BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 13 2007, pages 1683–1685
doi:10.1093/bioinformatics/btm157

Genome analysis

SciRoKo: a new tool for whole genome microsatellite search and investigation

Robert Kofler[1,*], Christian Schlötterer[2,3] and Tamas Lelley[1]

[1]University of Natural Resources and Applied Life Sciences, Department for Agrobiotechnology IFA-Tulln, Institute of Biotechnology in Plant Production, Konrad Lonrenz Straße 20, 3430 Tulln, [2]Institut für Ökologie, Universität Innsbruck, Technikerstraße 25, 6020 Innsbruck, Austria and [3]Institut für Tierzucht and Genetik, Veterinärmedizinische Universitat Wien, Josef Baumann Gasse 1, 1210 Wien, Austria

Received on February 22, 2007; revised on April 12, 2007; accepted on April 17, 2007

Advance Access publication April 26, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: SciRoKo is a user-friendly software tool for the identification of microsatellites in genomic sequences. The combination of an extremely fast search algorithm with a built-in summary statistic tool makes SciRoKo an excellent tool for full genome analysis. Compared to other already existing tools, SciRoKo also allows the analysis of compound microsatellites.

Availability: free for use: www.kofler.or.at/Bioinformatics

Contact: robert.kofler@boku.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Simple sequence repeats (SSRs) are known for more than 20 years, but their evolution is still not fully understood [for reviews see: Buschiazzo and Gemmell (2006); Ellegren (2004); Schlötterer (2000)]. Little is known about compound microsatellites. It is assumed that they account for 10% of all microsatellites (Weber, 1990) and that some compound microsatellites are involved in transcription fine regulation (Kashi and King, 2006).

Several bioinformatic tools for the identification of microsatellites in genomic sequences have been developed [see also Varshney *et al*. (2002) and Table S1 for a comparison]. The most commonly used tools for SSR search are: MISA (Thiel *et al*., 2003), SSRFinder (Gao *et al*., 2003), SSRIT (Temnykh *et al*., 2001), TRF (Benson, 1999), TROLL Castelo *et al*. (2002), Sputnik (http://espressosoftware.com/pages/sputnik.jsp), Modified Sputnik I (Morgante *et al*., 2002) and (Modified Sputnik II (La Rota *et al*., 2005).

Despite the large number of available tools, none combines a user-friendly interface with a statistical analysis of genomic microsatellites. Here, we present a novel software for the identification of SSRs in genomic sequences that fills this gap.

SciRoKo (SSR Classification and Investigation by Robert Kofler) contains two main modules: an SSR search module, which supports five different SSR search modes and a module for SSR-statistics, notably for mismatch frequency and compound microsatellite analysis. The program is considerably faster than the presently available tools allowing search for imperfect microsatellites (Table 1).

2 THE SCIROKO SSR-SEARCH MODULE

The SciRoKo SSR-search module is based on a scoring system, which considers the length of a microsatellite. Since a previous study identified this characteristic as to be the most informative variable describing microsatellites (Dieringer and Schlötterer, 2003). Five search modes are available, three for perfect and two for mismatched SSR search.

In the three perfect SSR search modes, a nucleotide at position i is tested for identity with the nucleotide at position $i+t$, where t is the motif length (1–6). Upon identity i is increased $i=i+1$ until no further identity can be found. If an identified SSR meets the specified minimum length (score), the SSR is saved to the output file. The three different perfect SSR search modes are based either on a specified minimum number of repeats or a minimum length in bp.

In the two mismatched SSR search modes, perfect SSRs (SSR-seeds) act as origin for subsequent 5′ and 3′ extensions. The perfect microsatellite seed is extended in the 5′ and 3′ direction by allowing for mutations (indels or base substitutions) in the microsatellite. The SciRoKo scores are calculated according to the two equations:

$$S_{\text{fixP}} = \text{hits} - mm\text{P} * mm \quad (1)$$

$$S_{\text{varP}} = \text{hits} - mm * (m_{\text{L}} * mm\text{P}) \quad (2)$$

The parameters are: hits (identity with a virtual perfect microsatellite, see Manual: File S8), number of mismatches (mm), mismatch penalty (mmP) and the length of the SSR motif (m_{L}). Equation (1) is used in the 'Fixed mismatch penalty' mode, Equation (2) in the 'Variable mismatch penalty' mode.

The SSR seed is iteratively extended in both directions until the score decreases. If the maximum obtained score exceeds a defined threshold, the microsatellite is saved.

The benchmark results (Table 1) clearly demonstrate that this algorithm is much faster than the currently available search

*To whom correspondence should be addressed.

tools, allowing imperfections in the microsatellite repeat. We used an elaborate set of test SSRs (File S1) to asses the SciRoKo SSR search module. Additionally, we compared the SSR-search results for *S.cerevisiae* obtained by using Modified Sputnik I–II with the search results from SciRoKo. Details of this search results as well as the summary statistics are available as Supplementary Material (Files S2–S5).

3 THE SCIROKO SSR-STATISTICS MODULE

The SciRoKo SSR-statistics module provides three different SSR classification and statistics options. An SSR motif length statistic, an SSR motif statistic and a motif association (compound SSR) statistic. SciRoKo combines microsatellite motifs that are related to each other (e.g.: AT and TA, File S6 and S7; full and partial standardization: Table S2)

3.1 Motif length statistics

The motif length statistics provides detailed information for mono-, di-, tri-, tetra-, penta- and hexanucleotide SSRs, i.e. total counts, average number of mismatches, average length of the SSRs and counts per Mb.

3.2 Motif statistics

Motif statistics provides information for each fully and partially standardized SSR motif, i.e. total counts, average number of mismatches, average length, counts per Mb and GC content.

3.3 Motif association statistics

Frequently, two different microsatellite motifs are found in close proximity. If only a small number of bases separates these two microsatellites, they are referred to as compound microsatellites. SciRoKo is the first search tool that allows a systematic survey for associated repeat motifs. All repeats that are observed within a specified distance ('*Max distance for association*') are reported. To account for the similarity of microsatellite motifs, four different types of associations between SSR repeats are reported:

- Type 1 motif associations: associations of fully standardized SSR–motifs.
- Type 2 motif associations: associations of partially standardized SSR–motifs, including the complementary strand compound microsatellite and not considering the 5′–3′ arrangement.
- Type 3 motif associations: associations of partially standardized SSR–motifs, considering the complementary strand compound microsatellite and the 5′–3′ arrangement.
- Type 4 motif associations: association of partially standardized SSR–motifs, excluding the complementary strand compound microsatellite and considering the 5′–3′ arrangement.

Table 1. Benchmark tests of SciRoKo, tandem repeat finder (TRF), Sputnik and Modified Sputnik I–II, with different DNA sequences

	rye ESTs	*S. cerevisiae*	*G. zeae*	*O. sativa*
Sequence count	9195	16	43	12
total size	4.3 Mb	12.1Mb	36.5 Mb	370.8 Mb
SciRoKo	1.27 s	2.75 s	7.3 s	1 min 32 s
TRF	5 min 24 s	40 s	1 min 37 s	48 min 37 s
Sputnik	45 s	Failed	Failed	Failed
M. Sputnik I	60 s	4 min 18 s	4 min 30 s	$\gg$ 2 h
M. Sputnik II	27 s	3 min 30 s	3 min 11 s	$\gg$ 2 h

4 COMPATIBILITY

The SciRoKo SSR-statistic module accepts also input files from Sputnik and Modified Sputnik I–II. SciRoKo provides an export option which allows saving the microsatellites in the Sputnik and Modified Sputnik I–II file formats. For Sputnik and Modified Sputnik II, tools allowing automated PCR-primer design for the SSR-search results have been created [SSRPrimer (Jewell *et al*., 2006) for Sputnik and a Perl script for Modified Sputnik II (LaRota *et al*., 2005)]. These tools might require some minor adjustments to use the exported SSR-search results for automated primer design.

5 METHODS

The Sputnik, Modified Sputnik I–II programs have been obtained at: http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota, the latest version of TRF 4.00 for Windows at: http://tandem.bu.edu/trf/trf.html

All SSR-search time measurements have been done on a common Laptop with an Intel©Pentium©M Processor 1.8 GHz and 1 GB RAM. Used SciRoKo settings: mismatched search; fixed mismatch penalty 4; minimum score 14; SSR-seed length 8; SSR-seed repeats 3; maximum mismatches at once 3. Default settings have been used for the Modified Sputnik I–II SSR-search with the following exceptions: minimum unit length 1; points for a mismatch −4; minimum score 11; used TRF settings: match–mismatch–indel 2–7–7; minimum alignment score to report 20; maximum period size 5. The Sputnik search has been done with the default settings. For accurate time assessment a small console program has been written. The sequences of *Oryza sativa*, *Saccharomyces cerevisiae*, *Gibberella zeae* and the *Secale cereale* (rye) ESTs have been obtained from the NCBI homepage (http://www.ncbi.nlm.nih.gov).

ACKNOWLEDGEMENTS

We thank Thomas Kofler for providing the web space and for helpful comments on programming in C#. This work was financially supported by the Austrian Science Fund (FWF, No P18414-B14). C.S. is supported by the Fonds zur Förderung der wissenschaftlichen Forschung (FWF).

Conflict of Interest: none declared.

REFERENCES

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

Buschiazzo,E. and Gemmell,N.J. (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays*, **28**, 1040–1050.

Castelo,A.T. *et al.* (2002) TROLL – tandem repeat occurrence locator. *Bioinformatics*, **18**, 634–636.

Dieringer,D. and Schlötterer,C. (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res.*, **13**, 2242–2251.

Ellegren,H. (2004) Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.

Gao,L. *et al.* (2003). Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol. Breed.*, **V12**, 245–261.

Jewell,E. *et al.* (2006) SSRPrimer and SSR taxonomy tree: Biome SSR discovery. *Nucleic Acids Res.*, **34**, W656–W659.

Kashi,Y. and King,D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends in Genet.*, **22**, 253–259.

La Rota,M. *et al.* (2005) Nonrandom distribution and frequencies of genomic and est-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics*, **6**, 23.

Morgante,M. *et al* (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.*, **30**, 194–200.

Schlötterer,C. (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma*, **109**, 365–371.

Temnykh,S. *et al.* (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. *et al.* (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. Epub 2002 Sep 14.

Varshney,R.K. *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.

Weber,J.L. (1990) Informativeness of human (dC-dA)n.(dG-dT)n polumorphisms. *Genomics*, **7**, 524–530.