APPLICATION NOTE

MICAS: a fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences

Vattipally B Sreenu,[1] Gundu Ranjitkumar,[1] Sugavanam Swaminathan,[1] Sasidharan Priya,[1] Buddhaditta Bose,[1] Mogili N Pavan,[1] Geeta Thanu,[1] Javaregowda Nagaraju,[2] Hampapathalu A Nagarajaram[1]

[1]Laboratory of Computational Biology and Bioinformatics Facility (EMBnet India node), Centre for DNA Fingerprinting and Diagnostics, Nacharam, India; [2]Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, Nacharam, India

Abstract: MICAS is a web server for extracting microsatellite information from completely sequenced prokaryote and viral genomes, or user-submitted sequences. This server provides an integrated platform for MICdb (database of prokaryote and viral microsatellites), W-SSRF (simple sequence repeat finding program) and Autoprimer (primer design software). MICAS, through dynamic HTML page generation, helps in the systematic extraction of microsatellite information from selected genomes hosted on MICdb or from user-submitted sequences. Further, it assists in the design of primers with the help of Autoprimer, for sequences containing selected microsatellite tracts.

Keywords: microsatellites, genome analysis, primer design, simple sequence repeats

Availability: The MICAS server is available at http://www.cdfd.org.in/micas

Contact: Hampapathalu A Nagarajaram (han@cdfd.org.in)

Introduction

Microsatellites, also known as simple sequence repeats (SSRs), are the short, direct, tandem repeats of DNA sequences with repeating units comprising less than six nucleotides (Tautz and Renz 1984). They are known to cause phase variation in most of the pathogenic bacteria (Borst 1991), especially bacteria like *Neisseria gonorrhoeae* (Burch et al 1997), *Haemophilus influenzae* (Hood et al 1996) and *Moraxella catarrhalis* (Peak et al 1996). Microsatellites have been proven to be very useful genetic markers for strain differentiation (Andersen et al 1996), forensics (Jeffreys et al 1992), kinship (Morin et al 1994) and construction of phylogenetic relations (Meyer et al 1995).

The task of identification and characterisation of microsatellites, in the pre-genomic era, was time-consuming and cumbersome, involving a number of steps: isolation of flanking regions specific for each microsatellite locus, by the construction and screening of different libraries; detection of clones containing microsatellites; design of locus-specific polymerase chain reaction (PCR) primers; and amplification of the respective regions from different sources of genomic DNA. However, in the post-genomic era, availability of whole genome sequences has made the task of screening genomes for microsatellites simple and easy. A computer program specially developed for microsatellite screening is sufficient to scan the whole genome for sequence composition of motif, and the location and frequencies of microsatellite tracts.

As a part of our ongoing project on structural and functional characterisation of microsatellites, we developed both MICAS: a web-based, fully automated server for extraction and analysis of microsatellites; and MICdb: a database of microsatellites extracted from prokaryote and viral genomes (Sreenu et al 2003). MICAS also provides an interface to automated primer design software that can be used for designing primers for PCR amplification of genomic regions harbouring microsatellites. The MICAS server is available for public access at http://www.cdfd.org.in/micas.

MICAS: the web server

MICAS has a three-module architecture as illustrated in Figure 1. The first module is a processing unit, the second

Correspondence: Hampapathalu A Nagarajaram, Laboratory of Computational Biology and Bioinformatics Facility (EMBnet India node), Centre for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad-76, AP India; tel +91 40 2717 1502; fax +91 40 2715 5610; email han@cdfd.org.in

© 2003 Open Mind Journals Limited. All rights reserved.

AUTHOR PROOF COPY ONLY

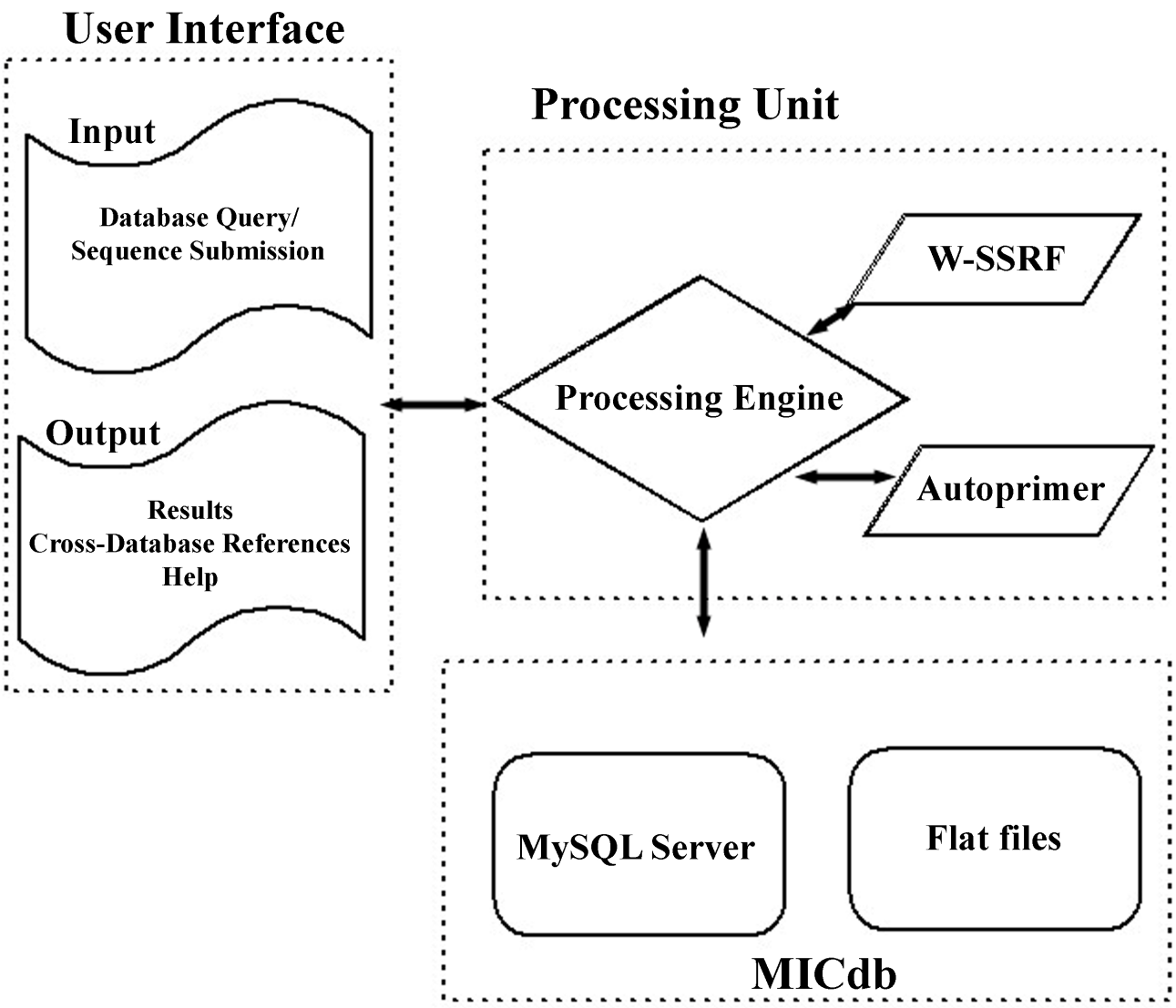

Figure 1 MICAS architecture.

module is a database and the third module is a user interface. The processing unit is made up of three programs: W-SSRF, Autoprimer and the processing engine.

W-SSRF: web-based simple sequence repeat finder

W-SSRF has been developed using Java™ programming language. The program scans a given nucleotide sequence for the presence of perfect simple sequence repeats from motif lengths of mono- to deca-nucleotides. The extracted information pertaining to SSRs includes the sequence content of the motif, repeat numbers, and start and end positions of the SSR tracts in the sequence.

Autoprimer: primer design software

Autoprimer has been developed using Java programming language. The essential parameters required for the optimum design of primers are namely: primer length, GC content, self-complementarity potential, primer dimer formation potential, self end-annealing potential and melting temperature (Tm). These have all been considered. Tm is calculated using the nearest-neighbour thermodynamics equation as given in Breslauer et al (1986). Default values for primer length and GC percentage in the primer sequences have been set to 20 nucleotides (optimum range is 18–30) and 50 nucleotides (optimum range is 45–65) respectively. Primers are examined for repetitive sequences. Dimerisation potential between the forward and reverse primers is checked by the extent of complementarity between them (Kampke et al 2001). Each primer is further tested for unintended hybridisation with itself, by testing for self-end annealing. By default, the program generates 10 sets of possible primer pairs for a given sequence. The forward and reverse primers are designed to suit the upstream and the downstream regions, respectively, of the microsatellite tract of interest.

MICdb: microsatellite database

The second module, a database, is completely constituted by MICdb, which is a combination of MySQL relational tables and flat files. The MySQL tables store precompiled, non-redundant microsatellites extracted from whole genomic sequences using SSRF – a stand-alone program similar to W-SSRF developed in the C programming language. The information stored includes the microsatellite motifs, their loci, repeat number, frequency of occurrence and location with regard to coding regions. The flat files contain the genome summary, genome sequences and the open reading frames (ORFs), along with predicted secondary structural information corresponding to their protein translations. The secondary structures for the translated proteins have been predicted using PSIPRED (Jones 1999). For precompilation of the database, the whole genome sequences were downloaded from the National Center for Biotechnology Information (NCBI) ftp site (ftp://ftp.ncbi.nih.gov).

User interface

The third module is the user interface, developed using Java servlets. The interface provides a platform for MICdb, W-SSRF and Autoprimer to input queries as well as to display query results by means of dynamically generated HTML pages. To query the MySQL® database, the MM.MySQL JDBC driver has been used.

Data retrieval using MICAS

MICAS has a user-friendly interactive front-end through which either MICdb can be queried for microsatellites from a selected genome or a sequence can be scanned for microsatellites. To extract microsatellite information from genomes hosted by MICdb, the user has to select a genome from the drop-down menu and query the database for occurrence of tandem repeats of microsatellite motifs of specified size (*S*) repeating at least a specified number of times (*N*). Following the query, MICAS outputs a table containing the complete list of microsatellite motifs satisfying *S* and *N*. This table also provides for each motif the number of times that it occurs in the whole genome. The motifs in this table are hyperlinked to their details: genomic locations (starting and ending nucleotide numbers) and regions of their occurrence (whether they are in coding or non-coding regions). The coding regions containing microsatellites are shown along with their predicted secondary structural information. Secondary structures of the translated proteins are predicted using PSIPRED (Jones 1999).

As mentioned earlier, MICAS also provides an interface to W-SSRF. A text box is provided where a sequence of interest can be pasted and submitted to W-SSRF to scan for microsatellites. The current version of W-SSRF can take a sequence of up to 20 kB file size. For user-submitted sequences, all microsatellite motifs, repeat number and their positions in the given sequence are displayed. If the user wishes, they can select a microsatellite tract (obtained either from database query or sequence scanning) along with its flanking sequence for Autoprimer to design primers for PCR. Provision has been made for the user to change the default settings of various parameters for primer design. By default, Autoprimer generates 10 sets of primer pairs; however, this can also be changed according to user choice.

Acknowledgements

This work was supported in part by core grants to the Centre for DNA Fingerprinting and Diagnostics from the Department of Biotechnology (to HAN) and by the Department of Biotechnology extramural grant (to JN). Authors acknowledge Mr Vishwanath, Ms Sushma and Ms Swapna for their assistance in the initial development of MICdb. VBS gratefully acknowledges the Council of Scientific and Industrial Research (CSIR), India for a senior research fellowship.

References

Andersen GL, Simchock JM, Wilson KH. 1996. Identification of a region of genetic variability among *Bacillus antharcis* strains and related species. *J Bacteriol*, 178:377–84.

Borst P. 1991. Molecular genetics of antigenic variation. *Immunol Today*, 12:29–33.

Breslauer KJ, Frank R, Blocker H, Marky LA. 1986. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci USA*, 83:3746–50.

Burch CL, Danaher RJ, Stein DC. 1997. Antigenic variation in *Neisseria gonorrhoeae*: production of multiple lipooligosaccharides. *J Bacteriol*, 179:982–6.

Hood DW, Deadman ME, Jennings MP, Bisercic M, Fleischmann RD, Venter JC, Moxom ER. 1996. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proc Natl Acad Sci USA*, 93: 11121–5.

Jeffreys AJ, Allen MJ, Hagelberg E, Sonnberg A. 1992. Identification of the skeletal remains of Josef Mengele by DNA analysis. *Forensic Sci Int*, 56:65–76.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292:195–202.

Kampke T, Kieninger M, Mecklenburg M. 2001. Efficient primer design algorithms. *Bioinformatics*, 17:214–25.

Meyer E, Wiegand P, Rand SP, Kuhlmann D, Brack M, Brinkmann B. 1995. Microsatellite polymorphisms reveal phylogenetic relationships in primates. *J Mol Evol*, 41:10–14.

Morin PA, Moore JJ, Chakraborty R, Jin L, Goodall J, Woodruff DS. 1994. Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science*, 265:1193–201.

Peak IR, Jennings MP, Hood DW, Bisercic M, Moxon ER. 1996. Tetrameric repeat units associated with virulence factor phased variation in *Haemophilus* also occur in *Neisseria* spp. and *Moraxella catarrhalis*. *FEMS Microbiol Lett*, 137:109–14.

Sreenu VB, Alevoor V, Nagaraju J, Nagarajaram HA. 2003. MICdb: database of prokaryotic microsatellites. *Nucleic Acids Res*, 31:106–8.

Tautz D, Renz M. 1984. Simple DNA sequences of *Drosophila virilis* isolated by screening with RNA. *J Mol Biol*, 172:229–35.