IJBPAS, January, Special Issue 2, 2022, 11(1): 40-53

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

JAVA BASED PACKAGE FOR MINING SIMPLE SEQUENCE REPEATS IN EXPRESSED SEQUENCE TAGS

UMANG*[1], PK BHARTI[2] AND AKHTAR HUSAIN[3]

1: Research scholar, School of Computer Science, Shri Venkateshwara University, Gajraula, 244236, Uttar Pradesh, India

2: Professor, School of Computer Science, Shri Venkateshwara University, Gajraula, 244236, Uttar Pradesh, India

3: Associate Professor, Department of Computer Science& IT, MJP Ruhelkhand University, Bareilly, 243006, Uttar Pradesh, India

***Corresponding Author: Umang: E Mail: anilumang@yahoo.co.in**

Received 15th June 2021; Revised 10th July 2021; Accepted 24th Aug. 2021; Available online 25th Jan. 2022

https://doi.org/10.31032/ijbpas/2022/11.1.2005

ABSTRACT

Objectives: The next-generation sequencing techniques have enabled us to retrieve massive pieces of information from online biological databases, but it has been challenging to process multiple and unlimited size expressed sequence tags files for identifying and characterizing simple sequences repeats. Many web-based tools were designed but with time due to lack of server maintenance, they become unusable; also few available stand-alone tools lack processing adequateness. Therefore with intent to process multiple expressed sequence tag files without size limits, using proper validations and ability to retrieve more genome-related features; a simple to use, speed efficient, and portable standalone tool has been developed. The front end has been designed using swing components in Java Net Beans and the entire algorithm was implemented in Java using modular object oriented approach. Interactive microsatellite search algorithm blended with dictionary based approach algorithm MISA **–a** Perl script was called via command line for data mining from expressed sequence tag files. Another parallel module retrieves additional information from GenBank files. In the pipeline primer 3 was invoked for designing batch primers. This algorithm with extended interface in Java Net Beans provides naïve users with a simple interactive tool for mining microsatellites, statistical analysis, primer designing, and options of

BLAST programs on one platform in the form of the stand-alone application. The number of repeats/ interruptions and BLAST algorithm parameters can be reset through the graphical interface. This tool has interactive modules that provides proper validation, batch processing and cost-effective analysis of simple sequence repeats in expressed sequence tags as compared to peers and the source code can be upgraded in future as per requirements.

Keywords: BLAST, Data mining, EST sequence, Java pipeline, Microsatellites, Primer Design

1. INTRODUCTION

Expressed sequence tags (ESTs) are short sequences of cDNA and are generated by high-throughput single-pass sequencing of complementary DNA (cDNA) clones. More than 74.2 million ESTs from all species are available in public databases. The explosive growth of next-generation sequencing data has resulted in ultra-large-scale datasets. These datasets are a source of prime importance for gene discovery, gene transcripts, and gene-sequence determination; also, EST containing simple sequence repeats (SSRs) have become an attractive choice for the development of SSR markers. Simple sequence repeats (SSRs) or "microsatellites" [1] are short tandem repeats (motifs) of length 1–6 nucleotides [2] and are found in genomes of both prokaryotes and eukaryotes [3]. Microsatellites can be genomic, if developed from genomic DNAs (gSSRs), or can be expressed, referred to as EST-SSRs, if developed from an expressed sequence database [4, 5]. EST- SSRs have high power because of their associations with expressed genes, directly contributing to a phenotype [6], also beneficial, being more conserved used as functional markers [7]. Massive information can be churned from these databases by comparing ESTs from multiple species. They are likely to be more capable of cross-genera transferability and are used for associating conserved genomic regions amid species, genus, for comparative genomics studies and establishing evolutionary relationships [8-10]. EST-SSR markers are also useful in marker-assisted selection studies and facilitate in establishing genetic linkage maps [11-12]. The SSRs found in the coding region affect gene activation, resulting in the expression of protein and reflects lesser polymorphism in the coding part [13]. EST-SSR are directly associated with genes affecting a particular trait; therefore, they are considered as a better resource for their use in breeding improvements [14], and SSRs present in the non-coding region affect gene regulation [15]. In eukaryotes, EST-SSR markers are more profusely found in coding regions than in non-coding regions [2, 16]. EST sequences are likely to be conserved evolutionarily; therefore, the expected rate of success is likely to be more in cross-species polymerase chain reaction amplification of EST-SSRs than cross-species amplification of SSRs developed from genomic DNA [17]. The rapid discovery of known or unknown genes from ESTs may contribute to the understanding of complex adaptive mechanisms as sorghum EST were obtained to identify and analyze genes that could respond to biotic stress [18].

Considering the importance of EST analysis in studying genetic similarity/dissimilarity, various studies have been made to identify and characterize them *in vitro,* which is labor-intensive and time-consuming. Therefore with the advancement in next-generation sequencing technology and the easy availability of ESTs in public databases, many bioinformatics tools have been developed to analyze EST *in silico*. These bio computational tools aided the development of EST-SSR markers on a large scale in a cost-effective manner [2, 19]. In a survey from the literature, it was revealed that most of the studies made by researchers using expressed sequence tags in species all over made the use of MISA[20] based software for microsatellite analysis, Primer 3[21] to design primers and BLAST for homology search as observed in bioinformatics analysis of the ESTs from *Rhizophora stylosa* Griff. Genomic Library [22], Pomegranate [23], Cacao [24], *Euphorbia esula* [25], characterization of simple sequence repeats in watermelon genome[26] , Development of EST-SSRs in the genus Rubus [13] and many more. Species-specific databases using different technologies were developed as *coniferEST* [27]; a comprehensive annotated transcriptome data set in yam [28] and garlic; the GarlicESTdb [29] was developed to enrich the EST information in public databases. SpicEST, a comprehensive database, was developed for two spices plants ginger and turmeric [30].

These researchers made use of software that fulfills the particular objective; these tools were found using web-based technology, species-specific databases, and stand-alone software. All these tools have different features that cater to different needs as per study or objectives.

Limited web-based tools are available that are costly to deploy, maintain, and also have processing inadequateness as they deal only with limited range of genome sequences. Web-based EST analysis pipeline

ESTPiper streamlines typical large-scale EST analysis components [31]. EST mining tools like EST2uni [32], ParPEST [33] and ESAP plus [34] literature is available, but registration process for direct access fails, seems server maintenance problem. Bioinformatics Tool Kit for EST Analysis [35] was developed by using combination of five software that performs pre processing /cleaning ,clustering and can be called via command line on Linux platform is non interactive, lack batch processing and difficult for non programmers to use. None of the stand-alone software was found to provide microsatellite identification/analysis with more genome information, batch primer designing, and BLAST options on one platform.

Therefore considering the importance of *in silico* mining of ESTs and the privilege of server independency for desktop applications this standalone java package for mining expressed sequence tags has been designed with a rich interactive graphic user interface that bridges the gap between existing technologies for the benefit of learners and researchers by enabling them to save intensive labor and cost.

2. MATERIALS AND METHOD

2.1 Material

Publicly available Batch files of EST sequences were downloaded in FASTA format from GenBank at the National Center for Biotechnology Information- website (http://www.ncbi.nlm.nih.gov/).

2.2 Graphic User Interface (GUI)

An interactive, user-friendly graphical user interface is implemented using Java Net Beans IDE 8.0.2 swing components, Perl and Java. It displays the tutorials to use the software. It allows MISA parameters and interruption number to set and reset as per requirements and saves them in configuration file MISA.ini; the users are allowed to select multiple files for batch processing. The validation checks for EST files, name, path, and count in the repository are displayed after submission. Users can clear the text box and exit from the system as when required. Upon batch, submission users can simply click the Mine SSRs button to identify microsatellites and then press the Design Primers button or BLAST to choose BLAST options as per requirements. BLAST algorithm parameters can be changed by selecting them from the graphical interface.

2.3 The Algorithm with Regular Expressions

PSEUDOCODE

Module 1

```
//Input -
```

1. Reset repeat numbers
2. Reset interruptions /parameters in misa.ini via interface
3. Save.

Module 2

//Input -

1. upload EST files in FASTA format as a file stream;
2. open each file in read mode

check for">"FASTA FILE

if FASTA format

then Upload to fasta files directory /create directories if not present/ delete old files if present.

else

skip

3. Start process builder via CMD to call perl.exe misa.pl // a Perl script having dictionary based

// approach data mining algorithm to detect microsatellites

4. Execute each file serially on misa.pl

Close files

Create directories if not present/ delete old files if present.

//Output -

Mined text files saved to different directory and statistics text files saved to another directory

Module 3

//Input -

1. Upload Genbank file
2. Check for extension
3. Open file containing motifs
4. Match accession Id from file containing motifs of each motif from GenBank file
5. copy start position and end position of simple sequence repeats /motifs
6. find coding /non coding or overlapping region
7. Fetch flanking region of 200 nucleotides from upstream and downstream position for similarity search using BLAST//(200[motif]200)

   //output files with coding /non coding information of motifs and sequence information to generate primers

Module 4

1. install Primer3 software
2. With default parameters run output files with coding /non coding information of motifs and sequence information to generate left and right primers, left and right primer length, primer TM, primer GC %, product size

Module 5

Use BLAST to find common, unique and polymorphic simple sequence repeats by running files with flanking regions of motifs

2.4 Pipeline

The entire algorithm was implemented using Java programming language that performs a call to configuration file misa.ini, then misa.pl. Misa.pl (MISA, http://pgrc.ipk-gatersleben.de/misa/download/misa.pl) is a Perl script that was used for mining perfect and compound simple sequence repeats. Motif coding / non-coding region from corresponding GenBank files was mined using guava-18.0.jar. Primer3 software (http://primer3.org/releases.html) with default parameters was used to design primers. The flanking regions of 200 nucleotides were fetched in the pipeline from corresponding FASTA files to design batch primers for the identified microsatellites. BLAST GUI has been added for similarity search of flanking regions and provides Blast options like BlastN, BlastX, TBlastX, BlastP, and TBlastN. The entire workflow of the algorithm has been diagrammatically represented in **Figure 1**.

2.4 Transfer of Mined Data

Outputs are written in tab-delimited text files and saved in designated folders. These files can be imported to any relational database management system for downstream analysis **(Figure 2)**.

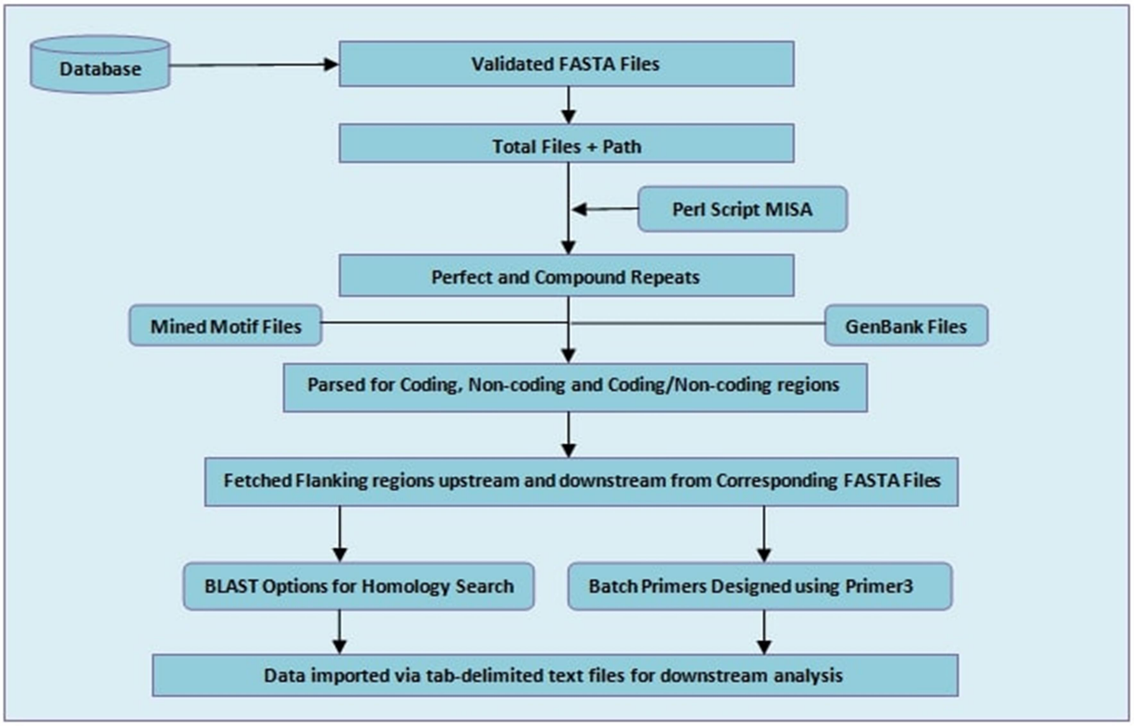
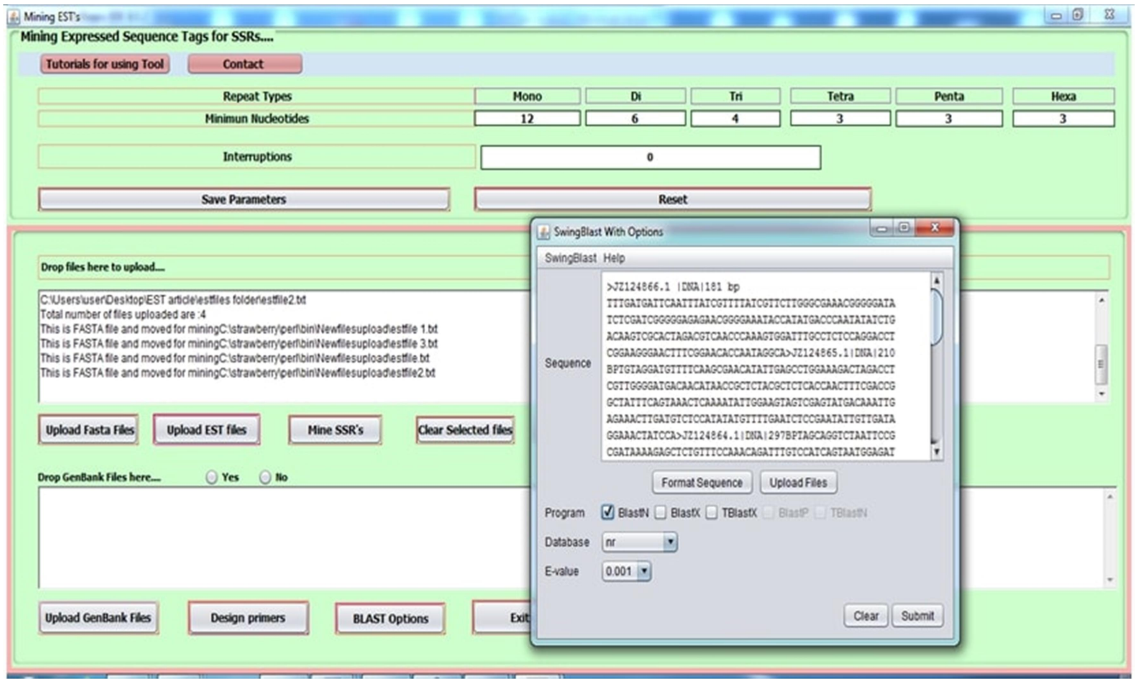


Figure 1: Diagrammatic representation of Java based tool



Repeat Types	Mono	Di	Tri	Tetra	Penta	Hexa
Minimun Nucleotides	12	6	4	3	3	3

Interruptions: 0

```
C:\Users\user\Desktop\EST article\estfiles folder\estfile2.txt
Total number of files uploaded are :4
This is FASTA file and moved for miningC:\strawberry\perl\bin\Newfilesupload\estfile 1.txt
This is FASTA file and moved for miningC:\strawberry\perl\bin\Newfilesupload\estfile 3.txt
This is FASTA file and moved for miningC:\strawberry\perl\bin\Newfilesupload\estfile.txt
This is FASTA file and moved for miningC:\strawberry\perl\bin\Newfilesupload\estfile2.txt
```

Sequence

```
>JZ124866.1 |DNA|181 bp
TTTGATGATTCAATTTATCGTTTTATCGTTCTTGGGCGAAACGGGGGATA
TCTCGATCGGGGGAGAGAACGGGGAAATACCATATGACCCAATATATCTG
ACAAGTCGCACTAGACGTCAACCCAAAGTGGATTTGCCTCTCCAGGACCT
CGGAAGGGAACTTTCGGAACACCAATAGGCA>JZ124865.1|DNA|210
BPTGTAGGATGTTTTCAAGCGAACATATTGAGCCTGGAAAGACTAGACCT
CGTTGGGGATGACAACATAACCGCTCTACGCTCTCACCAACTTTCGACCG
GCTATTTCAGTAAACTCAAAATATTGGAAGTAGTCGAGTATGACAAATTG
AGAAACTTGATGTCTCCATATATGTTTTGAATCTCCGAATATTGTTGATA
GGAAACTATCCA>JZ124864.1|DNA|297BPTAGCAGGTCTAATTCCG
CGATAAAAGAGCTCTGTTTCCAAACAGATTTGTCCATCAGTAATGGAGAT
```

Program ☑ BlastN ☐ BlastX ☐ TBlastX ☐ BlastP ☐ TBlastN

Database: nr

E-value: 0.001

Figure 2: Screenshot representation of Java tool showing ESTs files deposited to repositories; Identification, Primer Design and BLAST Options

3. RESULTS

This tool has been successfully tested and run on various size ESTs files for mining and analyzing perfect and compound microsatellites. Batch processing of EST format files results into information such as SSR type (perfect or compound), size, start position, end position, coding, and non-coding region, flanking sequences of 200 nucleotides for both upstream and downstream region of SSRs are provided with left and right primers, left and right primer length, primer TM, primer GC %, product size. These primers are essential for developing microsatellite-based markers. Details of null primers, i.e., microsatellites for which the primers are not designed due to insufficient flanking regions or poor melting temperature, are stored in separate files. Statistics details are displayed in separate files mentioning total number of sequences examined, the total size of examined sequences (base pair), the total number of identified SSRs, number of SSR containing sequences, number of sequences containing more than one motif, Number of SSRs present in the compound formation, distribution to different repeat type classes; having unit size with the corresponding number of SSR and frequency of identified SSR motifs. The data is automatically saved to tab-delimited text files in designated folders. BLAST options with different parameters can be used for a homology search.

4. DISCUSSION

This algorithm implemented via java package uses the dictionary based approach algorithm misa- a perl script and regular expressions for data mining in genome sequences; this tool has several unique features. The user can set, reset, and save the number of nucleotides and interruptions in the configuration file. Proper validations are applied to check valid file types and extensions. It checks the FASTA format sequences for processing. For batch processing, the number, names, and location of files send for mining are displayed **(Figure 2)**. Multiple files are processed with a single click, no need to attach or upload files again and again. The files after mining (simple sequence repeats) are saved to properly designated folders. ESTpiper [31], EST2uni [32], ParPest [33], PESTAS [36] pipelines are web-based, and we didn't found the link to access them, maybe due to server maintenance problems. However, these pipelines are using different technologies and are limited to cleaning, clustering, assembling, and BLAST comparisons and ESAP plus [34] seems challenging to access with a typical registration process. JaP-estmining tool has an advantage over them for being a standalone desktop application. We can also use other simple sequence mining standalone tools like GMATA [37]. But, it cannot process multiple files at one click, so each expressed sequence has to be copied to a different file for processing, thus creating a large number of records whereas JaP-estmining process single file containing as many sequences. No file size restrictions for processing the data as it was observed in MISA-web [38].The users can either detect microsatellites and related statistics or use corresponding GenBank files; they can identify coding, non-coding regions; with flanking sequence can go for primer designing set with default parameters. Alternatively, the user can BLAST flanking sequences for homology search with five BLAST options [39] as described. The users can either copy-paste sequence or upload preformatted ESTs files using the upload file button. The type of query sequence is identified, and then the appropriate BLAST option is enabled; this is an advantage over the Krait tool [40] that mines and designs multiple files. Output formats in tab-delimited text files are imported to any relational database management system for downstream analysis.

This java based tool varies in terms of file processing features and output and other utilities as shown in **Table 1**.

Table 1: Comparative features of EST processing tools

Program	Technology	Platform	Features	Remarks
ESAP Plus	PHP,HTML,CSS, Java Script, Apache HTTP server	Web based	Pre processing, clustering and assembly, SSR Mining and Primer design	Unable to register to access the tool online
PESTAS	JSP	Web based	Pre processing, clustering and assembling	Link Not available
ParPest	Perl, MySQL database using Red Hat Linux	Web based	EST cleaning, clustering, assembling and BLAST comparisons.	Link not available
ESTPiper	Perl, JavaScript, JAVA on Linux	Web based	Base calling, data cleaning, assembly, genome alignment, annotation, analysis of gene ontology.	Link not available
EST2uni	Apache HTTP Server, Perl scripting language, MySQL database management system, and PHP language on Linux platform	Web interface	Pre-processing, clustering, annotation, database creation, and data mining of EST collections	Link for code not available
GMATA	JavaScript, Perl script and R scripts	Standalone	No multiple files processing and BLAST options	Downloadable
Krait	Implemented in C and python	Standalone	Multiple files can be processed, SSR Mining, individual motif has to be clicked, no BLAST options	Downloadable and Performs very fast
MISA-Web	PHP and UNIX shell monitor server and Perl script	Web based	No flanking sequences, no coding and non-coding information.	none
Bioinformatics tool kit	Pherd, LUCY, RepeatsMasker, Cap3 tools ere integrated using java based pipeline.	Invoked via command line on Linux platform	Pre processing, clustering and assembly, EST Nucleotide database similarity searches	Downloadable
JaP-estmining	Java with misa.pl	Standalone tool	Multiple files can be processed; SSR Mining with coding and non-coding information and batch primer design with BLAST options and source code is modifiable.	Downloadable

5. CONCLUSION

The tool is developed and tested on AMD E-350 processor 1.60 GHz with 2.0 GB RAM and 32-bit operating system and performs much better with upgraded systems. Batch EST files have been processed to identify microsatellites, design primers, and use BLAST options within few seconds. These features all together are not available in peers. The tool is available on-demand and will be provided on open source platforms soon after adding more functionality.

6. CONFLICT OF INTEREST

The author declares no potential conflicts of interest.

7. REFERENCES

[1] Litt M, Luty JA. A hyper variable microsatellite revealed by invitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American journal of human genetics.*1989; 44(3), 397-401. PMID: 2563634; PMCID: PMC1715430.

[2] Toth G, Gaspari Z, Zurka, J. Microsatellited in different eukaryotic genome survey and analysis. Genome Research. 2000; 10 (7), 967-981.

[3] Field D, Wills C. Long, polymorphic microsatellites in simple organisms. *Proceedings of the Royal Society Biological Sciences*. 1996; 263, 209-215. https://doi.org/10.1098/rspb.1996.0033

[4] Vieira ML, Santini L, Diniz AL, Munhoz CF. Microsatellite markers: what they mean and why they are so useful. Genetics and Molecular Biology, 2016; 39(3), 312-328. http://dx.doi.org/10.1590/1678-4685-GMB-2016-0027

[5] Ellis J, Burke J. EST-SSRs as a resource for population genetic analyses. Heredity, 2007; 99, 125–132. https://doi.org/10.1038/sj.hdy.6801001

[6] Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. Trends in Biotechnology. 2005; 23(1), 48–55. DOI: 10.1016/j.tibtech.2004.11.005.

[7] Rohini Garg, Ravi K. Patel, Akhilesh K. Tyagi, Mukesh Jain, *De Novo* Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification. *DNA Research.*2011; 18(1), 53–63. https://doi.org/10.1093/dnares/dsq028

[8] Kantety RV, La Rota M, Matthews DE, Sorrells ME. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant* Molecular Biology. 2002; 48, 501-510. https://doi.org/10.1023/A:1014875206165.

[9] Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics*. 2003, 106(3), 411-422. DOI: 10.1007/s00122-002-1031-0.

[10] Feng SP, Li WG, Huang HS, Wang JY, Wu YT. Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (Hevea brasiliensis). *Molecular Breeding*. 2009, 23, 85-97.https://doi.org/10.1007/s11032-008-9216-0.

[11] Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK. Microsatellite markers: an overview of the recent progress in plants. *Euphytica*. 2010; 177(3), 309-334.https://doi.org/10.1007/s10681-010-0286-9.

[12] Sraphet S, Boonchanawiwat A, Thanyasiriwat T, Boonseng O, Tabata S, Sasamoto S, Shirasawa K, Isobe S, Lightfoot DA, Tangphatsornruang S, Triwitayakorn K. SSR and EST-SSR-based genetic linkage map of cassava (*Manihot esculenta* Crantz). *Theoretical and Applied Genetics*. 2011; 122(6), 1161-1170.**DOI**: 10.1007/s00122-010-1520-5.

[13] Hancock J.M. The contribution of slippage-like processes to genome evolution. *Journal of Molecular Evolution.* 1995; 41(6):1038-47. **DOI**: 10.1007/BF00173185.

[14] Bushakra JM, Lewers K S, Margaret E S, Tetyana Z,Christopher A S. Developing expressed sequence tag libraries and the discovery of simple sequence repeat markers for two species of raspberry (*Rubus* L.). *Biomed Central Plant Biology*. 2015; 15, 258. https://doi.org/10.1186/s12870-015-0629-8.

[15] Lawson MJ, L Zhang. Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. *Genome Biology*. 2006; 7(2), R14. **DOI**: 10.1186/gb-2006-7-2-r14. Epub 2006 Feb 21

[16] Andersen JR, Lubberstedt T. Functional markers in plants. *Trends in Plant Science*. 2003; 8(11), 554-60.**DOI**: 10.1016/j.tplants.2003.09.010.

[17] Arnold C, Rossetto M, McNally J, Henry RJ. The application of SSRs characterized for grape (*Vitis vinifera*) to conservation studies in Vitaceae. *American Journal of Botany*. 2002; 89(1), 22-8. **DOI**: 10.3732/ajb.89.1.22.

[18] Lee H. Pratt, Chun Liang, Manish Shah, Feng Sun, Haiming Wang, St. Patrick Reid, Alan R. Gingle, Andrew

H. Paterson, Rod Wing, Ralph Dean, Robert Klein, Henry T. Nguyen, Hong-mei Ma, Xin Zhao, Daryl T. Morishige, John E. Mullet, Marie-Michèle Cordonnier-Pratt. Sorghum Expressed Sequence Tags Identify Signature Genes for Drought, Pathogenesis, and *Skoto morphogenesis* from a Milestone Set of 16,801 Unique Transcripts. *Plant Physiology*. 2005; 139(2): 869–884. **DOI**: 10.1104/pp.105.066134.

[19] Karaoglu H, Lee CM, Meyer W. Survey of simple sequence repeats in completed fungal genomes. *Molecular Biology and Evolution.* 2005, 22(3),639-49. **DOI**: 10.1093/molbev/msi057

[20] MISA Perl Script .Available from: http://pgrc.ipk-gatersleben.de/misa/download/misa.pl. Date accessed: 01/2015.

[21] Primer3.Available from: http://primer3.org/releases.html . Date accessed: 09/2016.

[22] Ridha Wati, Mohammad Basyuni, Shigeyuki Baba, Hirosuke Oku. Bioinformatics Analysis of the Expressed Sequence Tags from *Rhizophora stylosa* Griff. Genomic Library. *AIP Conference Proceedings*. 2002; 020051. 2018. https://doi.org/10.1063/1.5050147

[23] Jian Z H, Liu XS, Hu JB, Chen YH, Feng JC. Mining microsatellite markers from public expressed sequence tag sequences for genetic diversity analysis in pomegranate. *Journal of Genetics*.2012; 91**,** 353–358. https://doi.org/10.1007/s12041-012-0185-z

[24] Riju A, Rajesh MK, Sherin PTPF, Chandrasekar A, Apshara SE, Arunachalam V. Mining of expressed sequence tag libraries of cacao for microsatellite markers using five computational tools. *Journal of Genetics*. 2009; 88, 217–225.**DOI**: 10.1007/s12041-009-0030-1.

[25] Surojit Sen, B Dehury, J Sahu, S Rathi, R N S Yadav. *In silico* mining and characterization of simple sequence repeats (SSRs) from *Euphorbia esula* expressed sequence tags (ESTs): A potential crop for biofuel. *Plants Omics Journal*. 2017; 10(02), 53-63. **DOI**: 10.21475/poj.10.02.17.pne370

[26] Zhu H, Song P, Koo DH, Guo L, Li Y, Sun S, Weng Y, Yang L. Genome wide characterization of simple sequence repeats in watermelon genome and their application in comparative mapping and genetic diversity analysis. *BMC Genomics*. 2016; 17, 557. https://doi.org/10.1186/s12864-016-2870-4

[27] Chun Liang, Gang Wang, Lin Liu, Guoli Ji, Lin Fang, Yuansheng Liu, Kikia Carter, Jason S Webb, Jeffrey FD Dean. ConiferEST: an integrated bioinformatics system for data reprocessing and mining of conifer expressed sequence tags (ESTs). *BMC Genomics*.2007; 8:134. **DOI**: 10.1186/1471-2164-8-134.

[28] Narina S, Ramesh B, Kameshwara RK, Alieu MS, Mohamed IA, Asiedu R, Mignouna JDH, Brian LS, Brian LS. Generation and analysis of expressed sequence tags (ESTs) for marker development in yam (Dioscorea alata L.). *BMC Genomics*. 2011; 12, 100. doi: 10.1186/1471-2164-12-100.

[29] Dae-Won Kim, Tae-Sung Jung, Seong-Hyeuk Nam, Hyuk-Ryul Kwon, Aeri Kim, Sung-Hwa Chae, Sang-Haeng Choi, Dong-Wook Kim, Ryong Nam Kim & Hong-Seog Park. GarlicESTdb: an online database and mining tool for garlic EST sequences. *BMC Plant Biology*. 2009; **9,** 61. https://doi.org/10.1186/1471-2229-9-61

[30] A Channdrasekar, Rijju A,Sathyanath NV, Santhosh EJ. SpicEST-An Annotated database on Expressed Sequence tags of spices. *Genes, Genomes and Genomics*. 2009; 3 (Special issue 1) 50-53 2009@Global science book.

[31] Zuojian Tang, Jeong-Hyeon Choi, Chris Hemmerich, Ankita Sarangi, John K Colbourne and Qunfeng Dong. ESTPiper – a web-based analysis pipeline for expressed sequence tags. *BMC Genomics*.2009; 10,174. **DOI**: 10.1186/1471-2164-10-174.

[32] Forment, J., Gilabert, F., Robles, A. *et al.* EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. *BMC Bioinformatics*. 2008; **9,** 5. https://doi.org/10.1186/1471-2105-9-5

[33] D'Agostino, N., Aversano, M. & Chiusano, M.L. ParPEST: a pipeline for EST data analysis based on parallel computing. *BMC Bioinformatics*. 2005; **6,** S9.

https://doi.org/10.1186/1471-2105-6-S4-S9

[34] Ponyared, P., Ponsawat, J., Tongsima, S., Seresangtakul, P., Akkasaeng, C., & Tantisuwichwong, N. *ESAP plus: a web-based server for EST-SSR marker development. BMC Genomics,* 2016; *17(S13).* **DOI**: 10.1186/s12864-016-3328-4.

[35] Premathilake, Hasitha & Tirimanne, Shamala. Bioinformatics Tool Kit for Expressed Sequence Tag Analysis. 2015; 10.13140/RG.2.1.3220.1045.

[36] Seong-Hyeuk Nam, Dae-Won Kim, Tae-Sung Jung, Young-Sang Choi, Dong-Wook Kim, Han-Suk Choi, Sang-Haeng Choi, Hong-Seog Park, PESTAS: a web server for EST analysis and sequence mining, *Bioinformatics*. 2009; (25)14, 1846–1848. https://doi.org/10.1093/bioinformatics/btp293

[37] Wang X, Wang L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. *Frontiers in Plant Science*. 2016; 7, 1350. **DOI**: 10.3389/fpls.2016.01350.

[38] Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. *Bioinformatics*. 2017; 33(16):2583-2585. **DOI**:10.1093/bioinformatics/btx198

[39] Bal, Harshawardhan, Johnny Hujol. “Java for Bioinformatics and Biomedical Aplications.” Springer Publications. 2006.

[40] Lianming Du, Chi Zhang, Qin Liu, Xiuyue Zhang, Bisong Yue. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics. 2018;* 34(4), 681–683. https://doi.org/10.1093/bioinformatics/btx665