BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Views & challenges
Volume 8(4)

Towards an efficient computational mining approach to identify EST-SSR markers

Jagajjit Sahu[1, 2], Priyabrata Sen[1], Manabendra Dutta Choudhury[2], Madhumita Barooah[1], Mahendra Kumar Modi[1], Anupam Das Talukdar[2*]

[1]Department of Agricultural Biotechnology, AAU, Jorhat, Assam, India; [2]Department of Life Science and Bioinformatics, Assam University, Silchar, Assam, India; Anupam Das Talukdar - E-mail: anupam@bioinfoaus.ac.in; Phone: 03842- 270823; Fax: 03842-270823; *Corresponding author

Received February 09, 2012; Accepted February 11, 2012; Published February 28, 2012

Abstract:
Microsatellites are the markers of choice due to their high abundance reproducibility, degree of polymorphism and co-dominant nature. These are mainly used for studying the genetic variability in different species and Marker assisted selection. Expressed Sequence Tags (ESTs) serve as the main resource for Simple Sequence Repeats (SSRs). The computational approach for detecting SSRs and developing SSR markers from EST-SSRs is preferred over the conventional methods as it reduces time and cost to a great extent. The available EST sequence databases, various web interfaces and standalone tools provide the platform for an easy analysis of the EST sequences leading to the development of potential EST-SSR Markers. This paper is an overview of *in silico* approach to develop SSR Markers from the EST sequence using some of the most efficient tools that are available freely for academic purpose.

Keywords: Bioinformatics, EST-SSR, *in silico* approach, Microsatellites, SSR Markers

Description:
Microsatellites or simple sequence repeats (SSRs) are short (1-5 bp long) DNA elements which are repeated tandemly having extensive coverage over the entire genome. Compared to other neutral regions of DNA, Microsatellites show a high level of length polymorphism due to an increased rate of mutation of one or more repeats. The use of SSRs markers developed for one particular species is normally applicable across wide range of related species. ESTs are single-read sequences generated from partial random sequencing of cDNA libraries and particularly attractive for marker development as they represent the coding regions of the genome. Another importance of ESTs is that they are being deposited in the public databases at an extremely faster pace. Moreover, recent studies have observed that the frequency of microsatellites was significantly higher in ESTs than in genomic DNA. Moreover, SSRs derived from ESTs essentially represent expressed genic sequences and hence are potential candidates for markers for comparative genomic studies. EST-SSR markers are potential candidates for gene tagging and comparative studies in related species. Expressed Sequence Tags of many crop species has been generated and a number of SSRs have been detected using powerful bioinformatics tools leading to the development of EST-SSR Markers.

The *in silico* approach is the analysis of the available sequence data using various computational tools which helps in connecting more varied types of biological data to be integrated and stored. The traditional methods of developing SSR markers are usually time consuming and labor-intensive. SSRs can be rapidly and cheaply identified from ESTs and other genomic sequence data using computational methods. It takes very less time and low cost to design EST-SSR markers at a large scale using different computational facilities and large set of EST data available on the Web. Additionally, bioinformatics tools also supplement existing approaches by automating the task of SSR identification from available DNA sequences. The EST sequences can be mined using various bioinformatics tools to find the SSRs. The EST-SSR mining is done in three steps such as Preprocessing, Assembly and SSR detection. In the Preprocessing step, the EST sequences downloaded from the public databases like dbEST **[1]** are modified to find the high quality EST sequences. This step helps in reduction of the overall noise in EST data to improve the efficacy of subsequent analyses **[2].** The sequences are searched for the vector contaminated sequence and then removal of the contaminated sequences can be done using Cross_match tool. Cross_match is a program for rapid protein and nucleic acid sequence comparison and database search **[3].** UniVec is a database that stores the update vector sequences for all the species which can be obtained from the NCBI FTP directory **[4].** Using Trimest, another tool from EMBOSS, the poly A and poly T tails can be trimmed **[5].** Also discarding the low quality and very short ESTs improves the process. In general the Masking of low complexity regions is done for a better preprocessing of the EST sequences during EST analysis but in case of EST-SSR mining it is restricted as it may remove some of the repetitive portions in the sequence which contains valuable microsatellites.

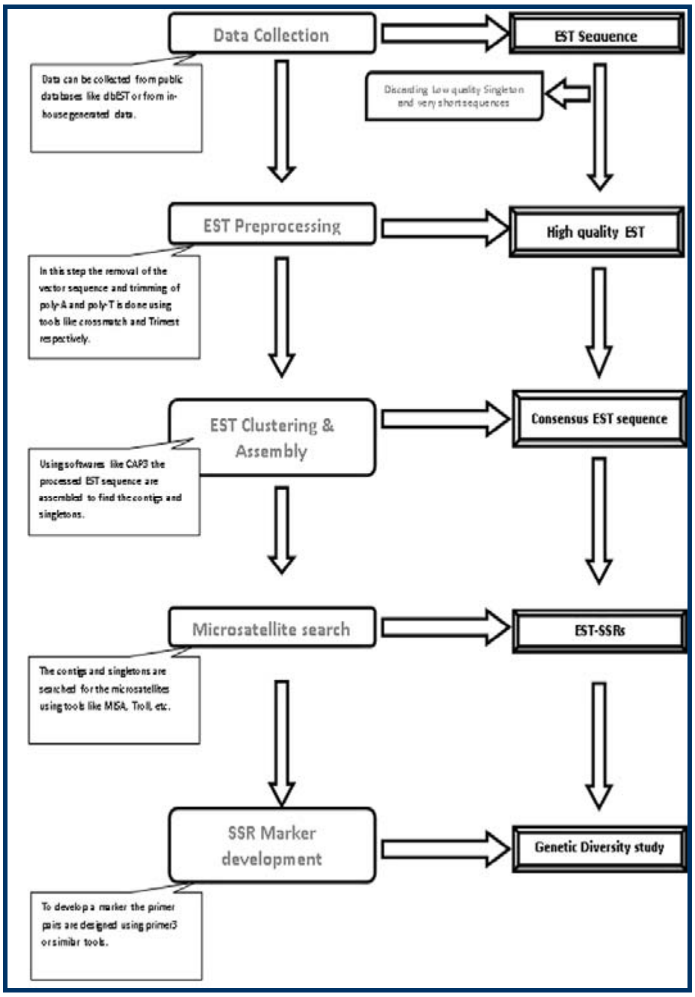
Data Collection
EST Sequence
Data can be collected from public databases like dbEST or from in-house generated data.
Discarding Low quality Singleton and very short sequences
EST Preprocessing
High quality EST
In this step the removal of the vector sequence and trimming of poly-A and poly-T is done using tools like crossmatch and Trimest respectively.
EST Clustering & Assembly
Consensus EST sequence
Using softwares like CAP3 the processed EST sequence are assembled to find the contigs and singletons.
Microsatellite search
EST-SSRs
The contigs and singletons are searched for the microsatellites using tools like MISA, Troll, etc.
SSR Marker development
Genetic Diversity study
To develop a marker the primer pairs are designed using primer3 or similar tools.

Figure 1: *In silico* approach to develop EST-SSR Markers

The purpose of EST clustering is to collect overlapping ESTs from the same transcript of a single gene into a unique cluster to reduce redundancy. Phrap and CAP3 are among the most extensively used programs for sequence clustering and assembly **[3, 6].** A simple way to cluster ESTs is by measuring the pair-wise sequence similarity between them. Then, these distances are converted into binary values, depending on whether there is a significant match or not, such that the sequence pair can be accepted or rejected from the cluster being assembled. The CAP3 program includes a number of improvements and new features and has a capability to clip 5' and 3' low-quality regions of reads. It uses base quality values in computation of overlaps between reads, construction of multiple sequence alignments of reads, and generation of consensus sequences. The program also uses forward-reverse constraints to correct assembly errors and link contigs. CAP3 is available to use both on the web and standalone.

There are several tools available for detection of SSRs from the nucleotide sequence with different parameters such as MISA (Micro Satellite identification tool by Thomas Thiel), TROLL etc **[7, 8].** MISA allows the identification and localization of perfect microsatellites as well as compound microsatellites which are interrupted by a certain number of bases. S. Rozen in 2000 developed Primer3 (Rozen and Skaletsky, 2000), a software which allows researchers to design the primers using a sequence with many different parameters **[9].** Also there are many other computational tools (both web based and standalone versions) available for primer designing. **(Figure 1)** shows a general methodology of *in silico* EST-SSR Marker development.

References:

[1] Boguski MS *et al. Nat Genet*. 1993 **4**: 332 [PMID: 8401577]
[2] Nagaraj SH *et al. Brief Bioinform*. 2007 **1**: 6 [PMID: 16772268]
[3] http://www.phrap.org
[4] ftp://ftp.ncbi.nih.gov/pub/UniVec/
[5] http://emboss.sourceforge.net/apps/#Apps
[6] Huang X & Madan A, *Genome Res*. 1999 **9**: 868 [PMID: 10508846]
[7] http://pgrc.ipk-gatersleben.de/misa
[8] Martins W *et al. Nucleic Acids Res*. 2006 **4**: e31 [PMID: 16493138]
[9] Rozen S & Skaletsky H, *Methods Mol Biol*. 2000 **132**: 365 [PMID: 10547847]

Edited by P Kangueane
Citation: Sahu *et al*. Bioinformation 8(4): 201-202 (2012)