DATA MINING FOR SIMPLE SEQUENCE REPEATS IN OIL PALM EXPRESSED SEQUENCE TAGS

Aikkal Riju[1] and Vadivel Arunachalam [2*]

[1]Aikkal , Kanul, Kannur , Kerala - 670564, India

Email: riju.bioinfo@gmail.com

[2]Central Plantation Crops Research Institute, (Indian Council of Agricultural Research), Kudlu P.O, Kasaragod – 671124, Kerala, India.

*corresponding author: vadivelarunachalam@yahoo.com fax: +91- 4994- 232894, Phone +91-4994232894

Abstract:

Expressed Sequence Tags or ESTs are small pieces of DNA sequence that are generated by sequencing either one or both ends of an expressed gene. ESTs provide researchers with a quick and inexpensive route for discovering new genes, for obtaining data on gene expression and regulation, and for constructing genome maps. Oil palm EST sequences as available in public domain are downloaded. They were grouped and made contigs using CAP3 and Phrap. Microsatellite repeats are located using 5 softwares (MISA, TRA, TROLL, SSRIT, SSR primer). Among the 5 methods MISA is found to be the best. It can elucidate the compound repeat also. Frequency and total number (202) of SSR were detected. Mononucleotide repeat is more abundant especially 'A/T' repeats in Oil palm. Flanking primers were designed using primer3, SSR primers. The results of the study are given as an online database 'MEMCO' to help Oil palm researchers.

Keywords: *Elaeis guineensis,* Expressed sequence tag, *in silico,* Mining, Oil palm, Simple sequence repeat.

Abbreviations : MISA- microsatellite identification tool,TRA - Tandem repeats analyzer, TROLL- Tandem Repeat Occurrence Locator, SSRIT - Simple Sequence Repeat Identification Tool, MEMCO – Mining expressed sequence tag markers in cocoa and oil palm.

INTRODUCTION:

The oil palm is a tropical palm tree. There are two species of oil palm, the better known one is the one originating from Guinea, Africa and was first illustrated by Nicholaas Jacquin in 1763, hence its name, *Elaeis guineensis* Jacq .(Its generic name derives from Greek *elaia* = olive, on account of its fruits rich in oil. Its specific name refers to the species area of origin.) The Oil palm originated in West Africa but has since been planted successfully in tropical regions within 20 degrees of the equator. The cultivated oil palm tree belongs to the species *Elaeis guineensis* from tropical Africa. It is allogamous and propagated via seeds. A related species from South America, *E. oleifera*, is a promising potential source of valuable characters, including reduced growth habit and resistance to diseases such like But Rot. The genome size oil palm is 1.95 Bbp and it has 2n = 2x = 32 chromosomes. Oil is extracted from both the pulp of the fruit (palm oil, an edible oil) and the kernel (palm kernel oil, used mainly for soap manufacture). Both palm oil and palm kernel oil are high in olefins, a potentially valuable chemical group that can be processed into many non- food products as well. Oil palm is normally monoecious; in other words, it has both male and female flowers on the same tree. It produces thousands of fruits, in compact bunches whose weight varies between 10 and 40 kilograms. Oil palm is a representative of the Arecaceae family and the Arecales order, which is a phylogenetically key clade of monocotyledons. In this paper 2413 tissue specific EST sequences downloaded to find simple sequence repeat using five different SSR finding software. The Molecular marker development procedure from EST resources follows these steps

- Download EST resources.
- Make contigs by Phrap and Cap3.
- Detect SSR from EST as well as Contigs by MISA, TRA, TROLL, SSRIT and SSR

  Primer.
- Design Primer for SSR site by suing Primer3 tool.

Materials and Methods

The GenBank accession numbers BM402088, BM402089, CN599371 to CN601781 ESTs were downloaded from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) of NCBI. These 2413 (dbEST release 012006) tissue specific EST sequences were grouped into 6 tissues. The problem of high redundancy in EST databases is now well understood. A powerful way to manage this redundancy is to assemble clusters of ESTs representing the same message into longer virtual cDNA sequences. There are several advantages to working with assemblies rather than individual ESTs: first, there are fewer sequences to analyze; second, the assembled sequences are longer and potentially contain more interpretable coding sequence than their individual component ESTs; third, sequencing errors present in individual ESTs may be corrected during the assembly process; fourth, the virtual cDNA sequences may extend to the 5' end of the mRNA, greatly facilitating cloning of the gene in the laboratory. The Program Phrap (Green 1999) and CAP3 (Huang and Madan 1999) were used to make assemble EST sequences into Contigs. The next level of the project was to find the Simple sequence repeats in Oil palm EST as well as contigs. Five different tools available on public domain were used for this purposed. We have also tried to assess the comparative efficiency of detecting microsatellites in EST sequences viz., MISA (Thiel et al. 2003), TRA (Bilgen. et al. 2004), TROLL (Martins et al. 2006), SSRIT (Temnykh et al. 2001) and SSR Primer (Robinson et al. 2004). We targeted to detect SSRs of size of 20 bp and above. Primer designing programs like Primer3 and SSR primer were used to design left as well as right flanking sequence of the detected microsatellite.

RESULT & DISCUSSION

Retrived EST sequences were separated by tissue wise like Abnormal apex, Normal apex, Male inflorescence, Female inflorescence, Immature Zygotic Embrayo, Lambda Zap II. Majority of the ESTs belongs to abnormal apex. The contigs making program Phrap and Cap3 produced 187 and 167 contigs respectively (Graph 1). Microsatellites or Simple Sequence Repeats (SSR) are

stretches of DNA containing tandem repeats of di, tri, tetra and above nucleotide units ubiquitously distributed throughout the eukaryotic genome. They are found to be abundant in plant genomes and are thought to be the major sources of genetic variation in quantitative traits. SSR originate from the unequal crossing over or replication errors resulting in formation of unusual secondary structures such as hairpins or slipped strands (Pearson and Sinden, 1998). Molecular genetic markers can be used to examine a group of individuals or populations to estimate various diversity measures and genetic distances, infer genetic structure and clustering patterns, test for Hardy- Weinberg equilibrium and multilocus equilibrium, and to test polymorphic loci for evidence of selective neutrality. This can be useful to plant breeders, germplasm managers, or others who are interested in population genetic properties of materials that they are working with. The three most common types of markers used today are RFLP, RAPD and Micro satellites. A wide variety of methods for construction of libraries enriched for microsatelite sequences have been reported, the most popular among those being the ones based on vectorette PCR using anchored primers. But these method is highly time consuming and expensive, alternative is to use bioinformatics, *ie* computational tools to screen the public database and find SSR. EST derived molecular markers especially SSR and SNP are highly useful in developing linkage maps and markers assisted breeding programs. These markers are also transferable to related genera. Harnessing the synteny, EST resources are tissue specific and useful in studying the gene expression. Recently ESTs have been developed in oil palm by IRD, (Jouannic et al. 2005) and MPOB (http:// palmoils.mpob.gov.my/palmgenes.html). ESTs are also a useful resource for designing probes for DNA micro arrays used to determine gene expression. Molecular marker techniques are advantageous as they directly reflect variations in the DNA sequences and therefore of independence of environment. Among many molecular marker techniques currently available, microsatellites and SSRs (Powell et al.1996) provides an improved technology in assessing genetic diversity and genetic relationships in plants as they are highly polymorphic, codominants, very informative and PCR based. EST- SSRs offers the following

advantages over other genome DNA- based, markers: 1) they should detect variation in the expressed portion of the genome, so that gene tagging should give "perfect" marker – trait associations; (2) they can be developed at no cost from the EST databases; and (3) once developed, these markers, unlike genomic SSRs, may be used across a number of related species. With the growth of sequence databases, several authors have reported an abundance of simple sequence repeats in different genomes. The Distribution of SSRs in the rice genome has also been studied on the basis of the two whole genome draft sequences released, respectively, by Syngenta and by the Beijing Genome Institute (BGI). In the draft sequence released by Syngenta, for instance, 48,351 SSRs (including di-, tri- and tera- nucleotide repeats) were available, giving a density of 8kb per SSR in the whole genome; SSRs represented by di-, tri-, and tetra- nucleotide repeats accounted respectively for 24%, 59% and 17% of the total SSRs.

We have used 5 different softwares to find out mono, di, tri and above microsatellite repeats. The result of detected SSR from Oil palm EST by MISA program showed the mononucleotide repeat as more abundant SSR. Among those mononucleotide repeat 'A/T' occurred 84 (49.7%) times, by using the program TRA, the Mononucleotide repeat 'A/T' has occurred 84 (63.63%) times. The Program TROLL also showed the more abundant SSR as the Mononucleotide repeat 'A/T' with a frequency of 95 (71.96%). The program SSRIT showed more abundant as 'AG / TC ' with 38.88% of the detected SSR. Because this program is designed to find dinucleotide and above repeat. The SSR detecting program, SSR Primer showed more abundant repeat as trimeric repeat with 25 (56.81%) times in Oil Plam EST. The resulted SSR contains perfect and imperfect repeat also, we found this program is very helpful for finding SNPs in Molecular marker site and primer design. The result of SSR detected by all the 5 method showed 160 perfect and 42 imperfect or compound type repeats, among the perfect repeats the mononucleotide repeat 'A/T' occurred 98 (61.25%) times, SSRs represented by mono- ,di-, tri- ,tetra- ,hexa- ,nano- and above decamer accounted respectively

for 61.25%, 21.25% ,10%, 3.13% , 3.13%, 0.62% and 0.62% of the total (160) SSRs (graph 2).

The result of detected SSR from Oil palm contigs by MISA program showed the mononucleotide repeat as more abundant SSR. Among those mononucleotide repeat 'A/T' occurs 9(56.25%) times. With the program TRA, the mononucleotide repeat 'A/T' has occurred 6 (37.5%) times. The Program TROLL also proved the more abundant SSR as the Mononucleotide repeat 'A/T' with a frequency of 8 (61.53%). The program SSRIT showed more abundant SSR as dinucleotide repeat. Among those detected SSR by SSRIT, the dinucleotide repeat 'AG / TC' has occurred 2(33.33%) times. By using the SSR detecting program, SSR Primer showed trinucleotide repeat was more abundant. The result of SSR detected by all the 5 method showed the mononucleotide repeat 'A/T' occurred 9 (45%) times. Computational ability of different tools used to find Oil Plam Contigs SSR has listed (Graph3). Frequency of Oil Plam SSR is listed in Table 1. A total of 64 microsatellite among the 202 detected SSR has successfully designed the forward and reverse primers.

Among the 5 methods MISA program has given maximum coverage of SSRs in both oil palm ESTs and Contigs . However it has the advantage of detecting the mono to decamer repeats and also compound repeats. But has the disadvantage of in ability to detect above decanucleotide repeats. The program SSR primer shows perfect and imperfect repeats. We have included the imperfect repeat with length 20bp or more in our data because imperfect repeats have its own value in the genetic studies. The TRA software was able to detect microsatellite like mono, di and above decamer repeat also. Coverage is next to MISA. Such comparative analysis of computational tools available for SSR detection has not been made so far. So we found TRA software to be extremely useful to molecular biologist interested in locating Nanonucleotide and above sized repeats. However those were interested in locating maximum repeats but less than penatamer size could use MISA software. We found 202 SSRs with length 20bp or more in oil palm EST libraries (dbEST release 012006) and SSR frequency was 1 / 4.44kbp (total length of the bp was 898029bp). In our study

we found more abundant group of repeats as mononucleotide repeats. It represents 61.25% of the detected Oil Palm EST SSRs. Mononucleotide repeats detected in our study belongs to 'A/T'. In the case of contigs also the mononucleotide repeat is more frequent than other types. Among that 'A/T' has represented 45% of the total SSRs. The results of the study are given as an online database 'MEMCO' to help Oil palm researchers (http://210.212.229.5/memco/about.htm). The result of this study has many practical implications for Oil palm breeders.

Database Availability

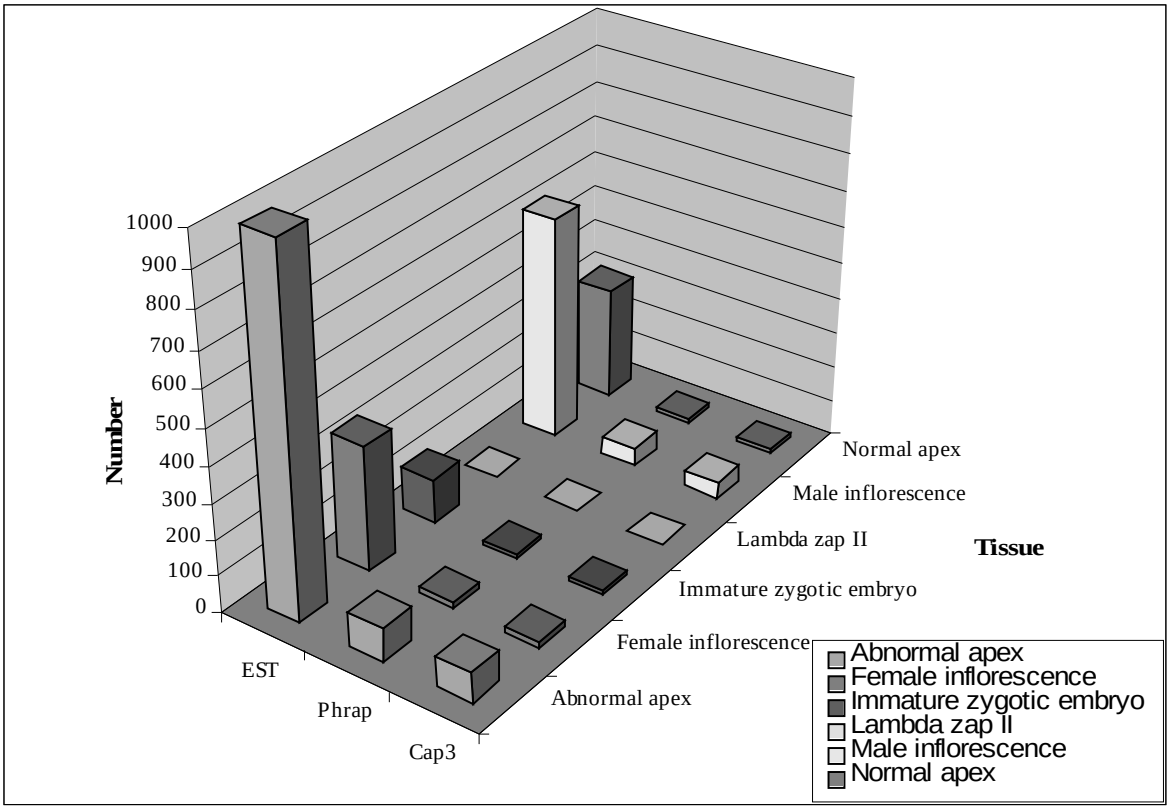
The results from the study have been compiled in the form of a database and are made available at **http://www.bioinfcpcri.org/databases** . The database contains the result of computational analysis of EST resources of Oil palm for Simple Sequence Repeats and Single Nucleotide Polymorphisms. The user can select any SSR finding method and view the detailed results. The URL of the site is http://210.212.229.5/memco/about.htm **.**

ACKNOWLEGEMENTS:

This work was supported by a grant from Department of biotechnology (BTISnet), New Delhi, India.

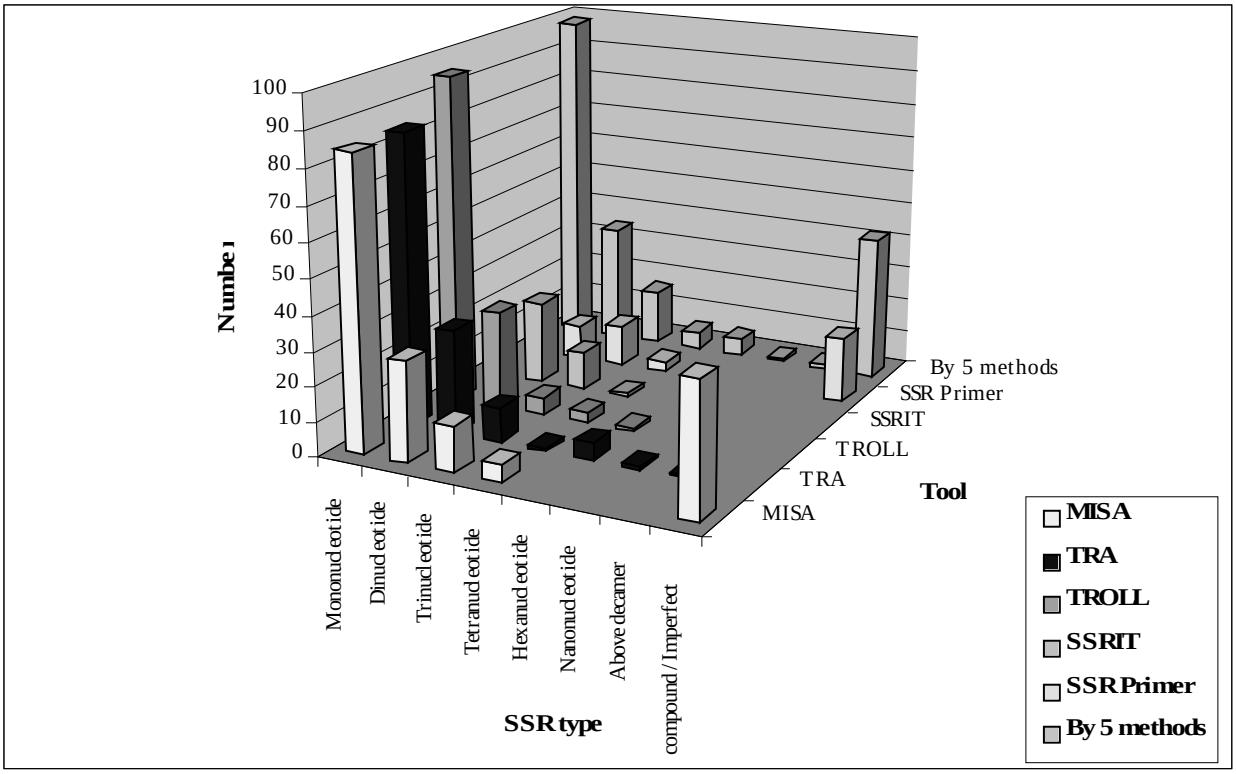
TABLES & GRAPHS

Graph 1. List of Oil palm tissue representing ESTs and Contigs



Graph 2. Computational analysis of Oil Palm EST - SSR by Different softwares

Total number of sequences analyzed 2413.



Graph 3. Computational analysis of Oil Palm Contigs- SSR by different softwares.

Total number of sequences analyzed 187

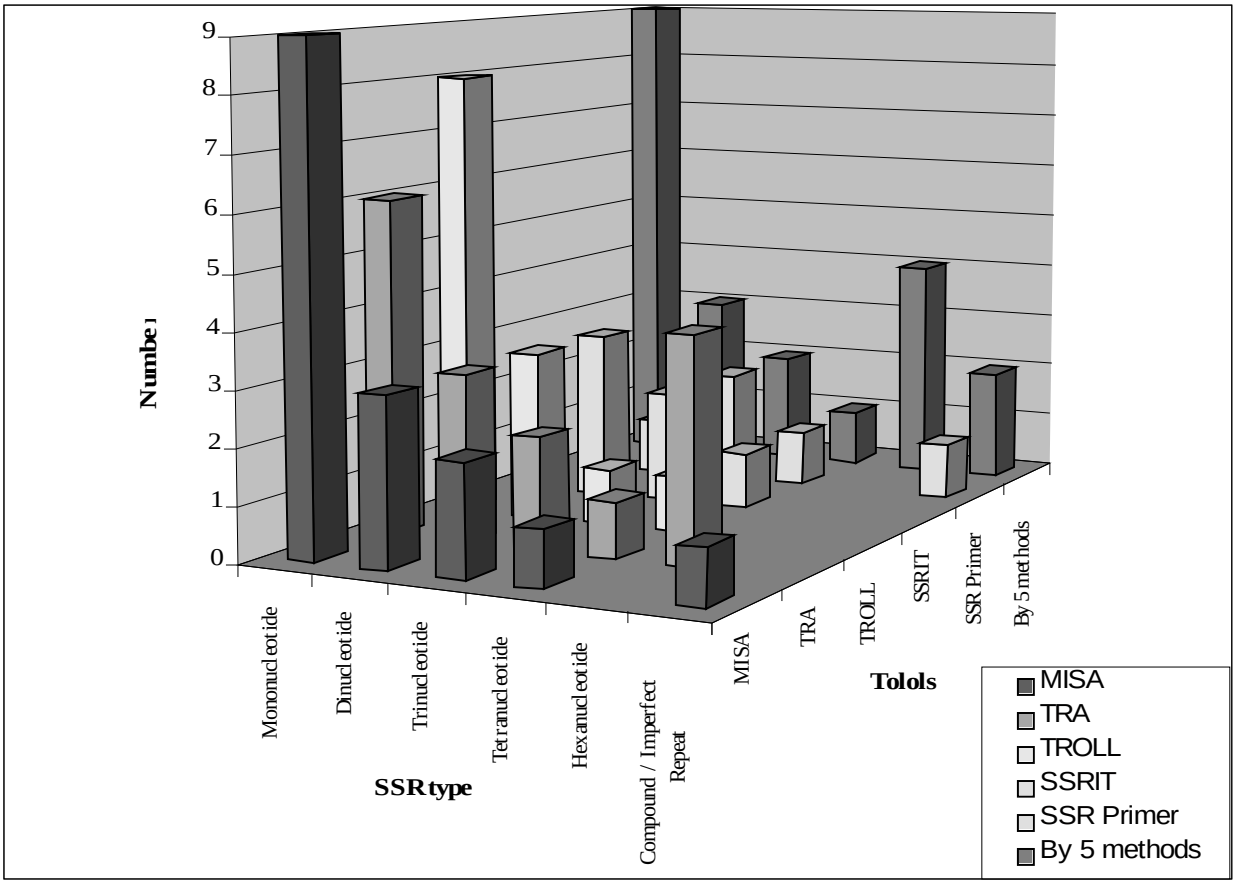
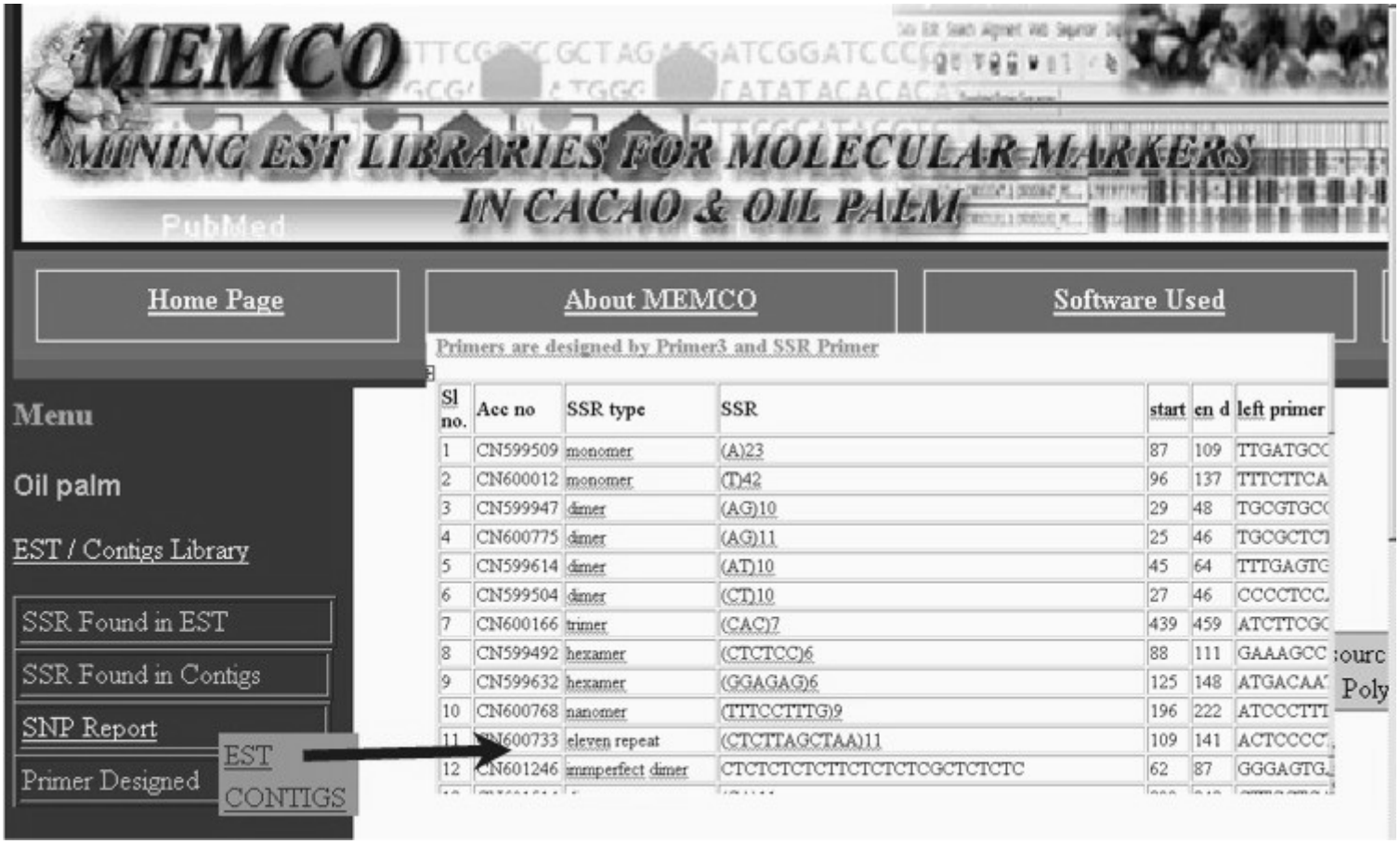


Table 1. Shows the frequency of SSR in Oil Plam EST

SSR	SSR type	Total
A / T	Mononucleotide	98
AG / TC	Dinucleotide	19
CT / GA	Dinucleotide	11
AT / TA	Dinucleotide	3
TG / AC	Dinucleotide	1
CCT / GGA	Trinculeotide	4
CAG / GTC	Trinculeotide	2
CAC / GTG	Trinculeotide	1
CCG / GGC	Trinculeotide	2
CGA / GCT	Trinculeotide	1
CTC / GAG	Trinculeotide	2
TTA / AAT	Trinculeotide	1
AGC / TCG	Trinculeotide	1
GTT / CAA	Trinculeotide	1
AGG / TCC	Trinculeotide	1
TATG / ATAC	Tetranucleotide	2
TTTA / AAAT	Tetranucleotide	1
GAGG / CTCC	Tetranucleotide	1
CTTC / GAAG	Tetranucleotide	1
CTCTCC / GAGAGG	Hexanucleotide	2
CAAGCC / GTTCGG	Hexanucleotide	1
TTTTTC / AAAAAG	Hexanucleotide	1
GGAGAG / CCTCTC	Hexanucleotide	1
TTTCCTTTG / AAAGGAAAC	Nanonucleotide	1
CTCTTAGCTAA / GAGAATCGATT	Above decamer (eleven repeat)	1

FIGURES

Figure 1. Showing the result of Oil Palm EST - SSR with Primer detection.



REFERENCE:

Bilgen M., Karaca M., Onus A. and Ince, A (2004). A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. *Bioinformatics*. 20 : 3379- 3386.

Green P. (1999) Phred, Phrap, Consed [Online] http://www.phrap.org/phredphrapconsed.html

Huang X. and Madan A (1999) CAP3: A DNA sequence assembly program. *Genome. Res.* 9**:** 868- 877.

Jouannic S, Argout X., Lechauve F, Fizames C, Borgel A, Morcillo F, Aberlenc-Bertossi F, Duval Y, and Treger J. (2005) Analysis of expressed sequences tags from oil palm (Elaeis guineensis) . *FEBS.* 579 : 2709 – 2714.

Martins W, De Sousa D, Proite K, Guimaraes P, Moretzsohn M. and Bertioli D.(2006). New softwares for automated microsatellite marker development. *Nucleic Acids Res*. 34 : e31.

Pearson, Christopher, Sinden E and Richard R (1998). Trinucleotide repeat DNA structure : Dynamic mutations from dynamic DNA. *Current opinion in Structural Biology*. 36 : 884- 889.

Powell W, Machray G.C.and Provan J. (1996) Polymorphism revealed by simple sequence repeats. *Trends in Plant Science* .1: 215- 222.

Robinson A.J, Love C.G, Batley J, Barker G. and Edwards D. (2004). Simple Sequence Repeat Marker Loci Discovery using SSR Primer. Bioinformatics . 20 :1475 – 1476.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour, S and McCouch S. (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.

Thiel T, Michalek V. and Graner, A. (2003). Exploiting EST data- bases for the development and characterization of gene- derived SSR- markers in barley (Hordeum vulgare L.). *Theoretical and .Applied Genetics*. 106 : 411–422.