Plant Omics Journal
2(6):228-237 (2009)
ISSN: 1836-3644

Southern Cross Journals©2009
www.pomics.com

SSR polymorphism in *Artemisia annua*: Recognition of hotspots for dynamics mutation

Kumar Parijat Tripathi, Sudeep Roy, Neha Maheshwari, Feroz Khan*, Abha Meena, Ashok Sharma

Bioinformatics & In Silico Biology Division, Central Institute of Medicinal & Aromatic Plants,
(Council of Scientific & Industrial Research), Lucknow-226015 (UP) India

* *Corresponding author: f.khan@cimap.res.in*

Abstract

Simple sequence repeats or micro-satellites form an important class of molecular markers for genomic and plant breeding applications. In the studied work, publicly available data of 85,252 expressed sequence tagged sites of *Artemisa annua* (an anti-malarial plant) have been assembled and clustered into 20,588 non-redundant sequences by using EGassembler program. Results indicate 75.8% reduction of data redundancy and simple sequence repeats in only 20.74% (*i.e.,* 4272) EST sequences. We have also identified the frequency, density, composition and distribution of SSR containing ESTs. Among all the SSR motifs, maximum frequency was showed by trinucleotide SSR motifs *i.e*., 47.84% while minimum by pentanucleotide SSR motifs *i.e*., 5.2%. The ESTs containing the unique SSRs were functionally annotated with the help of SwissProt protein database using BLASTx program. After GO based functional classification, only 18.8% of the total significant matches revealed the putative function, which is of agronomic importance. Approximately 4000 SSR-ESTs were analysed for their polymorphism both in ORF and non-ORF regions showing their importance in regulating the function of important genes.

Keywords: *Aartemisia annua*; Contig assembly; Expressed Sequence Tags; Microsatellites; Polymorphism; Simple Sequence Repeats

Abbreviations: EST_Expressed sequence tagged sites; GO_Gene ontology; SSR_Simple sequence repeats

Introduction

Artemisia annua L. is a source of artemisinin, an antimalarial agent (Klayman DL, 1985). The relatively low content of artemisinin in cultivated types of *A. annua* has been a limiting factor for its commercialization. Enhanced production of artemisinin either in cell/tissue culture or in the whole plant of *A. annua* is therefore highly desirable. To study the genomic constitution of plants and applying genetic engineering methodology to develop new variety, concept of molecular markers came into existence (Bennetzen, 2000). SSR markers quickly became the markers of choice for plant and animal genomes during the last decade because of the small sample size (genomic DNA) requirement for their analysis and their suitability for automation and high-throughput analysis (Hearne et al., 1992). The presence of SSRs in the transcripts of genes suggests that they may have a role in gene expression or function (Kashi and King, 2006). While di-, tri- or tetra-nucleotide SSRs are most commonly used for the construction of linkage maps of nuclear genomes, single nucleotide repeats have been used in the population genetic analyses of chloroplast genomes.

SSRs can be assayed using PCR technology and can also be screened using high-throughput platforms for molecular genetic linkage and population studies (Karlin et al., 1998). The number of repeats at a locus can change by mutation and the rate of mutation depends on the number of tandem units within the repeat (King DG, 1994). Another area where SSR markers are extremely valuable and are increasingly becoming popular is comparative genomics where SSR markers developed from one species could be utilized in a related or heterologous species towards genetic mapping, characterization, gene cloning, diversity, evolutionary studies of genetic variation, linkage mapping, gene tagging, establishment of genetic maps, integration of physical and genetic maps, determination of evolutionary relationships and comparative genome analyses (Karlin et al.,1998;

Katti et al., 2001). SSRs, once developed are extremely valuable, although their development is time consuming, laborious and expensive. Sequences from many genomes are continuously made freely available in the public databases and mining of these sources using modern computational approaches permits rapid and economical marker development. Moreover, SSRs derived from ESTs essentially represent expressed gene sequences and hence are potential candidates for markers for comparative genomic studies. ESTs are particularly attractive for marker development since they represent coding regions of the genome and are also being developed at an extremely faster pace for many genomes. In the present work, SSRs were mined from ESTs of the medicinal plant *A. annua* through pipeline of different bioinformatics approaches. They were further analyzed for abundance and distribution of various types in these plant ESTs. The study demonstrates the potential of computational mining approach for rapid discovery of SSRs for the development of markers for genetic analysis and related applications in genetics and plant breeding.

Materials and methods

Retrieval of EST sequences

A total of 85,282 ESTs were retrieved from dbEST database available at NCBI. Entire cDNA were utilized where full length cDNA sequences are available for mining purpose. Thus, all entries in the GenBank that belong to EST or cDNA categories were included in the training datasets.

Clustering and assembly of ESTs

All the 85,282 redundant ESTs of *A. annua* retrieved from NCBI were used to produce the non-redundant dataset and for clustering and assembly analysis. It was done through EGassembler webserver (Masoud-Nejad et al., 2006). An automated trimming and screening for various contaminants, low quality and low-complexity sequences was done. The masking of DNA sequences for repetitive elements including small RNA pseudo genes, LINEs, SINEs, LTR elements, vector sequences, organelle and other interspersed repeats was carried out. The server clustered and assembled the sequences into contigs and singletons using CAP3 (Huang and Madan, 1999) with the criterion of 80% overlap identity between one end of a read to another end. After the assembly, redundant data sets of 85,282 sequences were reduced to only 20,588 non-redundant sequences. The non-redundant dataset of contigs and singletons was further used for SSRs predictions.

Mining of SSRs from assembled ESTs

For SSR identification, 13,347 contig sequences and 7,241 singlet sequences were combined together so that to form non-redundant data set of 20,588 assembled sequences. SSRs were identified through CUGI's SSR webserver, for all possible combination of dinucleotide, trinucleotide, tetranucleotide and pentanucleotide repeats. The criteria used for analyzing minimum number of repeats were as follows: five for dinucleotide, four for trinucleotide, three for tetranucleotide and three for pentanucleotide.

Primers designing

Forward and reverse primers for predicted SSR-EST sequences were designed with the help of Primer-3 software. Mostly primers were designed for SSR-ESTs that have GC content between 40% and 60%; with at least 40 base pairs (bp) of sequence on either side of the SSR. From the total set of sequences analyzed, 4272 SSRs of different motif lengths were resulted.

Functional annotation of SSR-EST sequences

Gene ontology based functional annotation of SSR-ESTs was performed through BLASTx using Swiss-Prot protein database. BLAST best hit were retained meeting the following criteria: *E*-value < 1e-4, and similarity >80%. The most significant matches for the SSR-ESTs with unique SSR motif were considered. However, gene ontology descriptions were assigned to SSR-ESTs on the basis of Swiss-Prot protein sequence matches.

Results and Discussion

Redundancy in EST sequences

After ESTs sequence assembly, out of total 85,282 ESTs, 78,041 were successfully assembled into contigs. The remaining 7,241 showed no overlap with any ESTs, thus called as 'singletons'. After assembly the whole dataset was reduced to 20,588 sequences which showed 24.14% of data redundancy. Before assembly, out of total 85,282 sequences only 12,314 sequences showed presence of SSRs that comprises 14.43 % of the total sequences, while 496 SSR motifs were found unique. After assembly, dataset resulted in to 20,588 sequences (16.81%) with SSR repeats and 393 unique motifs. Total redundancy in SSR-ESTs and unique motifs were observed 28.11% and 79.23% respectively. Reduction in data redundancy in an assembled ESTs set of contigs and singletons was found 75.86%.

Table 1. Details of hotspots region for probable mutation in the form of trinucleotide SSRs present within contigs of assembled ESTs dataset of *A. annua*

Contigs	Length	Frequency	SSR Type	Motif	A.A coded	Repeats	Characterstics	Origin
Contig8665	1831	2	3	acc	T	4	small, turn like, polar	nonORF
Contig8665	1831	2	3	gcg	A	6	small, hydrophobic, aliphatic, turn like	nonORF
Contig261	1584	2	3	tgg	W	4	aromatic, hydrophobic	ORF
Contig261	1584	2	3	tgc	C	4	small, turn like, polar	ORF
Contig402	918	2	3	tta	L	6	aliphatic, hydrophobic	ORF
Contig402	918	2	3	tga	Nonsense	4	nonsense	ORF
Contig596	1832	2	3	tgt	C	5	small, turn like, polar	ORF
Contig596	1832	2	3	gtg	V	4	small, hydrophobic, aliphatic	ORF
Contig879	1157	2	3	tgg	W	4	aromatic, hydrophobic	ORF
Contig879	1157	2	3	gtg	V	4	small, hydrophobic, aliphatic	ORF
Contig1310	1507	2	3	aca	T	4	small, turn like, polar	ORF
Contig1310	1507	2	3	ctt	L	5	aliphatic, hydrophobic	ORF
Contig1366	1490	2	3	aac	N	4	polar, small, turn like	ORF
Contig1366	1490	2	3	taa	Nonsense	5	nonsense	ORF
Contig1920	769	2	3	aac	N	5	polar, hydrophilic, small, turn like	ORF
Contig1920	769	2	3	gat	D	5	small, turn like, polar, h, charged	ORF
Contig2783	1276	2	3	att	I	4	aliphatic, hydrophobic	ORF
Contig2783	1276	2	3	gag	E	6	turn like, hydrophilc, polar, charged	ORF
Contig5781	1592	2	3	tga	Nonsense	4	nonsense	ORF
Contig5781	1592	2	3	ggt	G	5	small, tiny, turn like, aliphatic	ORF
Contig6547	1242	2	3	tga	Nonsense	5	nonsense	ORF
Contig6547	1242	2	3	aac	N	4	polar, hydrophilic, small, turn like	ORF
Contig8774	1022	2	3	tca	S	4	small, turn like, polar, tiny	ORF
Contig8774	1022	2	3	gtg	V	4	small, hydrophobic, aliphatic	ORF
Contig10041	687	2	3	tgc	C	4	small, turn like, polar	ORF
Contig10041	687	2	3	ctg	L	5	aliphatic, hydrophobic	ORF
Contig10992	1559	2	3	gct	A	4	small,hydrophobic, aliphatic, turn like	ORF
Contig10992	1559	2	3	aga	L	5	polar, hydrophilic, charged, turn like	ORF
Contig11705	1396	2	3	ctt	L	4	aliphatic, hydrophobic	ORF
Contig11705	1396	2	3	gtg	V	7	small, hydrophobic, aliphatic	ORF
Contig11709	1604	2	3	acc	T	6	small, turn like, polar	ORF
Contig11709	1604	2	3	atc	I	4	aliphatic, hydrophobic	ORF
Contig12052	1474	2	3	tga	Nonsense	5	nonsense	ORF
Contig12052	1474	2	3	gat	D	4	small, turn like,hydrophilic, charged	ORF
Contig3005	1071	3	3	cat	H	4	turn like, hydrophilic,charged, aromatic	nonORF
Contig3005	1071	3	3	tgt	C	4	small, turn like, polar	nonORF
Contig6379	810	3	3	tat	Y	5	aromatic, hydrophobic	nonORF
Contig6379	810	3	3	tca	S	5	small, turn like, polar, tiny	nonORF
Contig7303	925	3	3	ggt	G	4	small, tiny, turn like, aliphatic	nonORF
Contig7303	925	3	3	gtg	V	5	small, hydrophobic, aliphatic	nonORF
Contig3373	1864	3	3	tca	S	4	small, turn like, polar, tiny	ORF
Contig3373	1864	3	3	att	I	6	aliphatic, hydrophobic	ORF
Contig3373	1864	3	3	aca	T	5	small, turn like, polar	ORF
Contig12298	884	3	3	gat	D	5	small, turn like,hydrophilic, charged	ORF
Contig12298	884	3	3	caa	Q	4	hydrophilic, polar, turn like	ORF
Contig1981	1498	4	3	att	I	5	aliphatic, hydrophobic	ORF
Contig1981	1498	4	3	aag	K	4	turn like, hydrophilic, polar, charged	ORF
Contig2219	1756	4	3	tgc	C	4	small, turn like, polar	ORF
Contig2219	1756	4	3	att	I	4	aliphatic, hydrophobic	ORF

Table1. Continued

Contig2219	1756	4	3	att	I	5	aliphatic, hydrophobic	ORF
Contig10648	1215	4	3	atg	M	4	hydrophobic	ORF
Contig10648	1215	4	3	taa	Nonsense	4	nonsense	ORF
Contig10648	1215	4	3	caa	Q	5	hydrophilic, polar, turn like	ORF
Contig10648	1215	4	3	gag	E	4	turn like, hydrophilic, polar, charged	ORF
Contig11943	2659	4	3	ctc	L	6	aliphatic, hydrophobic	ORF
Contig11943	2659	4	3	cag	Q	5	hydrophilic, polar, turn like	ORF
Contig570	618	4	3	tag	Nonsense	4	nonsense	nonORF
Contig570	618	4	3	ccg	P	5	small	nonORF
Contig570	618	4	3	agc	S	5	small, turn like, polar, tiny	nonORF
Contig3815	1428	4	3	gaa	E	4	turn like, hydrophilic, polar, charged	nonORF
Contig3815	1428	4	3	ttg	L	5	aliphatic, hydrophobic	nonORF
Contig7157	1367	4	3	tat	Y	4	aromatic, hydrophobic	nonORF
Contig7157	1367	4	3	att	I	4	aliphatic, hydrophobic	nonORF
Contig7157	1367	4	3	tca	S	6	small, turn like, polar, tiny	nonORF
Contig9955	1287	5	3	taa	Nonsense	4	nonsense	ORF
Contig9955	1287	5	3	tga	Nonsense	4	nonsense	ORF
Contig9955	1287	5	3	gat	D	5	small, turn like,hydrophilic, charged	ORF
Contig11481	1273	5	3	ctc	L	4	aliphatic, hydrophobic	ORF
Contig11481	1273	5	3	cca	P	4	small	ORF
Contig9955	1287	5	3	gtt	V	5	small, hydrophobic, aliphatic	nonORF
Contig9955	1287	5	3	tct	S	4	small, turn like, polar, tiny	nonORF
Contig7814	1436	8	3	ttc	F	4	hydrophobic	nonORF
Contig7814	1436	8	3	taa	Nonsense	4	nonsense	nonORF

Frequency of SSRs during sequence assembly

It was observed that before assembly the frequency percentage of dinucleotide motif was 20.12 % of the total SSRs, which showed increase upto 23.87% after assembly. Similarly, in the case of tetranucleotide SSRs, the frequency percentage ranged from 22.94 to 22.98%. This indicates that contigs assembly enhances the frequency of di- and tetra-nucleotide SSR motifs in the studied dataset. In case of tri and penta-nucleotide SSRs frequency, during the process of assembly drops down percentage was observed *i.e*., from 51.22 to 47.84% and 5.7 to 5.2% respectively. This indicates the possibility of higher redundancy in the ESTs comprising these motifs. Results suggest that after assembly trinucleotide SSR's showed maximum frequency *i.e*., 47.84%. Dinucleotide SSR comprises the second most common motifs present in *A. annua* with 23.87% of total studied dataset. It is very interesting to observe that before assembly, dinucleotide SSR motifs were found at third position with 20.12% of the total SSR's; the statistics changed after the assembly, as it increases to 23.87%.

Average distance between the SSRs

The distribution of SSRs in putative coding regions and UTRs were analyzed properly. Results showed that a significant proportion of SSRs was present in both coding as well as UTR region. The total length of SSR-ESTs present in whole data set was 4325827 Kb. In *A. annua* the average distance between SSRs was 4.8 Mb. The average distance between two dinucleotide, trinucleotide, tetranucleotide and pentanucleotide SSRs were 20 Mb, 10 Mb, 21 Mb and 92 Mb respectively.

Statistics of SSRs physical distances

No. of sequences with SSRs= 3024 (contigs) +1034 (singlets) = 4272
Total length of SSR-ESTs = 4325827 Kb
Average length of SSR-ESTs=4325827/4272 = 1012 Kb
Total number of sequences searched for SSRs = 20588
Average distance between two SSR (20588 × 1012)/4272 = 4877 Kb
Average distance between two di-SSR = (20588 × 1012)/1020 = 20426 Kb

Results showed minimum distance between adjacent trinucleotide SSRs *i.e*, 10.193 Kb, while maximum in two pentanucleotide SSR motifs *i.e*, 92.19 Kb. Trinucleotide SSRs showed maximum density in the studied SSR-EST dataset of *A. annua*. These results are in accordance with the previous known findings on cereal crops (Varshney et al., 2002).

Frequency of single and multiple SSRs

Results also showed that frequency of ESTs with single SSR was higher than multiple SSR-ESTs. However, at genetic level the importance of multiple SSR-ESTs is higher since one or more SSRs might be variable in nature, and could be used for developing a genetic marker.

SSR-ESTs polymorphism

SSRs have a major role in genetic variation underlying adaptive evolution by virtue of their specific mutation and functional qualities (McCarthy and Hilfiker, 2000; Pfost et al., 2000). In the present work, prescribed role of SSR mutability in *A. annua* and its aftereffects on its genetic variation have been studied. The idea is to distinguish the whole SSR –ESTs dataset into two groups on the basis of their occurrence in ORF and non-ORF regions of the expressed genome. Distribution of each variant type in ORF and non-ORF frame was calculated. It was observed that percentage of trinucleotide SSR motifs were higher in ORF regions as compared to non-ORF in total datasets. The findings support the hypothesis of Hancock and Simon (2005), which state that repeats in triplet SSR can be accommodated more readily within coding regions. Further, change in their length simply result in gain or loss of single amino acid from a protein which must be associated in some way with protein function. Percentage of trinucleotide repeat in non-ORF region was less *i.e.*, 26 % as compared to ORF region *i.e.*, 65 % [Figure 1]. It was observed that in case of ORF SSR-ESTs, distribution of trinucleotide SSR was higher. On the other hand the percentage distribution of di, tetra and pentanucleotide SSR motifs was high in non-ORF datasets. This supports the theory that dinucleotide give rise to frameshifts within coding regions and are therefore strongly rejected during evolution. In non-ORF regions the higher percentage of di-, tetra- and penta- showed that these motifs will be controlled through natural evolution to maintain the conservation of functionality of genes and their products. The number of tandem repeats for these SSR motifs lies broadly in a range of three to nine times. It was observed that SSR motifs which were present in ORF regions showed length polymorphism range from three to nine tandem repeats. This provides good source of variation among length and amino acid sequence in ORF. This observation support the hypothesis that length polymorphism within the ORF region of genes could lead to variation in phenotypic characters. It will also lead change in the functionality of genes and results into dynamic mutation process (Hancock and Simon, 2005).

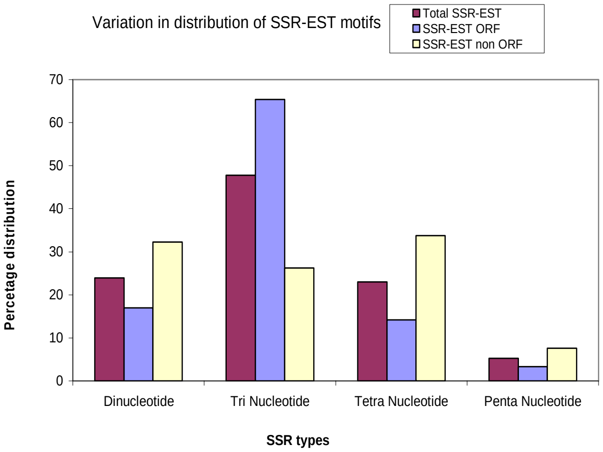

Fig 1. Variation in distribution of SSR motifs in ORF and non-ORF regions

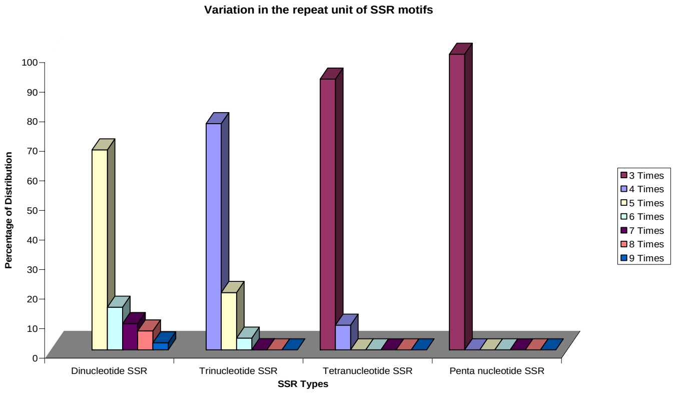

Fig 2. Variation in the repeats units of SSR motifs

In the present studied work, it was found that 67.7% of predicted SSRs were monomorphic in nature, while remaining 31.3% were polymorphic. Dinucleotide SSR motifs showed larger variation within the ORF region, generally it occurs with tandem repeats of five (66%), six (14.42%), seven (8.9%), eight (6.4%) and nine (2.4%) [Figure 2]. It is important to note that dinucleotide repeat sequences are preferential sites for recombination because of their high affinity for recombination enzymes (Biet et al., 1999). Some SSR sequences may influence recombination directly by their effects on DNA structures. It has been proposed that GT,CA,CT,GA,GC or AT repeat binding proteins could participate in recombination process by inducing Z conformation or other alternative secondary DNA structures (Korol et al.,1994; Karlin et al., 1998; Biet et al., 1999). Our results also support the above view. Trinucleotide SSR did not show a larger variation in distribution which was approx. 77% with tandem repeat four and 20% for tandem repeats five, others being negligible. Within ORF region pentanucleotide didn't show length polymorphism. Due to high abundance and

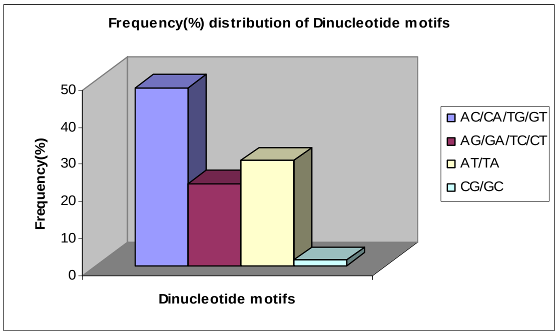

Fig 3. Percentage frequency distribution of Dinucleotide motifs

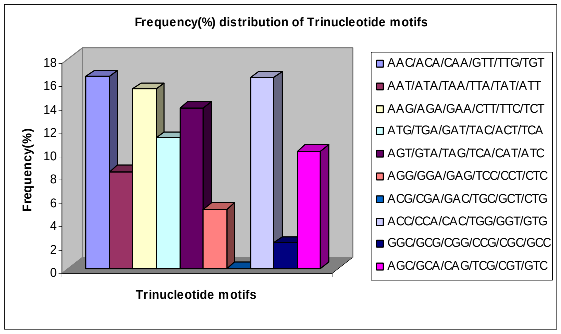

Fig 4. Percentage frequency distribution of Trinucleotide motifs

polymorphism, the predicted SSR markers could be beneficial in the development of genetic markers for the construction of linkage maps. Since the predicted SSRs were localized in upstream sequences of putative functional genes, they might be responsible for gene regulation. However the reviews published till date, have not clearly discussed the function and role of SSR polymorphism. Also the available information about SSRs location on chromosomes has been very limited. Thus polymorphism in SSRs could be useful to study evolutionary relationship and regulation of biochemical pathways. Results suggest that polymorphism in SSRs may have some functional role in the genetic evolution of *A. annua* variants (Cullis CA, 2002).

Frequency distribution of SSRs

Analysis of the dinucleotide SSRs showed that frequency distribution of AC, AT, TA, TG was maximum, while CG and GC was minimum (Figure 3). In case of trinucleotide repeats analysis motifs such as AAC/ACA/CAA/GTT/TGT/TTG showed the maximum frequency *i.e.*, 16.5% followed by ACC/CCA/CAC/TGG/GGT/GTG with 16.38% frequency (Figure 4). In tetranucleotide SSR motifs, the maximum frequency of 18.7% was showed by AAAC/AACA/ACAA/CAAA/ and TTTG/TTGT/TGTT/GTTT. The second most populous set was AAAT/AATA/ATAA/TAAA/TTTA /TTAT/TATT/ATTT with 16.08% frequency. On the other, hand three tetramer motifs showed very low level of occurrence with frequency from 0.1-0.3% (Figure 5). Similarly, in pentanucleotide motif set, maximum of 12.83% and minimum of 4.8% frequencies distribution was observed for one motif set each (Figure 6).

Variation in frequency of codon repeats in A. annua

Considering the trinucleotides SSR motifs in ORF showing variation in their tandem repeats, multiple occurrences within same ORF and change in their third base position give a special account of polymorphism within the ORF of contig datasets forming hotspots for dynamic mutation within the coding sequence of Artemisia genome. To identify these hotspots for mutation, 1542 trinucleotide SSR motifs have been taken, out of which 958 (62% of trinucleotide SSR) were present only ones within the ORF region of contiguous sequence. They do not show expansion of codon repeats with the change of third base of codon, thus lacking polymorphic characteristics. Motifs which showed multiple occurrences in ORF region of contigs showed expansion or increase in their tandem repeats with the change of third base. This occurs due to base change at third position of codon, thus the nature of its coding amino acid in protein sequence will also change, resulting into different stretch of protein's amino acid sequence. It can be hypothetically viewed as a precursor of genetic evolution in *A. annua* species resulting diversification in the genus Artemisia, as it is apparent that artemisinin (an anti-malarial bioactive compound) is only produced in species *A. annua*. The similar analysis was also carried out in non-ORF region considering 502 trinucleotide SSR motifs into account.

Some of these hotspots are also seen within the unique sequences (Table. 1 & 2). These SSR motifs could be utilized as molecular markers. The data showed that the hotspots for probable mutation lies both in ORF and non-ORF regions irrespective of their motif types. The findings support the theory that SSRs are randomly distributed in genomes and generally showed direct or indirect role in protein regulation (Morgante et al., 2002). It was observed that contig 8665 showed the variation in its sequence due to transformation of its trinucleotide SSR motif such as "ACC" to "GCG" with the change in their tandem repeat unit from four to six. It seems that stretch of SSR constituting "ACC" when repeated up to four times, translated into small polar amino acid *i.e.*, 'Threonine'. The motif when mutated to "GCG" and tandem repeated six times, thus allocated six hydrophobic amino acid *i.e.*, 'Alanine' within the non-ORF. This could lead

Table 2. Hotspots region for probable mutation in form of trinucleotide SSRs present within singletons of assembled ESTs dataset of *A. annua*

Singletons	Length	Frequency	SSR Type	Motif	AA coded	Repeats	Characterstics	Origin
gi|159702088	732	2	3	atg	M	4	hydrophobic	nonORF
gi|159702088	732	2	3	atc	I	4	aliphatic, hydrophobic	nonORF
gi|159668825	535	2	3	tga	Nonsense	5	nonsense	nonORF
gi|159668825	535	2	3	agt	S	4	small, turn like, polar, tiny	nonORF
gi|159667200	671	2	3	caa	Q	4	hydrophilic, polar, turn like	nonORF
gi|159667200	671	2	3	ttg	L	5	aliphatic, hydrophobic	nonORF
gi|159694358	714	2	3	atc	I	5	aliphatic, hydrophobic	nonORF
gi|159694358	714	2	3	tcc	S	4	small, turn like, polar, tiny	nonORF
gi|159707194	568	3	3	taa	Nonsense	5	Nonsense	ORF
gi|159707194	568	3	3	caa	Q	4	hydrophilic, polar, turn like	ORF
gi|159692582	709	3	3	cta	L	4	aliphatic, hydrophobic	ORF
gi|159692582	709	3	3	caa	Q	5	hydrophilic, polar, turn like	ORF
gi|159705256	756	4	3	gaa	E	4	turn like, hydrophilic, polar, charged	ORF
gi|159705256	756	4	3	tgg	W	5	aromatic, hydrophobic	ORF
gi|159669003	812	4	3	tca	S	4	small, turn like, polar, tiny	ORF
gi|159669003	812	4	3	atg	M	6	hydrophobic	ORF
gi|159630073	732	4	3	gtg	V	5	small, hydrophobic, aliphatic	ORF
gi|159630073	732	4	3	ggt	G	4	small, tiny, turn like, aliphatic	ORF
gi|159630073	732	4	3	gct	A	6	small, tiny, hydrophobic, aliphatic, turn like	ORF
gi|159621138	709	4	3	aac	N	4	polar, hydrophilic, small, turn like	nonORF
gi|159621138	709	4	3	cca	P	5	small	nonORF
gi|159621138	709	4	3	ggt	Q	4	hydrophilic, polar, turn like	nonORF

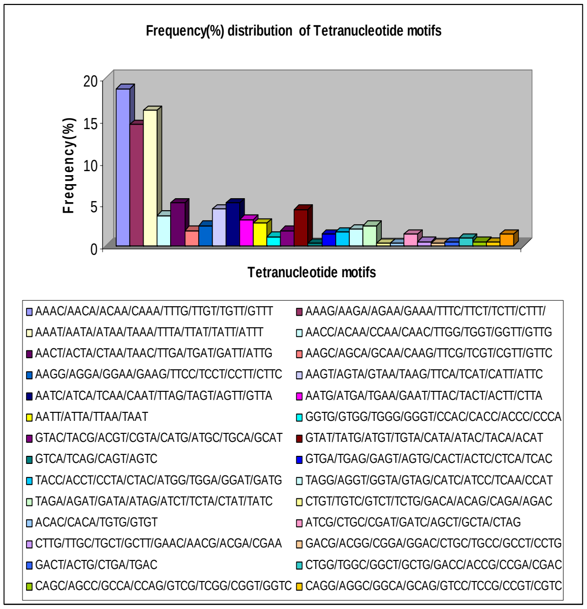

Fig 5. **Percentage frequency distribution of Tetranucleotide motifs.**

to some changes in regulatory regions resulting into phenotypic variations. Approximately, 45 instances were observed within assembled dataset of *A. annua* showing variation in their SSR motifs along with change in their tandem repeats.

To identify those genes and their products which are affected by mutagenic variation, BlastX program was carried out to obtain most probable hits for these highly mutation prone stretches of SSR within the contigs and singletons of assembled datasets. Contigs and singletons showed the presence of variation in their SSR motif both in their tandem repeats as well as amino acid. Most homologous entries in SwissProt protein database was found matched through BlastX for some of these hotspots carrying contigs and singletons. These findings are very important as expected values ranges from 0 to 4e-93, except in the case of contig 10992 where it matches homologous entries of coiled-coil domain containing protein with score 44 and E-value 0.002. Important proteins identified as close homolog were 2-cys peroxiredoxin BAS-1 like chloroplast precursor, serine/threonine protein kinase PBS1, chlorophyll a-b binding protein, methionyl t-RNAsynthetase, methyl tranferases, ketoacyl-CoA synthase, RNA binding proteins. All these proteins showcase the functional association with polymorphism in SSRs, hence regulating the metabolism in *A. annua*. Analysis related to the codon usage of trinucleotide SSR motifs indicates that large number of amino acids can be encoded through these motifs localized in the SSR-ESTs. The frequency of amino acid repeats encoded by the trinucleotide repeats was carried out successfully, which suggest that 'Serine' (10.9%) and 'Leucine' (9.7%) were the major amino acids. On the other hand 'Histidine' and 'Glutamine'

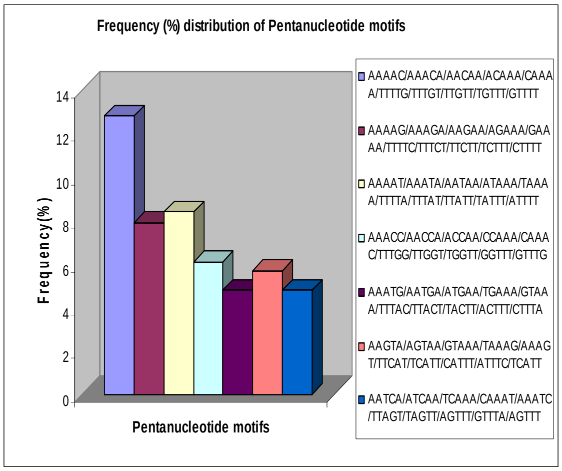

Fig 6. Percentage frequency distribution of Pentanucleotide motifs

showed distribution range of 7.2 and 5.72% respectively (Figure 4). Thus, we conclude that presence of trinucleotide motifs showing codons characteristics revealed the role of SSRs in codon usage within non-coding and coding genomic sequences (ESTs). This hypothesis is in accordance with the view of Katti et al. (2001) research work.

Functional annotation of SSR-ESTs and gene ontology based classification

SSR-ESTs of *A. annua* were gene ontologically assigned with their major molecular function, biological process and cellular component categories. Gene ontology for the corresponding SSR's was determined on the basis of sequence, domain and motif similarity. It was found that most of the SSR-ESTs belong to metabolism category of biological processes [Table 3]. This indicates an indirect role of above SSR-ESTs in development of genetic markers for the protein and enzymes involved in metabolite synthesis. The EST sequences were compared against known databases. Annotation of the SSR motifs present in *A. annua* was done by performing sequence similarity searches against the SwissProt protein database and assigned gene ontology descriptors based on best matched protein sequence features. Only 123 contigs and 36 singleton containing SSR motif showed similarity with SwissProt protein database constituting 20.4% and 13.8% of the total SSR-EST. A large number of proteins were reported which played enzymatic roles in a number of metabolic pathways related to secondary metabolite synthesis. Identified transcription factors such as HMR1 protein, HMG I/Y like protein, T1N15.25 associated with these SSR-ESTs seems, may have some role in gene regulation process. There were other proteins identified as regulatory gene such as chromatin-remodeling complex ATPase chain, Zinc finger protein 99, Troponin C, isoform 3, probable global transcription activator SNF2L1. These results are in accordance with previous findings of Hancock & Simmons (2005) and Elton et al. (2000). Putative chromatin-remodeling complex ATPase chain (ISW2-like) (sucrose non-fermenting protein 2 homolog) and SWI/SNF-related matrix-associated actin-dependent regulator (hSNF2H) showed response to osmotic stress and salt stress. Some of the important observations are as follows: mRNA cap guanine-N7 methyltransferase plays an important role in mRNA capping and mRNA processing, Berbamunine synthase forms the bisbenzylisoquinoline alkaloid berbamunine by phenol oxidation of N-methylcoclaurine without the incorporation of oxygen into the product during the process of berbamunine biosynthesis, Trans-cinnamate 4-monooxygenase enzymes were also

Table 3. Gene ontology based functional annotation and classification of SSR-ESTs of *A. annua*

S.No.	Gene Ontology	ESTs
	1. Biological process	
1.	Metabolism	18
2.	Stress	2
3.	Transport Regulation	4
4.	Cellular Respiration & ETC	3
5.	Regulation of Transcription	3
6.	Signaling	2
7.	Defence Response	2
8.	tRNA processing	2
9.	Biogenesis	1
10.	Unknown	2
	Total =	39
	2. Cellular Component	
1.	Cytoplasm	6
2.	Plasma Membrane	1
3.	Mitochondrion	7
4.	Chloroplast	4
5.	Synapse	1
6.	Endoplasmic reticulum membrane	3
7.	Cytoskeleton	2
8.	Sarcolemma	1
9.	Synapse	1
10.	Nucleus	2
11.	Nucleolus	1
12.	Golgi Apparatus	1
13.	Proteasome core complex	1
14.	Peroxisome	1
	Total =	32
	3. Molecular function (29)	
1.	Oxidase activity	1
2.	Oxidoreductase activity	7
3.	Kinase	6
4.	Synthase	1
5.	Ligase	1
6.	Synthetase	1
7.	Transferase	1
8.	Helicase	1
9.	Hydrolase	1
10.	Binding protein (ATP,GTP,CTP)	2
11.	Protein Binding	1
12.	DNA Binding	2
13.	Zinc ion Binding	1
14.	Binding protein (FAD,FMN,NADP)	3
15.	Rotational Mechanism	1
	Total =	30

identified which are associated with phenylpropanoid metabolism, Stilbene synthase 1 (resveratrol synthase-1) (RS1) are involved in phytoalexin and trihydroxystilbene biosynthesis, 40S ribosomal protein S19 and mitochondrial precursor were involved in RNA binding and formed a structural constituent of ribosomes. Beside this, SSR-ESTs corresponding to enzymes with catalytic role in chloroplast, plastid and cytosol were also identified. Some SSR-ESTs corresponding to cell metabolism showed their localization in endoplasmic reticulum, cytosol, plastid, trichome and chloroplast, thus indicating possible role in regulation of secondary metabolites biosynthesis.

It was also revealed that predicted SSR-ESTs possess some important role in following molecular functions *i.e*., regulation of transcription, transport regulation, signaling, defense response, stress and tRNA processing. Some of them have unknown putative function. This study demonstrate the utility of computational approaches for mining SSRs from ever increasing repertoire of publicly available medicinal plant EST sequences present in different databases. Computational approaches provide an attractive alternative way to conventional laboratory methods for rapid and economical development of SSR markers, by utilizing freely available genomic sequences in public databases. During study it was observed that only 4272 EST sequences had SSRs (20.74%), out of a total of 20,588 ESTs. Out of the total SSRs mined, 1020 (23.8%) were the dinucleotide, 2044 (47.8%) were the trinucleotide, 982 (22.9%) were tetra nucleotides and rest 226 (5.2%) were the pentanucleotide motifs. Moreover, out of 32 significant matches against SwissProt protein database only 18.8% were predicted to have some cellular function [Table 3]. Microsatellite distribution revealed the underlying mutational processes, evolutionary selection constraints as well as development in DNA repair mechanisms throughout the evolution. Knowledge of the occurrence and composition of SSRs across the number of species also helps a great deal in targeting specific SSRs for genetic marker development process. Using microsatellite markers will greatly benefit the genetic dissection of complex and quantitative traits in order to map genes and eventually clone and characterize the candidate genes controlling economically important traits. In order to effectively utilize the studied information and to extend the utility of SSRs in plant genomics, future work should be focused on both computational and molecular biology fronts.

References

Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 12:1021-9.

Biet E, Sun J, Dutriex M (1999) Conserved sequence preference in DNA binding among recombination protein: an effect of ssDNA secondary structure. Nucleic Acids Research 27:596-600.

Cullis CA (2002) Use of DNA polymorphisms in genetic mapping. Genet Eng (NY) 24:179-89.

Elton TY, James SS, Kristen VR (2000) Trinucleotides repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. Genetics 154:1053-1068.

Hancock JM, Simon M (2005) Simple sequence repeats in proteins and their significance for network evolution. Gene 345:113-118.

Hearne CM, Ghosh S, Todd JA (1992) Microsatellites for linkage analysis of genetic traits. Trends Genet 8:288-94.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877.

Karlin S, Campbell AM, Mrazek J (1998) Comparative DNA analysis across diverse genomes. Annual Rev Genet 32:185-225.

Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends in Genet 22:254-257.

King DG (1994) Triplet repeats DNA as highly mutable regulatory mechanism. Science 263:595-596.

Klayman DL (1985) Qinghaosu (artemisinin): an antimalarial drug from China. Science 228:1049-55.

Korol AB, Preygel IA, Preygel SI (1994) Recombination variability and evolution. Chapman and Hall Press, London.

Masoud-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S (2006) EGassembler: Online bioinformatics service for large scale processing, clustering, assembling ESTs and genomic DNA fragments. Nucleic Acids Res 34:459-462.

McCarthy JJ, Hilfiker R (2000) Use of single-nucleotide polymorphism maps in pharmacogenomics. Nat Biotechnol 18: 505-8.

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with non-repetetive DNA in plants Genome. Nat Genet 30:194 -200.

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution 18:1161-1167.

Pfost DR, Boyce-Jacino MT, Grant DM (2000) A SNPshot: pharmacogenetics and the future of drug therapy. Trends Biotechnol 18: 334-8.

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

Websites:

http://www.ncbi.nlm.nih.gov/dbEST (EST database at NCBI)

www.egassembler.hgc.jp (EGassembler webserver)

http://www.genome.clemson.edu/ (Clemson University Genomics Institute, USA)

http://www.genome.clemson.edu/cgi-bin/cugi_ssr (CUGI's SSR webserver)

http://frodo.wi.mit.edu/primer3/ (Primer3 designing web tool)