Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

SSR - SIMPLE SEQUENCE REPEATS IDENTIFICATION FROM ESTS

Riju A
Bioinformtics Centre, IISR
riju.aikkal@gmail.com

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

- **Microsatellites**, or **Simple Sequence Repeats** (SSRs), are polymorphic loci present in nuclear and organellar DNA that consist of repeating units of 1-6 base pairs in length.

- They are typically neutral, co-dominant and are used as molecular markers which have wide-ranging applications in the field of genetics, including kinship and population studies.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

- SNP and SSR markers can be rapidly and cheaply identified thorough computational method from EST and other genomic sequence data.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

SSRs are widely distributed throughout the genomes

Found in all prokaryotic and eukaryotic genomes

Three types of classes.

Perfect , imperfect and compound repeats

Perfect repeats are without interruptions.
Eg. ATATATATATATATATATATATATAT

Imperfect repeats are interrupted by non-repeat nucleotides
Eg. ATATATATATATATATATCTATAT

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Compound repeats :
Two or more SSRs are found adjacent to one another.
Eg.
ATATATATATATATATATGCGCGCGCGCATGATGGGGGGGGGGGGG

There may be also combination of this three.

Eg. Imperfect compound repeats.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Types of SSR

1. $(A)_n$, $(T)_n$, $(C)_n$, $(G)_n$ – mononucleotide SSRs
2. $(AT)_n$, $(CG)_n$, $(GT)_n$ – dinucleotide SSRs
3. $(ATT)_n$, $(CCG)_n$, $(GTA)_n$ – trinucleotide SSRs
4. $(CCGG)_n$, $(TATC)_n$ – tetranucleotide SSRs

Compound SSRs:

ATATATATCACACAATATATATCACACA - $(AT)_4(CA)_3$

CCGCCGATATATATCCGCCGATATATAT - $(CCG)_3(AT)_4$

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Types of SSR

Perfect SSR motifs:

AAAAAAAAAAAA - $(A)_{12}$

ATATATATATATATATAT - $(AT)_9$

CCGCCGCCGCCGCCGCCG - $(CCG)_7$

Imperfect SSR motifs:

AAAAA**T**AAAAAAAA - $(A)_{14}$

ATATATATA**C**ATATATAT - $(AT)_9$

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Perfect Microsatellites:

- (ATT)21
- (TTC)32
- (AACAT)11

Compound Microsatellites:

- (AAGA)15(GAAA)14(GGAG)6
- (TCTT)18(TCCT)7(TCTT)7

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Two types of markers

Dominant markers and Codominant markers

Allow the analysis of many loci per experiment within requiring previous information about their sequence.
Eg. RAPD, AFLPs and SMPLs

Codominant markers

allow the analysis of only a locus per experiment

More informative because the allelic variations of that locus can be distinguished
Eg. RFLPs, microsatellites

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Characteristic features of SSR

It's a powerful genetic marker

Codominat, Abundance, dispersal throughout the genome, multi-allelic variation,
high reproducibility and high level of polymorphisms.

High level of polymorphism is due to the mutation affecting the number of repeat units.

It requires small amount of DNA for screening.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

- **SSR demonstrate a high degree of transferability between species**

- **PCR primers designed to an SSR within one species frequently**
- **amplify a corresponding locus in related species**

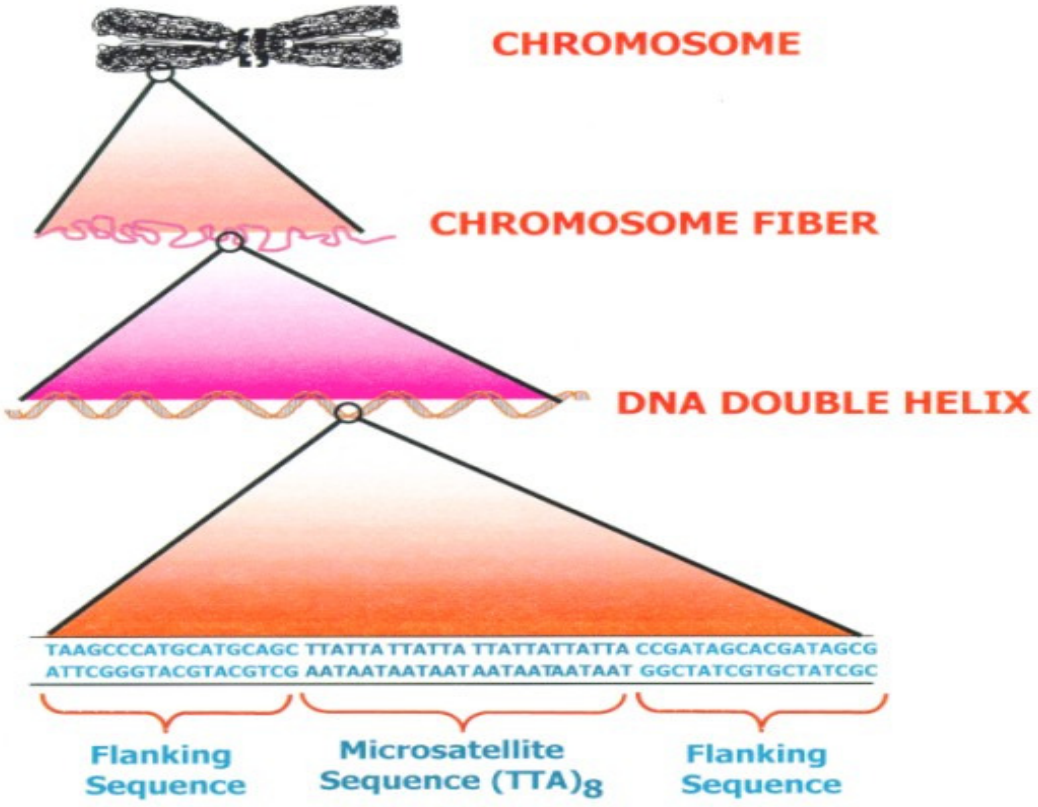
- **Enables comparative genetic and genomic analysis**
- **Used to study gene dosage-looking for duplications or deletions of a particular genetic region**

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

- SSR were initially considered to be evolutionally neutral
- Plays an important role in genome evolution and hotspot of recombination
- Functional roles have been attributed to some SSRs.
- They are thought to be invovled in gene expression, regulation, function, found to bind neclear proteins and function as transcriptional activating elements
- Resides in non-coding regions but has functional significance.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

MICROSATELLITES….



Majority are in non-coding region

Why do they exist?

- "junk" DNA
- A necessary source of genetic variation
- Regulate gene expression and protein function

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Depends on the size of SSR .

Class I – hypervariable markers ..
>20bp or more

Class II - Ranges from >12 but <20

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

MICROSATELLITE MUTATIONS

- **10^{-3} to 10^{-6} events per locus per generation (point mutation 10^{-9} to 10^{-10})**
- **Varies by**
  - **repeat type**
  - **base composition of the repeat**
  - **taxonomic group**
- **most common - addition or deletion of a single repeat**

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Mutation Mechanisms

1. Slippage in DNA at Replication (Slip-Strand Mispairing- SSM)
   - **increases or decreases the repeat by one unit**
2. Recombination
   A. **Unequal crossing over (UCO)**
   B. **Gene conversion**

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

POLYMORPHISM IN SSRs

1. ACGCATATATATATATATATATATATATATATGGATCCA $(TA)_{11}$
2. ACTACGCATATATATATATATATATATATATATGGATCCA $(TA)_{10}$
3. ACTACGCATATATATATATATATATATATATATATGGATCCA $(TA)_{12}$
4. ACTACGCATATATATATATATATATATATGGATCCA $(TA)_{9}$

SIMPLE SEQUENCE REPEAT

How do microsatellites mutate?

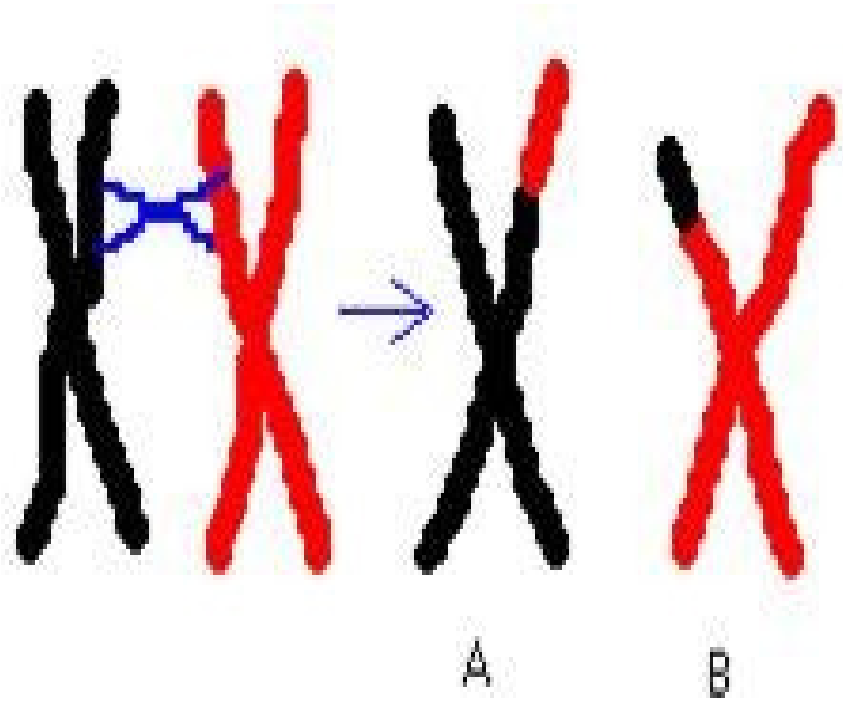
- Replication Slippage
- Unequal crossing-over during meiosis

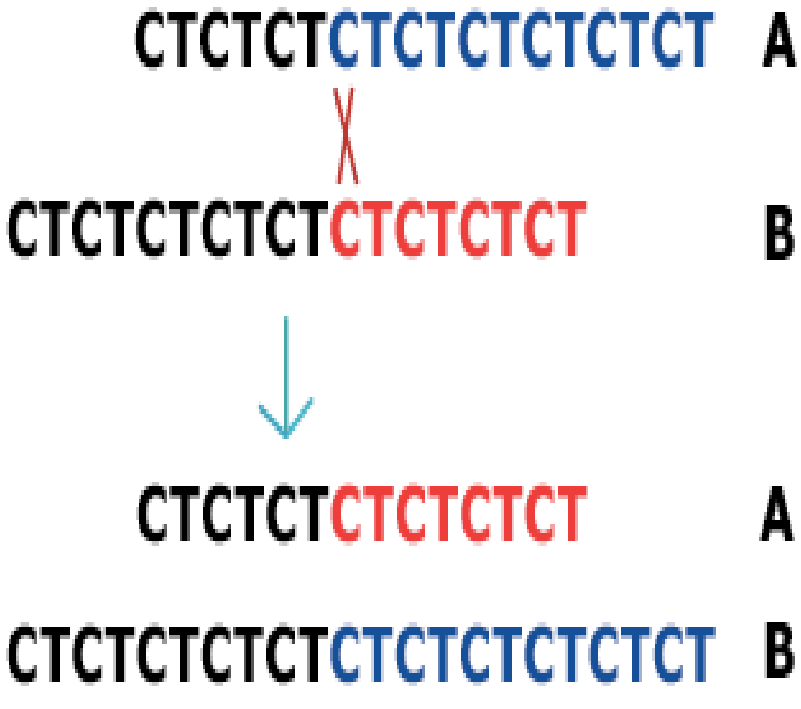
Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

SIMPLE SEQUENCE REPEAT

Unequal crossing-over during meiosis





Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Why SSRs are Preferred Genetic Markers

- Rapid processing is attainable
- Abundant throughout the genome
- Highly variable within various populations
- Small size range allows multiplex development
- Discrete alleles allow digital record of data
- Allelic ladders simplify interpretation
- PCR allows use of small amounts of DNA material
- Small product size compatible with degraded DNA

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Microsatellite Discovery

1. Sequence databases screened for the presence of repeat motifs.

2. Homologous loci can often be amplified with primers from related species.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

SSR tools and softwares

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

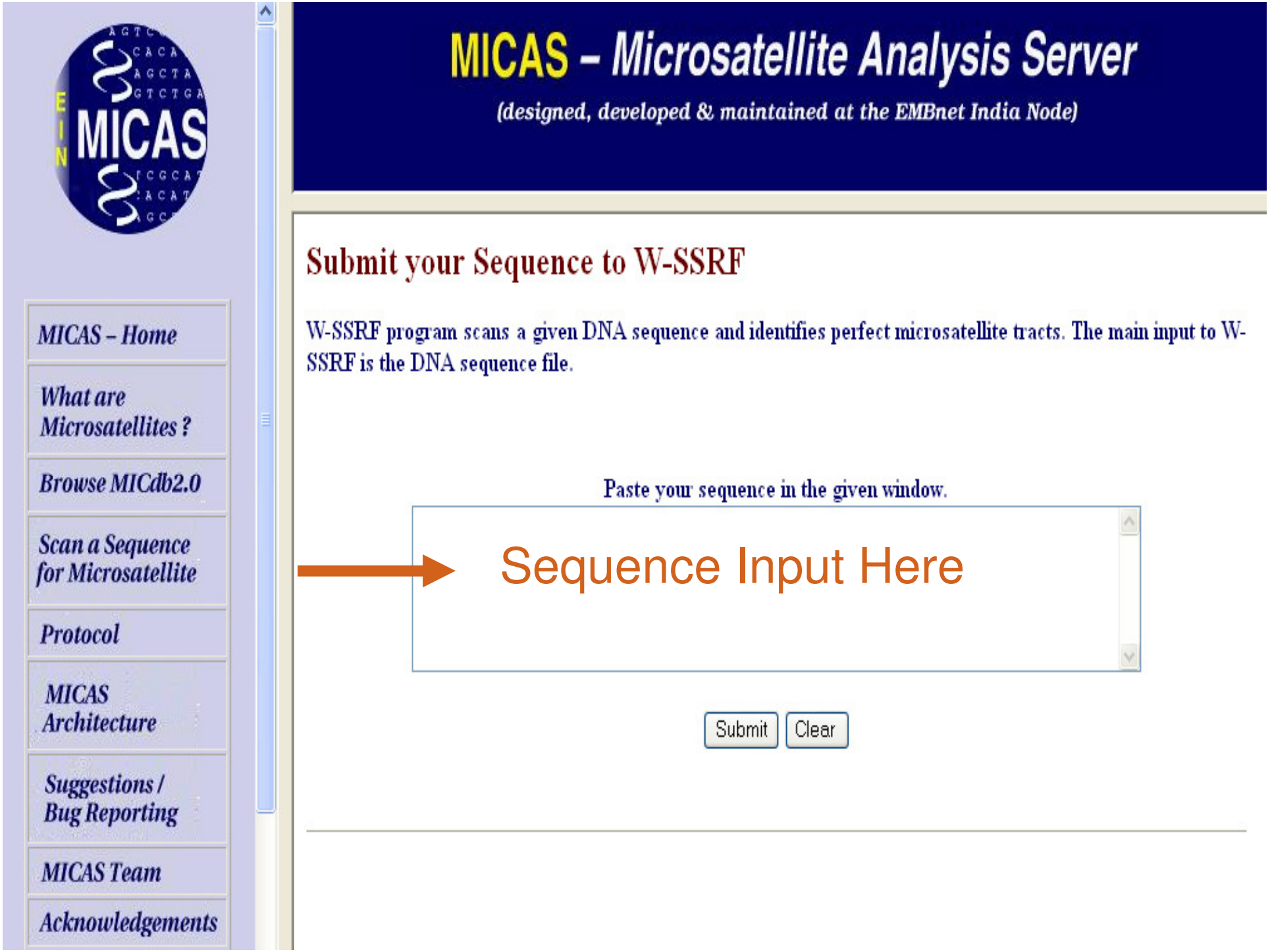
MICAS – MICROSATELLITE ANALYSIS SERVER

http://sunserver.cdfd.org.in:8080/MIC/index.html

- **MICAS is an Interactive user-friendly web-based server to find non-redundant microsatellites in a given nucleotide sequence/genome sequence.**

CDFD, Hyderabad

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009



Microsatellite Analysis Server MICAS

.	Motif	Repeats	Starting Position	EndingPosition
◉	TG	4	45	52
○	GT	2	64	67
○	GT	2	75	78
○	AT	6	81	92
○	TA	4	93	100
○	AT	8	101	116
○	TCT	2	67	72
○	TAT	2	90	95
○	ATA	2	98	103
○	ATATA	2	96	105
○	TATATAT	2	86	99

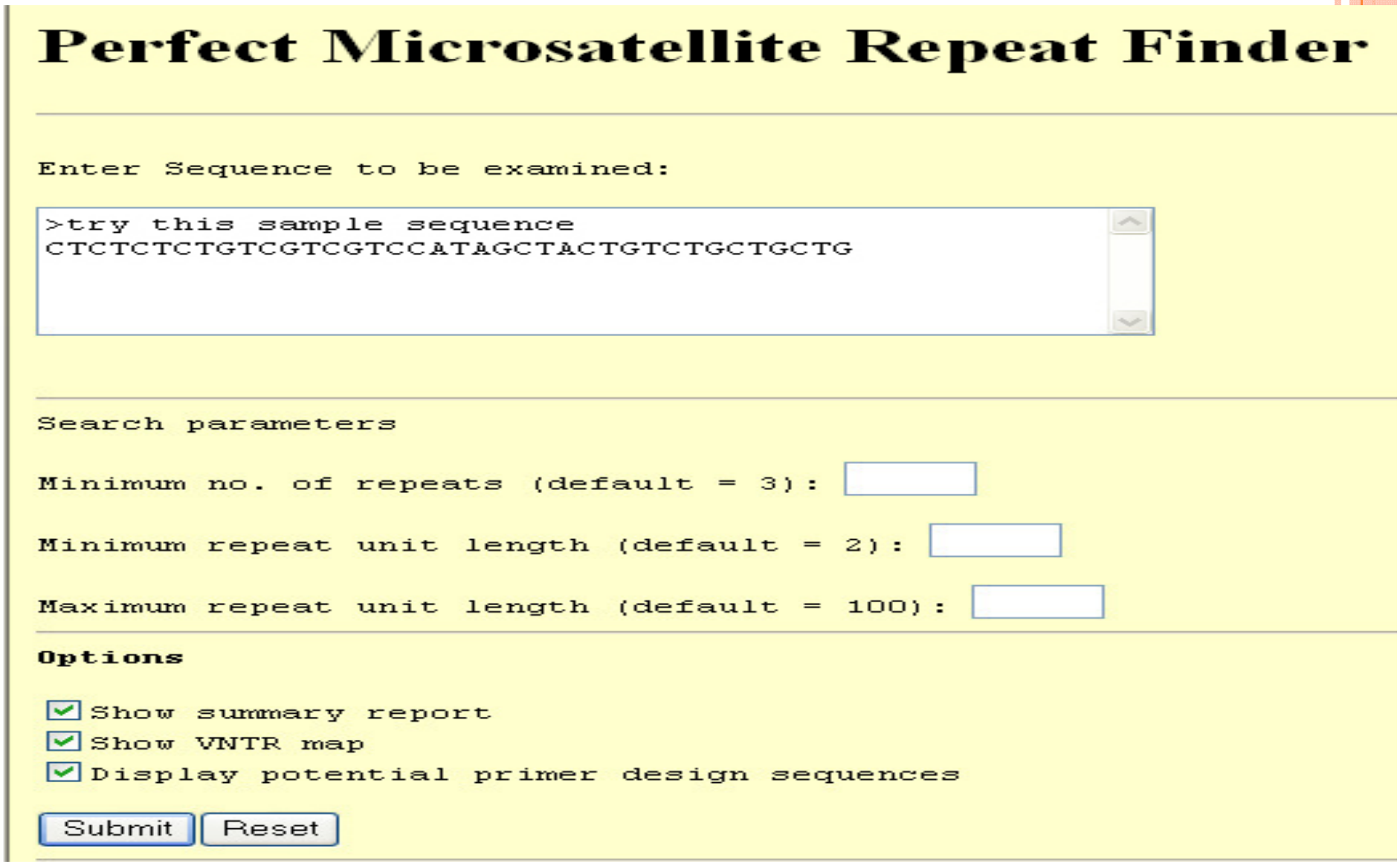
AUTOPRIMER → Get Primer | Reset

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

PERFECT MICROSATELLITE REPEAT FINDER

HTTP://SGDP.IOP.KCL.AC.UK/NIKAMMAR/REPEATFINDER.HTML

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009



Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Search parameters:
Minimum number of repeats: 3
Minimum repeat unit length: 2
Maximum repeat unit length: 100

Summary report

Start	Stop	Unit	Repeat
1	8	CT	4
9	17	GTC	3
30	38	CTG	3

Microsatellite map (VNTR's in UPPERCASE)

CTCTCTCTGTCGTCGTCcatagctactgtCTGCTGCTG

Potential Primer Sequences

VNTR #	Start	Stop	Sequence
1	1	8	CTCTCTCTgtcgtcgtccatagctactgtctgctgctg
2	9	17	ctctctctGTCGTCGTCcatagctactgtctgctgctg
3	30	38	ctctctctgtcgtcgtccatagctactgtCTGCTGCTG

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

MICROSATELLITE REPEATS FINDER

HTTP://WWW.BIOPHP.ORG/MINITOOLS/MICROSATELLITE_REPEATS_FINDER/DEMO.PHP

Microsatellite repeats finder

Sequence:

```
AACAATGCCATGATGATGATTATTACGACACAACAACACCGCGCTTGACGGCGGCGGATGGATGCCG
CGATCAGACGTTCAACGCCCACGTAACGTAACGCAACGTAACCTAACGACACTGTTAACGGTACGAT
```

Length of repeated sequence: Minimum: 2 Maximum: 6

Minimum number of repeats: 3

Minimum length of tanden repeat: 6

Allowed percentaje of mismatches: 0

Find Microsatellite repeats

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Posición	Cicle	Repeats	Sequence
9	3	3	ATGATGATG
29	3	3	ACAACAACA
48	3	3	CGGCGGCGG

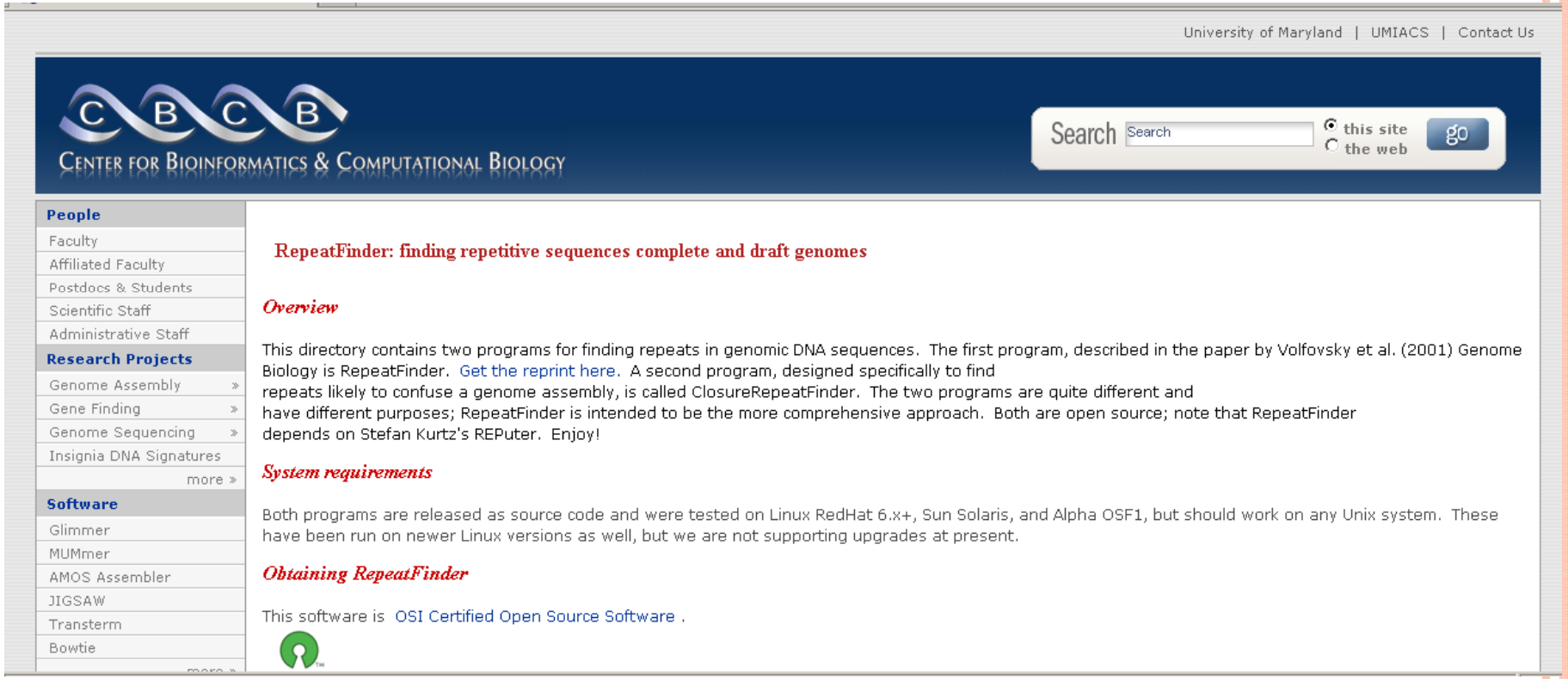
Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Welcome to Tandem Repeats Finder

New TRDB is a public database of tandem repeats that allows users to run their own sequences. It has many added features not available in TRF.

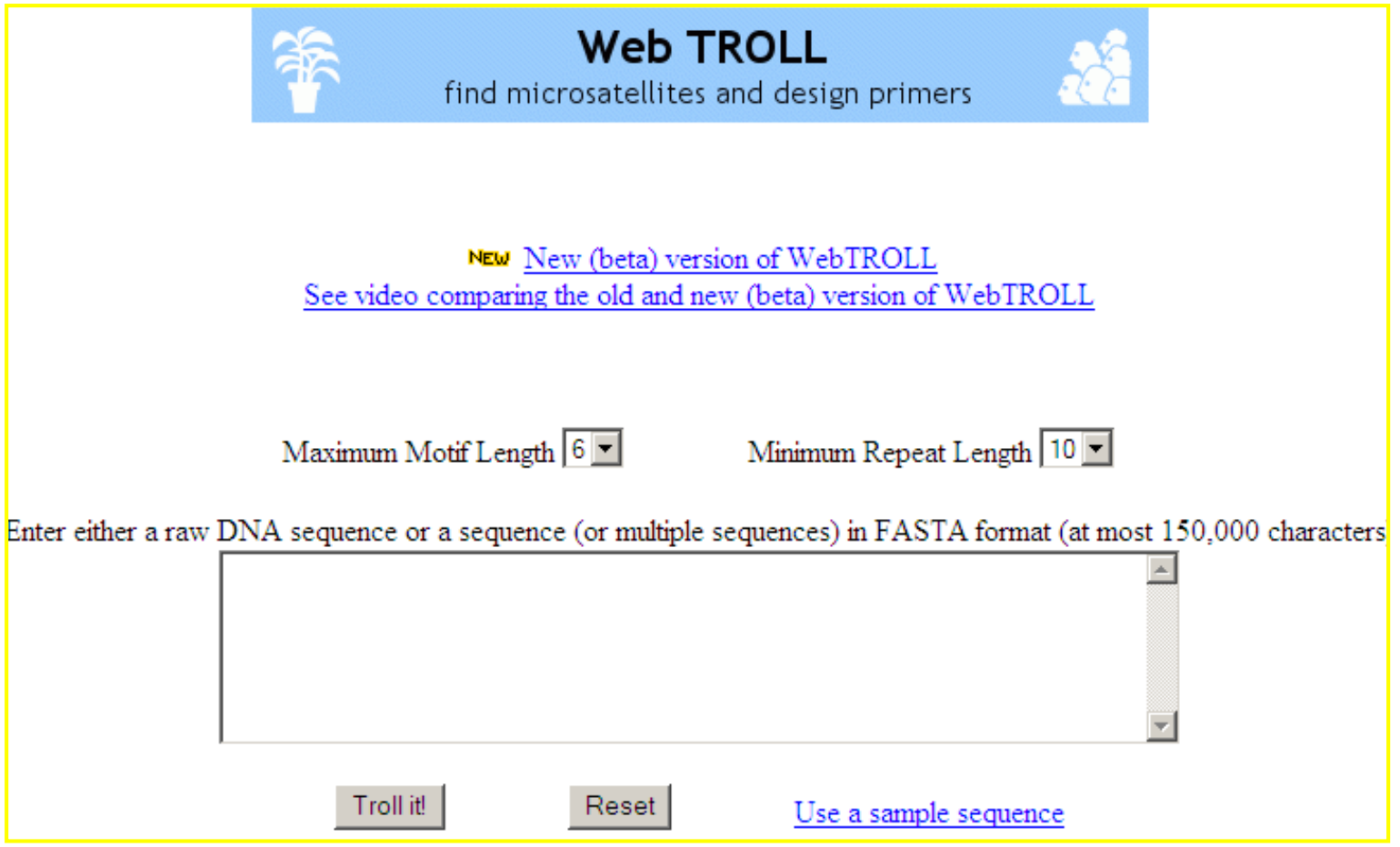
- Submit a Sequence for Analysis
- Download Your Own Copy of the Program
- Find Out How TRF Algorithm Works
- Find Out What's New
- Add Your Email to Our Notification List

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009



Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

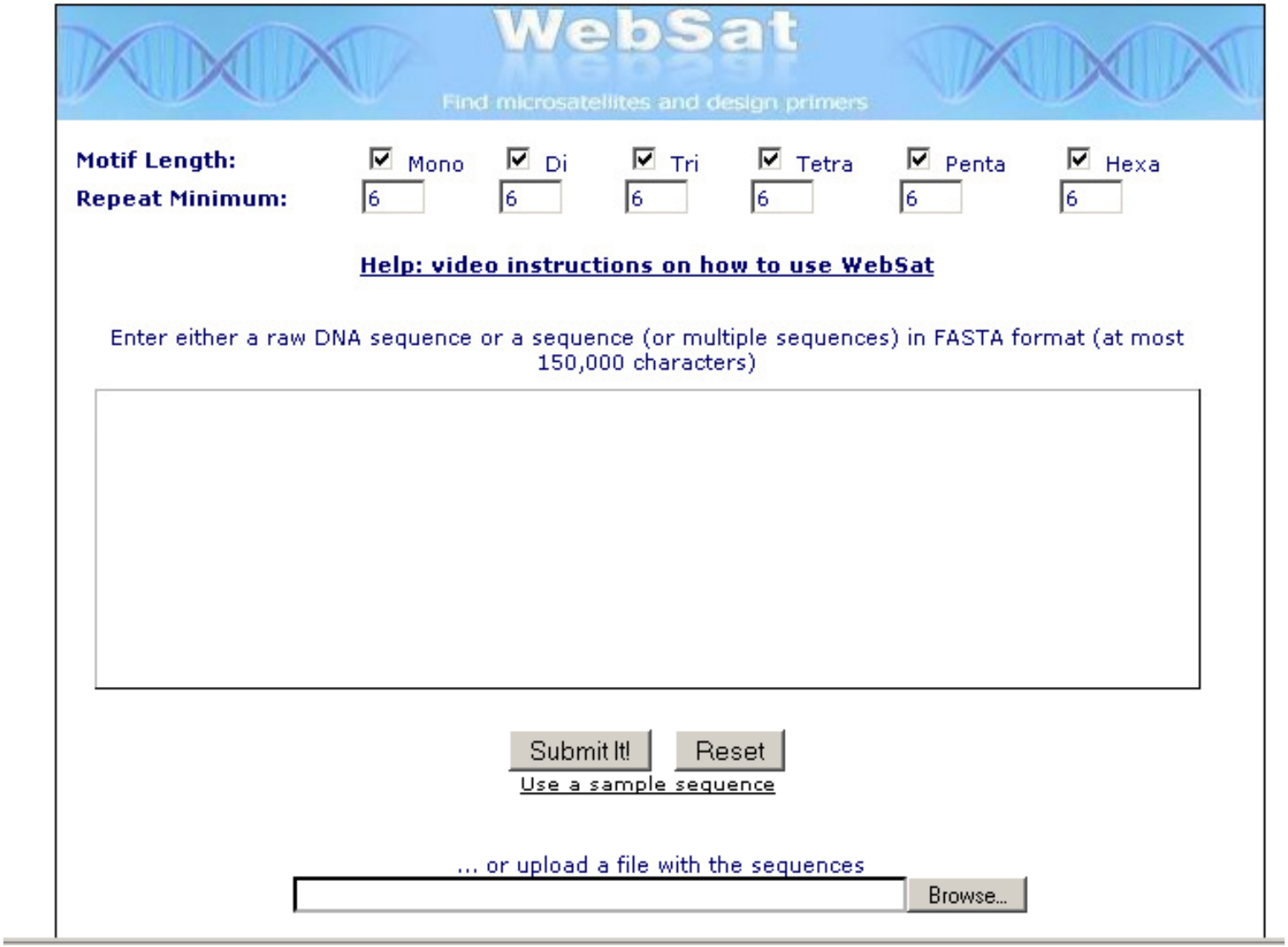
http://wsmartins.net/webtroll/troll.html



A. **Castelo, W.Martins and G. Gao (2002). TROLL - Tandem Repeat Ocurrence Locator, Bioinformatics Journal.**

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009



Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

msatcommander

locate microsatellite repeats, design locus-specific primers, and tag 'em

News

8/21/09: 0.8.2 binaries are in the download area. See Changelog for fixes and an enhancement.

8/21/09: New changes uploaded to SVN, updating the code (mostly for windows) to deal with Python 2.5, BioPython 1.51, and wxPython. Additionally, added mispriming library to remove issues where microsatellite repeats were selected as proper primer binding sites.

Feedback

If you have constructive feedback (likes, dislikes, enhancements), please send an email to the list-serve. All suggestions are welcomed and will be considered.

Purpose

msatcommander is a python program written to locate microsatellite (SSR, VNTR, &c) repeats within fasta-formatted sequence or consensus files. msatfinder will search for all di-, tri-, tetra-, penta-, and hexa-nucleotide repeats (with options to search for fewer repeat types and combinations of repeat types).

msatcommander will also design and tag primers using primer3 as its primer design engine (see References for citations). Many thanks to the primer3 team, in general, and Steve Rozen, in particular. Additional information regarding primer3 may also be found on the primer3 wiki.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

University School of Biotechnology

Guru Gobind Singh Indraprastha University, Delhi

EuMicroSat*DB*
(Eukaryotic Microsatellite Database)

How to use EuMicroSat*db*

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

At present EuMicroSat*db* has microsatellite data of the following species:

Saccharomyces cerevisiae	*Apis mellifera*
Schizosaccharomyces pombe	*Tribolium castaneum*
Aspergillus oryzae RIB40	*Oryza sativa* ssp. *japonica*
Aspergillus fumigatus	*Oryza sativa* ssp. *indica*
Cryptococcus neoformans var *JEC21*	*Arabidopsis thaliana*
Encephalitozoon cuniculi	*Ciona intestinalis*
Eremothecium gossypii	*Tetraodon nigroviridis*
Candida glabrata CBS138	*Danio rerio*
Debaryomyces hansenii	*Rattus norvegicus*
Kluyveromyces lactis	*Mus musculus*
Yarrowia lipolytica	*Gallus gallus*
Caenorhabditis elegans	*Canis familiaris*
Plasmodium falciparum	*Macaca mulatta*
Anopheles gambiae	*Bos taurus*
Drosophila melanogaster	*Pan troglodytes*
	Homo sapiens

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Searching Parameters

- **Repeat Unit Length (mono-, di-, tri-, tetra-, penta-, hexa-)**
- **Repeat Sequence**

  The sequence of the microsatellite e.g. GCC
- **Repeat No.**

  $(AT)_{10}$ has a repeat number 10
- **Microsatellite Length**

  (ATT)20 has a motif length 60
- **Position**

  Where the microsatellite is located on the chromosome
- **Interruption Size (Microsatellite Cluster)**

  Difference between two adjacent microsatellites

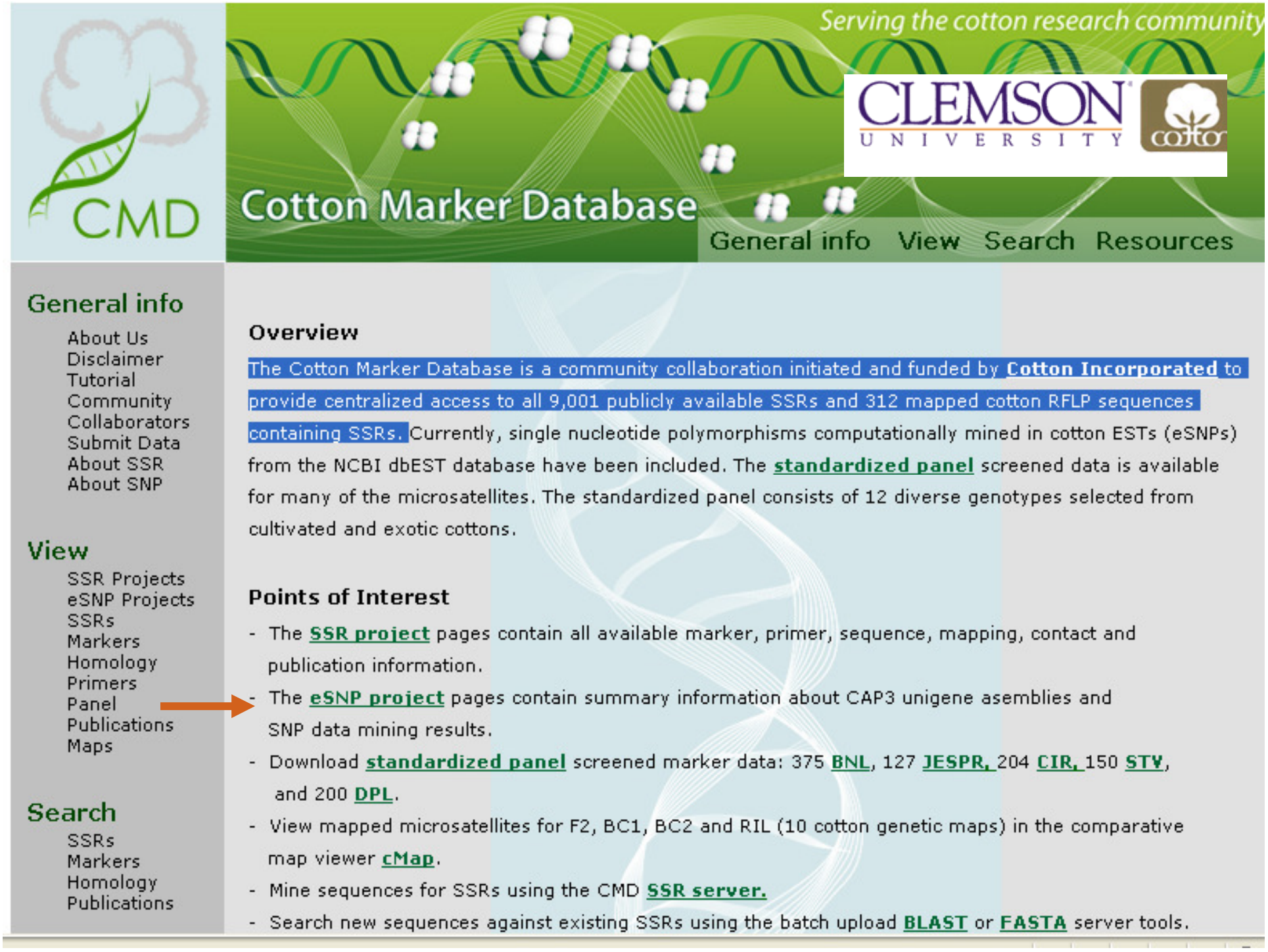
COTTON MARKER DATABASE

http://www.cottonmarker.org/

- The Cotton Marker Database is a community collaboration to provide centralized access to all 9,001 publicly available SSRs and 312 mapped cotton RFLP sequences containing SSRs.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009



INSATDB: INSECT MICROSATELLITE DATABASE

HTTP://210.212.212.8/PHP/INSATDB/MSATFORM.PHP

- **InSatDb**, presents an interactive interface to query information regarding microsatellite characteristics *per se* of five fully sequenced insect genomes (fruit-fly, honeybee, malarial mosquito, red-flour beetle and silkworm).
- **InSatDb** allows users to obtain microsatellites annotated with size (in bp and repeat units); genomic location (exon, intron, up-stream or transposon); nature (perfect or imperfect); and sequence composition (repeat motif and GC%).
- **InSatDb** is complete with the insects information, web links to find details, methodology and a tutorial. A separate *Analysis* section illustrates the comparative genomic analysis that can be carried out using the InSatDb output.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Insect :

○ *Bombyx mori* ○ *Anopheles gambiae* ○ *Drosophila melanogaster* ○ *Apis mellifera* ○ *Tribolium castaneum*

Location :

☐ Intron ☐ Exon ☐ I-E boundary ☐ Upstream(1000 bp) ☐ Intergenic ☐ Repeat elements ☐ All

Microsatellite characteristics :

1. ◉ Repeat type All **OR** ○ Repeat motif [ ] Enter motif and Click

[ ]

2. GC% [ ] ≥ [ ] ≤ [ ]

UGMICROSAT*DB*

DATABASE FOR MINING MICROSATELLITES FROM UNIGENES

HTTP://WWW.VEENUASH.INFO/WEB1/INDEX.HTM

- **UgMicroSat*db* (Unigene MicroSatellite *d*ata*b*ase), a web based relational database of microsatellites present in unigene sequences covering 80 genomes.**
- **UgMicroSat*db* allows microsatellite search using multiple parameters like microsatellite type simple (perfect) and compound (perfect and imperfect), repeat unit length (mono- to hexa-nucleotide), repeat number, microsatellite length and repeat sequence class.**
- **Microsatellites can also be retrieved by specifying EST, cDNA, CDS identity or by using Gene Index, GenBank, UniGene IDs.**
- **The database also provides information about trinucleotide repeats encoding various amino acids. Such codon repeats can be searched by specifying characteristics of coded amino acids like charge (basic, acidic or neutral), polarity (polar or non-polar) and their hydrophobic or hydrophilic nature.**
- **The nucleotide sequences of the target UniGenes are also provided to facilitate primer designing for PCR amplification of any desired microsatellite. UgMicroSat*db* is available at http://ipu.ac.in/usbt/UgMicroSat*db*.htm.**

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

	|_____Search_____|	
☐	**Repeat Unit Length**	Mononucleotide
☐	**Repeat Sequence**	
☐	**EST/mRNA/cDNA/CDS name (Functional annotation)**	
	Structural annotation	☐ EST ☐ mRNA ☐ cDNA ☐ CDS
☐	**GI/GB/UG IDs**	
☐	**Repeat Sequence Class (Tri-repeats_only)**	
☐	**Amino Acid [codon repeat]**	Alanine
☐	**Amino Acid (Charge) [codon repeat]**	Basic
☐	**Amino Acid (Polarity) [codon repeat]**	Polar
☐	**Amino Acid (Hydro-) [codon repeat]**	Hydrophobic
☐	**Other types [codon repeat]**	Amino Acids with Aliphatic R-Groups
☐	**Microsatellite Length**	Equal To

submit Reset

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

|_UgMicroSat_ID_|	Microsatellite	Microsatellite Length(bp)	|UG_Len|	|UG_Seq|	|IDS|	|_____
UMD-7-1-1-264	(CAG)5	15	883	Sequence	gb=BY955066 /gi=100403716 /ug=Ppa.1090	"gnl|UG|Ppa# Physcomitrella mRNA seque
UMD-7-1-1-490	(CAG)6	18	798	Sequence	gb=BY947547 /gi=100362825 /ug=Ppa.3571	"gnl|UG|Ppa# Physcomitrella mRNA seque
UMD-7-1-1-589	(CAG)7	21	861	Sequence	gb=BY948133 /gi=100369525 /ug=Ppa.4339	"gnl|UG|Ppa# Physcomitrella mRNA seque
UMD-7-1-1-601	(CAG)5	15	677	Sequence	gb=BY963326 /gi=100442831 /ug=Ppa.4425	"gnl|UG|Ppa# Physcomitrella mRNA seque
UMD-7-1-1-673	(CAG)5	15	846	Sequence	gb=BY949607 /gi=100377956 /ug=Ppa.5102	"gnl|UG|Ppa# Physcomitrella mRNA seque
UMD-7-1-1-754	(CAG)5	15	886	Sequence	gb=BY946150 /gi=100356009 /ug=Ppa.6222	"gnl|UG|Ppa# Physcomitrella mRNA seque
UMD-7-1-1-1105	(CAG)6	18	810	Sequence	gb=BY993357 /gi=100418147 /ug=Ppa.8239	"gnl|UG|Ppa# Physcomitrella mRNA seque

MICROSATELLITE REPEATS DATABASE

HTTP://INSILICO.EHU.ES/MICROSATELLITES/INFO.HTML

- Microsatellite Repeats finder has been used for detection of microsatellites in all up to date sequenced prokaryotic chromosomes.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

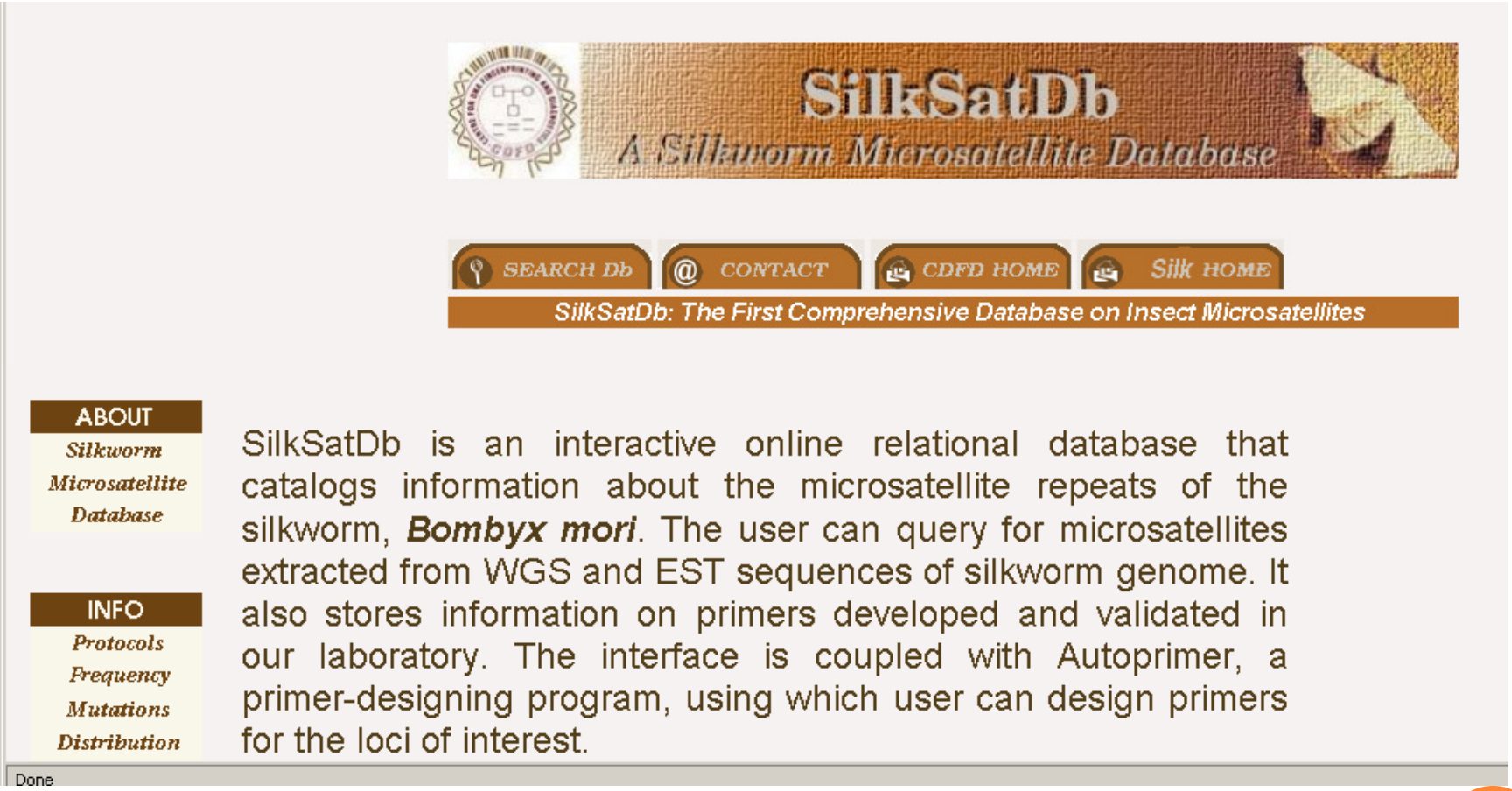
MMDBJ: Mouse Microsatellite Data Base of Japan

http://www.shigen.nig.ac.jp/mouse/mmdbj/top.jsp

- Mouse Microsatellite Data Base of Japan (MMDBJ) provides information about SSLPs among different mouse strains, focusing on strains derived from Japanese wild mouse, *Mus musculus molossinus*, which are genetically remote from the standard laboratory strains.
- Two strains, MSM and JF1, are now widely used in gene mapping because of the high level of microsatellite polymorphisms for the standard laboratory strains and high reproduction ability in crosses with other strains.
- This data base includes PCR conditions for all entries of primer sets and keyword searches for the information.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009



Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

SilkSatDb: A Silkworm Microsatellite Database

http://210.212.212.7:9999/PHP/SILKSAT/index.php

- SilkSatDb is an interactive online relational database that catalogs information about the microsatellite repeats of the silkworm, ***Bombyx mori***.
- The user can query for microsatellites extracted from WGS and EST sequences of silkworm genome.
- It also stores information on primers developed and validated in our laboratory.
- The interface is coupled with Autoprimer, a primer-designing program, using which user can design primers for the loci of interest.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NATURAL ENVIRONMENT RESEARCH COUNCIL

Small Genomes Microsatellite Database

Molecular Evolution and Bioinformatics Section

CEH Bioinformatics Centre

Introduction

These data were generated as part of a survey of microsatellite sequences in small genomes from GenBank. A companion database, GeneSwytch, contains curated information on microsatellites thought to be potential contingency loci. Information found in this database was generated with the Msatfinder program.

Data on genomes and microsatellites are included for:

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Small Genomes Microsatellite Database

Molecular Evolution and Bioinformatics Section

CEH Bioinformatics Centre

Database Log (microsatellites)

Categories of files in the Small Genomes Microsatellite Database

- ☐ Bacterial genomes
- ☐ Plasmid genomes
- ☐ Viral genomes
- ☐ Viroid genomes
- ☐ Phage genomes
- ☐ Organellar genomes

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

MISA

MIcroSAtellite identification tool (MISA)

Tool allows the identification and localization of perfect microsatellites.

Find Compound microsatellites also.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

MISA

MISA - MIcroSAtellite identification tool

This tool allows the identification and localization of perfect microsatellites as well as compound microsatellites which are interrupted by a certain number of bases.

In order to design primers flanking the microsatellite loci, two perl scripts serve as interface modules for the program-to-program data interchange between MISA and the primer modelling software Primer3 (Whitehead Institute). For installation instructions of Primer3 see http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Get MISA

Supplemental perl tools:

Get est_trimmer.pl - perl script useful for pre-preprocessing sequences (invocation without parameters is explaining the syntax)

Get Primer 3 interface modules - perl scripts allowing the interaction with Primer3

Questions and comments: Thomas Thiel

Last updated: 5/14/02

http://pgrc.ipk-gatersleben.de/misa

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

SSRIT

SSRIT - Simple Sequence Repeat Identification Tool

This tool finds all *perfect* simple sequence repeats (SSRs) in a given sequence.

For source code for a stand-alone version please click here.

For citation, please use this reference *Temnykh et al.* (2001).

Note: Netscape 2.0 or greater, or IE 4.0 or greater required.

---

1) Select search parameters

a) Select the maximum motif-length group you wish to find. For example, if you want to search for all SSRs up to and including pentamers (meaning, you'd like to search for dimers, trimers, tetramers, and pentamers), you should select 'pentamer' from the drop-down menu.

tetramer

b) Enter the minimum number of repeats you will allow. Entering a '5,' for example, will match SSRs with five or more motif repeats, such as ag-5 ('agagagagag').

5

2) Paste/Enter your sequence of interest into the textarea

The sequence(s) must be in FASTA format - meaning, there must be a title line with a '>' at the beginning for each sequence.

FOR EXAMPLE,

```
>seq1
agagattaggatcgatcgcgctctctctctctctcgatcgagatcgat
ggccatcatcatcatcatcattgagatatagcgcgatatcgagagatctc
agaatagatatcgcgctatagagagatcgagagagagtaga
>seq2
agagataggaatatgagatagcgggggggggggggcgctatacgcgctcg
gagagagatctctctctctcttatagagatcgatcgactagctagatata
agactcactcactcactcactcagcgcgat
```

Paste/Enter your sequence(s) here:

http://www.gramene.org/db/searches/ssrtool

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

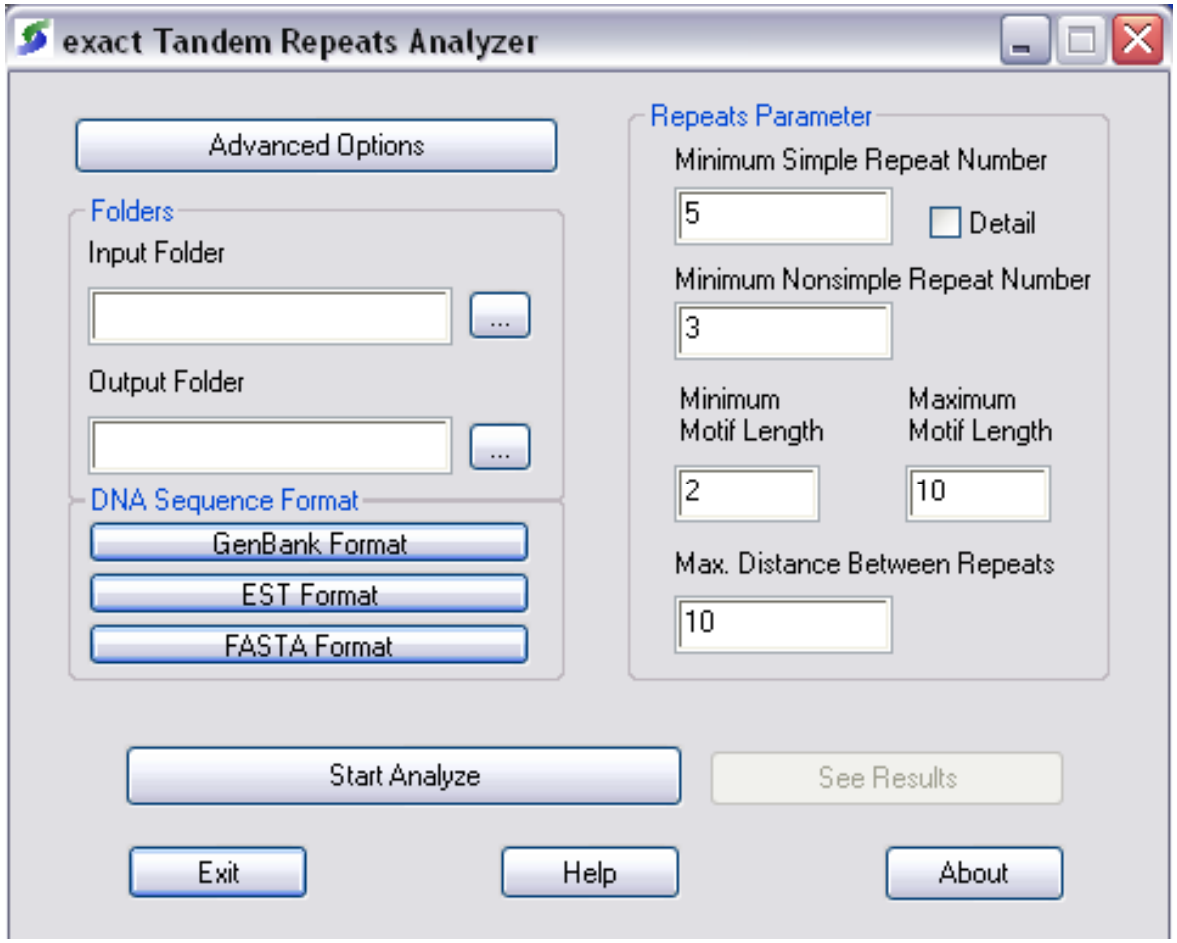
SSR IDENTIFICATION TOOLS

ETRA

Exact Tandem Repeats Analyzer 1.0 (E-TRA) combines sequence motif searches with keywords such as 'organs', 'tissues', 'cell lines' and 'development stages' for finding simple exact tandem repeats as well as non-simple repeats.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

ETRA



Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

ACPFG Bioinformatics

SSRPrimerII

Home
Help

Welcome to SSRPrimerII

SSRPrimerII accepts FASTA format sequences, searches for SSRs using SPUTNIK and designs PCR primers for the identified SSRs using Primer3. The process is described in:

Jewell E, Robinson A, Savage D, Erwin T, Love CG, Lim GAC, Li X, Batley J, Spangenberg GC and Edwards D. (2006) SSR Primer and SSR Taxonomy Tree: Biome SSR discovery. *Nucleic Acids Research* **34**: W656.W659.

Sequence entry is currently limited to 4Kbp. Longer sequences are discarded without analysis.

Please contact Dave.Edwards@uq.edu.au for further details or to report problems.

1

Sequence data:

Either: Select the sequence file to upload:

Browse...

Otherwise: Enter a sequence in FASTA format:

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Applications of EST SSR

EST -SSR represent a potentially valuable source of gene-based markers for population genetic analyses.

Rapid and inexpensive development and high levels of cross-taxon portability.

Potential to facilitate evolutionary analyses in a wide variety of taxa, and may well represent the best way forward for the analysis of species for which only limited resources are available.

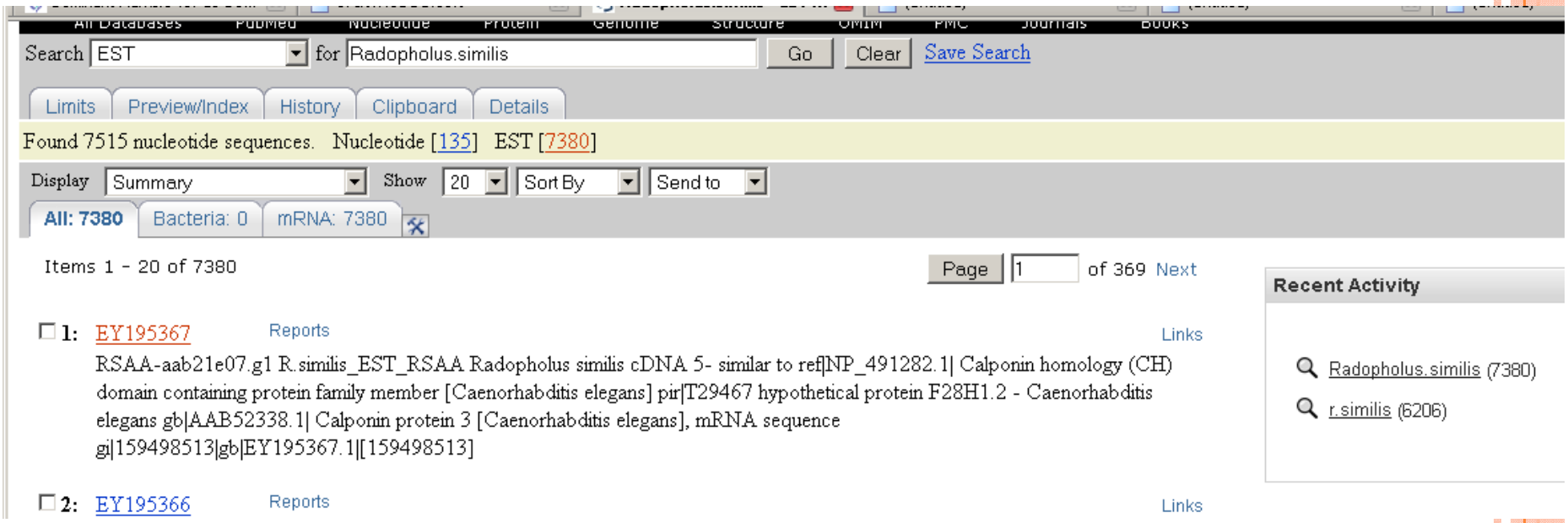
EST-SSR is the best resource for developing chromosome mapping of several crops. Linkage map can create with the help of MAPMAKER software

EST-SSR is used in the development of markers in phylogenetic analysis.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Case study using Radopholus similis ESTs

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009



7380 ESTs in dbEST

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

>gi|159498513|gb|EY195367.1|EY195367 RSAA-aab21e07.g1 R.similis_EST_RSAA Radopholus similis cDNA 5' similar to ref|NP_
GGCCGGGTCGCGAAGAGCCAGGGAGTGCCCACCGAGGAGACCTTCCAGAGCGTGGACTTGTTCGAGGCAC
GTGACCTCTACTCCGTGTGCATGACCCTGTTGTCGTTGGGCCGAATTATGGAGAAGAAGGGAAAGCCGAA
CCCATTCTCTGGATGAAGAGTAGAAGTGAGTGCAGCAAAAGACGGACAGAGCATGTGCTATCGCTCCCAT
TCGAGAACTCCCCGTTTTGCGAATTTTCTCCCCGTGTGCGACCTTGCAACAATACCGAGGCGTTAACTGT
TTTCCCCTCTCTCTCCTCTCAAACTGTGGGCATTTGAAAATAGCGATGCCGGAATAAATGGCCAATTCCA

>gi|159498512|gb|EY195366.1|EY195366 RSAA-aab21e06.g1 R.similis_EST_RSAA Radopholus similis cDNA 5', mRNA sequence
GGCCGGGATTGGCACGTAATTCAACTGGTCTTTTTGCTGGGCGGCAATTGGCAACTTTATCCGACAGAAG
CATGATAATTGGTGGAGCCGCCCTTTGCCGACGCACTTTTGCAGCCCCACCCCCCCATCGGGACACGGTA
TTCGAGGTAGAATCGGACGAGGACTTCGAGAATGGGGTATACCATGCCGAAAAACCAGTGCTCATGCAGT
TTTACGCCGATTGGTGTGGACCCTGTCAGAATTTGGCACCGAGGTTGATTGCCAAAGTGAATGGACAGGA
TGGGAAGGTATTGCTAGCGAGAGTGAATGTAGAGGGTTCCGCCGGATTTCTCGCGGAACAGTTTGATGTA
AGCTCGATTCCCACTGTGATGTGCTGGCTGCGAGGAGAGGTGGTTGACCGTTTTGAGGGCGACGTGGAAG
ACACGAAGATTGACCAAATCATATCCAAATTGGTGGAATATCAAACTGGAAATGAATGATAAAGGGCTTG
AACAGCCATTTAGTAGACGAAAAAAAAAAAAAAAA

>gi|159498511|gb|EY195365.1|EY195365 RSAA-aab21e05.g1 R.similis_EST_RSAA Radopholus similis cDNA 5', mRNA sequence
GGCCGGGTGCTGCAGGCATGTCGCACCGTCCCGCCACTCATCATCGTCCCTTCCCTGTCCCGCCCCGGCA
CACCCTTCCTGTCCAGACATTCAACTGCGGTGTGGTCGAAGTCTCTCCCAAGTCAATACCCGATGCTCCG
CCGCCATACGAGGAGTTTGTGCGTGTTCCACCACCACCGCCACAAAGGGCACCGCCCATTCTGACGCGGG
AAGAGGATGAGGAGTTGCAGAACAGACTGAACTCGGAGAGAGAGAGGGAACTGAGCGACTGGTGACATTT
GGTTTTGTCGAGTGCTGCAGCTTCGCACCATTTCCCTTTATATACGGGACTTTCTTCATTTCTTTTGTTC
CTGACTTAACAACAATTAATAGACCA

>gi|159498510|gb|EY195364.1|EY195364 RSAA-aab21e04.g1 R.similis_EST_RSAA Radopholus similis cDNA 5' similar to ref|NP_
GGCCGGGGGACGCCATTGTATTTTTGGTCGATGTAGCCGACCTGGAACGTATTCAGGAAGCAAGGGAGGA
ATTGTGGAGTCTGATGCAGGATGAACAGGTGGCAAGTGCACCTGTGCTTGTTTTGGGCAATAAGATCGAC
AAGCCGAATGCTCTCAGCGAAGACCAGCTCAAGTACTACCTCGGCATCCAACAATACTGCACAGGAAAAG
GCCAAGTTGCGCGCTCAGATCTGGCCACTCGTCCTTTGGATGTGTTCATGTGCTCAGTCCTTAGGCGACA
GGGTTACGGCGAAGGATTCTGTTGGCTCTCACAATATCTGGACTGATTGAACGCGCCTCGGAAGTTGAAA

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

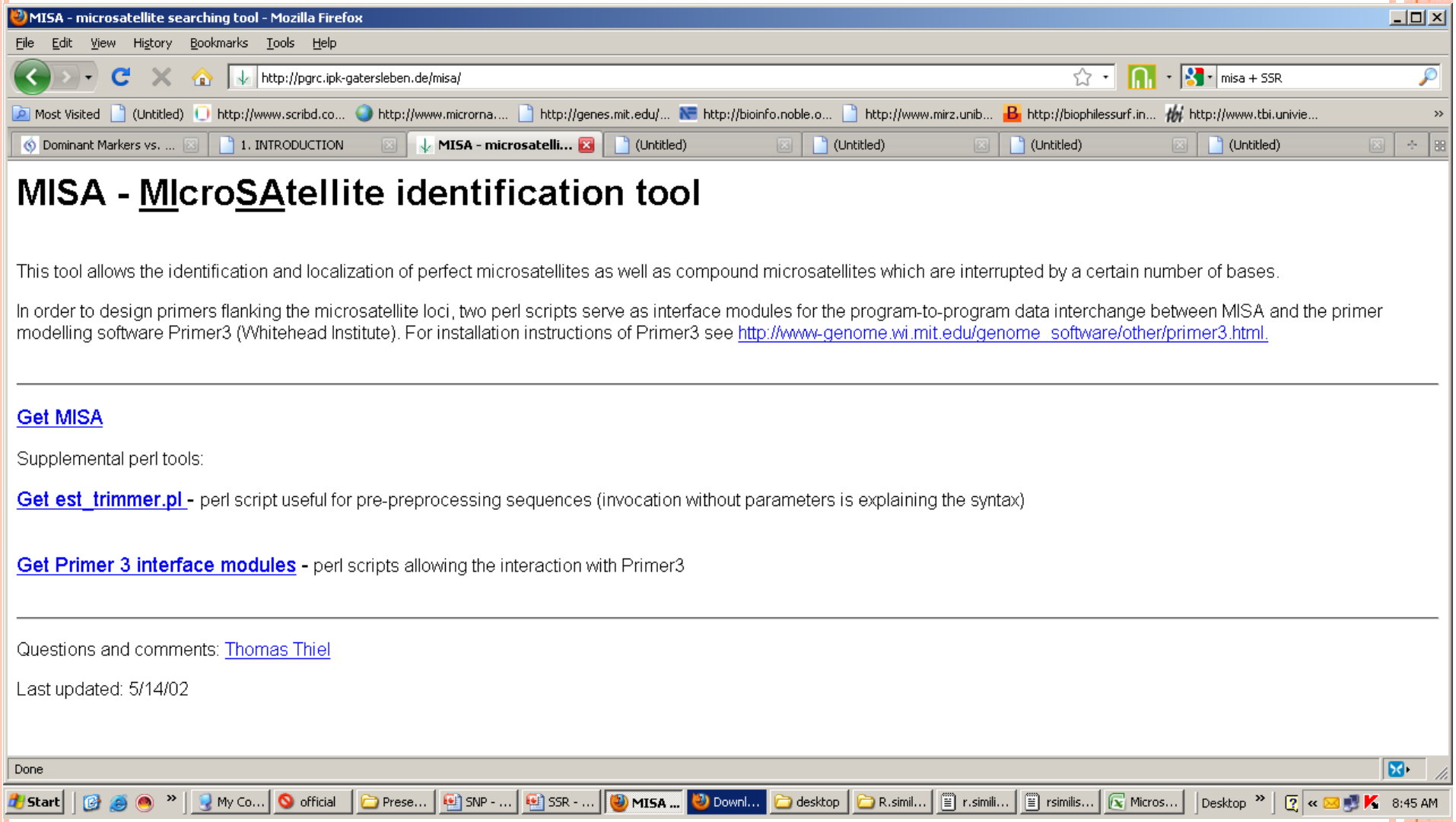
Total No of ESTs - 7380
Contigs -
Signlets -

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

SSR detection using MISA

MISA - MIcroSAtellite identification tool

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009



Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

MISA - MIcroSAtellite identification tool

Download MISA

Download misa.ini - example of a specification file

Syntax: misa.pl **<FASTAfile>**

Declarations:

- <FASTAfile> is the filename containing DNA sequences in FASTA format.
- An additional file containing the search parameters is required named "misa.ini".
  - In a single line beginning with "def" a sequence of number pairs is expected, whereas the first number defines the unit sizes and the second number the lower threshold of repeats for that specific unit.
  - In a single line beginning with "int" a single number is expected defining the maximal number of bases between two adjacent microsatellites to be recognised as being a compound microsatellite.

http://pgrc.ipk-gatersleben.de/misa/misa.html

misa.ini Identification parameters

definition(unit_size, min_repeats): 1-10 2-6 3-5 4-5 5-5 6-5

interruptions(max_difference_between_2_SSRs): 100

definition(unit_size,min_repeats): 1-12 2-6 3-5 4-3 5-3 6-2

interruptions(max_difference_between_2_SSRs): 100

Monomer minimum 12 bp

Dimer minimum 12 bp

Trimer 15bp
Tetra 12 bp
Penta 15bp
Hexa 12 repeats

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

syntax

misa.pl **<FASTAfile>**

misa.pl rsimilis.fasta

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

A1 ID

	A	B	C	D	E	F	G	H	I
1	ID	SSR nr.	SSR type	SSR	size	start	end		
2	EY194446	1	c	(GGCCGG)	48	1	48		
3	EY194450	1	c	(CGGGAA)	164	4	167		
4	EY190719	1	c	(GGGAAC)	102	7	108		
5	EY191483	1	c	(GCAACA)	121	8	128		
6	Contig56	1	c	(ATCACA)	90	9	98		
7	EY192307	1	c	(GGA)5cga	36	10	45		
8	EY191149	1	c	(AAATTA)	34	11	44		
9	EY190548	1	c	(T)18catcc	106	11	116		
10	Contig812	1	c	(T)17caact	49	12	60		
11	EY191543	1	c	(A)15ctaaa	46	13	58		
12	Contig727	1	c	(GTCATT)2	50	15	64		
13	EY195239	1	c	(TCAAAT)1	85	15	99		
14	EY191977	1	c	(AAGGCA)	89	19	107		
15	EY195150	1	p6	(TTTAAA)2	12	20	31		
16	EY191512	1	p6	(TGGATG)	12	20	31		
17	EY191257	1	p6	(TGATGC)1	12	21	32		
18	EY190935	1	p6	(TCCCAA)1	12	21	32		
19	CO897777	1	p6	(TAAGCT)1	12	21	32		
20	EY193889	1	p6	(TCTTAC)2	12	23	34		
21	EY192676	1	p6	(TAATTT)2	12	23	34		
22	EY194162	1	p5	(TAAAT)3	15	24	38		
23	EY190902	1	p5	(AAGCT)3	15	24	38		

Chart4 Chart3 Chart2 Chart1 rsimilis.fasta

Ready

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

```
RESULTS OF MICROSATELLITE SEARCH
================================

Total number of sequences examined:                  7380
Total size of examined sequences (bp):               1706771
Total number of identified SSRs:                     465
Number of SSR containing sequences:                  413
Number of sequences containing more than 1 SSR:      41
Number of SSRs present in compound formation:        32


Distribution to different repeat type classes
---------------------------------------------

Unit size    Number of SSRs
1      408
2      15
3      37
4      5
```

Frequency and most abundant repeats were found out.

For mononucleotides, although A, T, C and G are possible,

A and T are grouped into a single category, G and C since an A repeat on a strand is the same as a T repeat on the opposite strand, and a C on a strand is the same as a G on the opposite strand, resulting in two unique classes of mononucleotides,

A/T and C/G

all dinucleotide motifs were grouped into the following four unique classes

(i) AT/TA, (ii) AG/GA/CT/TC, (iii) AC/CA/TG/GT, and (iv) GC/CG.

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

The trinucleotide repeats are grouped into 10 unique classes

AAG/TTC contains AAG/AGA/GAA/CTT/TTC/TCT SSRs

(Jurka and Pethiygoda 1995; Katti et al. 2001)

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009

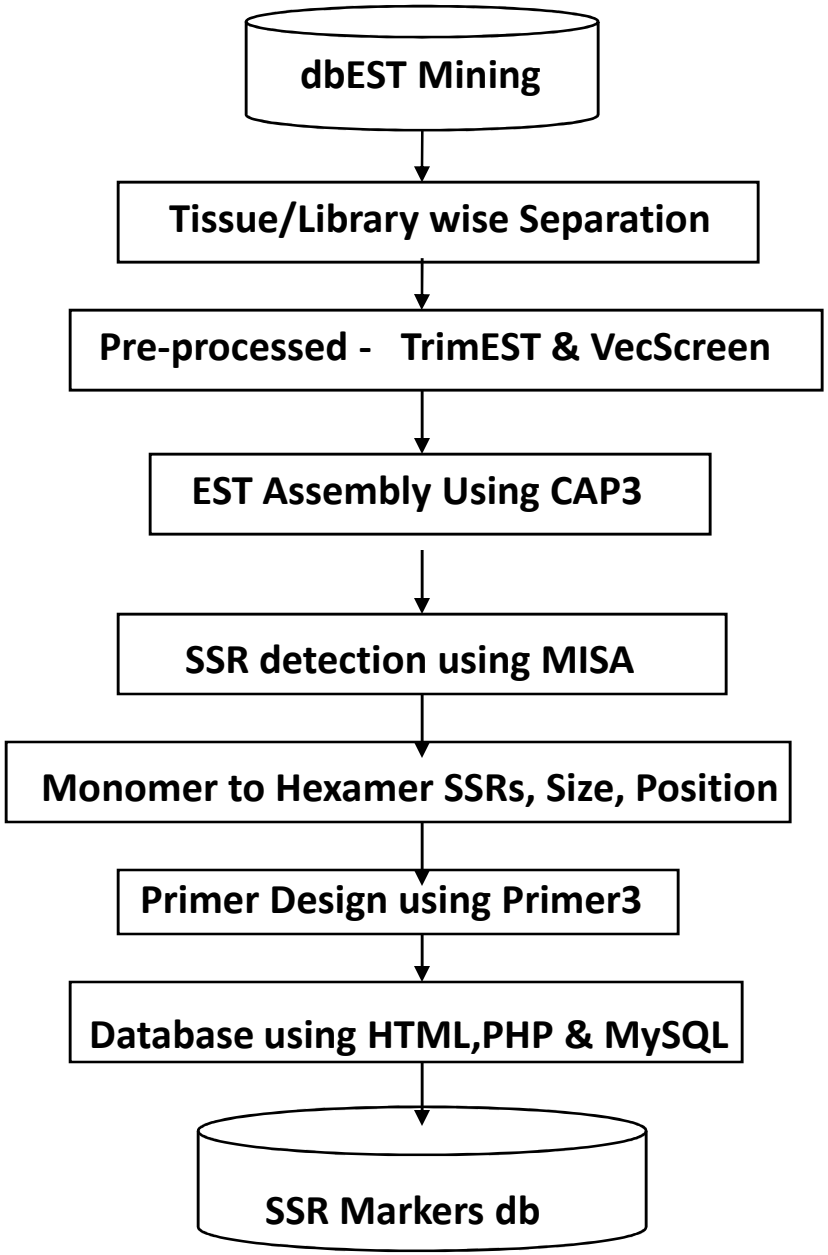


Fig. 1. Flowchart of various steps involved in *in silico* SSR discovery and database development

Nature Precedings : doi:10.1038/npre.2009.3908.1 : Posted 28 Oct 2009