BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 6(4)

MfSAT: Detect simple sequence repeats in viral genomes

Ming Chen[2, 3], Zhongyang Tan[1]*, Guangming Zeng[2, 3]

[1]College of Biology, State Key Laboratary for Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082, China; [2]College of Environmental Science and Engineering, Hunan University, Changsha 410082, China; [3]Key Laboratory of Environmental Biology and Pollution Control (Hunan University), Ministry of Education, Changsha 410082, China; Zhongyang Tan - Email: zhongyang@hnu.cn; *Corresponding author

Received April 05, 2011; Accepted April 25, 2011; Published May 07, 2011

Abstract:
Simple sequence repeats (SSRs) are ubiquitous short tandem repeats, which are associated with various regulatory mechanisms and have been found in viral genomes. Herein, we develop MfSAT (Multi-functional SSRs Analytical Tool), a new powerful tool which can fast identify SSRs in multiple short viral genomes and then automatically calculate the numbers and proportions of various SSR types (mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats). Furthermore, it also can detect codon repeats and report the corresponding amino acid.

Keywords: comparative genomics, simple sequence repeat, software, microsatellite, codon repeat.

Background:

Simple sequence repeats (SSRs) or microsatellites are tandemly repeated tracts consisting of 1-6 base pair (bp) long units **[1, 2]**. Comprehensive analysis of SSRs in 8619 pre-miRNAs indicates SSRs are widely present in these very small non-coding RNA sequences **[3]**. It has been demonstrated that SSRs can affect gene expression and the corresponding gene products and even cause phenotypic changes or diseases **[4, 5]**. Correspondingly, computational tools for detection of SSRs and their related information from whole genome sequences are increasing as well **[6]**. The growing number of analytical tools for SSRs has greatly assisted the understanding of SSRs at the genome-wide level. Our examination of the available tools reveals certain faults. In order to efficiently screen viral genome sequences for SSRs, we have developed a new tool called MfSAT.

Methodology:

Consider a sequence or multiple sequences over a finite alphabet {(a, t, g, c) or (a, u, g, c)}. A tract at a given locus will be defined as a microsatellite if that tract can be expressed as a tandem repeat of a motif of 1–6 bp size **[6]**. Our goal is to efficiently detect SSRs in a sequence or multiple sequences given an arbitrary motif size or minimum repeat number. The proposed algorithm has two parameters, maximum motif and minimum repeat number which are independent. When you run according to the first parameter, the minimum number is three, whereas if you run by use of another parameter, the maximum motif is "hexa". If users select the "Hexa→mono" tag, MfSAT progressively scans for nucleation sites starting from hexanucleotide repeat to mononucleotide repeat at a given locus. If no hexanucleotide repeat tract is detected, then pentanucleotide repeat nucleation site will be searched for and so on. This algorithm is the same with IMEx **[6, 7]**. However, if users select another tag, "Mono→hexa", in contrast to above step, in this section we assume the algorithm advances the shortest repeats. Given a candidate trinucleotide repeat motif k and its starting position j together with the starting position d of coding sequence of analyzed genome sequences, the verification formula determines whether an SSR is a codon repeat. The formula is as follows:

$S = (j\text{-}d)/3 \qquad (1)$

If S is an integer, the trinucleotide repeat is a codon repeats. It remains to judge what its corresponding amino acid is.

Software Requirements:

MfSAT can be used in any computer with windows system.

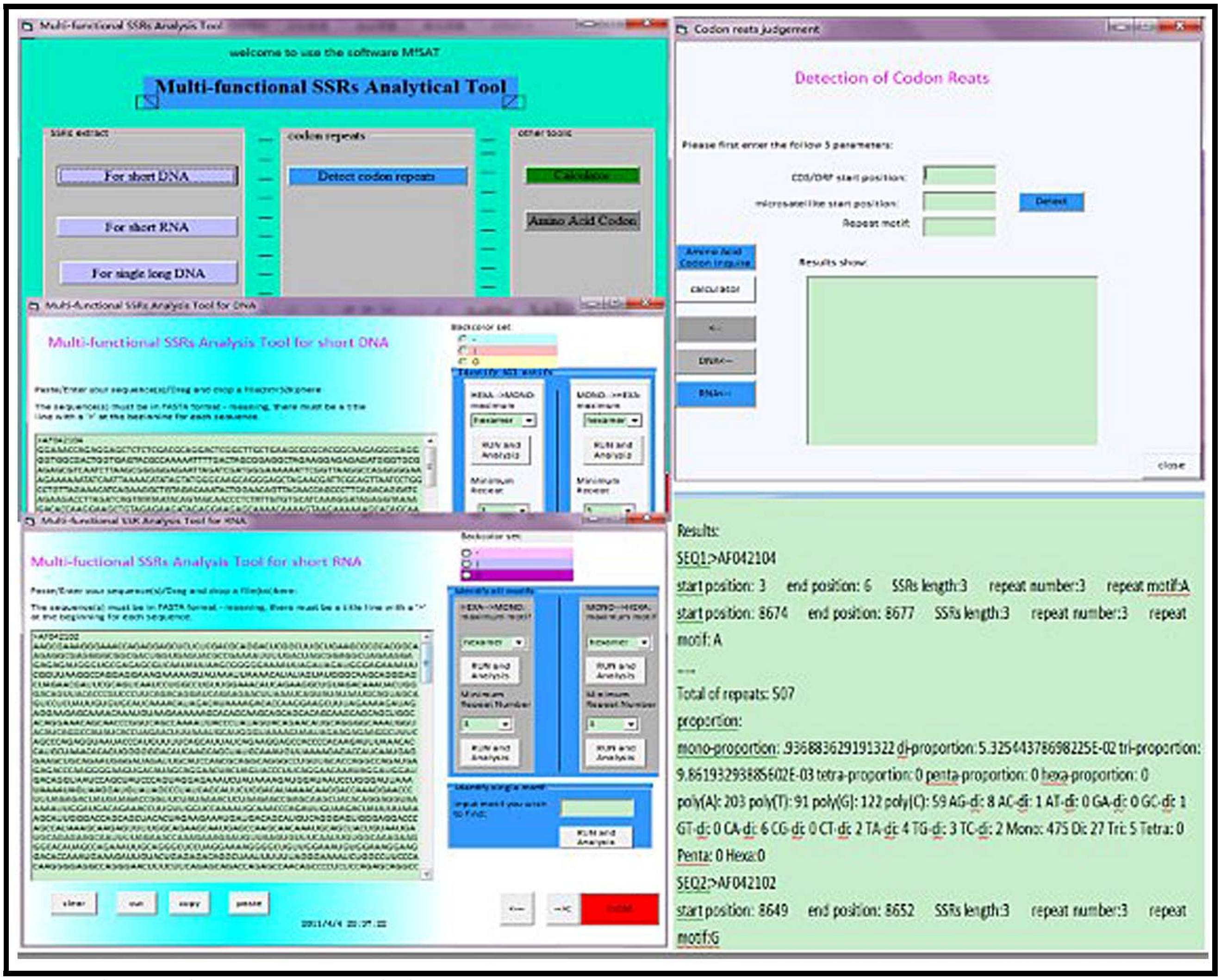
Input:

MfSAT uses a advanced and power algorithm 'regular expressions' to screen one or multiple viral DNA/RNA sequences in fast format for SSRs and reports the motif, repeat number, genomic location, abundance of each of six classes SSRs and many other features useful for SSRs' studies.

Output:

We have developed a new tool that can be successfully used to identify SSRs in viral genomes consisting of viral DNA or RNA sequences for escaping statistical troubles. Judging according to its performance, MfSAT is a definite advance compared to other available tools. A stand-alone software with several videos is available online at http://hudacm11.mysinamail.com/hunan.html. This tool is also available from authors Zhongyang Tan and Guangming Zeng on request (zhongyang@hnu.cn; zgming@hnu.cn). The output is composed of three parts: the first part consists of a list of SSRs, each with information such as repeat motif content, repeat number, starting position, end position, SSR length; the second part is the numbers of proportions of each of the six classes of SSRs (mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats); the third part comprises the numbers of poly(A), poly (T/U), poly(G), poly(C), and 12 classes of dinucleotide repeats including AG, GA, GT (GU), TG (UG), AC, CA, CT (CU), TC (UC), AT (AU), TA (UA), GC and CG repeats. It is clear from the results that MfSAT is more attractive in terms of consideration. **Figure 1** shows the software interface and output results of MfSAT.

ISSN 0973-2063 (online) 0973-8894 (print)



```
Results:
SEQ1:>AF042104
start position: 3    end position: 6    SSRs length:3    repeat number:3    repeat motif:A
start position: 8674    end position: 8677    SSRs length:3    repeat number:3    repeat
motif: A
----
Total of repeats: 507
proportion:
mono-proportion: .936883629191322 di-proportion: 5.32544378698225E-02 tri-proportion:
9.86193293885602E-03 tetra-proportion: 0 penta-proportion: 0 hexa-proportion: 0
poly(A): 203 poly(T): 91 poly(G): 122 poly(C): 59 AG-di: 8 AC-di: 1 AT-di: 0 GA-di: 0 GC-di: 1
GT-di: 0 CA-di: 6 CG-di: 0 CT-di: 2 TA-di: 4 TG-di: 3 TC-di: 2 Mono: 475 Di: 27 Tri: 5 Tetra: 0
Penta: 0 Hexa:0
SEQ2:>AF042102
start position: 8649    end position: 8652    SSRs length:3    repeat number:3    repeat
motif:G
```

Figure 1: Software Interface and Output Results of MfSAT.

Future Work:

Development of a linux version of MfSAT is in process.

Acknowledgments:

The authors sincerely thank Editor and anonymous reviewer for suggestions on improving the paper. The study was financially supported by Production, Education and Research guiding project, Guangdong Province (2010B090400439), Great program for GMO, Ministry of Agriculture of the people Republic of China (2009ZX08015-003A), the National Natural Science Foundation of China (No. 50608029, No.50978088, No. 50808073, No.51039001), Hunan Provincial Innovation Foundation for Postgraduate, the National Basic Research Program (973 Program) (No. 2005CB724203), Program for Changjiang Scholars and Innovative Research Team in University (IRT0719), the Hunan Provincial Natural Science Foundation of China (10JJ7005), the Hunan Key Scientific Research Project (2009FJ1010), and Hunan Provincial Innovation Foundation For Postgraduate (CX2010B157).

References:

[1] Chen M *et al. FEBS Lett*. 2009 583: 2959 [PMID: 19679131]
[2] Chen M *et al. FEBS Lett*. 2011 585: 1072 [PMID: 21382371]
[3] Chen M *et al. Mol Biol Evol*. 2010 27: 2227 [PMID: 20395311]
[4] Usdin K. *Genome Res*. 2008 18: 1011 [PMID: 18593815]
[5] Li YC *et al. Mol Biol Evol*. 2004 21: 991 [PMID: 14963101]
[6] Mudunuri SB & Nagarajaram HA. *Bioinformatics* 2007 23: 1181 [PMID: 17379689]
[7] Mudunuri SB *et al. Bioinformation* 2010 5: 221 [PMID: 21364802]

Edited by P Kangueane
Citation: Chen *et al.* Bioinformation 6(4): 171-172 (2011)