HUPATHOMICROSATDB:

A COMPREHENSIVE DATABASE OF SIMPLE, COMPOUND AND CLUSTURE OF MICROSATELLITE REPEATS OF HUMAN PATHOGEN

Dissertation

Submitted to the Orissa University of Agriculture & Technology, Bhubaneswar in partial fulfillment of the requirement for the award of degree of

MASTER OF SCIENCE IN BIOINFORMATICS

BY

JASHWANT KUMAR SARDAR
ADM NO: 05BI/08

DEPARTMENT OF BIOINFORMATICS
CENTRE FOR POST GRADUATE STUDIES
ORISSA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY
BHUBANESWAR-751003

2010

Name of the Advisor **Mr. Sukanta Kumar Pradhan**

DEDICATED TO MY

BELOVED PARENTS

&

My Brother and Sisters

'To raise new questions, new possibilities, to regard old problems from a new angle, require creative imagination and marks real advance in science'.

.....*Albert Einstein*

Orissa University of Agriculture & Technology
Department Of Bioinformatics
Centre for Post Graduate Studies
Bhubaneswar

Mr. Sukanta Kumar Pradhan
Head

CERTIFICATE –I

This is to certify that thesis entitled **"HuPathoMicroSatDb:** ***A Comprehensive Database of Simple, Compound, and Clusture of Microsatellite repeats of Human Pathogen"*** submitted for award for the degree of **Master of Science** in the subject of **Bioinformatics** embodies a faithful bonafied research work carried out by **Jashwant Kumar Sardar (Adm. No. 05BI/08)** under my guidance & supervision. No part of this thesis has been submitted by him for any other degree or diploma.

I further certify that any help or information received during the course of investigation have been duly acknowledged by him.

Place- Bhubaneswar

Date 21·8·10

SKPradhan
21·8·10
Mr. Sukanta Kumar Pradhan

HOD
Chairman
Advisory committee

CERTIFICATE –II

This is to certify that the dissertation entitled **"HuPathoMicroSatDb:** ***A Comprehensive Database of Simple, Compound, and Clusture of Microsatellite repeats of Human Pathogen"*** submitted by **Jashwant Kumar Sardar (Adm. No. 05BI/08)** to the Orissa University of Agriculture & Technology, Bhubaneswar in the partial fulfillment of the requirements for the award of the degree of **Master of Science** in **Bioinformatics** has been approved by the students advisory committee after an oral examination of the same in collaboration with external examiner.

ADVISORY COMMITTEE

1.	Mr. Sukanta Kumar Pradhan Head Department of Bioinformatics	Chairman	.....
2.	Mr. S. N Rath Asst. Professor Department of Bioinformatics	Member	.....
3.	Ms. S. Balabantray Asst. Professor Department of Bioinformatics	Member	.....
4.	Mr. A. Dash Head Department of CSA	Member	.....

External Examiner .....

ACKNOWLEDGEMENT

I do feel great pleasure in expressing my deep sense of gratitude, sincerest respect and cordial thanks to ***Mr.Sukanta Kumar Pradhan(HOD)****, Dept of Bioinformatics, OUAT, Bhubaneswar, Orissa for his Support, Advise and Co-operation.*

I am thankful to the member of the advisory committee ***Mr.S.N Rath , Miss S.Balabantray*** *Asst. Prof. dept Of Bioinformatics ,* ***Mr. A. Dash HOD*** *Dept. Of CSA for their support and encouragement to carry out this work successfully.*

I am grateful to ***Dr. Pradeep Das, Director, RMIRIMS*** *and other authorities of RMRIMS for providing the Institutional facilities.*

I also express my profound gratitude to my guide, ***Mr. Manas Ranjan Dikhit, Scientist B*** *and co-guide,* ***Dr. Ganesh Chandra Sahoo, Senior Scientist (HOD),*** *BioMedical Informatics Centre, RMRIMS for providing me an opportunity to work under their guidance. I am grateful to them for all the encouragement, unrelenting support and Brotherly concern during the period of my project*

I would like to owe my sincere thanks to the Respected ***Dean Dr. H.K. Senapati*** *for the encouragement.*

I feel great pleasure in expressing deepest regards and cordial thanks to ***Miss. S.R.Martha*** *mam for Aware me about career Using her Experience which help me in Great way to focus on my career. Also I owe sincere thanks to* ***Miss. Sudipta Mohanty*** *for her Moral support.*

I would like to owe my sincere thanks ***Mr. Lelin Patel, Miss. Mukta Rani & Miss.Chanda Jha*** *of RMRIMS, Patna for their help and motivation.*

I would like to express my heartiest and cordial regards to my beloved ***Mom & Dad*** *and My Loving* ***Brother & Sisters*** *whose unbound emotional support and attachment always inspire me to face all challenges.*

I also thanks to my Great supporting Friends ***Miss.Sonali Pradhan, Miss. Nisha Singh, Mr.Suraj Harichandan, Mr Sadasib Suranjan Mohanto , Mr. Sudhanshu Shekhar Parida,*** *My Loving Juniors* ***Mr.Smurti Ranjan Sahoo , Mr.Jitendra Maharana*** *and Lastly My senior* ***Mr. Bishwaranjan Jena*** *for their Co-operation and Help to Fulfill This Project.*

Lastly I express my gratitude to my Mentor for database design ***Mr.Vikram Kumar Parida.***

Jashwant Sardar

Mr Jashwant Kumar Sardar

Name of the Student : **Jashwant Kumar Sardar**

Admission No : **05BI/08**

Title Of Thesis : **HuPathoMicroSatDB:** ***A Comprehensive Database of Simple, Compound, and Clusture Of Microsatellite Repeats of Human Pathogen***

Degree for which thesis submitted : **Master of Science in Bioinformatics**

Name of the Dept, College, & University : **Department of Bioinformatics, Cetre for Post Graduate Studies, Orissa University Of Agriculture &Technology, Bhubaneswar, Orissa, 751003**

Year of submission : **2010**

Name of the advisor : **Mr. Sukanta Kumar Pradhan**

ABSTRACT

Microsatellites, also called as simple sequence repeats (SSRs) or simple tandem repeats (STRs) are ubiquitous component of genomes. A microsatellite consists of a specific sequence of DNA which contains 1–6 bp long (mono- to hexa- nucleotide) tandem repeats viz. $(A)_{16}$, $(GA)_{20}$, $(GATA)_{30}$. Microsatellites serve as excellent molecular markers for genotyping, strain differentiation, epidemiological analysis and genome analysis. With the whole genome sequencing initiatives of various organisms, large amount of genomic sequence data has accumulated over the last few years. These sequence resources available in the public domain have also served as an attractive source of *in silico* mining of microsatellite sequences. *In silico* mining of these sequences offers advantage in terms of time, labour and cost over conventional isolation from genomic libraries. However, finding potentially useful microsatellites occupying specific genomic regions still remains a challenge for the molecular biologists. Availability of this information can facilitate molecular mapping of desired traits and preparation of linkage maps saturated with evenly distributed SSR markers.

HuPathoMicroSatDb is database of microsatellite repeats of *Human Pathogen* for which whole genome sequencing has been completed and sequence data is available in public domain. These whole genome sequences are assembled as Species, Sub-Species, Strain and Chromosome. *HuPathoMicroSatDb* contains di to hexa nucleotide repeats of 233 species of *Human Pathogen* which include 929 Strain or 1070 Sample. These Pathogens are Bacteria, Viruses, Parasites and Fungi. Precise need based microsatellites data retrieval is possible using different input parameters like microsatellite type (simple perfect), repeat unit length (mono- to hexa-nucleotide), repeat number, microsatellite length and chromosomal location in the genome. Furthermore, information about clustering of different microsatellites in the genome can also be retrieved. Finally, to facilitate primer designing for PCR amplification of any desired microsatellite locus, 200 bp upstream and downstream sequences are provided.

Key Words: Microsatellites, Whole Genome, Genomic libraries, Repeats.

SK Pradhan 21.8.10

Mr. Sukanta Kumar Pradhan

ADVISOR

Jashwant Sardar 21.08.10

Jashwant Kumar Sardar

AUTHOR

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBRIVIATIONS

SSR	Simple Sequences Repeats
STR	Short Tandem Repeats
BMC	BioMed Central
CDFD	Centre For DNA Fingerprinting and Diagnostics
PGRS	Polymorphic GC Repitetive Sequences
UTR	Untranslated Region
BAC	Bacterial Artificial Chromosome
YAC	Yeast Artificial Chromosome
EST	Expressed Sequence Tag
GSS	Genome Survey Sequence
VNTR	Variable Number Tandem Repeats
SINEs	Short Interspersed Tandem Repeats
LINEs	Long Interspersed Tandem Repeats
RAPD	Random Amplified Polymorphic DNA
RFLP	Replication Fragment Length Polymorphism
XAMPP	Window Xp, Apache, MySql, PHP, PERL
PHP	Hypertext PreProcessor
PERL	Practical Extraction Report Language
HTTP	Hyper Text Transfer Protocol
URL	Universal Resource Locator
GUI	Graphical User Interface
SQL	Structured Query Language
MISA	Microsatellite Identification Tool
HTML	Hyper Text Markup Language
CSS	Cascading Style Sheets

Chapter - I
Introduction

CHAPTER I

1. INTRODUCTION

Genomes are scattered with simple repeats called microsatellites. Microsatellites, also called as simple sequence repeats (SSRs) or simple tandem repeats (STRs) are ubiquitous component of eukaryotic genomes. A microsatellite consists of a specific sequence of DNA which contains 1–6 bp long (mono- to hexa- nucleotide) tandem repeats. Over the years, molecular biologists have increasingly exploited these sequences for diverse applications (Forensics, diagnosis and Identification of Human Diseases, genetic structure analysis of subpopulations and populations, phylogenetic study, genome mapping etc.).

With the whole genome sequencing initiatives of various eukaryotic organisms, large amount of genomic sequence data has accumulated over the last few years. Availability of complete and annotated genome sequences of a number of organisms has provided an excellent opportunity to analyze microsatellites in a very great detail for their genomic locations, distributions and frequencies. Results from such analysis provide a useful basis for carrying out further investigations into the structural and functional characteristics of microsatellites. In silico mining of these sequences offers advantage in terms of time, labour and cost over conventional isolation from genomic libraries. Similarly, Chromosome has also been screened for the presence of microsatellites.

However, finding potentially useful microsatellites occupying specific genomic regions still remains a challenge for the molecular biologists. Availability of this information can facilitate molecular mapping of desired traits and preparation of linkage maps saturated with evenly distributed SSR markers. Popularity of in silico mining methods has led to the construction of various microsatellite databases in recent years, each with a different emphasis. We have developed a database of microsatellites extracted from genomic sequences of all the Human Pathogen which have been sequenced completely such Human Pathogens are *Bacteria, Viruses, Parasites and Fungi.* The feature of this database is that it also stores data on different status of various

microsatellite loci. *HuPathMicrsatDB* is a web-based relational database providing centralized access to publicly available *Human Pathogen* microsatellites.

There are so many SSR databases that stores information on microsatellite repeats size, type (perfect and compound) and location (intron, exon, upstream or transposons) of microsatellites. Some other databases, although published earlier are currently inaccessible. In conclusion, existing microsatellite databases either are very specific in their content and application or have limited utility to a wider audience. Thus, a collection of whole *Human Pathogen* genome microsatellite data at a single platform is still not available. Recognizing this gap, we have developed a comprehensive database for easy retrieval of information on microsatellites distributed in the sequenced *Human Pathogen* genomes. The database named as *HuPathoMicroSatDB* (*Human Pathogen* Micro-Satellite database) presents a web-based user friendly interface for the extraction of simple microsatellites from *233 Species* genomes including their Sub-Species, Strain, and Chromosome collectively 1070 sample of Fasta file being extracted for microsatellite repeats assembled in the database.

HuPathoMicroSatDB is database of microsatellite repeats in the genome of those *Pathogen(Bacteria, Viruses, Parasites and Fungi)* for which whole genome sequencing has been completed and sequences data is available in public domain. These whole genome sequences are assembled as Strain and Chromosome. *HuPathoMicroSatDB* contains di to hexa nucleotide repeats of 233 species of *Human Pathogen.* Precise need based microsatellites data retrieval is possible using different input parameters like microsatellite type (simple perfect), repeat unit length (mono- to hexa-nucleotide), repeat number, microsatellite length and chromosomal location in the genome. Furthermore, information about clustering of different microsatellites in the genome can also be retrieved, to facilitate primer designing for PCR amplification of any desired microsatellite locus, 200 bp upstream and downstream sequences to the desired microsatellite repeat are provided.

OBJECTIVES

- To find out different types of microsatellite repeats present in the genomes of any human pathogen by using microsatellite repeat finding algorithm.
- To develop a database (HuPathoMicroSatDB) of microsatellite repeats found in genomes of different Human Pathogen.
- To develop an algorithm to find out microsatellite clusters in the genomes of any Human Pathogen.
- To develop an algorithm to find out upstream and downstream flanking sequences to the respective microsatellite repeats.

Chapter – II
Review of literature

CHAPTER II

2. REVIEW OF LITERATURE

2.1 *Human Pathogen*

A Pathogen is a biological agent that causes disease to its host. There are several substrates and pathways whereby pathogens can invade a host. Body contains many natural orders of defense against some of the common pathogens in the form of the human immune system and by some helpful agent present in the human body. We come in contact with pathogens everyday but due to our immunity we get protected up to the normal. Most of the time our body's immune system destroys them before they can cause harm. We are considered exposed when we have been in contact with a pathogen but when a pathogen has entered the body and resulted in disease we are infected. Pathogens have been categorized in four categories:-

- Bacteria
- Viruses
- Parasites
- Fungi

These four Pathogens are the reason of infectious disease to human which results in major disorders or death. A very early prediction of correct pathogen causing disease is the first step of prescription.

2.1.1 *Bacteria*

Pathogenic bacteria are the only bacteria that cause infectious diseases. The vast majority of bacteria are harmless or beneficial but quite few are pathogenic still they are in small amount but very harmful. One of the bacterial diseases with highest disease burden is tuberculosis, caused by the bacterium Mycobacterium tuberculosis, which kills about 2 million people a year, India contribute one third patient of tuberculosis to the world. Pathogenic bacteria contribute to other diseases such as pneumonia which can be caused by bacteria such as Streptococcus and Pseudomonas. Pathogenic bacteria also

cause infections such as tetanus, typhoid fever, diphtheria, syphilis and leprosy. Not all Bacteria are pathogenic in nature or cause disease to human. Some bacteria which are found in milk are Streptococcus lactis or the Lactobacillus species. These organisms do not produce disease in man but ferment the carbohydrate lactose. Pathogenic bacteria is pathogenic in certain condition, such as a wound that allows for entry into the blood, or a decrease in immune function. Staphylococcus or Streptococcus are usually exist on the skin or in the nose without causing disease but can potentially cause skin infection, pneumonia, meningitis and even overwhelming sepsis, a systemic inflammatory response producing shock, massive vasodilation and death. Some species of bacteria, such as Pseudomonas aeruginosa, Burkholderia cenocepacia, and Mycobacterium avium, are opportunistic pathogens and cause disease mainly in people suffering from immunosuppression or cystic fibrosis.

Microsatellites are generally absent from bacterial genomes except in locations where they provide adaptive functional variability, and these appear to have evolved under selection. These repeats show length polymorphism characterized by either insertion or deletion (indels) of the repeat units, which in and around the coding regions affect transcription and translation of genes. After studying article of BMC Genomics by CDFD on Mycobacterium tuberculosis H37Rv, Mycobacterium tuberculosis CDC1551 and Mycobacterium bovis, revealed for the first time the presence of several polymorphic microsatellites. The coding regions affected by frame-shifts owing to microsatellite indels have undergone changes indicative of gene fission/fusion, premature termination and length variation. Interestingly, the genes affected by frame-shift mutations code for membrane proteins, transporters, PPE, PE_PGRS, cell-wall synthesis proteins and hypothetical proteins. Microsatellite indel mutations have novel functions and a certain degree of plasticity to the mycobacterial genomes. There is some correlation between microsatellite polymorphism and the variations in virulence, host-pathogen interactions mediated by surface antigen variation and adaptation of the pathogens.

2.1.2 *Viruses*

A virus is a small infectious agent than can replicate only inside the living cells of organisms. Most viruses are too small to be seen directly with a light microscope. Viruses infect all types of organisms, from animals and plants to bacteria and archaea. Virus particles (known as *virions*) consist of two or three parts: the genetic material made from either DNA or RNA, long molecules that carry genetic information; a protein coat that protects these genes; and in some cases an envelope of lipids that surrounds the protein coat when they are outside a cell. The shapes of viruses range from simple helical and icosahedral forms to more complex structures. The average virus is about one one-hundredth the size of the average bacterium. The origins of viruses in the evolutionary history of life are unclear: some may have evolved from plasmids pieces of DNA that can move between cells while others may have evolved from bacteria. In evolution, viruses are an important means of horizontal gene transfer, which increases genetic diversity. Viruses spread in many ways; plant viruses are often transmitted from plant to plant by insects that feed on sap, such as aphids, while animal viruses can be carried by blood-sucking insects. These disease-bearing organisms are known as vectors. Influenza viruses are spread by coughing and sneezing. The norovirus and rotavirus, common causes of viral gastroenteritis, are transmitted by the faecal-oral route and are passed from person to person by contact, entering the body in food or water. HIV is one of several viruses transmitted through sexual contact and by exposure to infected blood. Viruses can infect only a limited range of host cells called the "host range". This can be narrow or, as when a virus is capable of infecting many species, broad.

2.1.3 *Parasites*

Human parasites are organisms that live inside us so that we become their hosts. Since these parasites are unable to produce food for themselves, they depend on us for their survival. Unfortunately, parasites harm human beings because they consume our food and nutrients, they can destroy our tissues and cells, and they produce toxic waste products that can make people very ill. In some underdeveloped countries, human parasite infections are epidemic, sickening and killing thousands upon thousands of

people each year. In India, parasitic infections are basically found in Punjab, Bihar, and Uttar Pradesh etc. Cryptosporidium parvum, Plasmodium Falciparum and Trypanosoma brucei are found in Bihar, Uttar Pradesh and Orissa. Cryptosporidiosis is mostly a water and food-borne infection. The infective stage is the fully formed oocysts of the parasite, which are passed in the faeces and transmitted to a second individual via the faecal–oral route. The disease is seen in young calves. Clinically, the older calves are asymptomatic, but continuously shed oocysts in faeces contaminate the environment and become the source of infection to man. Our way of life can also contribute to the spread of parasites. A large percentage of children contract parasites from their day care centers. Children and adults with dogs and cats at home are at risk for getting parasites. Also, those people that eat at restaurants are at a higher risk because food handlers have been known to spread parasites.

2.1.4 *Fungi*

All fungi are eukaryotic and they are chemoheterotrophs. The study of microsatellite of fungi which are disease causing or Pathogenic in nature to human is very crucial for understanding fungal diseases. Fungal infection is of many types depending on the mode of entry into the host such as superficial, subcutaneous, systemic and opportunistic. In Superficial, infection is localized to skin, hair and nail while in case of subcutaneous it is dermis and subcutaneous tissue. There are some internal deep infection occurs in organs due to systemic i.e These are invasive infections of the internal organs with the organism gaining entry by the lungs, gastrointestinal tract or through intravenous lines. Opportunistic infection occurs in usually to the person having immune an metabolic defect or have undergone surgery. Our motto is to study these pathogenic fungi and their effect on human with the help of the study of their microsatellite repeats.

2.2 *Human Pathogen Genome*

2.2.1 *Bacterial Genome*

The first bacterial genome to be completed was that of Haemophilus influenzae, completed by a team at The Institute for Genomic Research in 1995. It has been more than 10 years since the first bacterial genome sequence was published. Hundreds of bacterial genome sequences are now available for comparative genomics, and searching a given protein against more than a thousand genomes will soon be possible. The subject of this review will address a relatively straightforward question: "What have we learned from this vast amount of new genomic data?" Perhaps one of the most important lessons has been that genetic diversity, at the level of large-scale variation amongst even genomes of the same species, is far greater than was thought. The classical textbook view of evolution relying on the relatively slow accumulation of mutational events at the level of individual bases scattered throughout the genome has changed. One of the most obvious conclusions from examining the sequences from several hundred bacterial genomes is the enormous amount of diversity even in different genomes from the same bacterial species. This diversity is generated by a variety of mechanisms, including mobile genetic elements and bacteriophages. An examination of the 20 Escherichia coli genomes sequenced so far dramatically illustrates this, with the genome size ranging from 4.6 to 5.5 Mbp; much of the variation appears to be of phage origin. This review also addresses mobile genetic elements, including pathogenicity and the structure of transposable elements. There are at least 20 different methods available to compare bacterial genomes. Study of microsatellite offers the chance to study genomic sequences found in ecosystems, including genomes of species that are difficult to culture. It has become clear that a genome sequence represents more than just a collection of gene sequences for an organism and that information concerning the environment and growth conditions for the organism are important for interpretation of the genomic data.

2.2.2 *Viral Genome*

The complete DNA sequence of a viral genome was reported by Sanger in 1977. However, Sanger's technique of DNA sequencing was still very slow. The nucleic acid

comprising the genome may be single-stranded or double-stranded, & in a linear, circular or segmented configuration. But Single-Stranded virus genomes may be

- Positive(+) Sense
- Negative(-) Sense
- Ambisense

Virus genomes range in size from approximately 3,200 nucleotides to approximately 1.2 million base pairs. Unlike the genomes of all cells, which are composed of DNA, viral genomes may contain their genetic information encoded in either DNA or RNA.
Purified (+) sense vRNA is directly infectious when applied to susceptible host cells in the absence of any virus proteins. There is an untranslated region(UTR) at the 5' end of the genome which does not encode any proteins & a shorter UTR at the 3' end. These regions are functionally important in virus replication & are thus conserved in spite of the pressure to reduce genome size. Both ends of (+) stranded eukaryotic virus genomes are often modified, the 5'end by a small, covalently attached protein or a methylated nucleotide 'cap' structure & the 3' end by polyadenylation. These signals allow vRNA to be recognized by host cells & to function as mRNA. Negative-sense RNA genomes are not infectious as purified RNA. This is because such virus particles contain a virus-specific-polymerase. The Herpesviruses are a large family containing more than 100 different members, at least one for most animal species which have been examined to date, including seven human herpesviruses. Herpesviruses have very large genomes composed of up to 230kbp.
The genomes of adenoviruses consist of linear, double-stranded DNA of 30-38kbp.

2.2.3 *Parasite Genome*

During 1993-94, scientists from developing and developed countries planned and initiated a number of parasite genome projects and several consortiums for the mapping and sequencing of these medium-sized genomes were established, often based on already ongoing scientific collaborations. Financial and other support came from WHO/TDR, welcome Trust and other funding agencies. Thus, the genomes of Plasmodium falciparum, Schistosoma mansoni, Trypanosoma cruzi, Leishmania major, Trypanosoma

brucei, Brugia malayi and other pathogenic nematodes are now under study. From an initial phase of network formation, mapping efforts and resource building(EST, GSS, phage, cosmid, BAC and YAC library constructions), sequencing was initiated in gene discovery projects but soon on a small chromosomes, and now on a fully fledged genome scale. Proteomics, functional analysis, genetic manipulation and microarray are ongoing to different degrees in the respective genome initiatives, and as the funding for the whole genome sequencing becomes secured, most of the participating laboratories, apart from larger sequencing centres, become oriented to post-genomics. Bioinformatics networks are being expanded, including in developing countries, for data mining, annotation and in-depth analysis. The genomes of the parasite are sequenced with the aim of learning more about how the parasite works, and with the hope that this would reveal potential drug targets.

2.2.4 *Fungi Genome*

The Fungi present serious threats to human health, serve as important models for biomedical research, and provide a wide range of evolutionary comparisons at key branch points in the 1 billion years spanned by the fungal evolutionary tree. The Fungal Genome Initiative has provide the sequence of key organisms across the fungal kingdom and thereby lay the foundation for work in medicine, agriculture, and industry. The fungal and genomics communities have worked together for over 2 years to choose the most informative organisms to sequence from the more than 1.5 million species that comprise this kingdom. Fungal infections have lethal consequences for the growing population of patients immunocompromised with AIDS or the therapeutically immunosuppressed after cancer chemotherapy or transplantation surgery. Fungal disease now represents as much as 15% of all hospital-acquired infections. One of the greatest needs clinically is the availability of diagnostics that can provide facile and accurate identification of particular fungal species. Genome sequences provide the opportunity for unique DNA probes that could be used for identification.

2.3 MICROSATELLITE

A microsatellite consists of a specific sequence of DNA bases or nucleotides which contains mono, di, tri, tetra, penta or hexa tandem repeats. For example,

AAAAAAAAAAA would be referred to as $(A)_{11}$

GTGTGTGTGTGT would be referred to as $(GT)_6$

CTGCTGCTGCTG would be referred to as $(CTG)_4$

ACTCACTCACTCACTC would be referred to as $(ACTC)_4$

or variable number tandem repeats (VNTR). The usable repeat lengths are from 8-40 copies--for example AC_{22} is a 'necklace' of 22 repeat units of the dinucleotide repeat AC. Microsatellites occur in many places (*loci*) throughout the genome, but in almost all cases they are in non-coding regions of the DNA.

Repeats of longer units form minisatellites or, in the extreme case satellite DNA. The term satellite DNA originates from the observation in the 1960s of a fraction of sheared DNA that showed a distinct buoyant density, detectable as a 'satellite peak' in density gradient centrifugation, and that was subsequently identified as large centromeric tandem repeats. When shorter (10–30-bp) tandem repeats were later identified, they came to be known as minisatellites. Finally, with the discovery of tandem iterations of simple sequence motifs, the term microsatellites were coined.

It is appropriate to study association of microsatellites with coding sequence as this is related to the mutational and selective forces that operate on different types of repeat. The bulk of simple repeats are embedded in non-coding DNA, either in the intergenic sequence or in the introns. Microsatellites that are used as genetic markers are usually of this type and are generally assumed to evolve neutrally. Their frequency and distribution reflects the underlying mutation process. In coding DNA, selection against frameshift mutations effectively hinders the expansion of everything other than trinucleotide repeats for which there might be further length constraints related to protein function.

Trinucleotide repeats associated with human disease comprise a special class of microsatellites in coding DNA. These loci undergo extensive repeat expansions, the mutational mechanism of which is thought to differ from that of most microsatellites in the genome. For instance, the establishment of hairpin structures with a relatively high amount of base-pair complementarities might stabilize loops that are generated during replication slippage. Microsatellites are also frequently found in the proximity of interspersed repetitive elements such as short interspersed repeats (SINEs) and long interspersed elements (LINEs).

Microsatellites owe their variability to an increased rate of mutation compared to other neutral regions of DNA. These high rates of mutation can be explained most frequently by slipped strand mispairing (slippage) during DNA replication on a single DNA double helix. Mutation may also occur during recombination during meiosis . Some errors in slippage are rectified by proofreading mechanisms within the nucleus, but some mutations can escape repair(The size of the repeat unit, the number of repeats and the presence of variant repeats are all factors, as well as the frequency of transcription in the area of the DNA repeat). Interruption of microsatellites, perhaps due to mutation, can result in reduced polymorphism. However, this same mechanism can occasionally lead to incorrect amplification of microsatellites; if slippage occurs early on during PCR, microsatellites of incorrect lengths can be amplified.

2.3.1 Development of microsatellite primers

If searching for microsatellite markers in specific regions of a genome; for example within a particular exon of a gene, primers can be designed manually. This involves searching the genomic DNA sequence for microsatellite repeats, which can be done by eye or by using various automated tools which are present in public domain. Once the potentially useful microsatellites are determined (removing non-useful ones such as those with random inserts within the repeat region), the flanking sequences can be used to design oligonucleotide primers which will amplify the specific microsatellite repeat in a PCR reaction. Random microsatellite primers can be developed by cloning random segments of DNA from the focal species. These are inserted into a plasmid or

phage vector, which is in turn, implanted into *Escherichia coli* bacteria. Colonies are then developed, and screened with fluorescently labelled oligonucleotide sequences that will hybridize to a microsatellite repeat, if present on the DNA segment. If positive clones can be obtained from this procedure, the DNA is sequenced and PCR primers are chosen from sequences flanking such regions to determine a specific locus. This process involves significant trial and error on the part of researchers, as microsatellite repeat sequences must be predicted and primers that are randomly isolated may not display significant polymorphism. Microsatellite loci are widely distributed throughout the genome and can be isolated from semi-degraded DNA of older specimens, as all that is needed is a suitable substrate for amplification through PCR.

2.3.2 Microsatellite detection

The most common way to detect microsatellites is to design PCR primers that are unique to one locus in the genome and that base pair on either side of the repeated portion (Figure1). Therefore, a single pair of PCR primers will work for every individual in the species and produce different sized products for each of the different length microsatellites.

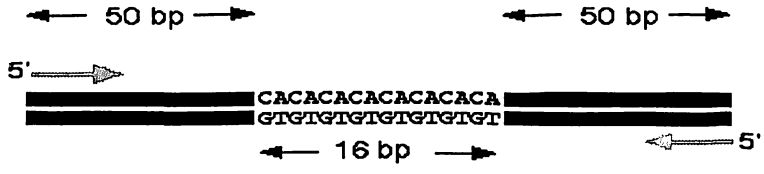


Figure 1. Detecting microsatellites from genomic DNA. Two PCR primers (forward and reverse gray arrows) are designed to flank the microsatellite region.

The PCR products are then separated by either gel electrophoresis or capillary electrophoresis. Either way, the investigator can determine the size of the PCR product and thus how many times the dinucleotide "CA" was repeated for each allele (Figure2). It would be nice if microsatellite data produced only two bands but often there are minor bands in addition to the major bands; they are called stutter bands and they usually differ from the major bands by two nucleotides.

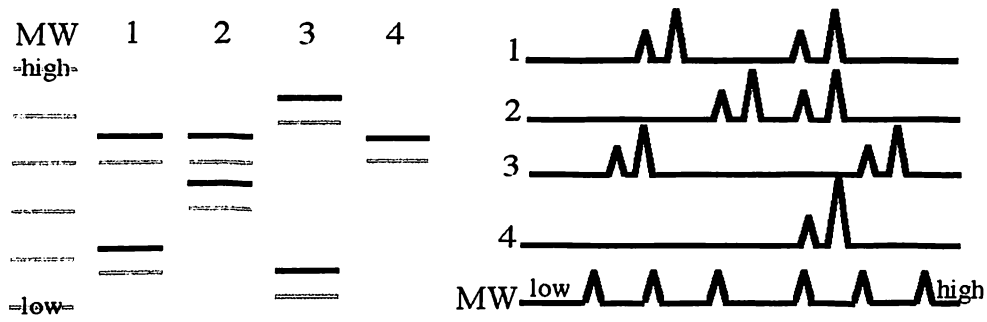


Figure 2. Stylized examples of microsatellite data. Left half: four sets of data were produced by gel electrophoresis, major (black) and stutter (gray) bands. MW; molecular weight standards. Right half: These data were produced by analysis on an automated capillary electrophoresis-based DNA sequencer. The data are line graphs with the location of each peak on the X-axis representing a different sized PCR product and the height of each peak indicates the amount of PCR product. The major bands produce higher peaks than the stutter peaks.

2.4 MICROSATELLITE IMPORTANCE IN HUMANPATHOGEN

The current rapid spread of Diseases caused by Human Pathogen including Bacteria, Viruses, Parasites and Fungi has brought the complexity of its clinical spectrum call for epidemiological and evolutionary investigation. The universally accepted standard procedure for characterizing and indentifying strains of different species under Bacteria, viruses, parasites and fungi isoenzyme analysis. However, this is performed only in a few laboratories and is done for specific species depending on the number of enzymes examined, very labour intensive and time consuming, thus, insufficient discriminative power of isoenzyme typing methods prevents researchers from establishing correlations between clinical feature, preferential host and particular group of strains. In the same way, the lack of discrimination between strains also prevents genetic studies on Pathogen populations. Several molecular biological typing methods has been developed to improve the discriminative powers of typing methods for the Pathogen. These include amplification of a pathogen DNA sequence by either a specific PCR or a random amplified polymorphic DNA (RAPD) PCR or detection of restriction fragment length polymorphisms (RFLPs) by Southern hybridization with DNA-specific

probes. The last two methods had drawbacks. RFLP analysis was a time-consuming technique and large amounts of purified DNA were needed, whereas RAPD analysis required strict conditions to obtain reproducibility between different laboratories and to generate complex patterns. In contrast, specific PCR-based methods were attractive because of their rapidity and because culturing pathogen could be avoided. However, in most cases, the level of polymorphism found with coding or repeated noncoding PCR-amplified sequences was not refined enough to distinguish between closely related strains. Microsatellite DNA sequences, tandem repeats of a simple nucleotide motif, are distributed abundantly in the genomes and may reveal important strain polymorphisms.

2.5 MICROSATELLITE DATABASES

2.5.1 MICDB

The MICdb (Microsatellites Database) is a comprehensive relational database of non-redundant microsatellites extracted from fully sequenced prokaryotic genomes. The current version of the database has been compiled from 83 genomes belonging to different phylogenetic groups. This database has been linked to MICAS, the web-based Microstatellite Analysis Server. MICAS provides a user-friendly front-end to systematically extract data on microsatellite tracts from genomes. The database contains the following information pertaining to the microsatellites: the regions (coding/non-coding, if coding, their GenBank annotations) containing microsatellite tracts; the frequencies of their occurrences, the size and the number of repeating motifs; and the sequences of the tracts. MICAS also provides an interface to Autoprimer, a primer design program to automatically design primers for selected microsatellite loci.

2.5.2 SILKSATDB

The SilkSatDb (silkmoth microsatellite database) is a relational database of microsatellites extracted from the available expressed sequence tags and wholegenome shotgun sequences of the silkmoth, *Bombyx mori*. The SilkSatDb also stored information on primers developed and validated in the laboratory. Users can retrieve information on the microsatellite and the protocols used, along with informative figures and

polymorphism status of those microsatellites. In addition, the interface is coupled with Autoprimer, a primer designing program, using which users can design primers for the loci of interest.

2.5.3 MMDBJ

Mouse Microsatellite Data Base of Japan (MMDBJ) provides information about simple sequence length polymorphisms (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 *molossinus*-derived 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.

2.5.4 CMD

The Cotton Microsatellite Database (CMD) is a curated and integrated web-based relational database providing centralized access to publicly available cotton microsatellites, an invaluable resource for basic and applied research in cotton breeding. At present CMD contains publication, sequence, primer, mapping and homology data for nine major cotton microsatellite projects, collectively representing 5,484 microsatellites. In addition, CMD displays data for three of the microsatellite projects that have been screened against a panel of core germplasm. The standardized panel consists of 12 diverse genotypes including genetic standards, mapping parents, BAC donors, subgenome representatives, unique breeding lines, exotic introgression sources, and contemporary Upland cottons with significant acreage. A suite of online microsatellite data mining tools are accessible at CMD. These include an SSR server which identifies microsatellites, primers, open reading frames, and GC-content of uploaded sequences; BLAST and FASTA servers providing sequence similarity searches against the existing cotton SSR sequences and primers, a CAP3 server to assemble EST sequences into longer transcripts prior to mining for SSRs, and CMap, a viewer for comparing cotton

SSR maps. The collection of publicly available cotton SSR markers in a centralized, readily accessible and curated web-enabled database provides a more efficient utilization of microsatellite resources and will help accelerate basic and applied research in molecular breeding and genetic mapping in *Gossypium* spp.

2.5.5 SATELLOG

Satellog, a database that catalogs all pure 1–16 repeat unit satellite repeats in the human genome along with supplementary data. Satellog analyzes each pure repeat in UniGene clusters for evidence of repeat polymorphism. A total of 5,546 such repeats were identified, providing the first indication of many novel polymorphic sites in the genome. Overall, polymorphic repeats were over-represented within 3'- UTR sequence relative to 5'-UTR and coding sequence. Interestingly, they observed that repeat polymorphism within coding sequence is restricted to trinucleotide repeats whereas UTR sequence tolerated a wider range of repeat period polymorphisms. For each pure repeat they also calculate its repeat length percentile rank, its location either within or adjacent to EnsEMBL genes, and its expression profile in normal tissues according to the GeneNote database. Satellog provides the ability to dynamically prioritize repeats based on any of their characteristics (i.e. repeat unit, class, period, length, repeat length percentile rank, genomic coordinates), polymorphism profile within UniGene, proximity to or presence within gene regions (i.e. cds, UTR, 15 kb upstream etc.), metadata of the genes they are detected within and gene expression profiles within normal human tissues. Unstable repeats associated with 31 diseases were analyzed in Satellog to evaluate their common repeat properties. The utility of Satellog was highlighted by prioritizing repeats for Huntington's disease and schizophrenia.

2.5.6 MRD

MRD is a database system to access the microsatellite repeats information of genomes such as archea, eubacteria, and other eukaryotic genomes whose sequence information is available in public domains. MRD stores information about simple tandemly repeated k-mer sequences where k= 1 to 6, i.e. monomer to hexamer. The web

interface allows the users to search for the repeat of their interest and to know about the association of the repeat with genes and genomic regions in the specific organism. The data contains the abundance and distribution of microsatellites in the coding and non-coding regions of the genome. The exact location of repeats with respect to genomic regions of interest (such as UTR, exon, intron or intergenic regions) whichever is applicable to organism is highlighted.

2.5.7 UgMicroSatdb

UgMicroSatdb (Unigene MicroSatellite database), a web-based relational database of microsatellites present in unigene sequences covering 80 genomes. UgMicroSatdb allows microsatellite search using multiple parameters like microsatellite type (simple perfect, 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 the desired microsatellite .

2.5.8 EuMicroSatdb

EuMicroSat*db* (**Euk**aryotic **MicroSat**ellite database) a web based relational database for easy and efficient positional mining of microsatellites from sequenced eukaryotic genomes. A user friendly web interface has developed for microsatellite data retrieval using Active Server Pages (ASP). The backend database codes for data extraction and assembly has written using Perl based scripts and C++. Precise need based microsatellites data retrieval is possible using different input parameters like microsatellite type (simple perfect or compound perfect), repeat unit length (mono- to hexa-nucleotide), repeat number, microsatellite length and chromosomal location in the

genome. Furthermore, information about clustering of different microsatellites in the genome can also be retrieved. Finally, to facilitate primer designing for PCR amplification of any desired microsatellite locus, 200 bp upstream and downstream sequences are provided. The database allows easy systematic retrieval of comprehensive information about simple and compound microsatellites, microsatellite clusters and their locus coordinates in 31 sequenced eukaryotic genomes. The information content of the database is useful in different areas of research like gene tagging, genome mapping, population genetics, germplasm characterization and in understanding microsatellite dynamics in eukaryotic genomes.

2.5.9 TPMD

Taiwan Polymorphic Marker Database (TPMD) is a marker database designed to provide experimental details and useful marker information allelotyped in Taiwanese populations accompanied by resources and technical supports. The current version deposited more than 3,72,000 allelotyping data from 1425 frequently used and fluorescent-labeled microsatellite markers with variation types of dinucleotide, trinucleotide and tetranucleotide. TPMD contains text and map displays with searchable and retrievable options for marker names, chromosomal location in various human genome maps and marker heterozygosity in populations of Taiwanese, Japanese and Caucasian. The integration of marker information in map display is useful for the selection of high heterozygosity and commonly used microsatellite markers to refine mapping of diseases locus followed by identification of disease gene by positional candidate cloning.

2.5.10 SSRD

Simple sequence repeats are predominantly found in most organisms. They play a major role in studies of genetic diversity, and are useful as diagnostic markers for many diseases. The Simple Sequence Repeats Database (SSRD) for the human genome was created for easy access to such repeats, for analysis, and to be used to understand their biological significance. The data includes the abundance and distribution of SSRs in the

coding and non-coding regions of the genome, as well as their association with the UTRs of genes. The exact locations of repeats with respect to genomic regions (such as UTRs, exons, introns or intergenic regions) and their association with STS markers are also highlighted. The resource will facilitate repeat sequence analysis in the human genome and the understanding of the functional and evolutionary significance of simple sequence repeats.

2.5.11 INSATDB

InSatDb presents an interactive interface to query information regarding microsatellite characteristics per se of five fully sequenced insect genomes (fruitfly, honeybee, malarial mosquito, red-flour beetle and silkworm). InSatDb allows users to obtain microsatellites annotated with size (in base pairs and repeat units); genomic location (exon, intron, up-stream or transposon); nature (perfect or imperfect); and sequence composition (repeat motif and GC %). One can access microsatellite cluster (compound repeats) information and a list of microsatellites with conserved flanking sequences (microsatellite family or paralogs).

2.5.12 FUNGAL GENOME

Fungal genome database is the study of examining and comparing SSRs in completely sequenced fungal genomes. They analyzed and compared the occurrences, relative abundance, relative density, most common and longest SSRs in nine taxonomically different fungal species: *Aspergillus nidulans, Cryptococcus neoformans, Encephalitozoon cuniculi, Fusarium graminearum, Magnaporthe grisea, Neurospora crassa, Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Ustilago maydis.* Their analysis revealed that, in all of the genomes studied, the occurrence, abundance, and relative density of SSRs varied and was not influenced by the genome sizes. No correlation between relative abundance and the genome sizes was observed, but it was shown that *N. crassa*, the largest genome analyzed had the highest relative abundance of SSRs. In most genomes, mononucleotide, dinucleotide, and trinucleotide repeats were more abundant than the longer repeated SSRs. Generally, in each organism, the

occurrence, relative abundance, and relative density of SSRs decreased as the repeat unit increased. Furthermore, each organism had its own common and longest SSRs. Their analysis showed that the relative abundance of SSRs in fungi is low compared with the human genome and that longer SSRs in fungi are rare. In addition to providing new information concerning the abundance of SSRs for each of these fungi, the results provide a general source of molecular markers that could be useful for a variety of applications such as population genetics and strain identification of fungal organisms.

Database	Details on					Coverage
	Simple Repeats	Compound Repeats	Clustering information	Genomic Positions	Flanking Sequences	
MICdb	Y	Y	N	Y	Y	19 archeal, 155 bacterial and 287 viral genomes
SilkSatDb	Y	Y	N	N	Y	Silkworm
MMDBJ	Y	Y	N	N	N	Mouse
CMD	Y	Y	N	N	Y	Cotton
Satellog	Y	N	N	Y	N	Human
Database of Molecular Mycology Research Lab.	Y	N	N	Y	N	9 fungal genomes
InSatDb	Y	Y	N	Y	Y	5 insect genomes
MRD	Y	N	N	N	N	8 eukaryotic genomes
SSRD	Y	N	N	N	N	Human
EuMicroSatdb	Y	Y	Y	Y	Y	31 eukaryotic genomes

Table 1: Showing different databases which are available in public domain and comparison of features covered by them.

So, to develop new microsatellite markers there is need to know various microsatellite positions. One can know the different microsatellite positions in the desired Species by using various microsatellite finding tools available in the public domain but for that one has to go through various multiple steps. Till date, no such database on microsatellites of all Human Pathogen is available which can fetch detailed information of all four pathogen such as Bacteria, Viruses, Parasites and Fungi. So to facilitate the scientists or people who are involved in genomic research of any human disease caused by various pathogen we have developed *HuPathoMicroSatDB* database, which finds immediate and accurate microsatellites in any of the Human Pathogen. *HuPathoMicroSatDB* database would be helpful to know new microsatellite positions

and as well as the flanking region information (200bp each side of the microsatellite sequence) to design primers for the development of microsatellite markers. This *HuPathoMicroSatDB* would be beneficial for microsatellite analysis of any pathogen whether its Bacterial, Viral, Parasites or Fungi in any research area of disease or drug design for any disease.

Chapter - III
Materials & Methods

CHAPTER III

3. MATERIALS AND METHODS

3.1 OVERALL ARCHITECTURE OF THE DATABASE

The overall architecture of database is based on three tier system:

- User interface (Client side)
- Application layer
- Data store (Server side)

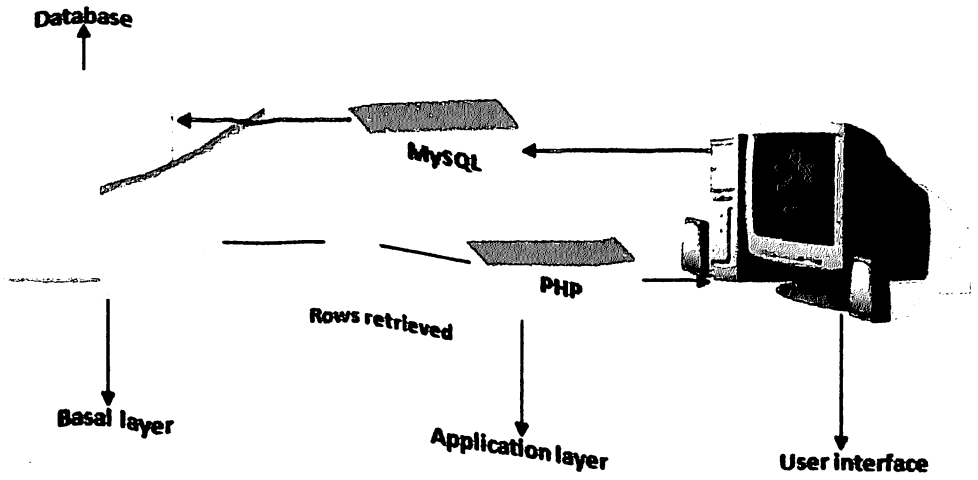


Figure 3: Architecture of database

This architecture Is Implemented Using **XAMPP** System

The XAMPP Technologies (Windows XP, Apache, MySQL, Perl, PHP/Python) is refers to a solution stack of software, usually free and open source software, used to run dynamic Web sites or servers. The original expansion is as follows:

- Windows, an operating system;
- Apache, for Web server;
- MySQL, it is for database management system.
- PHP programming language
- Python and Perl, for the programming language

The combination of these technologies is used primarily to define a web server infrastructure, define a programming paradigm of developing software, and establish a software distribution package for any application. This really is the programming framework of choice of a multitude of programmers worldwide.

3.2 XAMPP COMPONENTS

3.2.1 *Window Xp:* This is a source operating system that is based on Window and runs on a wide variety of hardware have been used for database creation.

3.2.2 *Apache:* Apache is a free industrial strength, open source stable, that's completely flexible and extensible. Apache on Linux is the web server used in 95 % of all web sites across the globe to deliver web content. Apache also runs on the windows O/s. Commonly known as Apache HTTP Server, it is a web server that can be used to serve both static as well as dynamic web pages. Apache is available for a variety of operating systems like: Microsoft Windows, Novell Netware and Unix. It provides different modules to add functionality to the basic server. Apache also provides a host of features that assist in application development such as: GUI based server configuration interfaces, URL rewriter enabling the creation of search engine friendly URLs and running of multiple websites from a single Apache installation etc.

3.2.3 *MySQL:* It is one of the most popular open source relational database management systems that make the best use of SQL to process data in a database. In addition to this it also provides great mechanisms ensuring that only authorized users have entry to the database server. Some of its features that make this database management system the most dominant transactional database engines on the market are: unlimited row-level locking, distributed transaction capability, and multi-version transaction support.

3.2.4 *PHP:* It is an open source server side scripting language that can be used to develop a whole range of dynamic web applications. PHP has built-in, native and very robust connectivity to MySQL and other databases. It is based on an Object Oriented Architecture and most of its features, concepts and syntax are based on the C and Perl

programming languages. Widely used as scripting language, it generally runs on a web server, taking PHP code as its input and creating Web pages as an output.

3.2.5 *Python and Perl:* It is also an open source server side scripting language that can be used for the data retrieval from the database. Fasta file are retrieved and extraction of microsatellite is done through coding perl and python.

3.3 COMMUNICATION BETWEEN THREE LAYERS

At the front end is user interface which is client side. This interface is designed to receive query and search data. Designed using html and PHP embedded in it. The initiation begins with query entered by user in the form of search term. These are received by html form (first layer) and using get or post method these request are send to the server side. At the middle level is application layer which is PHP which is embedded in html and also present in the form of PHP scripts at the server side to respond to these request. Apache is the web server responsible for responding to requests received from client browsers for information. MySQL is the database in which such information is stored. PHP is the middle layer programming environment of choice

1. That responds to such information requests being processed by web server.
2. Access the MySQL database where the information requested is stored.
3. Converts this to HTML.
4. Return this to the client browser via apache, thus servicing the client's request for information.

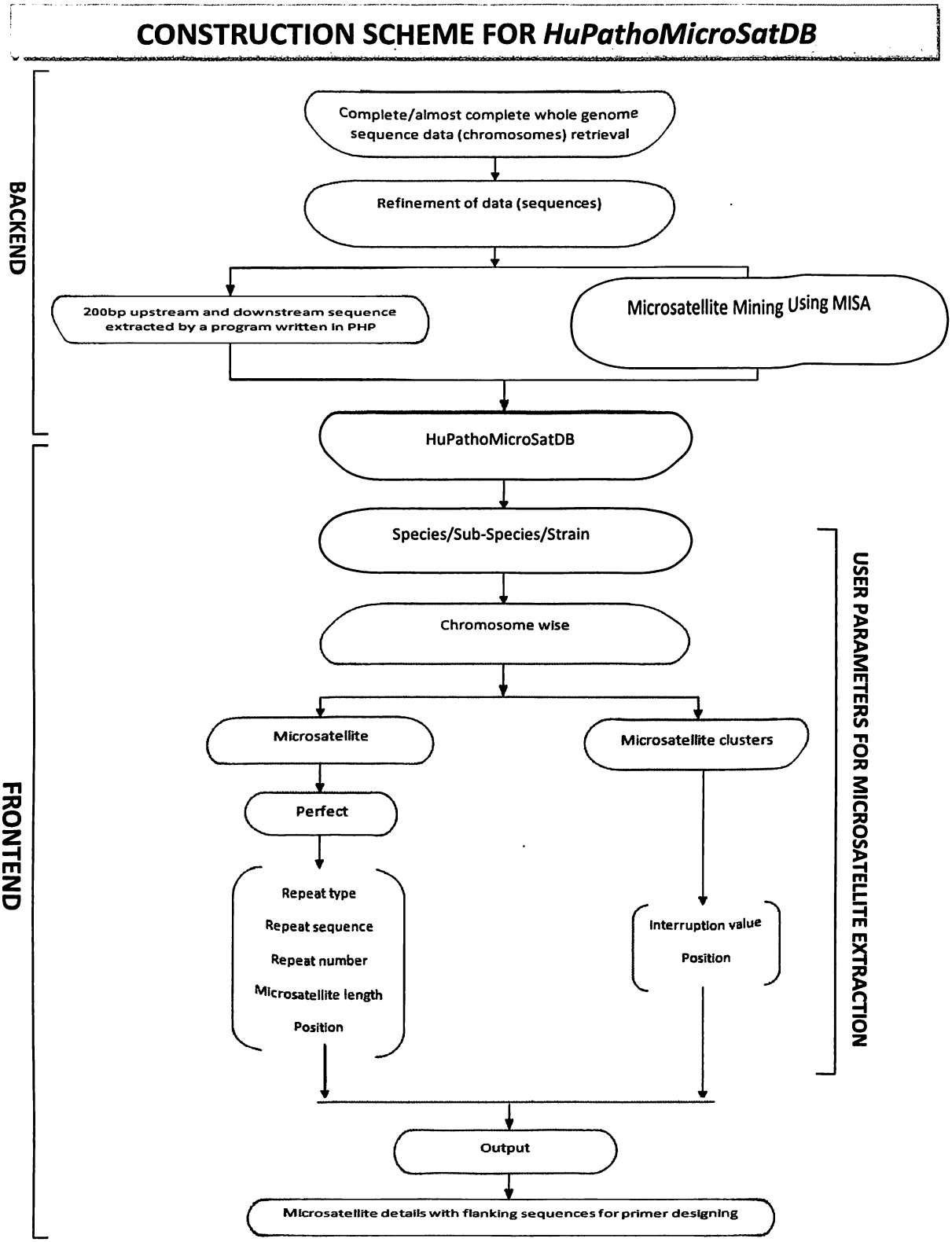


Figure 4: Construction scheme for *HuPathoMicroSatDB*

3.4 DATA COLLECTION

The data (Strain & Chromosome sequences) being most important thing for database, the approach was data to be accurate and complete. The sequence collection has been done from two sources:

- National Center for Biotechnology Information **(NCBI).**
- EMBL-EBI.

These Strain & Chromosome sequences collected which were then refined by removing special characters (paragraph marks) and checked the data redundancy.

3.5 MICROSATELLITE REPEAT FINDING

- After that all microsatellite repeats (di, tri, tetra, penta and hexa) were found by using MISA which is a repeat finding tool and freely available. We had modified this tool according to our convenience in such a way that all the generated data (microsatellite repeats) are entered to the table automatically. This automation saved our time of data entry and secondly, there were no chance of errors of data entry.

- Modifications which had done to the MISA tool lead to extraction of 200bp upstream and downstream flanking sequences.

- Microsatellite clustering information was also generated by using a program which was written in PHP.

- The data which was generated by using MISA tool :

    a) Perfect Microsatellite repeats.
    b) Compound Microsatellite repeats
    c) Microsatellite clusters.

- This data then entered in the database for perfect, compound and clusture microsatellite repeats.

3.6 DATABASE ARCHITECTURE AND DESIGN

- The design of database is very simple. We have created 233 tables for 233 species for perfect microsatellites and compound (mismatch/imperfect) microsatellites. The data which was generated by modified MISA algorithm was saved in the excel table. After that each excel data is imported into the 233 table for their specific species.
- For the creation of database MySQL was chosen as database environment because
    a) MySQL is ideal for both small and large applications
    b) MySQL supports standard SQL
    c) MySQL compiles on a number of platforms
    d) MySQL is free to download and use
    e) MySQL is one of component of XAMPP

3.6.1 Structure of Tables:

Database Name: ***PathoData***

a) Table name: **Species Name**

I. Purpose: Holds all types of microsatellite repeats, upstream and downstream sequences.

II. Primary key: Species

Field Name	Data Type	Description
Species	Varchar	Stores the name of the Species
Sub-Species	Varchar	Stores the Sub-Species of corresponding Species

Strain	Varchar	Stores the strain of Species/Sub-Species
Chromosome	Varchar	Stores the Chromosome of Species/Sub-Species/Strain
SSR Type	Varchar	Stores the type of SSR
Repeated Sequence	Varchar	Stores microsatellite repeat sequence
Repeat Sequence Length	Varchar	Stores the length of Microsatellite repeat sequence
Start	Int	Stores Position where SSR repeat Start
End	Int	Stores Position where SSR repeat End
Upstream/Downstream Bases	Varchar	Stores bases for Primer Design

Table 2: Structure of PATHODATA Database 233 table

3.7 DATA ENTRY

This part is the most crucial step of database development. The data entry started with entry of data obtained by parsing of sequence entries files and entry of data into main table. This entry of data was facilitated by phpMyAdmin tool of XAMPP(This whole project was developed using XAMPP tool which provide integrated environment of PHP, MySQL, Apache on windows system) which allows for data entry in a form manner. All data (microsatellite repeats) generated by MISA microsatellite mining tool automatically entered in to tables due to modification done by us. Further 200bp upstream and downstream were also entered automatically to the respective microsatellite repeat sequence in the tables.

3.8 DESIGN OF INTERFACE

This is the interface which is accessible by user. It includes main pages and other pages navigated by user. It was designed by mainly using HTML, java script and CSS to bring the uniformity. For designing mainly Dreamweaver was used a tool to hand code this interface. This interface is kept simple and elegant. This contains a horizontal menu with options for home, search, about, help, contact us. This provides navigation to other pages. This interface also contains some basic detail about the database.

3.9 DATABASE CONNECTIVITY

This was done by writing scripts in PHP programming language. For this purpose PHP is very suitable being it can be easily embedded in HTML, secondly it is free and also this is server side scripting language and on user request only pure html is sent back to browser not PHP coding so safeguarding the code from misuse. Until now before connectivity both steps were complete that is data entry and interface design of main page and other pages. So by using PHP the interface was connected with MySQL tables. Fields in the table were placed specifically on pages and tested that the data are fetched correctly.

Chapter - IV
Results & Discussion

CHAPTER IV

4. RESULTS AND DISCUSSION

The *HuPathoMicroSatDB* allows mining of different microsatellites along with their physical location on chromosomes in completely sequenced *Bacterial, Viruses, Parasite and Fungi* genomes. These whole genome sequences are assembled as Species, Sub-Species, Strain and Chromosome. At present, the database has over million entries of microsatellites covering 233 Species, 929 Strain and 1070 Sample. More genomes will be included in the database as and when their whole genome sequences are published and made available in the public domain. User can search for perfect repeats, compound microsatellite and microsatellite clusters. *HuPathoMicroSatDB* database can be searched using following need based input parameters:

Repeat unit length: the basic unit that is tandemly repeated in the microsatellite ranging from mononucleotide to hexanucleotide.

Repeat sequence: this parameter allows the user to search microsatellite for a specific base sequence, for example, AT, GCG, etc.

Repeat number: is used to search microsatellites on the basis of repeat number of the microsatellite e.g. $(CCT)_9$ has a repeat number of 9, $(AGAGG)_{10}$ has a repeat number of 10.

Microsatellite length: searches microsatellites on the basis of their total length in base pairs e.g. $(TTGCA)_5$ has a length of 25 bp.

Position: defined locations on the chromosome in terms of base pairs can be specified.

Microsatellite cluster: search can also be performed to look for adjacent microsatellites.

Further, if the user wants to design primers for PCR amplification of the desired microsatellite locus, the database also provides 200 bp upstream and downstream regions of all the microsatellite loci. The search options are further explained with the help of some case studies given in a power point tutorial available on the database

4.1 WEB INTERFACE

Figure 5: Snapshot of home page of *HuPathoMicroSatDB*

This is the home page of HuPathoMicroSatDB database and user will see this page. Many options are there on menu bar (Home, Search, About, Help and Contact us). Here we have given an additional feature of Google search, so that if user wants to search anything else can directly search from our home page.

4.2 VARIOUS SEARCH CRITERIA IN HuPathoMicroSatDB

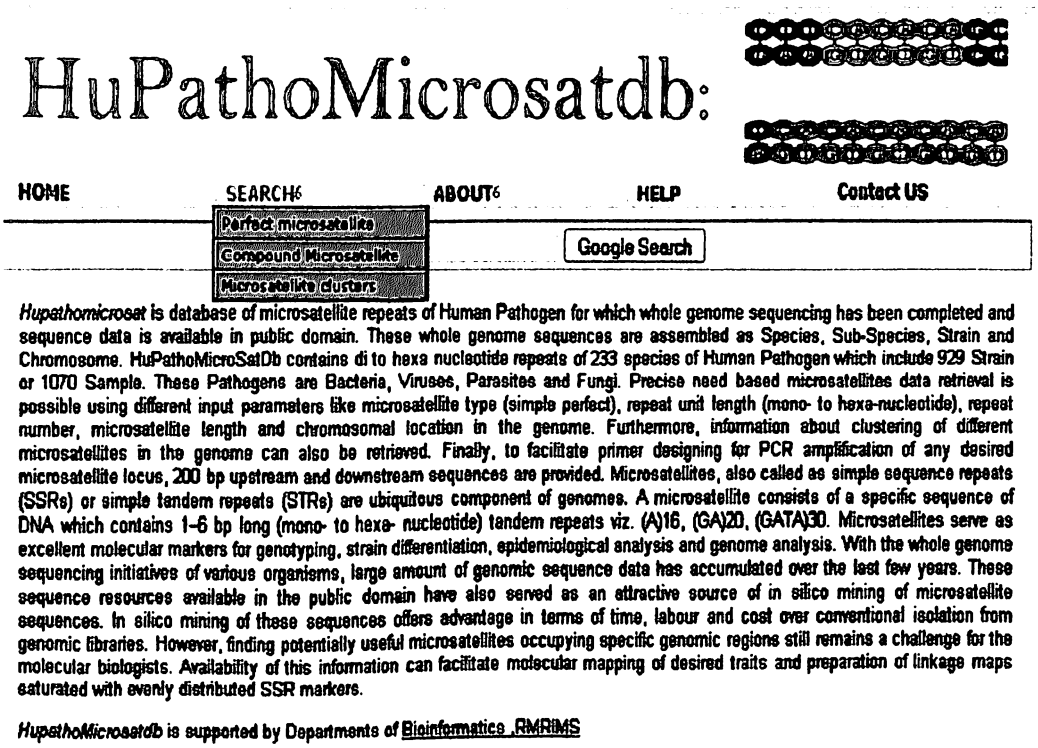


Figure 6: Snapshot of search menu bar option of *HuPathoMicroSatDB*

4.2.1 PERFECT MICROSATELLITE SEARCH.

4.2.2 MICROSATELLITE COMPOUND SEARCH.

After selecting search option from menu bar above two options will appear as shown in figure 6 and user can select the option for which user wants to know the microsatellites. Here we have given examples of different search cases for Different Pathogen. When user selects the perfect microsatellite search option from the above menu bar then the following search window will appear in front of user.

4.2.1 Perfect microsatellite search criteria

Case 1: If user want to search perfect microsatellite repeats in Virus for their strain, firstly select the Pathogen for which user wants to see microsatellite repeats and then select the Species, than select the Sub-Species if available if not please select No-Sub-

Species for next step for the selection of Strain and then ultimately select Chromosome if for the required species chromosome is available.

Query:

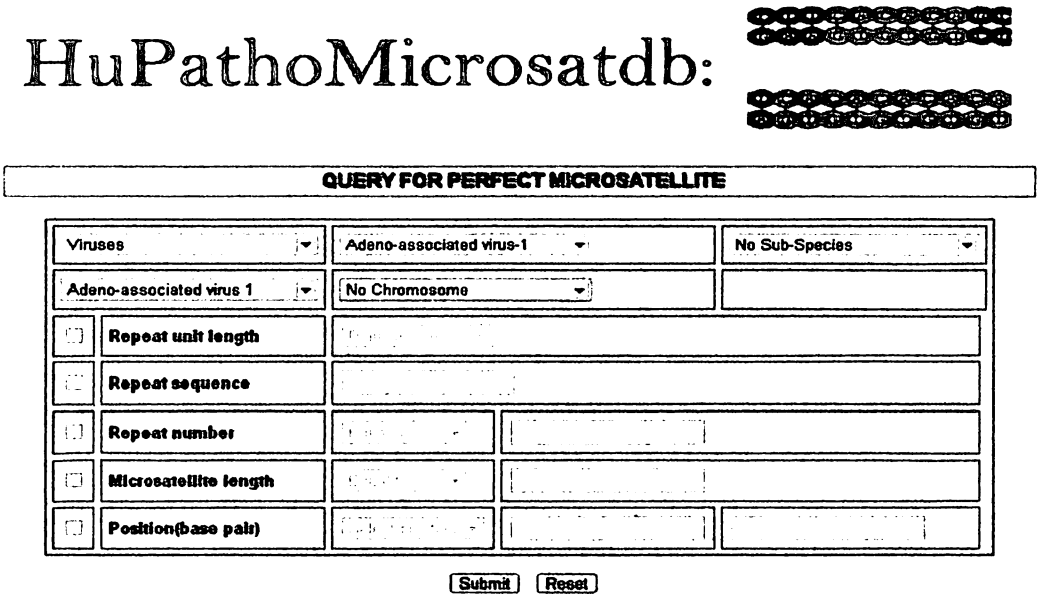


Figure 7: Snapshot of perfect microsatellite search page

OUTPUT:

HuPathoMicrosatdb:

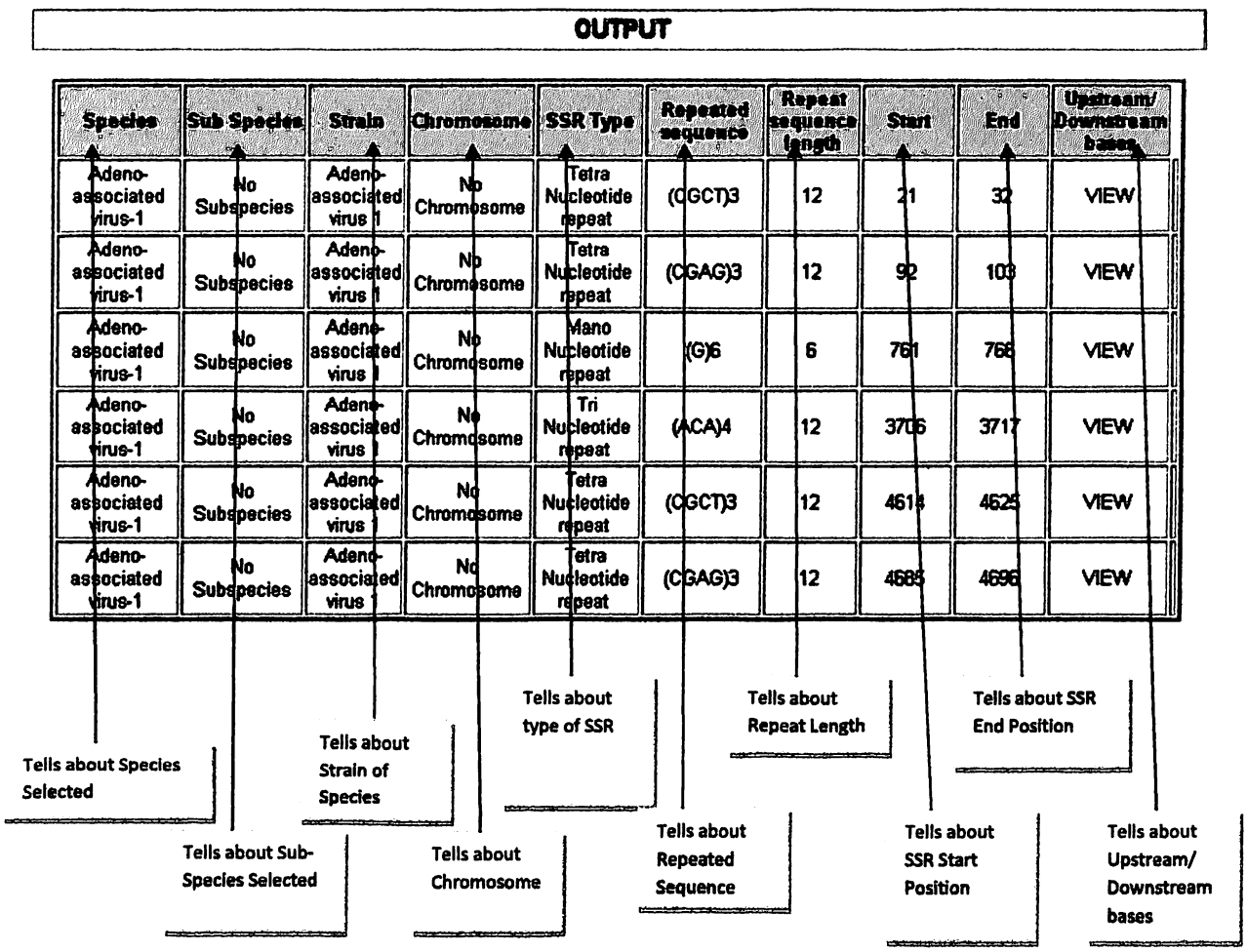


Figure 8: Snapshot of output of CASE1 perfect microsatellite search query

Here, the fig. 8 represents the microsatellite repeats in Species Adeno-Associated virus-1 for Strain Adeno-Associated virus Type 1 . There are so many columns, for which we have explained on the diagram itself. In above diagram, last column is for finding upstream and downstream bases.

When user will click on **VIEW** on the above window page, the following window will appear which contains 200bp upstream and downstream of the respective microsatellite repeat.

Case 2: If user wants to search for perfect microsatellite repeats of a Virus for particular species and different strain as given in case 1, Here we will search for same species but for different strain. Firstly select the Pathogen i.e.Virus for which user wants to see microsatellite repeats and then select Strain other then case 1. Let it be Adeno-Associated virus type 3.

Query:

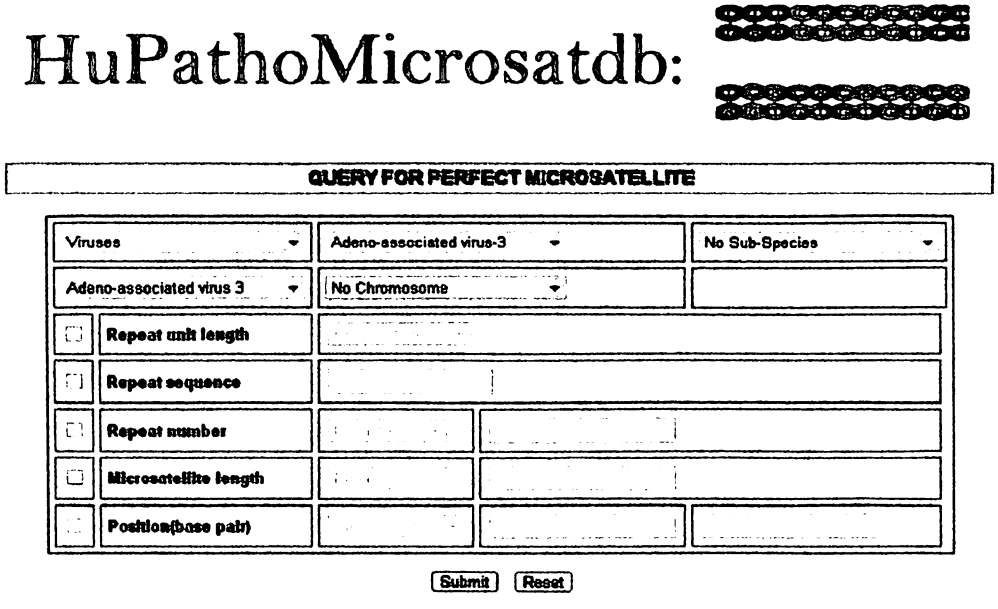
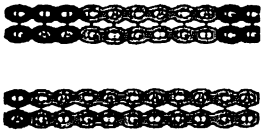


Figure 9: Snapshot of Case 2 perfect microsatellite search query

OUTPUT:

HuPathoMicrosatdb:



OUTPUT

Species	Sub Species	Strain	Chromosome	SSR Type	Repeated sequence	Repeat sequence length	Start	End	Upstream/ Downstream bases
Adeno-associated virus-3	No Subspecies	Adeno-associated virus 3	No Chromosome		(CTCG)3	12	23	34	VIEW
Adeno-associated virus-3	No Subspecies	Adeno-associated virus 3	No Chromosome		(CGAG)3	12	92	103	VIEW
Adeno-associated virus-3	No Subspecies	Adeno-associated virus 3	No Chromosome		(T)6	6	549	554	VIEW
Adeno-associated virus-3	No Subspecies	Adeno-associated virus 3	No Chromosome		(ACA)4	12	3692	3703	VIEW
Adeno-associated virus-3	No Subspecies	Adeno-associated virus 3	No Chromosome		(CTCG)3	12	4624	4635	VIEW
Adeno-associated virus-3	No Subspecies	Adeno-associated virus 3	No Chromosome		(CGAG)3	12	4693	4704	VIEW

Figure 10: Snapshot of Output of Case 2 perfect microsatellite search query

After selecting the above parameter we have seen that we have select the Species Adeno-Assocaited virus type 3 but we have selected Strain Adeno-Associated virus type 3. We got the result for same pathogen (virus) but different species and strain.

Case 3: If user wants to search for perfect microsatellite repeats of particular Strain but same species then please select pathogen (e.g Bacteria) and Species Bacillus Anthracis. After that user has a choice to select for particular Strain (Bacillus anthracis str A0248).

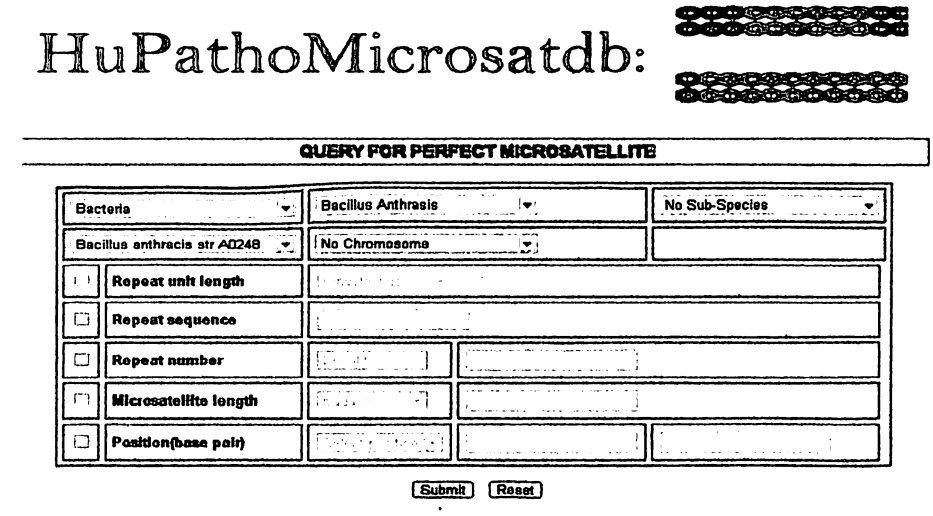


Figure 11: Snapshot of Case 3 perfect microsatellite search query

OUTPUT:

After selecting the above parameters when user clicks the submit button then all the repeats for particular strain will be developed.

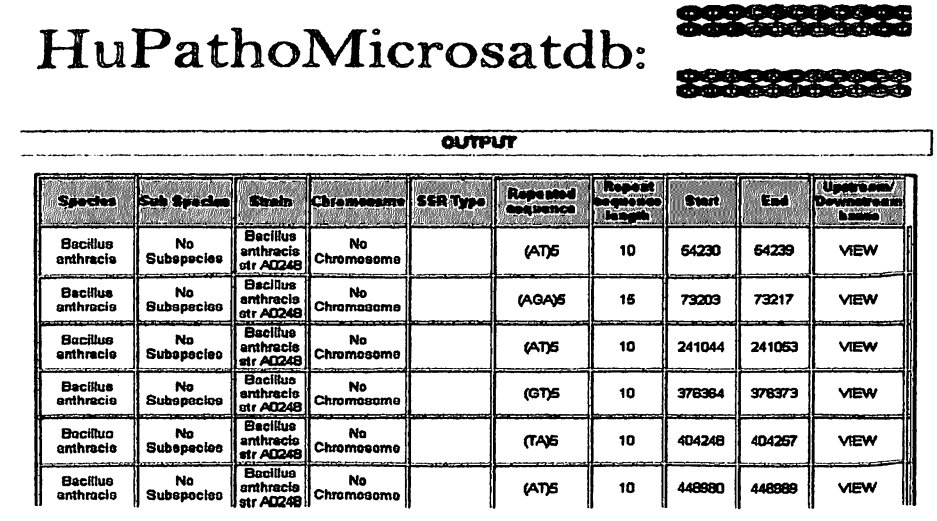


Figure 12: Snapshot of Output of Case 3 perfect microsatellite search query

After selecting the parameter we have selected we have got the output. Which show that we can select for any particular species with different strain.

4.2.2 *Microsatellite clustures search criteria*

If more than one type or even similar types of microsatellite repeats are present adjacent to each other, then these all are called microsatellite cluster. e.g.

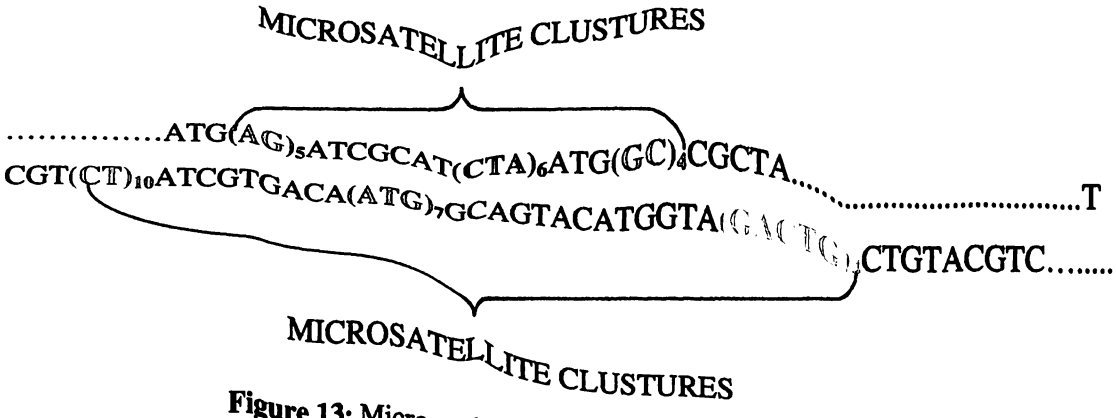


Figure 13: Microsatellite clustures in nucleotides sequence

User can identify the genomic regions showing high microsatellite density to study microsatellite clusturing in the genome by defining the size of interruption between neighbouring microsatellites.

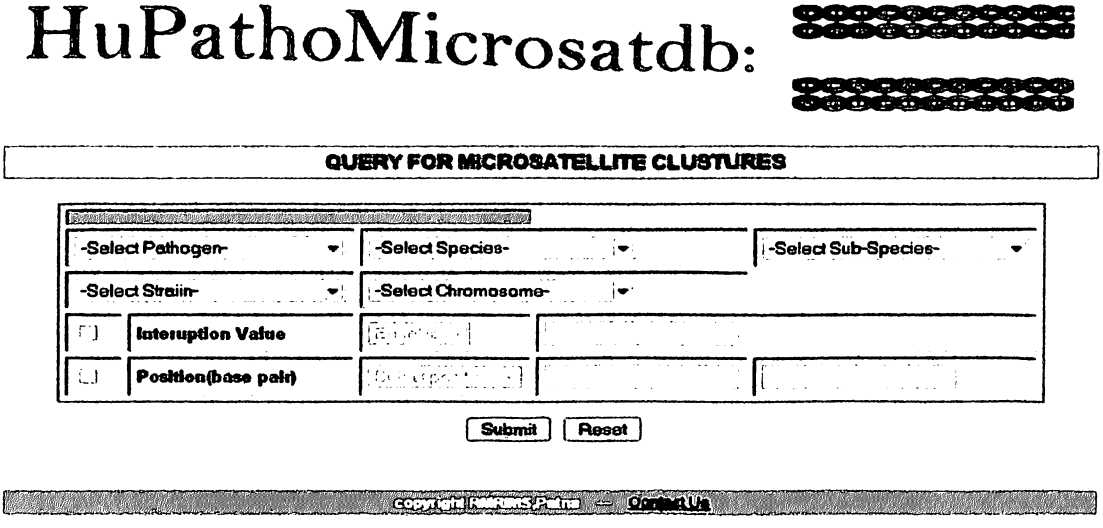


Figure 14: Snapshot of microsatellite clusture search form

Case 4: For searching microsatellite clustures interruption value is very important. Interruption value is defined as the difference between adjacent microsatellite repeats. Here we have given the example of Pathogen is Parasites, Species= Cryptosporidium parvum, Sub-Species= No, Strain= Cryptosporidium parvum Iowa II and chromosome = 1 with default interruption value.

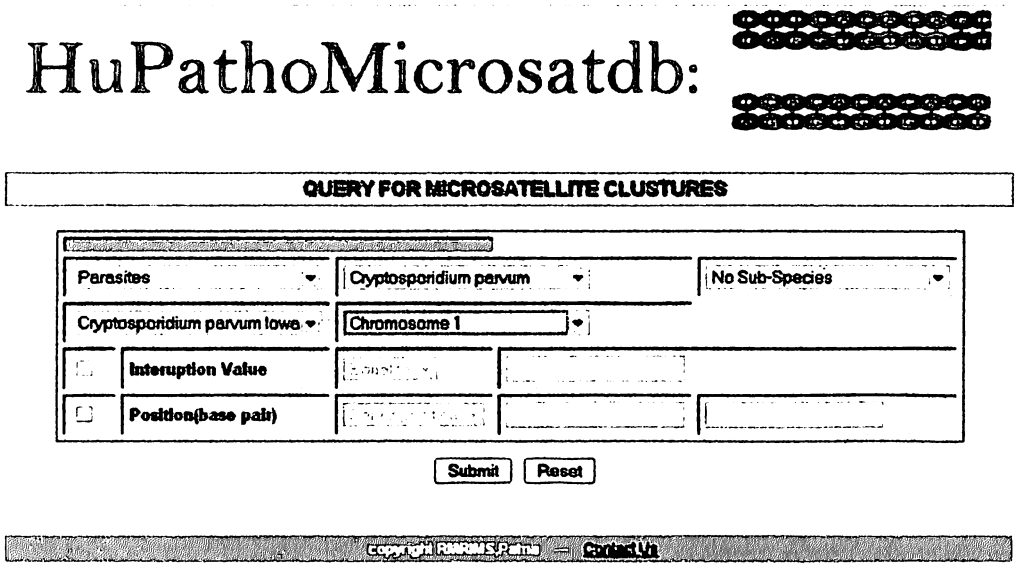


Figure 15: Snapshot of Case 4 microsatellite clustures search query

This search window will appear in front of user when user selects the microsatellite clustures search option from menu bar. User can put query to search for microsatellite clustures.

OUTPUT:

Species	Sub Species	Strain	Chromosome	SSR Type	Repeated sequence	Repeat sequence length	Start	End	Upstream Downstream bases
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum	No Chromosome	Cluster1	(ATT)11(AGT)6	51	72925	72975	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum	No Chromosome	c*	(GA)7(A)10*	23	122933	122955	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum	No Chromosome	Cluster2	(A)10(TA)7	24	135176	135199	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum	No Chromosome	c	(T)11(A)11	22	733345	733366	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum	No Chromosome	c*	(A)15(AT)9*	32	961244	961275	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum	No Chromosome	c	(GTAGTG)3(GTA)3	72	1137378	1137449	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum	No Chromosome	c	(GTG)11(GTA)7	54	1139684	1139737	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum Iowa II	No Chromosome	c*	(T)15(TA)6*	26	81176	81203	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum Iowa II	No Chromosome	c*	(AT)21(ATTT)3*	72	239600	239671	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum Iowa II	No Chromosome	c	(TAC)6(TAA)9	45	692637	692681	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum Iowa II	No Chromosome	c*	(TCATTA)3(ATC)10*	77	694659	694735	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum Iowa II	No Chromosome	c*	(AT)13(T)15*	40	731312	731351	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum Iowa II	No Chromosome	c*	(CTTT)3(TC)7*	49	823633	823681	VIEW
Cryptosporidium parvum	No Subspecies	Cryptosporidium parvum Iowa II	No Chromosome	c*	(AAAT)3(AT)16*	54	17901	17954	VIEW

Figure 16: Snapshot of output of Case 4 microsatellite clustures search query

This is the output of above query. Here in above figure two microsatellite clusters are present one is from Start 72925 to 72975 and another one is from 122933 to 122955. Similar to this user can search the various microsatellite clusters at specific positions (by selecting desired position in query) on all chromosomes and of all species.

Chapter - V
Conclusion

CHAPTER V

CONCLUSION

HuPathoMicroSatDB is a database of microsatellite repeats of different types (di, tri, tetra, penta and hexa repeats) of repeats in different *Pathogen*. Microsatellites have immense utility as molecular markers in different fields like genome characterization and mapping, phylogeny and evolutionary biology. Recently, analysis of the length polymorphisms of microsatellite containing regions has become an important tool for population and genetic studies for many different species. To develop new microsatellite markers there is need to know various microsatellite positions, database which is developed by us would be helpful to know new microsatellite positions and as well as give the flanking region information(200bp each side of the microsatellite sequence) to design primer for development of microsatellite markers. Till date, no such database is available so this *HuPathoMicroSatDB* would be beneficial for microsatellite analysis of different *Pathogen*.

FUTURE PERSPECTIVES

HuPathoMicroSatDb is the database of microsatellite repeats of Pathogen for which whole genome sequencing has been completed and sequences data is available in public domain. Currently HuPathoMicroSatDb contains microsatellite information extracted from 929 strains or 1070 sample. As the database creation has been made fully automated the database can be updated for any number of genomes. We have planned to incorporate microsatellite repeats information of various species related to tropical diseases. In the future version, hyperlinks to other useful databases will also be provided thereby increasing the information content associated with the microsatellites.