BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Server

Volume 8(4)

ESMP: A high-throughput computational pipeline for mining SSR markers from ESTs

Ranjan Sarmah, Jagajjit Sahu, Budheswar Dehury, Kishore Sarma, Smita Sahoo, Mousumi Sahu, Madhumita Barooah, Priyabrata Sen and Mahendra Kumar Modi*

Agri-Bioinformatics Promotion Programme, Department of Agricultural Biotechnology, Assam Agricultural University, Jorhat-785013, Assam, India; Mahendra Kumar Modi - Email: mkmodi@gmail.com; Phone: +91-(376)-2340001 (O); *Corresponding author

Received February 01, 2012; Accepted February 11, 2012; Published February 28, 2012

Abstract:
With the advent of high-throughput sequencing technology, sequences from many genomes are being deposited to public databases at a brisk rate. Open access to large amount of expressed sequence tag (EST) data in the public databases has provided a powerful platform for simple sequence repeat (SSR) development in species where sequence information is not available. SSRs are markers of choice for their high reproducibility, abundant polymorphism and high inter-specific transferability. The mining of SSRs from ESTs requires different high-throughput computational tools that need to be executed individually which are computationally intensive and time consuming. To reduce the time lag and to streamline the cumbersome process of SSR mining from ESTs, we have developed a user-friendly, web-based EST-SSR pipeline "EST-SSR-MARKER PIPELINE (ESMP)". This pipeline integrates EST pre-processing, clustering, assembly and subsequently mining of SSRs from assembled EST sequences. The mining of SSRs from ESTs provides valuable information on the abundance of SSRs in ESTs and will facilitate the development of markers for genetic analysis and related applications such as marker-assisted breeding.

Availability: http://bioinfo.aau.ac.in/ESMP.

Keywords: Expressed Sequence Tag, Simple Sequence Repeats, ESMP, Single Nucleotide Polymorphism

Background:
Expressed sequence tags (ESTs) represents short, unedited and randomly selected single-pass reads derived from cDNA libraries provides an alternative to whole genome sequencing of organisms. The analysis of EST data enable gene discovery, complete genome annotation, gene structure identification, establish the viability of alternative transcripts, guide single nucleotide polymorphism (SNPs) characterization and facilitate in proteomic exploration **[1].**

The ubiquity of microsatellite or simple sequence repeats (SSRs) in eukaryotic genomes and their usefulness as genetic markers has been well established over the last decade. SSRs are short (1-6 bp) repeat DNA motifs that are usually single locus markers with characteristics of hypervariability, abundance, reproducibility and ease of detection by polymerase chain reaction with unique primer pairs that flank the repeat motif **[2].** The availability of ESTs greatly accelerates the systematic identification of SSRs and corresponding marker development based on computer analytical approaches **[3]**. EST-SSR and genomic SSR markers are considered as complementary to plant genome mapping, with EST-SSR being less polymorphic but concentrated in the gene-rich regions **[4].**

Several EST assembly and annotation pipelines *viz.* EST analysis pipelines (ESTAP) **[5]**, EST pipeline system **[6]**, ParPEST **[7]** etc.

are available with their own objectives, provides cleansing EST sequences and annotating them using public databases. The mining of SSRs from ESTs requires different high-throughput computational tools that need to be executed individually which are computationally intensive and time consuming. To reduce the time lag and to streamline the cumbersome process of SSR mining from ESTs, we have developed **EST-SSR Marker Pipeline: ESMP** for mining of putative SSRs from EST sequences. ESMP accomplish EST pre-processing, clustering, assembly and subsequently mining of SSRs from assembled EST sequences. Cross_match **[8]**, Trimest **[9]**, CAP3 **[10]** and MISA **[11]** analytical tools has been integrated into ESMP for their respective applications to perform the process of ESTs assembly and mining of putative SSRs.

ESMP has a three-tier architecture system. Presentation tier helps the user interact with ESMP through a web browser, whereas the business tier performs different analytical services associated with user specific options. The data generated in the business tier is then deposited into the data tier. For the use of this pipeline it does not require any database or any application installation on user machine. Instead the user simply uploads the fasta formatted EST sequence data into the server to run the pipeline with default parameters. It also has the options to choose the user defined parameters which makes the pipeline more interactive, user friendly and flexible.

Implementation:
ESMP interface has been developed using HTML, CSS, JavaScript and PHP. MySQL has been used to store input EST data, intermediate data of the pipeline and mined SSRs statistics. The database schema is available at ESMP website. The backend system is a Linux machine with Intel ® Core(TM) 2Duo@3.33GHz CPU and 3GB RAM. Architecture and workflow of the pipeline is depicted in **Figure1**.

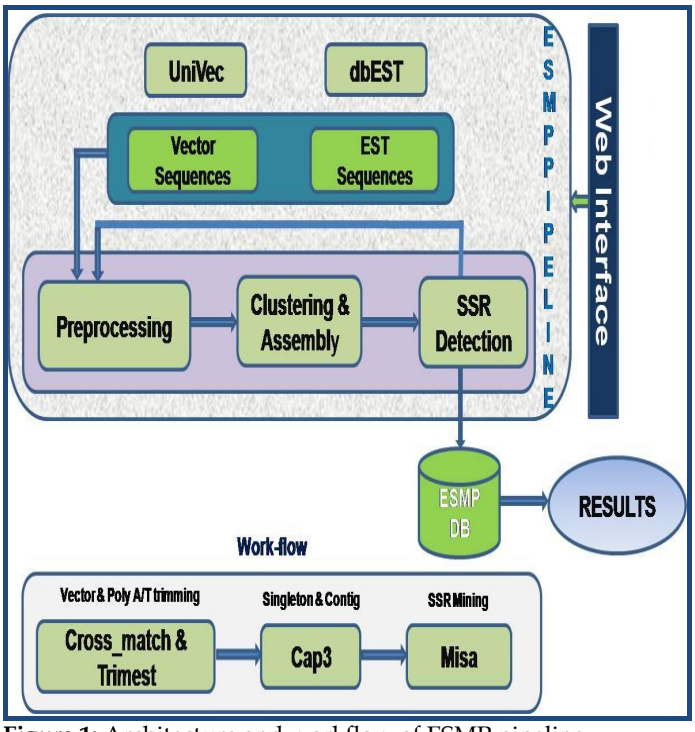

Figure 1: Architecture and workflow of ESMP pipeline.

Software input:
The ESMP web interface allows the user to submit EST sequences in the fasta format with ".reads" extension. It also asks the user to upload vector sequences in a plain text format which can be obtained from FTP site UniVec database (ftp://ftp.ncbi.nih.gov/pub/UniVec/) of NCBI. Although most EST projects produces a large number of chromatogram files, ESMP cannot accept chromatogram files due to file-size limitations of web-based uploading. Accordingly, chromatogram files have to be converted into DNA sequence files using a base-calling program such as phred **[8]**.

Software output:
The ESMP output is stored in a MySQL database. All the output files are stored in ".rar" extension which can be downloaded by the user as well as can be viewed in the current web page. The statistics files contains the statistics of putative SSRs *i.e.*, total number of sequence examined, total size of examined sequences (bp), total number of identified SSRs, number of SSRs containing sequences and number of sequence containing more than one sequences, number of SSR present in compound formation about the putative SSRs mined in the run. The statistics file can be transferred into an excel file for better visualisation of putative SSRs.

Conclusion:
ESMP pipeline is the integration of multiple tools which are individually used for their respective applications to accomplish the mining of putative SSRs from ESTs.

Caveat and future development:
ESMP currently supports pre-processing, assembly and putative SSR detection from EST datasets. This web-based EST-SSR pipeline reduces time lag and streamline the cumbersome process of SSR mining from ESTs, which is user-friendly. Our goal is not just to limit this pipeline for EST-SSR mining but to extend further for annotation and detection of suitable primer pairs which will flank the repeat motif. The mining of SSRs from ESTs provides valuable information on the abundance of SSRs in ESTs and will facilitate the development of markers for genetic analysis and related applications.

Acknowledgement:
The authors thankfully acknowledge the financial support for Agri-Bioinformatics Promotion Program by Bioinformatics Initiative Division, Department of Information Technology, Ministry of Communications & Information Technology, Government of India, New Delhi as well as to BTIS Net, DBT, and Government of India. Also the authors are grateful to Assam Agricultural University, Jorhat, Assam for providing the necessary facilities, constant support and encouragement throughout the study.

References:
[1] Nagaraj SH *et al. Brief Bioinform* 2007 **8**: 6 [PMID: 16772268].
[2] Thiel T *et al. Theor Appl Genet.* 2003 **3**: 411 [PMID: 12589540].
[3] Varshney RK *et al.* Cell Mol Biol Lett. 2002 **7:** 537 [PMID: 12378259]

[4] Varshney RK *et al*. *Trends Biotech*. 2006 **23**: 48 [PMID: 15629858]
[5] Mao C *et al*. *Bioinformatics*. 2003 **19**: 1720 [PMID: 15593407].
[6] Xu H *et al*. *Genomics Proteomics Bioinformatics*. 2003 **1**:236 [PMID: 15629036].
[7] D'Agostino N *et al*. BMC Bioinformatics. 2005 **4**: S9 [PMID: 16351758]
[8] Ewing B & Green P, *Genome Res*. 1998; **8**: 186 [PMID: 9521922].
[9] http://150.185.138.86/cgi-bin/emboss/trimest.
[10] Huang X & Madan A, *Genome Res*. 1999, **9**: 868. [PMID: 10508846].
[11] http://pgrc.ipk-gatersleben.de/misa/.

Edited by P Kangueane
Citation: Sarmah *et al*. Bioinformation 8(4): 206-208 (2012)