International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 11, Issue 08 (August 2015), PP.30-39

"MS-Extractor: An Innovative Approach to Extract Microsatellites on 'Y' Chromosome"

Ch. Uma Maheswari[1], Prof. G.V. Padma Raju[2]
[1]2/2 M.Tech(C.S.T) S.R.K.R Engineering College, Bhimavaram ,[2]Professor and Head of C.S.E, S.R.K.R Engineering College, Bhimavaram,Andhra Pradesh, India.

Abstract:- Simple Sequence Repeats (SSR), also known as Microsatellites, have been extensively used as molecular markers due to their abundance and high degree of polymorphism. The nucleotide sequences of polymorphic forms of the same gene should be 99.9% identical. So, Microsatellites extraction from the Gene is crucial. However, Microsatellites repeat count is compared, if they differ largely, he has some disorder. The Y chromosome likely contains 50 to 60 genes that provide instructions for making proteins. Because only males have the Y chromosome, the genes on this chromosome tend to be involved in male sex determination and development. Several Microsatellite Extractors exist and they fail to extract microsatellites on large data sets of giga bytes and tera bytes in size. The proposed tool "MS-Extractor: An Innovative Approach to extract Microsatellites on 'Y' Chromosome" can extract both Perfect as well as Imperfect Microsatellites from large data sets of human genome 'Y'. The proposed system uses string matching with sliding window approach to locate Microsatellites and extracts them.

I. INTRODUCTION

1.1. Microsatellites

Microsatellites are tandem repeats of 1-6 nucleotides found at high frequency in the nuclear genomes of most taxa [2]. As such, they are also known as simple sequence repeats (SSR), variable number tandem repeats (VNTR) and short tandem repeats (STR).

***Example of microsatellites*:**

a) Repeat units
AAAAAAAAAAA= (A) 11 = mononucleotide (11bp)
GTGTGTGTGTGT= (GT) 6 = dinucleotide (12bp)
CTGCTGCTGCTG= (CTG) 4 = trinucleotide (12bp)
ACTCACTCACTC= (ACTC) 4 =tetranucleotide(16bp)

b) Homozygous microsatellite
…CGTAGCCTTGCATCCTT**CTCTCTCTCTCTCT**ATCGGTCTACGTGG… (46 bp)
…CGTAGCCTTGCATCCTT**CTCTCTCTCTCTCT**ATCGGTACTACGTGG… (46 bp)

c) Heterozygous microsatellite
…CGTAGCCTTGCATCCTT**CTCTCTCTCTCTCT** ATCGGTACTACGTGG… (46 bp)
…CGTAGCCTTGCATCCTT**CTCTCTCTCTCTCTCTCT**ATCGGTACTACGTGG… (50 bp)

A microsatellite locus typically varies in length between 5 and 40 repeats, but longer strings of repeats are possible. Dinucleotide, tri nucleotide and tetranucleotide repeats are the most common choices for molecular genetic studies. Dinucleotides are the dominant type of microsatellite repeats in most vertebrates characterized so far, although trinucleotide repeats are most abundant in plants [2], [3], [4]. Despitethe fact that the mechanism of microsatellite evolution and function remains unclear, SRs were being widely employed in many fields soon after their first description [6], [13], [14] because of the high variability which makes them very powerful genetic markers.

Apart from repeat copy number variation, a microsatellite tract (e.g. GCGCGCGCGC) also suffers from substitutions and indels of nucleotides thereby becoming an 'Imperfect' tract (e.g. GCGCGCAGCGC: GC repeat with an insertion of A). Imperfect microsatellites are more stable than perfect microsatellites as they are less prone to slippage mutations [11] and are known to play a role in gene regulation [8].

1.2 Y-STR's

The Y chromosome is one of the smallest human chromosomes with an average size of 60 million base pairs (Mb).Y chromosome-specific STRs have been proved to be an important tool in paternity cases, especially when the alleged father is deceased. Y-STRs are also useful for analysis of stains in forensic investigations when a male suspect is involved. Y-STRs are the most used Y chromosome markers in the forensic field due to their typing simplicity and high level of diversity. STR typing involves simple and reliable polymerase chain reaction (PCR)(a) techniques and is tolerant of very degraded samples. Of all Y chromosome polymorphic STRs in fig. 1 [7] described to date, DYS19, DYS385, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 and YCAII have more data accumulated, being the most used in population and forensic genetics. A number of multiplex reactions have been reported in the literature but Y STR multiplexes have not reached their potential…

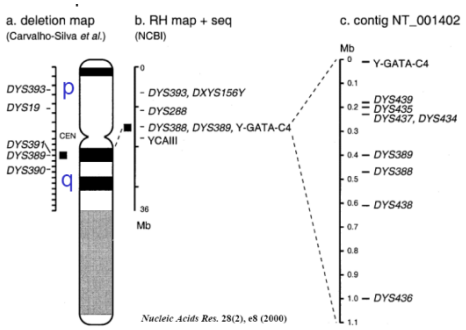

Fig. 1 Y chromosome STR's

Very little PCR optimization to-date (most work has been done with the original PCR primer sequences).
Summary of Y DNA Population Variation

- Fairly significant discrimination powers can be achieved when using many Y STR markers…very dependent on the population samples selected
- Population sub-structure exists and is more significant for Y SNPs
- We will need larger databases of Y STRs and Y SNPs for calculating powers of discrimination for Y haplotypes (for the same reasons as mt DNA).

No commercial Y STR kit exists yet (therefore these markers remain inaccessible to the general forensic DNA community). The proposed tool MS-Extractor can extract Microsatellite in Y chromosome.

II. LITERATURE SURVEY

In the due course of our studies on microsatellites, we made a survey of existing software tools for identification and extraction of microsatellites from nucleotide sequences. All these tools can be broadly classified into three categories: those who can identify

(i) Only perfect microsatellites (e.g. SSRF [10], GMATo [16]).
(ii) Both perfect and imperfect microsatellites (e.g. TRF [3]).
(iii) And can search particular motif in the genome sequence (e.g. SSR scanner [1]).

In our survey we also found some tools that extracts both perfect and imperfect but only considers substitutions but not indels. TRF tool identifies tandem repeats in larger homologous regions. It uses simple probabilistic model for tandem repeats, an overview and the set of criteria that guide the recognition process. This algorithm works without the need of specifying the pattern or pattern sizes. The algorithms of TRF [3],ATR Hunter [15] and STRING [9] have been designed to find tandem repeats of large-size motifs as large as 2000 bases and hence large numbers of microsatellites go unidentified by these methods. Many of these programs do not generate alignments between imperfect microsatellites and their expected perfect counter parts, and therefore require additional post-processing in order to study the mutational events in microsatellites. Imperfect Microsatellite Extractor [12] also extracts both perfect and imperfect microsatellites but it fails to extract on large data files. The proposed tool MS-Extractor is fast, highly sensitive and is also flexible where user can set the limits for imperfection (thus can be used for both perfect and imperfect microsatellites). The output comprises of a list of microsatellites each of which with information such as its total imperfection content, point mutations, sequence alignment with its perfect counterpart, whether the locus lies in the coding or non-coding region along with corresponding known details.

III. METHODOLOGY

The algorithm presented in this paper takes the human genome sequence of 'Y' chromosome line by line, process the line and extracts microsatellites present in that line. As a result, large no of Microsatellites are extracted and it works even on the file of large size. And also, for every motif identified it determines a consensus pattern. The algorithm identifies microsatellite if that sequence can be expressed as tandem repeat of size 1-6bp. The repeating motif at every iteration can harbor up to 'k' number of point mutations (substitutions or indels of nucleotides). For e.g. CAGTAGCATCAG ('CAG' repeat with substitutions=2).
The MS-Extractor works based on this definition and employs string matching algorithm with sliding window approach. MS-Extractor works on three step procedure.

Step 1: preprocessing

Identification of exact location of a microsatellite where the motif is repeated either adjacent to it (type 1 Extraction) (fig 1) or at certain intervals (type 2 Extraction) (fig 2). In this step the number of point mutations is set to zero.

GCCCAGCAG ***CAGCAG***CAGGCA

Fig. 2 Type 1 Extraction. The motif 'CAG' is identified tandemly with zero edit operations (k=0). This type can extend on both sides of the motif.

GCCCAG***CAG***CAC***CAG***CAGGCA

Fig. 3 Type 2 Extraction. The motif 'CAG' is identified after some intervention. The intervened sequence CAC is an iteration of CAG with C->G operation (k=1).

Step 2: Extending motif search

Extension of the microsatellite on both sides of the repeat as long as the imperfection limit is satisfied. Imperfection limit can be determined using two parameters k (substitutions and indels) and p (percentage of imperfection) set by the user requirement. The user can set a value for 'k' between 0 and m where m=repeat motif size. The number of imperfections between the individual copy and perfect motif is more than the limit. The percentage imperfection is calculated by

$$\text{Pencentage of imperefection(p)} = \frac{\text{number of mutation in the motif}}{\text{number of bases in the motif}} * 100$$

Step 3: classification

Given the repeat map of motifs identified, it determines a consensus pattern for the smallest unit in the tandem repeat. It creates the alignment with left flanking and right flanking sequence of the motif. And also classifies the motif is identified in coding region or non-coding region of the genome sequence. MS-Extractor uses .ptt file to classify the identified motifs. These details are shown as follows.
Consensus: TGGTC
Start:[57] End:[66] No. of iterations: 2
Total Imperfections: 0 (Substitutions: 0 Indels: 0) Tract-length: 10

Left Flanking Sequence:
CCCTCTCTAG

TGGTCTGGTC
****
TGGTCTGGTC
Right Flanking Sequence:

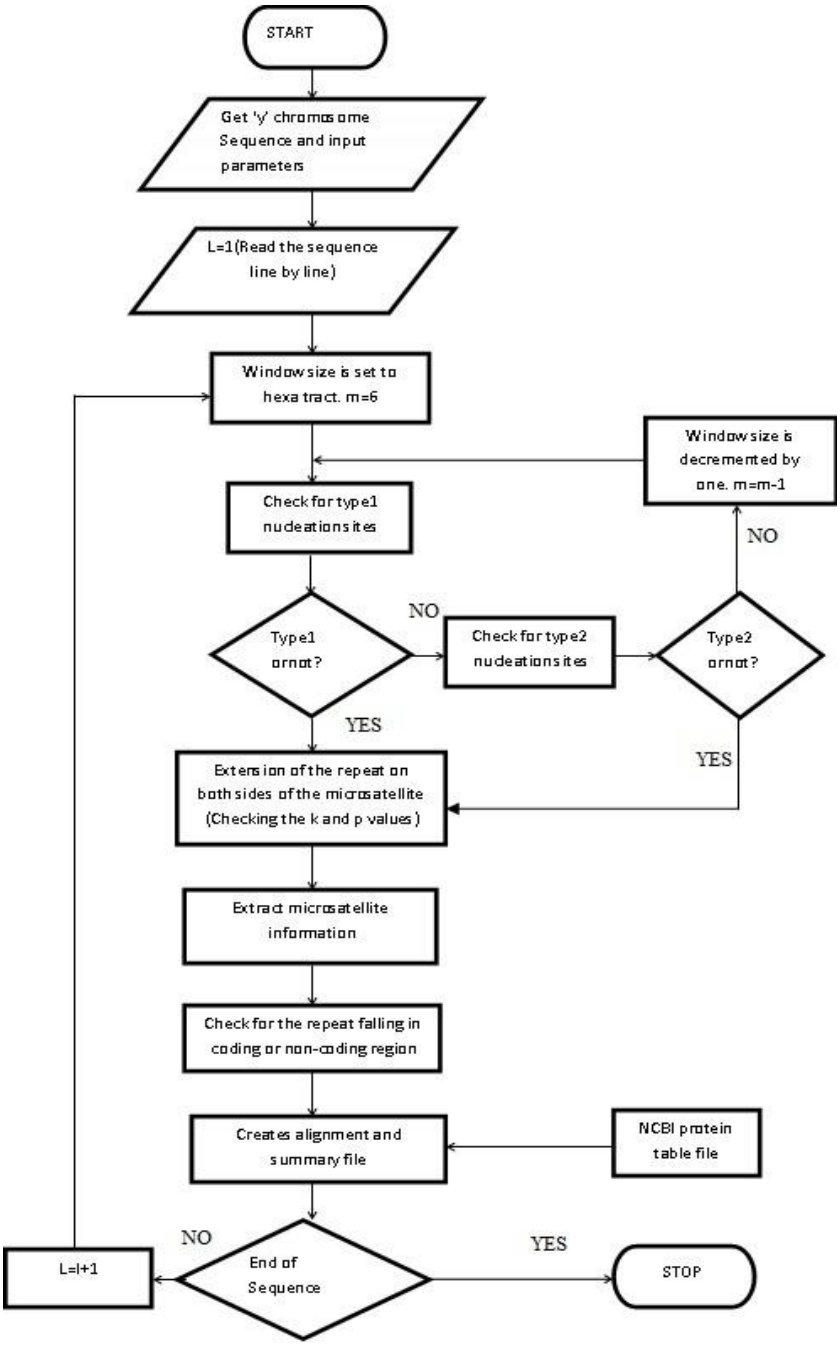

Fig. 3 Flowchart to extract microsatellites

The flowchart of MS-Extractor is shown in fig 3. MS-Extractor scans the genome sequence line by line and process the line for microsatellites starting from hexa nucleotide (i.e. tract size is set to 6). For e.g. $(ATGCGC)_3$ is a hexa nucleotide with repeat count =3.If hexa nucleotide is not detected in the sequence, then tract size is decremented by one and search for the penta nucleotide (i.e. m=5) and so on.

For each nucleotide it search for type 1 and type 2 microsatellites. And then extracts microsatellites extending both side of the motif. MS-Extractor eliminates redundancy of the motifs as it starts searching from hexa nucleotide. For e.g. in the sequence CGGCAACGGCAAA, the motif identified is $(CGGCAA)_2$ and the internal repeat of A within the tract is ignored.

MS-Extractor also creates the summary file with all the identified microsatellites and the repeat number, tract size, mismatch information (both substitutions and indels), and imperfection percentage for each base. It also creates alignments for each motif in the summary table.

Algorithm

The algorithm used in MS-Extractor is simple string matching algorithm with sliding window approach. MS-Extractor uses the following modules.

I. Checks whether the pattern is repeated no of times the user asked

Input: k (size set by user), size(motif size)
Output: 0 for match, negative value for mismatch.
Module: compare (k, size)
Step 1: for (j<size && k< array.size-1; j++, k++)
Read pattern 1
Step 2: for (j<size && k< array.size-1; j++, k++)
Read pattern 2
Step 3: returns the value of Pattern1. compareTo (pattern2).

II. Search for Type 2 Extraction:

Input: txt (sequence), motif size
Output: type 1 motif
Module: motif_check2 (txt, m)
Step 1: initially motif length=6.
Step 2: check whether the motif of length=6 is a mono, di, tri, tetra, penta or hexa nucleotide.
Step 3: expand on both side of motif with the length of the motif identified.
Step 4: compare two regions to check whether they match.
Step 5: if they match store the motif information.
Step 6: else decrement the length and goto step 2.
Step 7: stop.

III. Checks whether the motif is mono, di, tri, tetra, penta or hexa nucleotide.

Input: pat **(**motif)
Output: type of motif
Module: type_chckr (pat)
Step 1: iis initialized to 0.
Step 2: length of the motif is identified.
Step 3: if i=1, return mono.
Step 4: else if i=2, return di.
Step 5 else if i=3, return tri.
Step 6 else if i=4, return tetra.
Step 7 else if i=5, return penta.
Step 8 else return hexa.

IV. Checks for substitutions in 2 patterns

Input: sub (repeated copy), pat (perfect motif), k (limit)
Output: 0 for no substitutions or r
Module: sub_chckr (sub, pat, k)
Step 1: r, i, j initialized to 0.
Step 2: until pat is not null repeat following 2 steps.
Step 3: pat is compared to sub
Step 4: if equal r++, i++, j++
Step 5: compare r and k, if r>k return 0 else return r.

V. Checks for Indels on the right side pattern

Input: Ind (repeated copy), pat (perfect motif), k (limit)
Output: 0 for no indels or r
Module: indl_chckr (Ind, pat, k)
Step 1: repeat k number of times
Step 2: repeat until r<m (motif size)
Step 3: Ind is copied to temp
Step 4: call sub_chckr (temp, pat, k)
Step 5: if r value is not zero return r else -1.

VI. Search for Type 2 Nucleation sites:

Input: txt (sequence), motif size
Output: type 1 motif
Module: motif_check2 (txt, m)
Step 1: initially motif length=6.
Step 2: check whether the motif of length=6 is a mono, di, tri, tetra, penta or hexa nucleotide.
Step 3: expand on both side of motif with the length of the motif identified.
Step 4: compare two regions to check whether they match.
Step 5: if they match store the motif information.
Step 6: else decrement the length and goto step 2.
Step 7: stop.

These modules are used to identify motifs of size 1-6bp within the limit of substitutions and indels set by the user. And also we can identify that the motif is of type1 or type 2. MS-Extractor can apply these modules on both sides of the motif. And also there are other modules for creating primer classification and to create summary file. The module for imperfection percentage check is also included in this system.

IV. IMPLEMENTATION

MS-Extractor has been developed by using java language, which is platform independent.MS-Extractor is tested and compared with other tools that extract microsatellites both perfect and imperfect. This tool extracts motifs from large data sets from mb's to gb's. User can set parameters through the interface. HTML forms are used to develop an interface.

The input to the MS-Extractor given by the user involves the human genome sequence of 'Y' chromosome and also the following parameters; (i) number of repeats, (ii) number of substitutions/indels, (iii) imperfection percentage,(iv) coding table file. These parameters changed as per user requirements. Once these parameters set the batch file is executed. The output of this system contains two files. One of them is summary file which contain information about all the motifs identified along with their number of iterations, tract size, start & end locations, imperfection percentages for each base. Second file contains primer classification, alignments of each motif with consensus. These two files can be available in both text format and in HTML forms. The summary file can be used for further classification of motifs as input to other files. HTML forms contains links for each motif, these links contains a file attached to it. The file contains the alignment and consensus.

V. RESULTS AND DISCUSSION

To describe the system capability, we analyze the human genome 'Y' chromosome and extract microsatellites. The obtained results are compared with other tools. MS-Extractor is accurate when compared to those tools. MS-Extractor is fast and flexible and extracts more number of microsatellites. The proposed system can process the large data sets where as some tools fails to extract from them. To demonstrate the whole process we run the tool with some parameters in basic mode. The imperfection percentage(p) is set to 10% for all nucleotides; the imperfection limit is set as follows, (for mono=1, di=1, tri=1, tetra=2, penta=2, hexa=3), and the repeat number also set (for mono=10, di=5, tri=4, tetra=3, penta=2, hexa=2). Human 'Y' chromosome is given as input which is of size 25mb, and a protein information table also provided to classify the coding and no-coding regions. MS-Extractor identified many more microsatellites from the sequence and created a summary file as follows;

Table1. Summary file

Consensus		Iteration	Tract-size		Start	End	imperfection %
CTAACC	6	36	1	36	0		
TGGTC	2	10	57	66	0		
GCACCT		2	12	3	14	0	
CCCTT	2	10	8	17	0		
AAT	6	18	1	18	0		
TCTGT	2	10	22	31	0		
AACCCT		2	12	1	12	0	
TTCC	4	16	11	26	6		
TCCC	3	12	23	34	8		

TCCC	4	16	44	59	6
T	10	10	6	15	0
GAGG	3	12	9	20	8
GGCG	3	12	25	36	8
GGGGA	2	10	60	69	0
T	19	19	7	25	0
T	21	21	35	55	9
AATACA	2	12	4	15	0
CCCTT	2	10	27	36	0
A	21	21	24	44	9
ATA	7	21	29	49	4
TCTGT	2	10	52	61	0
CCTAAC	2	12	55	66	0
TTCC	3	12	6	17	0
TCCC	4	16	14	29	6
TCCC	4	16	30	45	6
T	10	10	46	55	0
CTTTC	2	10	23	32	0
CTC	4	13	1	13	7
CTC	4	12	54	65	8
CTC	4	12	13	24	8
CTC	4	12	50	61	8
CTC	4	12	57	68	8
GACA	3	12	29	40	8
GAGG	3	12	54	65	8
GCGGG	4	20	34	53	5
CTCC	4	16	12	27	6
TCCAA	2	10	53	62	0
TCTCC	2	10	1	10	0
TAAAAA	2	12	45	56	0
ATAA	4	16	54	69	6
TTATG	2	10	22	31	0
ATA	4	12	53	64	8
TATGT	2	10	28	37	0
TAAAAA	2	12	53	64	0
A	15	15	5	19	0
GTG	4	11	34	44	9
AGGCCA	2	12	1	12	0
TAA	6	18	29	46	0
TAGGTC	2	12	41	52	0
TAGGTC	2	12	19	30	0
TAGGTC	2	12	45	56	0
TAGGTC	2	12	23	34	0
TGGGTC	2	12	4	15	0
CTGT	3	12	36	47	8
CTGT	3	12	9	20	8
CTGT	3	12	38	49	8
TAGACC	2	12	31	42	0
CTGT	3	12	55	66	8
CTGT	3	12	1	12	8
AGGCCC	2	12	1	12	0
GACCTA	2	12	14	25	0
GACCTA	2	12	53	64	0

Table 1: continued

Consensus	Iteration	Tract-size	Start	End	imperfection %
CTGT	3	12	44	55	8
AGGCCC	2	12	44	55	0
GACCT	2	12	57	68	0
GACCTA	2	12	26	37	0
TAGACC	2	12	50	61	0
CTGT	3	12	43	54	8
AGGCC	2	12	43	54	0
GACCTA	2	12	27	38	0
CTGT	3	12	18	29	8

Table 1 refers to the summary file of MS-Extractor. Summary file contains each and every motif identified using this marker along with the information of the motif that includes iterations, tract size, starting and ending positions of the motif. MS-Extractor not only creates summary file. It is more flexible when compared to other tools presented in literature survey of this paper. It also creates alignment for each motif in the summary table and these alignments are stored in both text format and in HTML forms, provided as a link to the particular motif of the consensus. The following is the alignment file;

Table 2. Alignment table
Sequence:
Sequence Length :bp
0Imperfection %c : Mono: 10.0,Di: 10.0, Tri:10.0, Tetra:10.0, Penta: 10.0, Hexa:10.0
Mismatch in Pattern : Mono: 1,Di: 1, Tri:1, Tetra:2, Penta: 2, Hexa:3
Repeat Number : Mono: 10,Di: 5, Tri:4, Tetra:3, Penta: 2, Hexa:2
Composition: A: -100, T: -100, C:-100, G:-100

ALIGNMENTS:
Consensus: CTAACC
Start:[1] End:[36] No. of iterations: 6
Total Imperfections: 0 (Substitutions: 0 Indels: 0) Tract-length: 36

Left Flanking Sequence:

CTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
****
CTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
Right Flanking Sequence:

Consensus: TGGTC
Start:[57] End:[66] No. of iterations: 2
Total Imperfections: 0 (Substitutions: 0 Indels: 0) Tract-length: 10

Left Flanking Sequence:
CCCTCTCTAG

TGGTCTGGTC
****
TGGTCTGGTC
Right Flanking Sequence:

Consensus: GCACCT
Start:[3] End:[14] No. of iterations: 2
Total Imperfections: 0 (Substitutions: 0 Indels: 0) Tract-length: 12

Left Flanking Sequence:
--------AA

GCACCTGCACCT
****
GCACCTGCACCT
Right Flanking Sequence:

Table 2 contains the alignment in the form of consensus for each motif. These alignments are linked to the corresponding motif in the summary file. These links are provided in HTML forms.The fig 4 shows the graph that is drawn between TRF, sputnik and MS-Extractor. The input sequence, human genome 'Y' chromosome is given as input to the all three markers. Both TRF and GMATo failed to extract all motifs in the given sequence. MS-Extractor showed the utmost performance and extracted all the microsatellites in the 'Y' chromosome.

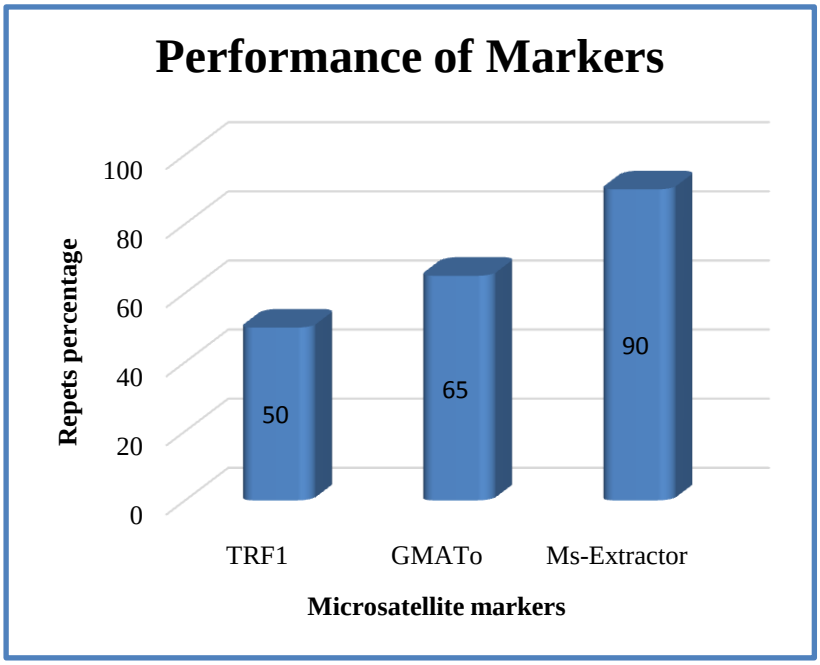

Fig 4: A graph that shows performance of markers. TRF extracted up to 50%, whereasGMATo identified 65%, and MS-Extractor identified upto 90% and more for human genome 'Y' chromosome.

MS-Extractor embodies all the required features for a systematic analysis of microsatellites which are not readily available in the other tools, as MS-Extractor has been designed keeping in view of the limitations we encountered with the other available tools. Using MS-Extractor, the users can: (i) search only perfect as well as imperfect microsatellites; (ii) get the coding/non-coding information of the microsatellite tracts;(iii) generate alignments with their perfect counter parts to know about substitutions and indels; (iv) restrict the imperfection limit for repeat unit of each size; (v) set the imperfection percentage threshold of the entire tract of each repeat size; (vi) restrict the minimum number of repeat units of a tract of each size;

VI. CONCLUSION

In this paper, we proposed a tool MS-Extractor to process the human genome sequences line by line and extract the microsatellites in single run. As it process the sequence line by line it occupies less memory space and it is fast. MS-Extractor is accurate and extracts 99% of all microsatellites and creates summary file with all the information as repeat number. It also provides information about coding/non-coding region. MS-Extractor is compared with other markers available and our marker showed best results among all other markers with an effective user interface.MS-Extractor is flexible and user interactive, as it provides a flexible environment for user by letting user to set the mutation limits and other parameters.The future work for this marker includes adding SSR standardization and compound microsatellite extraction.

REFERENCES

[1]. Anwar,T. and Khan,A.U., (2006). SSRscanner: a program for reporting distribution and exact location of simple sequence repeats. Bio information,1, 89–91.

[2]. Beckmann J.S. &Weber J.L., (1992). Survey of human and rat microsatellites. Genomics, 12, 627–631.

[3]. Benson,G., (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res., 27, 573–580.

[4]. Chen, C.X., Zhou, P., Choi, Y.A., Huang, S., Gmitter, F.G., 2006. Mining and characterizing microsatellites from citrus ESTs. Theor. Appl. Genet. 112, 1248-1257.

[5]. Kantety, R.V., Rota, M. L., Matthews, D.E., Sorrells, M.E., 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol. Biol. 48, 501-510.

[6]. Litt, M. &Luty, J.A. (1989) Ahypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal ofHuman Genetics*, 44, 397-401.

[7]. Mark A. Jobling and Chris Tyler Smith, (2003). The Human 'Y' Chromosome: An Evolutionary Marker Comes of Age. Doi: 10. 1038.

[8]. Meloni,R. et al., (1998). A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. Hum. Mol. Genet., 7, 423–428.

[9]. Parisi,V. et al., (2003). STRING: finding tandem repeats in DNA sequences. Bioinformatics, 19, 1733–1738.

[10]. Sreenu,V.B. et al., (2003). MICAS: a fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences.Appl. Bioinformatics, 2, 165–168.

[11]. Sturzeneker,R. et al., (1998). Polarity of mutation in tumor-associated microsatellite instability. Hum. Genet., 102, 231–235.

[12]. Suresh B. Mudunuri and Hampapathalu A. Nagarajaram, (2007).IMEX(Imperfect Microsatellite Extractor). Vol .23no .10.

[13]. Tautz D. (1989) Hypervariability of simple sequences as a general source for polymorphicDNA markers. *Nucleic Acids Research*, 17, 6463-6471.

[14]. Weber J.L. & May P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, 44, 388-396.

[15]. Wexler,Y. et al., (2004). Finding approximate tandem repeats in genomic sequences. RECOMB 2004.

[16]. Xuewen Wang*, Peng Lu &Zhaopeng Luo, 2013. GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. ISSN 0973-2063 (online) 0973-8894.

[17]. You-Chun Li, Abraham B. Korol, TzionFahima and EviatarNevo, (2004). Microsatellites with in genes: Structure, function and evolutionMol. Biol. Evol 21(6):991-1007.