Journal of Computer Science & Systems Biology - Open Access
www.omicsonline.com JCSB/Vol.2 November-December 2009
Research Article OPEN ACCESS Freely available online doi:10.4172/jcsb.1000043

VMD: Viral Microsatellite Database-A Comprehensive Resource for all Viral Microsatellites

Suresh B. Mudunuri[1]*, Allam Appa Rao[2], S Pallamsetty[3], Priyatosh Mishra[1] and H.A.Nagarajaram[4]

[1]Department of Computer Science and Engineering, Aditya Engineering College, Surampalem, E.G.Dist, Andhra Pradesh -533437, India
[2]Vice-Chancellor, Jawaharlal Nehru Technological University, Kakinada, E.G.Dist, Andhra Pradesh - 533003, India
[3]Department of Computer Science and Systems Engineering, Andhra University College of Engineering (AUCE), Visakhapatnam, Andhra Pradesh - 530003, India
[4]Laboratory of Computational Biology, Centre for DNA Fingerprinting and Diagnostics (CDFD), Nampally, Hyderabad, Andhra Pradesh – 500001, India

Abstract

Microsatellites are the small DNA sequences with a tandem repetition of a particular motif of size 1-6. Microsatellites are found in all known genomes and play a significant role in many fields including DNA Fingerprinting, Population Studies, Forensics, Paternity Studies, Gene Regulation, Genetic Disorder Studies, and Evolution of Genomes. They are extensively used as genetic markers for identifying pathogenic bacteria and viruses. More over, they are found to be associated with the plasticity, adaptation and virulence of bacteria and viruses. Insilico analysis of microsatellites in various viruses would reveal many interesting facts about their evolution and adaptation. To the best of our knowledge, there is no comprehensive and exclusive database of all viral microsatellites that extracts all types of microsatellites with flexible extraction options. In this paper, we describe the details of a relational database named Viral Microsatellite Database (VMD). VMD currently hosts microsatellites of around 3500 viral genomes along with their alignments, locus information, imperfection info, protein info etc. The database can be accessed and downloaded for free for academic / research purposes from http://www.mcr.org.in/vmd.

Keywords: Database; Microsatellite; Virus; Genome; Tandem repeat; Web-interface

Introduction

Microsatellites or Simple Sequence Repeats (SSRs) or Short Tandem Repeats (STRs) are tandem repeats of motifs of size 1-6 nucleotides long (mono to hexa nucleotides) viz. $(T)_9$, $(ATGC)_4$, $(ATCGAT)_3$ (Schlotterer, 2000). These microsatellites can be classified into three main categories; (i) perfect (ii) imperfect and (iii) compound (Merkel and Gemmel, 2008). A 'perfect' microsatellite tract is a tandem repetition of exact copies of a particular motif. E.g.CAGCAGCAGCAG. Here, the motif 'CAG' is repeated 4 times (represented as $(CAG)_4$. The perfect repeats are often interrupted by mismatches such as insertions, deletions or substitutions resulting in an 'imperfect' tract of microsatellite. Eg. CAGC*T*GCAGCAG is an imperfect microsatellite tract with a substitution A->T at 5th position. Finally, a 'compound' microsatellite tract is one that contains multiple motifs with in the same tract separated by 0 or more intervening nucleotides. Eg. $(CAG)_4$ctgca $(GC)_3$ is a compound microsatellite tract with two microsatellite tracts of motifs CAG and GC separated by 4 nucleotides. Over the years, these ubiquitous repeats are of great interest for the researchers due to their application and significance in various fields. They are found in all known genomes including bacteria and viruses and are distributed through out the genome in both coding and non-coding regions (Toth et al., 2007). These repeats also play an important role in gene regulation and are also responsible for causing changes in protein products (Li et al., 2004; Martin et al., 2005). Mutations in these microsatellite tracts have been implicated to be responsible for certain neurodegenerative diseases in humans (Tautz and Schlotterer, 1994). They are known to be used in several areas such as DNA Fingerprinting, Forensics, Paternity studies, Linkage analysis etc.

Apart from all these, microsatellites are known to be highly polymorphic by nature as they gain/lose repeat units (motifs) in course of time, thus, making them highly important in the studies of genome evolution (Jarne and Lagoda, 1996). They are thought to be one of the sources of genetic diversity (Kashi and King, 2006). More recently, studies revealed that microsatellites are imparting a certain degree of plasticity in bacterial genomes indicating their significance in the context of pathogen adaptability, virulence and survival (Sreenu et al., 2006). Studies in certain viral genomes have also shown that polymorphism does exist among the viruses (Davis et al., 1999) and proven to be useful as markers in epidemiological and virulence studies (Hood et al., 1996). Studies also show that microsatellites play an important role in the evolution of new virus strains (Deback et al., 2009).

The recent epidemics/pandemics of viruses such as AIDS (HIV), Avian influenza (bird flu), Chikungunya, SARS, H1N1

***Corresponding author:** Suresh B. Mudunuri, Department of Computer Science and Engineering, Aditya Engineering College, Surampalem, E.G.Dist, Andhra Pradesh -533437, India E-mail: sureshverma@gmail.com

Received December 05, 2009; **Accepted** December 19, 2009; **Published** December 21, 2009

Citation: Mudunuri SB, Rao AA, Pallamsetty S, Mishra P and Nagarajaram HA (2009) VMD: Viral Microsatellite Database-A Comprehensive Resource for all Viral Microsatellites. J Comput Sci Syst Biol 2: 283-286. doi:10.4172/jcsb.1000043

influenza (swine flu) alerted the researchers and triggered the need for the study of viral evolution to understand how new strains evolve. Viruses have high mutation and reproduction rates and can adapt to changing environments quite well. Viruses infect the host cells and reproduce themselves very rapidly (20-30 generations per day). Microsatellite studies in viruses would reveal many interesting facts about the evolution of many new viruses. In this regard, we developed a comprehensive, curated, web-based relational database of all viral microsatellites called Viral Microsatellite Database (VMD).

A few microsatellite databases such as MICdb2.0 (Sreenu et al., 2003) and Small Genome Microsatellite Database (SGMD) (http://www.genomics.ceh.ac.uk/cgibin/sgmd/index.cgi) have been developed and are available online. MICdb2.0 hosts only perfect microsatellites extracted from bacterial and viral genome sequences by using a repeat finder program called 'SSRF'. SGMD hosts imperfect microsatellites of various bacteria, organelles, plasmids and viruses extracted using a program called 'Msatfinder'. However, these databases are not specifically designed for virus microsatellites and does not include complete microsatellite information of all available viral genomes such as repeat statistics, alignments, motif specific search, coding/non-coding locus information etc. IMEx-web [unpublished] is another comprehensive resource available till date from which one can extract microsatellites of all types with varying degrees of imperfection. IMEx-web does not store the microsatellites in a database but extracts the repeats 'on the fly' using the program IMEx (Mudunuri and Nagarajaram, 2007) over the web. Storing the repeats in a database and providing a better search facility would be very advantageous rather than extracting the microsatellites using a program. Using a database is very flexible in terms of filtering the required output such as repeats of particular type/size; repeats in coding regions only; sort the results by a particular field; restrict the output data by various fields etc. More over, the database VMD is made available for download so that one can write their own SQL queries to extract re-

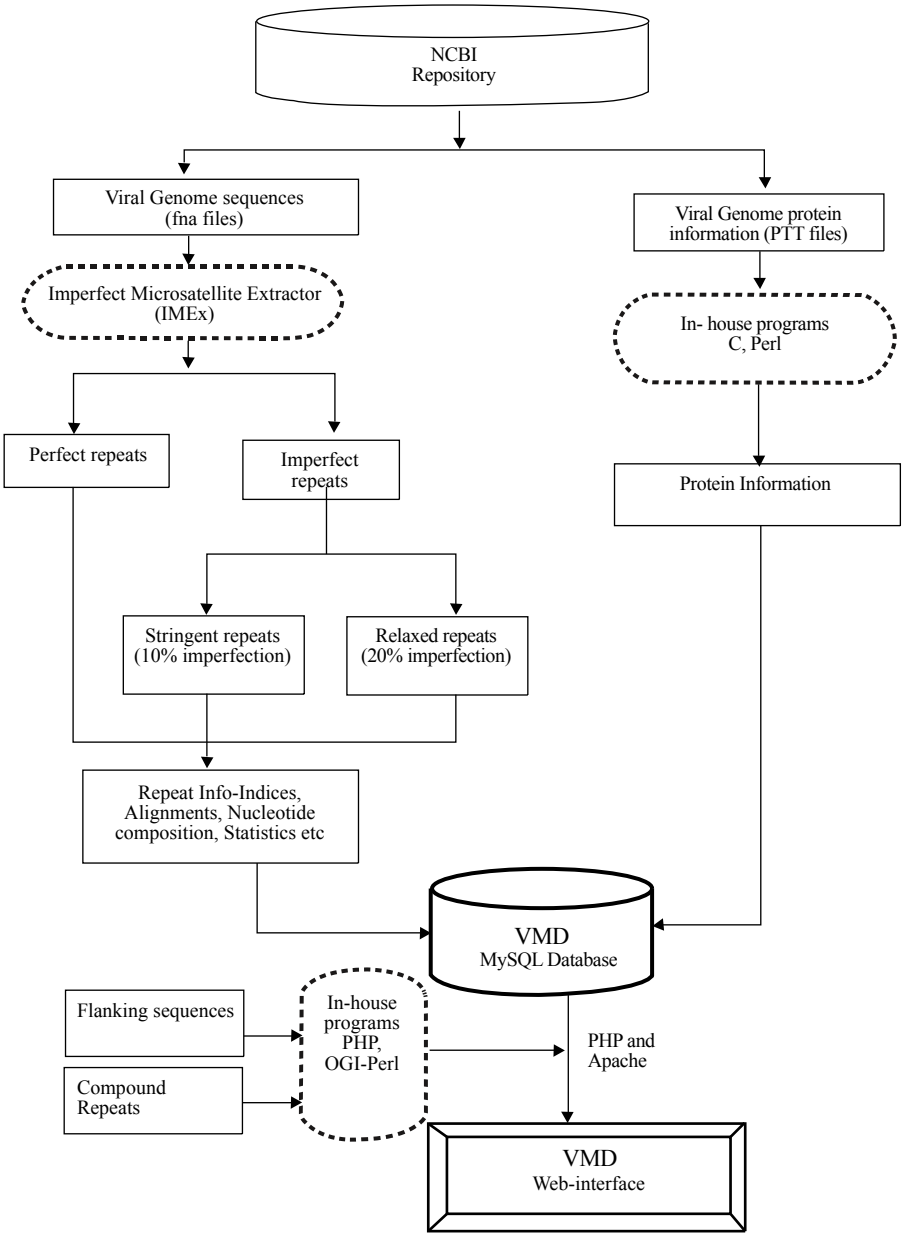

Figure 1: Architecture and Design of VMD.

quired information. This would be very advantageous for the comparative genomic studies of various viruses. The next sections cover the details of the design, development and implementation of VMD.

Database Descripton

Database architecture

VMD follows a multi-tier architecture where the data is extracted and curated using various programs/tools at different levels and compiled into a relational database. The database is connected to a flexible and user-friendly web-interface with many easy-to-use options. Perfect and Imperfect microsatellites are extracted from all the viral genomes using IMEx (Mudunuri and Nagarajaram, 2007) and their corresponding information is stored in the database. The flanking sequences and compound microsatellites are not stored in the database. As said before, compound microsatellites contain more than one microsatellite tract separated by 0 or more nucleotides. The user sets the maximum allowable distance (dMax) between any two microsatellites. In-order to reduce redundancy, compound microsatellites are not stored in the database but are extracted by calculating the distance between the adjacent microsatellites in the database and report them in the output as compound microsatellites if the distance is less than or equal to dMax. Similarly, flanking sequences of the repeats are also not stored in the database rather they are extracted from the sequence when required. The user will set the flanking limit that can vary for every query of the database. Hence, flanking sequences are extracted from the genome sequences 'on the fly' and are incorporated in the output. The detailed architecture and design of VMD is shown in Figure 1.

Microsatellite extraction

All the sequenced viral genomes were downloaded from the National Centre for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/) repository. We downloaded both the sequence (.fna files) and the protein information (.ptt files) files of all sequenced viral genomes. Microsatellites were extracted using Imperfect Microsatellite Extractor (IMEx) from all the sequences with the following parameters: Repeat numbers (n): Mono: 6, Di: 3, Tri-Hexa: 2; Imperfection limit / unit (k): Mono-Tri: 1, Tetra-Penta: 2 and Hexa: 3; Microsatellites are extracted from each sequence with three different levels of imperfection: Perfect microsatellites (with 0% imperfection), stringent imperfect microsatellites (with 10% imperfection, means one mismatch per every 10 nucleotides) and relaxed imperfect microsatellites (with 20% imperfection). The corresponding information of each microsatellite including the repeat motif, start and end co-ordinates, tract-length, number of iterations, imperfection % in the tract (p%), alignments, nucleotide composition etc are also stored in the database.

Web-interface

A user-friendly and easy-to-use web-interface has been designed to facilitate the users to extract the microsatellite information. In order to provide better navigation, the genomes are organized in alphabetical order and the user can browse through the genomes names easily. VMD search is provided in 2 different modes. Basic mode, with limited options, is designed specifically for beginners where the user can select a viral genome of his interest and the type of the repeat and extract the results. Advanced mode is for the expert users, which include many options to filter repeats of his interest. Using the advanced mode, apart from the options in basic mode, users can filter the repeats of a particular motif (e.g. CAG, GT, ATG) or particular size(s) (e.g. mono, tri, penta); restrict the maximum and minimum number of repeat units (no. of iterations); find microsatellites only in coding regions / non-coding regions / both; sort the results in a particular format; select the fields to be present in the output file; export the results in various formats etc. In-order to design primers for the microsatellites of interest, a popular primer design module Primer3 (Rozen and Skaletsky, 2000) has been incorporated in the web-interface.

Output options

VMD has a wide range of output options to facilitate the users to analyze the results. The output of the VMD queries can be exported in different formats and the user has a choice to select more than one of these. The results of the query can be exported into HTML, Excel, CSV and Text formats so that the users can use them for post-processing in their convenient way and analyze the results. The users of VMD can also choose what information should be part of the output such as repeat number, flanking sequence, coding info, alignments, nucleotide composition, etc. An option to send the results via an automated email delivery module is also present so that the users need not wait for the system to extract results.

Implementation

In-house programs (in C language) have been developed to submit each viral genome as input automatically to IMEx program for extraction of microsatellites. The database has been constructed using the popular relational database management system MySQL and the extracted microsatellite information has been stored in various tables in the database using Perl scripts developed in-house. Similarly, the protein information is extracted and stored in the database using in-house C and Perl programs. Web server using Apache software has been installed on the IBM Xeon server with Linux operating system. A comprehensive web-interface has been developed using HTML and CSS and hosted on the server. The forms are validated thoroughly using JavaScript. The connection between the web-interface and the MySQL database has been achieved using the server side scripting language PHP. The mapping of microsatellite repeats to their corresponding coding/non-coding regions as well as the extraction of compound microsatellites and flanking sequences has been achieved using the PHP programs. The primer3 module was incorporated via CGI-Perl programming.

Keeping in mind the importance of viruses and their effect in the current era, Viral Microsatellite Database (VMD) can be a very useful resource for microsatellite studies of thousands of viral genomes.

Availability

The current release (version 1.1, December 2009) archives microsatellites and their corresponding repeat information of 3465 complete viral genomes. A bimonthly updating of the database is planned. VMD has been made available for free and can be accessed from http://www.mcr.org.in/vmd. The database can be downloaded for free upon request purely for non-commercial use (for academicians and researchers).

Future plans

A comparative analysis web-interface will be developed to compare the microsatellite distribution of more than one virus at the same time, which would be very helpful for the researchers to observe the variations in various viral strains with regard to microsatellites. A module to automatically update the microsatellites of all complete viral genomes will be developed and their available protein information will be included in the future releases of VMD.

Acknowledgements

SBM and PM would like to thank Dr. Girendranath, Dean (CSE & IT) and all the staff members CSE Dept. of Aditya Engineering College who supported them during the development of the database and also the management of Aditya Engineering College for providing necessary resources and healthy environment during the entire work. SBM and HAN would also like to thank all the members of Lab of Computational Biology, CDFD for their valuable suggestions and assistance.

References

1. Davis CL, Field D, Metzgar D, Saiz R, Morin PA, et al. (1999) Numerous length polymorphisms at short tandem repeats in human cytomegalovirus. J Virol 73: 6265-70. » CrossRef » PubMed » Google Scholar
2. Deback C, Boutolleau D, Depienne C, Luyt CE, Bonnafous P, et al. (2009) Utilization of microsatellite polymorphism for differentiating herpes simplex virus type 1 strains. J Clin Microbiol 47: 533-40. » CrossRef » PubMed » Google Scholar
3. Hood DW, Deadman ME, Jennings MP, Bisercic M, Fleischmann RD, et al. (1996) DNA repeats identify novel virulence genes in haemophilus influenzae. Proc Natl Acad Sci USA 93: 11121-5. » CrossRef » PubMed » Google Scholar
4. Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11: 424-429. » CrossRef » PubMed » Google Scholar
5. Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22: 253-9. » CrossRef » PubMed » Google Scholar
6. Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21: 991-1007. » CrossRef » PubMed » Google Scholar
7. Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER (2005) Microsatellite instability regulates transcription factor binding and gene expression. Proc Natl Acad Sci USA 102: 3800-4. » CrossRef » PubMed » Google Scholar
8. Merkel A, Gemmell N (2008) Detecting short tandem repeats from genome data: opening the software black box. Brief Bioinform 9: 355-66. » CrossRef » PubMed » Google Scholar
9. Mudunuri SB, Nagarajaram HA (2007) IMEx: Imperfect Microsatellite Extractor. Bioinformatics 23: 1181-7. » CrossRef » PubMed » Google Scholar
10. Rozen S, Skaletsky H (2000) Primer3 on the www for general users and for biologist programmers. Methods Mol Biol 132: 365-8. » CrossRef » PubMed » Google Scholar
11. Schlotterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109: 365-71. » CrossRef » PubMed » Google Scholar
12. Sreenu VB, Alevoor V, Nagaraju J, Nagarajaram HA (2003) MICdb: database of prokaryotic microsatellites. Nucleic Acids Res 31: 106-8. » CrossRef » PubMed » Google Scholar
13. Sreenu VB, Kumar P, Nagaraju J, Nagarajaram HA (2006) Microsatellite polymorphism across the M. tuberculosis and M. bovis genomes: implications on genome evolution and plasticity. BMC Genomics 7: 78. » CrossRef » PubMed » Google Scholar