DEVELOPMENT OF ONLINE RELATIONAL DATABASE FOR LEISHMANIASIS

A PROJECT REPORT

Submitted by

JANAKI.P
KISHORE.N.G

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

BIOTECHNOLOGY

KUMARAGURU COLLEGE OF TECHNOLOGY, COIMBATORE

ANNA UNIVERSITY : CHENNAI 600 025

APRIL 2009

ANNA UNIVERSITY : CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report **"DEVELOPMENT OF ONLINE RELATIONAL DATABASE FOR LEISHMANIASIS"** is the bonafide work of **"JANAKI.P AND KISHORE.N.G"** who carried out the project work under my supervision.

SIGNATURE

Dr. P. Rajasekaran

Professor & Head

Department of Biotechnology,
Kumaraguru College of Technology,
Coimbatore – 641 006

SIGNATURE

Dr. Stephen V. Rapheal

Supervisor

Senior Lecturer
Department of Biotechnology,
Kumaraguru College of Technology,
Coimbatore – 641 006

CERTIFICATE OF EVALUATION

College : KUMARAGURU COLLEGE OF TECHNOLOGY

Branch : BIOTECHNOLOGY

Semester : Eighth Semester

S. No.	Name of the Students	Title of the Project	Name of the Supervisor with Designation
01	JANAKI.P Reg.No:71205214008	**"DEVELOPMENT OF ONLINE RELATIONAL DATABASE FOR LEISHMANIASIS"**	Dr. Stephen V. Rapheal Senior Lecturer
02	KISHORE.N.G Reg.No:71205214012		

The report of the project work submitted by the above students in partial fulfillment for the award of Bachelor of Technology degree in Biotechnology of Anna University was confirmed to be the report of the work done by the above students and then evaluated.

(INTERNAL EXAMINER) **(EXTERNAL EXAMINER)**

Dedicated to our Beloved Parents

&

Respected Guide

Acknowledgements

ACKNOWLEDGEMENTS:

With our deepest sense of gratitude we extend our heartfelt thanks to **Dr.V.StephenRapheal** Senior Lecturer Department of Biotechnology, Kumaraguru college of Technology, for his relentless support, masterly guidance, creative ideas and patient efforts for successful completion of the project.

Our sincere thanks to **Dr.P.Rajasekaran**, Proffesor & Head, Department of Biotechnology, Kumaraguru College of Technology. His gracious and ungrudging guidance all through our project work is highly acknowledged with gratitude.

We are happy to thank **Dr.V.Stephen Rapheal**, our class advisor and Lecturer, Department of Biotechnology, Kumaraguru College of Technology for his unsolicited and timely help encouragement without any hesitation.

We wish to extend our thanks to all **Teaching and Non-Teaching staffs** of the Department of Biotechnology for their kind and patient help throught the project work.

We are happy to thank our **Principal**, **Dr.Joseph.V.Thanikal and management**, Kumaraguru college of Technology for providing us all the facilities to carry out the project work.

We are happy to thank **Dr.S.Sadasivam**, Dean Academics, Kumaraguru College of Technology for his motivation and encouragement without any hesitation.

We thank all our friends who physically and emotionally helped us to bring out the work successfully.

Finally, we wish to express our deep sense of gratitude to our beloved parents and family members for their constant encouragement and love, without whose inspiration this study would not have seen the dawn of day.

[P.JANAKI] [KISHORE.N.G]

Abstract

ABSTRACT:

A database is a structured collection of records or data that is stored in a computer system. The structure is achieved by organizing the data according to a relational database model. A relational database, simply defined, is a database that is made up of tables and columns that relate to one another. These relationships are based on a key value that is contained in a column. The relational database model was developed by E.F.Codd back in the early 1970's and is the most widely used format for most biological databases. Biological database design, development, and long-term management is a core area of the discipline of Bioinformatics. Data contents include gene sequences, textual descriptions, attributes and ontology classifications, citations, and tabular data. These are often described as semi-structured data. Leishmaniasis is a disease caused by protozoan parasites that belong to the genus *Leishmania* and is transmitted by the bite of certain species of sand fly, including flies in the genus *Lutzomyia* and *Phlebotomus*. The chemotherapeutic treatments currently available have a number of limitations due to poor efficacy, unacceptable host toxicity and drug resistance, and new targets drug targets are required. Leishmaniasis research information systematically obtained over the period of several decades is scattered in various journals, databases and genome projects.Genomes of Leishmania species are actively being sequenced.

Genomes of *Leishmania donovi* and *Leishmania infantum* have been completely sequenced. In the present study relevant genomic and scientific data about the various aspects of Leishmaniasis were collected and organized to create a searchable highly organized database which will be hosted on the internet.

Table of Contents

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

RDBMS	Relational Database Management System
MOD	Model Organism Database
WHO	World Health Organisation
CGSC	Coli Genetic Stock Center
SGD	Saccharomyces Genome Database
BDGP	Berkeley Drosophila Genome Project
TAIR	The Arabidopsis Information Resource
PHP	Hypertext Processor
WAMP	Windows-Apache-Mysql-Php
LAMP	Linux-Apache-Mysql-Php
BLAST	Basic Local Alignment Search Tool
HSP	High Scoring Segment Pair
PMC	Pubmed Central

Introduction

1. INTRODUCTION:

1.1 LEISHMANIASIS:

Leishmaniasis is a disease caused by protozoan parasites that belong to the genus *Leishmania* and is transmitted by the bite of certain species of sand fly, including flies in the genus *Lutzomyia* and *Phlebotomus*. The disease was named in 1901 for the Scottish pathologist William Boog Leishman. This disease is also known as Leichmaniosis, Leishmaniose, leishmaniose, and formerly, Orient Boils, Baghdad Boil, kala azar, black fever, sandfly disease, Dum-Dum fever or espundia(H.P. Pandey *et al.*,2006).

Most forms of the disease are transmissible only from animals (zoonosis), but some can be spread between humans. Human infection is caused by about 21 of 30 species that infect mammals. These include the *L. donovani* complex with three species (*L. donovani, L. infantum*, and *L. chagasi*); the *L. mexicana* complex with 3 main species (*L. mexicana, L. amazonensis*, and *L. venezuelensis*); *L. tropica; L. major; L. aethiopica*; and the subgenus *Viannia* with four main species (*L. (V.) braziliensis, L. (V.) guyanensis, L. (V.) panamensis*, and *L. (V.) peruviana*). The different species are morphologically indistinguishable, but they can be differentiated by isoenzyme analysis, DNA sequence analysis, or monoclonal antibodies.(Francois chappuis *et al.*,2007)

The three primary clinical forms of leishmaniasis are cutaneous, mucacutaneous and visceral leishamaniasis. Cutaneous leishmaniasis can be further divided into localized, diffuse cutaneous, recidivans, and post–kala azar dermal leishmaniasis (PKADL).

More than 12 million people in 88 countries are known to be infected with leishmaniasis, but the true burden remains largely hidden. Two million new cases – 1.5 million of cutaneous leishmaniasis, 500 000 of the visceral form of the disease – occur annually, but declaration of the disease is compulsory in only 32 countries and a substantial number of cases are never recorded. (Nancy Malla & R.C. Mahajan, 2006)

Leishmaniasis is a disease of poverty, and its victims are among the poorest. In India, a country with a high leishmaniasis burden, 88% of leishmaniasis patients have a daily income of less than US$ 2, poor socioeconomic environment and low educational level; they live in either remote rural areas or poor suburbs. There is social stigma associated with the deformities and disfiguring scars caused by some forms of leishmaniasis, and disease-related disabilities impose a great social burden, hampering productivity and socio economic development.(Sarman Singh.,2005)

Leishmaniasis presents a spectrum of clinical manifestations. The visceral disease is particularly prevalent in Bangladesh, India, Nepal, Sudan and Brazil. These countries together account for 90% of the global visceral leishmaniasis (VL) burden. Malnutrition is a well-known risk factor in the development of this form of leishmaniasis, and epidemics flourish under conditions of famine, complex emergency and mass population movement. The cutaneous disease is particularly prevalent in Afghanistan, Algeria, Brazil, Iran, Peru, Saudi Arabia and Syria, which together account for 90% of the global cutaneous leishmaniasis (CL) burden. Though far less lethal, epidemics of the cutaneous form are of particular concern in some countries and difficult to control. Other manifestations include post-kala-azar dermal leishmaniasis (PKDL), which can follow recovery from a VL infection and occurs in India and Africa (mainly in Sudan and Kenya) (R.K. Singh *et al.*, 2006).

Mucocutaneous leishmaniasis (ML), which can follow cutaneous leishmaniasis and is endemic in Mexico and Central and South America, produces lesions which can lead to extensive and disfiguring destruction of the mucous membranes of the mouth, nose and throat cavities.

More than 20 species of *Leishmania* can infect humans, and other species are emerging, especially in association with HIV/AIDS. Thirty species of sandfly have been incriminated in transmission of the disease. In some areas, leishmaniasis is a zoonotic infection involving various animal reservoirs, while in other areas humans are the sole reservoir of infection, making vector and reservoir control costly and often impractical. It is planned to eliminate visceral leishmaniasis (i.e. to eliminate the disease as a public health problem) from the Indian subcontinent by 2015. The tools under discussion for this include vector control (by insecticide spraying and

impregnated bednets), rapid diagnostic tests for active case detection, and the new drugs miltefosine and paromomycin. A vaccine for leishmaniasis would also be a boon for global disease control, but no effective vaccine is yet on the market.(Javier Nieto *et al.*,2006)

Visceral leishmaniasis is a severe form in which the parasites have migrated to the vital organs. Visceral leishmaniasis or Kala Azar, found in tropical countries, is the most lethal form of the disease and is caused by *Leishmania donovani*. If untreated, visceral leishmaniasis may cause mortality rates as high as 90% within 4 to 24 months of the infection. Several epidemic outbreaks of visceral leishmaniasis have been reported over the past 10 years, with the majority of them occurring in the tropical countries such as Sudan, Brazil, India, Bangladesh, and Afghanistan.

Cutaneous and mucocutaneous leishmaniases are the other variants of this disease and are caused by *Leishmania major* and *Leishmania mexicana*, respectively. An estimated 2 million new infections are caused every year due to cutaneous and visceral leishmaniases. A newer cause of concern against these diseases is their development as opportunistic infections in immunocompromised patients, such as those suffering from AIDS.(S.K. Bhattacharya *et al.*,2005)

1.1.1 LIFECYCLE OF LEISHMANIASIS:

Leishmaniasis is transmitted by the bite of female phlebotomine sandflies. The sandflies inject the infective stage, metacyclic promastigotes, during blood meals. Metacyclic promastigotes that reach the puncture wound are phagocytized by macrophages and transform into amastigotes. Amastigotes multiply in infected cells and affect different tissues, depending in part on which *Leishmania* species is involved. These differing tissue specificities cause the differing clinical manifestations of the various forms of leishmaniasis. Sandflies become infected during blood meals on an infected host when they ingest macrophages infected with amastigotes. In the sandfly's midgut, the parasites differentiate into promastigotes, which multiply, differentiate into metacyclic promastigotes and migrate to the proboscis (Alam M.Z *et al*, 2008)

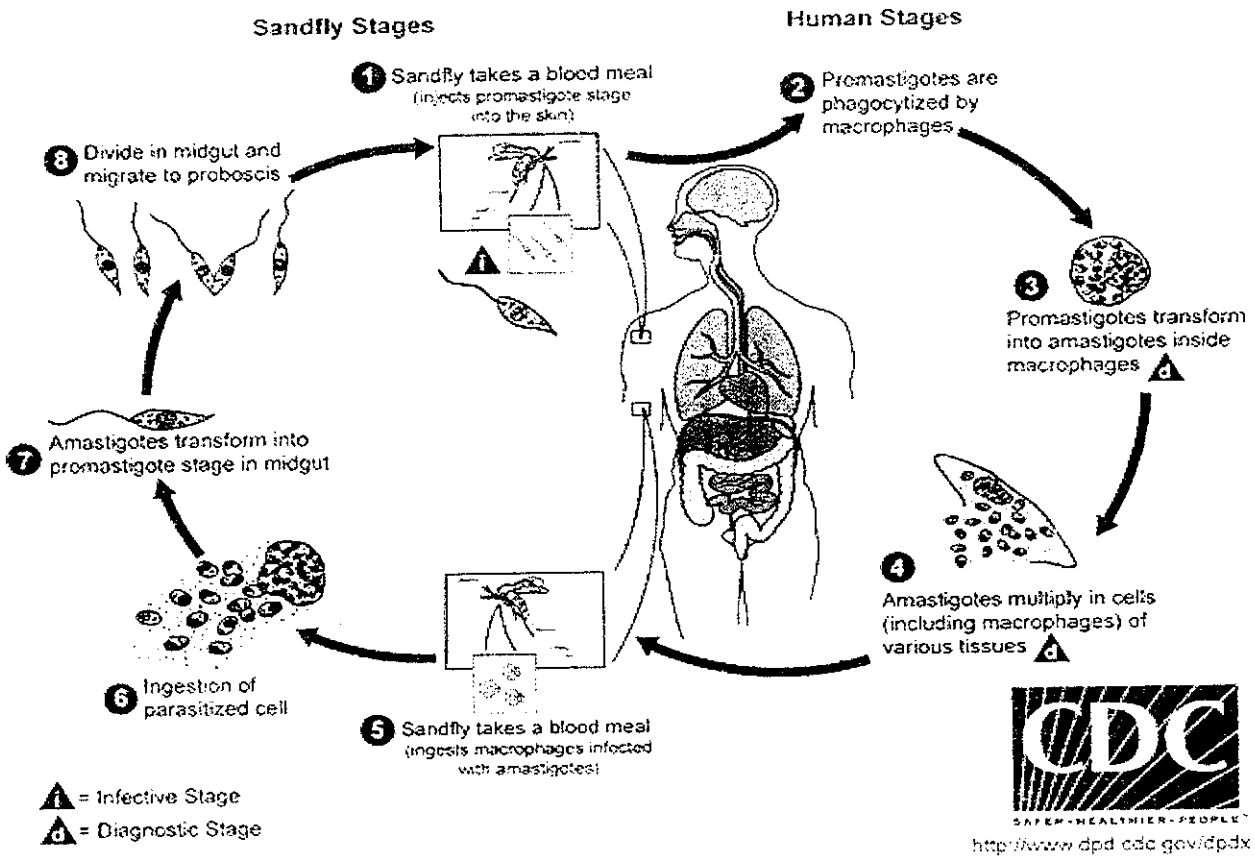


Fig 1.1 Life Cycle of Leishmania

1.1.2 LEISHMANIASIS IN AIDS PATIENTS:

Leishmania-HIV co-infection has emerged as a major complication of leishmaniasis. Most of the *Leishmania* endemic countries are also facing HIV epidemic and resulting into high rate of co-infection. Of the first 1700 cases of *Leishmania*-HIV coinfection reported to the World Health Organization from 33 countries up to 1998, 1440 cases were from the Mediterranean region: Spain (835); Italy (229); France (259); and Portugal (117). It is very important to know that HIV modifies the clinical presentation of Leishmaniasis in the co-infected patients. Several atypical aetiologic strains and species have been described in HIV-infected subjects. In HIV associated leishmaniasis caused by non-visceralizing species the parasite may disseminate to reticuloendothelial system and various other organs, and conversely the visceralizing species can manifest in atypical manner. Fulminant presentation of VL is possible in patients with AIDS, and relapses are usual. VL is now the fourth most

common opportunistic parasitic disease in HIV-positive individuals in Spain and 20-40 per cent cases had absence of splenomegaly. In Africa particularly Ethiopia and Sudan and Southern Europe, HIV-*Leishmania* co-infection is regarded as emerging disease, and as many as 70 per cent adults with VL also have HIV infection. Recently this co-infection has been noticed in Asia also. At least 10 cases of HIV-*L.donovani* have been reported from India. Dissemination is common and gastrointestinal involvement is commonest. The *Leishmania* amastigotes can be seen in gastrointestinal mucosal biopsy specimen and are commonly found in Kaposi's sarcoma cutaneous lesions concomitant with VL. *Leishmania* parasites were recently found in herpes zoster lesions in an HIV-positive patient (P.K. Sinha et al., 2005 & Israel Cruz *et al.*, 2006).

1.1.3 TREATMENT OF LEISHMANIASIS:

The chemotherapeutic treatments currently available have a number of limitations due to poor efficacy, unacceptable host toxicity and drug resistance, and new targets drug targets are required. Trypanothione reductase may be an ideal target for structure based drug design. In the current report, we describe the structure of trypanothione reductase of *Leishmania mexicana* and *L. amazonensis.* This species is considered in this report since it causes cutaneous leishmaniasis and vicerial leishmaniasis, which occurs as a coinfection with AIDS, so this species, has gained importance and a potent drug target against this organism is of current interest.

Newer serological tests for determining Leishmaniasis infection (i.e. ELISA) do not function as well in immunocompromised patients who aren't making antibodies to infections. In these situations, two or more tests must be used making the diagnostic procedure more expensive and less reliable.

The current drugs of choice have been on the market since the 1940s, well before the emergence of HIV. Little is known about if and how the drugs for Leishmaniasis cross react with the drugs for HIV. Furthermore, the drug regimens were designed for use on immunocompetent patients and it is unknown what the dosages should be for immunocompromised individuals. (Madeira da Silva L *et al.*, 2009)

1.2 DATABASES:

A database is a structured collection of records or data that is stored in a computer system. The structure is achieved by organizing the data according to a database model. The model in most common use today is the relational model. Other models such as the hierarchical model and the network model use a more explicit representation of relationships. Database contains a number of tables. Each table is made up of a series of columns. Data is stored in rows, and the place where each row intersects a column is known as a field.(Date, C. J *et al.*,2003)

1.2.1 PURPOSE OF DATABASES:

Databases are designed for three main purposes. These are to organize, store, and retrieve information as efficiently and effectively as possible. Probably the retrieval of information is the most important of these three. (Kroenke *et al.*, 1997)

1.2.2 MODELS OF DATABASES:

1.2.2.1 Hierarchical Databases: As its name implies, the Hierarchical Database Model defines hierarchically-arranged data.Perhaps the most intuitive way to visualize this type of relationship is by visualizing an upside down tree of data. In this tree, a single table acts as the "root" of the database from which other tables "branch" out.Relationships in such a system are thought of in terms of children and parents such that a child may only have one parent but a parent can have multiple children. Parents and children are tied together by links called "pointers" (perhaps physical addresses inside the file system). A parent will have a list of pointers to each children.

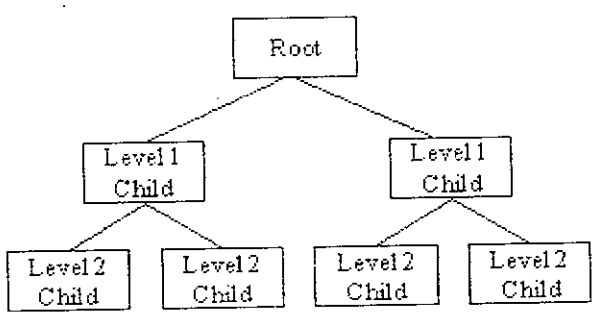


Fig 1.2 Representation of Hierarchial Database Model

This child/parent rule assures that data is systematically accessible. To get to a low-level table, you start at the root and work your way down through the tree until you reach your target. The hierarchical model however, is much more efficient than the flat-file model we discussed earlier because there is not as much need for redundant data. If a change in the data is necessary, the change might only need to be processed once

1.2.2.2 Network Database: Network Database model was designed to solve some of the more serious problems with the Hierarchical Database Model. Specifically, the Network model solves the problem of data redundancy by representing relationships in terms of sets rather than hierarchy. The model had its origins in the Conference on Data Systems Languages (CODASYL) which had created the Data Base Task Group to explore and design a method to replace the hierarchical model.

The network model is very similar to the hierarchical model actually. In fact, the hierarchical model is a subset of the network model. However, instead of using a single-parent tree hierarchy, the network model uses set theory to provide a tree-like hierarchy with the exception that child tables were allowed to have more than one parent. This allowed the network model to support many-to-many relationships.Visually, a Network Database looks like a hierarchical Database in that you can see it as a type of tree. However, in the case of a Network Database, the look is more like several trees which share branches. Thus, children can have multiple parents and parents can have multiple children.

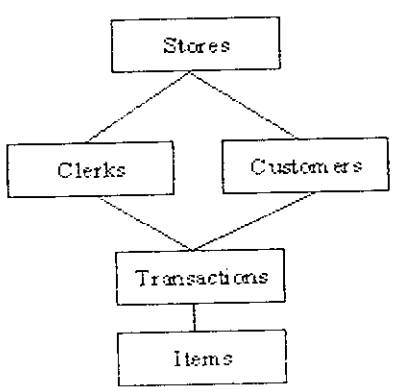


Fig 1.3 Representation Of The Network Database Model

Nevertheless, though it was a dramatic improvement, the network model was far from perfect. Most profoundly, the model was difficult to implement and maintain. Most implementations of the network model were used by computer programmers rather than real users. (Mark Whitehorn *et al.*, 2003)

1.2.2.3 Relational Databases: A relational database, simply defined, is a database that is made up of tables and columns that relate to one another. These relationships are based on a key value that is contained in a column. The relational database model was developed by E.F.Codd back in the early 1970's.Data access methodology makes the relational model a lot different from and better than the earlier database models because it is a much simpler model to understand. This is probably the main reason for the popularity of relational database systems today.

Another benefit of the relational system is that it provides extremely useful tools for database administration. Essentially, tables can not only store actual data but they can also be used as the means for generating meta-data (data about the table and field names which form the database structure, access rights to the database, integrity and data validation rules etc)

1.2.2.4 Post-Relational Database Models: Products offering a more general data model than the relational model are sometimes classified as post-relational. The data model in such products incorporates relations but is not constrained by the Information Principle, which requires that all information is represented by data values in relations.Some of these extensions to the relational model actually integrate concepts from technologies that pre-date the relational model. For example, they allow representation of a directed graph with trees on the nodes.Some products implementing such models have been built by extending relational database systems with non-relational features. Others, however, have arrived in much the same place by adding relational features to pre-relational systems.

1.2.2.5 Object Database Models: In recent years, the object-oriented paradigm has been applied to database technology, creating a new programming model known as object databases. These databases attempt to bring the database world and the application programming world closer together, in particular by ensuring that the

database uses the same type system as the application program. This aims to avoid the overhead (sometimes referred to as the *impedance mismatch*) of converting information between its representation in the database (for example as rows in tables) and its representation in the application program (typically as objects). At the same time, object databases attempt to introduce the key ideas of object programming, such as encapsulation and polymorphism, into the world of databases.(Galindo, J *et al.*, 2006)

1.2.3 BIOLOGICAL DATABASE:

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures. (Date, C. J *et al.*, 2003)

Relational database concepts of computer science and Information retrieval concepts of digital libraries are important for understanding biological databases. Biological database design, development, and long-term management is a core area of the discipline of Bioinformatics. Data contents include gene sequences, textual descriptions, attributes and ontology classifications, citations, and tabular data. These are often described as semi-structured data, and can be represented as tables, key delimited records, and XML structures. Cross-references among databases are common, using database accession numbers.

Biological databases have become an important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life.

Leishmaniasis research information systematically obtained over the period of several decades is scattered in various journals, databases and genome projects. Recently the genomes of *Leishmania donovi* and *Leishmania infantum* have been sequenced.

1.2.4 RELATIONAL DATABASE MANAGEMENT SYSTEM:

The RDBMS stores data logically in the form of table spaces and physically in the form of data files. Table spaces can contain various types of memory segments; for example, Data Segments, Index Segments etc. Segments in turn comprise one or more extents. Extents comprise groups of contiguous data blocks. Data blocks form the basic units of data storage. (Mark Maslakowski,2000)

1.2.5 DATABASE SCHEMA:

The schema of a database system is its structure described in a formal language supported by the database management system (DBMS). In a relational database, the schema defines the tables, the fields in each table, and the relationships between fields and tables.Schemas are generally stored in a data dictionary. Although a schema is defined in text database language, the term is often used to refer to a graphical depiction of the database structure.

1.3 DATABASE FOR SPECIFIC ORGANISMS:

Model Organism Databases (MODs) describe genome and other information about important experimental organisms in the life sciences. Also called organism-specific databases, these databases capture the large volumes of data and information being generated by modern biology. Behind every MOD is a software system that is designed to help manage the data within the MOD, and to help users query and access those data. In the past, every MOD project developed its own software tools.More recently, the Generic Model Organism Database (GMOD) Project began as an effort to create reusable software tools for developing MODs. GMOD is a loose federation of software applications (components) aimed at providing functionality that is needed by many or all model organism databases. (Arnaiz O *et al.*, 2007)

1.3.1 DATABASES FROM THE PERSPECTIVE OF MODEL ORGANISM RESEARCH:

In the last 100 years, research on a handful of organisms has played a profound role in advancing our understanding of the biological and biomedical sciences. The need to capture, organize, and access data from these model organisms has driven the creation of organism-specific databases. These model organism databases have allowed researchers to sift through masses of data, to gain access to information or materials they might have missed, and to go in new research directions. Comparative analysis has proven to be valuable in increasing our understanding of biological processes, including those in humans. Because these MODs are of immense value, offer tremendous opportunities, and represent a significant fiscal investment, it is timely to examine issues pertaining to the establishment, maintenance, evaluation, and future directions of model organism databases.(Stein LD *et al.*, 2002)

1.3.2 THE MODS AS RESEARCH RESOURCES:

MODs deal with two sets of research communities, with different needs and expectations:

- **Model organism community:**

  This community provides the data to a MOD, adds value by contributing to the curation of the data, and comprises a major set of users who need access to a great deal of specialized information, such as strain collections.

- **General research community**:

  This community uses but does not directly contribute information to MODs. Unlike the model organism community, the general community does not usually understand the specialized jargon and nomenclature for a model organism. The MODs should provide accessible summaries of genomic, functional, and phenotypic information in addition to full access to the

underlying datasets. (William M. Gelbart, Lansdowne Conference Center,1998).

The project addresses the need to develop a database solely devoted to Leishmaniasis.

Literature Review

2. LITERATURE REVIEW:

2.1 LEISHMANIA INFORMATION ON WEB:

A vast amount of leishmanial information has been collected by various research groups and medical experts. The data is distributed in various books,journals and internet.some major leishmanial information sources include

Books:

1. Leishmania: After The Genome
Publisher: Caister Academic Press
Edited by: 1, Peter J. Myler and 2, Nicolas Fasel 1-Seattle Biomedical Research Institute (SBRI), Seattle, WA, USA; 2- University of Lausanne, 1066 Epalinges, Switzerland
Publication year: 2008
ISBN: 978-1-904455-28-8
Pages: xiv + 306 + colour plate, Hardcover

2. Leishmania
Edited by: Farrell, Jay P. (Ed.)
Publisher:Series: World Class Parasites , Vol. 4
Publication year:2002,
ISBN: 978-1-4020-7036-5
Pages: 208 p., Hardcover

TABLE 2.1 Some Major Sources For Leishmanial Information on Web

S.No	Institute/Organisation	TYPE OF INFORMATION
1	Sanger Institute (www.sanger.ac.uk/Projects/L_major/)	General information about Leishmania and leishmaniasis and genome information
2	Carlo Denegri Foundation (www.cdfound.to.it/html/lei1.htm)	General information about Visceral Leishmania and genome sequencing
3	World Health Organisation (www.who.int/tdr/svc/diseases/leishmaniasis)	General information about Leishmania
4	Genedb (www.genedb.org/genedb/leish/)	Genome information about Leishmaniasis
5	CDC(Division of Parasitc information) (www.cdc.gov/ncidod/dpd/parasites/leishmania/factsht_leishmania.htm)	General information about Leishmania

2.1.1 SANGER INSTITUTE: (www.sanger.ac.uk/Projects/L_major/)

The Leishmania major Friedlin Genome Project is being carried out in sanger institute at Cambridge. The *L. major* Friedlin genome is 32.8Mb in size, with a karyotype of 36 chromosomes. The G+C content is approximately 63%.

The Pathogen Genomics group at the Wellcome Trust Sanger Institute has played a major role in the genome sequencing of Leishmania major. The sequence data were obtained by adopting several parallel approaches, including complete cosmid sequencing, whole chromosome shotguns and/or BAC sequencing/skimming.

Project Status: Sequencing complete, 2 small gaps remaining.Latterly, the PSU has also undertaken whole genome shotguns, to ~5x coverage, of both *Leishmania infantum* and *Leishmania braziliensis.*

2.1.2 CARLO DENEGRI FOUNDATION: (www.cdfound.to.it/html/lei1.htm)

Visceral leishmaniasis has a wide geographic distribution. North-Eastern China, India, Middle-East, Southern Europe (Mediterranean basin), Northern Africa, Central-East Africa and, in foci, Central and South America (especially Brazil and Honduras).

2.1.3 GENEDB: (www.genedb.org/genedb/leish/):

The genome of *L. major* Friedlin, the reference strain (MHOM/IL/80/Friedlin, zymodeme MON-103), was sequenced as part of a multi-centre collaboration. The genome has been manually annotated and so far more than 8,000 genes have been identified in the ~33.6Mb haploid genome, which is spread over 36 chromosomes. The database is undergoing continual manual annotation and curation.

2.1.4 WORLD HEALTH ORGANISATION:

(www.who.int/tdr/svc/diseases/leishmaniasis)

World health organization carries out project on leishmaniasis, monitors its existence throughout the world. For a long time, little was known about the transmission cycles of the disease, but over the last few years, field research carried out by who and the application of molecular biology have enabled substantial progress to be made in understanding the different links in the transmission chain. Moreover, simple new diagnostic tests have recently been developed which are practical, reliable and inexpensive. These techniques are available to concerned countries for the early detection and rapid treatment of the disease. The World Health Organization has been collecting leishmaniasis data electronically on a yearly questionnaire through Leishnet.

2.1.5 DEPARTMENT OF DEFENSE (WALTER REED ARMY INSTITUTE OF RESEARCH):

The focus of this research is to develop a diagnostic device to detect Leishmaniasis infection. The Overall project objective is to develop FDA approved diagnostic devices to detect active and latent infection caused by Leishmania sp. Extend the scope of the program to the development of a systemically active agent or biological therapy efficacious against Leishmaniasis

2.1.6 DEPARTMENT OF DEFENSE (Wrair,Brazil):

The project aims to identify the Genetic Factors Which Control Tropism in Leishmania.This study was designed to identify the genes in Leishmania related to temperature sensitivity. The Overall project objective is to identify the gene(s) that control tropism in Leishmania and determine its (their) sequence and function.

2.2 MODEL ORGANISM DATABASES:

Databases specific for single organisms have emerged to facilitate understanding of biological processes and disease states.Some of the model organism databases are given below,

2.2.1 MOUSE GENOME DATABASES (NERVENET): (www.nervenet.org)

Mouse genome database contains information about

- current symbols, old symbols, names, and aliases
- three sets of data on the positions of loci
    1. from the November 1993 GBASE Locus Map
    2. from the 1993 and 1994 Chromosome Committee Reports
    3. data from the April 1995 MIT SSLP database release
- sequence accession numbers associated with gene loci
- locations of homologous genes in human and 10 other mammalian species
- recombinant inbred strain data
- strain distribution data of alleles
- probe and PCR primer data
- enzyme commission numbers
- phenotype codes
- references (in progress)

2.2.2 SACCHAROMYCES GENOME DATABASE(SGD): (www.yeastgenome.org)

SGD™ is a scientific database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae*, which is commonly known as baker's or budding yeast.

Objectives

3. OBJECTIVES:

The project has been carried out with the following specific objectives:

- Creation of model literature search database in access
- Physical Creation of the Relational Database on Leishmaniasis using Mysql and testing of the database.
- Incorporation of sequence analysis tools
- Creation of web pages for remote server to access Mysql database
- Online deployment of the database

Materials & Methods

4.MATERIALS AND METHODS:

4.1 HARDWARE:

The present study was carried out in a system with following specifications:

System Manufacturer	: Hewlett-Packard
System Model	: HP Pavilion dv5 Notebook PC
System Type	: X86-based PC
Processor	:AMD Turion(tm)X2Dual-Core,2 Logical Processor(s)
BIOS Version/Date(HP)	:F.11, 06-08-2008
SMBIOS Version	: 2.4
Installed Physical Memory	: 3.00 GB
Total Physical Memory	: 3.00 GB
Available Physical Memory	: 1.85 GB
Total Virtual Memory	: 6.22 GB
Available Virtual Memory	: 4.92 GB
Page File Space	: 3.29 GB

4.2 SOFTWARE:

4.2.1 OPERATING SYSTEM:

An operating system is an interface between hardware and user; it is responsible for the management and coordination of activities and the sharing of the limited resources of the computer. The operating system acts as a host for applications that are run on the machine. As a host, one of the purposes of an operating system is to handle the details of the operation of the hardware.

4.2.1.1 Linux : Fedora is an RPM-based, general purpose operating system built on top of the Linux kernel, developed by the community-supported Fedora Project and sponsored by Red Hat. Fedora's mission statement is: "Fedora is about the rapid progress of Free and Open Source software."

One of Fedora's main objectives is not only to contain free and open source software, but also to be on the leading edge of such technologies. Fedora developers prefer to make upstream changes instead of applying fixes specifically for Fedora—this ensures that updates are available to all Linux distributions. (fedora.org)

History: The Fedora Project was created in late 2003, when Red Hat Linux was discontinued. Red Hat Enterprise Linux was to be Red Hat's only officially supported Linux distribution, while Fedora was to be a community distribution. Red Hat Enterprise Linux branches its releases from versions of Fedora.
The name of Fedora derives from Fedora Linux, a volunteer project that provided extra software for the Red Hat Linux distribution, Fedora Linux was eventually absorbed into the Fedora Project. Fedora is a trademark of Red Hat.(redhat.com)

Table 4.1 Fedora Developmental History

Project Name	Version	Code name	Release date	Linux version
Fedora Core	1	Yarrow	2003-11-05	2.4.19
	2	Tettnang	2004-05-18	2.6.5
	3	Heidelberg	2004-11-08	2.6.9
	4	Stentz	2005-06-13	2.6.11
	5	Bordeaux	2006-03-20	2.6.15
	6	Zod	2006-10-24	2.6.18
Fedora	7	Moonshine	2007-05-31	2.6.21
	8	Werewolf	2007-11-08	2.6.23.1
	9	Sulphur	2008-05-13	2.6.25
	10	Cambridge*	2008-11-25	2.6.27
	11	Leonidas	Possibly 2009-05-26	TBA

* used for current study

4.2.1.2 Windows:

Microsoft Windows is a series of software operating systems and graphical user interfaces produced by Microsoft. Microsoft first introduced an operating environment named *Windows* in November 1985 as an add-on to MS-DOS in response to the growing interest in graphical user interfaces (GUIs).[1] Microsoft Windows came to dominate the world's personal computer market, overtaking Mac

OS, which had been introduced previously. At the 2004 IDC Directions conference, it was stated that Windows had approximately 90% of the client operating system market. The most recent client version of Windows is Windows Vista; the most recent server version is Windows Server 2008. Vista's successor, Windows 7 (currently in public beta) is slated to be released between July 1, 2009 and June 30, 2010.

History: Microsoft has taken two parallel routes in its operating systems. One route has been for the home user and the other has been for the professional IT user. The dual routes have generally led to home versions having greater multimedia support and less functionality in networking and security, and professional versions having inferior multimedia support and better networking and security. (windows.com)

TABLE 4.2 Windows Developmental History

Release date	Product name	Current Version / Build	Notes
November 1985	Windows 1.01	1.01	Unsupported
November 1987	Windows 2.03	2.03	Unsupported
March 1989	Windows 2.11	2.11	Unsupported
May 1990	Windows 3.0	3.0	Unsupported
March 1992	Windows 3.1x	3.1	Unsupported
October 1992	Windows For Workgroups 3.1	3.1	Unsupported
July 1993	Windows NT 3.1	NT 3.1	Unsupported
December 1993	Windows For Workgroups 3.11	3.11	Unsupported
January 1994	Windows 3.2 (released in Simplified Chinese only)	3.2	Unsupported
September 1994	Windows NT 3.5	NT 3.5	Unsupported
May 1995	Windows NT 3.51	NT 3.51	Unsupported
August 1995	Windows 95	4.0.950	Unsupported
July 1996	Windows NT 4.0	NT 4.0.1381	Unsupported
June 1998	Windows 98	4.10.1998	Unsupported
May 1999	Windows 98 SE	4.10.2222	Unsupported
February 2000	Windows 2000	NT 5.0.2195	Extended Support until July 13, 2010[22]

September 2000	Windows Me	4.90.3000	Unsupported
October 2001	Windows XP	NT 5.1.2600	Extended Support until April 8, 2014 for SP2 and SP3 (RTM and SP1 unsupported).
March 2003	Windows XP 64-bit Edition 2003	NT 5.2.3790	Unsupported
April 2003	Windows Server 2003	NT 5.2.3790	Current for SP1, R2, SP2 (RTM unsupported).
April 2005	Windows XP Professional x64 Edition	NT 5.2.3790	Current
July 2006	Windows Fundamentals for Legacy PCs	NT 5.1.2600	Current
November 2006 (volume licensing) January 2007 (retail)	Windows Vista*	NT 6.0.6001	Current. Version Changed to NT 6.0.6001 with SP1 (February 4, 2008)
July 2007	Windows Home Server	NT 5.2.4500	Current
February 2008	Windows Server 2008	NT 6.0.6001	Current
TBA	Windows 7	NT 6.1.7000	Beta release

* Used for current study

Windows Vista: Windows Vista is a line of operating systems developed by Microsoft for use on personal computers, including home and business desktops, laptops, Tablet PCs, and media center PCs. Prior to its announcement on July 22, 2005, Windows Vista was known by its codename "Longhorn."

Development was completed on November 8, 2006; over the following three months it was released in stages to computer hardware and software manufacturers, business customers, and retail channels. On January 30, 2007, it was released worldwide, and was made available for purchase and download from Microsoft's website. The release of Windows Vista came more than five years after the introduction of its predecessor, Windows XP, the longest time span between successive releases of Microsoft Windows desktop operating systems.

Windows Vista contains many changes and new features, including an updated graphical user interface and visual style dubbed Windows Aero, improved searching features, new multimedia creation tools such as Windows DVD Maker, and redesigned networking, audio, print, and display sub-systems. Microsoft's primary stated objective with Windows Vista, however, has been to improve the state of security in the Windows operating system.

4.2.2 RELATIONAL DATABASE MANAGEMENT SYSTEM:

A Relational database management system (RDBMS) is a database management system (DBMS) that is based on the relational model as introduced by E. F. Codd. Most popular commercial and open source databases currently in use are based on the relational model. A short definition of an RDBMS may be a DBMS in which data is stored in the form of tables and the relationship among the data is also stored in the form of tables.

4.2.2.1 Microsoft Access: Microsoft Office Access, previously known as Microsoft Access, is a relational database management system from Microsoft that combines the relational Microsoft Jet Database Engine with a graphical user interface and software development tools. It is a member of the Microsoft Office suite of applications and is included in the Professional and higher versions for Windows and also sold separately. Access stores data in its own format based on the Access Jet Database Engine. It can also import or link directly to data stored in other Access databases, Excel, SharePoint lists, text, XML, Outlook, HTML, dBase, Paradox, Lotus 1-2-3, or any ODBC-compliant data container including Microsoft SQL Server, Oracle, MySQL and PostgreSQL. Software developers and data architects can use it to develop application software and non-programmer "power users" can use it to build simple applications (msdn.com,2008)

Tables: A table is a collection of data about a specific topic, such as students or contacts. Using a separate table for each topic means that you store that data only once, which makes your database more efficient, and reduces data-entry errors.Tables organize data into columns (called fields) and rows (called records).

Primary Key : One or more fields (columns) whose value or values uniquely identify each record in a table. A primary key does not allow Null values and must always have a unique value. A primary key is used to relate a table to foreign keys in other table.

Relationship: After you've set up multiple tables in your Microsoft Access database, you need a way of telling Access how to bring that information back together again. The first step in this process is to define relationships between your tables. After you've done that, you can create queries, forms, and reports to display information from several tables .A relationship works by matching data in key fields - usually a field with the same name in both tables. In most cases, these matching fields are the primary key from one table, which provides a unique identifier for each record, and a foreign key in the other table.

Forms: A form is nothing more than a graphical representation of a table. You can add, update, delete records in your table by using a form. A form is very good to use when you have numerous fields in a table. This way you can see all the fields in one screen, whereas if you were in the table view (datasheet) you would have to keep scrolling to get the field you desire. (Mary Ann Richardson,2008)

4.2.2.2 Mysql: MySQL was developed by a consulting firm in Sweden called TcX in 1996. They created it because they needed a relational database that could handle large amounts of data on relatively cheap hardware. MySQL is the fastest relational database on the market. It outperforms all the leading databases in almost every category. It has almost all the functionality the leading databases have. MySQL is an open source, Enterprise-level, multi-threaded, relational database management system.

MySQL is a program that manages databases, much like Microsoft's Excel manages spreadsheets. SQL is a programming language that is used by MySQL to accomplish tasks within a database, just as Excel uses VBA (Visual Basic for Applications) to handle tasks with spreadsheets and workbooks.

Some of the basic Mysql operations are as follows:

Show databases;

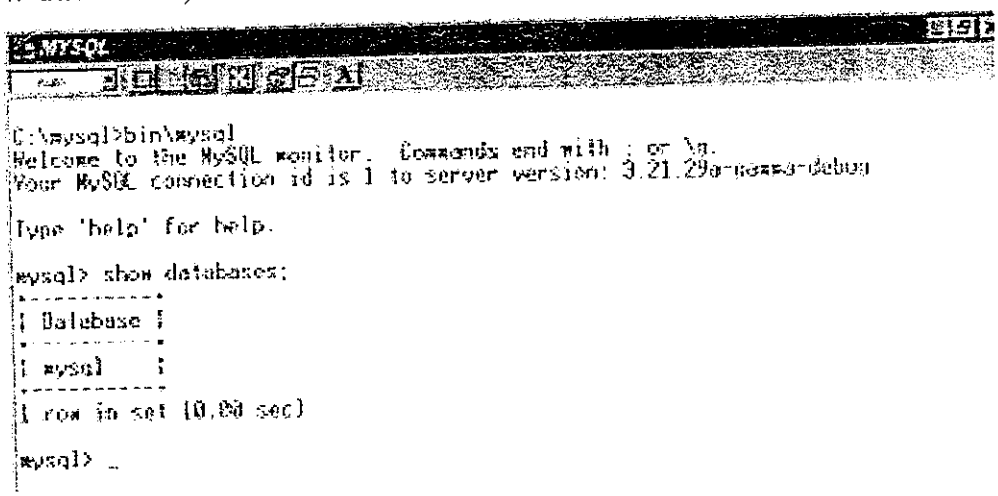


Fig 4.1 Mysql Console

To see the structure or schema of a database, the following command is issued:

Show tables from mysql;

To create a successful database, some thought has to be given to its design. A well-designed database will grow well. Retrieving and maintaining the information in a well designed database is a breeze.

Creating a Database: Use the CREATE DATABASE command to create a new database. To create a new database named newdb, issue the following command. The output will be as shown if the database is created successfully.

mysql> CREATE DATABASE newdb;

Query OK, 1 row affected (0.02 sec)

Dropping a Database: Use the DROP DATABASE command to drop a database completely.

mysql> DROP DATABASE newdb;

Query OK, 0 rows affected (0.00 sec)

Creating a New Table: Use the CREATE TABLE command to create a new database table. A table definition consists of a number of columns and a set of table options. Each column in the table definition is given a data type and can also be given a

constraint.A column name can be up to 64 characters long and can contain any character. Enclose the column name in quotes (") or backticks (`) if it contains a space. A column name cannot be one of the reserved keywords.

The INSERT Statement : The INSERT statement adds a new row of data to a table. At its simplest, INSERT is followed by a table name, the keyword VALUES, and a list of values in parentheses that correspond to each column in the table in turn. Several rows can be inserted in a single INSERT statement by supplying multiple lists of values. Each list must be enclosed in parentheses and separated by a comma

The DELETE Statement: The DELETE statement is used to remove data rows from a table. Its syntax is similar to a SELECT statement: You supply a table name after the keyword FROM and use a WHERE clause to filter the rows that are to be deleted. To delete only a single row from a table, ensure that the WHERE clause will match only that row. Check the value of the table's PRIMARY KEY column to ensure that an exact match is found.

The UPDATE Statement: The UPDATE statement is used to change one or some of the values in a data row. As with DELETE, include a WHERE clause to indicate that this row or rows are to be updated.

The REPLACE Statement: The REPLACE statement works just like INSERT, except that if a row already exists in the table with the same PRIMARY KEY value as the new data, the new row replaces it. Therefore, REPLACE never causes a PRIMARY KEY violation.

Loading Data from a File: Data from an external file can be loaded into MySQL using the LOAD DATA INFILE command. The data in the file must be in a structured format the default format is one record on each line, with values separated by tabs.If your data file is not tab-separated, must specify the separation method in the LOAD DATA INFILE statement using the TERMINATED BY and ENCLOSED BY keywords. (Mark Maslakowski.,2000)

4.2.3 WEB SERVER:

4.2.3.1 Apache Http Server: The Apache HTTP Server, commonly referred to simply as Apache, is a web server notable for playing a key role in the initial growth of the World Wide Web and in 2009 became the first web server to surpass the 100 million web site milestone. Apache was the first viable alternative to the Netscape Communications Corporation web server (currently known as Sun Java System Web Server), and has since evolved to rival other Unix-based web servers in terms of functionality and performance. The majority of all web servers using Apache are Linux web servers.

Apache is developed and maintained by an open community of developers under the auspices of the Apache Software Foundation. The application is available for a wide variety of operating systems, including Unix, GNU, FreeBSD, Linux, Solaris, Novell NetWare, Mac OS X, Microsoft Windows, OS/2, TPF, and eComStation. Released under the Apache License, Apache is characterized as free software and open source software.

4.2.4 PHP MAKER:

PHPMaker is a powerful automation tool that can generate a full set of PHP quickly from MySQL database. Using PHPMaker, you can instantly create Web sites that allow users to view, edit, search, add and delete records on the Web. PHPMaker is designed for high flexibility, numerous options enable you to generate PHP applications that best suits your needs. The generated codes are clean, straightforward and easy-to-customize. The PHP scripts can be run on both Windows or Linux/Unix servers. PHPMaker can save you tons of time and is suitable for both beginners and experienced developers alike.

4.2.5 DATABASE MANAGEMENT TOOLS:

4.2.5.1 WAMP Tool (Windows-Apache-Mysql-Php): WAMPs are packages of independently-created programs installed on computers that use a Microsoft Windows

operating system. The interaction of these programs enables dynamic web pages to be served over a computer network, such as the internet or a private network.

"WAMP" is an acronym formed from the initials of the operating system (Windows) and the package's principal components: Apache, MySQL and PHP (or Perl or Python). Apache is a web server, which allows people with web browsers like Internet Explorer or Firefox to connect to a computer and see information there as web pages. MySQL is a database manager (that is, it keeps track of data in a highly organized way). PHP is a scripting language which can manipulate information held in a database and generate web pages afresh each time an element of content is requested from a browser. Other programs may also be included in a package, such as phpMyAdmin which provides a graphical interface for the MySQL database manager, or the alternative scripting languages Python or Perl.(Wikipedia.org)

4.2.5.2 LAMP TOOL (Linux-Apache-Mysql-Php): The acronym LAMP 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:

- Linux, referring to the operating system;
- Apache, the Web server;
- MySQL, the database management system (or database server);
- one of several scripting languages: Perl, PHP or Python.

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.

Though the originators of these open source programs did not design them all to work specifically with each other, the combination has become popular because of its low acquisition cost and because of the ubiquity of its components (which come bundled with most current Linux distributions). When used in combination they represent a solution stack of technologies that support application servers.(_Dale Dougherty.,2001)

4.2.5.3 PhpMyAdmin: PhpMyAdmin is a free software tool written in PHP intended to handle the administration of MySQL over the World Wide Web. phpMyAdmin supports a wide range of operations with MySQL. The most frequently used operations are supported by the user interface (managing databases, tables, fields, relations, indexes, users, permissions, etc), while you still have the ability to directly execute any SQL statement. PhpMyAdmin has won several awards. Among others, it was chosen as the best PHP application in various awards and every year wins the SourceForge.net Community Choice Awards as "Best Tool or Utility for SysAdmins". phpMyAdmin is a more than ten years old project with stable and flexible code base.Advantages of PhpMyAdmin includes:

- Intuitive web interface
- Support for most MySQL features:
    - browse and drop databases, tables, views, fields and indexes
    - create, copy, drop, rename and alter databases, tables, fields and indexes
    - maintenance server, databases and tables, with proposals on server configuration
    - execute, edit and bookmark any SQL-statement, even batch-queries
    - manage MySQL users and privileges
    - manage stored procedures and triggers
- Import data from CSV and SQL
- Export data to various formats: CSV, SQL, XML, PDF, ISO/IEC 26300 - Open Document Text and Spreadsheet, Word, Excel, and others
- Administering multiple servers
- Creating PDF graphics of your database layout
- Creating complex queries using Query-by-example (QBE)
- Searching globally in a database or a subset of it
- Transforming stored data into any format using a set of predefined functions

4.2.6 SEQUENCE ANALYSIS TOOL:

The term "sequence analysis" in biology implies subjecting a DNA or peptide sequence to sequence alignment, sequence databases, repeated sequence searches, or other bioinformatics methods on a computer.

4.2.6.1 Basic Local Alignment Search Tool (BLAST): Basic Local Alignment Search Tool, or BLAST, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

The BLAST program was designed by Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman and Webb Miller at the NIH and was published in J. Mol. Biol. in 1990.(Altschul *et al.*,1990)

BLAST ALGORITHM:

An overview of the BLASTP algorithm (a protein to protein search) is as follows:

1. Remove low-complexity region or sequence repeats in the query sequence. Low-complexity region means a region of a sequence is composed of few kinds of elements. These regions might give high scores that confuse the program to find the actual significant sequences in the database, so they should be filtered out. The regions will be marked with an X (protein sequences) or N (nucleic acid sequences) and then be ignored by the BLAST program. To filter out the low-complexity regions, the SEG program is used for protein sequences and the program DUST is used for DNA sequences. On the other hand, the program XNU is used to mask off the tandem repeats in protein sequences.
2. Make a k-letter word list of the query sequence. Take k=3 for example, we list the words of length 3 in the query protein sequence (k is usually 11 for a DNA sequence) "sequentially", until the last letter of the query sequence is included. The method can be illustrated in figure

Query sequence: PQGEFG

Word 1: PQG

Word 2: QGE

FIG 4.2 The Method To Establish The K-Letter Query Word List During Blast Query

3. List the possible matching words. This step is one of the main differences between BLAST and FASTA. FASTA cares about all of the common words in the database and query sequences that are listed in step 2; however, BLAST cares about only the high-scoring words. The scores are created by comparing the word in the list in step 2 with all the 3-letter words. By using the scoring matrix (substitution matrix) to score the comparison of each residue pair, there are 20^3 possible match scores for a 3-letter word. For example, the score obtained by comparing PQG with PEG and PQA is 15 and 12, respectively. For DNA words, a match is scored as +5 and a mismatch as -4. After that, a neighborhood word score threshold T is used to reduce the number of possible matching words. The words whose scores are greater than the threshold T will remain in the possible matching words list,

   while those with lower scores will be discarded. For example, PEG is kept, but PQA is abandoned when T is 13.

4. Organize the remaining high-scoring words into an efficient search tree. This is for the purpose that the program can rapidly compare the high-scoring words to the database sequences.
5. Repeat step 1 to 4 for each 3-letter word in the query sequence.
6. Scan the database sequences for exact match with the remaining high-scoring words.

The BLAST program scans the database sequences for the remaining high-scoring word, such as PEG, of each position. If an exact match is found, this match is used to seed a possible ungapped alignment between the query and database sequences.

7. Extend the exact matches to high-scoring segment pair (HSP).

The original version of BLAST stretches a longer alignment between the query and the database sequence in left and right direction, from the position where exact match is scanned. The extension doesn't stop until the accumulated total score of the HSP begins to decrease. A simplified example is presented in figure

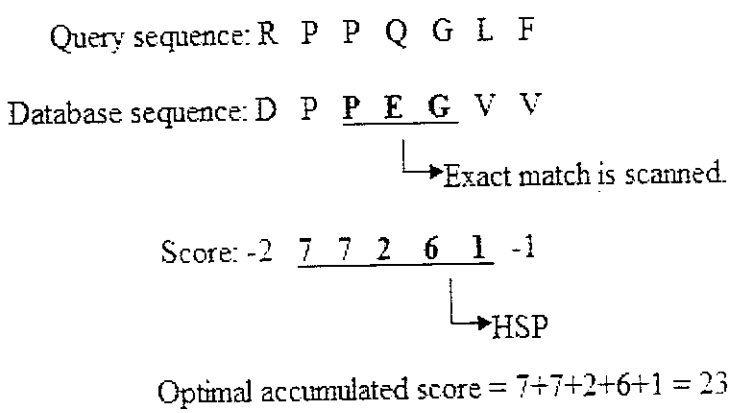


FIG 4.3 The Process To Extend The Exact Match During Blast Query

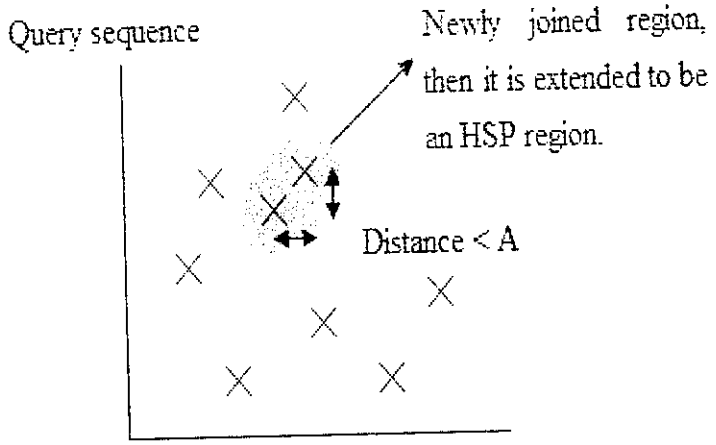


FIG 4.4 The Position Of Exact Matches During Blast Query

8. List all of the HSPs in the database whose score is high enough to be considered.

   We list the HSPs whose scores are greater than the empirically determined cutoff score S. By examining the distribution of the alignment scores modeled by comparing random sequences, a cutoff score S can be determined such that its value is large enough to guarantee the significance of the remained HSPs.

9. Evaluate the significance of the HSP score.

   BLAST next assesses the statistical significance of each HSP score by exploiting the Gumbel extreme value distribution (EVD). (It is proved that the distribution of Smith-Waterman local alignment scores between two random sequences follows the Gumbel EVD, regardless of whether gaps are allowed in the alignment). In accordance with the Gumbel EVD, the probability p of observing a score S equal to or greater than x is given by the equation

$$p(S \geq x) = 1 - \exp\left(-e^{-\lambda(x-\mu)}\right)$$

   Where,

$$\mu = {[\log(Km'n')]}/{\lambda}$$

The statistical parameters λ and K are estimated by fitting the distribution of the ungapped local alignment scores, of the query sequence and a lot of shuffled versions (Global or local shuffling) of a database sequence, to the Gumbel extreme value distribution. Note that λ and K depend upon the substitution matrix, gap penalties, and sequence composition (the letter frequencies).The m' and n' is the effective length of the query and database sequence, respectively. The original sequence length is shortened to the effective length to compensate for the edge effect (an alignment start near the end of one of the query or database sequence is likely not to have enough sequence to build an optimal alignment). They can be calculated as

$$n' \approx n - {(\ln Kmn)}/{H}$$

$$m' \approx m - {(\ln Kmn)}/{H}$$

where H is the average expected score per aligned pair of residues in an alignment of two random sequences. Altschul and Gish gave the typical values, λ = 0.318, K = 0.13, and H = 0.40, for ungapped local alignment using BLOSUM62 as the substitution matrix. Using the typical values for assessing the significance is called the lookup table methods, and is not accurate. The expect score E of a database match is the number of times that an unrelated database sequence would obtain a score S higher than x by chance. The expectation E obtained in a search for a database of D sequences is given by

$$E \approx 1 - e^{-p(s>x)D}$$

Furthermore, when p < 0.1, E could be approximated by the Poisson distribution as

$$E \approx pD$$

Note that the E value accessing the significance of the HSP score here (for ungapped local alignment) is not identical to the one in the later step to

evaluate the final gapped local alignment score, due to the variation of the statistical parameters.

10. Make two or more HSP regions into a longer alignment. Sometimes, we find two or more HSP regions in one database sequence that can be made into a longer alignment. This provides additional evidence of the relation between the query and database sequence. There are two methods, the Poisson method and the sum-of scores method, to compare the significance of the newly combined HSP regions. Suppose that there are two combined HSP regions with the sets of score (65, 40) and (52, 45), respectively. The Poisson method gives more significance to the set with the lower score of each set is higher (45>40). However, the sum-of-scores method prefers the first set, because 65+40 (105) is greater than 52+45(97). The original BLAST uses the Poisson method; gapped BLAST and the WU-BLAST use the sum-of scores method.
11. Show the gapped Smith-Waterman local alignments of the query and each of the matched database sequences.
    - The original BLAST only generates ungapped alignments including the initially found HSPs individually, even when there is more than one HSP found in one database sequence.
    - BLAST2 versions produce a single alignment with gaps that can include all of the initially found HSP regions. Note that the computation of the score and its corresponding E score is involved with the adequate gap penalties.
12. Report the matches whose expect score is lower than a threshold parameter E.(Altschul *et al.*,1990)

BLAST REQUIREMENTS:

- Requires an active internet connection to visit websites where molecular databases reside (e.g. http://www.ncbi.nlm.nih.gov)- have a lot of flexibility

working over the web (many different databases and informatics tools can be rapidly accessed)

- specify a target database to be searched using the website's BLAST server
- upload the query sequences (these are the sequences you want to learn more about) to a web-BLAST server; then these sequences are compared by the BLAST alignment algorithm to all sequences in the specified target database

BLAST Session Setup:

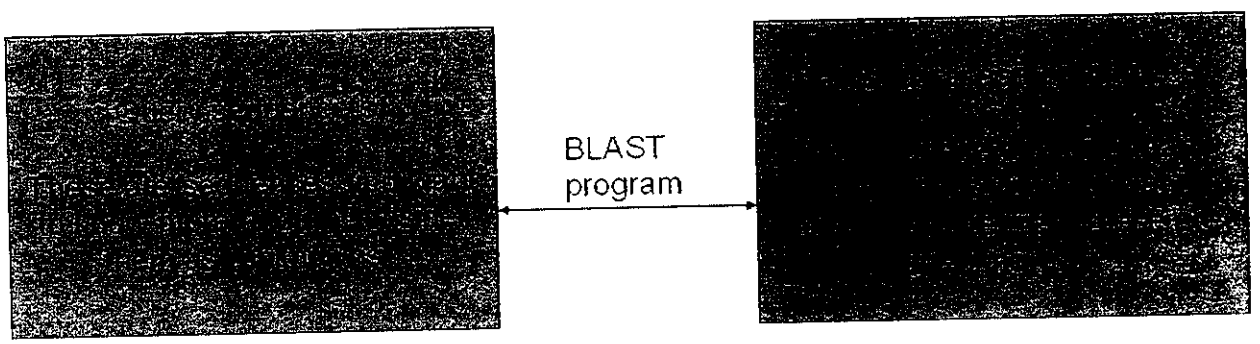


Fig 4.5 Blast Session

If BLAST detects a match between query sequences and database sequences, this indicates some meaningful relationship between the aligned sequences.(Doug Davis,2006)

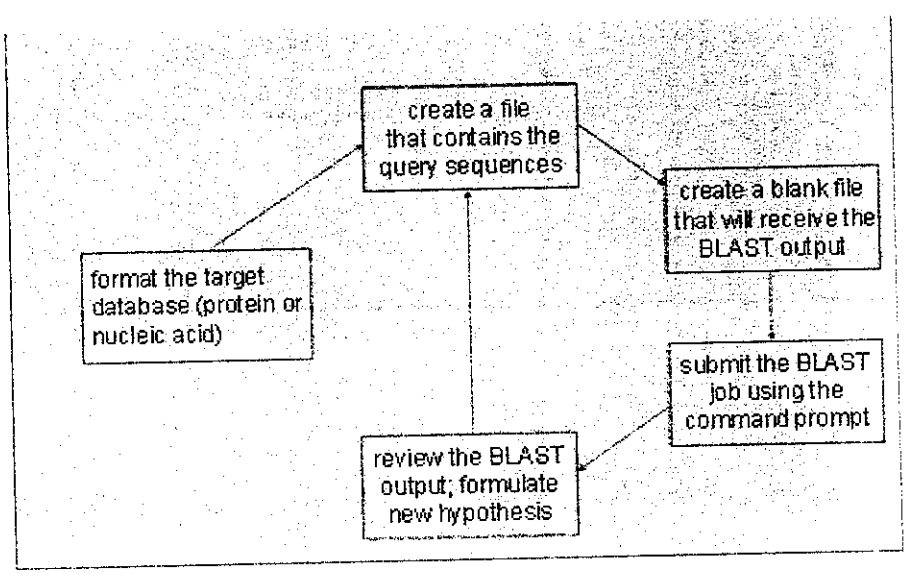


Fig 4.6 Flow Of Events In Blast

4.2.6.2 Standalone Www Blast Server: Standalone WWW BLAST server suite of programs was designed similar to the regular NCBI BLAST server and such command-line NCBI BLAST programs like "blastall", "blastpgp", "rpsblast" and "megablast". It incorporates most features, which exist in NCBI BLAST programs and should be relatively easy to use. This server does not support any request queuing and load balancing. As soon as the user hits a "Search" button, BLAST starts immediately if entered information is valid. So, this server is not intended to handle large load, which may exist in public service. Such queueing and loadbalancing however may be implemented using such products as Load Sharing Facility - "LSF" from Platform Computing Corporation. Interface to "LSF" was implemented in NCBI, however this was not included in this suite. Standalone server assumes that users have their own BLAST or RPS-BLAST database(s), that should be searched and want to have a simple WWW interface to such search. (Thomas Madden *et al.*,)

Configuration Of Blast Databases: To set up databases for the standalone WWW BLAST server, it is necessary to follow these steps:

1. Put a file with concatenated FASTA entries in the "./db" directory
2. Run "formatdb" program, available from the NCBI ftp site to format the database.
3. Add name of the database into server configuration file
4. Add name of the database into (PSI/PHI) WWW BLAST search form

Description Of Tags For The Main Blast Input Page: This standalone server has a tag convention analogous to the regular NCBI BLAST server. Sample BLAST search forms may be changed to accommodate particular needs of the user in the custom search. Here is the list of these tags and their meaning. If some tag is missing from the search input page, it will take a default value. Exceptions are tags PROGRAM, DATALIB and SEQUENCE (or SEQFILE), that should always be set.

- PROGRAM - name of the BLAST program. Supported values include programs: blastn, blastp, blastx, tblastx and tblastn

- DATALIB - name of the database(s) to search. This implementation includes possibility to use multiple databases. To use multiple databases several "DATALIB" tags should be used on the page for example using checkboxes (look for example at Microbial Genomes Blast Databases BLAST at NCBI). Note, that all of these databases should be properly written in the server configuration file.
- SEQUENCE and SEQFILE - these tags used to pass sequence. First SEQUENCE tag is used for the input sequence. If it is missing, the SEQFILE tag is used instead.
- UNGAPPED_ALIGNMENT - default BLAST search is a gapped search; this tag, if set, will turn gapped alignment off.
- MAT_PARAM used to set 3 parameters at the same time. Value for this tag should be in format " " where mat_name - string name of the matrix (BLOSUM62, etc), d1 - integer for cost to open gap and d2 - cost to extend gap (-G and -E parameters in blastall respectably)
- GAP_OPEN - set value for cost to open gap - 0 or missing tag invoked default behavior
- GAP_EXTEND - set value for cost to extend gap - 0 or missing tag invoked default behavior
- X_DROPOFF - Dropoff (X) for blast extensions in bits (default if zero) (-y parameter in "blastpgp" program)
- GENETIC_CODE - Query Genetic code to use (for blastx only)
- THRESHOLD_2 - Threshold for extending hits in second pass in multipass model search
- MATRIX - Matrix (default is BLOSUM62) (-M in blastall) EXPECT - Expectation value (-e in blastall)
- NUM_OF_BITS - Number of bits to trigger gapping (-N in blastpgp)
- NCBI_GI - If formatted database use SeqIds in the NCBI format this option will turn printing of gis together with accessions.
- FILTER - Multiple instances of values of this tag are concatenated and passed to the engine as "filter string" ("L" for low complexity and "m" if filter should

be set for lookup table only) - any letter will turn default filtering on - DUST for nucleotides and SEG for proteins (-F in blastall)

- DESCRIPTIONS - Number of one-line descriptions in the output (-v in blastall)
- ALIGNMENTS - Number of alignments to show (-b in blastall)

Functionality Offered By Blast Applications: The functionality offered by the BLAST+ applications has been organized by program type, as to more closely resemble Web BLAST. The following graph depicts a correspondence between the NCBI C Toolkit BLAST command line applications and the BLAST applications:

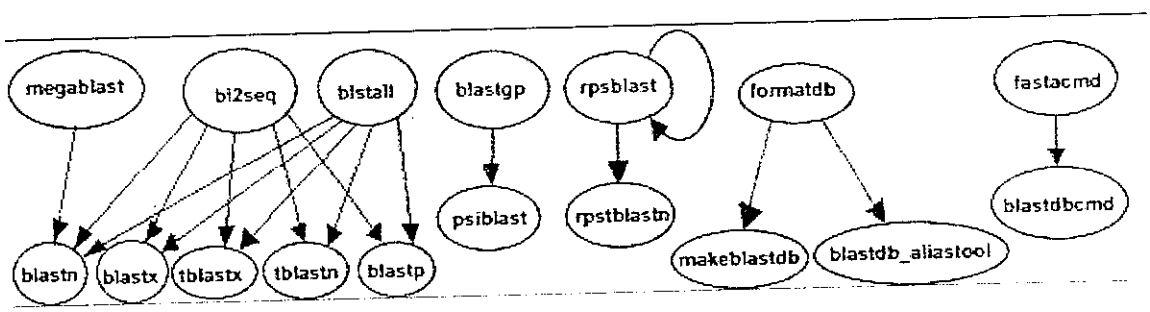


FIG 4.7 Various Blast Applications And Their Inter Relationship

As an example, to run a search of a nucleotide query (translated "on the fly" by BLAST) against a protein database one would use the blastx application instead of blastall. The blastx application will also work in "Blast2Sequences" mode (i.e.: accept FASTA sequences instead of a BLAST database as targets) and can also send BLAST searches over the network to the public NCBI server if desire. (Christiam Camacho et al.,)

Here's how the BLAST session looks in "Command Prompt" (this is the program you will use in Windows to run BLAST):

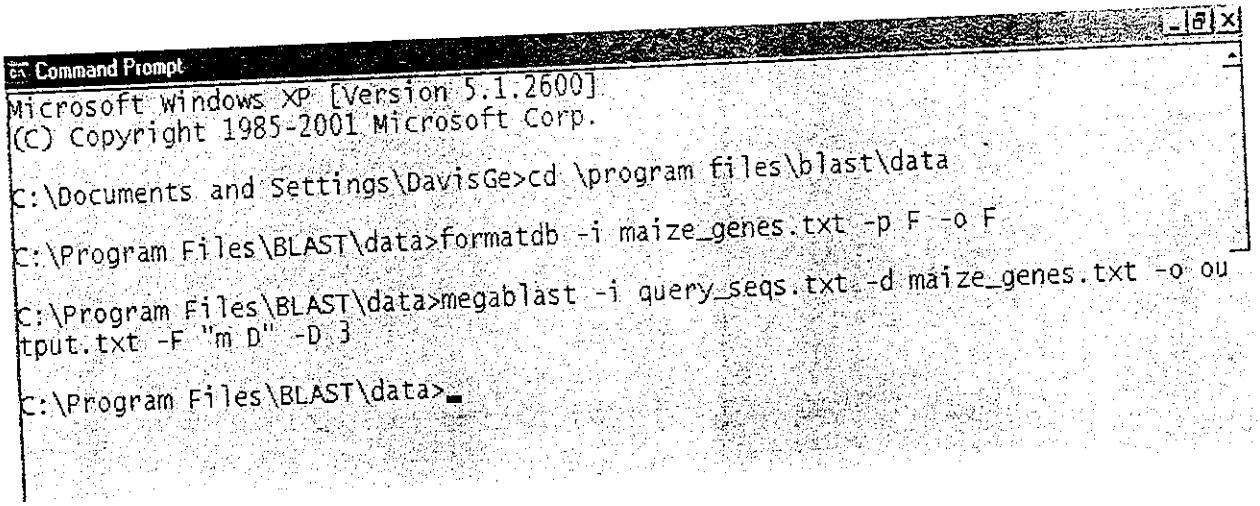


Fig 4.8 Blast In Command Line

4.2.7 LITERATURE SEARCH ENGINES:

A search engine is an information retrieval system designed to help find information stored on a computer system. The search results are usually presented in a list and are commonly called hits. Search engines help to minimize the time required to find information and the amount of information which must be consulted, akin to other techniques for managing information overload. A literature search is a detailed and organized, step by step search for all the material available on a topic.

4.2.7.1 Pubsearch: PubSearch is a literature curation management system designed to store and manage the available literature for an organism or system of interest. It provides database curators with a powerful literature search capability, stores relevant biological information, creates automatic associations between the biological information and the literature, and provides a user-friendly web interface for manual validation and curation. PubSearch is based on a simple MySQL relational database for the back-end, and Java Servlet and Java Server Pages for the API and front-end applications. PubSearch is extensible, allowing new biological objects of interest to be added, and flexible, allowing programming of different curation strategies.

It is currently used for curating biological information for the model organism database for Arabidopsis thaliana, TAIR. The system can be easily adapted for

extracting and curating information from the literature for any model organism database.

PubSearch dataflow and design overview:

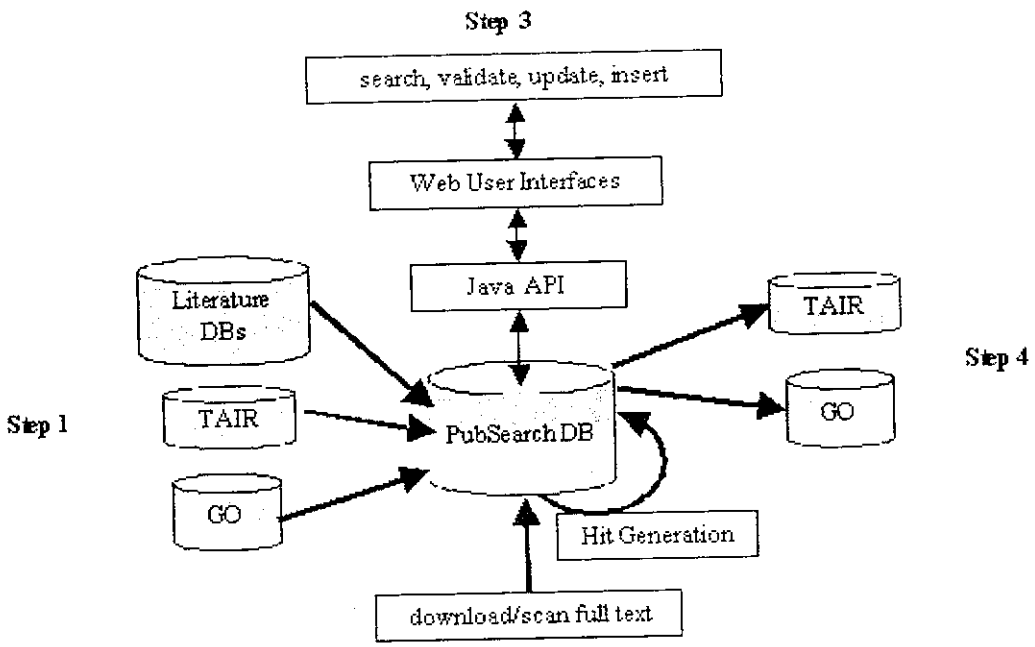


Fig 4.9 Pubsearch Dataflow

Step 1: The PubSearch database is loaded in batch mode using input from other databases such as literature databases (PubMed, Agricola, etc), TAIR (an example model organism database), and the Gene Ontology database.

Step 2: The PubSearch database The PubSearch database is the central component of the PubSearch system. The following operations are performed during PubSearch use: indexes the information by populating the Hit table.

Step 3: Through the Java API and the web-interface, curators search, update, validate and insert information, relying on the indexed data in the database.

Step 4: Finally, the data is exported to the TAIR production database and other databases such as the Gene Ontology database. (Yoo D et al.,2006)

4.2.7.2 TEXTPRESSO: Textpresso is a text-mining system for scientific literature. Textpresso's two major elements are (1) access to full text, so that entire articles can be searched, and (2) introduction of categories of biological concepts and classes that relate two objects (e.g., association, regulation, etc.) or describe one (e.g., methods, etc). A search engine enables the user to search for one or a combination of these categories and/or keywords within an entire literature

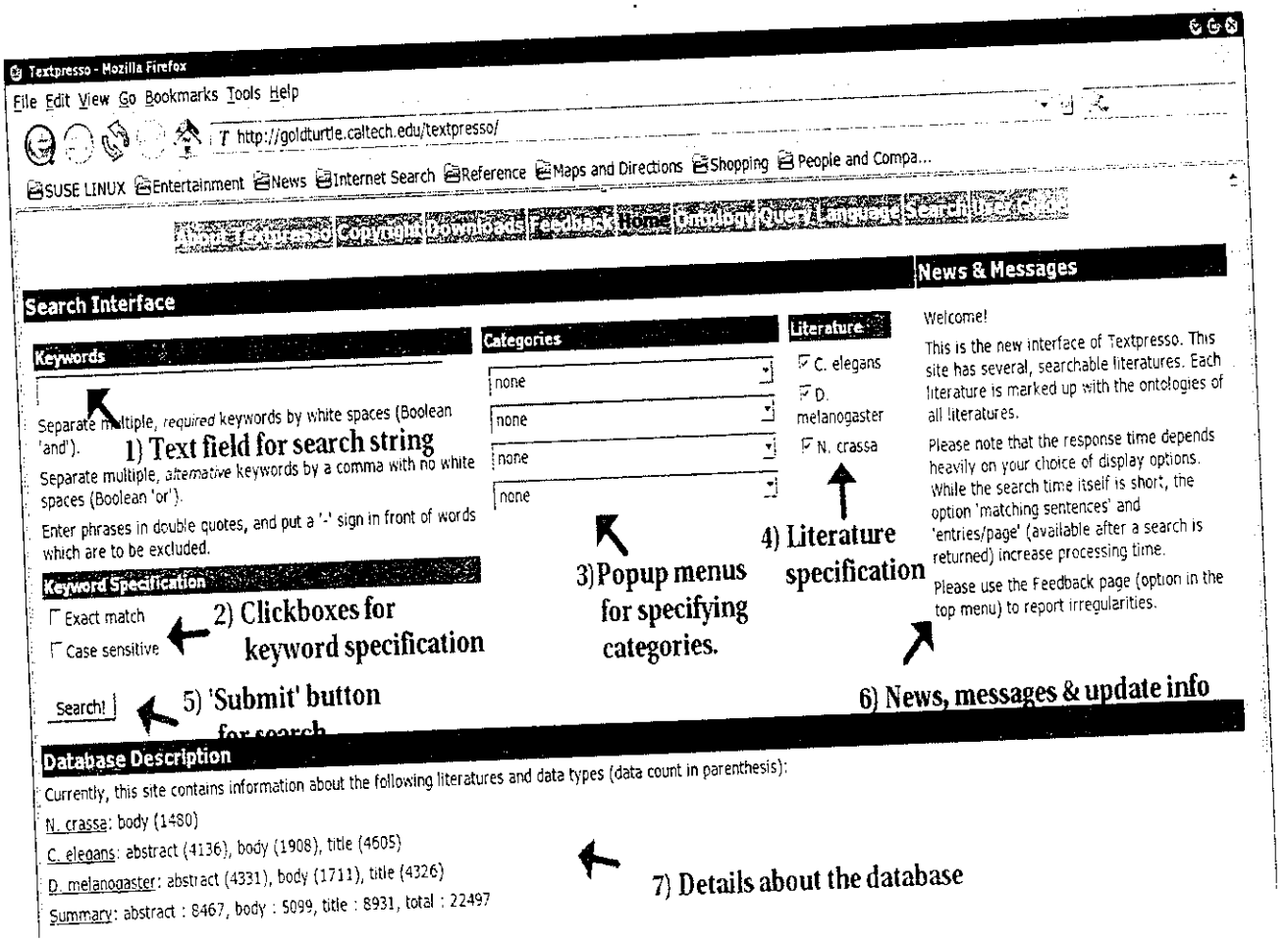


FIG 4.10 An Example Of Textpresso Search

Textpresso is useful as a search engine for researchers as well as a curation tool. It was developed as a part of WormBase and is used extensively by *C. elegans* curators. Textpresso has currently been implemented for 17 different literatures, and can readily be extended to other corpora of text. (Hans-Michael Muller., 2007.)

4.3 ACCESSING MYSQL DATABASE FROM THE WEB WITH PHP:

A user's web browser issues an HTTP request for a particular web page. The search results page is called results.php. Here are the steps involved:

1. The web server receives the request for `results.php`, retrieves the file, and passes it to the PHP engine for processing.
2. The PHP engine begins parsing the script. Inside the script is a command to connect to the database and execute a query .PHP opens a connection to the MySQL server and sends on the appropriate query.
3. The MySQL server receives the database query, processes it, and sends the results back to the PHP engine.
4. The PHP engine finishes running the script. This usually involves formatting the query results nicely in HTML. It then returns the resulting HTML to the web server.
5. The web server passes the HTML back to the browser, where the user can see the data.

Begin with the search form. The code for this plain HTML form is shown.

Database Search Page:

```
1 <html>
2 <head>
3 <title>Leishmania literature  Search</title>
4 </head>
5 <body>
6 <h1> Leishmania literature Search</h1>
7 <form action="results.php" method="post">
8 Choose Search Type:<br />
9 <select name="searchtype">
10<option value="author">Author</option>
11<option value="title">Title</option>
```

```
12<option value="abstract">Abstract</option>

13</select>

14<br />

15Enter Search Term:<br />

16<input name="searchterm" type=""text" size="40"/>

17<br />

18<input type="submit" name="submit" value="Search"/>

19</form>  20</body>  21</html>
```

This HTML form is reasonably straightforward.

The script that will be called when the Search button is clicked is `results.php`.

This Script Retrieves Search Results from the MySQL Database and Formats Them for Display

```
1 <html>
2 <head>
3 <title> Leishmania literature Search Results</title>
4 </head>
5 <body>
6 <h1> Leishmania literature Search Results</h1>
7 <?php
8 // create short variable names
9 $searchtype=$_POST['searchtype'];
10$searchterm=trim($_POST['searchterm']);
11if (!$searchtype || !$searchterm) {
12echo 'You have not entered search details. Please go back and try again.';
13exit;
14}
```

```
15if (!get_magic_quotes_gpc()){
16$searchtype = addslashes($searchtype);
17$searchterm = addslashes($searchterm);
18}
  @ $db = new mysqli('localhost', ' Leishmania literature ', ' Leishmania
19literature123 ', 'table');
20if (mysqli_connect_errno()) {
21echo 'Error: Could not connect to database. Please try again later.';
22exit;
23}
  $query = "select *  from table
24where ".$searchtype." like '%".$searchterm."%'";
25$result = $db->query($query);
26$num_results = $result->num_rows;
27echo "<p>Number of records found: ".$num_results."</p>";
28for ($i=0; $i <$num_results; $i++) {
29$row = $result->fetch_assoc();
30echo "<p><strong>".($i+1).". Title: ";
31echo htmlspecialchars(stripslashes($row['title']));
32echo "</strong><br />Author: ";
33echo stripslashes($row['author']);
34echo "<br />ABSTRACT: ";
35echo stripslashes($row['abstract']);
36echo "<br />ADDRESS: ";
37echo stripslashes($row['address']);
38echo "</p>";
39}
40$result->free();
41$db->close();
42?>    43</body> 44</html>
```

Note that this script allows you to enter the MySQL wildcard characters % and _ (underscore). This capability can be useful for the user,but you can escape these characters if they will cause a problem for your application.(Luke Welling et al.,)

4.4 CODD'S RULES FOR RELATIONAL DATABASES:

Codd's 12 rules are a set of thirteen rules (numbered zero to twelve) proposed by Edgar F. Codd, a pioneer of the relational model for databases, designed to define what is required from a database management system in order for it to be considered relational, i.e., an RDBMS

Codd produced these rules as part of a personal campaign to prevent his vision of the relational database being diluted, as database vendors scrambled in the early 1980s to repackage existing products with a relational veneer. Rule 12 was particularly designed to counter such a positioning. In fact, the rules are so strict that all popular so-called "relational" DBMSs fail on many of the criteria.

The rules are as follows:

Rule 0: The system must qualify as relational, as a database, and as a management system.For a system to qualify as a relational database management system (RDBMS), that system must use its relational facilities (exclusively) to manage the database.

Rule 1: The information rule:

> All information in the database is to be represented in one and only one way, namely by values in column positions within rows of tables.

Rule 2: The guaranteed access rule:

> All data must be accessible. This rule is essentially a restatement of the fundamental requirement for primary keys. It says that every individual scalar value in the database must be logically addressable by specifying the name of the containing table, the name of the containing column and the primary key value of the containing row.

Rule 3: Systematic treatment of null values:

The DBMS must allow each field to remain null (or empty). Specifically, it must support a representation of "missing information and inapplicable information" that is systematic, distinct from all regular values (for example, "distinct from zero or any other number", in the case of numeric values), and independent of data type. It is also implied that such representations must be manipulated by the DBMS in a systematic way.

Rule 4: Active online catalog based on the relational model:

The system must support an online, inline, relational catalog that is accessible to authorized users by means of their regular query language. That is, users must be able to access the database's structure (catalog) using the same query language that they use to access the database's data.

Rule 5: The comprehensive data sublanguage rule:

The system must support at least one relational language that

1. Has a linear syntax
2. Can be used both interactively and within application programs,
3. Supports data definition operations (including view definitions), data manipulation operations (update as well as retrieval), security and integrity constraints, and transaction management operations (begin, commit, and rollback).

Rule 6: The view updating rule:

All views that are theoretically updatable must be updatable by the system.

Rule 7: High-level insert, update, and delete:

The system must support set-at-a-time insert, update, and delete operators. This means that data can be retrieved from a relational database in sets constructed of data from multiple rows and/or multiple tables. This rule states that insert,

If the system provides a low-level (record-at-a-time) interface, then that interface cannot be used to subvert the system, for example, bypassing a relational security or integrity constraint.(Codd, Edgar Frank,1985).

The Leishmanial database was organized in a manner that the data was normalized & follows the Codd's rules for relational databases.

Results & Discussions

5. RESULTS AND DISCUSSION:

5.1 COLLECTION OF DATA:

5.1.1 Literature data:

Scientific literature comprises scientific publications that report original empirical and theoretical work in the natural and social sciences, and within a scientific field is often abbreviated as the literature. Academic publishing is the process of placing the results of one's research into the literature. Scientific research on original work initially published in scientific journals is called primary literature. Patents and technical reports, for minor research results and engineering and design work (including computer software) can also be considered primary literature. Secondary sources include articles in review journals (which provide a synthesis of research articles on a topic to highlight advances and new lines of research), and books for large projects, broad arguments, or compilations of articles.

Sources: The major sources for literature data in biological sciences include Pubmed, Science direct,Agricola. The pubmed stores medline records that are well structured.

5.1.1.1 PubMed (www.ncbi.nlm.nih.gov): PubMed Central (PMC) is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature. The PMC journal list comprises journals that deposit material in PMC on a routine basis and generally make all their published articles available here. PMC also has the author manuscripts of articles published by NIH-funded researchers in various non-PMC journals. Increasing free access to these articles is the goal of the NIH Public Access policy. Similar manuscripts from researchers funded by the Wellcome Trust are available in PMC as well. PMC's utilities include an OAI service that provides XML of the full-text of some articles, functions for scripting PMC searches. It's about preservation and access: digitizing the complete run of back issues of many of the journals in PMC

5.1.1.2 Agricola: AGRICOLA offers vital agricultural information from 1970 to the present. It contains over 3.7 million citations to journal articles, monographs, theses, patents, software, audio-visual materials, and technical reports related to agriculture. These records describe publications and resources encompassing all aspects of agriculture and allied disciplines, including animal and veterinary sciences, entomology, plant sciences, forestry, aquaculture and fisheries, farming and farming systems, agricultural economics and earth and environmental sciences. AGRICOLA serves as the document locator and bibliographic control system for the NAL collection. The extensive file provides comprehensive coverage of newly acquired worldwide publications in agriculture and related fields. in two year spreads; 1970-Present and 1984-Present.

Literature data collected from the database:

TABLE 5.1 LITERATURE DATA

S.No	SITE	NUMBER
1	PubMed	14550
2	Agricola	2765

5.1.2 Sequence data:

TABLE 5.2 SEQUENCE DATA

S.No	Type	NUMBER
1	EST	26794
2	Nucleotide sequence	88033
3	Protein sequence	56526
4	Gene sequence	26956

5.2 Curation of literature data:

Abstarcts collected from Pubmed & Agricola were semi automatically curated to extract relevant fields for incorporation into database. Data collected was parsed into different categories like author, journal name, published date, abstract, keywords and title respectively. Medline Abstract data format:

In Medline abstracts every discerning details about the articles are separated according to various categories. Example of medline abstract format is given below:

*PMID- 19120250
OWN - NLM
STAT- MEDLINE
DA - 20090105
DCOM- 20090128
IS - 1749-6632 (Electronic)
VI - 1149
DP - 2008 Dec
TI - Serologic and molecular evaluation of Leishmania infantum in cats from Central Spain.
PG - 361-4
*AB - Infection by different Leishmania spp. in cats has been reported in many countries. In Spain, since the first Leishmania infection described in 1933, sporadic clinical cases in cats have been reported. Various serologic studies performed in other areas of Spain have shown seroprevalences ranging between 1.7 and 60%. The aim of the present study was to determine the prevalence of leishmaniasis in cats from Central Spain (Madrid), and to assess the existence of associations between Leishmania infantum infection and relevant data obtained from each cat. Two-hundred thirty-three cats attended at the Veterinary Teaching Hospital in Madrid between September 2005 and June 2006 were tested for L. infantum using the indirect immunofluorescent antibody (IFA) test (cutoff: 1:100) and PCR. PCR testing was performed on the samples to detect Leishmania infection, targeting the kinetoplast DNA (kDNA). Our results showed a seroprevalence of 1.29% (3/233) using IFA test. Another seven cats were also seroreactive to L. infantum one dilution under the cutoff (1:50). Considering all the seroreactive samples, the percentage of positive animals to L. infantum was 4.29%. Only one of the cats (0.43%) included in the study was PCR-positive. Relative lymphocytosis and an increase in alanine aminotransferase (ALT) value were statistically associated with seroreactivity to L. infantum. Our results demonstrate the presence of cats seroreactive to L. infantum in Central Spain, an endemic area for this disease in dogs.
*AD - College of Veterinary Medicine, Complutense University of Madrid, Madrid, Spain.
*FAU - Ayllon, Tania
*AU - Ayllon T
*FAU - Tesouro, Miguel A
*AU - Tesouro MA
*FAU - Amusategui, Inmaculada
*AU - Amusategui I
*FAU - Villaescusa, Alejandra
*PT - Journal Article
*PL - United States
*TA - Ann N Y Acad Sci
*AID - 10.1196/annals.1428.019 [doi]
*SO - Ann N Y Acad Sci. 2008 Dec;1149:361-4.

* The fields extracted.

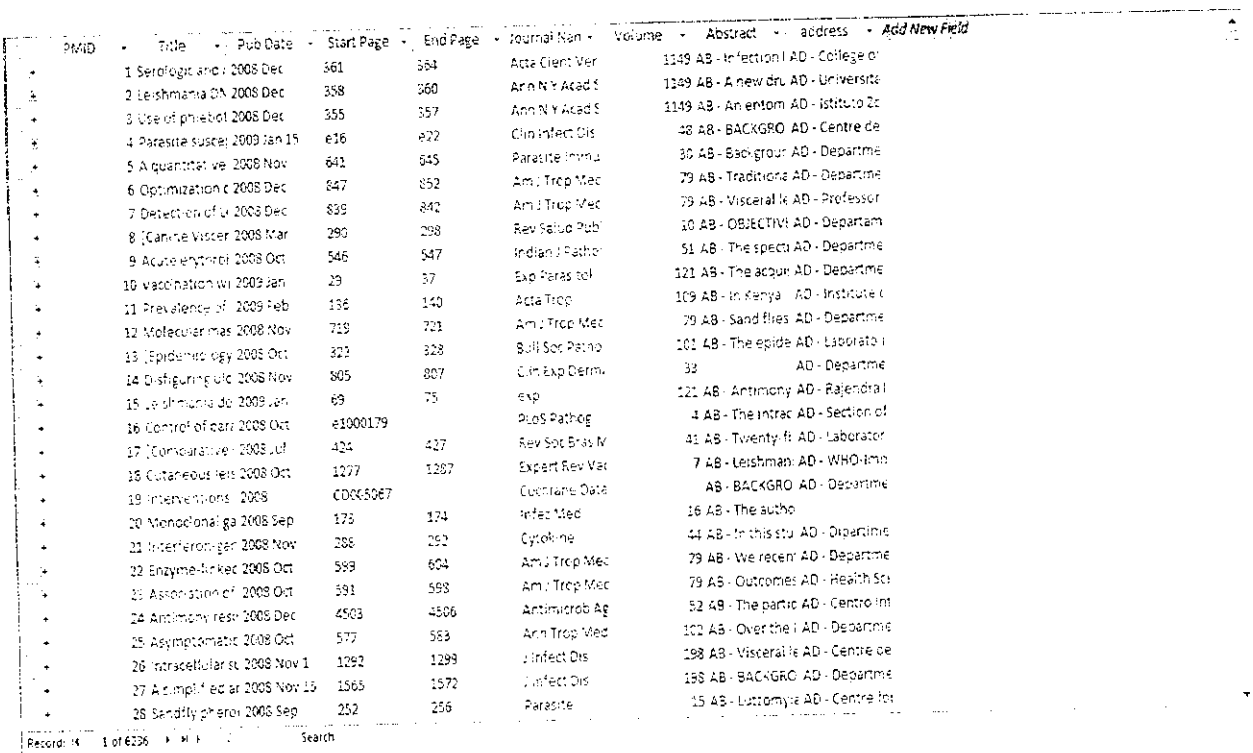
5.3 REMOVAL OF DUPLICATE DATA:

Journal articles relating to Leishmania and Leishmaniasis were collected from various web sites like PubMed, Agricola. The abstracts were carefully checked for duplicates by searching for combination of title character length and year fields. A perl script was used for this purpose.

5.4 CREATION OF DATABASE:

5.4.1 MICROSOFT ACCESS:

Incorporation Of Data Into Database: Parsed Data was loaded into different tables like author, journal name, published date, abstract, keywords and title respectively.



PMID	Title	Pub Date	Start Page	End Page	Journal Nan	Volume	Abstract	address	Add New Field
+	1 Serologic and /	2008 Dec	361	364	Acta Cient Ver	1149	AB - Infection I	AD - College o	
+	2 Leishmania DN	2008 Dec	358	360	Ann N Y Acad S	1149	AB - A new dru	AD - Universita	
+	3 Use of phlebot	2008 Dec	355	357	Ann N Y Acad S	1149	AB - An entom	AD - Istituto Zo	
+	4 Parasite suscep	2009 Jan 15	e16	e22	Clin Infect Dis	48	AB - BACKGRO	AD - Centre de	
+	5 A quantitative	2008 Nov	641	645	Parasite Immu	30	AB - Backgrour	AD - Departme	
+	6 Optimization c	2008 Dec	847	852	Am J Trop Med	79	AB - Traditiona	AD - Departme	
+	7 Detection of Le	2008 Dec	839	842	Am J Trop Med	79	AB - Visceral le	AD - Professor	
+	8 [Canine Visce	2008 Mar	290	298	Rev Salud Publ	10	AB - OBJECTIVI	AD - Departam	
+	9 Acute lymphoi	2008 Oct	546	547	Indian J Pathol	51	AB - The specti	AD - Departme	
+	10 vaccination wi	2009 Jan	29	37	Exp Parasitol	121	AB - The acqui	AD - Departme	
+	11 Prevalence of	2009 Feb	136	140	Acta Trop	109	AB - In Kenya	AD - Institute (	
+	12 Molecular mas	2008 Nov	719	721	Am J Trop Med	79	AB - Sand flies	AD - Departme	
+	13 [Epidemiology	2008 Oct	322	328	Bull Soc Pathol	101	AB - The epide	AD - Laboratoi	
+	14 Disfiguring ulc	2008 Nov	805	807	Clin Exp Derm	33		AD - Departme	
+	15 Leishmania do	2009 Jan	69	75	exp	121	AB - Antimony	AD - Rajendra I	
+	16 Control of para	2008 Oct	e1000179		PLoS Pathog	4	AB - The intrac	AD - Section of	
+	17 [Comparative	2008 Jul	424	427	Rev Soc Bras M	41	AB - Twenty-fi	AD - Laborator	
+	18 Cutaneous leis	2008 Oct	1277	1287	Expert Rev Vac	7	AB - Leishman	AD - WHO-Imn	
+	19 Interventions	2008	CD006067		Cochrane Data		AB - BACKGRO	AD - Departme	
+	20 Monoclonal ga	2008 Sep	178	174	Infect Med	16	AB - The autho		
+	21 Interferon-gan	2008 Nov	288	292	Cytokine	44	AB - In this stu	AD - Departme	
+	22 Enzyme-linked	2008 Oct	599	604	Am J Trop Med	79	AB - We recent	AD - Departme	
+	23 Association of	2008 Oct	591	598	Am J Trop Med	79	AB - Outcomes	AD - Health Sc	
+	24 Antimony resi	2008 Dec	4503	4506	Antimicrob Ag	52	AB - The partic	AD - Centre for	
+	25 Asymptomatic	2008 Oct	577	583	Ann Trop Med	102	AB - Over the l	AD - Departme	
+	26 Intracellular st	2008 Nov 1	1292	1299	J Infect Dis	198	AB - Visceral le	AD - Centre de	
+	27 A simplified ac	2008 Nov 15	1565	1572	J Infect Dis	198	AB - BACKGRO	AD - Departme	
+	28 Sandfly pheron	2008 Sep	252	256	Parasite	15	AB - Buttonyx	AD - Centre for	

Record: 1 of 6236 Search

FIG 5.1 Database Tables Containing Relevant Fields

After the parsed data is loaded in tables, common fields in tables must be related by defining relationship between them.this important to create queries and forms and to search the database.relationship among common fields in table was created as follows:

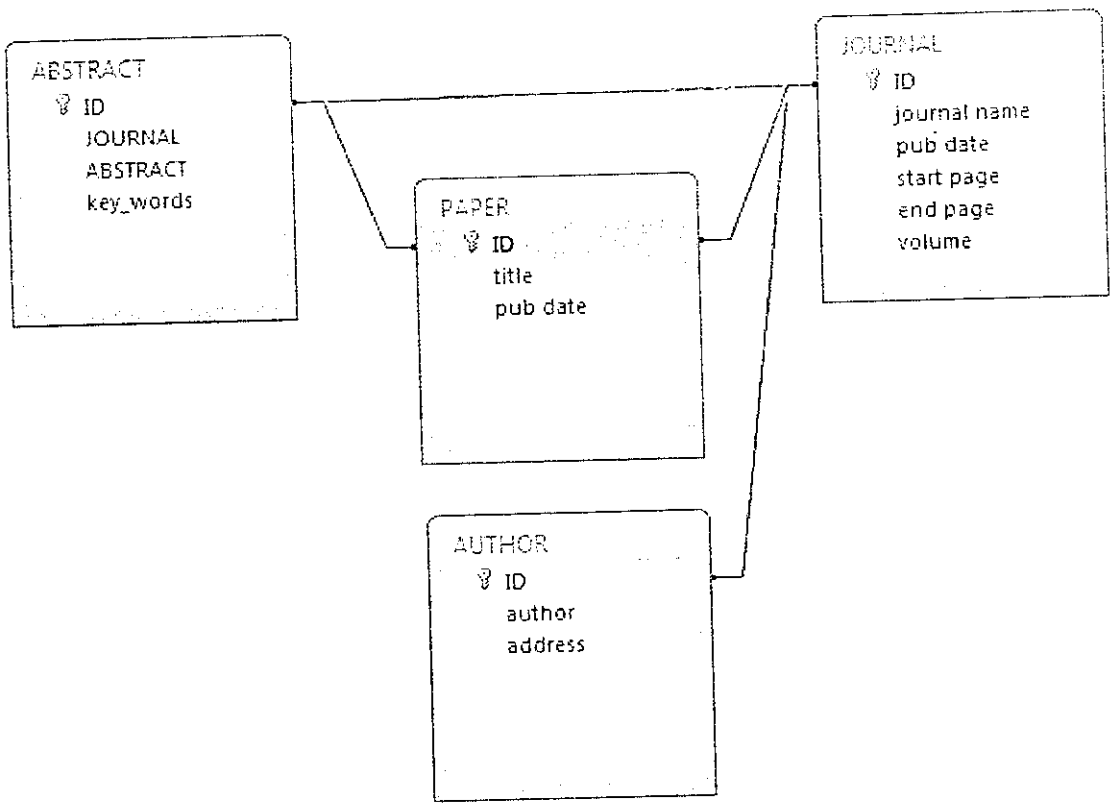


FIG 5.2 Relationship Among Tables

Form is a database object that must be created to enter, edit, or display data from a table or a query.. Search buttons were created in forms to trigger searching the database when the user clicks the button

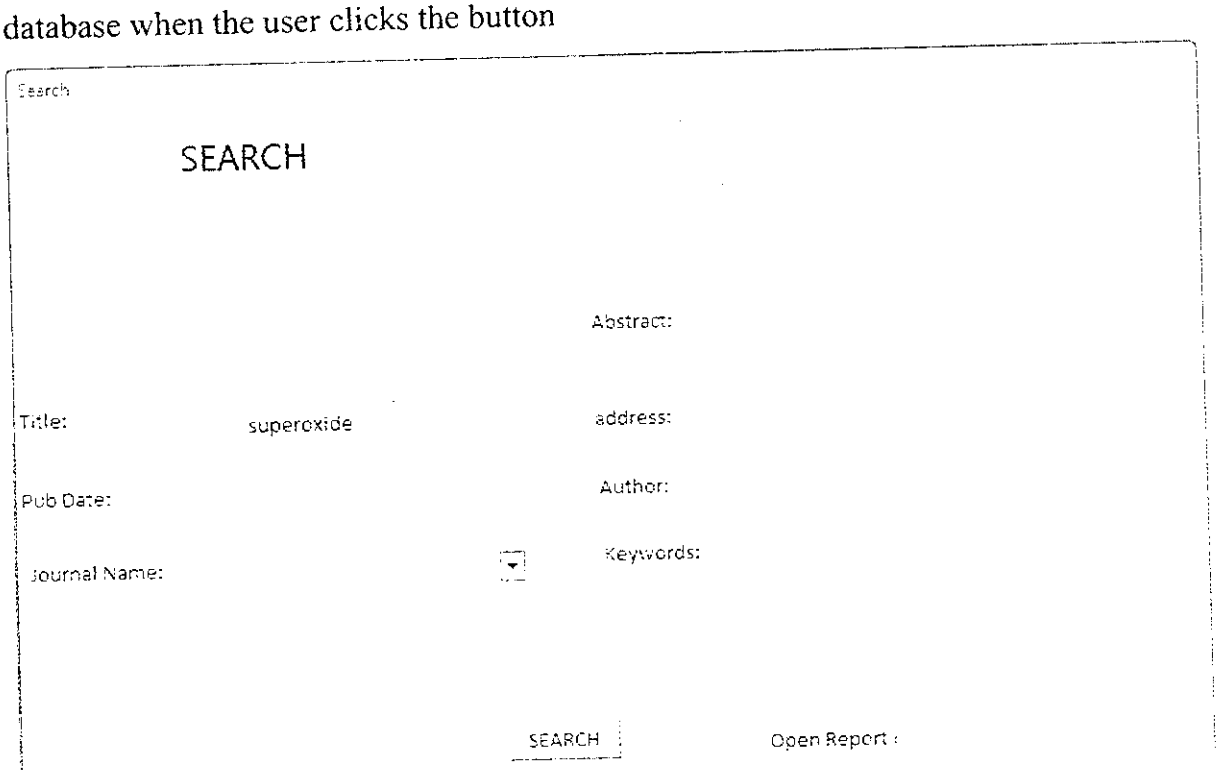
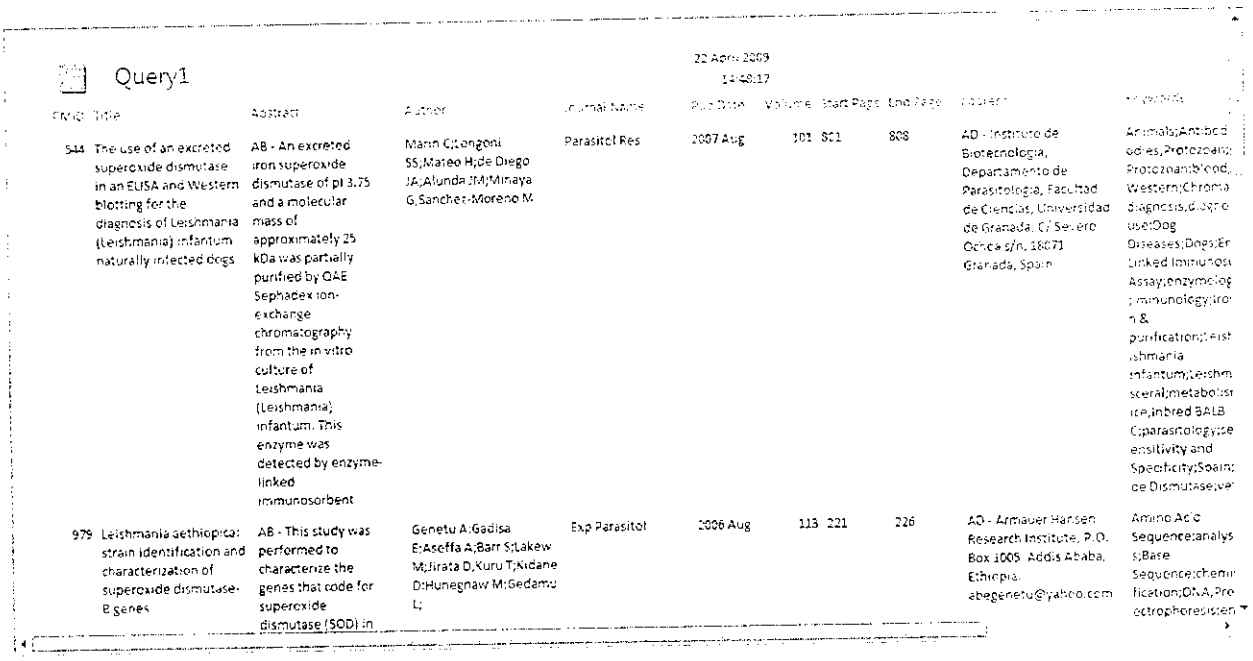


Fig 5.3 Search Form For The Database

Results of the search will be shown in report format as follows:



Query1

22 April 2009 14:48:17

EMID	Title	Abstract	Author	Journal Name	Pub Date	Volume	Start Page	End Page	Address	Keywords
544	The use of an excreted superoxide dismutase in an ELISA and Western blotting for the diagnosis of Leishmania (Leishmania) infantum naturally infected dogs	AB - An excreted iron superoxide dismutase of pI 3.75 and a molecular mass of approximately 25 kDa was partially purified by QAE Sephadex ion-exchange chromatography from the in vitro culture of Leishmania (Leishmania) infantum. This enzyme was detected by enzyme-linked immunosorbent	Marin C;Longoni SS;Mateo H;de Diego JA;Alunda JM;Minaya G;Sanchez-Moreno M	Parasitol Res	2007 Aug	101	821	828	AD - Instituto de Biotecnologia, Departamento de Parasitologia, Facultad de Ciencias, Universidad de Granada, C/ Severo Ochoa s/n, 18071 Granada, Spain	Animals;Antibodies,Protozoan;Protozoan;Blood;Western;Chromatography,Gel;Disease;Dog Diseases;Dogs;Enzyme-Linked Immunosorbent Assay;enzymology;immunology;Iron &... purification;Leishmania infantum;Leishmaniasis,Visceral;metabolism;Male;Mice,Inbred BALB C;parasitology;Sensitivity and Specificity;Spain;Superoxide Dismutase;ver...
979	Leishmania aethiopica: strain identification and characterization of superoxide dismutase-B genes	AB - This study was performed to characterize the genes that code for superoxide dismutase (SOD) in	Genetu A;Gadisa E;Aseffa A;Barr S;Lakew M;Jirata D;Kuru T;Kidane D;Hunegnaw M;Gedamu L;	Exp Parasitol	2008 Aug	119	221	226	AD - Armauer Hansen Research Institute, P.O. Box 1005 Addis Ababa, Ethiopia. abegenetu@yahoo.com	Amino Acid Sequence;analysis;Base Sequence;chemification;DNA;Pre... electrophoresis;Sen...

FIG 5.4 Example Report For Search

Thus a model literature search database was created in access.

5.4.2 Mysql:

WAMP (Windows-Apache-Mysql-Php) tool was set up in the system. Wamp server was started and was run in offline mode.

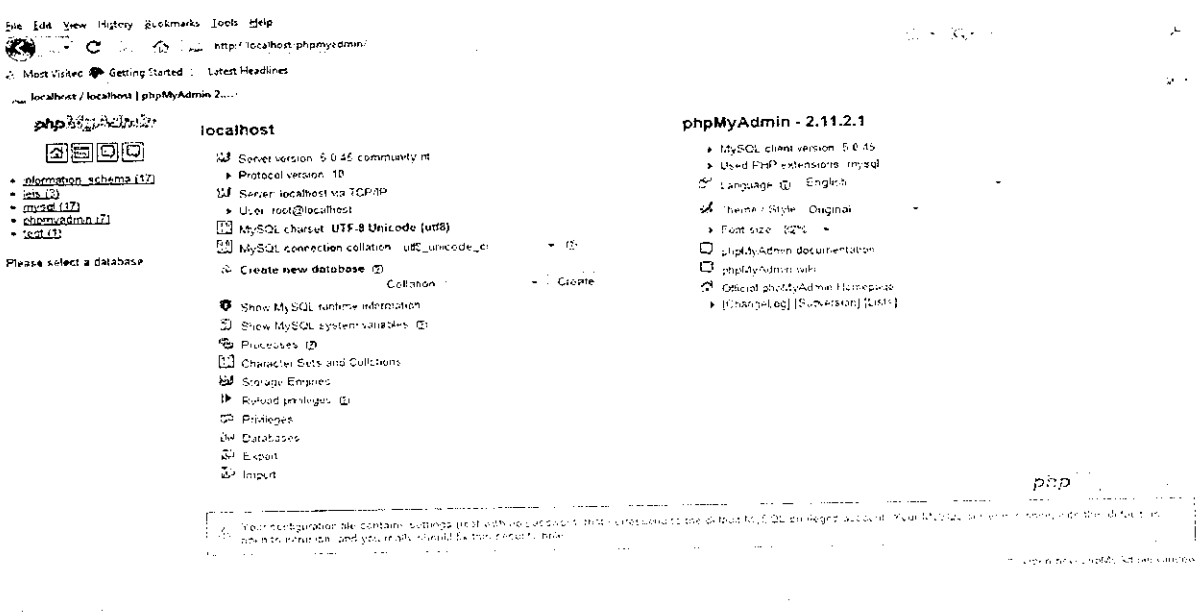


Fig 5.5 Wamp Server

Literature collected were grouped and loaded into mysql database using the Mysql commands.

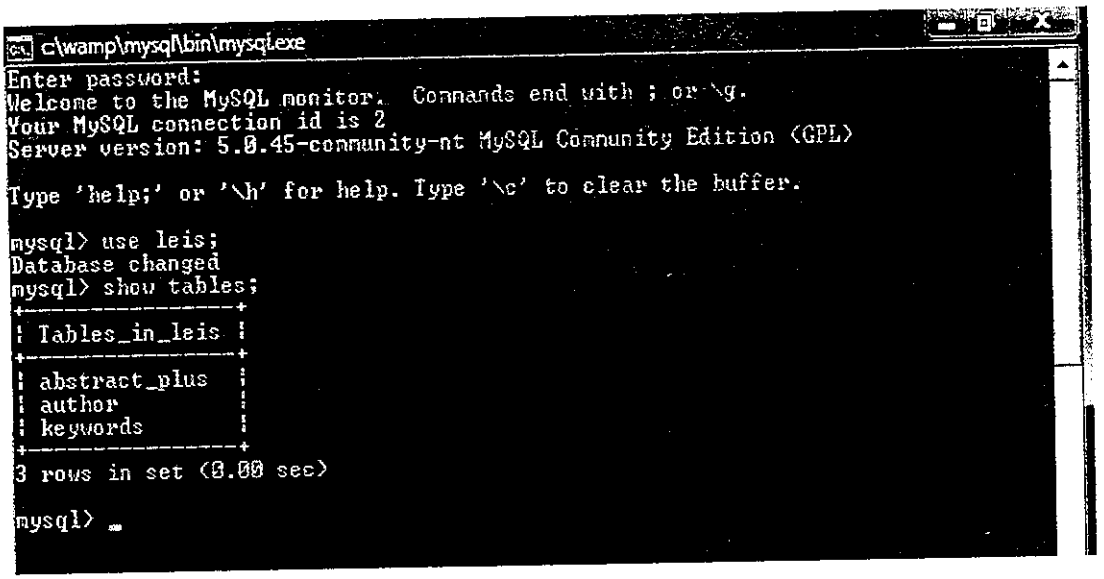


FIG 5.6 Mysql Console Displaying Tables

Php scripts are needed to connect the mysql database to the web.Connection was established between mysql database and Php using wamp server.

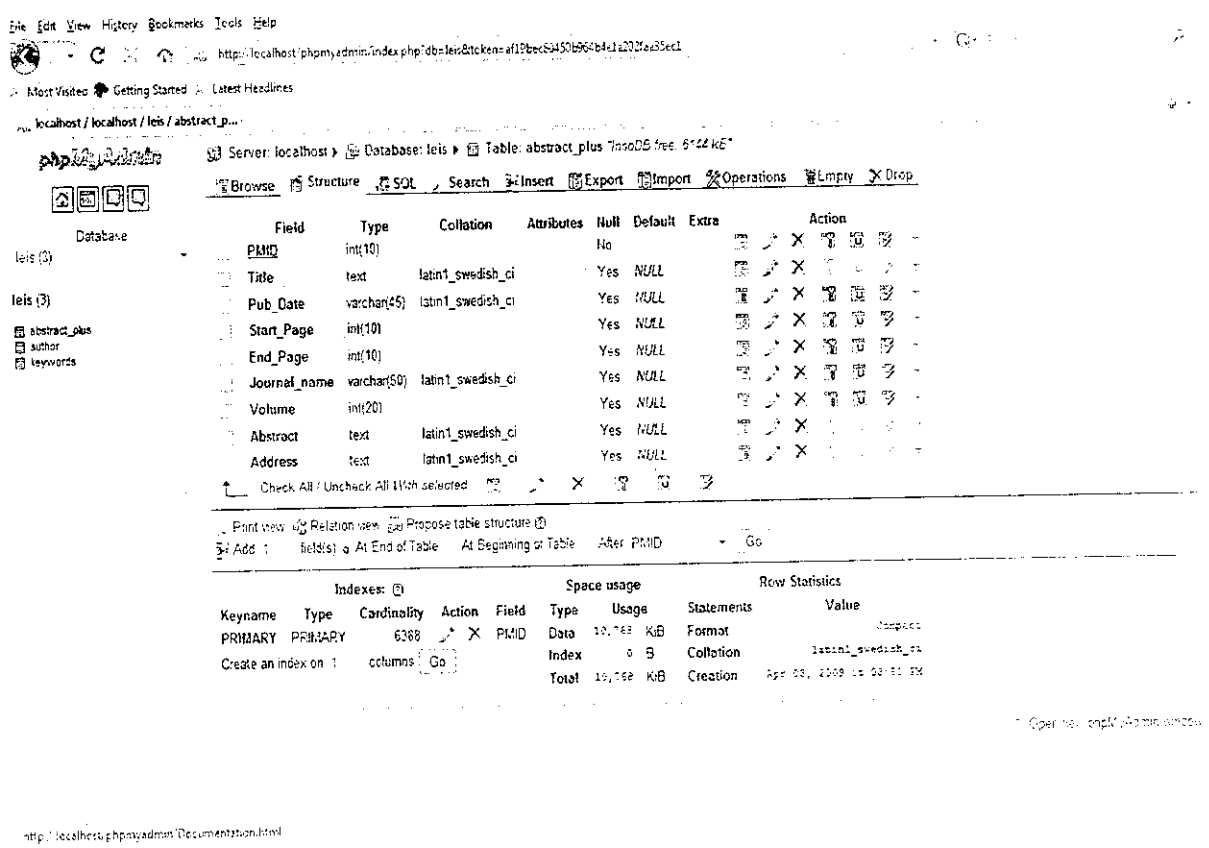


Fig 5.7 Phpmyadmin

5.5 INCORPORATING BLAST TOOL:

Basic Local Alignment Search Tool (BLAST) was downloaded from NCBI and was installed.

5.5.1 FORMATTING BLAST DATABASES:

Local blast databases were created using the formatdb commands.these databases were used to search against the desired sequences.

5.5.2 Implementing blast server:

Network blast server was started and the created databases were used to find hits against the desired sequences.

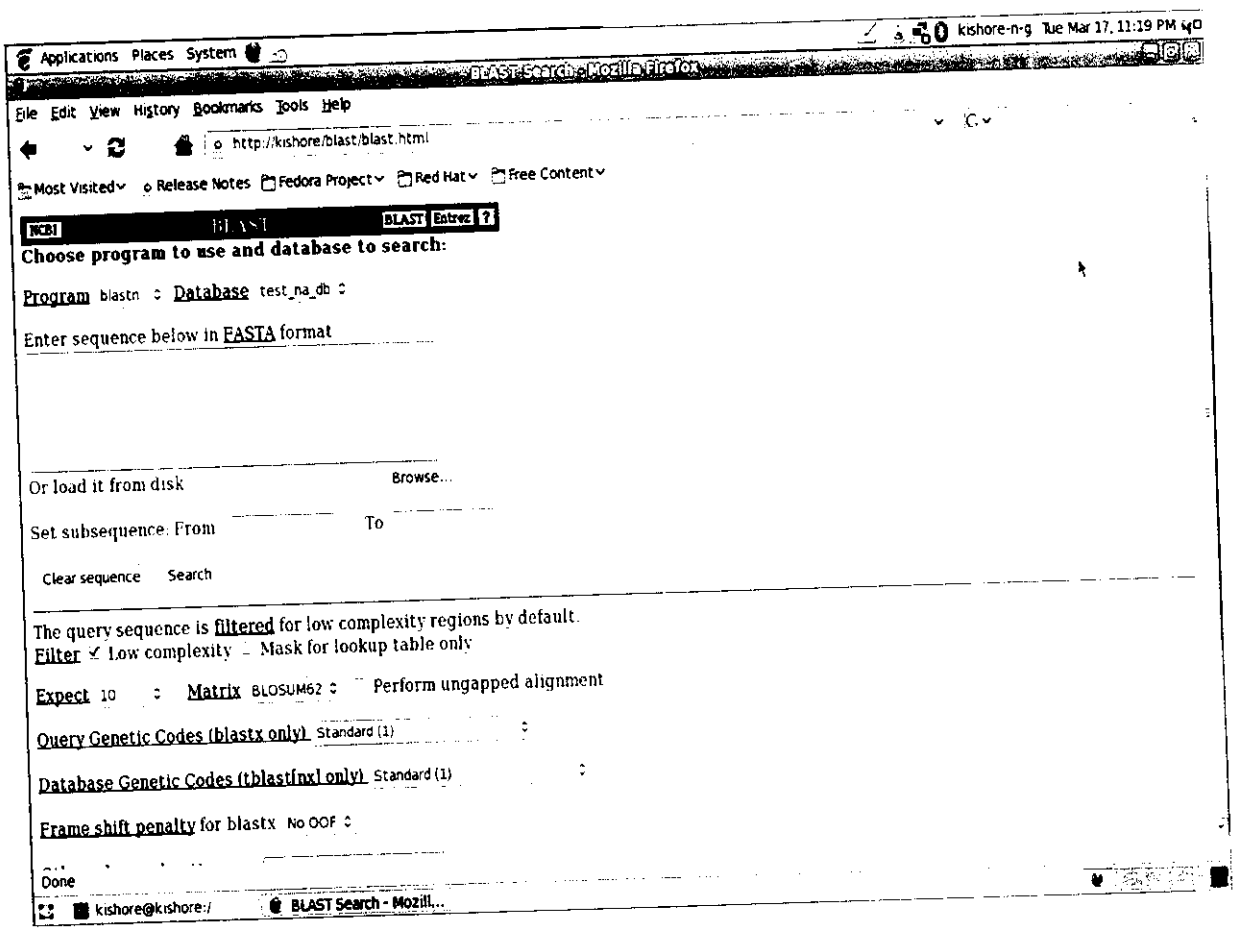


Fig 5.8 Blast Server Running In Local Host With A Sample Database

5.5.3 Blast result:

The results of the blast search were displayed in a weg page as follows,

FIG 5.9 Example Of Blast Result

5.6 DEVELOPMENT OF ONLINE ACCESS WEB FRONT END:

Web page was designed for remote user to access Leishmanial database and it was connected to database via Php scripts.

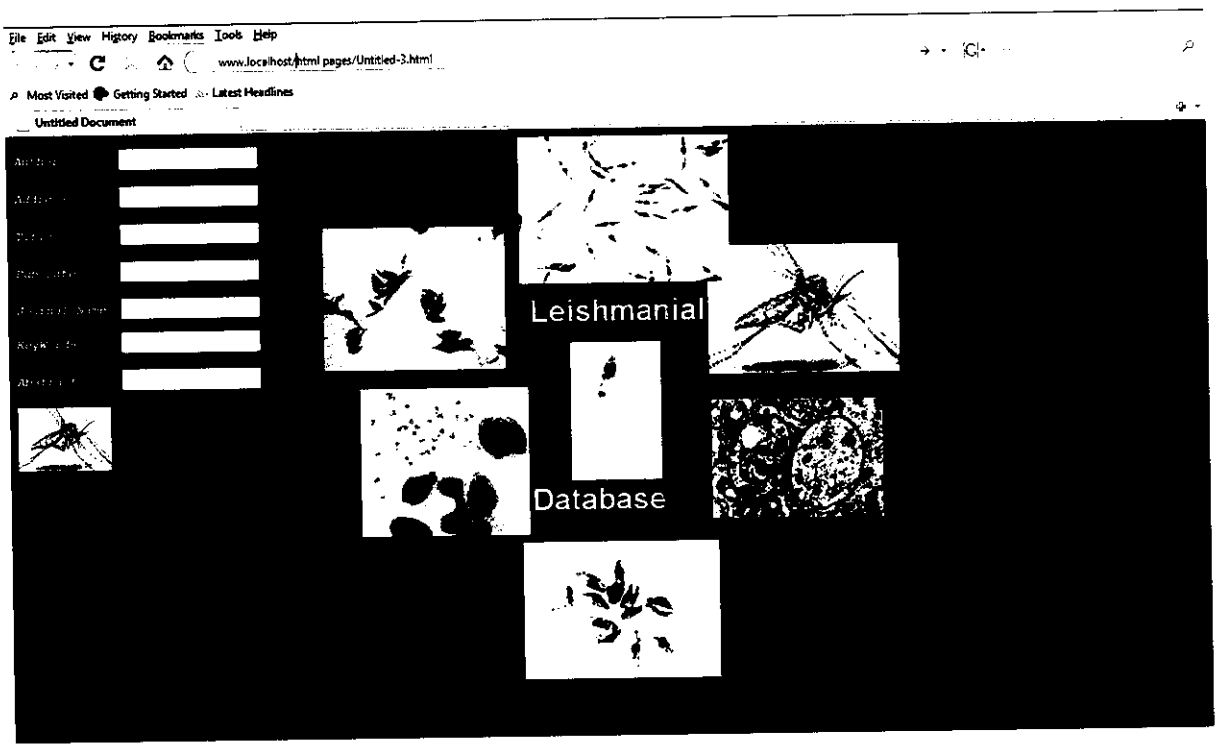


Fig 5.10 Front End For The Leishmanial Literature Database

Summary & Conclusion

SUMMARY AND CONCLUSION:

1. The present study was carried out to address the need for a localized online source for genetic and literature information on *Leishmania* and *Leishmaniasis*. This need arose on account of the increasing incidences of *Leishmaniasis* in the third world countries partly because of drug resistance and due to co-infection during AIDS. Hence accelerated research is needed to address this important disease. The relational model is the most preferred format to store and retrieve biological information. Biological data is voluminous and diverse. The database management systems used for the study included MS Access and the open source MySQL.
2. Literature data from public databases covering various aspects of Leishmaniasis was collected: Pubmed database (14550) and Agricola database (2765).
3. Sequence data related to Leishmania was also collected form NCBI databases: EST (26794), Nucleotide (88033), Protein (56526), and gene sequences (26956).
4. The data was curated for extracting relevant fields and duplicate records were removed.
5. The database was designed in MS Access and the proper relationships were created between the tables.
6. Forms containing relevant search fields and the Reports were designed and the database was tested and fine-tuned.
7. Scale-up of the database was carried out by designing the database in MySQL.
8. The sequence data collected were formatted as per the BLAST program requirements. BLAST server was setup in the local host server and the necessary configuration and database connections was carried out.
9. WAMP environments was created for making the database accessible online. Web-front end was designed for the Leishmanial database via the incorporation of server-end PHP scripts.

A functional and searchable database containing Leishmanial sequences and literature has been developed.

For further development of the database a dedicated web domain has to be created and the database hosted in that domain to enable researchers from across the world to access the data related to Leishmaniasis. Further functionality and collaboration with researchers would lead to expansion of the utility of the database.

References

REFERENCES:

1) Alam, M.Z., Haralambous, C., Kuhls, K., Gouzelou, E., Sgouras, D., Soteriadou, K., Schnur, L., Pratlong, F. and Schonian, G(2009). The paraphyletic composition of Leishmania donovani zymodeme MON-37 revealed by multilocus microsatellite typing. Microbes Infect. PubMed.

2) Andreini, C., Bertini, I., Cavallaro, G., Holliday, G.L. and Thornton, JM.(2009).Metal-MACiE: a database of metals involved in biological catalysis. Bioinformatics.Epub ahead of print ,PubMed .

3) Arnaiz, O., Cain, S., Cohen, J. and Sperling, L. (2007). "ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data.". *Nucleic Acids Res.* 35(Database issue): pp439–44.

4) Bhattacharya, S.K.. , Sinha, P.K ., and Pandey, K..(2005). Pre- & post-treatment evaluation of immunological features in
Indian visceral leishmaniasis (VL) patients with HIV co-infection

5) Chakour, R., Allenbach, C., Desgranges, F., Charmoy, M., Mauel, J., Garcia, I., Launois, P., Louis, J. and Tacchini-Cottier F. (2009)A new function of the Fas-FasL pathway in macrophage activation. J Leukoc Biol. PubMed.

6) Cheng, K.C. and Strömvik, M.V.(2008) SoyXpress: a database for exploring the soybean transcriptome. BMC Genomics.PubMed.

7) Colpitts, S.L., Dalton, N.M. and Scott P.(2009) IL-7 receptor expression provides the potential for long-term survival of both CD62L(high) central memory T cells and Th1 effector cells during Leishmania major infection. J Immunol.pp5702-11.PubMed.

8) Date, C. J.(2003).An Introduction to Database Systems, Eighth Edition, Addison Wesley.

9) Demeler, B., Brookes, E., Nagel-Steger, L.(2009) Analysis of heterogeneity in molecular weight and shape by analytical ultracentrifugation using parallel distributed computing. Methods Enzymol. pp87-113. PubMed.

10) Dr.Thakur, P., Diagnosis & management of leishmania/HIV co-infection. Indian J Med Res 121, pp 407-414.

11) Dursun, O., Erisir, S. and Yesilipek, A.(2009). Visceral childhood leishmaniasis in southern Turkey: experience of twenty years. Turk J Pediatr. pp1-5.PubMed .

12) E. F. CODD,(1970)." A Relational Model of Data for Large Shared Data Banks". Communications of the ACM,pp377-387

13) Francois Chappuis, Shyam Sundar, Asrat Hailu, Hashim Ghalib, Suman Rijal,Rosanna W. Peeling, Jorge Alvar and Marleen Boelaert (2007). Visceral leishmaniasis:what are the needs for diagnosis,treatment and control? IEEE Trans Inf Technol Biomed,pp314-345.

14) Galindo, J., Urrutia, A. And Piattini, M. Fuzzy(2006) Databases: Modeling, Design and Implementation (FSQL guide). Idea Group Publishing Hershey, USA.

15) Gil-Redondo, R., Estrada, J., Morreale, A., Herranz, F., Sancho J. and Ortiz, A.R.(2009). Virtual screening data management on an integrated platform. J Comput Aided Mol,pp171-84. PubMed.

16) González,U., Pinart M., Rengifo-Pardo, M., Macaya, A., Alvar J. and Tweed, J.A.(2009). Interventions for American cutaneous and mucocutaneous leishmaniasis. Cochrane Database Syst Rev. PubMed.

17) Hariharan, M., Scaria, V. and Brahmachari S.K.(2009) dbSMR: a novel resource of genome-wide "SNPs affecting microRNA mediated regulation. BMC Bioinformatics.Epub ahead of print, PubMed.

18) Hidalgo, C.A., Blumm, N., Barabási, A.L. and Christakis, N.A.(2009) A dynamic network approach for the study of human phenotypes. PLoS Comput Biol. PubMed.

19) Javier Nieto ,Israel Cruz, Javier Moreno , Carmen Cañavate, Philippe Desjeux and Jorge Alvar(2006).*Leishmania*/HIV co-infections in the second decade.Indian J Med Res 123, pp 357-388.

20) Jeffery, K.M., Maggio, L., Blanchard, M.(2009) Making generic tutorials content specific:recycling evidence-based practice (EBP) tutorials for two disciplines. Med Ref Serv Q. pp1-9. PubMed.

21) Kroenke and David, M., (1997).Database Processing: Fundamentals, Design, and Implementation .Prentice-Hall, Inc., pp 130-144.

22) Lu T., Huang, X., Zhu, C., Huang, T., Zhao, Q., Xie, K., Xiong, L., Zhang, Q. and Han B.(2008) RICD-a rice indica cDNA database resource for rice functional genomics. BMC Plant Biol. pp100-118. PubMed

23) Madeira da Silva L., Owens, K.L., Murta, S.M. and Beverley S.M.(2009) Regulated expression of the Leishmania major surface virulence factor lipophosphoglycan using conditionally destabilized fusion proteins. Proc Natl Acad Sci U S A.PubMed.

24) Mahmoud, M.A., Al-Khafaji, J.T., Al-Shorbaji, N., Sara, K., Al-Ubaydli, M. and Ghazzaoui, R., Liu, F. and Fontelo, P. (2008) BabelMeSH and PICO Linguist in Arabic. AMIA Annu Symp Proc. pp916:944. PubMed.

25) Mandal G., Sarkar A., Saha P., Singh N., Sundar S, Chatterjee (2009)M. Functionality of drug efflux pumps in antimonial resistant Leishmania donovani field isolates.Indian J Biochem Biophys. pp86-92. PubMed.

26) Mandal, A., Asthana, A.K., Aggarwal, L.M.(2008) Development of an electronic radiation oncology patient information management system. J Cancer Res Ther, pp178-85. PubMed.

27) Mark Maslakowski(2000)Sam's teach yourself MySQL. ISBN: 0672319144

28) Mark Whitehorn and Date C. J. (2003). An Introduction to Database Systems. Eighth Edition, Addison Wesley.

29) Mary Ann Richardson (2008). Run a parameter query in access. www.tech republic.com

30) McDonald, A.G., Boyce, S. and Tipton, K.F.(2009) ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res. PubMed.

31) Meng, S., Brown, D.E., Ebbole, D.J., Torto-Alalibo, T., Oh YY, Deng, J., Mitchell, T.K. and Dean, R.A.(2009) Gene Ontology annotation of the rice blast fungus, Magnaporthe oryzae. BMC Microbiol. Suppl 1:S8. Review. PubMed.

32) Milon, G.(2009) Perpetuation of Leishmania: some novel insight into elegant developmental programs. Vet Res.PubMed.

33) Mo F., Hong, X., Gao, F., Du L., Wang, J., Omenn, G.S. and Lin, B.(2008). A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. BMC Bioinformatics. pp169-537. PubMed.

34) Nancy Malla and Mahajan, R.C.,(2006). Pathophysiology of visceral leishmaniasis - some recent concepts. Indian J Med Res 123, pp 267-274.

35) Owens J.(2009) Organizing, exploring, and analyzing antibody sequence data: the case for relational-database managers. Methods Mol Biol. pp525:569-80,PubMed.

36) Ozensoy Töz S., Sakru, N., Ertabaklar, H., Demir, S. And Sengul M.(2009) Serological and entomological survey of zoonotic visceral leishmaniasis in Denizli Province,Turkey. New Microbiol.pp93-100. PubMed.

37) Pretz, J.E. and Link, J.A.(2008) The Creative task Creator: a tool for the generation of customized,Web-based creativity tasks. Behav Res Methods.pp1129-33. PubMed.

38) Robinson, T.J., Duvall, S. and Wiggins R.(2008). Creation and Usability Testing of a Web-Based Pre-Scanning Radiology Patient Safety and History Questionnaire Set. J Digit Imaging. PubMed.

39) Sarman singh (2006). New developments in diagnosis of Leishmaniasis. Indian J Med Res 123, pp 311-330.

40) Shah, N. and Musen, M.(2008) UMLS-Query: A Perl Module for Querying the UMLS. AMIA Annu Symp Proc. pp652-6. PubMed.

41) Sheng C., Ji H., Miao, Z., Che X., Yao J., Wang W., Dong G., Guo W.and Zhang W.(2009)Homology modeling and molecular dynamics simulation of N-myristoyltransferase from protozoan parasites: active site characterization and insights into rational inhibitor design. J Comput Aided Mol Des. PubMed.

42) Singh, R.K.., Pandey, H.P., and Sundar, S., (2006).Visceral leishmaniasis (kala-azar): Challenges ahead. Indian J Med Res 123, pp 331-344.

43) Sinha, P.K., Sanjiva Bimal, Singh, S.K., Krishna Pandey, Gangopadhyay, D.N. and Bhattacharya, S.K.(2006).Pre- & post-treatment evaluation of immunological features in Indian visceral leishmaniasis (VL) patients with HIV co-infection. Indian J Med Res 123, pp 197-202.

44) Souza W., Attias M. and Rodrigues J.C. (2009)Particularities of Mitochondrial Structure in Parasitic Protozoa (Apicomplexa and Kinetoplastida). Int J Biochem Cell Biol. PubMed.

45) Spath, G.F., Schlesinger, P., Schreiber, R. and Beverley S.M.(2009) A novel role for Stat1 in phagosome acidification and natural host resistance to intracellular infection by Leishmania major. PLoS Pathog.PubMed.

46) Stefanov, S., Lautenberger, J. and Gold, B.(2008) An Analysis Pipeline for Genome-wide Association Studies. Cancer Inform. pp455-61. Epub PubMed.

47) Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. (2002). "The generic genome browser: a building block for a model organism system database." *Genome Res.*pp 1599–610.

48) Stephen F. Altschul, Warren Gish, Webb Miller,Eugene W. Myers and David J. Lipmanl.(1990)." Basic Local Alignment Search Tool" J.mol.biol,pp403-410.

49) Suryapriya, P., Snehalatha, A., Kayalvili, U., Krishna, R., Singh, S. and Ulaganathan, K.(2009).Genome-wide analyses of rice root development QTLs and development of an online resource, Rootbrowse. Bioinformation. pp279-81. PubMed.

50) Syed-Mohamad SM.(2009) Development and implementation of a web-based system to study children with malnutrition. Comput Methods Programs Biomed.pp83-92.PubMed.

51) Ul Bari, A. and Ejaz, A.(2009) Rhinophymous leishmaniasis: A new variant. Dermatol Online J. PubMed.

52) William M. Gelbart, and Joe Nadeau,(1998).Report of the NIH model organism Database Workshop. Lansdowne Conference Center . Virginia.

53) Wong I.L., Chan K.F., Zhao Y., Chan T.H. and Chow LM. (2009)Quinacrine and a novel apigenin dimer can synergistically increase the pentamidine susceptibility of the protozoan parasite Leishmania. J Antimicrob Chemother. PubMed

54) Zhou, T. and Caflisch, A.(2009) Data management system for distributed virtual screening. J Chem Inf Model pp145-52.PubMed