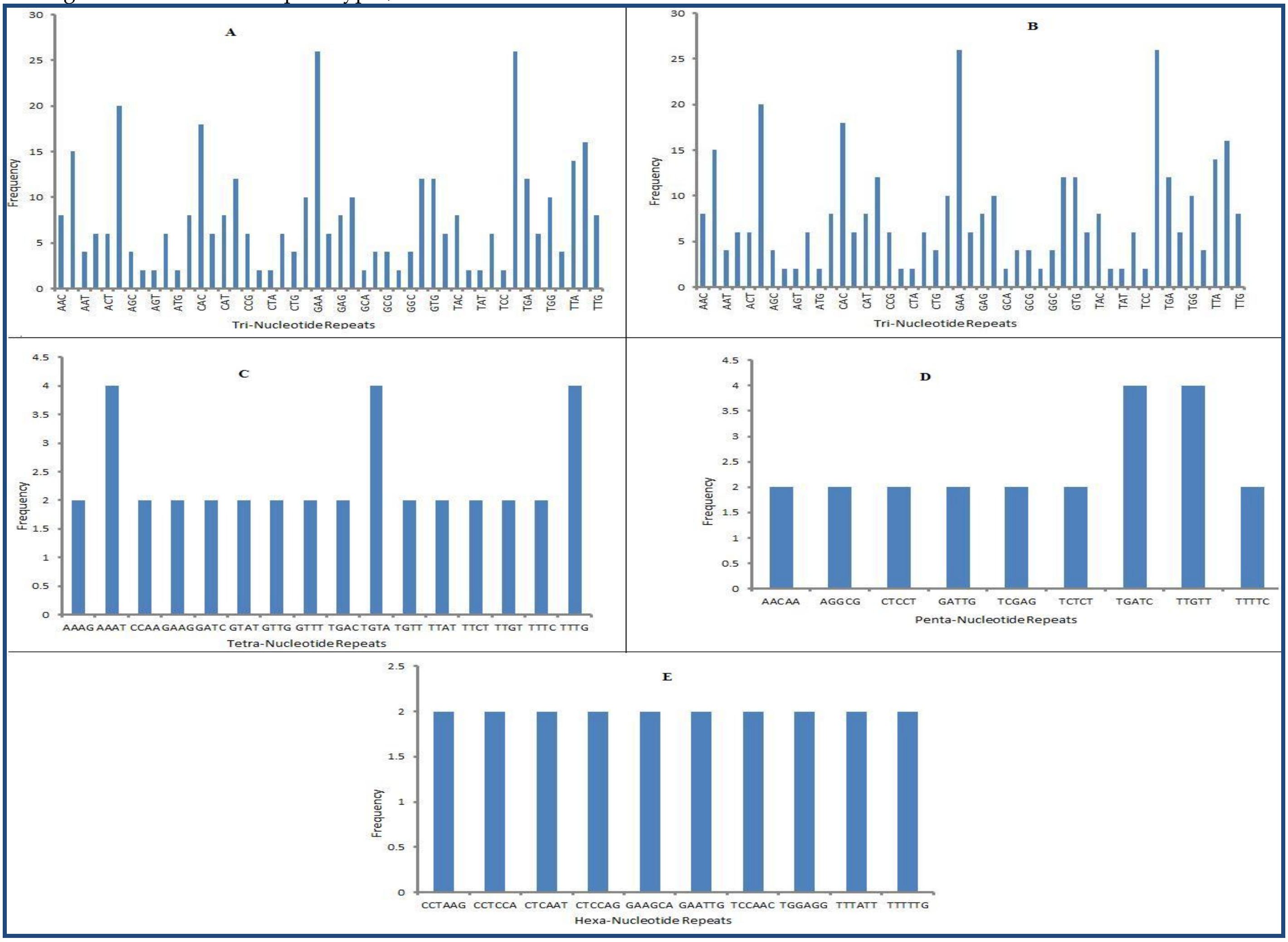

**Figure 2**: Frequency distribution of **(A)** mono- and di-, **(B)** tri-, (C) tetra-, **(D)** penta- and hexa-nucleotide repeat motifs in EST sequences of *Camellia sinensis*.

In Tea the length of the SSRs detected varried from 10 to 392. Polymerphic markers are mostly found in lengthy repeats. So the ESTs containing SSRs more than length of 100 were choosen for the design of primer pairs. Out of 97 such SSR-ESTs, 245 primer pairs were designed. (**Figure 3)** describess the flowchart of the stepwise procedure involved with *in sillico* mining of SSRs from *Camellia sinensis* EST sequences. From 469 SSR containing EST contigs, the Mono nucleotide SSR containing sequences were not taken into account. Using InterProScan the rest 314 contigs were analyzed and total 935 differentfunctional domains were found **Table 3 (see supplementary material)**. 250 sequences were found to be SSR-FDMs containing both SSRs and FDMs. These sequences were then assigned to gene ontology terms in the Swissprot database.

All the 469 SSR containing sequences were annotated against the non-redundant protein database to know the functions using BLASTx. The BLASTx result summarizes (**Figure 4**) 54 (11.51%) putative proteins, 74 (15.77%) predicted proteins, 50 (10.67%) hypothetical proteins, 146 (31.13%) of different functional classes and for 145 (30.29%) sequences no functions were found as there was no specific similarity. The SSR-ESTs were also assigned to gene ontology terms. 59 out of 250 contigs containing FDMs were found to have vital role in various processes, and functions. The (**Figure 5)** describes the details about the biological Processes, the Cellular Components and the Molecular Function of gene ontology. The gene ontology provided information regarding ESTs related to many genes coding for secondary metabolite production, translation, lipid transportation, stress response, GTP binding proteins, iron and sulfur cluster binding proteins and many other functions.

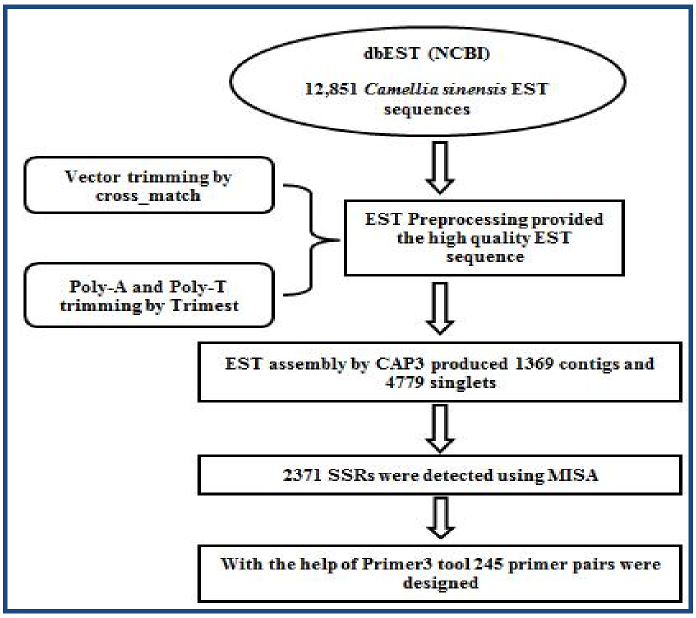
dbEST (NCBI)
12,851 *Camellia sinensis* EST sequences
Vector trimming by cross_match
Poly-A and Poly-T trimming by Trimest
EST Preprocessing provided the high quality EST sequence
EST assembly by CAP3 produced 1369 contigs and 4779 singlets
2371 SSRs were detected using MISA
With the help of Primer3 tool 245 primer pairs were designed

**Figure 3**: Work Flow Chart

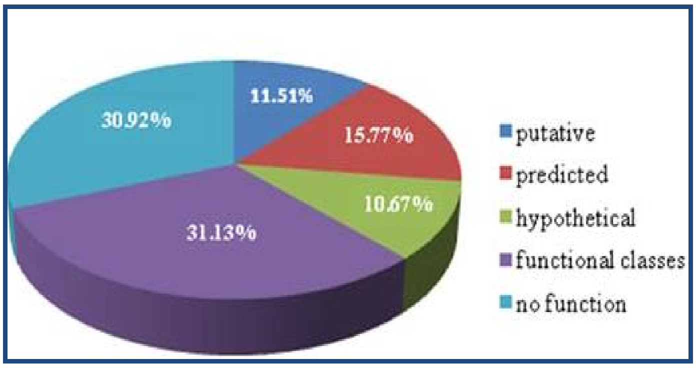

**Figure 4**: BLASTX results analysis.

Currently a number of studies are being reported regarding the development of EST-SSRs in vast number of plant species using the *computational tools*. Microsatellite markers are very important for studing genetic variability and understanding the genetic information. In the present study, 2371 microsatellites were detected from the available EST data for *Camellia sinensis*. The redundancy in the EST sequences was reduced by preprocessing and assembly of the EST sequences which helped in detecting unique SSRs. The primer pairs designed can be checked for the transferability in the related species. BLASTx results helped to know the putative functions of the ESTs and the GO annotation revealed lots of functional domains. This will lead to the development of the specific markers to find different secondary metabolites, stress responsible gene and many other gene information that may present in the plants.

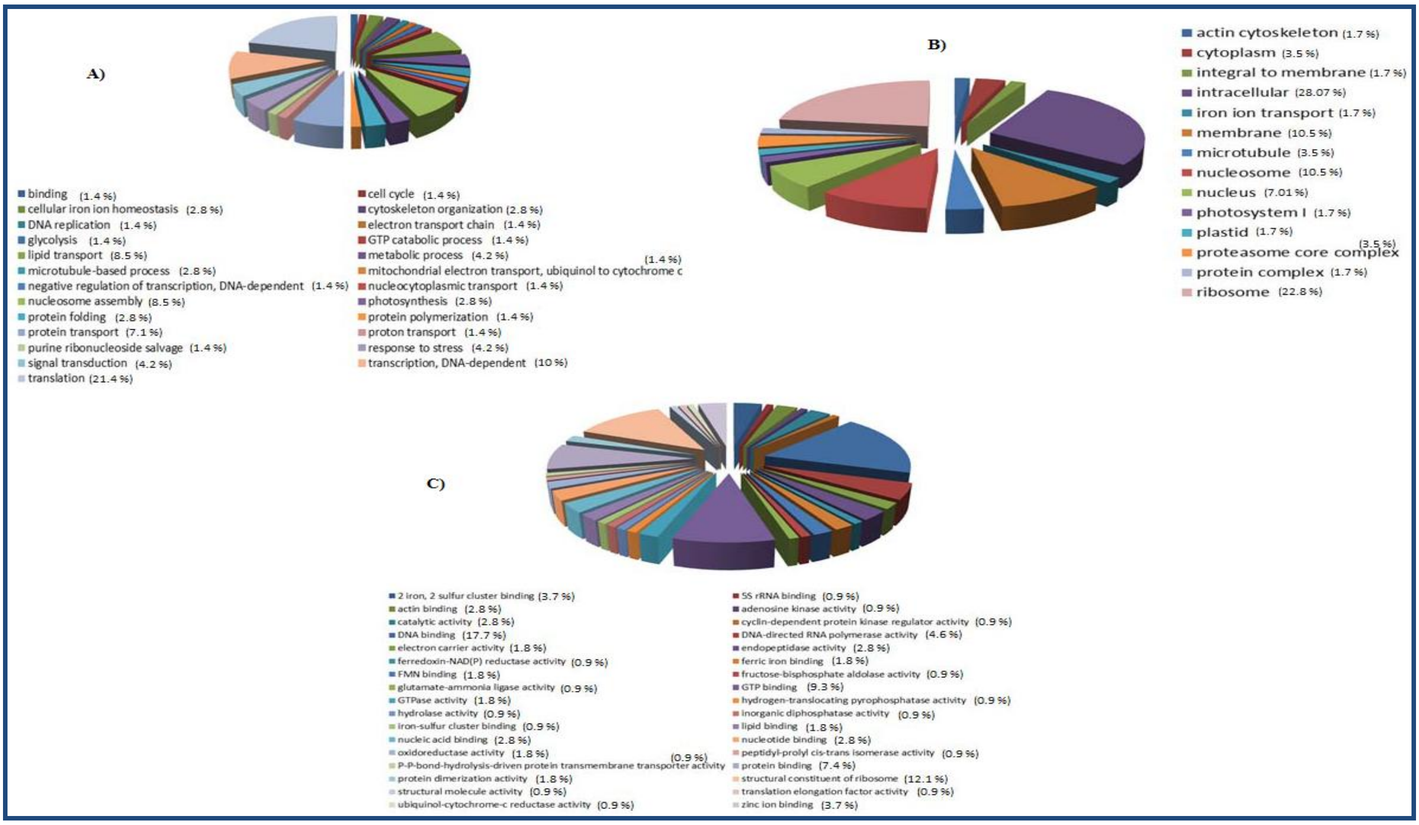

**Figure 5:** Gene Ontology **(A)** Biological Process; **(B)** Cellular omponent; **(C)** Molecular Function

## Conclusion:

*Camellia sinensis* commonly known as Tea is a plant of great commercial value. Till date not many work has been reported regarding the application of molecular markers in this plant mainly because of unavailability of suitable markers. Simple Sequence Repeats (SSRs) are the most powerful genetic markers for genetic linkage analysis, diversity study and marker assisted selection. The microsatellite markers are mainly being used in plant genetic analysis due to the reducing cost of DNA sequencing and increasing availability of EST sequence data in different plants. To look into the genetic make-up of *Camellia sinensis*, inter-species variability, evolutionary relationship study, development and application of molecular markers are of immense importance. The EST-SSRs developed in the present study will shed light into the discovery of the information. These SSRs can also be used as molecular markers to identify gene function, if the SSRs are found to be linked to a gene of importance. The SSRs derived from the ESTs can be used in the related species for which very less number of sequences is available because of high cross-species transferability nature of EST-SSRs. These can enhance the cross species applications to develop conserved orthologous marker sets. Also this study provides a brief idea about the approach to develop computationally mined SSRs from ESTs.

## Acknowledgement:

This project is supported by the Agri-Bioinformatics Promotion Program, Bioinformatics Initiative Division, Department of Information Technology, Ministry of Communications & Information Technology, Government of India, New Delhi. The authors are grateful to Assam Agricultural University, Jorhat, Assam for providing the necessary facilities as well as to BTIS Net, DBT, Government of India for constant support.

## References:

**[1]** Kantety RV *et al. Plant Mol Biol*. 2002 **48**: 501 [PMID: 11999831].
**[2]** Nagaraj SH *et al. Brief Bioinform*. 2007 **1**: 6 [PMID: 16772268].
**[3]** http://www.ncbi.nlm.nih.gov.
**[4]** http://www.phrap.org.
**[5]** ftp://ftp.ncbi.nih.gov/pub/UniVec/.
**[6]** http://emboss.sourceforge.net/apps/#Apps.
**[7]** Huang X & Madan A, *GenomeRes* 1999 **9**: 868 [PMID: 10508846].
**[8]** http://pgrc.ipk-gatersleben.de/misa.
**[9]** Rozen S & Skaletsky HJ, *Methods Mol Biol*. 2000 **132**: 365 [PMID: 10547847]
**[10]** Quevillon E *et al. Nucleic Acids Res*. 2005 **33**: W116 [PMID: 15980438].
**[11]** http://blast.ncbi.nlm.nih.gov/Blast.cgi.
**[12]** Binns D *et al. Bioinformatics*. 2009 **25**: 3045 [PMID: 19744993].
**[13]** Li B *et al. Yi Chuan Xue Bao*. 2004 **31**: 1089 [PMID: 15552043].
**[14]** Li B *et al. Genomics Proteomics Bioinformatics*. 2004 **2**: 24 [PMID: 15629040].
**[15]** Subramanian S *et al. Genome Biol*. 2003 **4**: R13 [PMID: 12620123].
**[16]** Astolfi P *et al. Gene*. 2003 **317**: 117 [PMID: 14604799].
**[17]** Li YC Korol *et al. Mol Biol*. 2004 **21**: 991[PMID: 14963101]

**Edited by P Kangueane**
**Citation: Sahu** *et al*. Bioinformation 8(6): 260-266 (2012)

## Supplementary material:

**Table 1**: The summary of the SSR detection using MISA

| **RESULTS OF MICROSATELLITE SEARCH** | |
|---|---|
| Total number of sequences examined: | 148 |
| Total size of examined sequences (bp): | 822676 |
| Total number of identified SSRs: | 371 |
| Number of SSR containing sequences: | 636 |
| Number of sequences containing more than 1 SSR | 08 |
| Number of SSRs present in compound formation | 40 |

**Table 2**: Frequency of identified SSR motifs (Except Mono-Nucleotides)

| **Repeat type** | **No of SSRs found** | **Repeat motif units** | **No of SSRs** |
|---|---|---|---|
| Di-Nucleotide | 583 | AC | 12 |
| | | AG | 110 |
| | | AT | 31 |
| | | CA | 4 |
| | | CT | 137 |
| | | GA | 108 |
| | | GT | 8 |
| | | TA | 24 |
| | | TC | 122 |
| | | TG | 27 |
| Tri-Nucleotide | 185 | AAC | 4 |
| | | AAG | 8 |
| | | AAT | 2 |
| | | ACC | 3 |
| | | ACT | 3 |
| | | AGA | 10 |
| | | AGC | 2 |
| | | AGG | 1 |
| | | AGT | 1 |
| | | ATC | 3 |
| | | ATG | 1 |
| | | CAA | 4 |
| | | CAC | 9 |
| | | CAG | 3 |
| | | CAT | 4 |
| | | CCA | 6 |
| | | CCG | 3 |
| | | CGC | 1 |
| | | CTA | 1 |
| | | CTC | 3 |
| | | CTG | 2 |
| | | CTT | 5 |
| | | GAA | 13 |
| | | GAC | 3 |
| | | GAG | 4 |
| | | GAT | 5 |
| | | GCA | 1 |
| | | GCC | 2 |
| | | GCG | 2 |
| | | GGA | 1 |
| | | GGC | 2 |
| | | GGT | 6 |
| | | GTG | 6 |
| | | TAA | 3 |
| | | TAC | 4 |
| | | TAG | 1 |
| | | TAT | 1 |
| | | TCA | 3 |
| | | TCC | 1 |
| | | TCT | 13 |
| | | TGA | 6 |
| | | TGC | 3 |
| | | TGG | 5 |
| | | TGT | 2 |
| | | TTA | 7 |
| | | TTC | 8 |
| | | TTG | 4 |
| Tetra-Nucleotide | | AAAG | 1 |
| | | AAAT | 2 |
| | | CCAA | 1 |
| | | GAAG | 1 |
| | | GATC | 1 |
| | | GTAT | 1 |
| | | GTTG | 1 |
| | | GTTT | 1 |
| | | TGAC | 1 |

| | | | |
|---|---|---|---|
| | | TGTA | 2 |
| | | TGTT | 1 |
| | | TTAT | 1 |
| | | TTCT | 1 |
| | | TTGT | 1 |
| | | TTTC | 1 |
| | | TTTG | 2 |
| Penta-Nucleotide | 11 | AACAA | 1 |
| | | AGGCG | 1 |
| | | CTCCT | 1 |
| | | GATTG | 1 |
| | | TCGAG | 1 |
| | | TCTCT | 1 |
| | | TGATC | 2 |
| | | TTGTT | 2 |
| | | TTTTC | 1 |
| Hexa-Nucleotide | 10 | CCTAAG | 1 |
| | | CCTCCA | 1 |
| | | CTCAAT | 1 |
| | | CTCCAG | 1 |
| | | GAAGCA | 1 |
| | | GAATTG | 1 |
| | | TCCAAC | 1 |
| | | TGGAGG | 1 |
| | | TTTATT | 1 |
| | | TTTTTG | 1 |

**Table 3**: FDM Analysis

| FDM_NAME | NO | FDM_NAME | No |
|---|---|---|---|
| Actin depolymerizing proteins | 1 | Yip1 | 1 |
| adh_short | 1 | GroES-like | 1 |
| ALDOLASE_CLASS_I | 1 | ADENOSINE KINASE | 2 |
| ALPHATUBULIN | 1 | ADF_H | 2 |
| Arm | 1 | AP2 | 2 |
| ARMADILLO/BETA-CATENIN REPEAT FAMILY PROTEIN / U-BOX DOMAIN-CONTAINING PROTEIN | 1 | ATS3 | 2 |
| Cell cycle regulatory proteins | 1 | AvrRpt-cleavage | 2 |
| CHAPERONIN10 | 1 | CCT | 2 |
| Cpn10 | 1 | CKS | 2 |
| CSA_PPIASE_2 | 1 | FLAVODOXIN_LIKE | 2 |
| CSAPPISMRASE | 1 | Glutamine synthetase/guanido kinase | 2 |
| Cyclophilin-like | 1 | LEA_3 | 2 |
| DNA_pol_delta_4 | 1 | Lipase/lipooxygenase domain (PLAT/LH2 domain) | 2 |
| Electron transport accessory proteins | 1 | LRR | 2 |
| ETHRSPELEMNT | 1 | NC | 2 |
| FeThRed_A | 1 | PLANT_LTP | 2 |
| FMN_red | 1 | P-loop containing nucleoside triphosphate hydrolases | 2 |
| GLNA_ATP | 1 | PREDICTED PROTEIN | 2 |
| GTPRANTC4 | 1 | PRELI | 2 |
| H_PPase | 1 | RNA-BINDING PROTEIN | 2 |
| L domain-like | 1 | Thioredoxin-like | 2 |
| L-aspartase-like | 1 | Tryp_alpha_amyl | 2 |
| LIGHT-INDUCIBLE PROTEIN ATLS1 | 1 | Ubiquinol-cytochrome c reductase | 2 |
| LIPOYL SYNTHASE | 1 | ORMDL | 2 |
| LSM | 1 | DNA-binding domain | 3 |
| MIF | 1 | LIPIDTRNSFER | 3 |
| N-terminal nucleophile aminohydrolases (Ntn hydrolases) | 1 | MSF1/PX19 RELATED | 3 |
| NUCLEAR ACID BINDING PROTEIN | 1 | PROFILIN | 3 |
| PfkB | 1 | PSI_PSAK | 4 |
| Pollen_Ole_e_I | 1 | RNA_pol | 4 |
| Preprotein translocase SecE subunit | 1 | RRM | 4 |
| Pro_isomerase | 1 | DUF | 5 |
| RAN | 1 | Ferritin | 5 |
| Ras | 1 | Glutathione S-transferase (GST), C-terminal domain | 5 |
| Remorin_C | 1 | Proteasome | 5 |
| small_GTP: small GTP-binding protein domain | 1 | Q3E953_ARATH_Q3E953/Q9FHL4_ARATH_Q9FHL4 | 5 |
| Sm-like ribonucleoproteins | 1 | SecE | 5 |
| Spt4 | 1 | ferredoxin-like | 8 |
| SUCROSE TRANSPORT | 1 | EFACTOR_GTP | 9 |
| SUGAR KINASE | 1 | ubiquitin | 9 |
| Tim17 | 1 | ZINC FINGER PROTEIN | 10 |
| Tmemb_14 | 1 | RIBOSOMAL PROTEIN | 17 |
| Translational machinery components | 1 | HISTONE | 20 |
| TRANSMEMBRANE PROTEIN | 1 | signal-peptide | 67 |
| TSP9 | 1 | Transmembrane_regions | 211 |