Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 5 **Software**

G-IMEx: A comprehensive software tool for detection of microsatellites from genome sequences

Suresh Babu Mudunuri[1], Pankaj Kumar[4], Allam Appa Rao[2], Pallam Setty Sanaboina[3], Hampapathalu Adimurthy Nagarajaram[4,*]

[1]Department of Computer Science and Engineering, Aditya Engineering College (AEC), Surampalem 533 437, India; [2]Jawaharlal Nehru Technological University (JNTU), Kakinada, 533 003, India; [3]Department of Computer Science and Systems Engineering, Andhra University College of Engineering (AUCE), Visakhapatnam 530 003, India; [4]Laboratory of Computational Biology, Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad 500 001, India; H.A.Nagarajaram - Email: han@cdfd.org.in; Phone: +91-40-24749367; +91-9908209193; Fax: +91-40-24749448; *Corresponding Author

Received June 28, 2010; Accepted August 25, 2010; Published November 1, 2010

Abstract:
Microsatellites are ubiquitous short tandem repeats found in all known genomes and are known to play a very important role in various studies and fields including DNA fingerprinting, paternity studies, evolutionary studies, virulence and adaptation of certain bacteria and viruses etc. Due to the sequencing of several genomes and the availability of enormous amounts of sequence data during the past few years, computational studies of microsatellites are of interest for many researchers. In this context, we developed a software tool called Imperfect Microsatellite Extractor (IMEx), to extract perfect, imperfect and compound microsatellites from genome sequences along with their complete statistics. Recently we developed a user-friendly graphical-interface using JAVA for IMEx to be used as a stand-alone software named G-IMEx. G-IMEx takes a nucleotide sequence as an input and the results are produced in both html and text formats. The Linux version of G-IMEx can be downloaded for free from http://www.cdfd.org.in/imex

Keywords: Bioinformatics Tool, Stand-alone program, Microsatellites, Simple Sequence Repeats, Genomes

Background:

Microsatellites, also known as Simple Sequence Repeats (SSRs) or Short Tandem Repeats (STRs), are tandem repetitions of a nucleotide motif of size 1-6 bp. They are distributed in both coding as well as non-coding regions of all known genomes. Because of their polymorphic nature, they are known to play an important role in gene regulation, pathogenesis, bacterial adaptation and in evolution of genomes **[1-5]**. They are also applied in various fields such as DNA fingerprinting, Paternity studies, Forensics, Evolutionary studies etc. As the sequencing of new genomes is increasing day-by-day, microsatellites of many genomes remain unexplored. Analysis of these microsatellites is important to understand their role in various studies. Computational analysis is a better alternative to the time-consuming and money-intensive traditional wet lab microsatellite studies. A software tool that can extract all types of microsatellites with greater sensitivity and provides flexible options to analyze the repeats detected is the need of the day.

Few tools **[6-9]** exist in the public domain for extracting microsatellites from genome sequences, but many of them suffer from certain lacunae in-terms of their features and their efficiency. In the course of our studies on evolution of microsatellites in prokaryotic genomes, we developed a novel algorithm **[10]** to detect imperfect microsatellites from nucleotide sequences. The algorithm has been implemented in the form of a stand-alone software with a user-friendly graphical user interface (GUI) called G-IMEx. The present communication gives the details of this software.

Methodology:

The algorithmic details of IMEx have been reported elsewhere **[10]**. For the sake of continuity we reiterate the method. IMEx scans the input sequence and looks for two consecutive exact repeat units or two alternate exact repeat units and considers them as the ‘candidate’ microsatellite repeat tract. The ‘candidate’ tract is expanded on both sides by allowing few mismatches in each individual repeat unit (‘k’ – imperfection limit / repeat unit) such that the percentage of imperfection of the entire tract does not cross the threshold set by the user. The expansion is also terminated if a repeat unit with more than ‘k’ mismatches is encountered. The program further collates and clusters equivalent microsatellite repeats into families. It also has an option to identify compound microsatellites, which are regions containing more than one microsatellite tract separated by a certain distance as defined by the user.

Software Requirements:

G-IMEx has been developed on the Linux platform and requires preinstalled C and Java (for graphical interface). An ideal environment for running G-IMEx would be a latest Fedora or other Linux distribution with a gcc compiler (version 3.4 or higher), Java version (1.6 or higher) and any browser software.

Input options:

G-IMEx offers several options for identification, extraction, collation, clustering and reporting of microsatellites from an input DNA sequence in FASTA format. The software can handle large sequences such as genomes easily and is comparatively faster than many other tools. Users can set the limits for repeat size, repeat number, repeat type and imperfection level. In addition users can set levels (0 to 4) for clustering of equivalent microsatellites and also to detect compound microsatellites i.e., those microsatellites which are close to each other sequentially. There is also an option to use the core IMEx program in batch mode for scanning multiple sequences.

Output options:

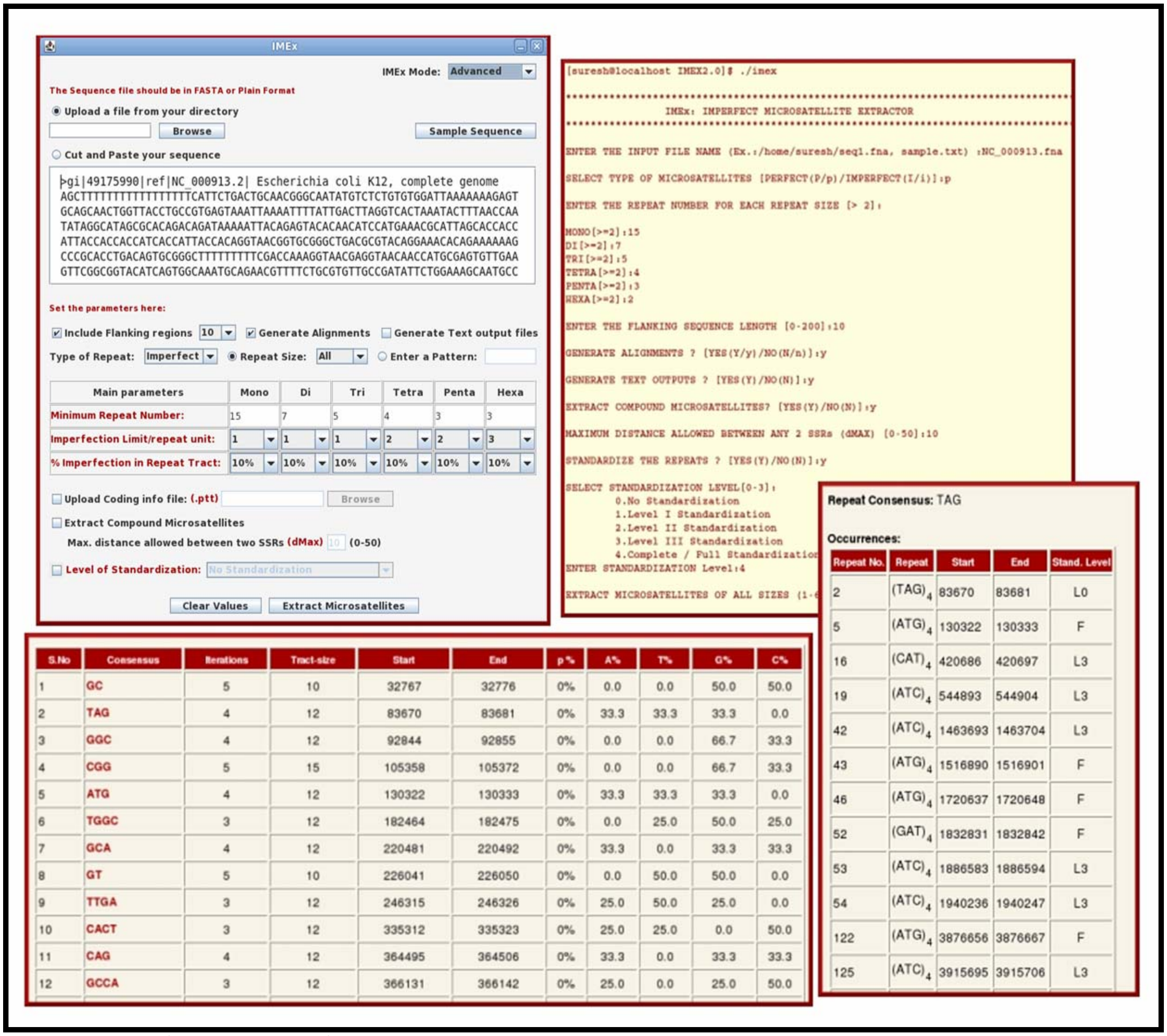
G-IMEx creates a folder with the name of the input sequence file and the results are stored in two formats – html and text. The text format of results is optional and separate directories are created for text and html results. The output includes a well-formatted summary table file with information such as the repeating motif, repeat number, imperfection %, tract size, nucleotide composition and protein information (if it falls in coding region) etc. Along with the information about the microsatellite extracted, its corresponding alignment with its perfect repeat counterpart is also produced automatically in a separate alignment file which facilitates analysis of mutational events in a microsatellite tract. **Figure 1** shows the snapshot of the GUI and the result pages of G-IMEx.

Future Work:

The current version of G-IMEx is available only for Linux users. Efforts are underway to develop versions compatible to Windows and Macintosh systems.

Acknowledgements:

The authors acknowledge the support from the lab members of computational biology and the core grant of CDFD. SBM would like to thank Mr. Priyatosh Mishra for his valuable help during the development of the graphical interface and the management of AEC for providing necessary facilities.



The Sequence file should be in FASTA or Plain Format

● Upload a file from your directory

○ Cut and Paste your sequence

```
>gi|49175990|ref|NC_000913.2| Escherichia coli K12, complete genome
AGCTTTTTTTTTTTTTTTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGT
GCAGCAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCC
```

Set the parameters here:

☑ Include Flanking regions 10 ☑ Generate Alignments □ Generate Text output files

Type of Repeat: Imperfect ● Repeat Size: All ○ Enter a Pattern:

Main parameters	Mono	Di	Tri	Tetra	Penta	Hexa
Minimum Repeat Number:	15	7	5	4	3	3
Imperfection Limit/repeat unit:	1	1	1	2	2	3
% Imperfection in Repeat Tract:	10%	10%	10%	10%	10%	10%

□ Upload Coding info file: (.ptt)

□ Extract Compound Microsatellites

Max. distance allowed between two SSRs (dMax) 10 (0-50)

□ Level of Standardization: No Standardization

```
[suresh@localhost IMEX2.0]$ ./imex

*****************************************************************************
                    IMEx: IMPERFECT MICROSATELLITE EXTRACTOR
*****************************************************************************

ENTER THE INPUT FILE NAME (Ex.:/home/suresh/seq1.fna, sample.txt) :NC_000913.fna

SELECT TYPE OF MICROSATELLITES [PERFECT(P/p)/IMPERFECT(I/i)]:p

ENTER THE REPEAT NUMBER FOR EACH REPEAT SIZE [> 2]:

MONO[>=2]:15
DI[>=2]:7
TRI[>=2]:5
TETRA[>=2]:4
PENTA[>=2]:3
HEXA[>=2]:2

ENTER THE FLANKING SEQUENCE LENGTH [0-200]:10

GENERATE ALIGNMENTS ? [YES(Y/y)/NO(N/n)]:y

GENERATE TEXT OUTPUTS ? [YES(Y)/NO(N)]:y

EXTRACT COMPOUND MICROSATELLITES? [YES(Y)/NO(N)]:y

MAXIMUM DISTANCE ALLOWED BETWEEN ANY 2 SSRs (dMAX) [0-50]:10

STANDARDIZE THE REPEATS ? [YES(Y)/NO(N)]:y

SELECT STANDARDIZATION LEVEL[0-3]:
        0.No Standardization
        1.Level I Standardization
        2.Level II Standardization
        3.Level III Standardization
        4.Complete / Full Standardizatio
ENTER STANDARDIZATION Level:4

EXTRACT MICROSATELLITES OF ALL SIZES (1-6
```

Repeat Consensus: TAG

Occurrences:

Repeat No.	Repeat	Start	End	Stand. Level
2	$(TAG)_4$	83670	83681	L0
5	$(ATG)_4$	130322	130333	F
16	$(CAT)_4$	420686	420697	L3
19	$(ATC)_4$	544893	544904	L3
42	$(ATC)_4$	1463693	1463704	L3
43	$(ATG)_4$	1516890	1516901	F
46	$(ATG)_4$	1720637	1720648	F
52	$(GAT)_4$	1832831	1832842	F
53	$(ATC)_4$	1886583	1886594	L3
54	$(ATC)_4$	1940236	1940247	L3
122	$(ATG)_4$	3876656	3876667	F
125	$(ATC)_4$	3915695	3915706	L3

S.No	Consensus	Iterations	Tract-size	Start	End	p %	A%	T%	G%	C%
1	GC	5	10	32767	32776	0%	0.0	0.0	50.0	50.0
2	TAG	4	12	83670	83681	0%	33.3	33.3	33.3	0.0
3	GGC	4	12	92844	92855	0%	0.0	0.0	66.7	33.3
4	CGG	5	15	105358	105372	0%	0.0	0.0	66.7	33.3
5	ATG	4	12	130322	130333	0%	33.3	33.3	33.3	0.0
6	TGGC	3	12	182464	182475	0%	0.0	25.0	50.0	25.0
7	GCA	4	12	220481	220492	0%	33.3	0.0	33.3	33.3
8	GT	5	10	226041	226050	0%	0.0	50.0	50.0	0.0
9	TTGA	3	12	246315	246326	0%	25.0	50.0	25.0	0.0
10	CACT	3	12	335312	335323	0%	25.0	25.0	0.0	50.0
11	CAG	4	12	364495	364506	0%	33.3	0.0	33.3	33.3
12	GCCA	3	12	366131	366142	0%	25.0	0.0	25.0	50.0

Figure 1: Snapshots of Graphical User-Interface and Results Pages.

References:

[1] AV Belkum *et al. Microbiol MolBiol Rev* **62**(2):275 (1998) [PMID: 9618442].

[2] P Martin *et al. Proc Natl. Acad Sci USA* **102**(10):3800(2005) [PMID:15728391]

[3] ER Moxon *et al. Curr Biol* **4**(1):24 (1994) [PMID: 7922307]

[4] VB Sreenu *et al. BMC Genomics* **7**:78 (2006) [PMID: 16603092]

[5] R Kofler *et al. BMC Genomics* **9**:612 (2008) [PMID: 19091106]

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

[7] R Kofler *et al. Bioinformatics* **23**: 1683 (2007) [PMID: 17463017]

[8] M LaRota *et al. BMC Genomics* **6**: 23(2005) [PMID: 15720707]

[9] http://www.genomics.ceh.ac.uk/msatfinder.

[10] SB Mudunuri & HA Nagarajaram *Bioinformatics* **23**(10):1181(2007) [PMID: 17379689]

Edited by P. Kangueane

Citation: Mudunuri *et al.* Bioinformation 5(5): 221-223 (2010)