Plant Omics Journal POJ

POJ 3(4):121-128 (2010) ISSN:1836-3644

In silico SSR and FDM analysis through EST sequences in *Ocimum basilicum*

Sanchita Gupta*, Rishi Shukla, Sudeep Roy, Naresh Sen, Ashok Sharma

Biotechnology Division, Central Institute of Medicinal & Aromatic Plants (Council of Scientific & Industrial Research), Post Office CIMAP, Lucknow- 226015, India

***Corresponding author: 0804sanchita@gmail.com**

Abstract

The simple sequence repeats (SSRs) were extracted from 23,260 expressed sequence tags (ESTs) of *Ocimum basilicum*, a plant of great medicinal value. The results revealed that 447 SSRs were identified, out of which number of di, tri, tetra, penta, and hexanucletide SSRs were 212, 215, 17, 1 and 2 respectively. The occurrence of SSRs in UTRs (Untranslated Regions) is slightly more than coding regions. The SSRs containing ESTs (SSR-ESTs) were functionally annotated with the help of BLASTx program at NCBI server. In this study, 21.44 % SSR-ESTs could be assigned a significant match of translated SSR-EST query with protein databases. 58 primer pairs associated with SSR-ESTs were designed. These primers will be used further as markers to study transferability to related species. The study of functional domain markers (FDM) can provide information of functional property of microsatellite markers and predicted protein domains. 398 SSR-FDM were designed for significant functional domains. These findings will help to analyze molecular markers that have functional importance and should also facilitate the analysis of genetic diversity in plants especially medicinal plants.

Keywords: *Ocimum basilicum,* Expressed Sequence Tags, Simple Sequence Repeats, Functional Domain Marker

Abbreviations: EST_Expressed Sequence Tag, SSR_Simple Sequence Repeat, FDM_Functional Domain Marker, UTR_ Untranslated Region, DNA_Deoxyribonucleic acid, SNP_Single nucleotide polymorphism

Introduction

Analysis of variation at DNA level is the key to modern genetics studies, which encompasses newer tools and methods like microsatellite analysis, single nucleotide polymorphism (SNP) studies and other DNA marker systems based on gross and specific DNA sequence variations. Due to their ability to reveal the unexplored enormous genetic variation in the genome, such DNA markers have become extremely important for the genetic analysis of crop plants. Among different classes of molecular markers, microsatellite or simple sequence repeat (SSR) markers are the most favored for a variety of applications in plant genetics and breeding because of their multi-allelic nature, reproducibility, codominant inheritance, high abundance and extensive genome coverage (Gupta and Varshney 2000). Studies to identify microsatellites have been undertaken in rice, barley, wheat, maize, soybean, tomato, grapevine, forest trees, sunflower and *Brassica* species. *In-silico* microsatellite marker studies have also been done in several medicinal and aromatic plants of commercial importance (Tripathi et al. 2008, 2009). Basil (*Ocimum basilicum* L.), a member of the Lamiaceae family, has been traditionally used as a medicinal plant in the treatment of headaches, coughs, diarrhea, constipation, warts, worms, and kidney malfunctions. Externally, basil can be used as an ointment for insect bites and its oil is applied directly to the skin to treat acne. Natural components from basil have long been used to flavor foods and dental and oral products. *Ocimum* contains monoterpene derivatives such as camphor, limonene, thymol, citral, geraniol, and linalool. Basic chromosome number in *Ocimum* species is x=12 (Carovic-Stanko et al. 2010), whereas *Ocimum basilicum* is reported to be tetraploid (Pushpangadan et al. 1975). Microsatellites developed from ESTs, popularly known as EST–SSRs or genic SSRs, represent functional molecular markers as a putative function for a majority of such markers can be deduced by database searches and other *in-silico* approaches. Furthermore, 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 (Scott et al. 2000; Kantety et al. 2002; Varshney et al. 2002). Due to the above advantages of SSR markers, and relatively easy accessibility of large EST resources, increasing numbers of SSR markers are now being identified and used for a variety of applications in a number of plant species like, grapes (Scott et al. 2000), sugarcane (Cordeiro et al. 2001), and cereals such as wheat, barley, rye, rice (Varshney et al. 2005). A simple sequence repeat-functional domain marker (SSR-FDM) relies on development of molecular markers for putative functional domains. The FDM represents the SSR-based molecular markers and predicted protein domain (Yu et al. 2010). For development of SSR markers for *Ocimum*, 23,260 ESTs available at NCBI (http://www.ncbi.nlm.nih.gov) were pooled and analyzed with the following objectives: (1) analysis of the frequency and distribution of SSRs in the expressed portion of the

Table 1. Reduction in redundancy

Total no. of ESTs	No. of ESTs forming contigs (%)	No. of contigs	No. of singletons (%)	No. of assembled sequences	Reduction in redundancy (%)
23260	20844 (89.6%)	3318	2416 (10.4%)	5734	75.35

Ocimum genome, (2) functional annotation of SSR containing ESTs (3) development of novel SSR markers for *Ocimum basilicum*.

Materials and methods

Retrieval of ESTs sequences

All the EST sequences of *Ocimum basilicum* were retrieved from dbEST database of NCBI. A total of 23,260 sequences were retrieved which were related to different plant tissues e.g. leaves, stem, root, etc. The downloaded sequences were obtained in FASTA format for sequence assembly and SSR analysis. A single text file was compiled containing all the 23,260 EST sequences.

EST sequence assembly

EST sequences were assembled using the contig assembly program CAP3 (Huang and Madan 1999) available at Clemson University. The basic CAP3 tool with default parameters was used for assembly of EST sequences. 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.

Identification of SSR motifs

To detect SSRs in the EST sequence data set, we used a modified version of a Perl script SSRIT program (Temnykh et al. 2001) available at CUGI new SSR server. We set the motif parameter in the script to define SSRs as mono- to hexanucleotide with minimum repeat number (mono-10, di-6, tri-5, tetra-5, penta-5 and hexa-5). The modified script (CUGISSR) takes a FASTA formatted sequence file as an input and produces an output file with sequence name, number of SSRs in the sequence, SSR type, SSR motif, repeat number, sequence coordinates for SSR and the length of the sequence. CUGISSR also analyzes the SSR data and produces another output file with the frequency of the occurrence of SSRs in the data set, the frequency of each of the possible SSR motif types and the range and the repeat numbers for each type of SSR. To examine the location of SSRs in the EST sequences in relation to the putative coding region, CUGISSR uses the FLIP program (http://www.bch.umontreal.ca/ogmp/manlinks/flip.txt), available through OGMP server (Organelle Genome Mega sequencing Project, University of Montreal) (http:// megasun.bch.umontreal.ca/aboutflip.html). The proportion of the UTR and coding region in the assembled ESTs was estimated from the average length of each region in the SSR-containing assembled ESTs. To calculate the average distance between SSRs, the total length of each region (calculated by multiplying the length of assembled ESTs by the proportion of each region) was divided by the number of SSRs in the region.

Functional domain analysis

The sequences containing SSRs were searched for functional domain markers (FDM) through InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/).

Primer designing

Primer sequence designing for SSR-EST sequences was performed through PRIMER3 software (http://frodo.wi.mit.edu/primer3/). The conditions for primer designing were set at default values.

SSR-EST similarity searches

Pairwise comparison of SSR-EST sequences against the GenBank non-redundant protein database and SwissProt database was performed through BLASTx program at NCBI server. The most significant matches (EXP <1e−6) for each sequence were recorded.

Functional annotation

SSR-ESTs sequences with significant matches to protein entries of SwissProt-UniprotKB database were functionally classified according to corresponding protein gene ontology (GO) terms. All the predicted proteins were classified according to following three GO descriptors: (i) biological process, (ii) molecular function and (iii) cellular localization.

Results

EST assembly and reduction in redundancy

ESTs are indispensable for gene discovery and for detecting sequence features, such as SSRs, in genes. ESTs, however, often represent partial and redundant cDNA sequences making it difficult to analyze them effectively for SSRs. To construct longer and less redundant sequences, the publicly available ESTs were assembled from CAP3 program. CAP3 is a commonly used program (Whitfield et al. 2002; Pertea et al. 2003) which identifies overlapping sequences and generates contigs with consensus sequences.

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. In pursuance of this objective we calculated the reduction in redundancy. (Table. 1)

The percentage of ESTs forming contigs was 89.6% indicating that the majority of the ESTs had overlapping sequences with other ESTs while only 10.4% sequences were unique and had no corresponding overlapping sequences. After assembly, a non redundant group of ESTs was assembled consisting of contigs and singletons which are hereafter referred to as assembled EST sequences. The reduction in redundancy was found to be 75.35% which means that the number of ESTs had been reduced by a sizeable proportion prior to the SSR analysis. Both these figures point to the excessive overlapping that exists in EST sequences belonging to the same genome.

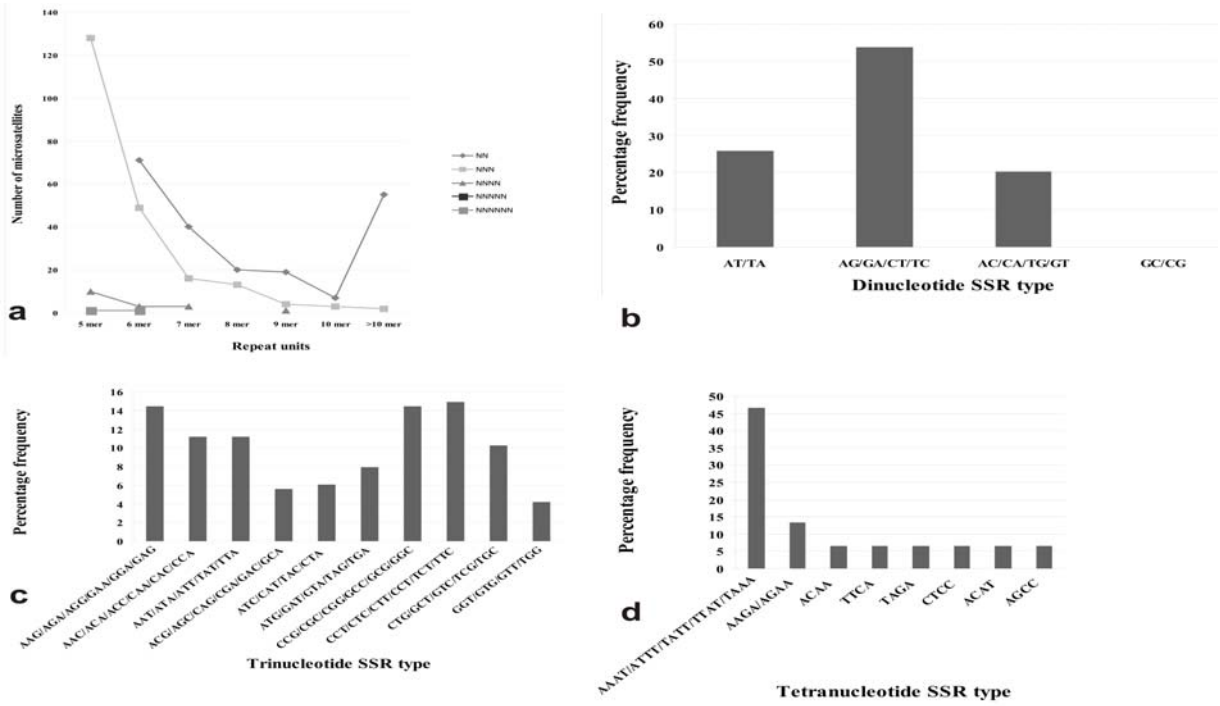

Fig 1. (a) Distribution of SSRs on the basis of repeat number. **(b)** Frequency distribution within dinucleotide SSRs. (**c)** Frequency distribution within trinucleotide SSRs. **(d)** Frequency distribution with tetranucleotide SSRs.

Table 2. Frequency of SSR-ESTs before and after assembly

	Before Assembly	After Assembly
No of sequences searched for SSRs	23260	5734
No of SSRs (%)	1378 (5.92%)	447 (7.79%)
No of unique sequences with SSRs	1303	402
No of motifs that are unique	99	85

Frequency of SSR-ESTs

SSRs in the ESTs were detected using the CUGISSR Perl Script. In this study SSRs are defined as 2-6 bp repeats equal or greater than 5 repeats. The frequency of SSRs was 7.79% in assembled sequences. (Table. 2) This suggests that *Ocimum* ESTs contain relatively high numbers of SSRs. The frequency of SSRs in EST sets was previously reported as 2.4% in *Arabidopsis*, 4.1% in Almond and Peach, and 4.8% in Rosa (Jung et al. 2004). The percentage of SSR in tissue specific ESTs in some medicinal plants responsible for secondary metabolite production are 4.5 % in *Papaver somniferum*, 10 % in *Phaseolus vulgaris*, 10.8 % in *Coptis japonica*, 12.9 % in *Catharanthus roseus* and 12.31 % in *Mentha piperita* (Tripathi et al. 2008). The result suggests that occurrence of SSRs in case of medicinal plants is relatively higher than other plants.

Frequency distribution of SSR-ESTs

Analysis of SSR-ESTs revealed that trinucleotide SSRs were the most common SSRs at 48.1%. They are closely followed by dinucleotide SSRs at 47.1%. These results are in agreement with previous findings that showed trinucleotide SSRs being the most abundant type in *Arabidopsis* ESTs (Cardle et al. 2000) and in exons of genomic DNA sequences in all eukaryotes studied (Toth et al. 2000**).** There is a sharp difference in the number of ESTs, between trinucleotide and tetranucleotide SSRs. Pentanucleotide and hexanucleotide SSRs are less than 1%. Table 3 illustrates the preference for SSR-ESTs in the form of dinucleotide and trinucleotide repeats.

Distribution of SSR repeats

Analysis of SSRs from the perspective of repeat number reveals that the frequency of occurrence of SSRs changes by the number of repeats (mer) for each type of SSRs from dinucleotide to hexanucleotide. In this analysis we have taken repeat numbers from 5 mer to 10 mer and a separate class of greater than 10 mer. It was observed that in case of trinucleotide SSRs, 5 mer was highest i.e. 128. Amongst the other type of repeats from 6 mer to more than 10 mer, frequency of dinucleotide was high. [Figure 1. a]

Frequency distribution within various SSR types

Among dinucleotide SSRs, AG/GA was the most frequent repeat at and CG/GC was the least frequent, similar to what had been previously observed in plant ESTs such as *Arabidopsis thaliana*, wheat, barley, rice, maize, almond, peach, rosa (Miyao et al. 1996; Cardle et al. 2000; Kantety et al. 2002; Jung et al. 2005) and also in *Catharanthus roseus*, *Coptis japonica*, *Papaver somniferum, Oryza satva, Phaseolus vulgaris, Capsicum annum* and *Mentha piperita* (Tripathi et. al. 2008) [Figure 1. b)]. Among trinucleotide SSRs, the most frequent motif was CCT/CTC/CTT/CCT /TCT/TTC at 14.88% and GGT/GTG/GTT/TGG was the least frequent at 4.18 % [Figure 1. c]. Among tetranucleotide SSRs, AAAT/ATTT/TAAA/TATT/TTAT motif was the most frequent at 46.67% [Figure 1. d]. AAGA/AGAA motif

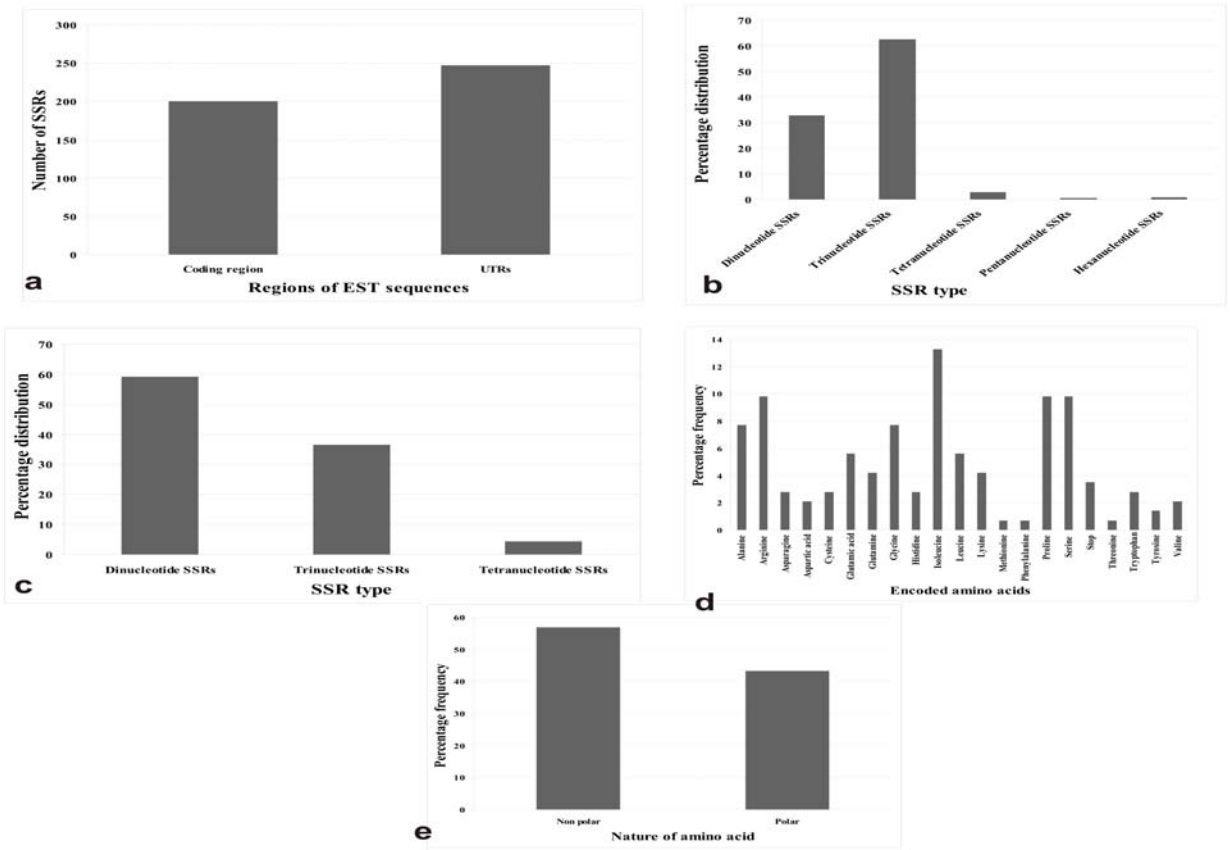

Fig 2. (a) Distribution of SSRs in coding regions and UTRs. **(b)** Percent distribution of SSRs in coding region. **(c)** Percent distribution of SSRs in UTRs. **(d)** Percentage distribution of encoded amino acids. **(e)** Frequency of trinucleotide SSRs encoding polar and non polar amino acids.

Table 3. Frequency distribution of various SSR-ESTs

	Before Assembly		After Assembly	
Motif Length	Frequency	Frequency (%)	Frequency	Frequency (%)
2 bp	543	16.31	212	47.48
3 bp	788	23.67	215	48.10
4 bp	37	1.11	17	3.8
5 bp	5	0.15	1	0.22
6 bp	5	0.15	2	0.45

was the second most frequent at 16.67% while AACA, ACAT, AGCC, CTCC, TAGA, TTCA motifs were the least frequent and shared equal frequency at 6.67%. In this study only one pentanucleotide SSR repeat was found i.e. CTGCC. Among hexanucleotide SSRs, two motifs ACCATG and CTTCTG were identified, both of which occurred once. Hence, both shared a frequency of 50%.

Distribution of various SSRs types in putative coding region and UTRs

Analysis of SSRs reveals a strong bias in the distribution of SSRs in the coding regions and UTRs. According to the analysis, frequency of SSRs was greater in UTRs (247) compared to Coding Region (200) [Figure 2. a]. The greater presence of SSR in UTRs points to their roles as binding sites for proteins and as regulatory elements. Further, distribution of SSRs in coding region reveals that trinucleotide SSRs are the most frequent (62.5%) while pentanucleotide SSRs are the least frequent (0.5%). Di, tetra and hexanucleotide SSRs have intermediate frequencies. (33%, 3% and 1% respectively) [Figure 2. b]. In the UTRs, contrary to the coding region, dinucleotides SSRs are the most frequent (59.11%). Penta and hexanucleotide SSRs do not occur in this region [Figure 2. c].

Distribution of trinucleotide SSRs and encoded amino acids

Each trinucleotide motif codes an amino acid which has putative roles in biological activity of protein molecules. Out of a total of 215 trinucleotides, 13.29% trinucleotides SSRs encoded Isoleucine, closely followed by trinucleotides SSRs encoding Arginine, Proline and Serine each has equal contribution (9.79 %) [Figure 2. d].

Nature of trinucleotide SSR encoded amino acids

Trinucleotide SSR encoded amino acids were classified on the basis of their polar and non polar nature. Non polar amino acids were more frequent (56.8 %) than polar amino acids

Table 4. Percentage distribution of various BLASTx results

BLASTx Results	No. of results
Significant matches	83 (21.44%)
Low sequence similarity	187 (48.3%)
Unknown/Hypothetical/Predicted Proteins/Gene ontology absent	75 (19.37%)
No Hits	42 (10.85%)

(43.2 %) [Figure 2. e]. It was observed that most of the amino acids are water insoluble i.e. hydrophobic.

Analysis of SSR-FDM

402 SSR containing sequences were analyzed for functional domain markers (FDM). Mono nucleotide SSRs containing sequences were not considered for this analysis. Through InterProScan 3924 functional domains were analyzed. The domains were analysed from interpro member databases such as pattern scan, SignalPHMM, TMHMM, HMMPanther, and FPrintScan. The functional domains were responsible for 2Fe-2S ferredoxin binding, iron-sulphur binding, 4Fe-4S ferredoxin binding and conserved site, and also function as von Willebrand factor type C, EGF-like region conserved site,alpha defensin, anaphylatoxin/fibulin, anaphylatoxin/ fibulin, Cystine knot, C-terminal, Insulin-like growth factor binding protein, N-terminal domain, Cys-rich conserved site, Integrin beta subunit, Alpha defensin, Agouti, Thiolase active site and Tubulin conserved site. Signal P domains searched through SignalPHMM database were unintegrated. As a result 398 SSR-FDMs were identified, the sequences containing both SSRs as well as FDMs. The SSR-FDM provide information regarding transcribed genetic markers having putative functions. (Appendix 1)

Analysis of BLASTx results

BLASTx was performed on SSR-ESTs to search proteins with significant match to translated SSR-EST nucleotide sequence. BLASTx was performed against non-redundant Genbank database. For this study, a significant match was defined as a sequence with E value $<=10^{-4}$ and identity>=70%. Out of 388 unique SSR-ESTs, 83 had significant match to proteins. Functional annotation of these SSR-ESTs was performed using Swissprot database. Also, specific primers were designed for such sequences. These primers amplify a gene of interest which produces a known protein product. The number of SSR-ESTs that produced no hit was 42 (10.85%). This indicates presence of sequences encoding proteins which are specific to *Ocimum* or proteins which are present in other plant/animal systems but are still not reported (Table 4).

Functional annotation of significant matches

For functional annotation, SSR-ESTs with significant matches were assigned gene ontology terms in Swissprot database. The result was tabulated with the gene ontology term and corresponding number of SSR-ESTs.

Biological process

A biological process is a series of events accomplished by one or more ordered assemblies of molecular functions. In a gamut of biological processes corresponding to SSR-ESTs, the most frequent was 'Response to stress' (18 SSR-ESTs) followed by 'Response to cadmium ion 'Oxidation reduction' and 'Regulation of transcription' (9 SSR-ESTs).

Molecular function

Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. In a gamut of molecular functions, the most frequent was 'ATP /GTP binding' (14 SSR-ESTs), followed by Transferase activity (13 SSR-ESTs) and DNA/RNA binding (10 SSR-ESTs).

Cellular component

A cellular component is a component of a cell, but with the provision that it is part of some larger object; this may be an anatomical structure or a gene product group. In a gamut of cellular components housing putative proteins, the most frequent was 'Plasma Membrane' (21 SSR-ESTs) followed by 'Chloroplast' (19 SSR-ESTs) and 'Nucleus' (15 SSR-ESTs). (Supplementary table 1).

Primer designing for SSR-ESTs with significant matches

Out of 83 SSR-ESTs with significant matches, primers were designed for 37 SSR-EST contigs and 21 SSR-EST singletons. Hence a total of 58 SSR primers were designed using Primer 3.0. (Appendix 2)

Discussion

Ocimum basilicum, commonly known as Tulsi, is a plant of great medicinal value. For genetic improvement of this genus only limited molecular markers are available. Simple sequence repeats are an important class of molecular markers for genomics and plant breeding applications due to their abundance, hyper variability, and suitability for high-throughput analysis, high polymorphism and transportability. Computational approaches have been used here to mine ever increasing EST sequences in public databases. The publicly available collection of 23,260 expressed sequences tags (ESTs) from *Ocimum basilicum* have been assembled and clustered using CAP3 assembly program. Assembly of EST sequences resulted in 5734 non-redundant EST sequences which were reported to have 447 EST-SSRs, distributed among 388 SSR containing ESTs (SSR-ESTs). Among all the SSR motifs the percentage frequency of trinucleotide SSRs is maximum (48.1%) and that of pentanucleotide SSR is minimum (0.2%). Out of 38 SSR-ESTs successful functional annotation of 83 SSR-ESTs was performed using Gene Ontology terms. These 83 SSR-ESTs were subjected to primer designing which yielded a total of 58 primer sets for *Ocimum basilicum.* The sequences having both SSRs and FDMs signifies that functional domains provide predicted functions to the molecular markers.

Acknowledgement

Financial grant from Department of Biotechnology, New Delhi under BTISnet programme is gratefully acknowledged.

References

Carovic-Stanko K, Liber Z, Besendorfer V, Javornik B, Bohanec B, Kolak I, Satovic Z (2010) Genetic relations among basil taxa *(Ocimum* L.) based on molecular markers, nuclear DNA content, and chromosome number. Plant Syst Evol 285:13-22

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156: 847–854

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to Erianthus and Sorghum. Plant Sci 160: 1115-1123

FLIP program - http://www.bch.umontreal.ca/ogmp/manlinks/flip.txt

Gupta PK, Varshney RK (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113:163-185

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9: 868-77

InterProScan - http://www.ebi.ac.uk/Tools/InterProScan

Jung S, Abbott A, Jesudurai C, Tomkins J, Main D (2005) Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. Funct Integr Genomics 5: 136–143

Jung S, Jesudurai C**,** Staton M**,** Du Z**,** Ficklin S**,** Cho I**,** Abbott A, Tomkins J, Main D (2004) GDR (Genome Database for Rosaceae): integrated web resources for rosaceae genomics and genetics research. BMC Bioinformatics 5: 130

Kantety RV, Rota ML, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Molecular Biology 48: 501–510

Miyao A, Zhong HS, Monna L, Yano M, Yamamoto K, Havukkala I, Minobe Y, Sasaki T (1996) Characterization and genetic mapping of simple sequence repeats in the rice genome. DNA Res 3: 233-238

NCBI- http://www.ncbi.nlm.nih.gov

OGMP server- http://megasun.bch.umontreal.ca/aboutflip.html

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651–652

Primer3 - http://frodo.wi.mit.edu/primer3

Pushpangadan P, Sobti SN, Khan R (1975) Karyomorphological studies in the genus Ocimum. I. Basilicum Group. Nucleus 18:177-182

Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100: 723–726

Temnykh S, Clerck GD, Lukashova A (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11: 1441–1452

Tripathi KP, Roy S, Khan F, Shasany AK, Sharma A, Khanuja SPS (2008) Identification of SSR-ESTs corresponding to alkaloid, phenylpropanoid and terpenoid biosynthesis in MAP's. Online J Bioinforma. 9: 78-91. doi:10.1093/bioinformatics/btn615

Tripathi KP, Roy S, Maheshwari N, Khan F, Meena A, Sharma A (2009) SSR polymorphism in *Artemisia annua*: Recognition of hotspots for dynamics mutation. Plant Omics J 2: 228-237

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10: 967-81

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cellular and Molecular Biology Letters 7: 537–546

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends in Biotech 23: 48-55

Whitfield W, Band R, Bonaldo F, Kumar G, Liu L, Pardinas R, Robertson M, Soares B, Robinson E (2002) Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. Genome Res 12: 555–566

Yu JK, Paik H, Choi JP, Han JH, Choe JK, Hur CG (2010) Functional Domain Marker (FDM): an *In silico* demonstration in solanaceae using simple sequence repeats (SSRs). Plant Molecular Biology Reporter 28: 352-356

Supplementary table1. Gene ontology classification of EST sequences containing SSRs

Gene Ontology	No. of SSR-ESTs
Cellular Component	
Plasma membrane	21
Chloroplast	19
Nucleus	15
Cytoplasm	10
Mitochondrion	10
Ribosome	6
Endoplasmic reticulum	6
Peroxisome	5
Vacuole	5
Cell Wall	4
Apoplast	4
Phragmoplast/Spindle	2
Myosin complex	1
Glyoxysome	1
Molecular Function	
ATP/GTP binding	14
Transferase activity	13
DNA/RNA binding	10
Protein binding	8
Zinc ion binding	7
Transcription factor activity	5
Copper ion binding	5
Synthase activity	5
Oxidoreductase activity	4
Iron ion binding	4
Heme binding	4
Hydrolase activity	3
Structural constituent of ribosome	3
Protein heterodimerization activity	3
Signal transduction	3
Dehydrogenase activity	3
Ligase activity	2
Magnesium ion binding	2
Monooxygenase activity	2
Thiamine-phosphate diphosphorylase activity	2
Catalase activity	2
Metal ion binding	2
Dehydratase activity	2
Lyase/Ligase activity	2
Ribulose phosphate-3 epimerase activity	1
Cadmium ion binding	1
Pyridoxal phosphate binding	1
Steroid binding	1
Cofactor binding	1
Phosphopathetheine binding	1
Cobalt ion binding	1
Potassium ion binding	1
Isomerase activity	1
FAD binding	1
Adenosyl homocysteinase activity	1
Transaldolase activity	1
Actin binding	1
Nucleoside-Triphosphatase activity	1
Endopeptidase activity	1
Biological Process	
Response to stress	18
Response to cadmium ion	9
Oxidation reduction homeostasis	9
Regulation of transcription	9
Plant parts development	8
Leaf morphogenesis	4
Cell differentiation	3
Gibberelic acid mediated signalling	3
Metabolic process	3
Methionine biosynthetic process	3
Translation	3
Auxin homeostasis	3
Cellular metal ion homeostasis	3
ATP/GTP metabolic process	3
Chloroplast organisation	2
Ethylene mediated signalling	2
Glutamine biosynthetic process	2
Lignin biosynthetic process	2
Lipid metabolic process	2
L-phenylalanine biosynthetic process	2
Pentose phosphate shunt	2
Terpenoid biosynthetic process	2
Thiamine biosynthetic process	2
Abscisic acid mediated signalling	1
Actin filament-based movement	1
Anti-apoptosis	1
Chlorophyll biosynthetic process	1
Chromosome segregation	1
Circadian rhythm	1
Cytokinin mediated signalling	1
DNA unwinding	1
ER overload response	1
Fatty acid biosynthetic process	1
Flavonoid biosynthetic process	1
Glycine metabolic process	1
Golgi organisation	1
Guard mother cell cytokinesis	1
Innate immune response	1
Isoprenoid biosynthetic process	1
L-phenylalanine catabolic process	1
L-seriene metabolic process	1
mRNA export form nucleus	1
Multidimensional cell growth	1
Negative regulation of cell growth	1
Negative regulation of DNA recombination	1
Nitrate assimilation/Nitrogen	1

fixation	
Oligosacchride biosynthetic process	1
One-carbon compound metabolic process	1
Phenylpropanoid metabolic process	1
Photorespiration	1
Photosynthesis	1
Protein folding	1
Protein import into chloroplast	1
Protein metabolic process	1
Proton transport	1
Regulation of stomatal movement	1
tRNA aminoacylation	1
Ubiquitin dependent protein catabolic process	1
Unidimensional cell growth	1
Vegetative to reproductive phase transition	1
Viral replication formation and maintainence	1
Wax biosynthetic process	1
Xylem//Phloem pattern formation	1