Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 4, April -2015

Processing Genome Data For Finding Exact Tandem Reapeats In DNA Sequence

Tushar N. Patil[1], Dipak Y. Patil[2], Swapnil A. Vichare[3], Jayant V. Sathe[4], Atul B. Chavan[5]

Assistant Prof. – Mr. S.S. More

Department of Computer Science & Engineering, Sou. Sushila Danchand Ghodawat Institutes Atigre, Kolhapur.

Abstract — *Tandem repeats are multiple duplications of substrings in the DNA that occur contiguously, or at a short distance, and may involve some mutations (such as substitutions, insertions, and deletions). Tandem repeats have been extensively studied also for their association with the class of repeat expansion diseases (mostly affecting the nervous system). Comparative studies on the output of different tools for finding tandem repeats highlighted significant differences among the sets of detected tandem repeats, while many authors pointed up how critical it is the right choice of parametersWe are developing tool which collects, sorts, splice and provides statistical Overview on DNA data files. Observing various types of ESTs those are useful for downstream applications such as mining microsatellites specific to an organ, tissue or development stage.This tool can accept or useful for input in multiple formats like Genebank, EST or FASTA file to invoke tandem repeats with its total count and it's percentage of existence. And also invoke microsatellite by assigning list of organism file as per user's requirements. It is very useful to researcher.*

I. INTRODUCTION

One of the most interesting features of prokaryotic and eukaryotic genomes (both coding and non-coding regions) is the presence of relatively short perfect tandemly repeated DNA sequences. These repeated DNA sequences are distributed almost at random throughout the genome. Repeats containing DNA sequences have attracted much attention from researchers as – (i) they play important roles in the formation of hairpin structures that may provide some structural or replication mechanism (ii) they are often associated with neurological disorders and (iii) they are used as DNA markers, such as microsatellites or Simple Sequence Repeats (SSR), Inter Simple Sequence Repeats(ISSR) and Directed Amplification of Minisatellite DNA (DAMD-PCR) in Marker Assisted Selection (MAS), positional cloning, identification of quantitative and qualitative loci and mapping for breeding and evolutionary studies. Recent evidence also suggests that some Variable Number of Tandem Repeats (VNTRs) and SSR sequences play significant roles in the regulation of transcription, and that some may also influence the translational efficiency or stability of mRNA, or modify the activity of proteins by altering their structure. Expressed Sequence Tags (ESTs) are single-pass DNA sequences, usually about 300–500 nucleotides in length, obtained from mRNA (cDNA) representing genes expressed in a given tissue and/or at a given development stage. A typical EST usually contains only a portion of the coding region (either translated or untranslated, or both) of the original gene transcript. One of the useful applications of ESTs is in the study of the gene expression pattern in a given organ, tissue or development stage in response to a particular treatment. The composition of a tissue specific EST population, therefore, offers an overall overview of the expressed genes and, consequently, is a novel tool in gene discovery and in understanding the biochemical pathways involved in physiological responses. ESTs have also been mined for Single Nucleotide Polymorphisms (SNP) and SSR. Microsatellites or SSRs are stretches of DNA consisting of exact simple tandemly repeated short motifs of 1–6 base pairs in length. SSRs are one of the best DNA markers because they are highly polymorphic, inherited in a co-dominant fashion, and highly abundant, being dispersed evenly throughout the genome. They can serve as sequence-tagged sites for an choring in genetic and physical maps. The standard procedure for developing SSRs involves the construction of a small-insert genomic library, its subsequent hybridization with tandemly repeated oligonucleotides, and the sequencing of candidate clones. Unfortunately, this process is time consuming and labourintensive. There are several programs to locate repeat strings in sequences, such as Tandem Repeats Finder, TRF (Benson1999), REPuter (Kurtz *et al*. 2001), Simple Sequence Repeat Identification Tool, SSRIT (Kantety *et al*.2002), Simple Sequence Repeat Finder, SSRF (Sreenu *et al*. 2003), Search for Tandem Repeats IN Genomes, STRING (Parisi *et al*. 2003), and Microsatellite Search, MISA (Thiel *et al*. 2003). Although these repeat finding programs are useful, they have several disadvantages that limit their use. Important limiting aspects of these programs are the number of sequences that programs accept the length of the repeats they find, and acceptable DNA sequence formats. None of these repeat finding programs informs researchers about the distribution of repeats among organisms, organs, tissues, cell types or development stages when multi-sequences or organs are used. Briefly, an exact tandem repeat is a single exact tandem repetition of a suitable motif. If an exact tandem repeat undergoes a small number of point mutations, it becomes an inexact tandem repeat. A third type of variation can occur at compound repeats that contain two or more different tandem repeats. Bilgen *et al*. (2004) reported two different kinds of compound repeats, exact compound and inexact compound repeats.

Our preliminary studies using Gen- Bank DNA databases indicated that other kinds of compound repeats in DNA sequences also existed. These repeats cannot be detected by TRA and other programs (Bilgen *et al*. 2004). These repeats are termed as compound, imperfect, and extended compound repeats. Compound repeats are those repeats with two or more repeat strings run of the same or different uninterrupted repeats shown as (AGAAG) r1 (AGATAA) r2. Imperfect repeats are those sequences having at least two or more exact simple repeats separated by non-repeated nucleotides varying in size, shown by (TCTTC) r1CACATAA- (AGAAG) r2 (CACATAA nucleotides are non-repeated sequences in the given example). Extended compound repeats are sequences having at least two or more compound repeats, but one of the compound repeats is interrupted by non-repeating units of adjacent sequences shown as (CTTCT) r1 (AGAAG) r2 TCTTATGA (TATA) r3. ESTs are a fast, inexpensive way to determine which genes are being actively transcribed in a tissue or organ at a given stage of development. Since ESTs represent the transcribed part of the genome.

II. LITERATURE SURVEY

Several search tools are available for mining microsatellite repeats in assembled genome sequences. Originally, microsatellite mining from sequence databases involved pattern-match searches using BLASTN (Basic Local Alignment Search Tool that compares a given nucleotide query sequence with sequences contained in a nucleotide database) or using tools with similar algorithms. Some algorithms, such as Repeat-Pattern Toolkit and Repeatmasker, which were developed for locating genomic repeats, later became popular for mining microsatellites. These days, more sophisticated, user-friendly microsatellite-specific software, such as MISA (MIcroSAtellite, a microsatellite mining tool) and Tandem-Repeats Finder, are preferred. In general, dedicated microsatellite- finding tools can be classified broadly into three subcategories based on their architecture. Tools, such as MISA and TROLL (tandem repeat occurrence locator), detect tandem repeats following certain specific construction rules and ensure an exhaustive survey of all repeats. TROLL is based on Aho-Corasick Algorithm (ACA) that follows the 'dictionary approach', drawing the keyword tree adapted for bibliographic search and attempts to match keywords exactly. A common drawback of such algorithms is that they are heavily biased to mine exact-tandem repeats (ETR) or perfect repeats. Tools, such as Tandem- Repeats Finder (TRF) and mreps, use a two-phase screening of genomic sequences for the detection of microsatellites. During the first screening (or scanning), certain sequences are listed tentatively as microsatellites on the basis of certain search parameters selected by the user or by default of the tool itself. These regions are next validated for the presence of desired sequences following some statistical rules. This pool of sequences might not be exhaustive because some sequences passing the validation tests might not be detected by statistical tests. The third approach is the most straightforward, in which algorithms are designed to align a given motif or library of motifs along genomic sequences. Regions showing an alignment score higher than a given threshold are considered as microsatellites. The latter two approaches are effective for screening of approximate-tandem repeats (ATRs) or imperfect repeats in addition to exact-tandem repeats (ETRs). Owing to the differences in the underlying algorithms and the search criteria adopted(by default or by the user), any two tools are not likely to yield completely identical set of results. Leclercq et al. observed that Repeat masker and STAR (search for tandem approximate repeats) detect fewer microsatellites than TRF and mreps. We also recorded some discrepancies while trying to compare output as well as efficiency of different search tools. TRF suffers from its incompatibility with genomic sequences longer than 5 Mb. Because most of the eukaryotic chromosome sequences exceed this length, the use of TRF becomes lmited. We have used MISA for scanning whole-genome sequences with ease and efficiency. Tools, such as E-TRA (exact tandem repeats analyzer) and MISA, are also additionally powered to present statistical analysis. STRING (search for tandem repeats in genomes) presents results in graphical form. MISA, however, suffers from the disadvantage of not detecting interrupted repeats with a good efficiency and also suffers from inappropriate classification of different repeat motifs. A tool released recently, SciRoKo, is fast and suitable for whole -genome scanning. Moreover, it also provides statistical analysis of simple as well as compound repeats. For PCR amplification of the desired microsatellite locus, some of the mining tools, such as MISA and msatminer, have an inbuilt facility or provide links to primer designing tools. In any case, the choice of a microsatellite-mining tool generally depends on the nature and ultimate aim of the research as well as personal preferences of the user.

III. PROBLEM STATEMENT

During recent decades, microsatellites have become the most popular source of genetic markers. More recently, the availability of enormous sequence data for a large number of eukaryotic genomes has accelerated research aimed at understanding the origin and functions of microsatellites and searching for new applications. DNA, the mystical sequence where life starts, has been one of the major research objects in bioinformatics. In those DNA sequences disperse iterations of nucleotide motifs are called Tandem Repeats (TRs). TRs' genetic and evolutionary mechanisms remain controversial; however it is believed that they are functionally important for gene transcription, translation, chromatin organization, recombination, DNA replication, cell cycle, etc. Currently TRs, as important genetic makers, has been prevalently applied in realms such as paternity testing, forensic investigations and so on. Therefore, it is essential to find and study those TRs by using Data Mining technologies. Advanced user defined parameters/options let the researchers use different minimum motif repeats search criteria for varying motif lengths simultaneously. One of the most interesting features of genomes is the presence of relatively short tandem repeats (TRs). These repeated DNA sequences are found in both prokaryotes and eukaryotes, distributed almost at random throughout the genome. Some of the tandem repeats play important roles in the regulation of gene expression whereas others do not have any known biological function as yet. Nevertheless, they have proven to be very beneficial in DNA profiling and genetic linkage analysis studies. In this project we report two new utility tools; Organism Miner and Keyword Finder. Organism Miner utility collects, sorts, splice and provides statistical overview on DNA data files. Keyword Finder analyses all the sequences in the input folder and extracts and collects keywords for each specific organism or for all organisms, which have the DNA sequence and generates statistical overview. To meet above requirements the goal of the system is to design the tool necessary to identify microsatellites present into the genome nucleotide sequences.

Working of various existing microsatellite mining utilities and their limitations are incorporated in this chapter. The results obtained from existing utilities are not satisfactory and not up to the mark. Above problem statement helps to identify the microsatellite and their functionalities in all respect.

OrgMiner utility:

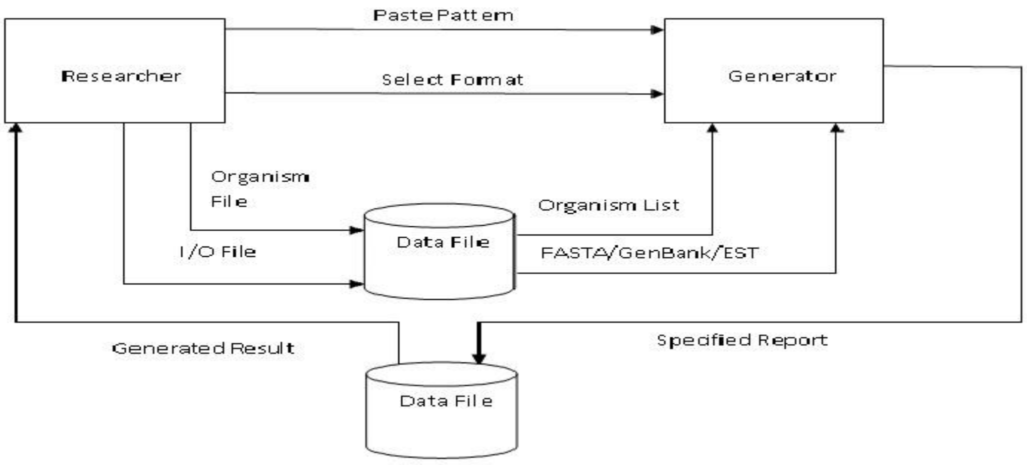

Fig: 1.1 System Architecture of OrgMinner

The utility contains following features.

A. Organism Miner (OrgMiner) utility sorts GenBank, EST or FASTA formatted DNA sequence files according to organism list which will be provided by the user. Input folder contains GenBank or EST formatted DNA sequence data that are under the investigation. Input folder may contain single file, multi files or multi files with multi sequences. Output folder is the folder of results containing the combined or spliced DNA sequence data which is generated by the OrgMiner upon completion of the analysis. OMETRs (Optimized Motifdivide Exact Tandem Repeats) algorithm is used to implement OrgMiner which gives the output in the form of chain of adjacently similar motifs (CASM).

B. By using CASM, tandem repeats will be captured in the form of di, tri, tetra format.

C. CASM (Chain of Adjacently Similar Motifs): For sequence $T = T1\ T2 \dots Tp \geq p$, if the similarity of any two adjacent motifs Ti and Ti+1 $(0 \leq i < p)$ meets the similarity threshold r, then T is a CASM, and p is its length.

D. Organism list file is a text file in which organisms are listed in each line. Name of the organisms can be given in the form of only ''genus name'' or ''genus and species name''. When the genus name is given (for instance Gossypium), OrgMiner will collect the data for all the Gossypium (Gossypium hirsutum, Gossypium raimondii, Gossypium arboreum

etc.). If a species; for instance, Gossypium barbadense is under the investigation, then organism list file should contain ‘‘Gossypium barbadense’’ in a single line.
E. Output file contains information about the total numbers of GenBank, EST or FASTA DNA sequences scanned, DBEST Id, Number of DNA sequences for each organism and Total percentage within whole sequences.
F. When OrgMiner utility results are compared with existing utility such as REPuter, Simple Sequence Repeat Finder(SSRF), Microsatellite Search(MISA), Simple Sequence Repeat Identification Tool(SSRIT) this will give better results in terms of motif is existing within file, its total count, type of file as input and percentage of EST with respect to specific organism. These results are varying according to input organism text file.
G. For existing tools such as REPuter, Simple Sequence Repeat Finder(SSRF), Microsatellite Search(MISA) it is mandatory to paste the whole sequences in their tool. So the researchers get certain limitations to use such tools. But OrgMiner does not have such limitation. It accepts raw file of genome database for processing which could be in any format.
H. System provides the facility to email results. It could be either pdf, excel or html format. User can email results by clicking on send mail link.
I. System facilitates both by selecting file and by paste option to check microsatellites in given nucleotide base sequences.
J. OrgMiner utility provides output in pdf, rtf and excel format as per the user demand.

OMETR (Optimize Motif Divide Exact Tandem Repeats) algorithm:
Algorithm: OMETRs (Optimized Motif-divide Exact Tandem Repeats)
Input: DNA sequence seq, motif length scopes (a, b), Similarity threshold r, minimum periods pmin;
Output: OMETRs (Optimized Motif Divide Exact Tandem Repeats) or CASM (Chain of Adjacently Similar Motifs)

```
Begin
        Set k=a;
        Repeat
        for (i=0; i<k; i++) do
        1) Starting from index i, divide seq by length k; let m0 be the first motif
        divided from seq, set length = 1, put m0 in the buffer, and for each motif
        mc divided from seq later(c starts from 1),
        do f = S(mc−1,mc , r) , and
        2) if (f == −1)
        buffer.add(mi);
        length ++;
        3) else if (length >= pmin)
        add buffer to CASMs;
        clear buffer;
        buffer. add (mi);
        length =1;
        End for
        Clear buffer, and Return all the CASMs;
        Until k = b;
End
```

Keyword Finder utility:

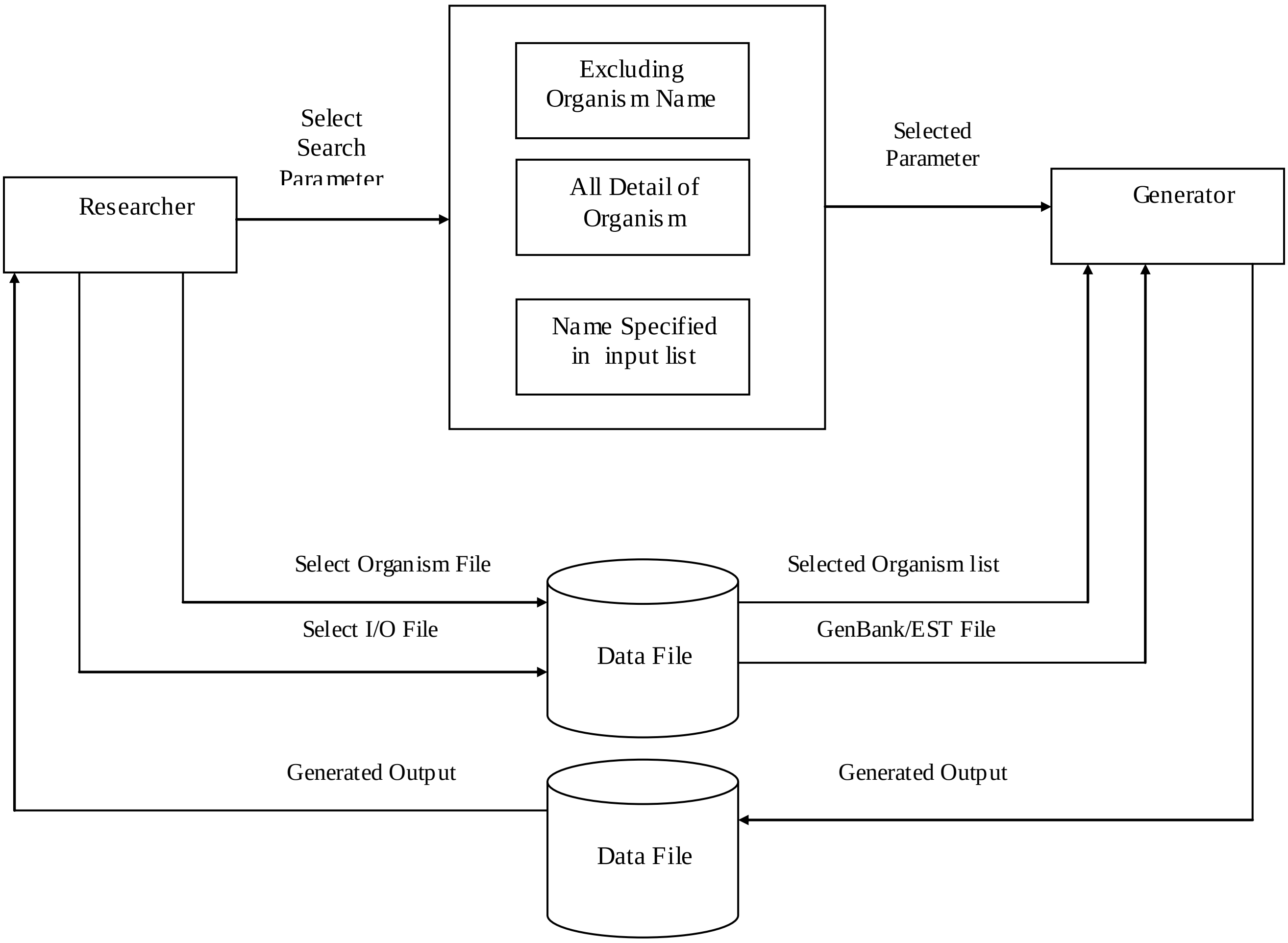

Fig: System Architecture of Keyword Finder.

The Keyword Finder utility contains following features:

A. Keyword Finder accepts list of organism file as an input file as shown in figure and according to the input file content it identifies keyword of type organ, development stage and tissue of that particular organism under investigation.

B. It incorporates statistical analysis of how many EST's are scanned and percentage of their existence with respect to total sequences.

C. Keyword Finder can analyze GenBank or EST formatted DNA data files.

IV. CONCLUSION

OrgMiner and Keyword Finder may enhance the use of these sequences for the development of tissue organ, development stage or stress specific microsatellite DNA markers, inter simple sequence repeats (ISSRs) and directed amplification of minisatellite DNA (DAMD-PCR). Utility programs reported here may provide great helps to splice DNA sequences data for a particular organism or a group of organisms, extract keywords for particular tissue, organ, or development stages for microsatellite data mining programs accepting Gen-Bank, EST and FASTA formatted DNA sequences. Reported programs could be useful for both gene mapping and association studies and discovering microsatellites and other type of tandem repeats located in coding regions of important genes that are expressed under various conditions of environment, stress, organ, tissue and development stage. Microsatellites and other tandem repeats located in coding regions of important genes that are expressed under various conditions of environment, stress, organ, tissue and development stage would also lead to the development of tissue/organ/development stage specific SSRs and that would be very valuable to understand the repeat function in gene and mapping for breeding and evolutionary studies. Furthermore these programs might be used to identify genes whose expression levels differ among the organs, tissues and development specific tissues.

V. FUTURE WORK

- To add Flanking sequence length option in OrgMiner and Keyword Finder utility. Flanking sequence consists of the 500 nucleotides on each side of a repeat. Flanking sequence is recorded in the alignment file. This may be useful for PCR primer determination.
- To provide option for motif selection in OrgMiner and Keyword Finder utility. A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains. Sequence motifs are short conserved regions of polypeptides. It will possible to select percentage of existence with respect to particular motif

VI. REFERENCES

[1] A.G. Ince, M. Karaca, M. Bilgen, A. N. Onus, Faculty of Agriculture, Akdeniz University, 07059 Antalya, Turkey."Digital differential display tools for mining microsatellite Containing organism, organ and tissue" Plant Cell Tissue Organ Cult (2008) 94:281–290 DOI 10.1007/s11240-008-9372-2.

[2] A.G. Ince, M. Karaca, M. Bilgen, A. N. Onus, Faculty of Agriculture, Akdeniz University, 07059 Antalya, Turkey."Digital differential display tools for mining microsatellite Containing organism, organ and tissue" Plant Cell Tissue Organ Cult (2008) 94:281–290 DOI 10.1007/s11240-008-9372-2

[3] Prakash C. Sharma1, Atul Grover2 and Gunter Kahl 3, 1. University School of Biotechnology, Guru Gobind Singh Indraprastha University, Kashmere Gate, Delhi 110 006, India. 2. Sri Radha Kissen Kanoria Center for Advanced Studies in Bioscience and Biotechnology, Banasthali Vidyapith, Banasthali 304 022, India. 3. Biocenter, University of Frankfurt, Frankfurt am Main, Germany GmbH, Frankfurt Innovation Center (FIZ), Frank furtam Main, Germany. "Mining microsatellites in eukaryotic genomes" TRENDS in Biotechnology Vol.25 No.11

[4] A. N. Pankratov, M. A. Gorchakov, F. F. Dedus, N. S. Dolotova,L. I. Kulikova, S. A.Makhortykh, N. N. Nazipova, D. A. Novikova, M. M. Olshevets, M. I. Pyatkov, V. R.Rudnev, R. K. Tetuev, and V. V. Filippov Institute of Mathematical Problems ofBiology, Russian Academy of Sciences, Institutskaia, Puschino, Moscowoblast,"Spectral Analysis for Identification and Visualization of Repeats in Genetic Sequences" 142290Russia. DOI: 10.1134/S 10 54 66180904018X.