Plant Archives Vol. 16 No. 2, 2016 pp. 510-514 *ISSN 0972-5210*

AN EFFICIENT BIOINFORMATICS TOOLS FOR IDENTIFICATION OF SSRS : A REVIEW

Shashi Prabha, Ashwani Yadav, Ashwani Kumar, Hemant Kumar Yadav[1], Sujit Kumar, R. S. Yadav[2] and Rajendra Kumar*

U. P. Council of Agricultural Research, Lucknow - 226 010 (Uttar Pradesh), India.
[1]National Botanical Research Institute, Lucknow - 226 001 (Uttar Pradesh), India.
[2]Department of Botany, D.A.V. College, Muzaffarnagar - 251 001 (Uttar Pradesh), India.

Abstract

During the last few decades, the use of molecular markers has played an increasing role in plant breeding and genetics. Among various molecular markers, SSR markers are the most popular and versatile molecular marker have been widely used because of their valuable features such as abundance, co-dominance inheritance, high polymorphism, reproducibility and ease of assay by PCR. All these feature which make them very suitable for physical mapping, genetic mapping, comparative mapping, genetic diversity, QTLs analysis and study of evolutionary relationship. Using molecular method for the development of SSRs might be laborious, costly and time-consuming. Using bioinformatics tools to mine sequences in public databases facilitates a cost-effective detection of SSRs. SSRs are useful as molecular markers because their development is inexpensive. In this review paper, introduction of SSRs marker and development of SSRs markers by bioinformatics tools are discussed.

Key words : Bioinformatics tools, expressed sequence tags, genetic diversity, polymorphism and simple sequence repeats markers.

Introduction

Marker is a gene and linked with a specific trait. The markers used in genetics and plant breeding can be classified into 3 different types of genetic markers such as morphological markers, biochemical markers and molecular markers. Morphological markers are characterized as observable traits such as seed size and flower color etc (Sumarani *et al*., 2004) and biochemical markers are protein produced by expression of genes that are identified by definite staining and electrophoresis methods (Pillai *et al*., 2000) such as alloenzymes etc. These two markers have several disadvantages such as they are limited in number and influenced by environmental factors (Varshney *et al*., 2005). The molecular marker (genetic marker) is a tiny part of DNA sequence showing polymorphism between dissimilar individuals (Ghori *et al*., 2015). DNA-based molecular markers as versatile tools and have applied in area such as plant breeding, genetic engineering and taxonomy etc (Joshi *et al*., 2011) and it is also a very potent genomic tools to raise the efficiency and accuracy of breeding methods for crop improvement (Varshney *et al*., 2012). Several molecular markers such as Restriction Fragment Length Polymorphisms (RFLP), Random Amplification of Polymorphic DNAs (RAPD), Sequence Tagged Sites (STS), Amplified Fragment Length Polymorphisms (AFLP), Simple Sequence Repeats (SSR) and Single Nucleotide Polymorphism (SNP) have been introduced and deployed in various ways in several plant breeding programmes. The various types of available DNA markers can be classified into three categories:

First generation DNA markers (Hybridization based)

The first generation DNA marker system is based on southern blotting techniques. RFLP (Restriction Fragment Length Polymorphism) is an example of the first generation DNA markers. RFLP technique based on hybridization of genomic DNA digested with restriction enzymes and DNA probes. Grodzicker *et al*. (1975) used for the first time to identify DNA sequence polymorphism for genetic mapping of a temperature sensitive mutation of adenovirus serotypes. Botstein *et al*. (1980) used RFLP for human genome mapping and afterward RFLP used for plant genomes (Helentjaris *et al*., 1986; Weber and

***Author for correspondence:** E-mail: rajendrak64@yahoo.co.in

Helentjaris, 1989).

Second generation (PCR based markers)

The second generation DNA marker is based on polymerase chain reaction (PCR). PCR is technique used for amplification of small DNA fragments. It is widely used because it requires a very small amount of DNA to analysis. All PCR based markers depend on the PCR primers which bind to specific sites in the genome.

Third generation DNA markers (DNA sequence based)

Currently, third generation DNA markers is cheap, non gel-based assays with high throughput detection systems. SNPs (Single Nucleotide Polymorphism) are an example of third generation DNA markers.

Distributions and Occurrences

The SSRs are distributed all over the genome as compared to other molecular markers (GousMiah *et al*., 2013). Microsatellites are present in the coding region as well as noncoding region (Tautz and Renz, 1984; Gupta *et al*., 1996; Toth *et al*., 2000). Coding region contain DNA sequence that encodes protein by expression of genes. Noncoding DNA sequences do not encode any protein sequences, but few noncoding DNA is transcribed into transfer RNA (t-RNA), regulatory RNAs and ribosomal RNA (r-RNA). The amount of total genomic DNA and proportion of coding and noncoding DNA varies between organisms. Approximately 85–90% region of prokaryotes genome contains non-repetitive (Koonin and Wolf, 2010) and 20% of a prokaryote genome is noncoding (Fabrico, 2012). Eukaryotes genome such as plants and mammalians, its most of genome contain repetitive DNA (Lewin, 2004) in case of human, over 98% of the genome is noncoding DNA (Elgar *et al*., 2008). Microsatellites are more abundant and longer in vertebrates as compared to invertebrates. Longer microsatellites are present in cold-blooded species (vertebrates) (Chambers and MacAvoy, 2000). Toth *et al*. (2000) studied on various taxa and found that while minimum number of microsatellites was exhibited by *C. elegans* and maximum number of microsatellites was exhibited by rodents. In the genomes of several organisms, repeated analysis of microsatellite frequency has revealed that occurrence of microsatellites are comparatively low in prokaryotes, bats, lepidopterans and birds but in most mammals and fishes have a high frequency of repeat motifs.

Role of simple sequence repeats (SSRs) varies in coding region and non cooding region. The SSRs located in a coding region can affect the activation of a gene and consequently, the expression of a protein. If the SSRs located in a noncoding region, it may effect on gene regulation (Lawson and Zhang, 2006). The untranslated region (UTRs) regions of different organisms contain different repeat of SSRs. Occurrence of SSRs in the 5' - UTRs are essential for expression of a few genes. Occurrence of SSRs in 3'–UTRs cause transcription slippage. Exon is the functional part of the mRNA that encodes expression of proteins. In many species, exons contain more tri and hexanucletide motifs and rarely contain di and tetranucleotides motifs. In human trinucleotides repeat reveal about a two times greater occurrence in exonic and intergenic region in all chromosomes except Y chromosome (Subramanian *et al*., 2003). SSRs motifs vary from species to species and within same species. In case of animal genome dinucleotides motifs (CA)n were normally found while dinucletodies motifs (AT)n were rarely found (Moore *et al*., 1991). In case of plant genome dinucleotides repeat are more common while mono and tetranucleotides repeats are lesser (Wang *et al*., 1994; Schug *et al*., 1998). In several species, dinucleotide repeats of SSRs are limited in coding region as compared to non-coding region (Li *et al*., 2002) whereas trinucleotides are found more abundant in the coding regions of the genome (Toth *et al*., 2000). Varshney *et al*. (2002) recognized trinucleotide repeats of SSRs were the most frequent found as compared to dinucleotides repeats and tetranucleotides repeats in cereals.

Types of Simple sequence repeat (SSRs)

Microsatellites arise in regions consisting of short numbers of tandem repeat of DNA sequences in genome and reveal extreme polymorphism (Shamjana *et al.,* 2015). Microsatellites have been classified into three classes depending upon their occurrence and source for development. First is Genomic (gSSRs), which isolated from the nuclear genome. Second is EST or genic microsatellites (EST-SSRs), which developed by exploiting EST sequences. Third is organellar microsatellite which developed from the chloroplast (cpSSRs) and mitochondrial (mtSSRs) genome of an organism (Siju *et al*., 2014). The number of repeats is characteristics of an allelic deviation at certain locus. The number of repeat depends on the type and of the size of the motifs. Based on type of repeat sequence, SSRs are classified into four categories (Oliveira *et al*., 2006). First is perfect microsatellite in which the repeat sequence is continuous and is not broken up by any base not belonging to the motif. Second is imperfect microsatellite in which a pair of bases is present between the repeat motif that does not match the motif sequence. Third is interrupted

microsatellite in which a few sequence inside the repeated sequence that does not match the motif sequence. Forth is compound microsatellite is also called as composite microsatellites in which two neighboring distinctive repeats present in the sequence. Based on the length of repeat motif, SSRs are classified into two categories. Class I microsatellites is perfect SSRs of ≥20 nucleotides in length. Class II microsatellites is perfect SSRs of ≥12 nucleotides and ≤20 nucleotides in length (Temnykh *et al*., 2001).

Tools to search for SSR in genome

The *in silico* tools required for searching SSRs from sequences have become efficient and inexpensive alternative for plant species. Bioinformatics tools, which detected SSR repeats and developed a PCR-based SSR markers are listed below:

Sputnik tool

Sputnik tool is fast, simple and easy to use. This tool searches DNA sequence files in FASTA format for SSRs (Abajian, 1994). Sputnik is used to search for repeated sequences of nucleotides of length between 2 and 5. It searches perfect, imperfect and compound repeats of SSRs. Sputnik has useful for SSRs identification in various species such as barley and *Arabidopsis* (Cardle *et al*., 2000). Disadvantage of sputnik is that it cannot identify mononucleotide repeats and at present it is not supported by a web interface (Duran *et al*., 2009).

Repeat finder

Repeat Finder tool used for finding of SSRs from small to medium datasets in which DNA sequences file used in FASTA format. It identify perfect, compound and imperfect repeat. Repeat Finder has been used for identifying SSRs in peanut (Jayashree *et al*., 2005).

SSR Locator

This tool is newly developed and it used as recognition and characterization of SSRs. Victoria *et al*. (2011) applied SSR Locator to reading the pattern of expressed sequence tags derived SSR markers for model plants.

SSRIT

Simple Sequence Repeats Identification Tool is used for the identification of perfect simple sequence repeats (SSRs). The outcome obtained in tabular format. Its output contain no. of repeats, SSR start and end, motif (repeat) type and sequence ID. SSRIT has been applied in rice for identification of SSRs (Temnykh *et al*., 2001). Singh *et al.* (2011) applied this tool to mine SSRs in wheat rust *Puccinia*sp. Kantety *et al*. (2002) applied SSRIT tool to mine SSR in ESTs from sorghum, maize, rice, wheat and barley.

TRF Tool

Tandem repeat finder can find large number of SSRs repeats (approximately 2,000 base pairs). TRF finds perfect, compound and imperfect repeats and has been used in cowpea for identification of SSRs (Chen *et al*., 2007).

MISA

MISA detects perfect, interrupted and compound SSRs. MISA can also design PCR amplification primers on either side of each SSR sequences. MISA has been used in wheat (Yu *et al*., 2004), peanut (Liang *et al*., 2009), barley (Thiel *et al*., 2003; Kota *et al*., 2001b) and rye (Khlestkina *et al*., 2004) for identification of SSR.

TROLL

It is Simple Sequence Repeat (SSR) finder based on a slight modification of the Aho-Corasick algorithm. It requires a standard personal computer (PC) to operate. It have used for SSRs identification in *Arabidopsis* genome (Duran *et al*., 2009).

PolySSR (Tang *et al.*, 2008)

It is tool used to identify polymorphic SSRs rather than just SSRs. It finds polymorphic short sequence repeats from EST sequences available on public databases. Poly SSR has been used in potato, tomato, rice, *Arabidopsis*, *Brassica* and chicken for identification of SSR.

Conclusion and Future Prospects

At present, several molecular markers are reported and effectively used in plant breeding programs. Amongst all molecular markers, SSRs marker is ideal marker due to their valuable characteristics like highly polymorphic nature, abundance, genome-wide distribution and co-dominanceetc. Therefore, it is widely used for identification of alleles linked with disease and also for physical mapping, genetic mapping, association mapping and comparative mapping, genetic diversity, QTLs analysis, marker assisted selection, study of evolutionary relationship and mapping of desired genes.The co-dominant nature of SSRs is appropriate for genetical analysis in segregating F_2 population. Hyper-variability nature of SSRs shows very high allelic variations even among very closely related varieties. The high reproducibility nature of SSR would be an important in genetic study. Reproducibility of the SSR profile is as strong as it is with RFLPs. Experimental measures for SSR analysis is easy and it requires only a small amount of DNA.

Acknowledgments

We gratefully acknowledge the financial support provided by U.P. Council of Agricultural Research, Lucknow (U.P.), India.

References

Abajian, C. (1994). Sputnik: (http://espressosoftware.com/sputnik /index.html).

Botstein, D., R. L. White, M. Skolnick and R. W. Davis (1980). Construction of a genetic map in man using restriction fragment length polymorphisms. *Amer J Hum Genet.,* **32** : 314–331.

Cardle, L., L. Ramsay, D. Milbourne, M. Macaulay, D. Marshall and R. Waugh (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genet*., **156** : 847-854.

Chambers, G. K. and E. S. MacAvoy (2000). Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, **126** : 455– 476.

Chen, X. F., T. W. Laudeman, P. J. Rushton, T. A. Spraggins and M. P. Timko (2007). CGKB: an annotation knowledge base for cowpea (*Vigna unguiculata* L.) methylation filtered genomic gene space sequences. *BMC Bioinf*, **8** : 112-116.

Duran, C., D. Edwards and J. Batley (2009). *Molecular Marker Discovery and Genetic map Visualization*. In: Applied Bioinformatics (Eds. Edwards D, Hanson D and Stajich J), Springer (USA). 165-189.

Elgar, G. and T. Vavouri (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.*, **24(7)** : 344–52.

Fabrico, C. (2012). 7 Non-coding RNAs, Epigenomics, and Complexity in Human Cells. In Morris, Kevin V. Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection.

Gous, M., M. Y. Rafii, M. R. Ismail, A. B. Puteh, H. A. Rahim, K. N. Islamand M. A. Latif (2013). A Review of Microsatellite Markers and Their Applications in Rice Breeding Programs to Improve Blast Disease Resistance. *Int. J. Mol.Sci.*, **14(11)** : 22499–22528.

Grodzicker, T., J. Williams, P. Sharp and J. Sambrook (1975). Physical mapping of temperature sensitive mutants of adenovirus. *Cold Spring Harbor Symposia on Quantitative Biology,* **39** : 439-446.

Gupta, P. K., H. S. Balyan, P. C. Sharma and B. Ramesh (1996). Microsatellites in plants : a new class of molecular markers. *Curr. Sci.*, **70** : 45-54.

Helentjaris, T., M. Slocum, S. Wright, A. Schaefer and J. Nienhuis (1986). Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. *Theoretical and Applied Genetics*, **61** : 650-658.

Jayashree, B., M. Ferguson, D. Ilut and J. H. Doyle and Crouch (2005). Analysis of genomic sequences from peanut (*Arachishypogaea*). *Electron. J. Biotech.*, **8** : 3.

Joshi, S. P., K. Prabhakar, P. K. Ranjekar and V. S. Gupta (2011). *Molecular markers in plant genome analysis*. **1**-19.

Kantety, R. V., M. L. Rota, D. E. Matthews and M. E. Sorrells (2002). Data mining for simple sequence repeats in expressed sequence tags from barely, maize, rice, sorghum and wheat. *Plant Mol. Biol.*, **48** : 501-510.

Khlestkina, E. K., M. H. M. Than, E. G. Pestsova, M. S. Röder, Malyshev, S.V. Korzun and A. Börner (2004). Mapping of 99 new microsatellite- derived loci in rye (*Secalecereale* L.) including 39 expressed sequence tags. *Theor. Appl. Genet.,* **109** : 725-732.

Koonin, E.V. and Y. I. Wolf (2010). Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics*, **11(7)** : 487–498.

Kota, R., R. K. Varshney, T. Thiel, K. J. Dehmer and A. Graner (2001). Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). *Hereditas,* **135** : 145-151.

Lawson, M. J. and L. Zhang (2006). Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol*., **7** : R14.1-11.

Lewin, Benjamin (2004). Genes VIII (8th ed.) Upper Saddle River, NJ : Pearson/Prentice Hall. ISBN 0-13-143981-2.

Li, Y. C., A. B. Korol, T. Fahima, A. Beiles and E. Nevo (2002). Microsatellites : genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.*, **11** : 2453-2465.

Liang, X., X. Chen, Y. Hong, H. Liu, G. Zhou, S. Li and B. Guo (2009). Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and Arachis wild species. *BMC Plant Biol.*, **9** : 35.

Moore, S. S., L. L. Sargeant, T. J. King, J. S. Mattick, M. George and J. S. Hetzel (1991). The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in closely related species. *Genomics*, **10** : 654-660.

Oliveira, E. J., J. G. Paìdua, M. I. Zucchi, R. Vencovsky and M. L. C. Vieira (2006). Origin, evolution and genome distribution of microsatellites. *Genet. Mol. Res.*, **29** : 294-307.

Pillai, S. V., P. Sundaresan, Harisankar and G. O. Sumarani (2000). *Molecular characterization of germplasm in tropical tuber crops*, DAE-BRNS Symposium, Mumbai.

Schug, M. D., K. A. Wetterstrand, M. S. Gaudette, R. H. Lim, C. M. Hutter and C. F. Aquadro (1998). The distribution and frequency of microsatellite loci in *Drosophila melanogaster*. *Mol. Ecol.,* **7(1)** : 57-70.

Shamjana, U., T. Bharadwaj and T. Grace (2015). Microsatellites: A versatile marker for genetic/evolutionary/ecological studies. *International Journal of Advanced Biological Research*, **5(2)** : 86-95.

Siju, S., K. Dhanya, S. Bhaskaran and S. Thotten Elampilay (2014). Methods for Development of Microsatellite Markers : An Overview. *Not SciBiol.*, **6(1)** : 1-13.

Singh, R., B. Pandey, M. Danishuddin, S. Sheoran, P. Sharma and R. Chatrath (2011). Mining and survey of simple sequence repeats in wheat rust *Puccinia* sp. *Bioinformation*, **7(6)** : 291–295.

Subramanian, S., R. M. Mishra and L. Singh (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biology*, **4(2)** : R13.1-10.

Sumarani, G. O., S. V. Pillai, P. Harisankar and S. Sundaresan (2004). Isozyme analysis of indigenous cassava germplasm for identification of duplicates. *Genetic Resources and Crop Evolution.,* **51** : 205-209.

Tang, J., J. A. M. Leunissen, R. E. Voorrips, C. G. Linden and B. Vosman (2008). HaploSNPer : a web-based allele and SNP detection tool. *BMC Genet.*, **9** : 23.

Tautz, D. and M. Renz (1984). Simple sequence repeats are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res.*, **12** : 4127-4138.

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

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

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

Varshney, R. K., H. Kudapa, M. Roorkiwal, M. Thudi, M.K. Pandey, R. K. Saxena and S. K. Chamarthi *et al.* (2012). Advances in genetics and molecular breeding of three legume crops of semi-arid tropics using next-generation sequencing and high-throughput genotyping technologies. *J. Biosci.*, **37(5)** : 811–820.

Varshney, R. K., R. Sigmund, A. Boerner, V. Korzun, N. Stein, M. Sorrells, P. Langridge and A. Graner (2005). Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Science*, **168** : 195-202.

Varshney, R. K., T. Thiel and N. Stein *et al.* (2002). *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell. Mol. Biol. Lett.*, **7** : 537–546.

Victoria, F. C., L. C. Maia and A. C. Oliveira (2011). *In silico* comparative analysis of SSR markers in plants. *BMC Plant Biol.*, **11** : 15.

Wang, Z., J. L. Weber, G. Zhong and S. D. Tanksley (1994). Survey of plant short tandem DNA repeats. *Theor. Appl. Genet*., **88** : 1-6.

Weber, D. and Helentjaris (1989). Mapping RFLP loci in Maize using B-A translocations. *Genetics*, **121** : 583-590.

Yu, J. K., T. M. Dake, S. Singh, D. Benscher, W. Li, B. S. Gill and M. E. Sorrells (2004). Development and mapping of EST-derived simple sequence repeat (SSR) markers for hexaploid wheat. *Genome,* **47** : 805-818.