Bioinformation Volume 5 open access
www.bioinformation.net Issue 9 Hypothesis

Exploiting EST databases for the mining and characterization of short sequence repeat (SSR) markers in *Catharanthus roseus* L.

Raj Kumar Joshi, Basudeba Kar, Sanghamitra Nayak*

Centre of Biotechnology, Siksha O Anusandhan University, Bhubaneswar-751003, India; Sanghamitra Nayak- Email: sanghamitran@yahoo.com; Phone: 09437061976; *Corresponding author.

Received January 05, 2011; Accepted January 06, 2011; Published February 07, 2011

Abstract:
Periwinkle (*Catharanthus roseus* L.) (Family: Apocyanaceae) is a ornamental plants with great medicinal properties. Although it is represented by seven species, little work has been carried out on its genetic characterization due to non-availability of reliable molecular markers. Simple sequence repeats (SSRs) have been widely applied as molecular markers in genetic studies. With the rapid increase in the deposition of nucleotide sequences in the public databases and advent of bioinformatics tools, it has become a cost effective and fast approach to scan for microsatellite repeats and exploit the possibility of converting it into potential genetic markers. Expressed sequence tags (EST's) from *Catharanthus roseus* were used for the screening of Class I (hyper variable) simple sequence repeats (SSR's). A total of 502 microsatellite repeats were detected from 21730 EST sequences of turmeric after redundancy elimination. The average density of Class I SSRs account to 1 SSR per 10.21 kb of EST. Mononucleotides was the most abundant class of microsatellite motifs. It accounted for 44.02% of the total, followed by the trinucleotide (26.09%) and dinucleotide repeats (14.34%). Among all the repeat motifs, (A/T)n accounted for the highest Proportion (36.25%) followed by (AAG)n. These detected SSRs can be used to design primers that have functional importance and should also facilitate the analysis of genetic diversity, variability, linkage mapping and evolutionary relationships in plants especially medicinal plants.

Keywords: *Catharanthus roseus*, Expresses sequence tags, short sequence repeats, SSR Locator.

Background:
Catharanthus roseus (L.) G. Don, commonly known as periwinkle, of the family Apocynaceae is an ornamental plant with great medicinal value. The genus Catharanthus includes seven species besides *C. roseus* namely *C. coriaceus, C. lanceus, C longifoleus, C ovalis, C pussilus, C scitulus* and *C trichophyllus*. *C pussilus* is endemic to India while the rest are abundantly present in Madagascar. *Catharanthus roseus* possesses a suitable genetic system, as it is biannual, seed cyclable, diploid (2n=16) and amenable to controlled pollination and Micropropagation **[1, 2]**. As such, various genetic, proteomic and biotechnological studies is in progress in Catharanthus thereby making it a highly investigative plant species. An important breeding objective for the improvement of C. roseus is the development of molecular markers. Variation at DNA level is the key to modern studies and several DNA marker systems has been developed including restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), inter simple sequence repeats (ISSR), simple sequence repeats (SSR), amplified fragment length polymorphism (AFLP) and their variants in the recent times to analyze gross and specific DNA sequence variations in different species. Among different classes of molecular markers, microsatellite markers are the most favored.

Microsatellites, or simple sequence repeats (SSRs), are stretches of DNA consisting of tandemly repeated short units of 1–6 base pairs in length. Compared with other molecular markers, simple sequence repeats (SSRs) are more advantageous and used for a variety of applications because of their multi-allelic nature, reproducibility, co-dominant inheritance, high abundance and extensive genome coverage **[3]**. Various studies has been undertaken to identify microsatellites in different crop species including rice, wheat, maize, barley, soybean several medicinal and aromatic plants of commercial importance **[4]**. In C. roseus, a number of microsatellite markers have recently been developed and deployed for the study of intraspecific and interspecific as well as intrageneric and intergeneric genetic polymorphism **[5]**. However, most of these are genomic SSRs whose development is highly laborious, cumbersome and cost-intensive.

Currently, with the accumulation of biological data originating from whole genome sequence initiatives, the use of bioinformatics tools helps to maximize the identification of these sequences and consequently, the efficiency in the number of generated markers. Advances in genomic technologies have generated a large number of expressed sequence tags (ESTs) that has been made available in public database, thereby offering an opportunity to develop EST derived SSR markers by data mining. ESTs are short and single pass sequences read from mRNA (cDNA) **[6]** representing a snapshot of genes expressed in a given tissue. EST–SSR markers are expected to possess high interspecific transferability as they belong to relatively conserved genic regions of the genome. With recent increasing emphasis on functional genomics, large datasets of ESTs are being developed, and with evolving bioinformatics tools, it is now possible to identify and develop EST- SSR markers at a large scale in a time and cost-effective manner **[7, 8]**. As of January 2011, GenBank had released 21730 EST sequences from *Catharanthus roseus* (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). In this context, the use of EST or cDNA-based SSRs has been reported for several species including grape **[9]**, sugarcane **[10]**, durum wheat **[11]**, rye **[12]** and medicinal plant like basil **[13]**. Keeping the above point of view, the present study aims to assess the suitability of existing public databases for the mining of simple sequence repeats. We have mined updated EST libraries of *Catharanthus roseus* for this analysis to find the SSR polymorphisms. There are various SSR detecting softwares such as MISA **[14]**, SSRFinder **[15]**, SSRIT **[16]**, TRF **[17]**, TROLL **[18]**, Sputnik (http://espressosoftware.com/pages/sputnik.jsp), Modified Sputnik I **[19]** and Modified Sputnik II **[20]**. However, we used a new SSR detecting software SSR Locator **[21]** to identify the SSR polymorphisms because of its user friendly Windows interface and ability to interpret results as html files.

Methodology:

EST database of NCBI contains 21730 *Catharanthus roseus* express sequence tag data. 21730 sequences retrieved were related to different plant tissues i.e leaves and root. The downloaded sequences were obtained in FASTA format for sequence assembly and SSR analysis. A single text file was compiled containing all the 21730 EST sequences.

The EST sequences were screened against the UniVec database from NCBI (ftp://ftp.ncbi.nih.gov/pub/ UniVec/) for detecting vector and adapter sequences by using the program Cross_Match **[22]**; the following parameters were used: minmatch ≥13 and minscore ≥20. Furthermore, polyA/T tails and X characters were removed using the EST_trimmer.pl script (http://pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl) until no stretch of (A/T)5 or (X)1 was present in a window of 100bp at the 5′ or 3′ end, respectively. dbEST has redundancy in EST sequences. In order to remove the redundancy, EST sequences were assembled using the contig assembly program CAP3 **[23]**. The sequences containing file was submitted in FASTA formatted text file. The results were in different output files e.g. unigenes, contigs and singlets. For the purpose of the SSR identification, we combined the contig and singlet sequences to form non-redundant sequence data set. The SSR detection tool SSR locator was used to detect EST-SSR loci. EST-derived SSRs were considered to contain motifs ranging in length from 2 to 6 nucleotides, with dinucleotide repeat numbers ≥9, trinucleotide repeat numbers ≥6, tetranucleotide repeat numbers ≥5, pentanucleotide repeat numbers ≥4, hexanucleotide repeat numbers ≥3, and compound SSR motif length ≥24 bp.

Results and Discussions:

ESTs are often represented by redundant cDNA sequences making it difficult to analyze them effectively for SSRs. To overcome this problem, the CAP3 program was used. The reduction in redundancy is used as a measure of degree of overlapping between EST sequences. The objective was the elimination of redundancy in EST sequences and arriving at a contiguous sequence (contigs) which can be used for analysis of SSRs. CAP3 is a commonly used program **[24]**, which identifies overlapping sequences and generates contigs with consensus sequences.

The 21730 redundant EST sequences retrieved from NCBI scanned for Class I microsatellite repeats represented approximately 9.25Mb of *Catharanthus roseus* genome. 632 SSR's were detected in this dataset corresponding to 1.0 SSR per 14.6 kb. During pre-processing, 835740 bp of empty vectors, low-quality sequences and Poly A/T tails were removed successfully. Trimming of poly A and poly T tails resulted in the removal of 9% of original dataset. Rest of the sequences were clustered and assembled into a non-redundant dataset of 5928 unique gene sequences (1227 contigs and 4704 singlets). Scanning of Class I microsatellites in this non-redundant dataset revealed 502 unique SSR containing sequences **(Table 3, see supplementary material)**. This accounts to 1SSR per 10.21kb of *Catharanthus roseus* genome **(Table 1, see supplementary material)**. The reduction in redundancy of Class I SSR's obtained by trimming and clustering of non-redundant dataset is shown in **Figure 1**. Seventeen cases were found where two microsatellites were immediately adjacent to each other; 57 ESTs contained two adjacent repeats at a distance of <10 bp to each other. Cardle *et al* 2000 **[25]** estimated the average distances (in kb) between SSRs in sets of non-redundant ESTs of various plants such as rice (3.4), soybean (7.4), tomato (11.1), Arabidopsis (13.8), poplar (14.0) and cotton (20.0) through a comprehensive computational study. Considering the same criteria in the present study, SSRs occurs with a frequency of 1SSR per 14.6 kb in Catharanthus ESTs. This suggests that the frequency of cDNA-SSRs in the expressed portion of the Catharanthus genome is low in comparison to rice, soybean, tomato and Arabidopsis and higher than other plant species.

The occurrence of the individual SSR motifs among the non-redundant set of 502 SSRs is summarized in **Table 2 (see supplementary material)**. Obviously, the proportion of SSR unit sizes was not evenly distributed: 221 (44%) were mononucleotide, 72 (14%) dinucleotide, 131 (26%) trinucleotide, 27 (5%) Tetranucleotide, 21 (4%) Pentanucleotide and 22 (4%) hexa-nucleotide microsatellites. 1.6% of microsatellites were of compound types. The relative abundance of mononucleotides even after the trimming of poly A/T tracts could clearly indicate their occurrence of genome rather than at the ends of mRNA. Regarding dimeric SSRs, the motifs AG (41%) and AC (24%) were by far the most common ones, whereas AT is present only at low abundance i.e 7% only. The deficiency of AT SSRs in EST sequences is in accordance with reports from rice **[26]**, Arabidopsis **[25]** and maize **[27]**. 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 **[28]** and in our present study CG/GC motif was completely absent. Among trimeric microsatellites, AAG (43%), AGG (22%) and AGC-CCG (19%) were the most-common motifs. ACG also accounted for 14%. In plants, common motif is AAG while CCG is a specific feature of monocot genome **[14]**. Lu *et al* 2010 **[29]**, have found (AAG)n to be the most abundant repeat motif in Gossypium barbadense. Siju *et al* **[30]** also found (AAG)n to be the most abundant in turmeric accounting to 8.2%. Similarly, trimeric motifs like AGG/CCT, AGC/CGT are also fairly well represented in rice **[26]**, maize **[27]**, pearl millet **[28]**, and barley **[14]**. ATT represented only 3% of the total trimeric SSRs. Most of the monocots and dicots used to have least amount of motif ATT owing to the fact that TAA-based variants code for stop codons that have a direct effect on protein synthesis **[27]**. The most-frequent tetrameric microsatellite motifs were AAAC/GTTT and AAAT/ATTT. The penta and hexa nucleotides were found to be in lowest frequency. In all the repeat motifs, most of the SSR repeat motifs derived from the ESTs were A/T (36.25%) followed by AAG/CTT (8.56%), AG/CT (8.16%) and G/C (7.76%) **(Figure 2)**. Rest of the repeat motifs accounted less than 5% contribution to the total SSR motifs. In the 1−6 repeat types, the most frequent repeat motifs were A/T, AG/CT, AAG/CTT, AAAC/GTTT, AAAAC/GTTTT, and AAAGCC/GGCTTT, which accounted for 56.94%, 57.20%, 32.82%, 37.03%, 42.85% and 22.72% of all types, respectively.

Conclusion:

Microsatellites serve for divergent roles in the field of plant genomics. EST databases provide a valuable resource for the development of microsatellite markers, which are associated with transcribed genes. Development of SSR markers from EST-databases saves both cost and time, once sufficient amounts of EST sequences are available. In the present study, we identified 502 non-redundant hyper variable microsatellites from EST data source of *Catharanthus roseus* using SSR identification tool SSR Locator. The frequency analysis described in the study could offer potential for the designing of suitable repeat probes for effective targeting and isolation of microsatellite repeats from Catharanthus. Moreover, these non-redundant SSR resources can be used to design informative SSR primers that can be applied in studies of genetic variation, linkage mapping, comparative genomics and characteristic distribution of genes on chromosomes of Catharanthus species.

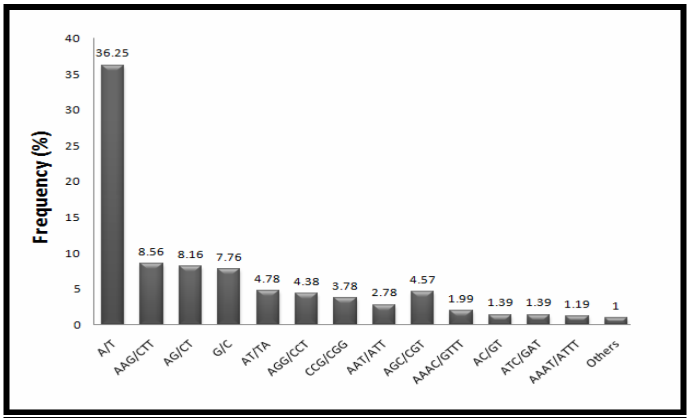
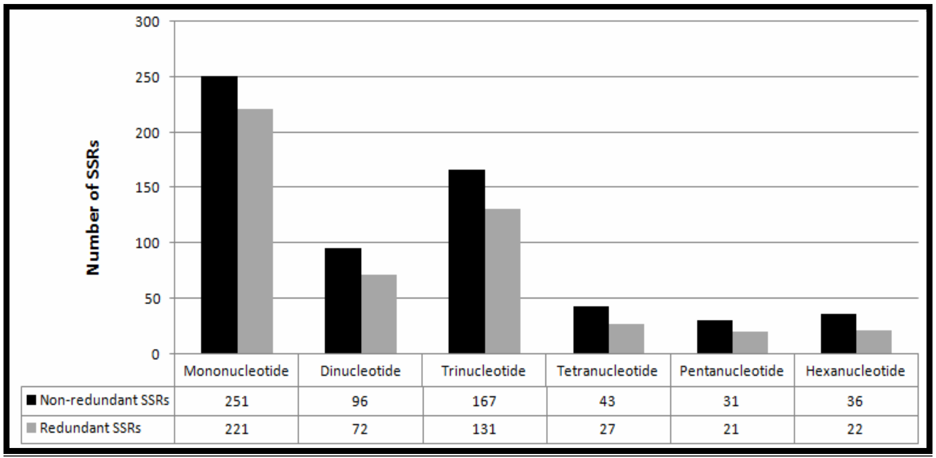


Figure 1: Distribution of EST-SSRs based on the motifs.



	Mononucleotide	Dinucleotide	Trinucleotide	Tetranucleotide	Pentanucleotide	Hexanucleotide
■ Non-redundant SSRs	251	96	167	43	31	36
■ Redundant SSRs	221	72	131	27	21	22

Figure 2: Reduction in redundancy by trimming (Poly A/T tails) and assembling *Catharanthus roseus* ESTs.

Acknowledgement:
The authors would like to thank Prof. Manoj Ranjan Nayak, President, Siksha O Anusandhan University for his able guidance and support.

References:

[1] P Mishra & S Kumar, *J. Med. Aromat. Plant Sci*. **22**: 306 (2000)
[2] Debanata *et al. Curr. Pharm.Biotech*. **7**: 35 (2006)
[3] PK Gupta & RK Varshney. *Euphytica* **113**: 163 (2000)
[4] KP Tripathi *et al. Plant Omics J* **2**: 228 (2009)
[5] B Shokeen *et al. Mol. Biol. Notes.* **5**: 818 (2007)
[6] MD Adams *et al. Science* **252**: 1651 (1991) [PMID: 2047873]
[7] RV Kantety *et al. Plant Mol Biol* **48**: 501 (2002) [PMID: 11999831]
[8] RK Varshney *et al. Cell Mol Biol Lett*. **7**: 537 (2002)
[9] KD Scott *et al. Theor Appl Genet* **100**: 723 (2000)
[10] G Cordeiro *et al. Plant Sci* **160**: 1115 (2001)
[11] I Eujayl *et al. Theor Appl Genet* **104**: 399 (2002) [PMID: 12582712]
[12] B Hackauf *et al. Plant Breed* **121**: 17 (2002)
[13] S Gupta *et al. Plant Omics J* **3**(4): 121 (2010)
[14] T Thiel *et al. Theor Appl Genet* **106**: 411 (2003) [PMID: 12589540]
[15] L Gao *et al. Mol Breed* **12**: 235 (2003)
[16] S Temnykh *et al. Genome Res* **11**: 1441 (2001)
[17] G Benson *et al. Nucleic. Acid Res* **27**: 573 (1999) [PMID: 9862982]
[18] AT Castelo *et al. Bioinformatics* **18**:634 (2002) [PMID: 12016062]
[19] M Morgante *et al. Nat. Genet*, **30**: 194 (2002) [PMID: 11799393]
[20] M La Rota *et al. BMC Genomics*, **6**: 23 (2005) [PMID: 15720707]
[21] CM Luciano *et al. Int J Plant Genom*. DOI:10.1155/2008/412696 (2008)
[22] W Li *et al. Bioinformatics* **22**: 1658 (2006) [PMID: 16731679]
[23] X Huang & A Madan. *Genome Res* **9**: 868 (1999) [PMID: 10508846]
[24] G Pertea *et al. Bioinformatics* **19**: 651 (2003) [PMID: 12651724]
[25] L Cardle *et al. Genetics* **156**: 847 (2000) [PMID: 11014830]
[26] S Temnykh *et al. Theor Appl Genet* **100**: 697 (2000)
[27] ECL Chin. *Genome* **39**: 866 (1996) [PMID: 8890517]
[28] S Senthilvel *et al. BMC Plant Biology* **8**: 119 (2008) [PMID: 19038016]
[29] YD Lu *et al. Chin Sci Bullet* **55** (18): 1889 (2008)
[30] S Siju *et al. Mol Biotechnol* **44**: 140 (2010)

Edited by P Kangueane
Citation: Joshi *et al.* Bioinformation 5(9): 378-381 (2011)

Supplementary material:

Table 1: Summary of EST-derived microsatellies from the EST database of *Catharanthus roseus* L.

Parameters	Values
Total number of ESTs	21730
Total sequence analyzed	9259222bp
Total number of SSR identified including Poly A/T	632
EST after vector and Poly A/T removal	8423482bp
Total gene sequences after assembly	5928 (5125782bp)
Total number of contigs	1227 (1818690bp)
Total number of singletons	4704 (3307092bp)
Total number of SSR loci located	502
Frequency of SSR loci in *Catharanthus* EST	1 per 10.210 kb

Table 2: Distribution of SSR motifs in *Catharanthus roseus*

Repeat motif type	Number	Frequency (%)	The most abundant motif	Frequency within their own repeat (%)
Mononucleotide	221	44.02	A/T	56.94
Dinucleotide	72	14.34	AG/CT	57.20
Trinucleotide	131	26.09	AAG/CTT	32.82
Tetranucleotide	27	5.37	AAAC/GTTTT	37.03
Pentanucleotide	21	4.18	AAAAC/GTTTT	42.85
Hexanucleotide	22	4.38	AAAGCC/GGCTTT	22.72
Compound	8	1.59	-	-
Total	502	100		

Table 3: Total number of detected SSR loci.

Motif	Number of loci	Motif	Number of loci
A/T	182	AAAAC/GTTTT	9
G/C	39	AAAGG/CCTTT	5
AT/TA	7	AATGG/CCATT	1
AC/GT	24	AAAAG/CTTTT	2
AG/CT	41	ACCGG/CCGGT	1
AAC/GTT	5	ACAGC/GCTGT	3
AAG/CTT	43	AAAGCC/GGCTTT	5
AAT/ATT	3	AACCCT/AGGGTT	2
ACC/GGT	3	AGCCTG/CAGGCT	3
ACG/CGT	14	AGCTCC/GGAGCT	2
AGC/CGT	19	AACAGG/CCTGTT	2
AGG/CCT	22	AAGATG/CATCTT	1
ATC/GAT	7	AAACCC/GGGTTT	1
CCG/CGG	19	AGCGGC/GCCGCT	1
AAAC/GTTT	10	AAACAT/ATGTTT	1
AAAG/CTTT	4	AAATCC/GGATTT	1
AAAT/ATTT	6	AAAAGG/CCTTTT	1
AGAT/ATCT	1	AGCCCC/GGGGCT	1
ATAG/TCTA	3	Compound SSR	8
ATCG/CGAT	3	TOTAL	502