Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

SSRscanner: a program for reporting distribution and exact location of simple sequence repeats

Tamanna Anwar[1] and Asad U Khan[1,2]*

[1]Distribution Information Sub-centre, [2]Interdisciplinary Biotechnology Unit Aligarh Muslim University, Aligarh -202002, India; Asad U Khan* - Email: huzzi99@hotmail.com; Phone: +91-571-2723088; Fax: +91-571-2721776;

* Corresponding author

received February 06 2006; accepted February 12 2006; published online February 20, 2006

Abstract:

Simple sequence repeats (SSRs) have become important molecular markers for a broad range of applications, such as genome mapping and characterization, phenotype mapping, marker assisted selection of crop plants and a range of molecular ecology and diversity studies. These repeated DNA sequences are found in both prokaryotes and eukaryotes. They are distributed almost at random throughout the genome, ranging from mononucleotide to trinucleotide repeats. They are also found at longer lengths (> 6 repeating units) of tracts. Most of the computer programs that find SSRs do not report its exact position. A computer program SSRscanner was written to find out distribution, frequency and exact location of each SSR in the genome. SSRscanner is user friendly. It can search repeats of any length and produce outputs with their exact position on chromosome and their frequency of occurrence in the sequence.

Availability:

This program has been written in PERL and is freely available for non-commercial users by request from the authors. Please contact the authors by E-mail: huzzi99@hotmail.com

Keywords: scanner; SSR; repeats; script; microsatellite

Background:

SSRs (simple sequence repeats) or microsatellites are the genetic loci where one or few bases are tandemly repeated for varying numbers of times. Such repetitions occur primarily due to slipped-strand mis-pairing and subsequent error(s) during DNA replication, repair, or recombination. **[1]** SSRs comprising 1–6 bp long, occur frequently and are ubiquitously interspersed in many genomes. **[2, 3]** The biological importance of SSR tracts has been clearly delineated. Microsatellite loci show extensive length polymorphism, and hence they are widely used in DNA fingerprinting and diversity studies. They are also considered as ideal genetic markers for the construction of high-density linkage maps. **[4, 5]** In spite of its high significance, a bioinformatics tool for the analysis of these regions is not available.

Available algorithms directly or indirectly detect tandem repeats. However, there are many limitations with these algorithms. The drawbacks are high computational time required by the algorithm and their inability to predict the positions of SSRs in the genome. The program Tandem Repeats Finder **[6]** locates repeats with motifs of any size and type, including repeats with insertions and deletions. The program Sputnik **[7]** (unpublished) uses recursion to search for both exact and approximate tandem repeats. Repeating unit lengths of 2 to 5 are sought, and a score is used to determine location. Tandem Repeat Occurrence Locator (TROLL) **[8]**, uses a keyword tree adapted from bibliographic searching techniques and attempts to match the keywords exactly but it does not specify the positions of repeats. In this work, we describe a program called SSRscanner (Simple sequence repeat scanner) that uses dictionary approach to find simple sequence repeats of pre-selected motifs.

SSRscanner is a PERL script developed for scanning genomes to find repeats of any length, their exact position on chromosome and frequency of occurrence. It is fast and requires a standard Personal Computer (PC) and PERL to operate. SSRscanner can accept large sequences as input and large number of motifs can be searched simultaneously. Thus, the running time of the program is greatly reduced. To demonstrate the use of SSRscanner, *Arabidopsis thaliana* genome was analyzed for finding out distribution, frequency and specific position of SSRs in the genome. **[9]** The advantages over many other programs developed for SSR identification includes its ability to search motifs of any length repeated for a number of times and to give the exact position of the motif in the genome.

Methodology:

Program Input:

SSRscanner (implemented in PERL) accepts two text files (.txt). Upon execution of the program, it prompts to enter the file name containing the DNA sequence data. It also prompts for a file containing motifs of different repeat types. It then prompts for the number of times for the motifs to be repeated (for example for searching tri-nucleotide repeats the user should enter 3) (Figure 1B).

Input sequence and motifs are parsed to SSRscanner that extracts the SSRs giving their distribution/frequency and specific location. The results from SSRscanner are appended into result files (Figure 1A).

Program output: SSRscanner gives two files in output (Figure 1A). They are (1) Motifposition.txt (gives the frequency of each repeat provided in the motif file) and (2) Motifresult.exe (gives the specific location of each repeat). The data obtained can then be arranged into desired formats.

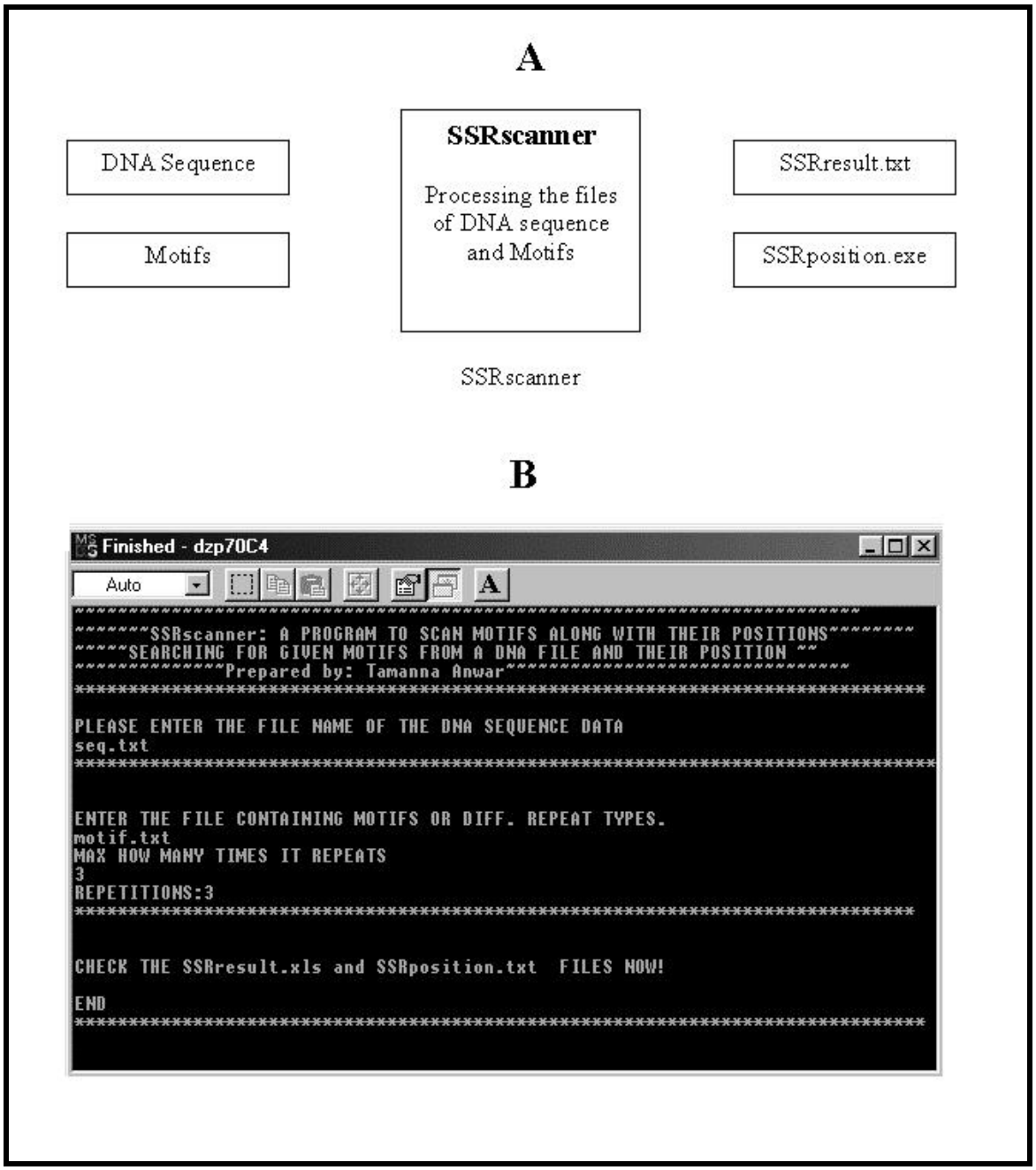

Figure 1: (A) Overview of SSRscanner operational structure; (B) SSRscanner command line showing the executed program.

Caveats and Future Development:
SSRscanner is a PERL script and it requires PERL to be installed on a PC before running the program. We are developing a web based CGI for SSRscanner.

Acknowledgement:
The authors are grateful to Professor M. Saleemuddin for providing facilities to carryout this work and for his support throughout this project. We also thanks to the Staff of the Distribution information sub-center for their technical help. Department of Biotechnology, Ministry of Science and Technology, Government of India is acknowledged for the financial support.

References:

[1] G. Levinson & G. A. Gutman, *Mol. Biol. Evol*., 4: 203 (1987) [PMID: 3328815]

[2] R. Gur-Arie, *et al*., *Genome Res.*, 10:62 (2000) [PMID: 10645951]

[3] G. Toth, *et al*., *Genome Res.*, 10:967 (2000) [PMID: 10899146]

[4] J. S. Beckmann & M. Soller, *Biotechnology*, 8: 930 (1990) [PMID: 1366775]

[5] M. Morgante & A. M. Olivieri, *Plant J.*, 3:175 (1993) [PMID: 8401603]

[6] G. Benson, *Nucleic Acids Res.*, 27:573 (1999) [PMID: 9862982]

[7] Sputnik [http://abajian.net/sputnik/]

[8] A. Castelo, *et al*., *Bioinformatics*, 18:634 (2002) [PMID: 12016062]

[9] T. Anwar & A. U. Khan, *Bioinformation*, 1:64 (2005) [Abstract]

Edited by P. Kangueane

Citation: Anwar & Khan, Bioinformation 1(3): 89-91 (2006)