Influence of Parameters and Differences in Results of Microsatellite Detection Software: A Bioinformatics Study

[*1]Sujan Patnana, [2]Ramasree Sripada
[*1]Dept. of Computer Science & Engineering, Aditya Engineering College, Surampalem;
[2]Dept. of Computer Science & Engineering, Aditya Engineering College, Surampalem, India
Phone:9985848485, Email Id: sujan.patnana@gmail.com

Received date: Dec 1st 2012, Accepted date: Dec 5th 2012, Published date: Jan 1st 2013

Abstract:
Microsatellites are small nucleotide repeats which are formed by tandem repetition of motifs of size 1-6 bp. These repeats play an important role in Genome evolution, Associated with various diseases, Used as molecular markers in DNA Fingerprinting, Population Genetics, Paternity Studies, Foresnsics, etc. Various bioinformatics tools are being used for extraction of STRs during the computational studies. However, computational studies of STRs can suffer from a significant bias depending on the software tool used in the study. We did a comparative study of various popular microsatellite extraction tools while detecting perfect and imperfect microsatellites separately. We found that the tools show similer efficiencies while detecting perfect repeats whereas differ a lot in efficiencies while detecting imperfect microsatellites.

Keywords:
Microsatellites; Perfect Repeats; Software Tools; Comparison; Short Tandem Repeats; Simple Sequence Repeats.

Introduction:

Microsatellites or Short Tandem Repeats (STRs) or Simple Sequence Repeats (SSRs) are tandem repeats of nucleotide motifs of the size 1-6bp [1]. Because of their polymorphic nature, abundance and distribution through out coding and non-coding regions, these repeats have been a major area of interest for the researchers. They have been used as genetic markers in DNA Fingerprinting and in paternity studies [2]. They are also associated with various diseases [3] and play an important role in various regulatory mechanisms and evolution [4] [5]. Recently, they are also found to be associated with the plasticity and evolution of bacterial genomes [6]

Microsatellites can be categorized into perfect, imperfect (with few mismatches) and compound microsatellites. A perfect microsatellite tract is one with 100% identical copies of motifs. For example, ATGATGATGATG is a perfect microsatellite tract with the motif 'ATG' repeating 4 times (represented as $(ATG)_4$). The 'Perfect' microsatellite tracts also suffer from point mutations such as indels and substitutions there by making it an 'Imperfect' microsatellite tract. For example, ATGATCATGATG is the imperfect tract with a substitution at 6th position. The other type of microsatellites called 'Compound' microsatellite includes multiple motifs in the same tract separated by certain distance among the individual repeats. For example, $(ATG)_n$gcctc$(GC)_m$ is a compound microsatellite tract with two microsatellite tracts of ATG and GC motifs separated by 5 nucleotides. Out of the three, imperfect microsatellites are of much interest for the researchers and are more stable when compared to perfect tracts [7]

Increasing availability of genome data has led to the development of various software tools for *in-silico* study of these microsatellites in various genomes in place of cost and time intensive laboratory methods. Till date, many software programs are available for extraction of tandem repeats from genome sequences. These tools have been used extensively by biologists who might not have the necessary knowledge about the inner implementation (algorithm) and functionality of the software. Even though many software tools are available for microsatellite extraction, they vary largely in-terms of their algorithm, efficiency and functionality. Recent studies [8, 9, 10] showed that there is a significant bias in the number of repeats detected by various microsatellite detection algorithms. This is a serious problem as the microsatellite studies involving these repeats suffer from inconsistent and incomplete results and the credibility of the studies may be questioned. This paper attempts a similar study to compare the efficiency of a set of microsatellite detecting tools in detecting perfect and imperfect repeats.

Materials and Methods:

Generally, computational studies on microsatellites use a tool of their choice with certain parameters for extracting imperfect microsatellites. The study by Merkel *et. al.*, 2008 [9] showed that there is a significant study bias while using different tools and different parameters. The study concluded that the minimum array length is the key factor that results in these biases in results and suggested the following thresholds for future studies: 12nt for mono, di and tri-nucleotides, 16nt for tetra-nucleotides, 20nt for penta-nucleotides and 24nt for hexa-nucleotides. However, the recent study by Mudunuri et., al., 2010 [10] revealed that the tools also differ in efficiency even after detecting imperfect repeats with the above suggested parameters. In this paper, we have used the same parameters

and compared the commonly used microsatellite extraction software in detecting perfect repeats and imperfect repeats separately. For comparison of tools in detecting perfect repeats, the tools MISA [11], mreps [12], Msatfinder [13], poly [14], SSRIT [15] and IMEx [16,17] have been used. We have extracted perfect microsatellites from NC_001136 (Saccharomyces cerevisiae chromosome IV) using all the above tools and the results have been compared. For detecting imperfect repeats, we have used the tools IMEx, SciRoKoCo [18], Sputnik [19] and TRF [20] with their default parameters. Imperfect microsatellites have been extracted from three different sequences namely *E.coli k12* bacterial genome, C *elegans* chromosome I, and *Drosophila melanogaster* chromosome X. We have used the minimum array length suggested for all tools by Merkel *et. al.*, 2008 (for mono,di,tri – 12, tetra – 16, penta – 20, hexa – 24) for both the experiments. For the tools, IMEx and mreps, we have used the following minimum repeat numbers: mono: 12, di: 6, tri: 4, tetra: 4, penta: 4, hexa: 4. For other tools, as there is no such option to set the limits, we have set 12 as the minimum array length for all repeats and then did some post-processing to filter out repeats less than the thresholds.

Results and Discussion:

The tools used in this study differ in terms of the algorithm, features and the type of repeats they detect. MISA (MicroSAtellite Identification Tool) [11] is a perl program designed to detect perfect and compound microsatellites. It uses the power of regular expressions of perl to detect microsatellites. mreps [12] uses a heuristic based seed detection technique to identify tandem repeats (including microsatellites and macrosatellites) in a given input sequence. Msatfinder [13] is a perl based program that detects perfect and interrupted microsatellites. It uses regular expression search at different levels of accuracy and speed. IMEx (Imperfect Microsatellite Extractor) [16, 17] uses a sliding window approach to detect perfect, imperfect and compound microsatellites from a given sequence. The tools Poly [14] and SSRIT [15] have been specifically developed to detect only perfect microsatellites. SciRoKoCo (SSR Classification and Investigation by Robert Kofler) [18] is a recently developed microsatellite detection software based on a statistical model and has been widely used for detecting imperfect and compound microsatellites. Sputnik [19] is a small C program that uses a recursive algorithm to detect imperfect microsatellites. TRF (Tandem Repeat Finder) [20] is the most popular tandem repeat detection tool that uses a probabilistic approach to detect tandem repeats of size as large as 2000bp. The remaining sections of this paper explain the results of the experiment conducted using the above mentioned tools.

A. Comparison of Perfect Repeat Tools

The perfect repeat detection tools have been used to detect repeats from *Saccharomyces cerevisiae* chromosome IV. The following table (Table I) shows the number of perfect repeats (overall and motif wise) when detected using all the tools.

Comparison of microsatellite extraction software on the sequence NC_001136 (Saccharomyces cerevisiae chromosome iv) considering only perfect repeats

Tool	No. of Perfect Repeats Detected						
	Mono	*Di*	*Tri*	*Tetra*	*Penta*	*Hexa*	*Total*
IMEx	119	54	134	7	2	3	319
MISA	119	54	134	7	3	4	321
Mreps	119	54	134	7	3	4	321
Msatfinder	119	54	134	7	3	3	320
Poly	110	43	83	0	0	9	255
SSRIT	0	54	134	7	3	4	202

Table I

The results indicate that there is less variation in terms of the number of repeats when only perfect repeats are considered. When the repeats are compared based on motif-size , there seems to be very little difference. SSRIT did not pickup any mono-repeats as the tool can not detect mono-nucleotide repeats. So, overall the tools are working with similar efficiencies when only perfect repeats are extracted.

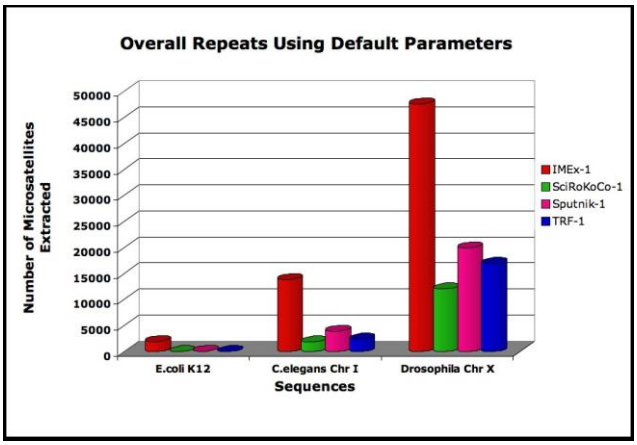
B. Comparison of Imperfect Repeat Tools

Further, we have extracted imperfect microsatellites from three different sequences namely *E.coli k12* bacterial genome, *C. elegans* chromosome I, and *Drosophila melanogaster* chromosome X. The sequences have been used earlier by Mudunuri *et., al.,* 2010 [10] and the experiment has been conducted again in this study. The tools TRF, Sputnik, SciRoKoCo and IMEx for these studies as the tools are specifically designed to detect imperfect repeats. When the number of imperfect repeats detected by these tools (using default parameters) was compared, we found that there is a huge variation among the results.

For *E. coli* genome, SciRoKoCo extracted only 3 repeats where as IMEx, TRF and Sputnik extracted 1899, 13 and 68 repeats respectively. For the *C. elegans* chromosome I, SciRoKoCo, TRF and IMEx extratced 1824, 2368, 3908 and 13827 repeats respectively. For the *Drosophila melanogaster* Chromosome X sequences, IMEx extracted 47625 repeats where as the tools SciRoKoCo, TRF, and Sputnik extracted 12070, 16946 and 19956 repeats respectively. The numbers indicate clearly that the tools differ in their efficiency when imperfect repeats are

considered. IMEx seems to have performed better among the tools followed by Sputnik, TRF and SciRoKoCo. The following bar graph depicts the variation in the number of imperfect repeats detected by the 4 tools.

FIGURE 1:



BAR GRAPH REPRESENTING THE DISTRIBUTION OF IMPERFECT MICROSATELLITES IN 3 DIFFERENT GENOMES WHEN COMPARED WITH IMEX, SCIROKOCO, SPUTNIK AND TRF.

The study clearly indicates that the tools show similer efficiencies while detecting perfect repeats. When imperfect repeats are considered, the tools differ a lot in terms of their efficiencies. The main reason for the differences in the efficiencies of the tools is due to the algorithm that defines the imperfection of the microsatellite. For perfect repeats, the tools are working almost the same as the definition of a perfect microsatellite is dependent only on the minimum repeat numbers. Where as for imperfect microsatellites, the level of imperfection (number of substitutions / insertions / deletions) allowed by the algorithm changes from tool to tool. This causes the tools to detect different number of repeats. Some tools that allow very few imperfections detect very few repeats where as the tools that allow more imperfections can detect more number of repeats. It should also be noted that the tools (except IMEx) lack provision to define the imperfection of the repeats. So, one should be careful in using microsatellite tools especially while detecting imperfect microsatellites from genome sequences.

Acknowledgment:

The authors would like to thank the management of Aditya Engineering College, Surampalem for their support and encouragement. Sujan Patnana thanks the TalentSprint Edu. Services, Hyderabad for providing necessary infrastructure for carrying out this work.

References:

[1] Schlotterer, C. (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma*, 109, pp.365–371.

[2] Tautz, D. and Schlotterer, C. (1994) Simple sequences. *Curr. Opin. Genet. Dev.*, 4, pp.832–837.

[3] Ellegren, H. (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* 24, pp. 400-402.

[4] Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER. (2005) Microsatellite instability regulates transcription factor binding and gene expression. *PNAS.*, 102, pp.3800–3804.

[5] Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. (2004) Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.*, 21, pp.991–1007.

[6] Sreenu,V.B., Kumar, P., and Nagarajaram, H.A. (2007) Simple sequence repeats in mycobacterial genomes. *J. Biosci.*, 32, pp.3–15.

[7] Fan, H. and Chu, J.Y. (2007) A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics.*, 5(1), pp.7-14.

[8] Leclercq, S., Rivals, E., and Jarne, P. (2007) Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics*, 8, pp.125.

[9] Merkel, A., and Gemmell, N. J. (2008) Detecting microsatellites in genome data: variance in definitions and bioinformatic approaches cause systematic bias. *Evol. Bioinform.*, 4, pp. 1-6.

[10] Mudunuri, S. B., Rao, A.A., Pallamsetty, S., and Nagarajaram, H.A. (2010). Comparative Analysis of Microsatellite Detecting Software: A Significant Variation in Results and Influence of Parameter. *Proceedings of ACM's International Symposium of Bio-computing (ISB 2010)*, 15-17, February 2010.

[11] T. Thiel, 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, pp. 411-422.

[12] Kolpakov, R., Bana, G., and Kucherov, G. (2003) mreps: Efficient and flexible detection of tandem repeats in DNA *Nucleic Acids Res*, 31, pp. 3672-2678.

[13] Thurston, M. I., and Field, D. (2005) Msatfinder: Detection and Characterization of Microsatellites. *Oxford: Centre for Ecology and Hydrology*.

[14] Bizzaro, W., and Marx, K. A. (2003) Poly: a quantitative analysis tool for Simple Sequence Repeat (SSR) tracts in DNA, *BMC Bioinformatics,* 4, p. 22.

[15] Temnykh, S., DeClerck, G., Lukashova, S., Lipovich, L., Cartinhour, S., and Mc-Couch, S. (2001) Computational and experimental analysis of microsatellites in rice (*oryza sativa l.*): frequency, length variation, transposon associations, and genetic marker potential, *Genome Res*, 11, pp. 1441-1452.

[16] Mudunuri, S.B., and Nagarajaram, H.A (2007) IMEx: Imperfect Microsatellite Extractor, *Bioinformatics,* 23, pp. 1181–1187.

[17] Mudunuri, S.B., Kumar, P., Rao, A.A., Pallamsetty, S., and Nagarajaram, HA (2010) G-IMEx: A comprehensive software tool for detection of microsatellites from genome sequences. *Bioinformation,* 5, pp. 001–003.

[18] Kofler, R., Schlotterer, C., and Lelley, T. (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation, *Bioinformatics*, 23, pp. 1683-1685.

*****