Short Scientific Report

Journal of Plantation Crops, 2015, 43(1):62-66

MAPS: A web-based tool for detection of microsatellites in whole genome sequences

T.P. Jamshinath, S. Naganeeswaran, Jomon K. Jose and M.K. Rajesh*

Bioinformatics Centre (DBT Sub-DIC), ICAR - Central Plantation Crops Research Institute, Kasaragod-671124, Kerala, India

(Manuscript Received: 19-12-14, Revised: 10-01-15, Accepted: 30-01-15)

Keywords: Coconut, genome, microsatellites, SSR

Microsatellites or simple sequence repeats (SSRs) are tandemly repetitive DNA sequences with very short nucleotide motif of 1-6 bp (Ellegren, 2004). Genomes of both eukaryotes and prokaryotes are scattered with these repeats (Gur-Arie *et al*., 2000; Bacolla *et al.*, 2008). SSRs are ubiquitously found in the genome, both in non-coding as well as protein-coding regions, 5'-UTRs and 3'-UTRs and introns (Madsen *et al*., 2008; Riley and Krieger, 2009). The expansions and/or contractions of SSRs can lead to gene gain or loss of function and in non-coding regions; these have been known to affect gene regulation (Li *et al*., 2004). Significantly high mutability at microsatellite loci, ranging from 10^{-6} to 10^{-2} per generation (Schlotterer, 2000) has a role in genome evolution by creating genetic variation within a gene pool. This genetic variation occurs primarily by slipped-strand mispairing and subsequent errors occurring during DNA replication/ repair/recombination (Toth *et al*., 2000). Many other factors such as the repeat number, sequence of the repeat motif, the genomic position of the microsatellite and the genetic-biochemical background of the cell may also possibly contribute to the observed divergence of microsatellite distribution in various organisms (Schlotterer, 2000). In prokaryotes, SSRs have been implicated to possess important roles in gene regulation, pathogenesis and bacterial adaptation, host interaction and genome evolution (Mrazek *et al.,* 2007).

Microsatellites have been developed into one of the most popular classes of genetic markers as they possess certain desirable properties like high reproducibility, multi-allelic nature, co-dominant mode of inheritance, abundance and wide genome coverage (Morgante and Olivieri, 1993; Toth *et al*., 2000). Because of these properties, they have found wide applications in various fields such as DNA fingerprinting, paternity and forensic studies, evolutionary studies *etc.* (Varshney *et al.*, 2005).

Isolation of SSR loci from a new species involves time-consuming, labor-intensive and extensive procedures. With the advent of next-generation sequencing (NGS) technologies, complete genome/transcriptome sequences are available for a number of organisms, which provides opportunities to examine the genomic locations, distributions and frequencies of microsatellites in these organisms. *In silico* methods, based on computational tools, have been developed to screen sequence data and produce a complete list of SSRs. Presently, over a dozen tools were available for mining microsatellite repeat from genome sequences. Some of these tools concentrate on finding SSRs, while the others include additional function of designing PCR primers flanking the SSRs, thus facilitating the marker development process. These tools are very useful, providing a standalone version and, in some case, a web online version as well. Some of the microsatellite-specific softwares available are MISA (Microsatellite) (Thiel *et al*., 2003), SSR Locator (da Maia *et al*., 2008), SPUTNICK (La Rota *et al*., 2005), tandem repeat finder (TRF) (Benson, 1999), TROLL

*Corresponding Author: mkraju.cpcri@gmail.com

(Castelo *et al*., 2002), simple sequence repeat identification tool (SSRIT) (Temnykh *et al*., 2001) and RISA (Kim *et al*., 2012). Each of the tools uses different algorithms, architecture and approaches for mining microsatellite. As a result, the microsatellite dataset identified by a given tool is often different from that obtained by using another tool.

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 (Kim *et al*., 2012). However, these tools possess some limitations like their capacity to handle only small data sets, requirement of analyzing the results obtained manually, complicated installation procedures, and their incapability to automatically design primers for the detected SSRs *etc*. (Kim *et al*., 2012)

In the present work, we present a new comprehensive computing tool with a web interface, named as MAPS (Microsatellite analysis and prediction software), which can successfully detect all types of microsatellites from whole genome sequences, report the motifs composition, frequency and genomic position of all microsatellites and also display the frequency of occurrence and distribution of microsatellites within coding and non-coding region of entire genome sequences. MAPS algorithm was developed using Java for the purpose of identification and analyzing microsatellite from nucleotide records. MAPS has the capability to extract perfect, imperfect and compound microsatellites from whole genome sequence effectively and accurately and is also able to differentiate microsatellites on the bases of coding and non-coding region from genome (only in GenBank file). The program can perform readings on fasta or multi-fasta files as well as GenBank files, searching all possible arrangements in each sequence. The front end was developed using JSP (Java Server Page) and MySQL database management system served as a back end. Apache web server has been used in web server application (Fig.1). The results are saved in MySQL for a certain

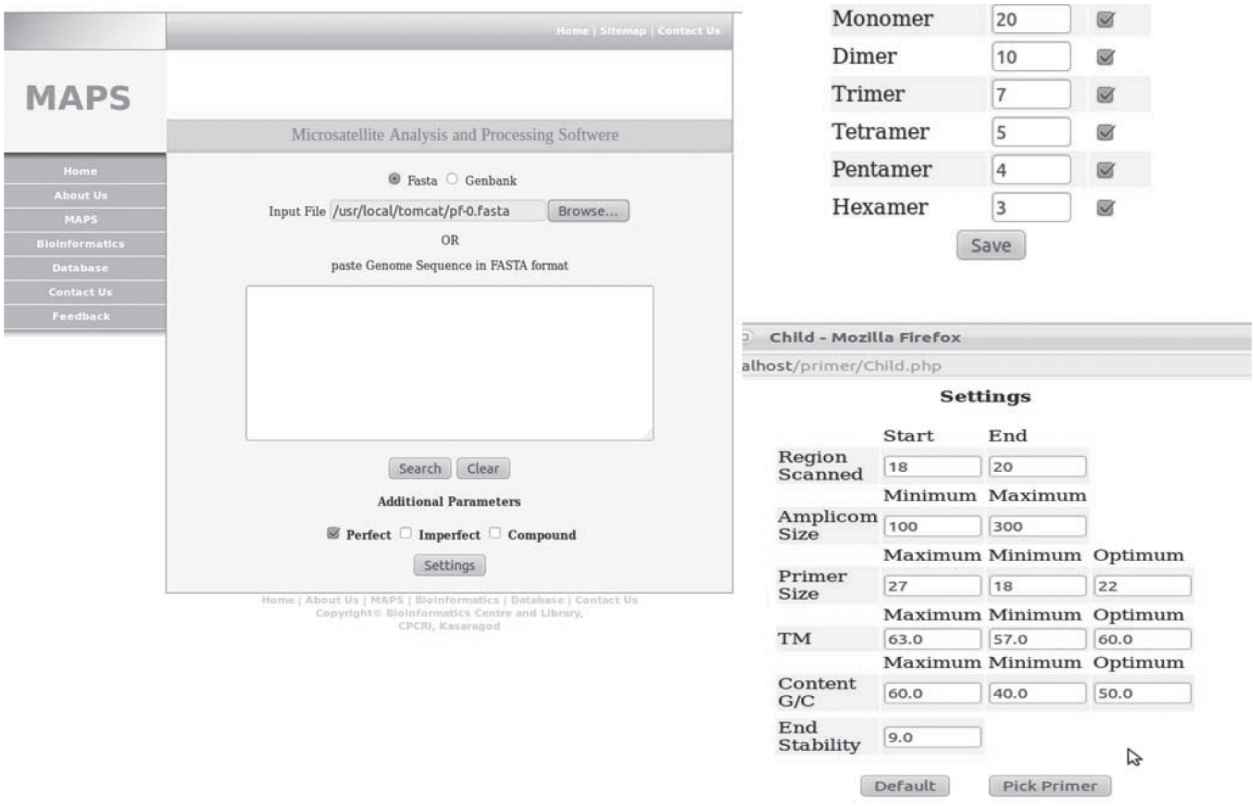

Fig. 1. The MAPS interface: search page, SSR parameter settings and primer parameter settings

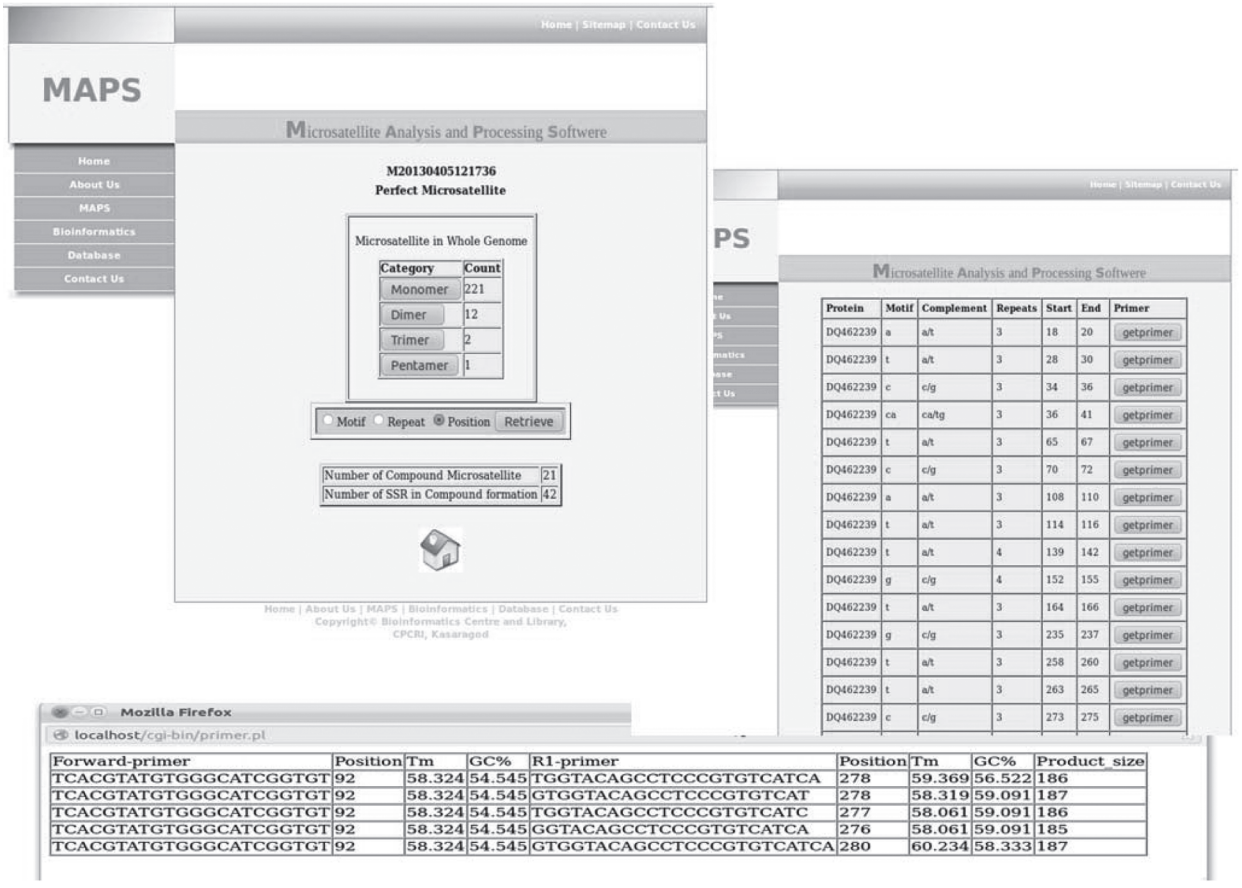

Fig. 2. SSR motif mining, SSR position and primer designing for particular SSR

period of time and user can retrieve the information by a unique ID given at the time of submission.

The MAPS SSR search is based on the motif-matching method, where each motif (mono- to hexa-mer) is scanned and the repeats are counted throughout the genome. MAPS has been configured to locate a minimum of 8 bp SSRs: Monomer (x8), 2-mers (x4), 3-mers (x4), 4-mers (x3), 5-mers (x3), 6-mers (x3). Three types of search options have been incorporated based on the type of SSRs *viz.*, perfect, imperfect and compound. In perfect search, a motif at position 'i' is tested for identity with the nucleotide at position i+t, where t is the motif length (1-6). If the motifs meet the specified minimum length, the particular SSR is saved to database. Identity 'i' is increased to i=i+1 each time until no further identity can be found. In case of imperfect SSRs, identity increases i=i+2 from i and m=m+1 in the case of compound SSR, where 'm' is the position of any motif found on nucleotide sequence.

Separate modules have been developed which execute programs to screen entire genome for finding all possible arrangements of motifs, frequency of motifs, combination of nucleotides and to locate position of microsatellites. MAPS can also identify all possible microsatellites within ORFs and

Table 1. Comparison of 'MAPS' with two commonly used SSR detection tools

Parameters	MAPS	MISA	SSR Locator
Perfect microsatellites	√	√	√
Compound and imperfect microsatellites	√	√	√
FASTA/ multi-FASTA	√	√	√
Genbank	√	*	*
Coding/non-coding regions	√	*	*
Amino acid composition	√	*	√
Speed of operation	Fast	Fast	Moderate

Table 2. Total microsatellite count in whole genome of *Pseudomonas fluorescens* Pf5 using three computational tools

Strain	MISA		SSR Locator		MAPS	
	Whole genome	Coding sequence	Whole genome	Coding sequence	Whole genome	Coding sequence
Pf-5	2,227,891	1,951,850	6,80,570	6,22,664	1,813,624	1,451,104

non-coding sequence from the whole genome sequence. The program can read a sequence of any size as memory is dynamically allocated. MAPS can store all information about microsatellite of given organism into a database according to the parameters specified by the user. User can also retrieve all microsatellite data from the database. The tool is user-friendly, can easily handle large data sets and fast.

Perl language-based scripts is integrated with MAPS which passes the predicted SSR information into Primer 3.0 and allows the user to design both forward and reverse primers for flanking region of a particular SSR. A window permits selection of Primer 3.0 parameters, such as range of primer and amplicon sizes, as well as optimum primer size, ranges of melting temperature (Tm) (minimum, maximum and optimum) and GC content (minimum and optimum) (Fig. 2).

A comparison of features of MAPS with two commonly used SSR detection tools, namely MISA and SSR Locator, is provided in Table 1. In order to evaluate the efficiency of MAPS, the whole genome of plant growth promoting rhizobacteria (PGPR) *Pseudomonas fluorescens* Pf5 was analyzed with two commonly used SSR detection tools, namely MISA and SSR Locator, using the same parameters for minimum number of repeats as MAPS. The results, provided in Table 2, highlight the efficiency of MAPS tool. The tool is hosted at http://14.139.158.118:8080/MSQL/Home.jsp.

To conclude, the combination of an extremely sensitive, but fast search algorithm with a built-in summary statistic tool makes MAPS an excellent tool of choice for detecting microsatellites from whole genome sequences. MAPS can be used successfully for mining microsatellites within the coding and non-coding region of genomes and also compound and imperfect microsatellites. The tool producing reports for frequency of motifs, nucleotide arrangement, position and also designs primers for the detected microsatellites.

Acknowledgments

The authors thank Dr. George V. Thomas, former Director, ICAR - CPCRI, Kasaragod for his support and encouragement. This work was supported by a grant from Department of Biotechnology (BTISnet), Government of India.

References

Bacolla, A., Larson, J.E., Collins, J.R., Li, J., Milosavljevic, A., Stenson, P.D.,Cooper, D.N. and Wells, R.D. 2008. Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. *Genome Research* **18**: 1545–1553.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573-580.

Castelo, A.T., Martins, W. and Gao, G.R. 2002. TROLL-tandem repeat occurrence locator. *Bioinformatics* **18**: 634-636.

da Maia, L., Palmieri, D., Queiroz, V., Marini, M., Félix, F.A. and Costa A. 2008. SSRLocator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International Journal of Plant Genomics* **1**: 1-9.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**: 435-445.

Gur-Arie, R., Cohen, C.J., Eitan, Y., Shelef, L., Hallerman, E.M. and Kashi, Y. 2000. Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. *Genome Research* **10:** 62–71.

Kim, J., Choi, J-P., Ahmad, R., Oh, S-K., Kwon S-Y. and Hur C-G. 2012. RISA: a new web-tool for rapid identification of SSRs and analysis of primers. *Genes and Genomics* **34**: 583-590.

La Rota, M., Kantety, R.V., Yu, J-K. and Sorrells, M.E. 2005. Non-random distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* **6**: 23.

Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. 2004. Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution* **21**: 991-1007.

Madsen, B.E., Villesen, P. and Wiuf, C. 2008. Short tandem repeats in human exons: a target for disease mutations. *BMC Genomics* **9**: 410.

Morgante, M. and Olivieri, A.M. 1993. PCR-amplified microsatellites as markers in plant genetics. *Plant Journal* **3**: 175-182.

Mrazek, J., Guo, X. and Shah, A. 2007. Simple sequence repeats in prokaryotic genomes. *Proceedings of National Academy of Sciences USA* **104**: 8472-8477.

Riley, D.E. and Krieger, J.N. 2009. UTR dinucleotide simple sequence repeat evolution exhibits recurring patterns including regulatory sequence motif replacements. *Gene* **429**: 80-86.

Schlötterer, C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**: 365-371

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S. and McCouch, S. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Research* **11**: 1441-1452.

Thiel, T., Michalek, W., Varshney, R.K. and Graner, A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* **106**: 411-422.

Toth, G., Gaspari, Z. and Jurka, J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research* **10**: 967-981.

Varshney, R.K., Graner, A. and Sorrells, M.E. 2005. Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* **23**: 48-55.