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

Suresh B Mudunuri[1], Pankaj Kumar[4], Allam Appa Rao[2], S Pallamsetty[3], H A 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

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.

limits for repeat size, repeat number, repeat type and imperfection level. In addition users can set levels (0 to 4) for clustering of equivalent

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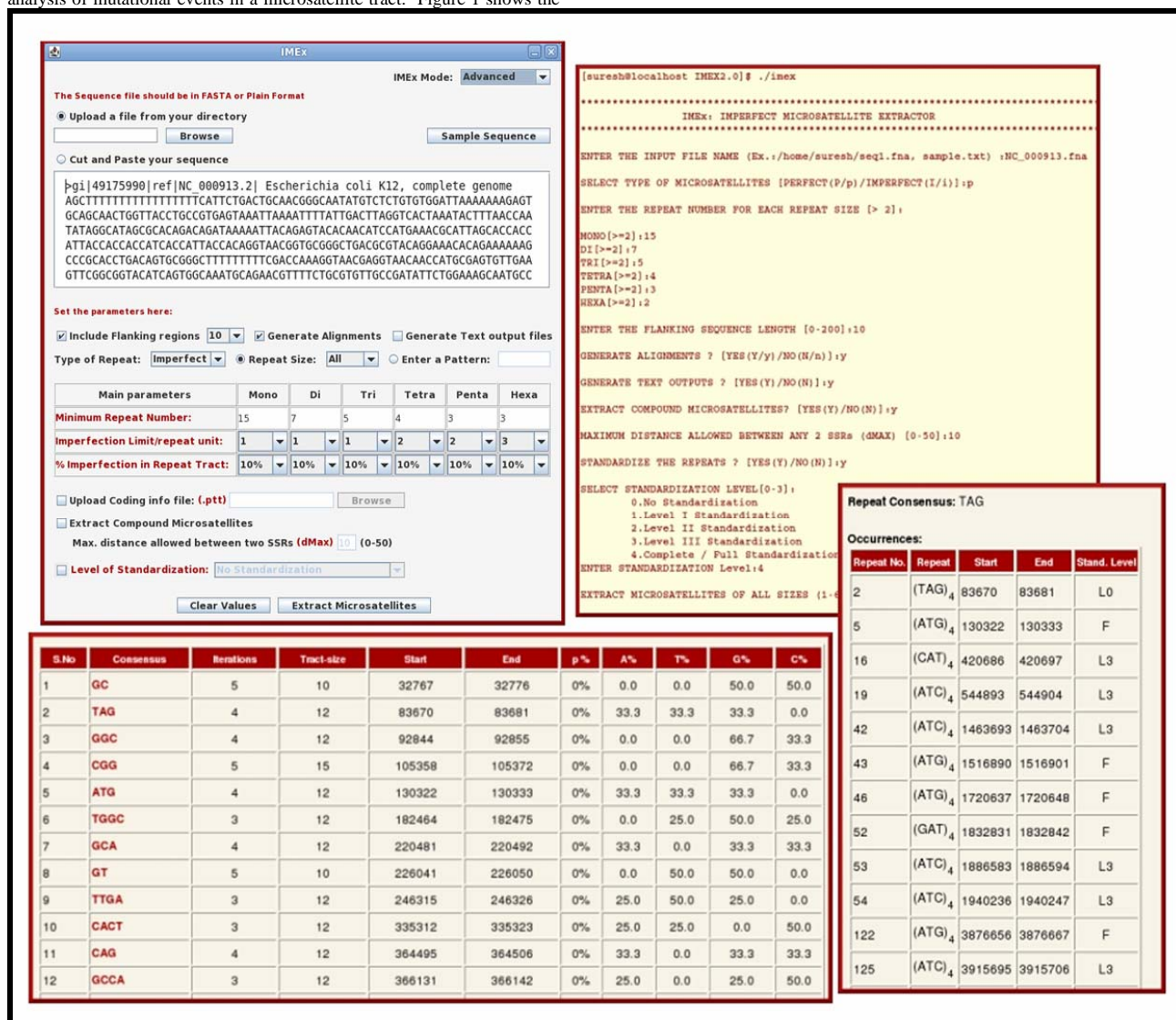
Figure 1: Snapshots of Graphical User-Interface and Results Pages.

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]